Tutorial

Identification of somatic variants in a matched tumor-normal pair

December 20, 2024

Sample to Insight

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@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 individual.

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] (https://onlinelibrary.wiley.com/doi/10.1155/2013/270362).

The example data used in this tutorial include only reads mapping to chromosome 22. The sequencing reads have already been trimmed for Illumina adapter sequence.

Prerequisites

- Installed *CLC Genomics Workbench* 24.0 on a machine meeting the system requirements: https://digitalinsights.qiagen.com/technical-support/system-requirements/.
- *Biomedical Genomics Analysis* plugin 24.0 (if you are using other versions you can expect slightly other views and results).

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 website: https://resources.qiagenbioinformatics.com/testdata/Example_data_tumor_normal.zip.
2. Start the 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** and click **Next**.
5. Choose where to save the example data in the **Navigation Area** and click **Finish**.

The data set includes the following files:

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.

normal reads
: Sequencing reads from the normal sample.

tumor reads
: Sequencing reads from the tumor sample

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)** template workflow. Among the generated output is a Genome Browser view showing the identified somatic variants together with the human reference sequence, the human genes, and variants found in the public database ClinVar.

1. To run the **Identify Somatic Variants from Tumor Normal Pair (WES)** template workflow:

   Template Workflows | **Biomedical 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.

