Tutorial

Analysis of SARS-CoV-2 Using MinION Sequences

January 10, 2025

Sample to Insight

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Analysis of SARS-CoV-2 Using MinION Sequences

Sequencing using MinION is a fast and cost-efficient way to track the evolution of SARS-CoV-2 as it allows sequencing to be done in a few hours with little investment in laboratory equipment. However, samples are often of low quality and contaminated with RNA from other sources. Here we demonstrate how to overcome the challenge of metatranscriptomics and get the most from Oxford Nanopore MinION reads using the *CLC Genomics Workbench*.

The tutorial covers the following:

- Import of data required for the analysis.
- Trimming MinION reads.
- Mapping MinION reads to a reference.
- Calling variants in the sample relative to a reference and visualizing variant calls and mappings.
- Extracting a consensus sequence from mapped reads.
- Using BLAST to identify a strain.
- Running the pipeline in a workflow.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 25.0 or higher.

If you are not already familiar with the tools for long read analysis, we suggest first completing the tutorial "De Novo Assembly Using Long Reads and Short Read Polishing".

Background information

The data we will be working with is a metatranscriptomic MinION data set from the Wuhan seafood market pneumonia outbreak [Chan et al., 2020]. Sampling the virus responsible for the disease involves obtaining sputum, throat or nasopharyngeal swabs, or bronchoalveolar lavage fluid (BALF) samples from a patient. Because of this, samples will contain RNA from non-viral sources. This data set was prepared using the Sequence-Independent, Single-Primer Amplification (SISPA) protocol for additional viral sequence enrichment.

Some basic tips

- In this tutorial, we refer to the tools in the Workbench menus, but tools can also be launched by clicking on the **Quick Launch** button (🚀) in the Workbench toolbar.
- We will run several of the tools in "Batch" mode. Checking the "Batch" option in launch wizards allows us to analyze several input elements individually, with the same parameter settings, while stepping through the launch wizard for the tool only once.

Import the data

1. The example data is available from our web site: https://resources.qiagenbioinformatics.com/testdata/SARS-CoV-2_MinION_example_data.zip.
   Download and unzip it.

2. Open the *CLC Genomics Workbench*.

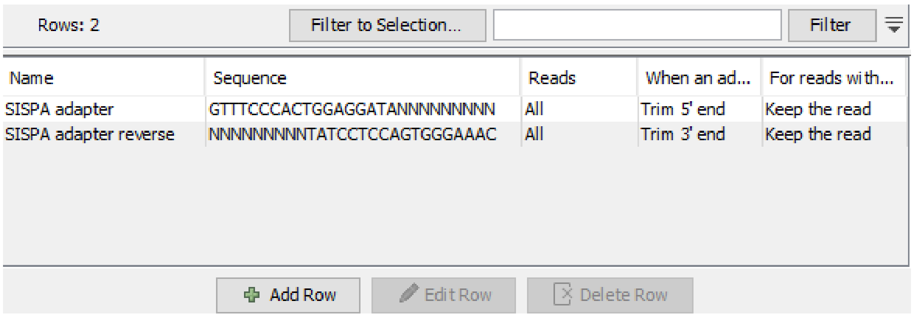
3. Create a new folder for the project with a relevant name, for example named "SARS-CoV-2 MinION Tutorial".

4. Import the reference sequence, trim adapter list, and custom database provided with the example data by going to:

   File | **Import** | **Standard Import**

   Choose the option **Automatic import** and select "MT135044.gbk", "SISPA adapter trim list.clc", and "Custom Betacoronavirus Database.clc". Save the items in the folder you just created.

   After import, you can open the "SISPA adapter trim list" and see that it looks like the one shown in figure 1.



