QIAGEN

Tutorial

Identification of Variants in a Tumor Sample

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 Variants in a Tumor Sample

This tutorial will guide you through the process of identifying variants and verifying them.

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 (http://www.hindawi.com/crim/oncological.medicine/2013/270362/).

The example data used in this tutorial include only reads mapping to a short fraction of chromosome 5. The reads have already been trimmed for Illumina adapter sequences.

Prerequisites For this tutorial, you must be working with the Biomedical Genomics Workbench 2.5 or higher. Minimum recommended machine specifications for working with human data sets are listed at http://www.qiagenbioinformatics.com/system-requirements/, but in this tutorial we are working with a reduced dataset and a standard desktop computer/laptop with 4 GB RAM will be sufficient. Note that this tutorial includes images, workflows, and tools from Biomedical Genomics Workbench 3.0

Overview The analyses carried out in this tutorial include:

- Mapping reads to a reference sequence
- Local realignment
- Detecting variants
- Mapping quality check
- How to check the identified variants for potential false positives

Importing the data and the references

First, we need to download and import the data.

1. Download the sample data from our web site: http://resources.qiagenbioinformatics.com/testdata/Example_data_tumor_25.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_25.zip*. Leave the Import type set to **Automatic import**.
5. Save the imported data.

The data set includes the following files:

tumor_reads_chr5
Illumina sequencing reads from the tumor sample

target_regions_chr5
Targeted regions from the exome enrichment (in our case, coding regions for a small fraction of chromosome 5)

Data configuration For demonstration purposes, we have chosen to run the analysis with only chr5 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.

Biomedical Genomics Workbench provides the necessary whole genome reference data that can be downloaded and configured using the **Data Management** button found at the top right corner of the Biomedical Genomics Workbench. The reference data just have to be download once, and this is typically done when the Biomedical Genomics Workbench is started for the first time:

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 **Identification of Variants in a Tumor Sample** Reference Data Set.
3. If you had not downloaded this data set before, click on the button labeled **Download**. You can check the progress of the download in the 'Processes' tile of the toolbox.
4. When the Reference Data Set is downloaded, click on the button labeled **Apply** and then click on the button labeled **Close**.

Note that we have added extra steps at the end of the tutorial that will ensure you are setting the reference data back to its original state after you are done with the tutorial. Remember to do these steps to make sure your next analyses will be executed against whole genomes versions of the data.

Variant identification

The first step in the analysis is the mapping of sequencing reads from the tumor sample to chr5, which is directly followed by the detection of indels. The detected indels serves as a guidance-variant track for the next step, the local realignment that is done to improve the mapping and enable a better detection of variants. After the variant detection step, potential false positives are filtered away based on average base quality. The outputs from the analysis are a read mapping, a quality report for the target regions and a Genome Browser View.

The quality report for the targeted regions should be checked to identify poorly covered regions, and to check the specificity of reads to the targeted regions. These could be indications that the enrichment was not successful or that the primers/oligos were not specific.

All these steps can be facilitated using the **Identify Variants (WES)** ready-to-use workflow.

1. Start the workflow with:

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

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

2. Select the sequencing reads. In our case, choose **tumor_reads_chr5** (figure 1). Click **Next**.

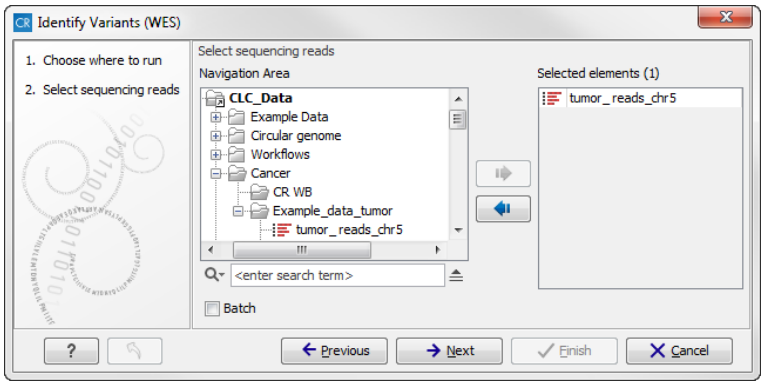


Figure 1: *Select the sequencing reads.*

3. In the Indels and structural variants window, select the target region. This can be done by clicking on the folder icon ( ) in the right side next to "Restrict calling to target regions" and selecting **target_regions_chr5** (figure 2). Click **Next**.