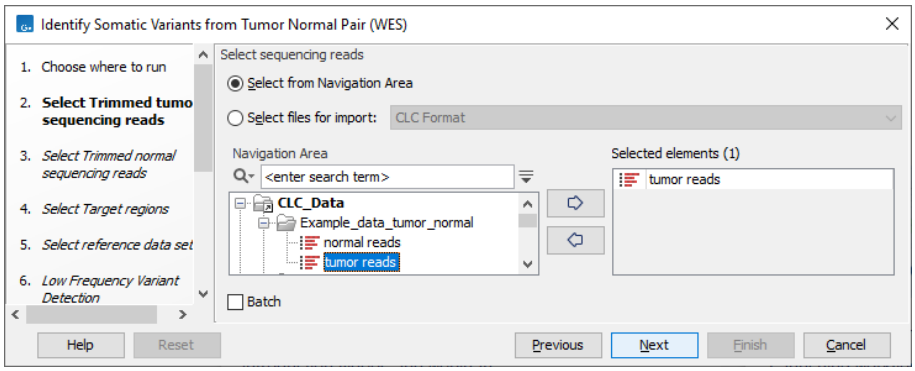

Figure 1: *Select the reads from the tumor sample.*

2. Select the tumor reads (figure 1).

3. In the next step, select the reads from the normal sample (figure 2). Click on **Next**.

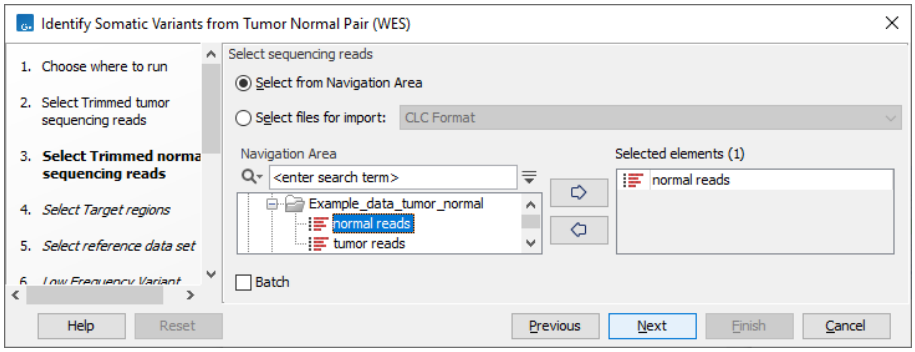

Figure 2: *Select the reads from the normal/control sample.*

4. Select the **target regions - chr22** file (figure 3). Click **Next**.

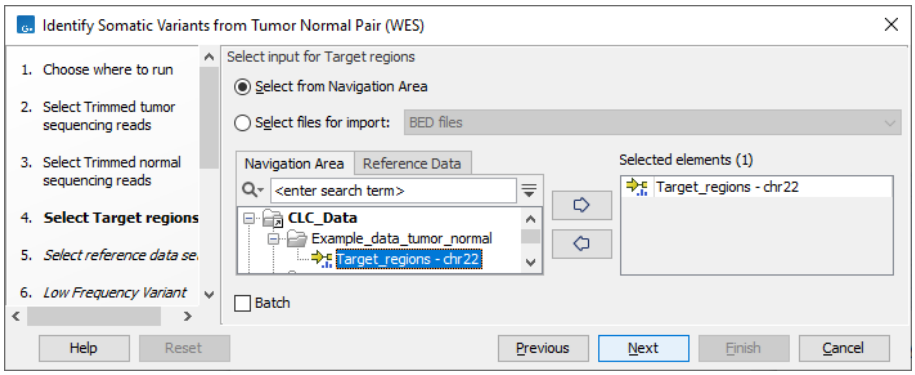

Figure 3: *Select the target file.*

5. For demonstration purposes, we have chosen to run the analysis with only chr22 of the human reference sequence (hg19). Typically, we would recommend to run the analysis on the complete human genome, and not only a part of it. First select the Tutorial Reference Data Set called **Identification of Somatic Variants in a Matched Tumor-Normal Pair** (figure 4). If you have not downloaded this data set before, click on the button labeled **Download to Workbench**. Note that you can choose where (Server or Workbench) you want to download the reference data if you are working connected to a Server. When the download is completed, click **Next**.

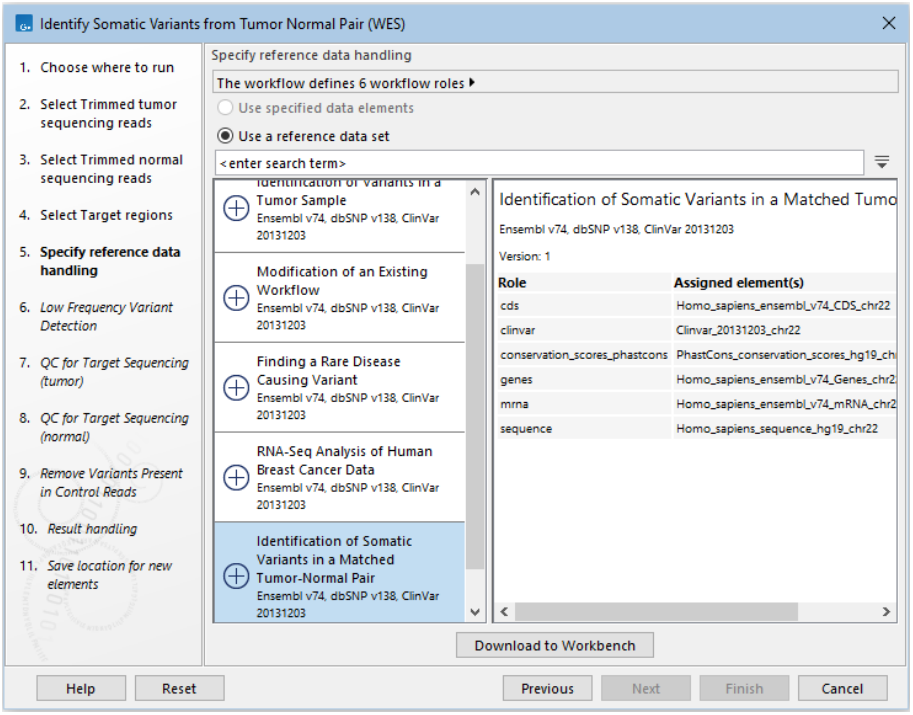

Figure 4: *Download the relevant tutorial reference databases.*

6. In the next dialog change the **Minimum frequency** to 15.0 for the **Low Frequency Variant Detection** tool (see figure 5).

7. The next 2 dialogs are quality checks for Target Sequencing of the tumor and the normal samples. Adjust the minimum coverage to 20 for both (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 variants present in control reads by leaving (*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 template workflow. Please note that some of the parameters are locked and cannot be 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 **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.

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 8.

