CIDR Metagenomics Hub

CIDR

None

Table of contents

1. Introduction

1.1 Metagenomics network hub

The Network Hub is a resource for users of the CIDR clinical metagenomics workflow. Here, you can find SOPs, technical and FAQ/troubleshooting information regarding the implementation of metagenomics in a clinical evaluation/research setting.

1.1.1 Network Sites

1.1.2 Lab protocols

The lab protocol is a same-day DNA/RNA extraction, host-depletion and ONT library preparation workflow for delivery of preliminary sequencing results in < 6 hours.

1.1.3 Informatics workflow

The workflow covers the end-to-end processing of respiratory samples sequencing data, delivering a metageonomic report describing the microbial communities within them. The workflow leverages ONT Nanopore sequencing at its core to produce real-time sequencing data on the GridION platform. The informatics workflow runs locally alongside the sequencing experiment, producing reports as early as 30 minutes after commencing sequencing.

1.1.4 Reporting framework

This SOP is followed to parse results from the informatics workflow for application in a clinical evaluation service setting.

2. Lab Resources

2.1 Panmetagenomics protocol

Note

Please reference use of this method in any presentation or publication as:

Alcolea-Medina, A., Alder, C., Snell, L.B. et al. Unified metagenomic method for rapid detection of microorganisms in clinical samples. Commun Med 4, 135 (2024). https://doi.org/10.1038/s43856-024-00554-3

Important

Risk assessment for handling respiratory samples needs to be performed by each laboratory.

The tools and documentation described here and on the CIDR GitHub are not validated for diagnostic use and are for research and evaluation purposes only.

2.1.1 Version 1.2

2.1.2 Version 1

3. Bioinformatics

3.1 Clinical metagenomics bioinformatics 3.8.1

3.1.1 Introduction

The principal output of the CIDR Metagenomics workflow is a HTML report listing organisms with detectable nucleic acids (RNA/DNA) and some additional information on AMR associated sequence data. The solution packages two applications - CIDR Metagenomics Workflow and Organism Query alongside a few scripts to help manage and analyse outputs. The Metagenomics Workflow runs ontop of MinKNOW, analysing sequencing data in real time producing an easily digested report. Organism Query can be used to scrutinise classifications contained within a report. It leverages the full NCBI nt and RefSeq databases producing a report similar to NCBI BLAST in ~15 minutes. The Organism Query report is designed to provide the user with appropriate information to scrutinise a significant taxonomic classification.

Recent updates

19/05/25 - v3.8.1 now runs Organism Query on all non-viral detections above reporting threshold (including exempt taxa). **See reporting SOP**

Getting started:

1. Install workflow (on first use).
2. Starting a MinKNOW experiment: -MinKNOW (MinION/GridION). -Gourami (>= Q-line V1.1).
3. Start the Metagenomics workflow.

Optional:

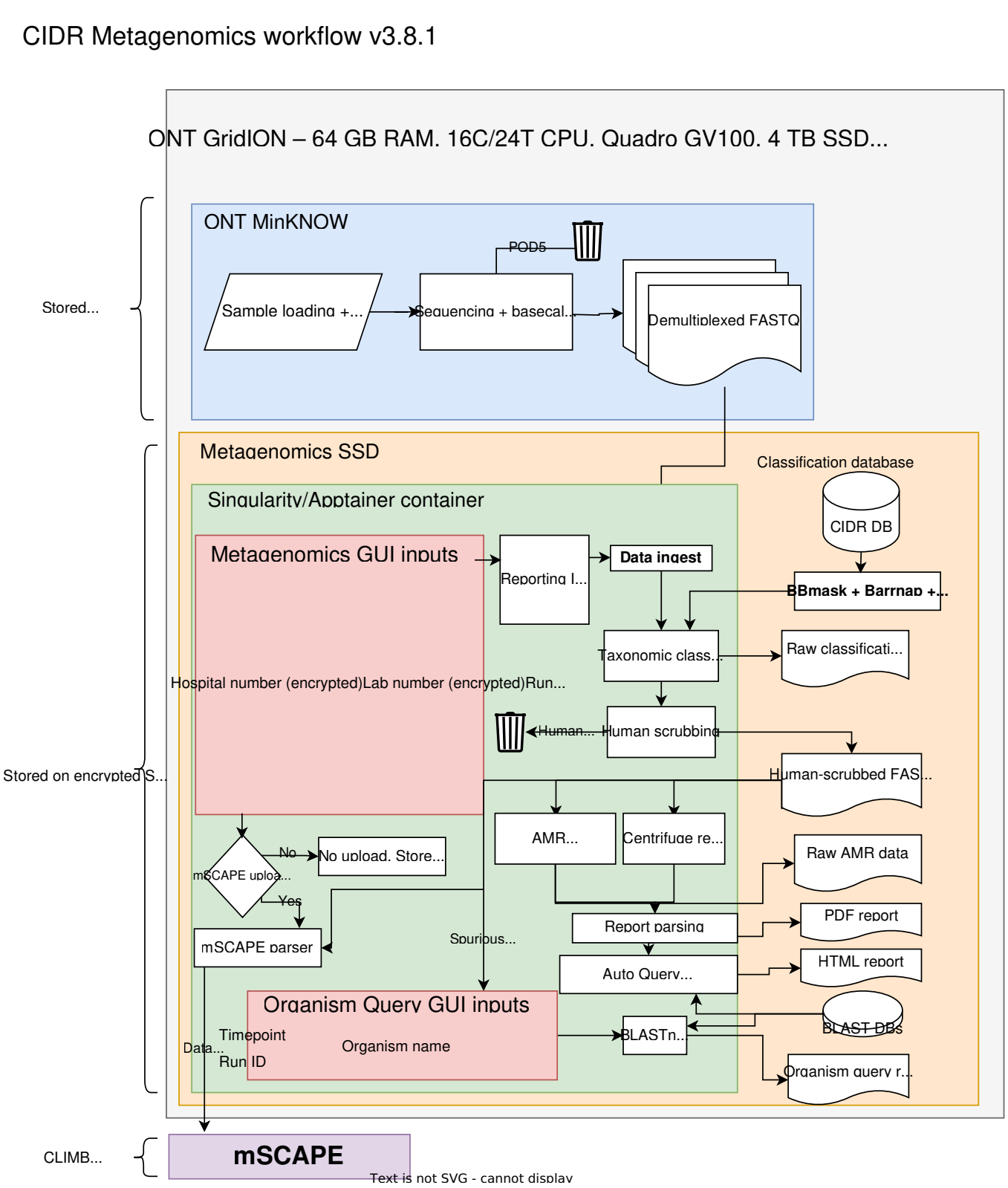
1. Query a classification.
2. Upload data to mSCAPE.
3. Generate a summary spreadsheet.

mSCAPE

Users can opt in to mSCAPE on an per-experiment basis for an automatic upload of sequencing data to UKHSA mSCAPE. No information leaves the sequencing device/ CIDR workflow without expressed case-by-case authorisation.

