QIAGEN

Tutorial

Analysis of Viral Hybrid Panel Data and Identification of Viral Integration Sites

August 11, 2021

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Tutorial

Analysis of Viral Hybrid Panel Data and Identification of Viral Integration Sites

This tutorial will take you through the tools available in CLC Microbial Genomics Module to identify the presence of viruses in hybrid panel capture data and search for viral integration sites.

Introduction Identifying the viral species/strain and possible integration sites is becoming increasingly important to understand the mechanisms behind disease patterns. Human papillomavirus (HPV), in particular, is recognised as the leading cause of cervical cancer. This tutorial will guide you through the viral tools included in CLC Microbial Genomics Module to analyze NGS data from hybrid capture panels.

Prerequisites For this tutorial, you must be working with CLC Genomics Workbench 21.0 or higher and you must have installed CLC Microbial Genomics Module 21.1 or higher.

Overview Using Hybrid Capture NGS data from *Human Papillomavirus (HPV)* positive samples collected among Gabonese women, this tutorial will guide you through the following:

- Customizing a provided template workflow in order to identify the HPV strains present in the samples.
- Using a database of HPV genomes and a human reference in order to identify potential viral integration sites.

Downloading and importing the data

For this tutorial we will use a *HPV* data set originally created and described by Nkili-Meyong et al., 2019. To ensure a reasonable analysis time for the tutorial, only 4 of 21 samples are included in this tutorial, and each read file has been reduced to include only 20% of the original reads.

To begin: **Download** the data from our website: `http://resources.qiagenbioinformatics.com/testdata/viral_tutorial/Viral_analysis_tutorial_data.zip`. Unzip and save the files locally.

The data for this tutorial includes the following files:

- **Homo_sapiens_sequence_hg38_chr11.clc**: Chromosome 11 from the human reference genome.
- **Homo_sapiens_refseq_GRCh38_chr11_CDS.clc**: CDS regions for chromosome 11.
- **HPV genomes database.clc** 192 HPV reference genomes
- **HPV genomes database (taxpro index).clc**: Taxonomic profiling index created from the HPV reference genomes
- **HPV type 16.clc** HPV type 16 reference genome.
- **Raw_reads**: 5 paired-end data files in fastq format from HPV target capture samples. Each sample has been downsampled to contain a maximum of 50,000 read pairs.

Now that the data has been downloaded, we can get started.

1. Open your CLC Workbench and go to **File** | **Import** | **Standard Import**. In the wizard, leave the import option to **Automatic Import**. Choose the five files ending in ".clc" as seen in figure 1.

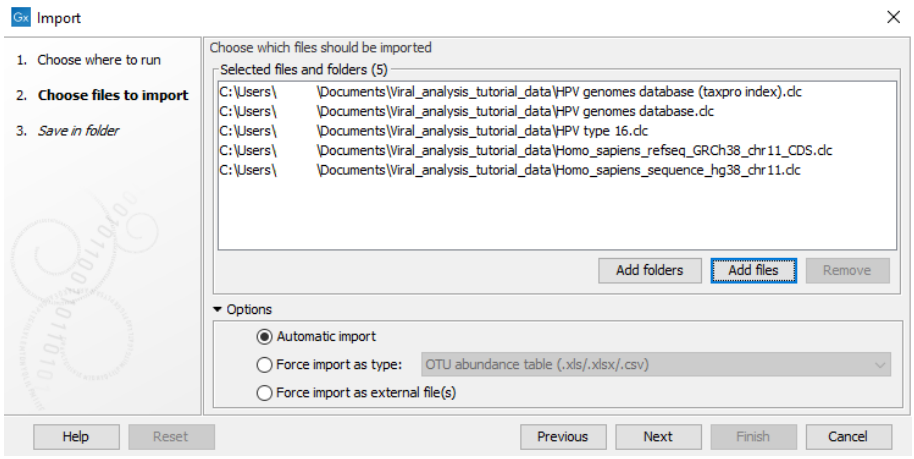

Figure 1: *Select the five files ending in ".clc".*

2. Save the imported files into a new folder you can call for example "Viral analysis tutorial".
3. Next, import the read files. Go to **File** | **Import** | **Illumina...**.
4. Click "Add folder" and select the "Raw_reads" folder from the download location (figure 2). Leave the import options as default and save the reads to a subfolder (for example titled "Raw reads") within the "Viral analysis tutorial" folder .