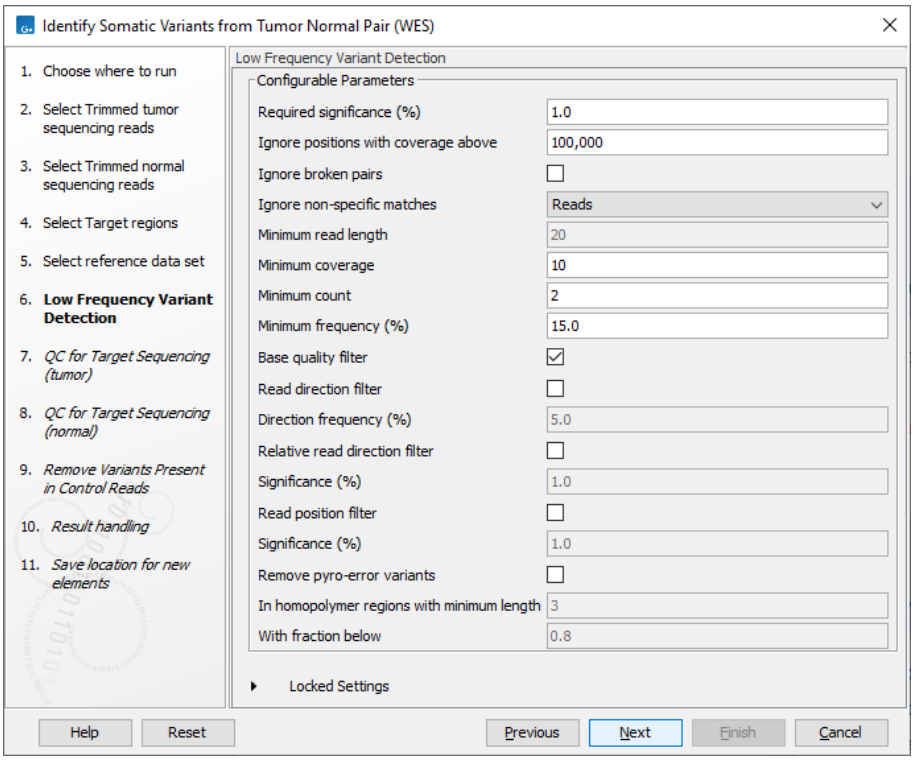

Figure 5: *Adjustment of minimum frequency*

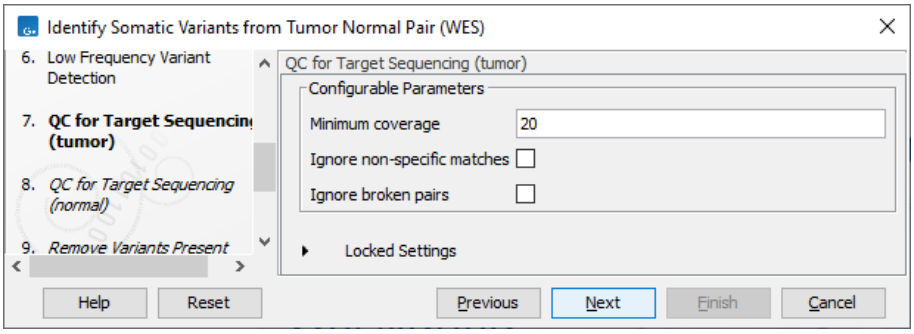

Figure 6: *Adjust the settings for quality check of the tumor sample.*

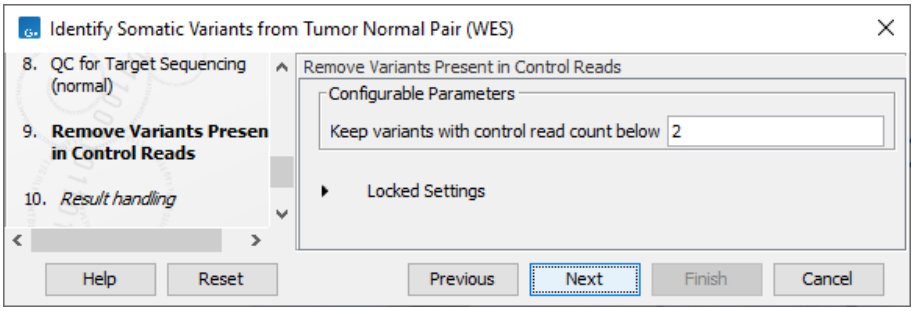

Figure 7: *In this wizard step you can adjust the cutoff for removal of germline variants.*

Inspection of potential somatic variants

Double click on the file *Genome Browser View Tumor Normal Comparison* in the **Navigation Area** generated with the template workflow. The file opens in a split view as a track list and variant table below. 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 Track

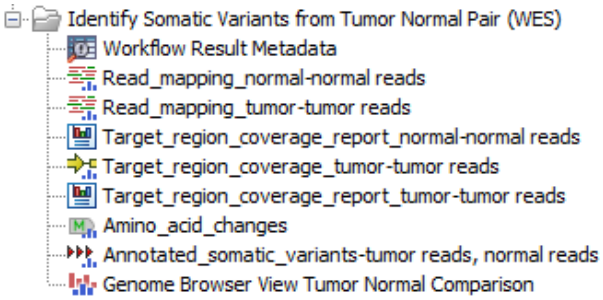

Figure 8: *The output from the workflow. Please note that in this example we have created a folder with the name "Identify Somatic Variants from Tumor Normal Pair (WES)".*

List 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 Track List. You can also use the zoom function in the Track List to zoom in and out on specific variants.

