Tutorial

Identification of Variants in a Tumor Sample

January 19, 2024

Sample to Insight

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 have installed the Biomedical Genomics Analysis plugin, preferable version 24.0. Expect slightly different views and results, if you are using other versions. Minimum recommended machine specifications for working with human data sets are listed at https://digitalinsights.qiagen.com/technical-support/system-requirements/, but in this tutorial we are working with a reduced data set and a standard desktop computer/laptop with 4 GB RAM will be sufficient.

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 website: http://resources.qiagenbioinformatics.com/testdata/Example_data_tumor_25.zip.

2. Open the *CLC Genomics Workbench* and go to:

   File | **Import (⬇)** | **Standard Import (⬇)**

3. Choose the zip file called *Example_data_tumor_25.zip*. Leave the Import type set to **Automatic import**.

4. **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)

Variant identification

The first step in the analysis is the mapping of sequencing reads from the tumor sample to chr5, which is 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, variant tracks and finally a Genome Browser view showing all tracks together in one 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)** template workflow.

1. Start the workflow with:

   Toolbox | Template Workflows | Biomedical 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 click **Next**.

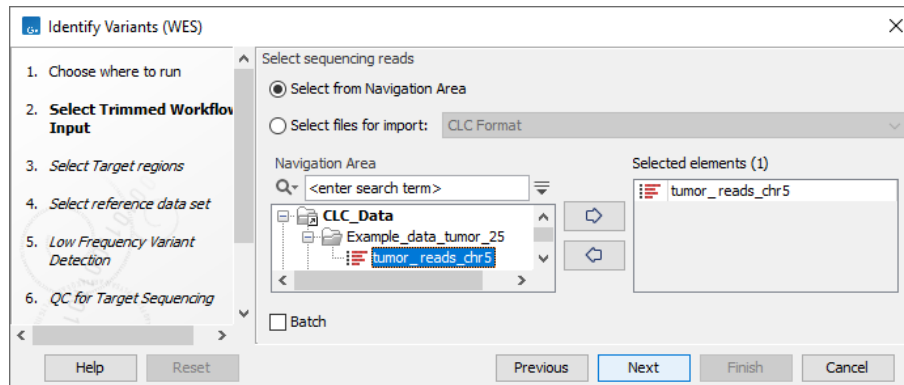2. Select the sequencing reads. In our case, choose **tumor_reads_chr5** (figure 1). Click **Next**.

Figure 1: *Select the sequencing reads.*

3. In the next dialog you select the target region you imported earlier. In our case select **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.