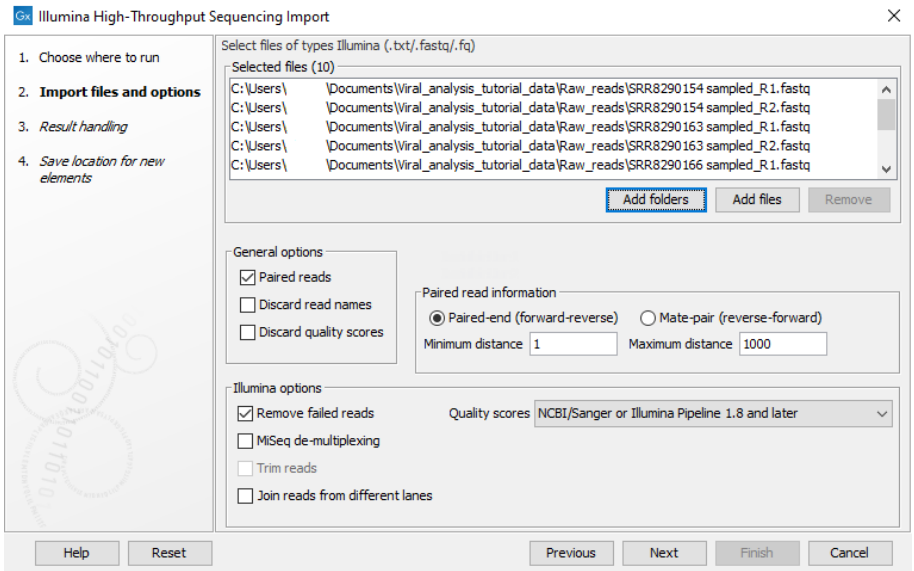

Figure 2: *Import the contents of the "Raw_reads" folder.*

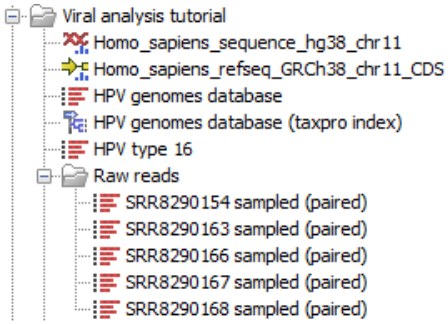

Figure 3: *Viral analysis folder after import of all files.*

Your "Viral analysis tutorial" folder should look like figure 3

Remember that when working with your own data, you can create, customize or download databases using the following tools found in the **Databases** ( ) | **Taxonomic Analysis** ( ) and the **Tools** ( ) folders of the **Microbial Genomics Module** Toolbox:

- **Create Annotated Sequence List** ( )
- **Download Custom Microbial Reference Database** ( )
- **Download Curated Microbial Reference Database** ( )

Taxonomic profiling indexes can be create with **Create Taxonomic Profiling Index** tool found in **Databases** ( ) | **Taxonomic Analysis** ( ).

Now that all the reference data, databases and sample data have been imported, we are ready to configure the provided template workflow so you can perform the analysis of the sample data.

Run the Analyze Viral Hybrid Capture Panel Data workflow

In this section, you will run the **Analyze Viral Hybrid Capture Panel Data** workflow. The workflow is designed for determining abundance of viruses present in a sample and identifying the closest matching reference among a user-specified reference list. The workflow also maps to the reference, calls variants and outputs a consensus sequence.

The default workflow includes a step with mapping host reads to a set of housekeeping genes. This is not relevant for the data used in the workflow and we will therefore create a modified copy of the workflow. How to do so is described in detail below.