Technical facets

After loading a metagenomic library on to an ONT sequencing device and launching the sequencing experiment in ONT MinKNOW the pipeline is initialised by the user through the Metagenomics Launcher graphical user interface (GUI). The software periodically ingests base called FASTQ data from the GridION `/data/` directory at set intervals - 0.5, 1, 2, 16 and 24 hours. At each interval, the pipeline performs human scrubbing, taxonomic classification, AMR detection and MLST which is then consolidated in to a PDF reports which are saved in the `/media/grid/metagenomics/results/` directory. The diagram below illustrates further details of how the pipeline works:



Taxonomic classification

At its core, the pipeline leverages the [Centrifuge](https://ccb.jhu.edu/software/centrifuge/manual.shtml k-mer taxonomic classification tool. There are numerous similar methods, the most common being Kraken2. We chose Centrifuge namely

because of its small memory footprint and existing deployment in our legacy workflows. Each read of the raw sequencing data is aligned against an the bespoke Centrifuge database, and is assigned a confidence score and a taxonomy. Our database comprises an optimised database of curated eukrayotic, prokaryotic, and viral reference sequences sourced primarily from NCBI RefSeq and FDA-ARGOS.

The Centrifuge database is QCed using additional taxonomy databases (where available) such as GTDB and scrubbed for regions of significant sequence identity with the our human database.

This database is provided on the SSD sent to each site. We recommend users have > 64 GB of memory available to run the workflow.

Visit the Sequence database for more details on the contents of the k-mer classification and Auto Query BLAST database.

Centrifuge score filtering

Prior to report generation, classifications/reads are subject to a scoring algorithm. This score is a function of read to reference database match length and sequence identity. Score thresholds are currently set at 5000 for all organisms except for viruses for which the score is 250. A function for modifying scoring for specific taxa is built into the workflow. Please do not modify this scoring unless instructed to do so. To add a new filtering threshold for specific taxa, a JSON file in `../metagenomics/db/ref/thresholds` with NCBI taxIDs can be provided (follow the structure of existing examples), then, add the name of the JSON file to ranking.txt file on a new line. This rank determines the priority of lists, given that across lists there may be multiple occurrences of the same taxIDs. Reads which fall below the defined thresholds are not included in the report and appear as ‘Below Threshold’ reads in the ‘Quality Control’ section of the PDF.

Thresholds and priority organisms

Taxa excluding viruses are subject to a 1% relative abundance threshold. This threshold was empirically determined through an assay performance evaluation at GSTT [1]. All classifications without the reporting threshold applied are available at the bottom of the report under the ‘Centrifuge Full’ section.

Specific detections will bypass this 1% threshold, meaning at any relative abundance, taxa will appear in the Above Threshold report section. Consult the Reporting SOP for more information on how to report these taxa. Priority organisms include: Aspergillus spp., Candida spp., Chlamydia spp., Pneumocystis spp. and Mycoplasma spp..

Human scrubbing

Human read removal is required as per the project's ethical approval. A bespoke database of human sequences from RefSeq, T2T CHM13v2.0, NCBI GenBank and HPRC have been screened for microbial sequence, and added to the classification database. Human reads are removed from the analysis workflow on the Metagenomics SSD as soon as they are identified and microbial FASTQ reads are stored in `./metagenomics/results/{sample_id}/{timepoint}/microbial`. Human reads are retained in the raw sequencing (/data) directory on the GridION local storage until deletion.

Antimicrobial resistance prediction

Reads classified as taxa with a relative abundance above 1% are screened using CARD [3] and VFDB [4] databases using ABRicate [5]. The ABRicate outputs are parsed using Scagaire [6], which associates gene predictions with an organism. The parsed outputs of AMR and virulence predictions are added to the PDF report. Common notable detections include ESBL genes, vanA-X, mecA/C. The absence of detected resistance genes ishould not be interpreted as an indication of antimicrobial susceptibility without further validation

3.1.2 Related code snippets

Masking a FASTA database using BBmask

```
bbmask.sh in=unmasked.fasta out=masked.fasta entropy=0.7 -Xmx80g maskrepeats=t
```

Building a centrifuge index -

```
# --bmax needs tuning based on available memory. centrifuge-build -p 10 --conversion-table accession2taxid.map --taxonomy-tree ./taxdump/nodes.dmp --name-
```

3.2 Setting up CIDR Metagenomics bioinformatics workflow

3.2.1 Overview

Each Network site will receive an ONT GridION sequencing platform and an external SSD containing the software and databases required for analysing metagenomic datasets. The software has been designed such that it will be easy for anybody to set up and use. Follow the instructions below to install the bioinformatics workflow.

3.2.2 Install instructions

1. Insert the USB SSD in to one of the blue USB ports at the rear of the GridION. Try to place the disk away from the warm exhaust as this may lead to overheating.

2. After logging in to the GridION Ubuntu operating system, modify the file browser setting by following the video below. This is to enable the running of scripts without using the terminal.

1. Using the file browser, on the taskbar on the left side of the screen, navigate to the **metagenomics** disk, which can be found in the navigation pane inside the file browser.

Info

As a security feature, the removable SSD has been encrypted. Enter the encyrption key provided to you and confirm that you'd like the key remembered.

1. Navigate to the `metagenomics` disk in the file explorer and double-click `launch_installer.sh`, selecting to 'Run in terminal'. When prompted to do so, type the password for the GridION device (not the encryption key). See below for a video guide.

Info

As you type, no lettering or symbols will appear. This is normal. If you mistype, press enter and try again.

There may be additional outputs in you terminal window compared to the video.

1. Some icons should appear on the desktop linking to each app. You will need right click on the icons and select `Allow launching` before continuing with the next stage.

Success!

We have now installed the CIDR metagenomics workflow. The next step will be to run through a control dataset to test the workflow has run sucessfully.

3.2.3 Install validation

Included with the software is a small dataset based on the Zymo community standard. In this step we will validate the function of the workflow with this dataset and generate a report.

1. Double click the Metagenomics Launcher icon on the desktop.

2. Fill out the fields, as indicated in the video below. More information on how to fill the fields and run the launcher can be found in the Starting the metagenomics workflow section.