4. In the next dialog, select the relevant reference data set needed to run this tutorial. 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. Select the **Identification of Variants in a Tumor Sample** Reference Data Set. If you had not downloaded this data set before, click on the button labeled **Download** (figure 3). If you are connected to a Server, you will be given the choice of where the reference data should be downloaded. When the download is completed, click **Next**.
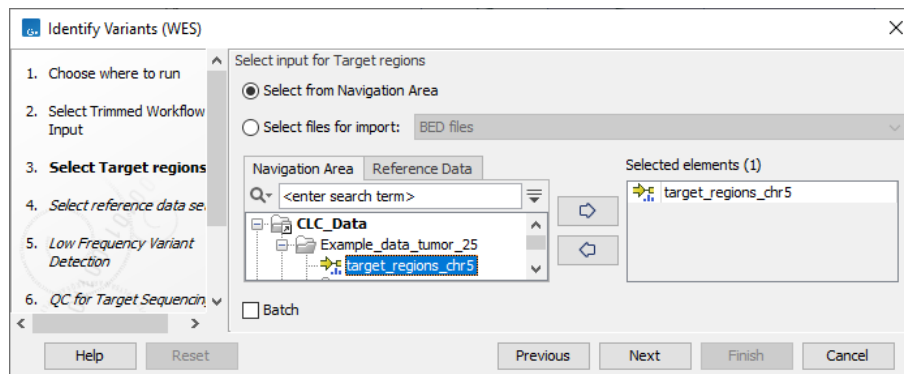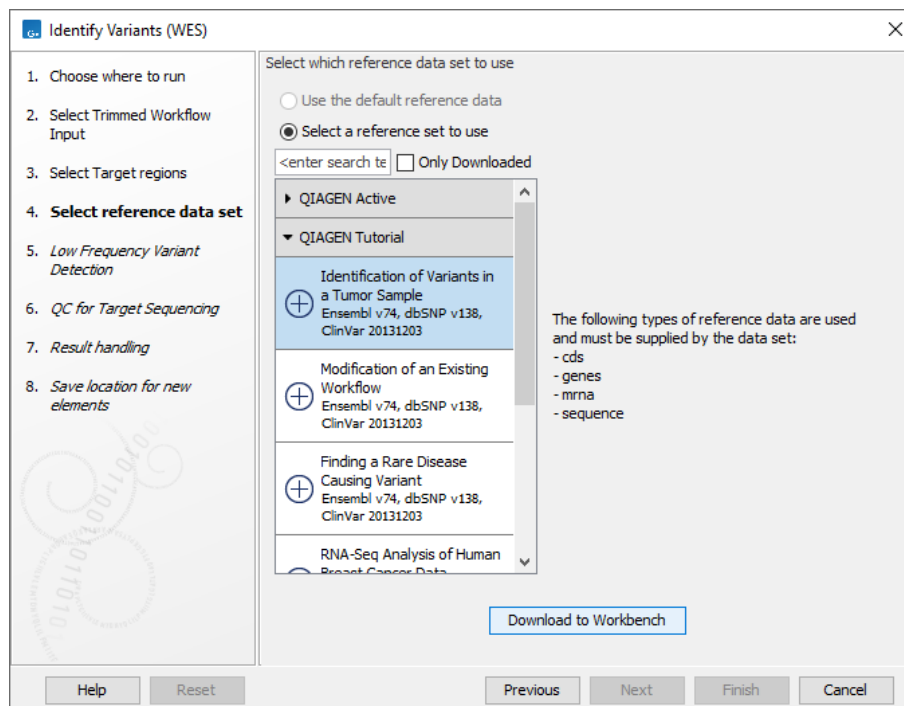
Figure 2: *Select the target regions.*

Figure 3: *Select the reference data set from the Tutorial section of the QIAGEN references.*

5. In the next step you can specify the parameters used for calling variants using the Low Frequency Variant Detection tool (figure 4). Check that the value for **Minimum frequency** is set to 5.0%. Click **Next**.

6. In the QC for Target Sequencing dialog, change the **Minimum coverage** setting to the value 10. See figure 5. Click **Next**.

7. Choose to **Save** the outputs of the workflow. Click **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, you can follow the progress in the **Processes** tab, which you will find in the **Toolbox** in the lower-left corner of the Workbench. The results will be placed in the location you specified when the job has finished.