1. In the Toolbox, go to **Metagenomics** ( ) | **Taxonomic analysis** ( ) | **Workflows** ( ). Select **Analyze Viral Hybrid Capture Panel Data** ( ) with one click (do not open the wizard yet with a double click). Right-click on the name of the workflow and choose the option **Open Copy of Workflow** (figure 4).
2. This opens a copy of the workflow in the view area of your workbench.
3. Locate the following five workflow elements: "Map Reads to Human Control Genes", "Read Mapping Human Control Genes", "Collect and Distribute Human Control Genes

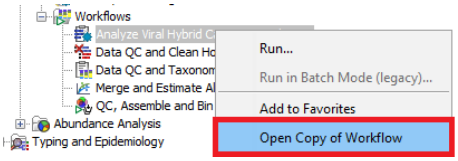

Figure 4: *Open a copy of the workflow.*

Reports", "Combine Human Control Genes Reports" and "Report (Human Control Genes Read Mapping Report)" (figure 5).

4. Select and delete the elements by pressing "Delete" or right-click and select "Remove".

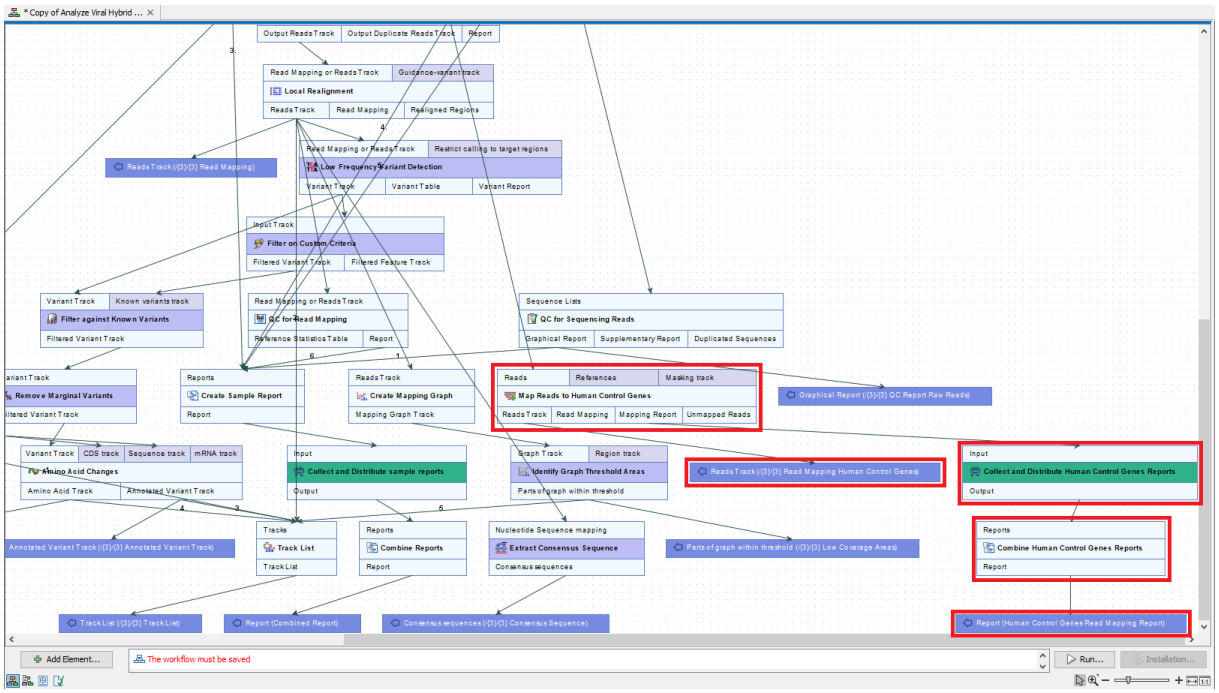

Figure 5: *Delete the workflow elements associated with human control genes.*

5. Save your workflow in your Navigation Area by dragging the workflow tab to the relevant location in your Navigation Area (here in the folder called "Viral analysis tutorial"). You can also right-click on the workflow copy tab and select "Save as...". Keep the workflow open.

We are now ready to run the modified workflow.

1. In the open workflow view, click on "Run..."
2. Select all 5 samples from the "Raw reads" folder (figure 6). Do not check the "Batch" option as the workflow already creates batch units. Click **Next**.
3. Here, you can configure batching. We will leave it as is so click **Next**.
4. The next wizard window gives you an overview of the samples selected. We want to analyze all 5 samples so we just click **Next**.