Name	Sequence	Reads	When an ad...	For reads with...
SISPA adapter	GTTTCCCACTGGAGGATANNNNNNNNN	All	Trim 5' end	Keep the read
SISPA adapter reverse	NNNNNNNNNTATCCTCCAGTGGGAAAC	All	Trim 3' end	Keep the read

Figure 1: *Trim adapter list.*

5. Import the MinION reads by going to:

   File | **Import** | **Oxford Nanopore. . .**

   Select "SRR10948550.fastq" and "SRR10948474.fastq". Do not tick the "Discard quality scores" box. Save to a new folder, for example named "Raw Reads".

   Note: the full dataset can also be downloaded from SRA by using **Download** | **Search for Reads in SRA. . .** ( ) and searching for "PRJNA601630".

Information on these read sets

Run	# of Spots	# of Bases	Size	Sampling method
SRR10948474	505,484	284.6M	250.1Mb	Sputum
SRR10948550	425,717	146.3M	126.6Mb	Nasopharyngeal swab

Trim adapters from the reads

The samples were prepared using a SISPA protocol, so the first thing we will do is trim the reads for adapters using the Trim Reads tool and the imported trim adapter list.

1. Run **Trim Reads** by going to:

   Tools | **Prepare Sequencing Data ()** | **Trim Reads ()**

2. Select the sequence lists containing the reads. Click **Next**.

3. Deselect both of the options in the *Quality trimming* step and then click **Next**. For error-prone Oxford Nanopore reads, quality trimming is generally not recommended.

4. In the *Adapter trimming* step of the wizard:

   - Click on the file selector icon () to the right of the **Trim adapter list** field then select the "SISPA adapter trim list" and click **OK**.
   - Click **Next**.

5. Leave options in the *Homopolymer trimming*, *Sequence trimming*, and *Sequence filtering* steps unchecked and click **Next** in each case.

6. In the *Results handling* steps, select the **Create report** option and choose to save the output to a new folder, for example named "Trimmed reads".

Map long reads to reference

In this section, we will map the reads to a closely related strain to create a new consensus sequence. We will use isolate "MT135044" as the reference. This was imported in an earlier step. We stress that this is an example, and any high-quality assembly of a closely related strain will work.
Note: Since we are dealing with metatranscriptomic data, we do not expect all reads to map and coverage may be quite low as a result.

1. Run **Map Long Reads to Reference** by going to:

   Tools | **Resequencing Analysis ()** | **Map Long Reads to Reference ()**

2. Choose to run in batch mode by ticking the **Batch** box and select the trimmed reads. Click **Next**.

3. Leave the *Batch overview* settings as is and click **Next**.

4. In *References*, use the file selector icon () to select the previously imported "MT135044" reference sequence and click **OK** then **Next**.

5. Leave the mapping options set to **Automatic** and click **Next**.

6. Choose **Create reads track** and check the **Create report** option. Choose to save the results in a specified folder and click **Next**. Save the results to a new folder, for example named "Reads tracks".

7. After the tool has finished running, inspect the mapping reports. The table below shows the "Mapped reads" statistics from the report:

   Read mapping report

Run	**Count**	**% reads**	**Average Length**	**# of bases**	**% bases**
SRR10948474	191,114	37.91%	663.11	126,730,061	62.34%
SRR10948550	996	0.23%	702.25	699,445	0.80%

The low percentage of mapped reads for sample SRR10948550 indicates that most reads in this sample are not from SARS-CoV-2.