   Note: When running a targeted sequencing workflow, please ensure that you obtain the correct target regions from the vendor of your target enrichment kit.

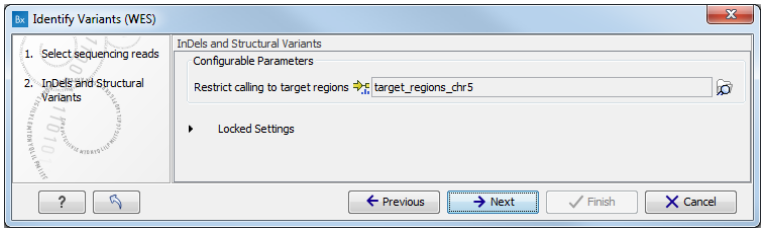


Figure 2: *Select the target regions.*

4. In the next step you can specify the parameters used for calling variants using the low frequency variant caller (figure 3). Again you must specify the target regions track (*target_regions_chr5*) by clicking on the folder icon ( ) in the right side next to "Target Regions". Check that the value for **Minimum frequency** is set to 5%. Click **Next**.

5. In the next window, you have to specify once more the targeted regions used for the enrichment experiment, **target_regions_chr5**.

   Additionally, change the **Minimum coverage** setting to the value 10. Figure 4 shows how this step should look, with the correctly set parameters. When you have selected the target regions track and adjusted the settings, click **Next**.

6. Choose to **Save** the outputs of the workflow. Click on the button labeled **Next** to specify the location where the outputs should be saved. We suggest you to create a folder called "Analyzed data", and click **Finish**.

After you have started the job(s), you can follow the progress in the **Processes** tab, which you will find in the **Toolbox** in the lower-left corner of the Workbench.

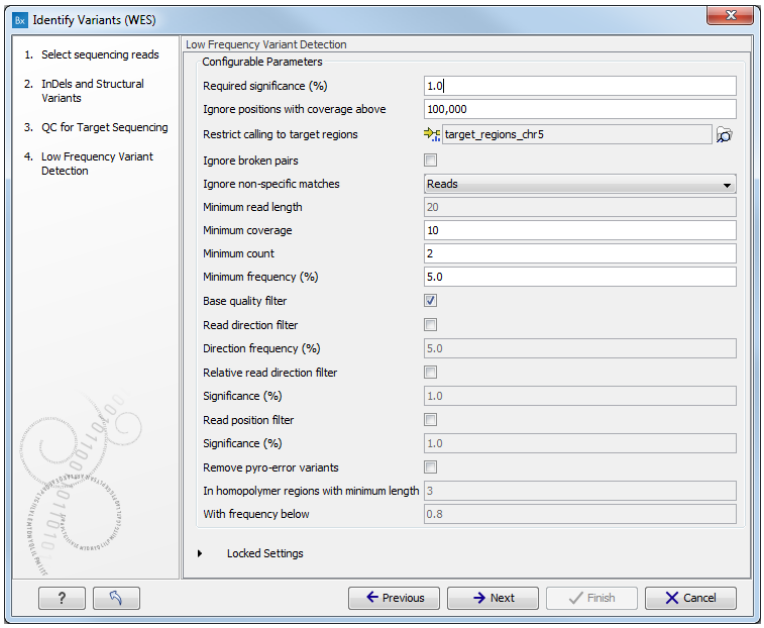


Figure 3: *The correct parameter settings in the "Low Frequencing Variant Detection" step of the wizard.*