1. Wait ~10 minutes for the workflow to complete. Open up the PDF report which can be found in the `reports` folder on the metagenomics disk in a folder corresponding to the name of the sample provided in the launcher eg. gstt_control_1. See video below for further information.

1. Inspect the `/metagenomics/reports/validation_sample/` PDF report at the **0.5 hr timepoint**, it should match the CIDR validation report provided here.

Success!

We have now tested the CIDR metagenomics workflow. The next step will be to run a sequencing experiment, running the workflow in real-time.

3.3 Launching a sequencing experiment using ONT software

3.3.1 Starting a sequencing experiment in MinKNOW/Gourami

Introduction

Recent updates

For users of Q-line V1.1 or later, ONT have replaced MinKNOW with the Gourami sequencing software. Go to the Starting a MinKNOW/Gourami experiment for instructions on using that.

Jump to the Gourami section for instructions on how to start a sequencing experiment.

This document instructs users on how to use ONT MinKNOW and Goruami interfaces. These tools are used to initialise and control sequencing experiments on the GridION device. Following completion of the relevant section, proceed to the Running the metagenomics workflow section to start the metagenomics analysis.

For MinKNOW users (GridION RUO/Q-line V1.0) users follow Running a flow cell check and Starting a sequencing experiment on MinKNOW sections.

Important

Before completing any protocols, users should check flow cells sent to them by ONT are above the warranty pore count. The pore check should be run:

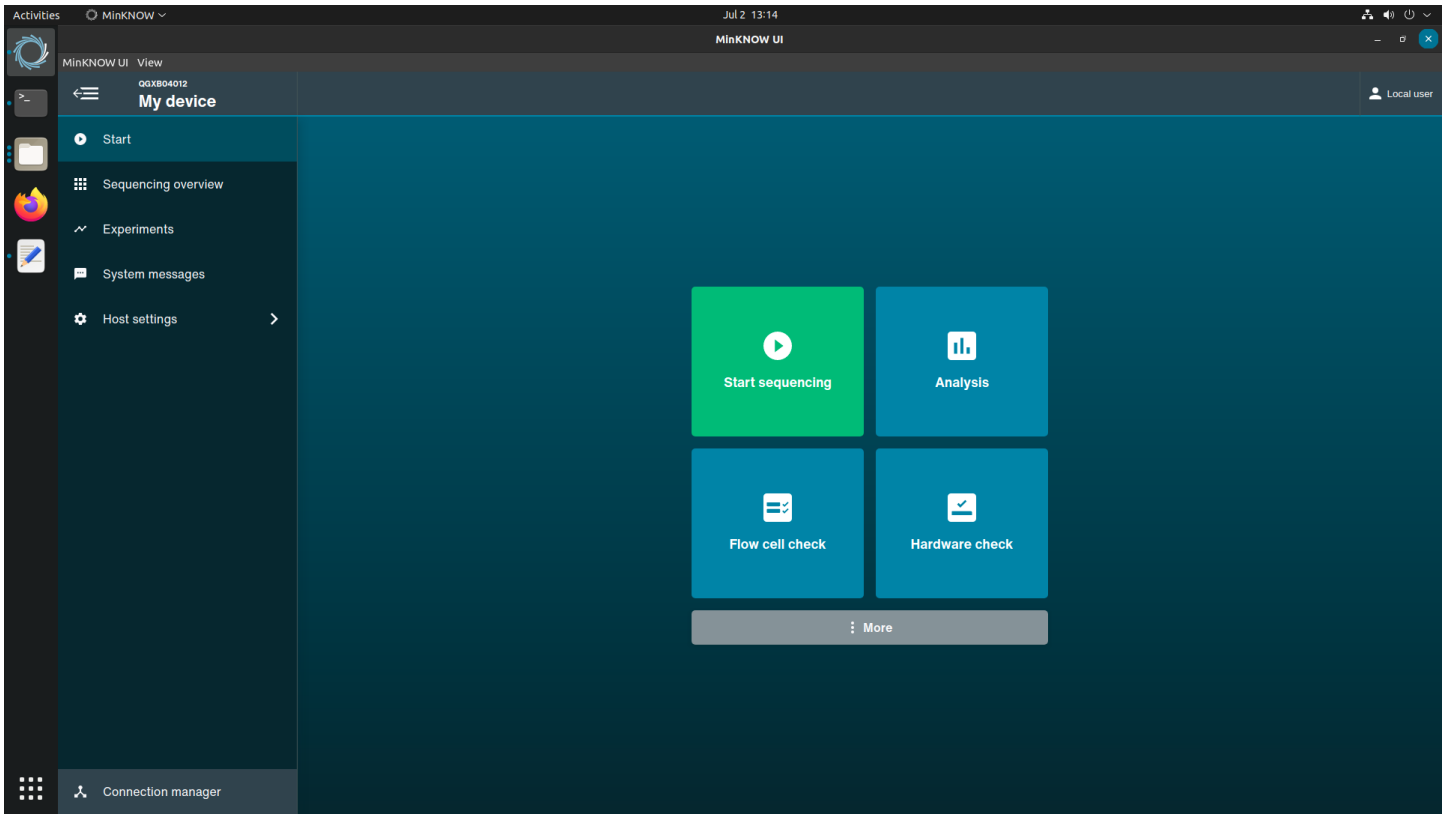
1. As soon as you have recieved a new batch from ONT.
2. Immediately before starting a sequencing experiment.