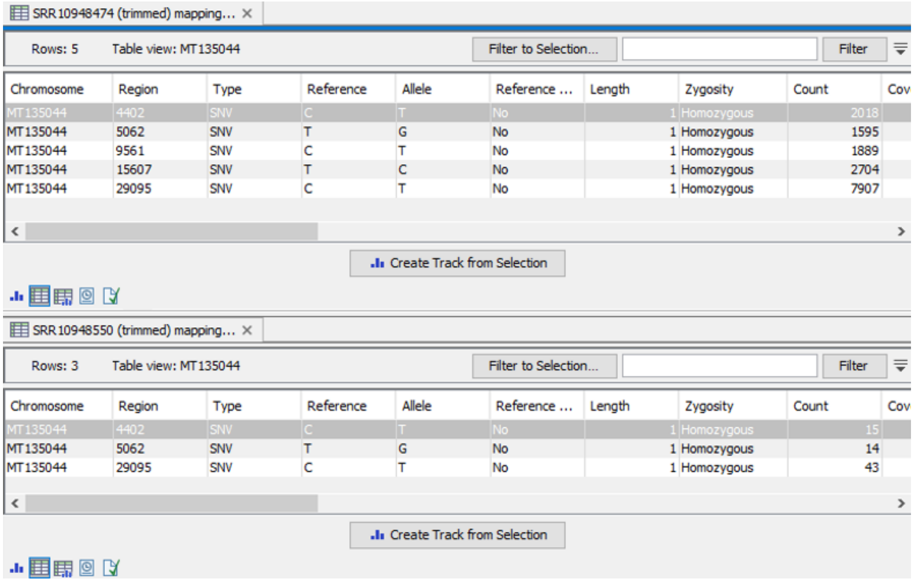
Call variants relative to the reference sequence

Next, we will call variants against the reference by using the Fixed Ploidy Variant Detection tool on the reads tracks.

1. Run **Fixed Ploidy Variant Detection** by going to:

   Tools | **Resequencing Analysis** ( ) | **Variant Detection** ( ) | **Fixed Ploidy Variant Detection** ( )

2. Choose to run in batch mode by ticking the **Batch** box and select the reads tracks. Click **Next**.
3. Leave the *Batch overview* settings as is and click **Next**.
4. Change the *Fixed ploidy variant parameters* to **Ploidy** = 1. Click **Next**.
5. Under *Coverage and count filters*, set **Minimum frequency (%)** = 75.0. Click **Next**.
6. The *Noise filters* can be left as default. Click **Next**.
7. Make sure the **Create track** option is selected. Choose to save the results in a specified folder and click **Next**. Save the results to a new folder, for example named "SARS-CoV-2 variant tracks".
8. Open the output variant tracks to observe the result. Both runs have variant calls relative to the MT135044 reference, three of which are present in both samples. Figure 2 shows the variant tracks in Table View for both samples.



SRR10948474 (trimmed) mapping...

Rows: 5 Table view: MT135044

Filter to Selection...
Filter

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Cov
MT135044	4402	SNV	C	T	No	1	Homozygous	2018	
MT135044	5062	SNV	T	G	No	1	Homozygous	1595	
MT135044	9561	SNV	C	T	No	1	Homozygous	1889	
MT135044	15607	SNV	T	C	No	1	Homozygous	2704	
MT135044	29095	SNV	C	T	No	1	Homozygous	7907	

Create Track from Selection
SRR10948550 (trimmed) mapping...

Rows: 3 Table view: MT135044

Filter to Selection...
Filter

Chromosome	Region	Type	Reference	Allele	Reference ...	Length	Zygosity	Count	Cov
MT135044	4402	SNV	C	T	No	1	Homozygous	15	
MT135044	5062	SNV	T	G	No	1	Homozygous	14	
MT135044	29095	SNV	C	T	No	1	Homozygous	43	

Figure 2: *Table view of the variant tracks. Top: SRR10948474, Bottom: SRR10948550.*

The sampling dates are January 28, 2020 for MT135044 and before January 21, 2020 for SRR10948474 and SRR10948550. This shows the ability to trace mutations over time during an active outbreak.

Viewing tracks in a track list

Now the results can be inspected together by creating a track list. Any track from the same reference (such as the reads tracks and variant tracks we've just created) can be viewed together in this way.

1. Create a track list by going to:

   File | New | Track List ()

2. Select the two reads tracks () and two variant tracks () for both samples and click **Finish**.

3. Keep the track list open for now. You can also choose to save it by pressing CTRL + S (Cmd + S on Mac).

4. Optional: You can include the reference and its annotations by first converting them to tracks. This will also allow for the creation of amino acid tracks.