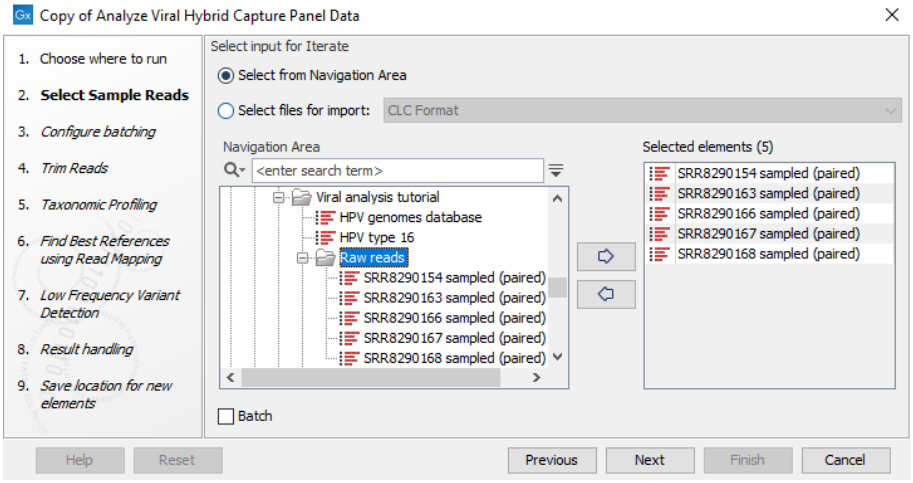

Figure 6: *Select the contents of the Raw reads folder*

5. In the next dialog, you can set the trimming options including adapters if you data contains adapter sequences. We will leave them as default (figure 7). Click **Next**.

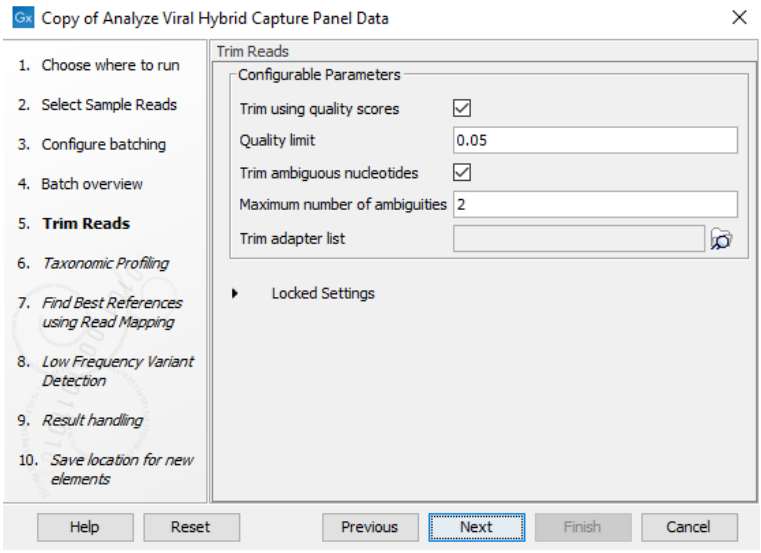

Figure 7: *Select trimming options and optionally add adapter lists.*

6. In the Taxonomic profiling wizard step, click ( ) and locate and select "HPV genomes database (taxpro index)". Disable "Filter host reads" as we are working with a virus which is potentially integrated into the human genome. The settings should look like figure 8.
7. In the next wizard step, select the HPV genomes database (figure 9) and click **Next**.
8. Leave the variant detection settings as default and click **Next**.
9. Save the Results in a new folder, for example called "Viral Analysis".

The workflow will now run. It may take a few moments to complete. You can monitor its progress in the Processes tab, located next to the Toolbox in the bottom left side of the Workbench. When it is done, we will take a look at the output.