Important

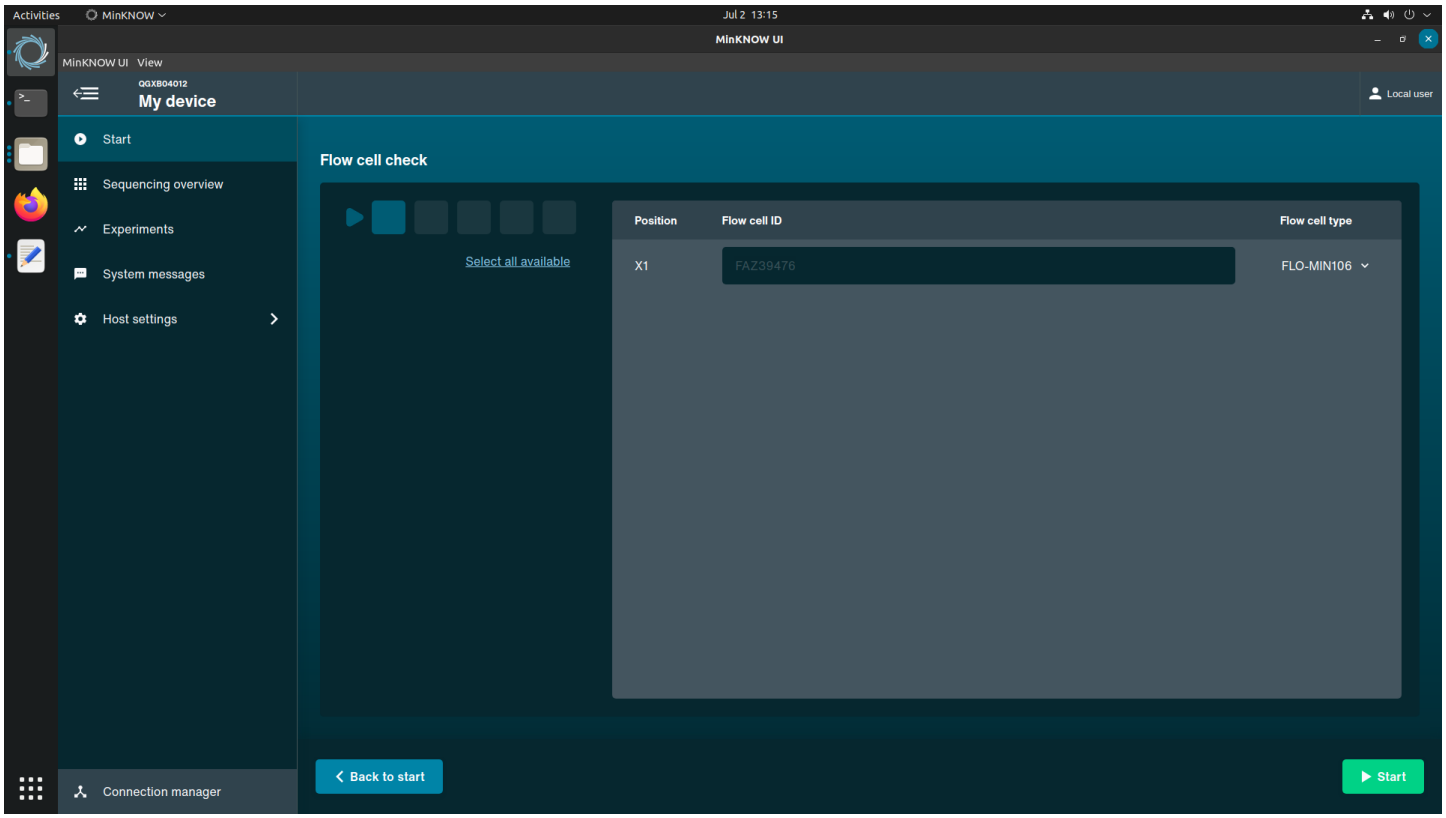
Only **after** starting a sequencing experiment and it has begun generating reads can you commence analysis using the CIDR workflow.

Running a flow cell check

Oxford Nanopore Technologies will replace any flow cell that falls below the warranty number of active pores within three months of purchase, provided that you report the results within two days of performing the flow cell check and you have followed the storage recommendations. A MinION flow cell (used also in the GridION) should have 800 pores.



1. Select **Flow Cell Check** from the MinKNOW Start screen.
2. Indicate the corresponding sequencing positions you'd like to check by selecting the square icons below the 'Flow Cell Check' title. (See image below)
3. Click on the green Start button and wait for the flow cell check to complete.
4. If the pore count is < 800 and the flow cell is still in warrant, contact ONT for a replacement within two days of completing the check.



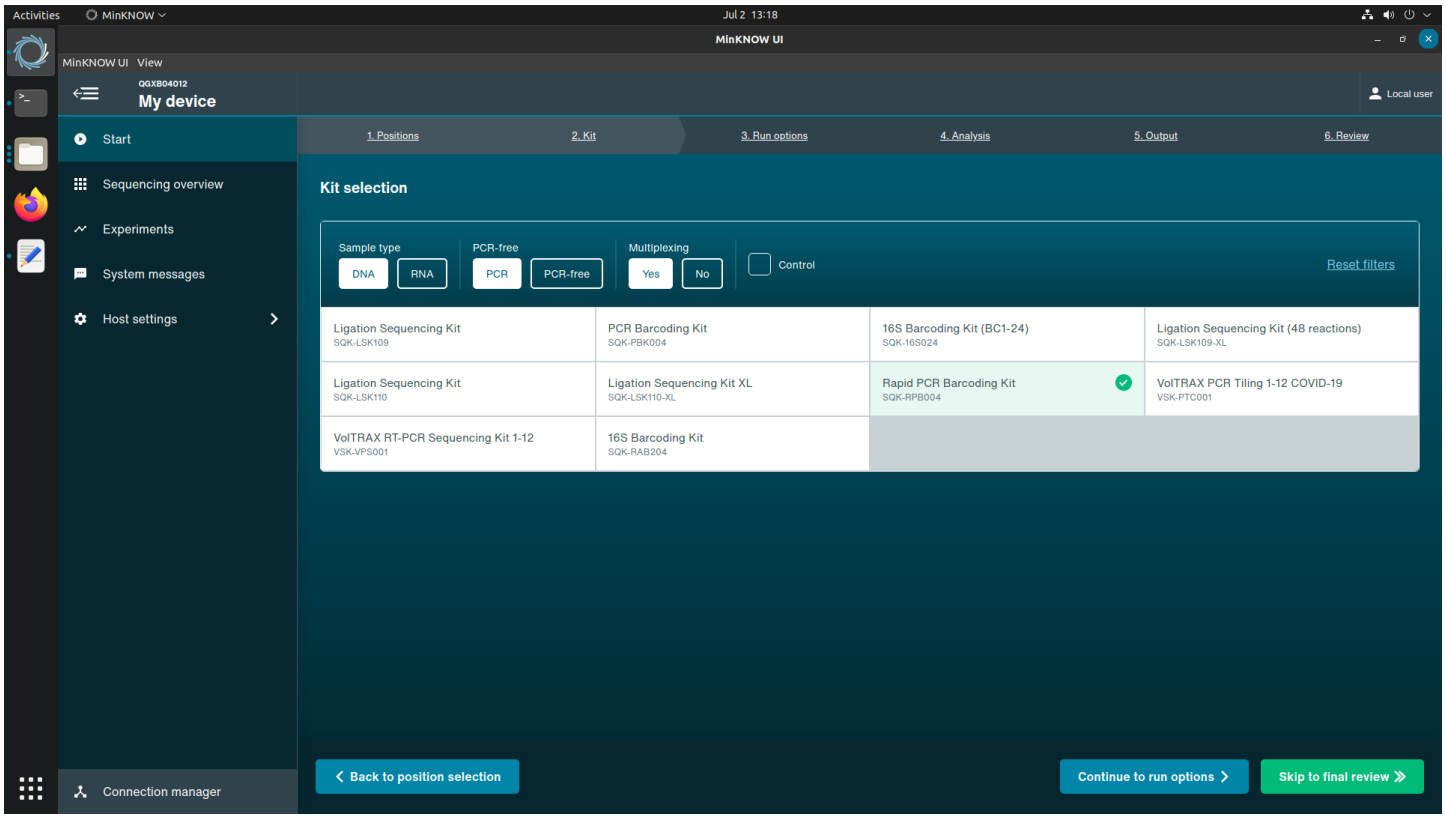
Starting a sequencing experiment on MinKNOW