   (a) Run **Convert to Tracks** by going to:

   Tools | Utility Tools () | Tracks () | Convert Tracks () | Convert to Tracks ( )

   (b) Select MT135044 as input and click **Next**.

   (c) Check **Create sequence track** and **Create annotation tracks**. In **Annotation types**, click on *Edit selection* (+), double-click on "CDS" to select it and click **Done**. Then click **Next**.

   (d) Choose to save the result, click **Next** and save to a new folder, for example named "Tracks".

   (e) You can now drag and drop the reference and CDS tracks into the open track list.

   (f) If you want to include amino acid changes in your investigation, run **Amino Acid Changes** by going to:

   Tools | Resequencing Analysis () | Functional Consequences () | Amino Acid Changes ()

   (g) Choose to run in batch mode by ticking the **Batch** box and select the variant tracks. Click **Next**.

   (h) Leave the *Batch overview* settings as is and click **Next**.

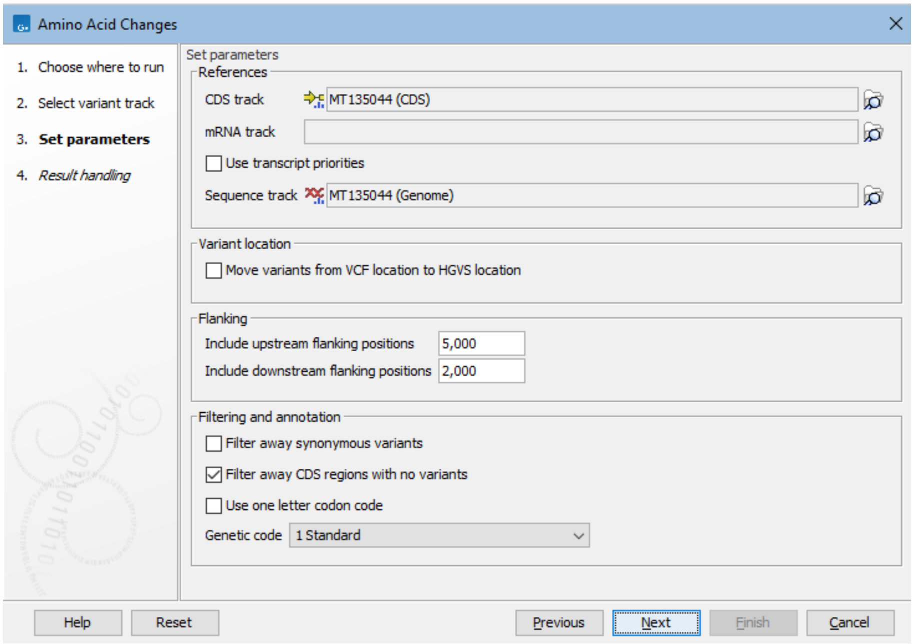
   (i) In *Set parameters* select the CDS and Genome tracks created in previous steps as shown on figure 3. Leave the other settings as default. Click **Next**.

   (j) Make sure **Create amino acids track** is ticked. Choose to save the results in a specified folder and click **Next**. Save the results to your preferred location.

   (k) You can now drag and drop the amino acids tracks into the open track list.

