Tutorial

Copy Number Variant Detection

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

Copy Number Variant Detection

Copy number variants are amplifications and deletions of exon and chromosome fragments as well as whole exons or chromosomes. The Copy Number Variant Detection algorithm is designed to detect copy number variations (CNVs) from targeted resequencing experiments. These can be either gene panels or whole exome sequencing. It identifies CNVs regions where the normalized coverage of the sample to be analyzed differs from the control samples in a statistically significant way.

The tool takes read mappings and target regions as input, and produces amplification and deletion annotations. The annotations are generated by a 'depth-of-coverage' method, where the target-level coverages of the case and the controls are compared in a statistical framework using a model based on 'selected' targets. Note that to be 'selected', a target has to have a coverage higher than the specified coverage cutoff AND must be found on a chromosome that was not identified as a coverage outlier in the chromosomal analysis step. If fewer than 50 'selected' targets are found suitable for setting up the statistical models, the CNV tool will terminate prematurely.

The algorithm implemented is based on the following papers:

- **Li et al.**, *CONTRA: copy number analysis for targeted resequencing*, Bioinformatics. 2012, 28(10):1307-1313. http://dx.doi.org/10.1093/bioinformatics/bts146
- **Niu and Zhang**, *The screening and ranking algorithm to detect DNA copy number variations*, Ann Appl Stat. 2012, 6(3):1306-1326. http://dx.doi.org/10.1214/12-AOAS539SUPP

To run the Copy Number Variant Detection tool at least one case sample and one control sample not containing the CNV to be detected are needed. It is beneficial if the control sample shares as many experimental parameters (e.g. gender or enrichment and sequencing method) with the case sample as possible.

The algorithm carries out the analysis in several steps.

- Base-level coverages are analyzed for all samples.
- A coverage baseline is generated using the control samples.
- A chromosome-level coverage analysis is carried out on the case sample, and any chromosomes with unexpectedly high or low coverages are identified. Targets on these chromosomes are not used to set up the statistical models.
- Sample coverages are normalized.
- Each chromosome is segmented into regions of similar coverage ratio.
- A statistical model for the variation in target-level coverage ratios is created using the results of the segmentation.
- Region-level CNVs are identified using the model from the previous step. The CNVs predicted in this way are both locally and globally supported by the data.
- If chosen in the parameter steps, gene-level CNV calls are also produced.

Overview The features presented in this tutorial include:

1. Data management configuration
2. Downloading and importing the example dataset
3. Running the Copy Number Variant Detection tool
4. Identifying target-specific, region-specific and gene-specific copy number variations
5. Creating a Genome Browser View for visual inspection of the results

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.

Example Dataset Sequence data:

We will use paired-end exome sequencing data from a massive acinic cell carcinoma sample and a matched normal tissue sample from the same patient. The samples were 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.**, *A case report and genetic characterization of a massive acinic cell carcinoma of the parotid with delayed distant metastases.* Case Rep Oncol Med. 2013, 2013:270362, http://dx.doi.org/10.1155/2013/270362

The example data provided for this tutorial include only the reads mapping of chromosome 14.

Data management configuration

In the last section of this tutorial the human genome reference will be used. The reference data can be downloaded using the Data Management function found in the upper right corner of the Workbench (figure 1). If you have not downloaded the reference data already, please download this sequence. If you have already downloaded the reference data, you can skip the steps below.

In this tutorial the following data is required: **Sequence - Human reference sequence hg19**. To download this data:

1. Click on the button **Data Management** in the top right corner of the Workbench to open the Data Management.
2. Select in the section Reference Data Elements the **Sequence Hg19** (highlighted in figure 1) in the list and click on the button labeled **Download** for this entry.
3. You will see a green progress bar labeled Download reference data in the very lower left corner of the Workbench indicating that the download is in progress. Wait for the download to finish. The process bar will disappear once all elements in the set are downloaded.

There is now a folder in your Navigation Area called CLC_References, and the sequence hg19 will be found in the homo_sapiens/sequence/hg19 folder.