1. From the Start screen in MinKNOW select 'Start Sequencing'.
2. Select the position occupied by the flow cell loaded for the sequencing experiment, enter the Experiment and Sample IDs and select 'Continue to kit selection' at the bottom of the screen.

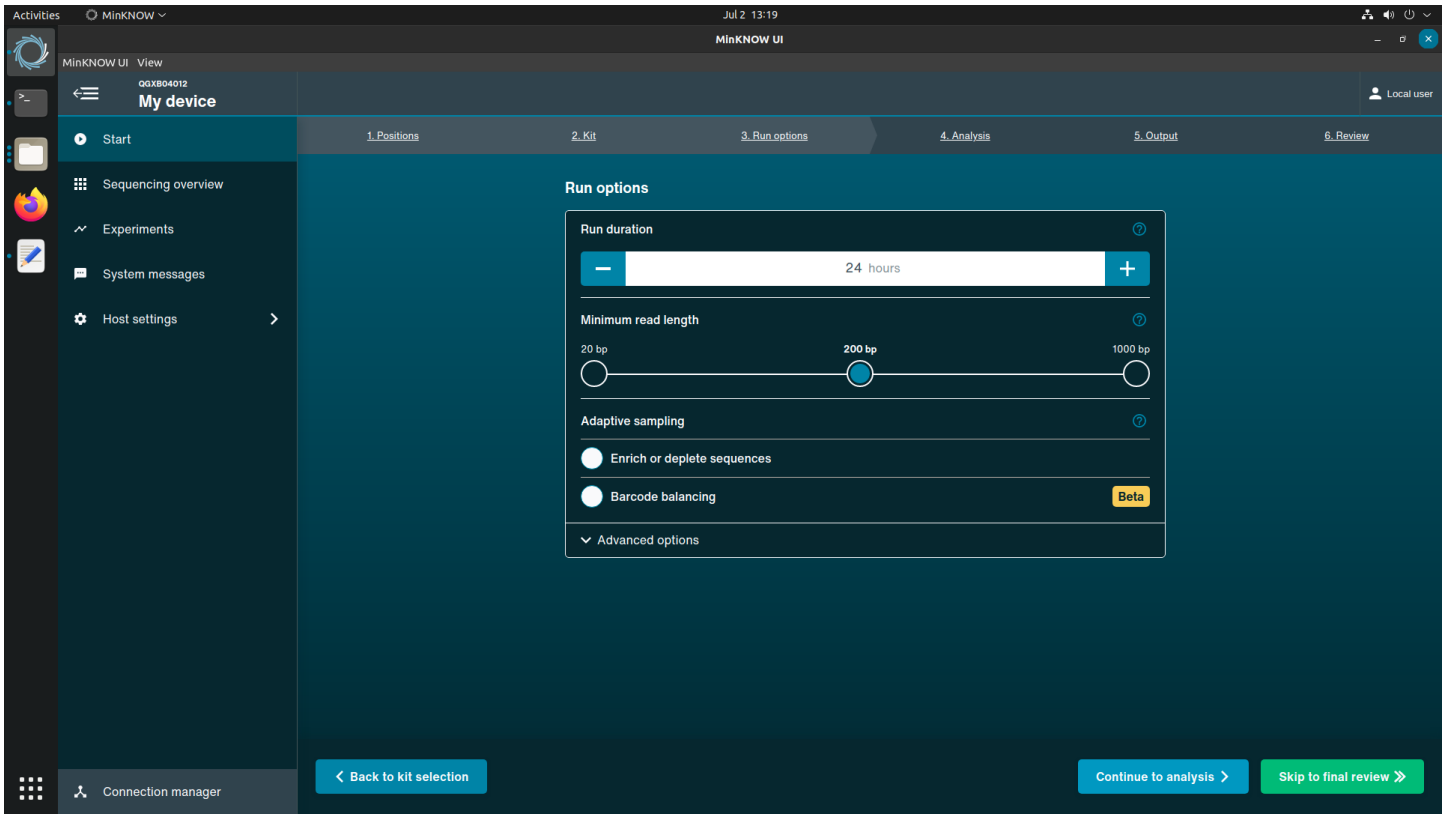
Note

You should enter a new and unique Experiment ID and Sample ID for each new sequencing library.
A summary of the configuration parameters is shown at the bottom of this section.