5. The order of tracks in the track list can be changed by dragging and dropping tracks directly in the track list view.

The track list view can be used to investigate the mapped reads, the read coverage across a variant position, the effect on amino acid sequence due to variants, etc. If you zoom in on the variant in position 9561, your track list should look like figure 4. For details on how



Amino Acid Changes

1. Choose where to run
2. Select variant track
3. **Set parameters**
4. *Result handling*

Set parameters

References

CDS track: MT135044 (CDS)

mRNA track:

□ Use transcript priorities

Sequence track: MT135044 (Genome)

Variant location

□ Move variants from VCF location to HGVS location

Flanking

Include upstream flanking positions: 5,000

Include downstream flanking positions: 2,000

Filtering and annotation

□ Filter away synonymous variants

☑ Filter away CDS regions with no variants

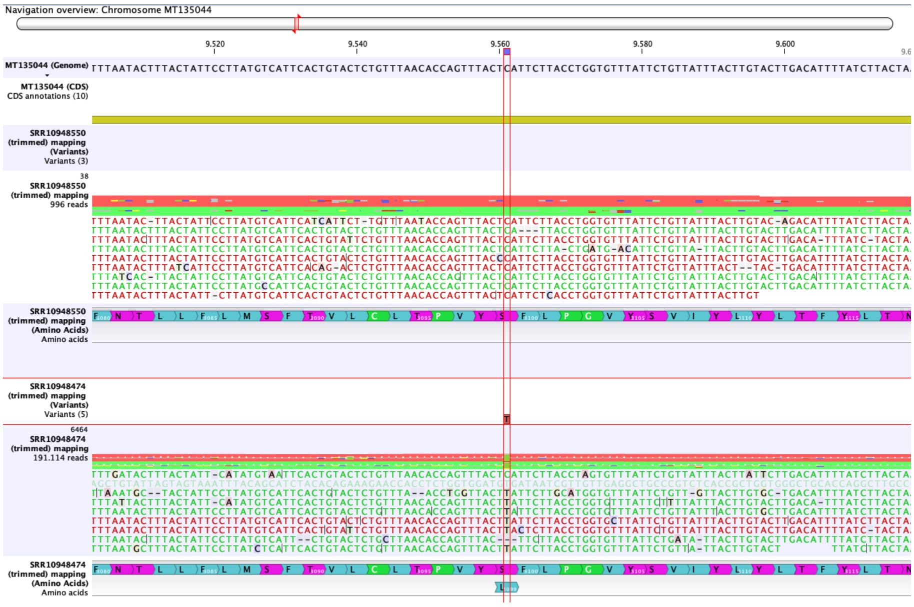
□ Use one letter codon code

Genetic code: 1 Standard

Figure 3: *Amino Acid Changes options.*

to work with tracks, see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Tracks.html.

Tip: As MinION reads have many errors, a more stringent display of the mapped reads gives a better experience. To achieve this, change the settings in the **Side Panel** to the right. Under *Track layout* | *Reads track* | *Hide insertions below (%)* change the value to e.g., 75.



Navigation overview: Chromosome MT135044
9,520
9,540
9,560
9,580
9,600
MT135044 (Genome)
MT135044 (CDS)
CDS annotations (10)
SRR10948550 (trimmed) mapping (Variants)
Variants (3)
SRR10948550 (trimmed) mapping
996 reads
SRR10948550 (trimmed) mapping (Amino Acids)
Amino acids
SRR10948474 (trimmed) mapping (Variants)
Variants (5)
SRR10948474 (trimmed) mapping
191.114 reads
SRR10948474 (trimmed) mapping (Amino Acids)
Amino acids

Figure 4: *Read mapping of MinION samples and variant call visualization.*

Extract consensus sequences from the reads tracks

We have shown how to use an existing reference to call variants. This section focuses on how to use the mapped reads to construct consensus sequences for these samples. The consensus sequences will be used in the next steps to search for similar sequences in a custom BLAST database.

1. Run **Extract Consensus Sequence** by going to:

   Tools | **Resequencing Analysis** () | **Extract Consensus Sequence** ()
2. Choose to run in batch mode by ticking the **Batch** box and select the reads tracks. Click **Next**.
3. Leave the options in *Handle low coverage and conflicts* set to their default values. Click **Next**.
4. Choose to save the results in a specified folder and click **Next**. Save the results to a new folder, for example named "Consensus sequences".

We now have two complete SARS-CoV-2 sequences.