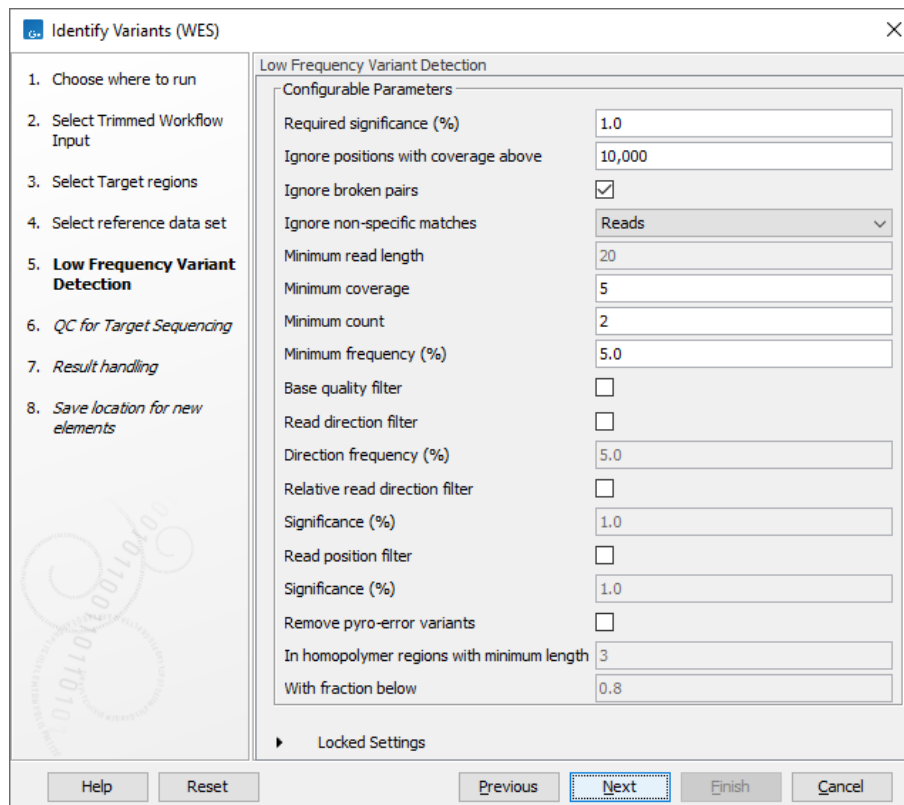The following results will be generated (see figure 6):

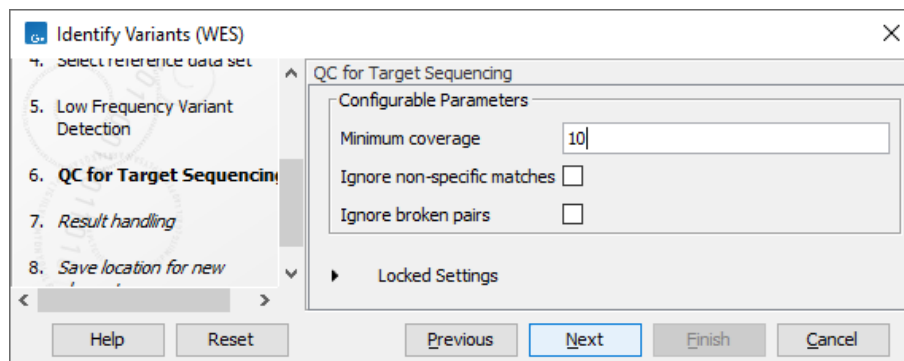Figure 4: *The correct parameter settings in the Low Frequencing Variant Detection step of the wizard.*

Figure 5: *The correct parameter settings in the QC for Target Sequencing step of the wizard.*

- (⬛) *Read_mapping-tumor_reads_chr5*: Mapped reads from the tumor sample to chromosome 5 of the human reference genome hg19

- (⬛) *Target_region_coverage-tumor_reads_chr5*: The targeted regions with information about the minimum, maximum and average coverage for each region.

- (⬛) *Target_region_coverage_report-tumor_reads_chr5*: Quality report for the mapping to the targeted regions

- (⬛) *Unfiltered_variants-tumor_reads_chr5*: Identified variants before any filtering.

- (⬛) *Identified_variants-tumor_reads_chr5*: Identified variants after filtering out variants with an average low base quality.

Figure 6: *The analysis outputs.*

- (▶▶▶) *Indels-indirect_evidence-tumor_reads_chr5*: This track is empty in this tutorial as no indels could be found in this reduced data set.