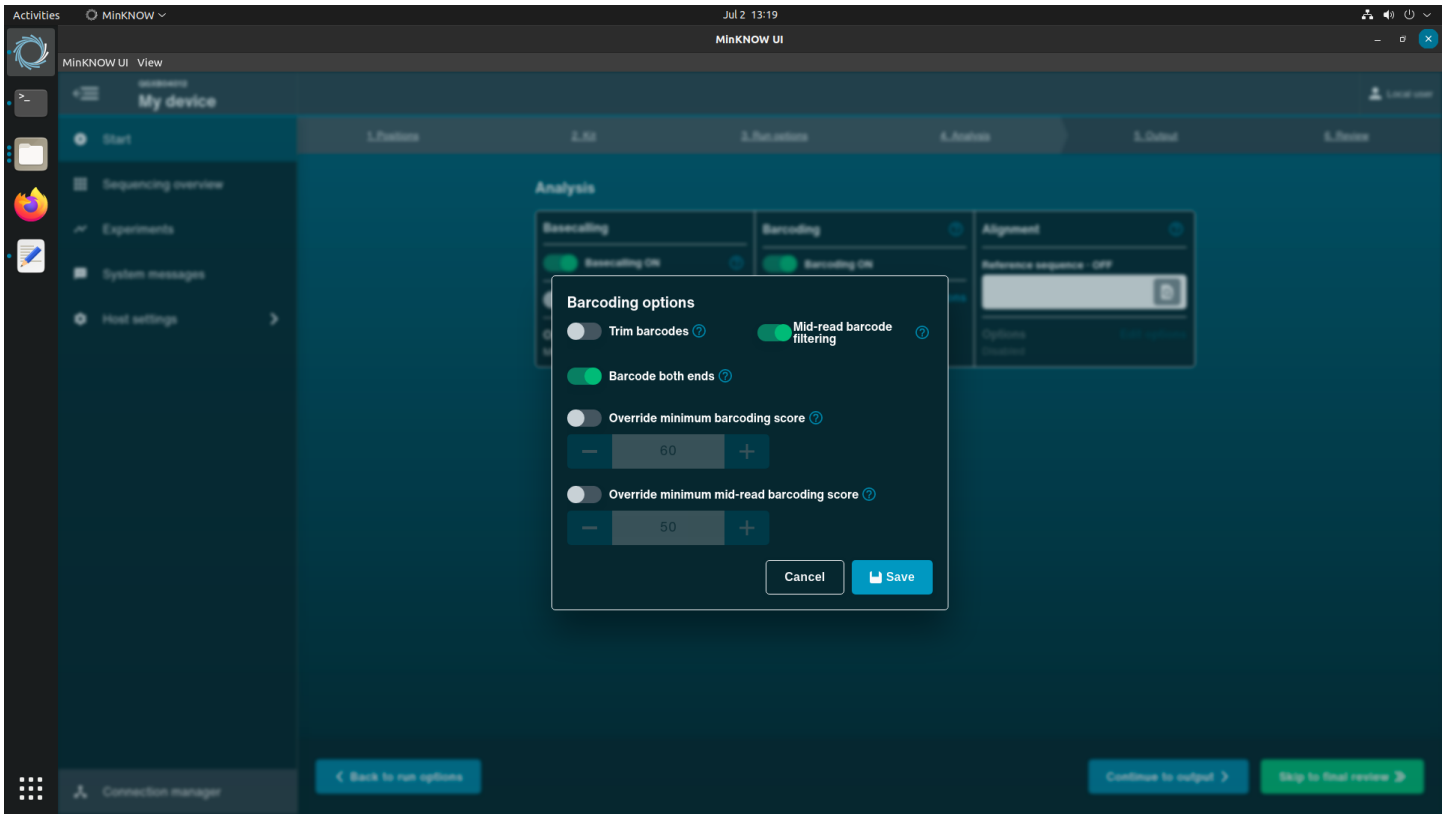
1. Select the RPB-004 library preparation kit from the Kit selection screen. Click 'Continue to run options' at the bottom of the window.



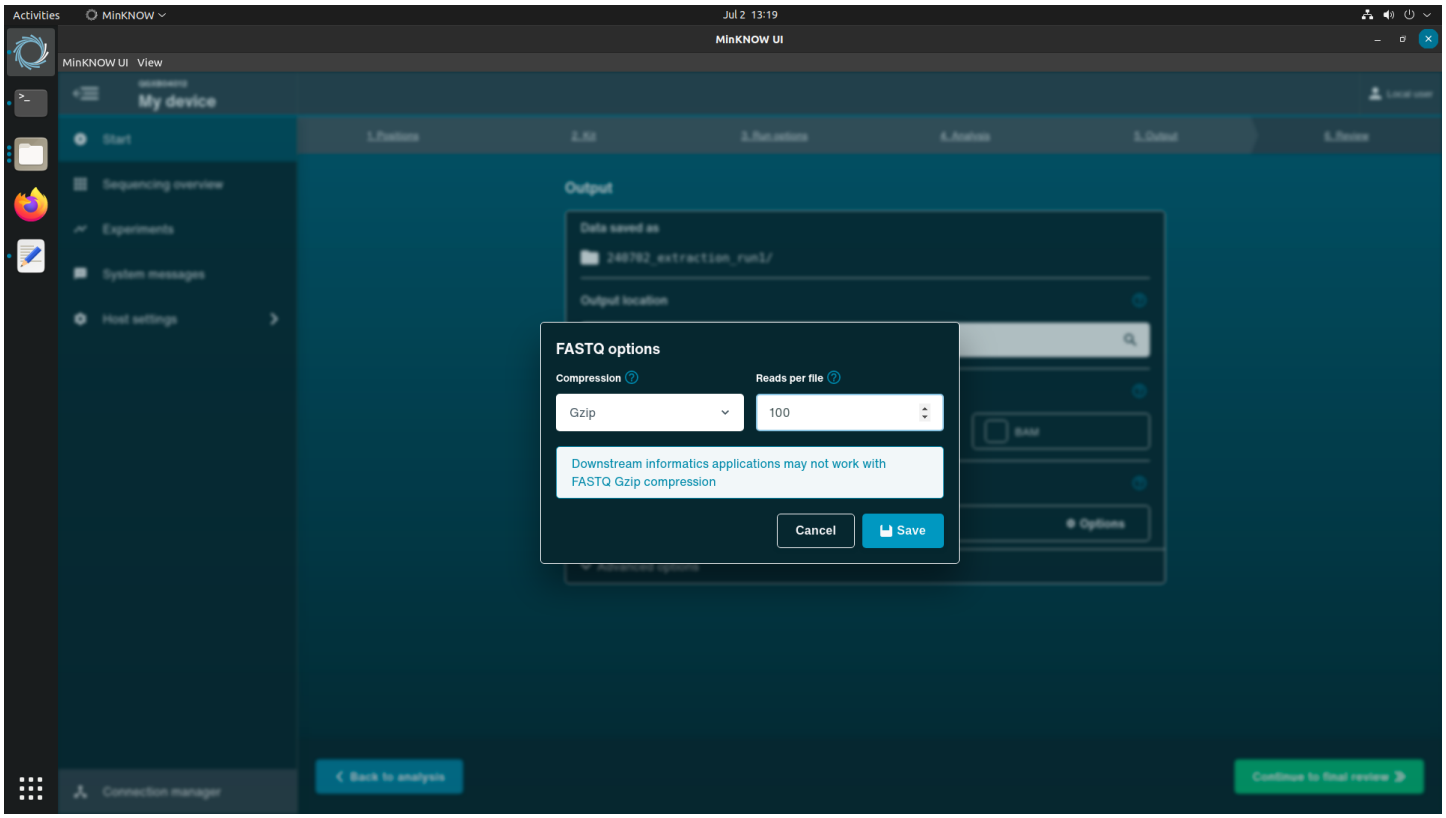
1. In the run options screen, set the sequencing experiment to last for 24 hours. Leave the read length at 200 bp and the other settings as default and select 'Continue to analysis' at the bottom of the window.