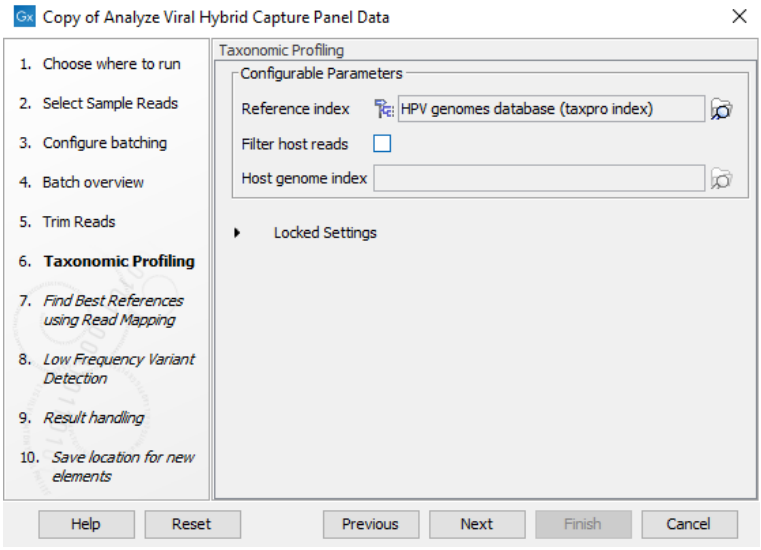

Figure 8: *Select the taxpro index and disable host read filtering.*

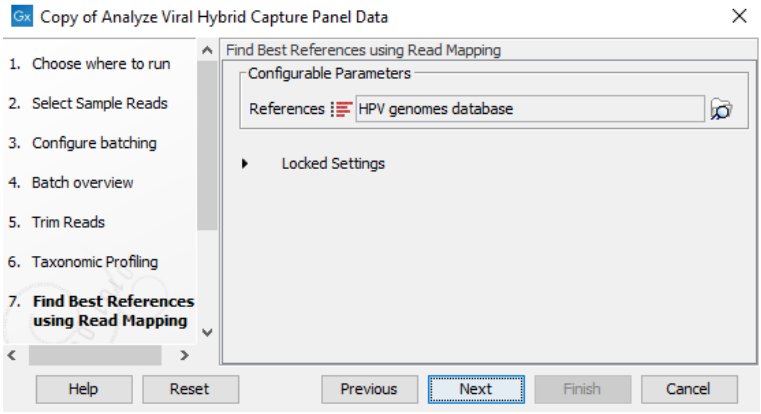

Figure 9: *Select the HPV genomes database.*

- First, open the **Combined report**. This contains summary statistics for all samples. For example, in the "QC for read mapping - Reference coverage" section, you can see that sample SRR8290166 only covered 86.30 % of the reference. When you are done, close the report.
- Open the **Merged Abundance Table**. Here, you can see that the samples all contain Human papillomavirus type 16. There are a few additional hits but the relative abundance is low and they can therefore likely be attributed to noise.

The workflow also generates a folder for each sample containing output files specific to that sample. We will look at a few of these files:

10. Open the folder for sample SRR8290163 (it should be called "SRR8290163 sampled (paired)").
11. Open the "Find Best Reference Report". Here you can see that matching reference is HPV type 16 and that the reference is completely covered by the viral reads.

12. Open the Track List. This view shows the read mapping to the best matching reference (figure 10). We can see in the mapping and the Low Coverage Areas track that there there is a dip in the coverage in a region of approximately 2000 bps indicating a part of the virus genome has been deleted. This is common in viral integration events. There is also a number of variants detected but these are are not covered in this tutorial.

Figure 10: *Track list output of Analyze Viral Hybrid Capture Panel Data*

13. Repeat this inspection for the other 4 samples. Sample SRR8290166 and SRR8290163 has a similarly sized areas of low coverage.

For a description of all output files not covered in this tutoeial, see https://resources.qiagenbioinformatics.com/manuals/clcmgm/current/index.php?manual=Analyze_Viral_Hybrid_Capture_Panel_Data.html.

Identifying Viral Integration Sites

In this section, you will look for integration of HPV into the human genome. To limit the runtime and memory required, we will only use two of the samples and chromosome 11 from the human genome. We will also limit the search to HPV type 16. When identifying viral break point in your own data, you can download the human reference with the **Reference data manager** (see https://resources.qiagenbioinformatics.com/manuals/clcserver/current/admin/index.php?manual=Reference_data_management.html)

1. To begin, from the Toolbox, choose:
Metagenomics | **Taxonomic analysis** | **Identify Viral Integration Sites** or use **Launch** to search and run.

2. Select samples SRR8290163 sampled (paired) and SRR8290166 sampled (paired) (figure 11). We are running the tool as standalone on multiple samples so we will enable **Batch**.