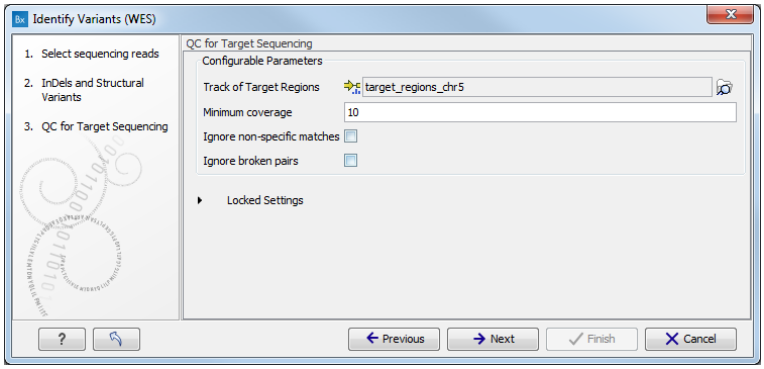


Figure 4: *The correct parameter settings in the 'QC for Target Sequencing' step of the wizard.*

The results will be placed in the location you specified when the job has finished.

The following results will be generated (see figure 5):

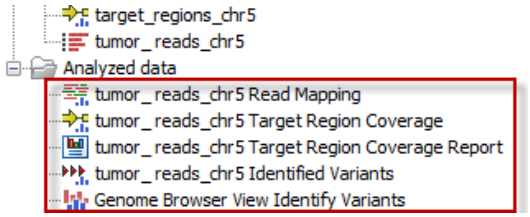


Figure 5: *The analysis outputs.*

- () *tumor_reads_chr5 Read Mapping*: Mapped reads from the tumor sample to chromosome 5 of the human reference genome hg19
- () *tumor_reads_chr5 Target Region Coverage*: The targeted regions with information about the minimum, maximum and average coverage for each region

- () *tumor_reads_chr5 Target Regions Coverage Report)*: Quality report for the mapping to the targeted regions
- () *tumor_reads_chr5 Identified Variants*: Identified variants after filtering out variants with an average low base quality.
- () *Genome Browser View Identify Variants*: Genome Browser View, which enables the direct comparison and validation of identified variants in the context of the mapped sequencing reads and targeted regions.

Checking the QC report for the targeted regions

The quality report for targeted regions should be checked to find out if the enrichment of the target regions was successful.

We would like to answer the following questions:

- Is the average coverage in the target regions sufficient?
- Are all specific targets sufficiently covered?
- Is the specificity of the reads mapping to the target regions in the expected range (e.g. above 50% for exome sequencing and above 90% for targeted amplicon sequencing)?

To answer these questions, open *tumor_reads_chr5 Target Region Coverage Report*.

Is the average coverage in the target regions sufficient?

Have a look at the **Summary** table in the report, where you will find the average coverage of reads in all target regions. This value should be minimum 10 for targeted data, as the minimum threshold for the variant caller was set to 10. For amplicon data, we expect it to be larger than 100.

See Figure 6 for the average coverage of the target regions in our example.

1.1 Summary

Number target regions	124
Total length of target regions	22,946
Average coverage	23.8
Number of target regions with coverage below 10	70
Total length of regions of targets with coverage below 10	12,660

Figure 6: *Average coverage of targeted regions*

We can see that the value is above the value of 10 that we need as minimum to facilitate an accurate variant calling.

Is the specificity of the reads mapping to the target regions within the expected range?

Before we proceed we should check the enrichment kit from the vendor. Normally, this just needs to be checked once, when a gene or exome panel is used for the first time.

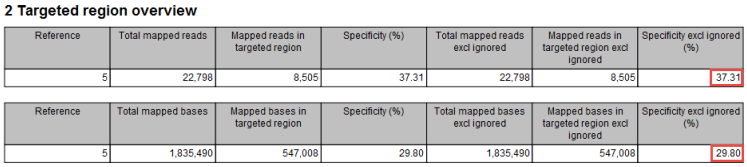
For a hybridization/array approach (most exome kits) we should have a minimum of 50% of the reads mapped specifically to the targeted region. For amplicon data we expect to have a minimum of 90% of reads on target.