Using BLAST to compare to other SARS-CoV-2 assemblies

The sequences can be used as queries in a BLAST search to find closely related assemblies. For this tutorial, we will run a local BLAST search against a custom BLAST database, which we will create first. It is also possible to download one of NCBI's public databases e.g., Betacoronavirus, for this step, but given the large size of the database and the length of the sequences, the download and search will take several hours.

For this example, we will use the sequence list "Custom Betacoronavirus Database" from the tutorial zip file, which contains a list of Genbank assemblies with the taxonomical genus level "Betacoronavirus".

1. Run **Create BLAST Database** by going to:

   Tools | **BLAST** ()| **Create BLAST Database** ()
2. Select the sequence list "Custom Betacoronavirus Database", which you imported in the beginning of the tutorial. Click **Next**.
3. If you want, you can change the name and description, but it is not necessary. Click **Finish**.
4. Now, run BLAST on the two consensus sequences we generated in the previous step. Run **BLAST** by going to:

   Tools | **BLAST** ()| **BLAST** ()
5. Select the consensus sequence lists. Click **Next**.
6. Choose "blastn" in the *BLAST program* drop-down menu and select the option **BLAST database** as the *Target*. In the selection field below, select the custom Betacoronavirus database. Click **Next**.

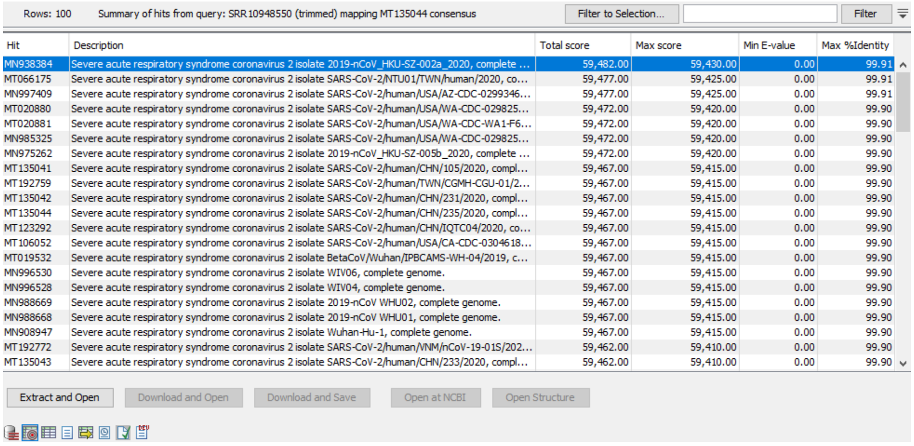
7. We are looking for sequences similar to our query sequence and are only interested in the top hits, so change **Expect** = 0.01 and **Max number of hit sequences** = 100. Leave the other search settings as default, and click **Next**.

8. Choose to **Open** the results rather than saving them and click **Finish**.

9. A Multi BLAST table containing the two queries and the top reported result will open. Take note of the best matching sequence in the *Accession (E-value)* column. For SRR10948550, we expect to see MN938384, and for SRR10948474, we expect to see MN975262. These are the sequences submitted by [Chan et al., 2020] that were assembled based on the same read sets we are using.

   More information about the best matches can be seen in this table by ticking the boxes in the *Show column* section in the **Side Panel**.

10. Select either of the rows in the table by clicking on them and then click the **Open BLAST Output** button in the bottom of the view. This will open an individual table for the selected query. In the bottom left of the new table, you can choose between three different views, *BLAST Graphics*, *BLAST Hit Table*, and *BLAST HSP Table*, which hold analysis values such as Max score, %Identity, %Positive, %Gaps, etc. Figure 5 shows the BLAST Hit Table view for the SRR10948550 query.