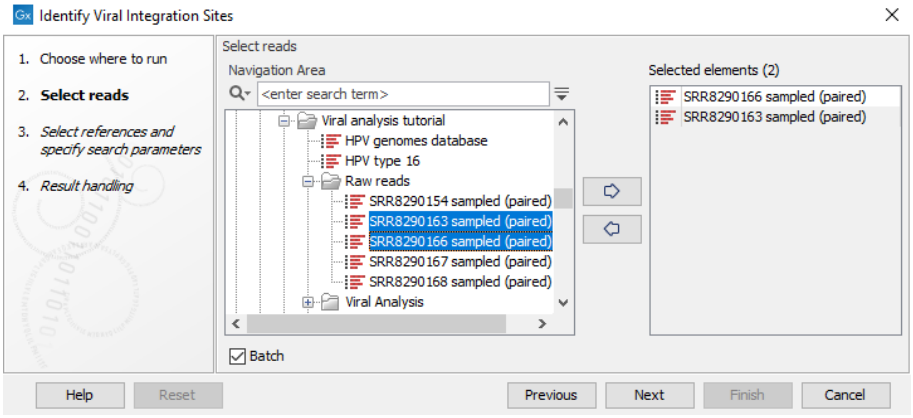

Figure 11: *Select sample SRR8290163 and SRR8290166*

3. In the next step, you should see an overview of the two batch units. If you do not see this step, you probably forgot to enable Batch. Click **Next**

4. In the next wizard step, select HPV type 16 as input for "Viral references". Select Homo_sapiens_sequence_hg38_chr11 and Homo_sapiens_sequence_hg38_chr11_CDS as "Host reference" and "Host annotations". The viral reference is already annotated and we will therefore leave "Viral annotations" empty. Leave the other settings as default. Your screen should look like figure 12.

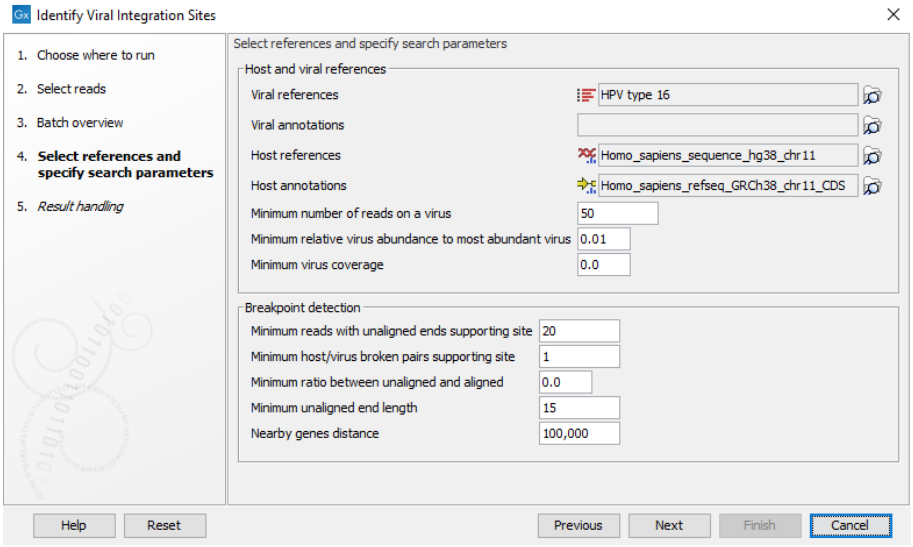

Figure 12: *Select the viral and host references.*

5. In Result handling, leave all the output options checked and select "Save to specified location" and "Create subfolders per batch unit". Save the results to a location of your chosing, for example a folder titled "Viral integration"

The tool will now run and may take a few minutes to complete. You can monitor its progress in the Processes tab, located next to the Toolbox in the bottom left side of the Workbench. When it is done, open the results folder for "SRR8290166 sampled (paired) batch".

1. We will first take a look at the host mappings in a track list.
   - Press Ctrl + L (⌘ + L on a Mac) and select the Host mappings () and Host breakpoints (). Click finish.
   - Double-click on the Host breakpoint in the track list to open a table view of the breakpoints.
2. Click on one of the two breakpoints. The track list viewer automatically zooms in on the region.
3. Observe that the integration is supported by both unaligned ends and broken pairs (figure 13.