Please have a look at the **Targeted Region Overview** section in the report.

In this section, you will find the total number of reads mapping to the target regions as well as the percentage that map to the targeted regions.

In figure 7 you can see that in our example, 37.3% of reads and 29.8% of bases map to the target regions. These numbers are quite low for a hybridization enrichment approach, which was used here.

In this tutorial, we are only considering a small fraction of the total target regions, so these numbers are not very accurate. If we looked at all targets and all reads, the values would be substantially higher.



2 Targeted region overview

Reference	Total mapped reads	Mapped reads in targeted region	Specificity (%)	Total mapped reads excl ignored	Mapped reads in targeted region excl ignored	Specificity excl ignored (%)
5	22,798	8,505	37.31	22,798	8,505	37.31

Reference	Total mapped bases	Mapped bases in targeted region	Specificity (%)	Total mapped bases excl ignored	Mapped bases in targeted region excl ignored	Specificity excl ignored (%)
5	1,835,490	547,008	29.80	1,835,490	547,008	29.80

Figure 7: *Specificity of the reads mapping to the targeted regions*

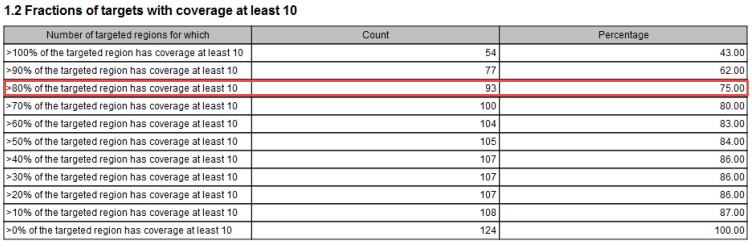
Are all targets sufficiently covered?

This is one of the most important questions when it comes to diagnostics, where you have to make sure that important regions are 100% covered and have coverage above a certain value (in most cases, this value is 30x). If the coverage is less than this, the enrichment and the sequencing have to be redone, or missing regions have to be sequenced using, for example, Sanger sequencing.

This question is also very important for research analyses, for example if you are interested in a particular region and wish to do comparisons between samples. Here, you should make sure that such a region is well covered in all samples.

We wish to check how many targets have more than 10x coverage in at least 80% of the total region of the target. To do this, go to the section of the report called **1.2 Fractions of targets with coverage at least 10**, and look at the value in the table **>80% of the targeted region has coverage at least 10**.

As you can see, the value is in the range of what acceptable and what would be expected for a hybridization experiment (see Figure 8).



1.2 Fractions of targets with coverage at least 10

Number of targeted regions for which	Count	Percentage
>100% of the targeted region has coverage at least 10	54	43.00
>90% of the targeted region has coverage at least 10	77	62.00
>80% of the targeted region has coverage at least 10	93	75.00
>70% of the targeted region has coverage at least 10	100	80.00
>60% of the targeted region has coverage at least 10	104	83.00
>50% of the targeted region has coverage at least 10	105	84.00
>40% of the targeted region has coverage at least 10	107	86.00
>30% of the targeted region has coverage at least 10	107	86.00
>20% of the targeted region has coverage at least 10	107	86.00
>10% of the targeted region has coverage at least 10	108	87.00
>0% of the targeted region has coverage at least 10	124	100.00

Figure 8: *73% of all targets are more than 80% covered with at least 10 reads.*

Is a particular target well covered with reads?

Let us now pretend that we are particularly interested in gene CCNB1. We would like to check if all the target regions of this gene are covered with at least 10 reads. To do this we will go to the output data found in the **Navigation Area**.