- (▮▮▮) *Genome Browser View Identify Variants*: Track List that enables the direct comparison and validation of identified variants in the context of the mapped sequencing reads and targeted regions.
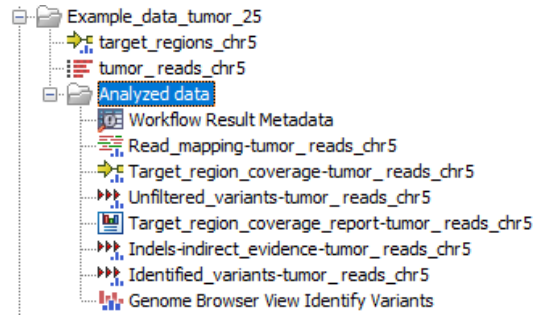
Checking the QC report for the target regions

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?

- 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)?

- Are all specific targets sufficiently covered?

To answer these questions, open *Target_region_coverage_report-tumor_reads_chr5*.

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 detection tool was set to 10. For amplicon data, we expect it to be larger than 100.

See figure 7 for the average coverage of the target regions in our example.

1.1 Summary

Number target regions	124
Total length of targeted regions	22,946
Minimum coverage	0
Maximum coverage	106
Average coverage	22.5
Median coverage	18.0
Number of target regions with coverage < 10	72
Total length of target regions containing positions with coverage < 10	13,020
Total length of target region positions with coverage < 10	3,992
Total length of target region positions with coverage ≥ 10	18,954
Percentage of target region positions with coverage ≥ 10 (%)	82.6

Figure 7: *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 8 you can see that in our example, 36.5% of reads and 29.1% 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	21,952	8,008	36.48	21,952	8,008	36.48

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	1,769,422	515,539	29.14	1,769,422	515,539	29.14

Figure 8: *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 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 is acceptable and what would be expected for a hybridization experiment (see figure 9).

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

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	52	41.94
≥90% of the targeted region has coverage at least 10	75	60.48
≥80% of the targeted region has coverage at least 10	88	70.97
≥70% of the targeted region has coverage at least 10	96	77.42
≥60% of the targeted region has coverage at least 10	100	80.65
≥50% of the targeted region has coverage at least 10	104	83.87
≥40% of the targeted region has coverage at least 10	107	86.29
≥30% of the targeted region has coverage at least 10	107	86.29
≥20% of the targeted region has coverage at least 10	107	86.29
≥10% of the targeted region has coverage at least 10	107	86.29
≥0% of the targeted region has coverage at least 10	124	100.00