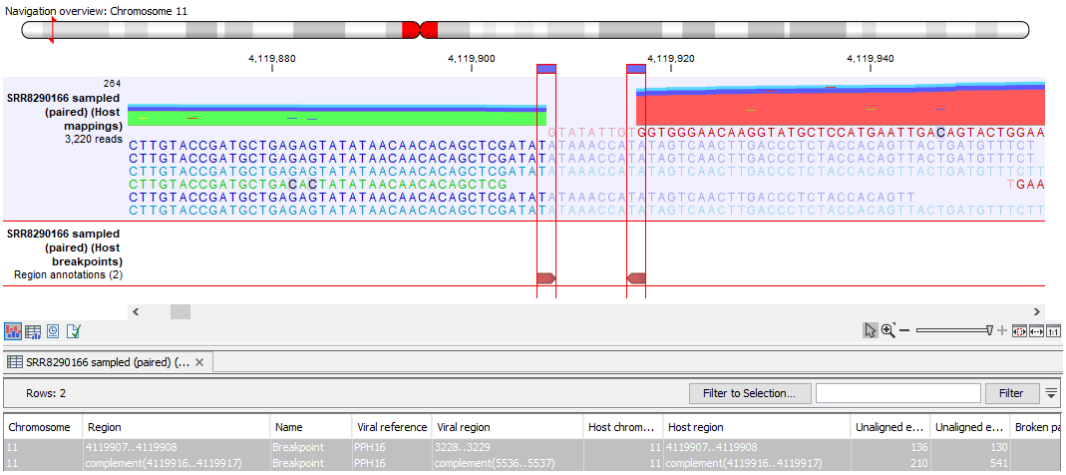

Figure 13: *Host read mapping with breakpoint regions selected.*

4. Optionally, you can repeat the above for sample SRR8290163.
5. Finally, open the viral integration outputs () for both samples. Here, you can see that the two integrations are different. The regions deleted in the virus during the integration are not the same. The viral regions CDS E4, E5 and L2 have been deleted in SRR8290166 while in sample SRR8290163 it is E1 and part of E2 that is deleted. They are also integrated in different regions of the host (figure 15).
6. In sample SRR8290163, zoom in to see that the viral integration event potentially impacts multiple CDS regions in the host genome. To recreate the view shown in figure 15, set the virus extent to 50 and use the mouse to zoom in on the host chromosome.
7. Repeat for SRR8290166 to further confirm that the integration events are different.