1. Open the data item called *Genome Browser View Identify Variants* (). Double-click on the file name in the **Navigation Area** to open it in the **View Area**. The opened file is split in two, with a track list containing the relevant input data and outputs from the workflow that was analyzed (shown in figure 9), as well as a table listing all found variants below.

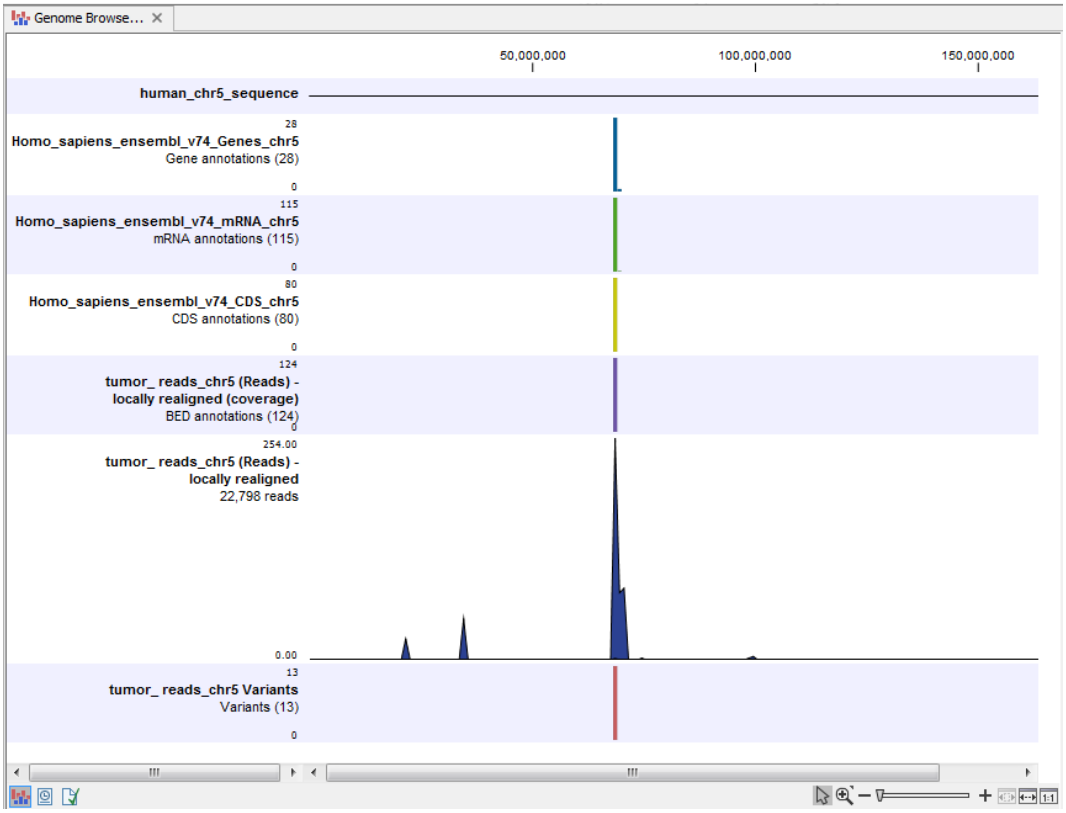


Figure 9: *The Genome Browser View shows the different tracks in one view and makes it easy to inspect the identified varients and correlated them to e.g. the reference sequence, genes, or other tracks.*

2. In the left side of the opened Genome Browser View, double-click on the name of the track *tumor_reads_chr5 Target Region Coverage*. This will open the table view of this track, with all target regions and information about coverage specifically for each region. You can deselect the column "Name" in the right hand side panel to make it easier to see the columns of the table you are interested in.

3. Filter the table entries to show only the targeted regions for gene CCNB1 by entering the text **CCNB1** in the search field (figure 10).

4. Have a look at the **Percentage with coverage above 10** column.

Here, all coding regions for the gene CCNB1 have at least 95% of the region covered with 10 or more reads. Eight of the target regions are full covered in their entirety (100%) (see figure 11).