Type *CYP2D6* into the filter field of the Table (highligted in red in figure 9)

- 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_20131203". 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 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.

Additional tips

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

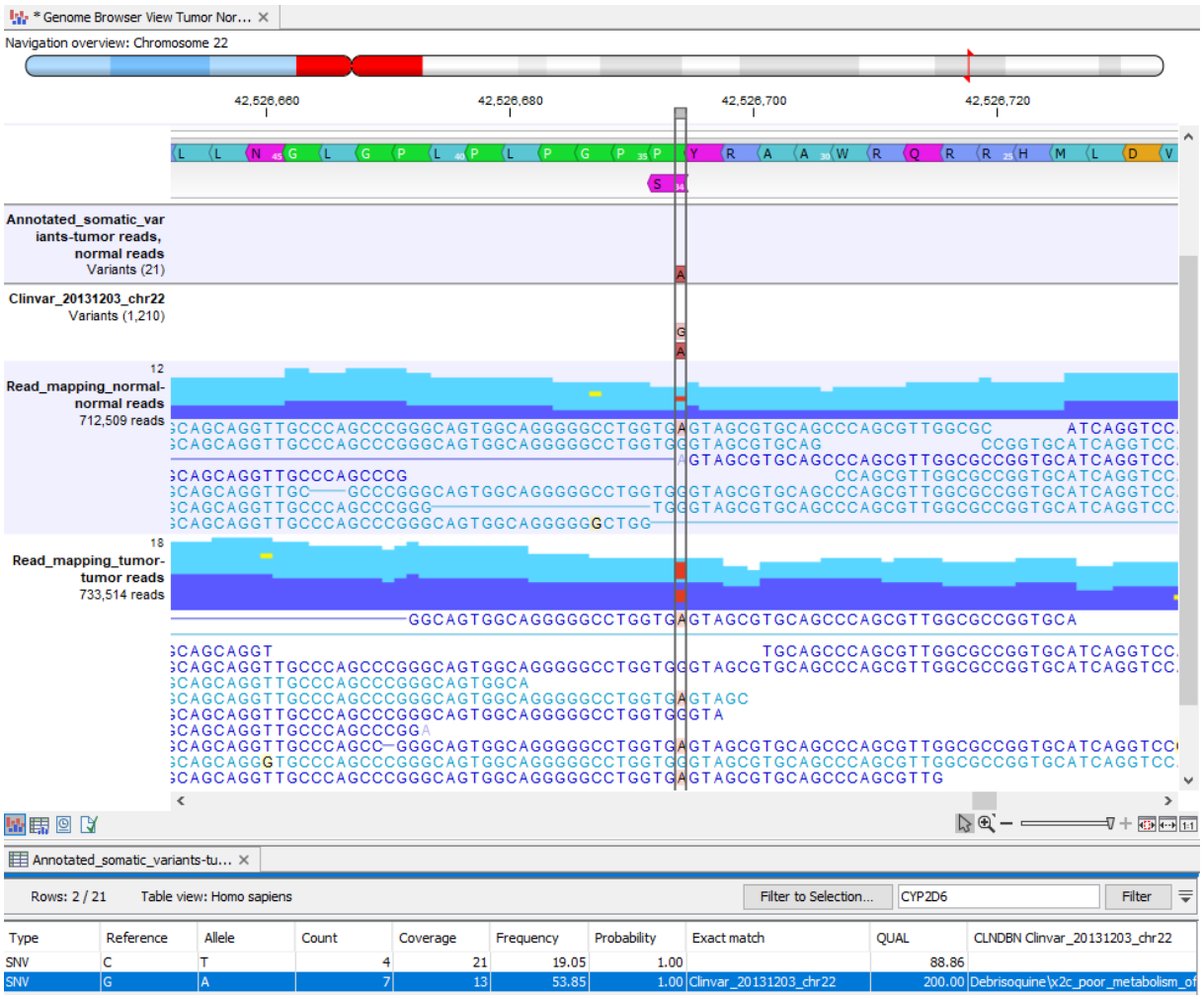

Figure 9: *Inspect variants and read mappings in the Genome Browser view. The table and the track view are linked so that if you select a row in the table, the track view above will jump to the position of the variant in the selected row. You can open additional tables by double-clicking on the name of a track in the left side of the track list, for example Clinvar_20131203_chr22. All tables opened from the track view will also be linked to the track. In the tables, the filter field allows easy filtering of the table (highlighted in red), the different columns in the table allows you to inspect information added to the variants (highlighted in red).*

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.

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.