Figure 9: *71% of all targets are more than 80% covered with at least 10 reads.*

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 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 10), as well as a table listing all found variants below.
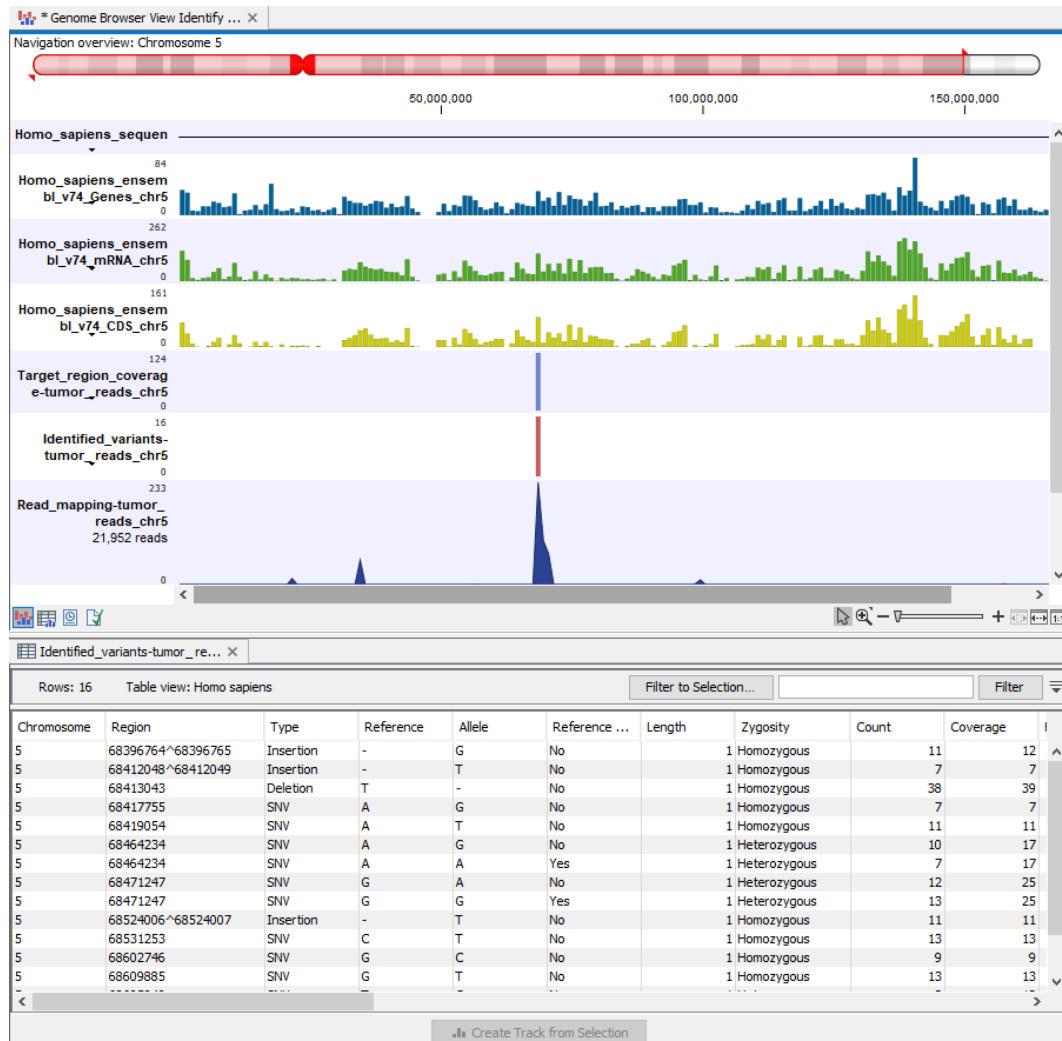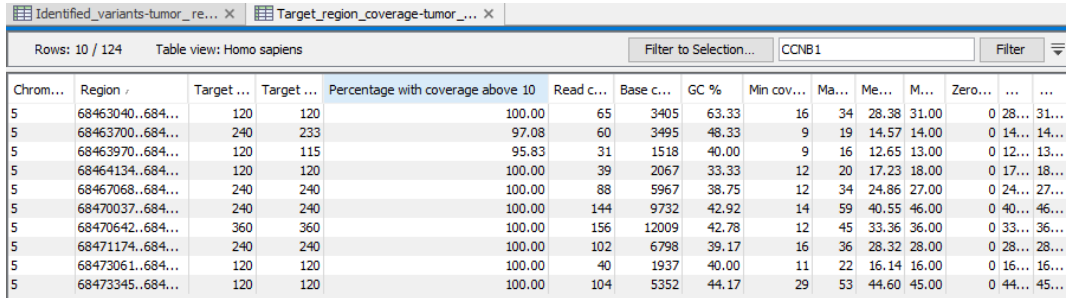
Figure 10: *The Genome Browser view shows the different tracks in one view and makes it easy to inspect the identified variants and their association with e.g., the reference sequences, genes or other tracks.*

2. In the left side of the opened Track List, double-click on the name of the track *Target_region_coverage-tumor_reads_chr5*. 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 11).