In conclusion we can say that our particular target (the CCNB1 gene) is sufficiently covered.

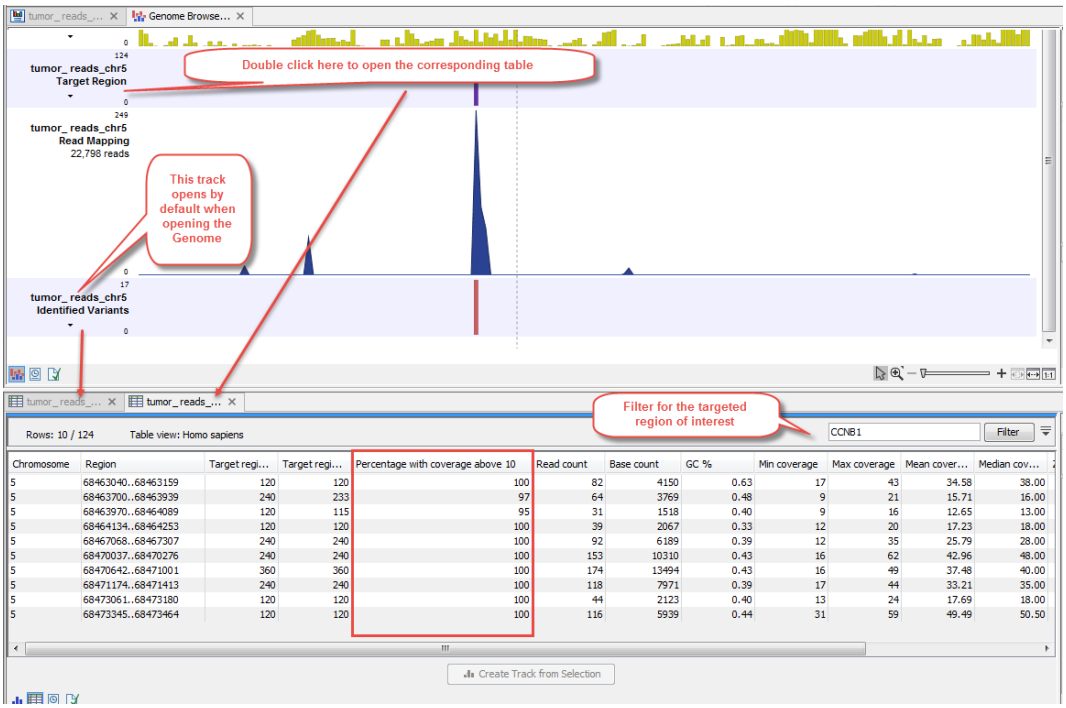


Figure 10: *From the Genome Browser View the "tumor_reads_chr5 Target Region Coverage" track has been opened in table view by clicking on the name of the variant track in the left side of the Genome Browser View. The filter has been used to show only CCNB1 entries.*

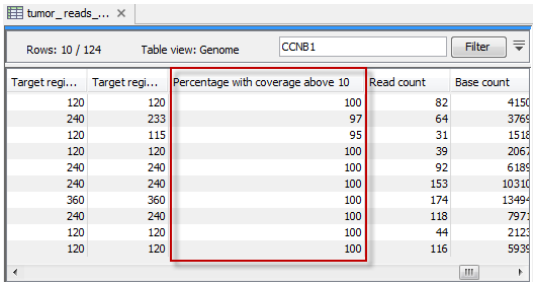


Figure 11: *In the column "Percentage with coverage above 10" you can see that eight of the ten target regions are covered in their entirety (100%).*

Check the identified variants for potential false positives

It is very likely that you will end up with a huge number of variants being reported for a tumor sample. There are several reasons for this.