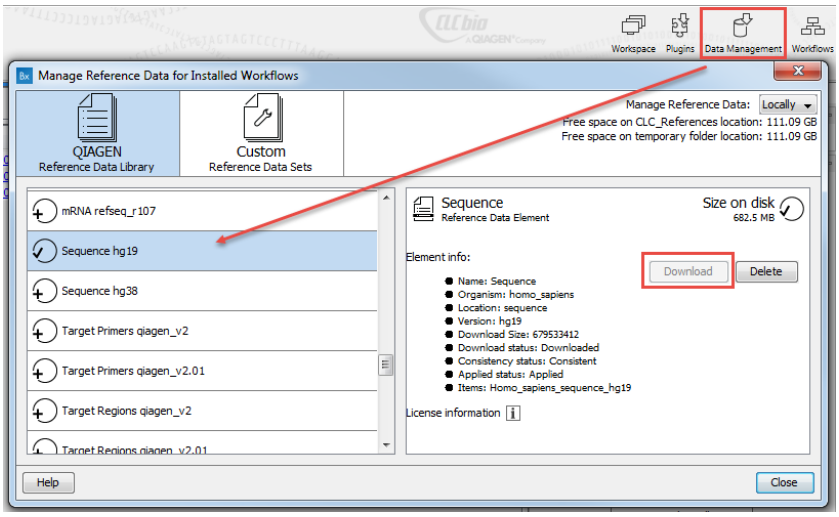

Figure 1: *View of the Reference Manager.*

Downloading and importing the data

1. Download the sample data from our web site: `http://resources.qiagenbioinformatics.com/testdata/normal_tumor_CNV.zip`
2. Click the button labeled **Import** located over the Navigation area, and choose the Standard Import option.
3. Choose the zip file called *normal_tumor_CNV.zip*. Leave the Import type set to **Automatic import**.
4. Click on the button labeled **Next**.
5. Choose where to save the example data in the Navigation Area by selecting a folder and click on the button labeled **Finish**.

The folder **normal_tumor_CNV** with the imported data is now located in the Navigation Area includes four files in the so-called track format:

- **Normal (Reads)** Sequencing reads from the normal tissue sample mapped to human hg19 chromosome 14.
- **Tumor (Reads)** Sequencing reads from the tumor sample mapped to human hg19 chromosome 14.
- **Target regions** Regions on chromosome 14 enriched for prior to sequencing.
- **Genes** Gene annotations located on chromosome 14.

Tracks contain data with coordinates based on a reference sequence. In our case, the reference is hg19. Typical track data types are variants and all kinds of annotations. Different track files based on the same reference can be visualized together in the Genome Browser View.

Running the Copy Number Variant Detection tool

Besides the case sample to be analyzed, at least one control sample is needed as the algorithm compares coverage in the case sample versus normal sample(s) to identify CNVs. During the copy number analysis the algorithm normalizes across the case sample as well as across all samples provided to determine the baseline and fold changes.

The first step for copy number detection is to run a read mapping on the sample and the control samples. This read mapping step has already been performed and the resulting read mappings are provided in the normal_tumor_CNV folder as described above.

The Copy Number Variant Detection tool is located in the Toolbox under

Tools | **Resequencing analysis** | **Copy Number Variant Detection**

Double-click on this tool to open the wizard. The wizard will guide you through the input options and parameters for this algorithm.

1. Select the read mapping of the sample to be analyzed for copy number variants. Click on **Tumor (Reads)** in the Navigation Area to highlight this file and click either on the blue arrow pointing to the right or double-click on the Tumor (Reads) file to select the tumor sample. The Tumor (Reads) file should now show up in the Selected elements field (see figure 2).