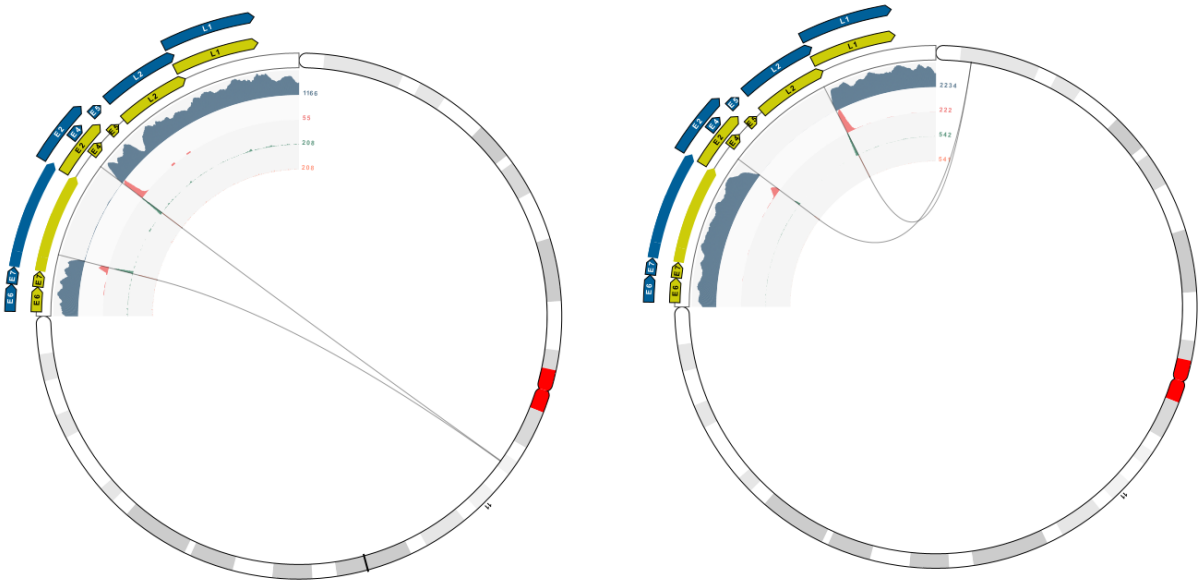

Figure 14: *Viral integration events in chromosome 11.*

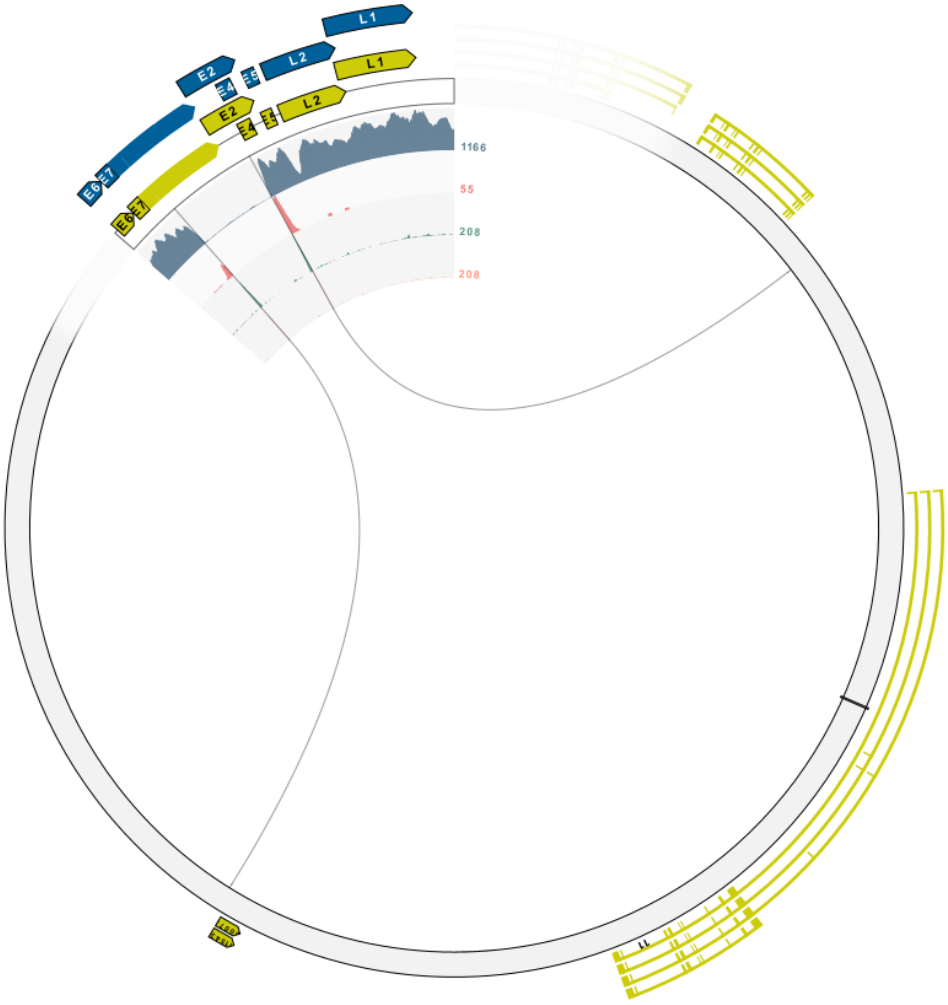

Figure 15: *Zoom in to see that the viral integration event impacts CDS regions.*

References

[Nkili-Meyong et al., 2019] Nkili-Meyong, A. A., Moussavou-Boundzanga, P., Labouba, I., Koumakpayi, I. H., Jeannot, E., Descorps-Declère, S., Sastre-Garau, X., Leroy, E. M., Belembaogo, E., and Berthet, N. (2019). Genome-wide profiling of human papillomavirus dna integration in liquid-based cytology specimens from a gabonese female population using hpv capture technology. *Scientific reports*, 9(1):1–11.