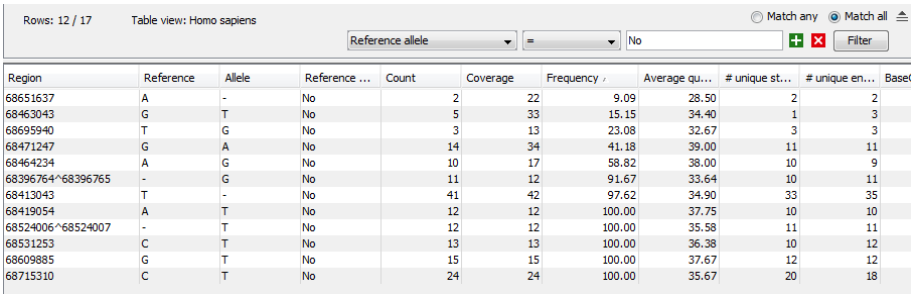
Firstly, in many cancers the DNA machinery does not work well, which leads to many variants. Moreover, genome rearrangements and aneuploidy are very common events in cancers. Many tumors also include many different cells with different mutation patterns. To be able to detect variants occurring only in a small fraction of cells (which can play an important role in tumor relapse), variants have to be detected at a very low frequency in the data. It is often a challenging task to distinguish these variants from sequencing errors.

In this section you will learn how variants can be filtered to get the best candidate variants for further analysis.

1. Go back to the variant table *tumor_reads_chr5 Identified Variants* that opened together with

the Genome Browser View earlier (see figure 12). You can see that 17 variants have been identified in this region.

2. Use the advanced filter option at the top of the table to filter for **Reference allele** contains **No**. You can do this by clicking the down arrow next to the **Filter** button, choosing **Reference allele** and **Contains** in the drop-down menus, typing "No" into the text field, and clicking **Filter**. This will leave above 12 variants that are different from the human reference sequence.



Rows: 12 / 17 Table view: Homo sapiens Match any Match all Reference allele = No Filter

Region	Reference	Allele	Reference ...	Count	Coverage	Frequency	Average qu...	# unique st...	# unique en...	Base
68651637	A	-	No	2	22	9.09	28.50	2	2	
68463043	G	T	No	5	33	15.15	34.40	1	3	
68695940	T	G	No	3	13	23.08	32.67	3	3	
68471247	G	A	No	14	34	41.18	39.00	11	11	
68464234	A	G	No	10	17	58.82	38.00	10	9	
68396764^68396765	-	G	No	11	12	91.67	33.64	10	11	
68413043	T	-	No	41	42	97.62	34.90	33	35	
68419054	A	T	No	12	12	100.00	37.75	10	10	
68524006^68524007	-	T	No	12	12	100.00	35.58	11	11	
68531253	C	T	No	13	13	100.00	36.38	10	12	
68609885	G	T	No	15	15	100.00	37.67	12	12	
68715310	C	T	No	24	24	100.00	35.67	20	18	

Figure 12: *Use the filter function to identify variants that differs from the human reference sequence. Note that we have changed here which columns were selected or not to fit the table display to the tutorial narrative.*

3. Look at the **Frequency** column. The frequency of most variants is very high, but for some variant it is less than 25%. If you look at the number of reads supporting it (look in the **Count** column), you can see that there are less than 10 reads, meaning that these variants are not supported by a lot of reads.

In general, if you would like to validate your variant results, you should take note of the following:

The average base quality for the variant A low average base quality (below 20) could suggest that this is a sequencing error.

The number of unique reads that support the variant Please check the value in the columns: **# unique start positions** and **# unique end positions**. These values should be greater than one. If they are not, the variant could be due to a PCR error during enrichment.

The regions surrounding the variant In the track list, look at the regions surrounding the variant in the reference sequence. Is it in a homopolymer region (e.g. in a stretch of As)? Is it a deletion or an insertion? If so, then the variant may well be a sequencing error.

The number of reads supporting the variant This value should be minimum 1, but preferably 5 or more.

Other things to note

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. That history includes information about how the data was created, what parameter settings were used, what 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.

This is a good way of double-checking what source data and parameters you have used for the analyses 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**.