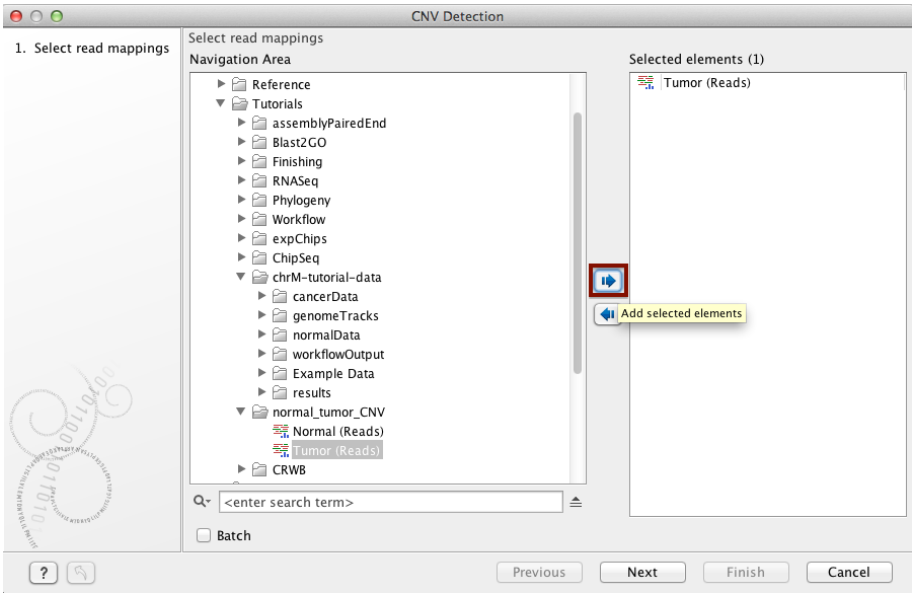

Figure 2: *Select reads mappings.*

2. Click on the button labeled **Next**.
3. In the next wizard step, the input parameters can be selected by clicking on the browse icon on the right side of each field (see figure 3). For Target regions track select the file called **Target regions**.
4. Similarly, select for Control mappings the **Normal (reads)** file and for the Gene track the **Genes** file (see figure 3).
5. Leave the **Read filters** on default settings to ignore non-specific matches and ignore broken pairs. Reads are specified as non-specific matches when there is more than one position

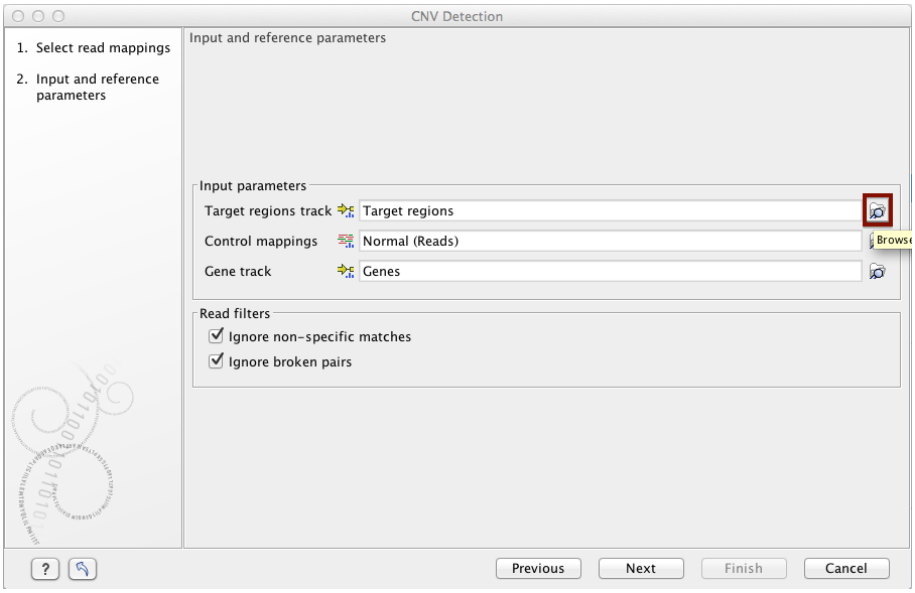

Figure 3: *Input and preference parameters.*

in the genome where the read can map to. Broken pairs are paired end reads where the distance between the two read partners does not fulfill the user-specified or calculated parameters.

6. Click the button labeled **Next**.

7. Now specify the algorithm specific parameters. As the Copy Number Variant Detection tool relies on statistics, the **Threshold for significance** is defined by a p-value. The default value used is 0.05. If you increase this value, more copy number variants will be reported. The value for **Minimum fold change** is set to 1.5 by default. The lower you set this value, the more copy number variants will be reported. Set the cutoff for the parameter **Low coverage cutoff** to 10, as the coverage for the samples under consideration is not very high (see figure 4).

   Advanced tip: You can investigate the coverage of your sample in the Biomedical Genomics Workbench 2.1 by running the **QC for Targeted Sequencing** tool.

8. In this sample, we would like to look for small copy number variations which span very few targets. Therefore, choose the **Graining level "Fine"** (see figure 4). If you want to learn about the difference between Coarse, Intermediate, and Fine, you can mouse-over these words and a tooltip with more details will show up. Alternatively, you can press the question mark button in the lower left corner of the tool wizard. This will link you to the corresponding manual entry.

9. To increase the sensitivity even further, down to the single target level, check the option **Enhance single-target sensitivity**. This option may lead to more false positive calls, but in this case, we prioritize a high sensitivity.