Rows: 100 Summary of hits from query: SRR10948550 (trimmed) mapping MT135044 consensus

Hit	Description	Total score	Max score	Min E-value	Max %Identity
MN938384	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-002a_2020, complete ...	59,482.00	59,430.00	0.00	99.91
MT066175	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/NTU01/TWN/human/2020, co...	59,477.00	59,425.00	0.00	99.91
MN997409	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/AZ-CDC-0299346...	59,477.00	59,425.00	0.00	99.91
MT020880	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-029825...	59,472.00	59,420.00	0.00	99.90
MT020881	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-WA1-F6...	59,472.00	59,420.00	0.00	99.90
MN985325	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/WA-CDC-029825...	59,472.00	59,420.00	0.00	99.90
MN975262	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV_HKU-SZ-005b_2020, complete ...	59,472.00	59,420.00	0.00	99.90
MT135041	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/105/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT192759	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/TWN/CGMH-CGU-01/2...	59,467.00	59,415.00	0.00	99.90
MT135042	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/231/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT135044	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/235/2020, compl...	59,467.00	59,415.00	0.00	99.90
MT123292	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/IQTC04/2020, co...	59,467.00	59,415.00	0.00	99.90
MT106052	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/CA-CDC-0304618...	59,467.00	59,415.00	0.00	99.90
MT019532	Severe acute respiratory syndrome coronavirus 2 isolate BetaCoV/Wuhan/IPBCAMS-WH-04/2019, c...	59,467.00	59,415.00	0.00	99.90
MN996530	Severe acute respiratory syndrome coronavirus 2 isolate WIV06, complete genome.	59,467.00	59,415.00	0.00	99.90
MN996528	Severe acute respiratory syndrome coronavirus 2 isolate WIV04, complete genome.	59,467.00	59,415.00	0.00	99.90
MN988669	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU02, complete genome.	59,467.00	59,415.00	0.00	99.90
MN988668	Severe acute respiratory syndrome coronavirus 2 isolate 2019-nCoV WHU01, complete genome.	59,467.00	59,415.00	0.00	99.90
MN908947	Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.	59,467.00	59,415.00	0.00	99.90
MT192772	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/VNM/nCoV-19-01S/202...	59,462.00	59,410.00	0.00	99.90
MT135043	Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/CHN/233/2020, compl...	59,462.00	59,410.00	0.00	99.90

Figure 5: *BLAST Hit Table view for SRR10948550.*

11. You can retreive the sequence of any of the BLAST hits by selecting the hit and clicking on the **Extract and Open** button. You could choose to save the best hit for each sample for further analysis.

We have now identified the known strains most closely related to the samples.

Run the analysis using a workflow

Analysis steps can be put into a workflow, allowing a complex pipeline to be executed simply and reproducibly. To illustrate this, a workflow containing the analysis steps from this tutorial is included in the tutorial zip file. It can be installed using the Workflow Manager, (see `https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Using_workflow_installation_files.html`) and a copy can be opened to inspect the workflow (see https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Managing_workflows.html).

The workflow is set up with **Iterate** and **Collect and Distribute** control flow elements that handles batching of the input MinION reads, including collecting results in sample and combined reports and creating a track list with all sample tracks. Outputs have also been named in a way that puts the results into subfolders. More on creating, editing, and batching workflows, can be found at https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html.

To support the evaluation of the analysis results, the following quality control tools were included in this workflow:

- **QC for Sequencing Reads.** See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkber current/index.php?manual=QC_Sequencing_Reads.html
- **QC for Read Mapping.** See https://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=QC_Read_Mapping.html

Note: Steps relating to BLAST (creating databases and running searches) cannot be done in the context of a workflow, and thus still need to be done separately, as described earlier in the tutorial.

Bibliography

[Chan et al., 2020] Chan, J., Yuan, S., Kok, K., To, K., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C., Poon, R., Tsoi, H., Lo, S., Chan, K., Poon, V., Chan, W., Ip, J., Cai, J., Cheng, V., Chen, H., Hui, C., and Yuen, K. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet*, 395(15):514–523.