1. On the Analysis window, under barcoding, select 'Edit options'. In the popup window, select 'Barcode at both ends' and 'Mid-read barcode filtering'. Select the 'Continue to output' button at the bottom of the window.



1. On the Output window, deselect the FAST5 option. Where the FASTQ checkbox is selected, click the gear icon and set 'Reads per file' to 100. Continue to final review.



1. Check the parameters below match what is indicated on-screen.

Parameter	Value
Selected Kit	SQK-RPB004
Run length	24 hours
Minimum read length	200 bp
Adaptive sampling	Off
Basecalling	On (High accuracy basecalling)
Barcoding	On
Require both ends	On
Detect mid-read barcodes	On
Alignment	Off
Location	/data
FAST5	Off
FASTQ	On (Gzip, 100 reads per file)
Read filtering	Qscore:9 Readlength: unfiltered, Read splitting: Disabled

1. Start the sequencing experiment. After the flow cell reaches temperature, navigate to the barcodes screen to verify data has been output.

Success!

After reads start to appear on the barcoding screen you can advance to Starting the metagenomics workflow.

3.3.2 Starting a sequencing experiment in Gourami

Introduction

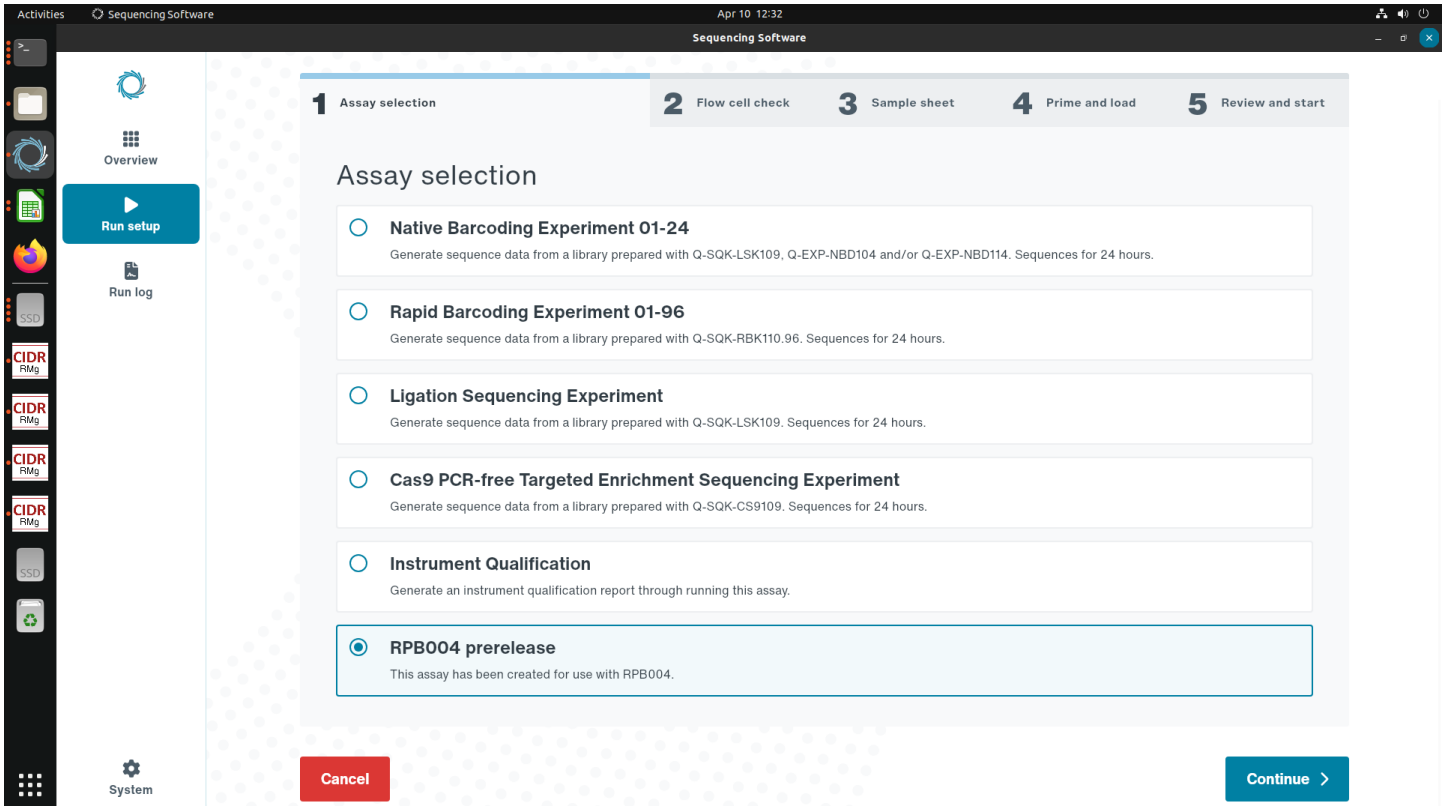
Important

This section applies only to users with GridION version 1.1 or greater. For users with MinION or GridION RUO/1.0 platforms, please refer to the MinKNOW sections.