10. Click the button labeled **Next**.

11. In the Result handling wizard step choose to **Save** your results and click **Next**.

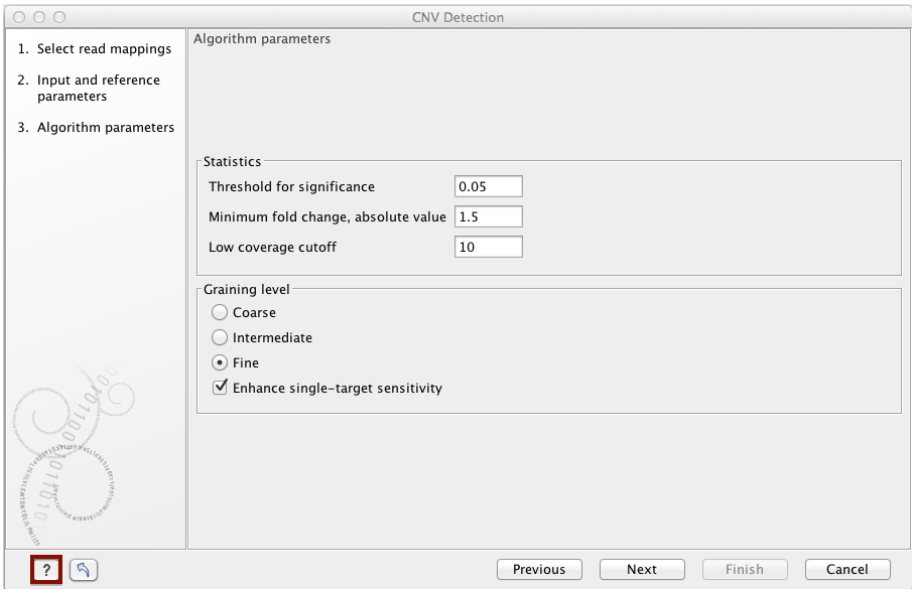

Figure 4: *Algorithm parameters.*

12. Select a location in the Navigation Area to save the results. You can create a new folder, e.g. called results, by using the New Folder button located over the file structure displayed. Click **Finish**.

The analysis of the copy number variants only takes a few minutes on this sample.

Results of the Copy Number Variant Detection tool

Five different output files are generated with the output options chosen for the Copy Number Variant Detection tool (see figure 5). Two are reports and the other three are track files.

- **Tumor (Gene CNVs)** Reports the copy number changes per gene. Only available if a gene list was provided as input.
- **Tumor (Region CNVs)** Reports copy number changes based on regions. Each region consists of a number of consecutive targets showing the same fold changes within the statistical error.
- **Tumor (Reads) CNV results report** Report that summarizes the results of the Copy Number Variant Detection tool.
- **Tumor (Target CNVs)** Optional: track that gives information about the predicted copy number status of each individual target.
- **Tumor (Reads) CNV algorithm report** Optional: reports the statistical model used. This information can be used to evaluate how well the assumption of the model were fulfilled.

To open any of the files double click on it in the **NAVIGATION AREA**. Open the **Tumor (Gene CNVs)** file for visualization of the Gene CNVs track (see figure 6).

