Virus detection with and without host genome using different amount of reads

The session course_VirusDetect_potato contains small RNA-seq reads from potato (Solanum tuberosum). It has three FASTQ files containing different amount of reads. We will detect viruses in all of them to see how the amount of reads affects the results. We also want to see what effect the host genome subtraction has on the results.

1. Select **Sample_7.fastq** and check how many reads it has. What length are the majority of the reads? Is the base quality ok?

2. Select **Sample_7.fastq** and run **VirusDetect** with default parameters. Repeat the run so that you set the host organism to potato (**Solanum_tuberosum.SolTub_3.0**). Open both log files and look at them side by side.

-How many reads mapped to reference virus database and how many mapped to the potato genome?

-How many reads were used for the de novo assembly?

-How did the host genome subtraction affect the number of de novo contigs assembled and why do you think that is?

-Were any host-derived contigs detected?

-What was the number of non-redundant contigs used for BLAST searches?

-How many viruses were detected by BLASTN? Were the same viruses identified by both VirusDetect searches?

-Was any BLASTN virus assignment filtered out?

3. Extract the html reports from each of the tar packages using the tool **Utilities / Extract .tar or .tar.gz file** so that you set the parameter **Extract by extension =.html**.

-Why does the tar package from the run with the host genome subtraction contain more html reports?

-Open both **Sample_7_blastn_matching_references.html** files side by side and inspect them for the details of the reported viruses. Did the host genome subtraction increase the virus coverage and depth?

-Compare the files **Sample_7_undetermined.html**. Was the same number of undetermined contigs found in both VirusDetect searches? Why?

4. Extract the **pdf** reports from the second tar package (made with the host genome).

-How does the contig length look like? Do both viruses have equally long contigs?

5. Repeat the run with the host genome so that you set the parameter **Minimum fraction of virus reference covered by contigs = 0.9** (90%) and **Return matching reference sequences = yes**.

-How many viruses do you detect now? Can you explain why M72416 is not reported anymore?

6. Inspect the reference virus coverage in the genome browser: Extract the files **Sample_7_blastn_matching_references.fa** and **Sample_7_blastn_matching_references.fa.fai**.

Select **Sample_7.fastq** and **Sample_7_blastn_matching_references.fa** and the tool **Bowtie for single end reads and own genome**, and check in the parameters that the fasta file is assigned as the genome.

-When the results come, check in the log file how many reads aligned.

Select **Sample_7.bam** and **Sample_7_blastn_matching_references.fa** and **Genome browser**. In the Genome pull-down menu, scroll to the bottom to find your reference. Zoom in to see the read alignments.

-Is the coverage uniform along the virus reference sequence?

7. Check if more viruses can be detected with more reads: Select **Sample_3.fastq** which has 250 000 reads, and analyze it with **VirusDetect** setting potato as the host organism.

-How does the number of contigs and detected viruses change when more reads are used?

Extract the **html** files from the tar package.

-Open **Sample_3_blastn_matching_references.html**. What has happened to the coverage and the number of contigs?

-Open **Sample_3_blastx_matching_references.html**. Do the hits look convincing in terms of the host organism?

-Open **Sample_3_undetermined.html**. Has the number of undetermined contigs increased? Why do you think that is?

Extract the **pdf** files from the tar package.

-Open **Sample_3_KJ534601.bn.pdf** and compare it to **Sample_7_KJ534601.bn.pdf**. What can you say about the number and length of the contigs?

8. BONUS EXERCISE What happens if you increase the number of reads further? Select the file **Sample_1.fastq** which has 2 500 000 reads, and analyze it with **VirusDetect** setting potato as the host organism.

-Do you detect more viruses?

-Were host-derived contigs assembled this time?

-How does the contig length for KJ534601 look now?