Chrom...	Region	Target ...	Target ...	Percentage with coverage above 10	Read c...	Base c...	GC %	Min cov...	Ma...	Me...	M...	Zero...	...	...
5	68463040..684...	120	120	100.00	65	3405	63.33	16	34	28.38	31.00	0	28...	31...
5	68463700..684...	240	233	97.08	60	3495	48.33	9	19	14.57	14.00	0	14...	14...
5	68463970..684...	120	115	95.83	31	1518	40.00	9	16	12.65	13.00	0	12...	13...
5	68464134..684...	120	120	100.00	39	2067	33.33	12	20	17.23	18.00	0	17...	18...
5	68467068..684...	240	240	100.00	88	5967	38.75	12	34	24.86	27.00	0	24...	27...
5	68470037..684...	240	240	100.00	144	9732	42.92	14	59	40.55	46.00	0	40...	46...
5	68470642..684...	360	360	100.00	156	12009	42.78	12	45	33.36	36.00	0	33...	36...
5	68471174..684...	240	240	100.00	102	6798	39.17	16	36	28.32	28.00	0	28...	28...
5	68473061..684...	120	120	100.00	40	1937	40.00	11	22	16.14	16.00	0	16...	16...
5	68473345..684...	120	120	100.00	104	5352	44.17	29	53	44.60	45.00	0	44...	45...

Figure 11: *From the Track List the "Target_region_coverage-tumor_reads_chr5" track has been opened in table view by clicking on the name of the variant track in the left side of the Track List. The filter has been used to show only CCNB1 entries.*

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 fully covered in their entirety (100%).

In conclusion we can say that our particular target (the CCNB1 gene) is sufficiently covered.
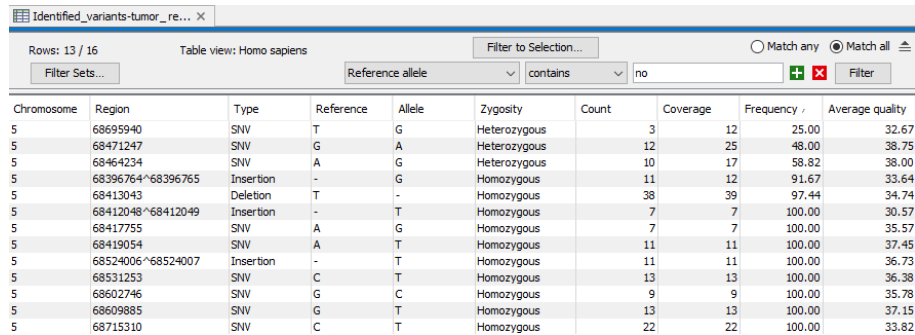
Check the identified variants for potential false positives

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 *Identified_variants-tumor_reads_chr5* that opened together with the Genome Browser view earlier (see figure 12). You can see that 16 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 13 variants that are different from the human reference sequence.

Chromosome	Region	Type	Reference	Allele	Zygosity	Count	Coverage	Frequency	Average quality
5	68695940	SNV	T	G	Heterozygous	3	12	25.00	32.67
5	68471247	SNV	G	A	Heterozygous	12	25	48.00	38.75
5	68464234	SNV	A	G	Heterozygous	10	17	58.82	38.00
5	68396764^68396765	Insertion	-	G	Homozygous	11	12	91.67	33.64
5	68413043	Deletion	T	-	Homozygous	38	39	97.44	34.74
5	68412048^68412049	Insertion	-	T	Homozygous	7	7	100.00	30.57
5	68417755	SNV	A	G	Homozygous	7	7	100.00	35.57
5	68419054	SNV	A	T	Homozygous	11	11	100.00	37.45
5	68524006^68524007	Insertion	-	T	Homozygous	11	11	100.00	36.73
5	68531253	SNV	C	T	Homozygous	13	13	100.00	36.38
5	68602746	SNV	G	C	Homozygous	9	9	100.00	35.78
5	68609885	SNV	G	T	Homozygous	13	13	100.00	37.15
5	68715310	SNV	C	T	Homozygous	22	22	100.00	33.82

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 it is as low as 25%. If you look at the number of reads supporting it (look in the **Count** column), you can see that there are only 3 reads, meaning that this variant is 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 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.

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.