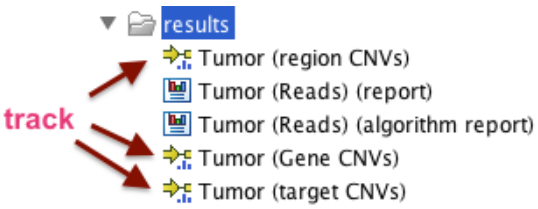

Figure 5: *Output files.*

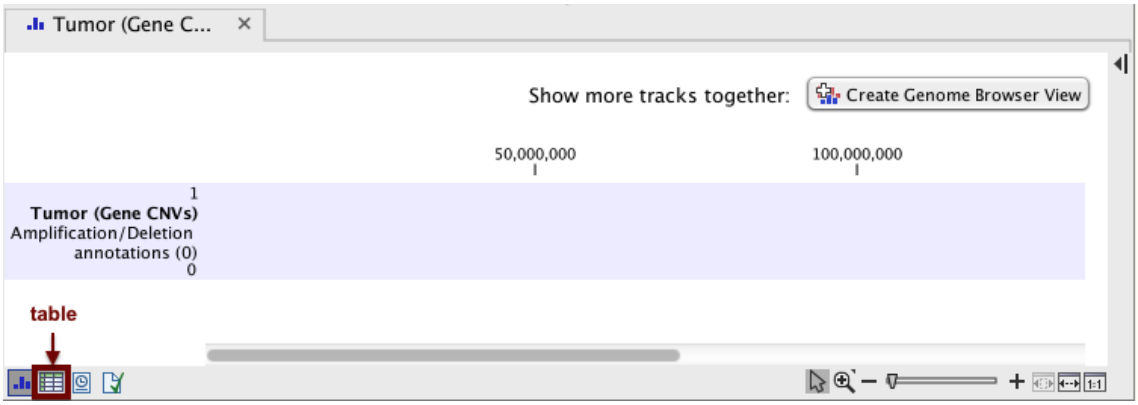

Figure 6: *Visualization of the Gene CNVs track.*

To access the data underlying the visualization, switch to the tabular view. Click on the **show table** button highlighted in figure 6. In the table showing up, three genes are reported to contain copy number variants together with details for each copy number variant. These details include the consequence (gain or loss), the fold change and the p-value (figure 7).

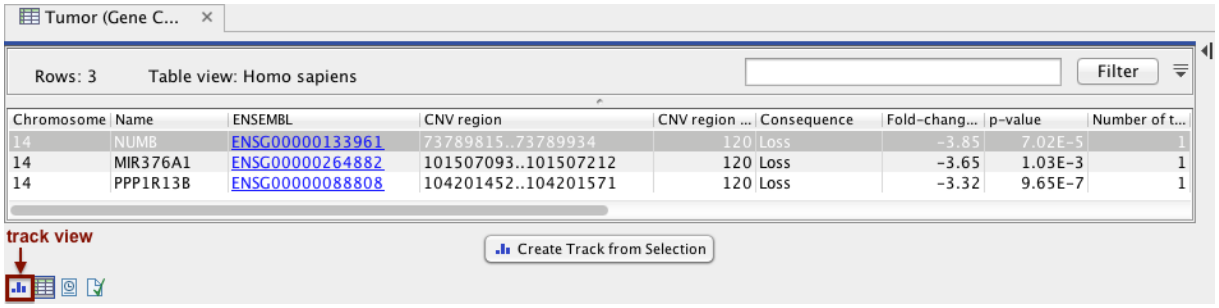

Figure 7: *Tabular view of the data.*

To go back to the visual representation of the copy number variants on the gene level Tumor (Gene CNVs), click on the **show track** button located left to the show table button.

Visualizing results in a Genome Browser View

To investigate the copy number variant in context of the read mapping from the case sample and the control sample we create a Genome Browser View. The Genome Browser View allows for the visualization of multiple track files within the same view. A detailed description of the Genome Browser View tools is provided in the manual: http://resources.qiagenbioinformatics.com/manuals/biomedicalgenomicsworkbench/current/index.

php?manual=Genome_browser.html.

1. Click on the button labeled **New** in the upper left corner of the Workbench and select **Genome Browser View**.
2. Select the file **Homo_sapiens_sequence_hg19** from the folder

   CLC_References | **homo_sapiens** | **sequence** | **hg19**

   from the Navigation Area.
3. In addition select the following files from the normal_tumor_CNV and results folders:

   - **Genes**
   - **Tumor (Reads)**
   - **Normal (Reads)**
   - **Tumor (Gene CNV)**

   Please note, that the tracks will show up in the view in the same order, as they have been selected in this step. You do, however, have the option to rearrange them by dragging them in the Genome Browser View.
4. Click **Finish**. A new Genome Browser View will open.

At this point you do not see much because you look at chromosome 1 where no reads are mapped. On the right side of the Workbench you will see a grey **SIDE PANEL**. Under Navigation select chromosome 14 to be displayed. The Genome Browser View now should look similar to figure 8.

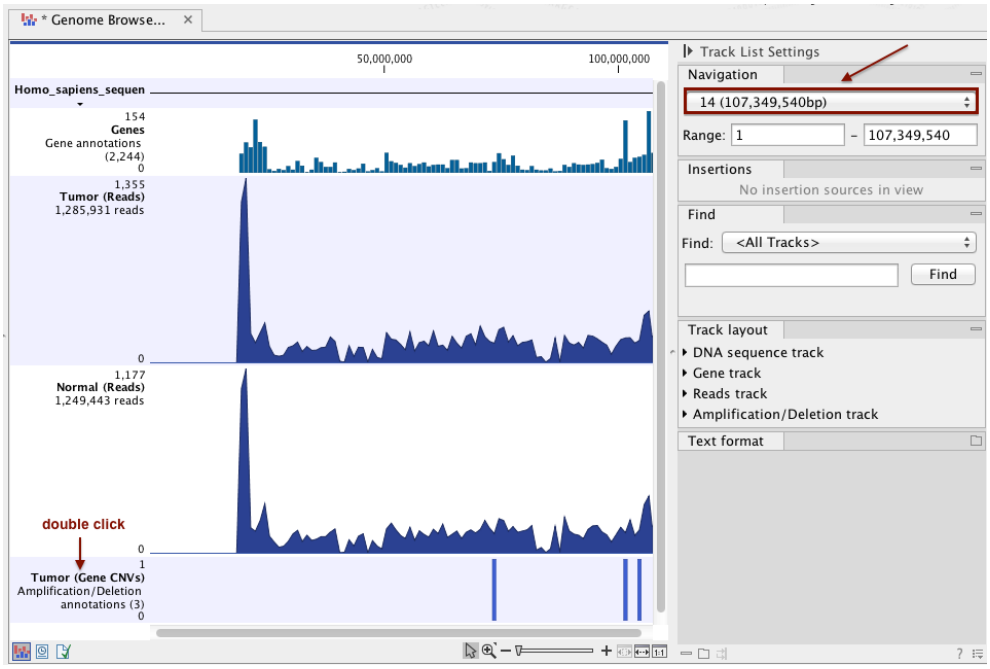

Figure 8: *Genome Browser View.*

By double clicking on the name in the track **Tumor (Gene CNVs)** (see figure 8) you will open the underlying table of these CNV annotations. The table can be used to navigate to a position on chromosome 14 where a CNV has been reported.

1. **Click** on the gene name **PPP1R13B** (figure 9).
2. In the **SIDE PANEL** on the right side go down to **Track layout**, click on **Read tracks**, check the box for **Fix maximum of coverage** graph and enter the value 200. The latter ensures that the y-axis for the read coverage is adjusted to the same values for all read mappings displayed in the Genome Browser View and therefore the coverage can be more easily compared.
3. Zooming options are located in the lower right corner beneath the Genome Browser View. Click on the **zooming option** with the red square to zoom into the selection.

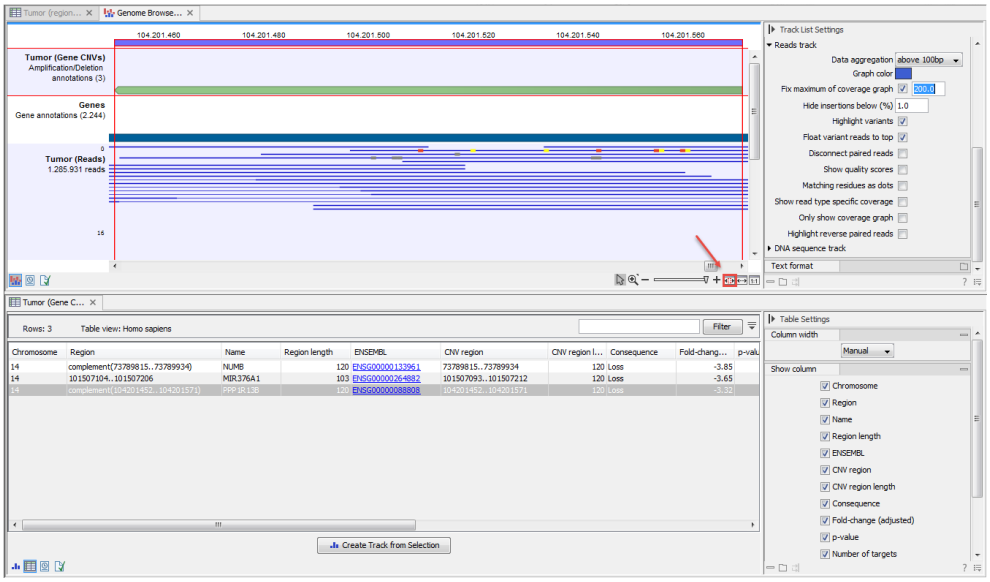

Figure 9: *Tabular view of CNV annotations.*

If you now compare the number of reads mapped back to this region of chromosome 14 in the tumor sample and the control sample, you will see that fewer reads are mapped for the tumor sample. This is in accordance with the CNV (loss) detected by the algorithm.

Finally, we want to save the **Genome Browser View**. You can do so by using the **Save** button in the upper left corner of the workbench.