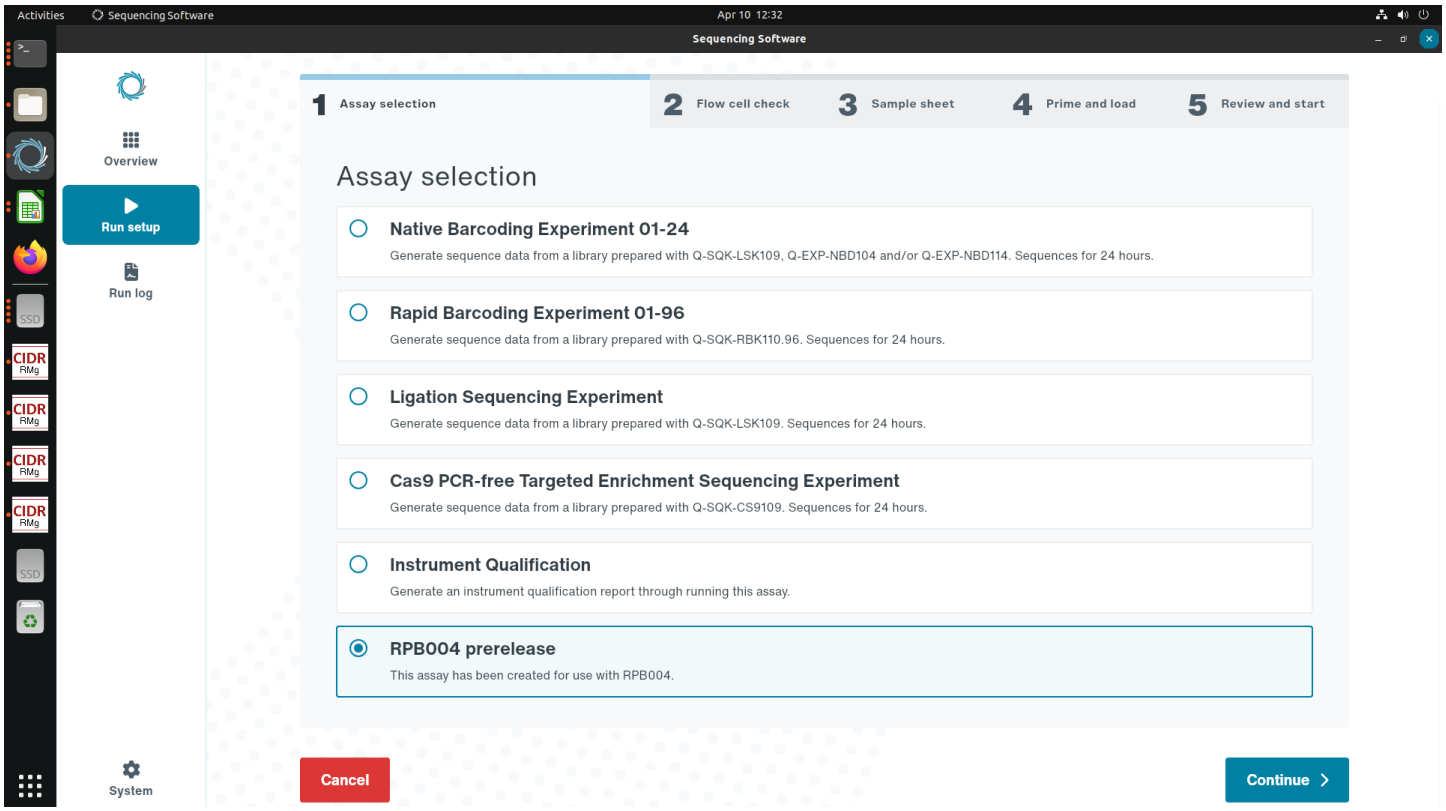
This document instructs users on how to use the Goruami interface. These tools are used to initialise and control sequencing experiments on the GridION device. Following completion of the relevant section, proceed to the Running the metagenomics workflow section to start the metagenomics analysis.

Launching a sequencing experiment using Gourami

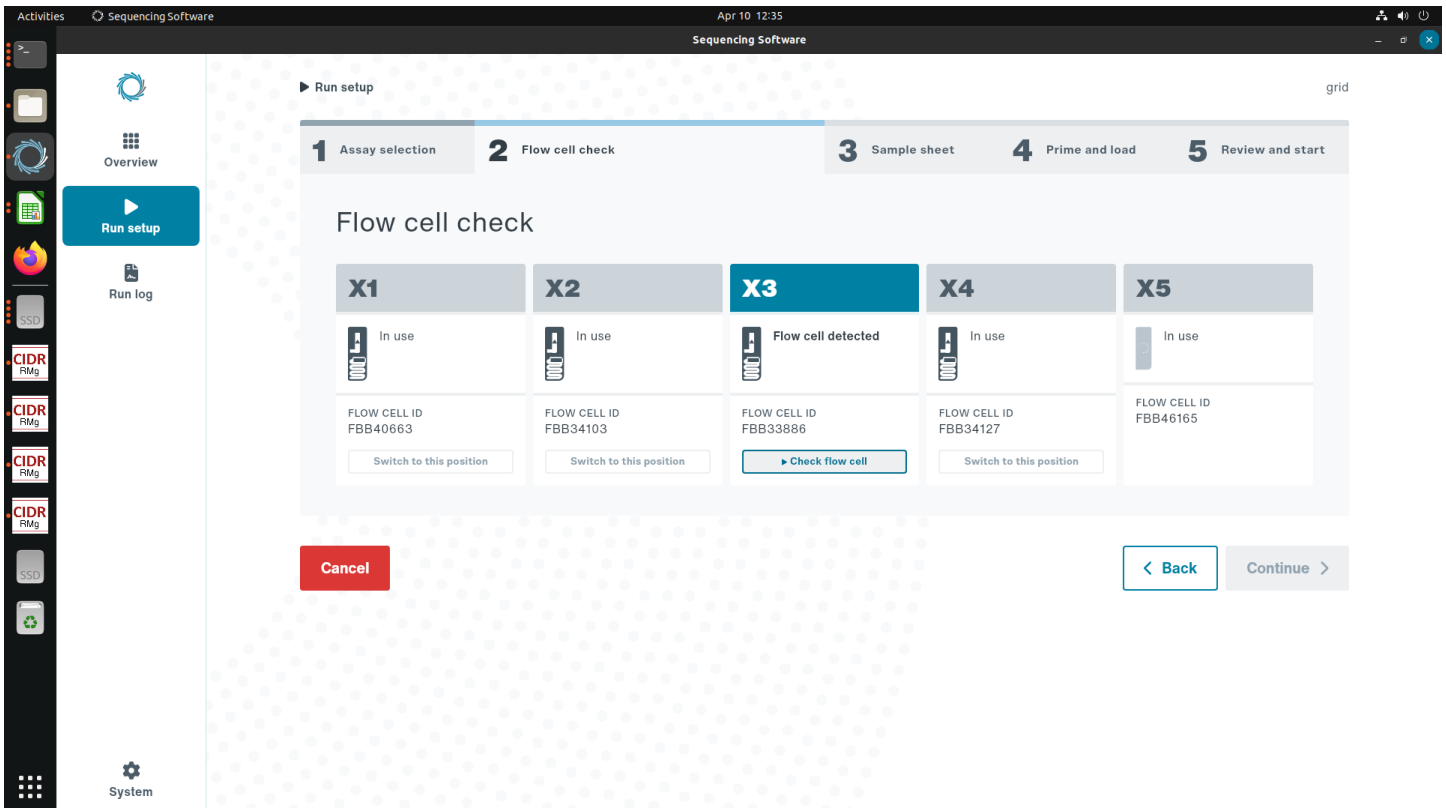
1. Open the Sequencing Software by selecting the icon on the taskbar to the right as shown in the screenshot below.



1. From assay selection, select “RPB004 prerelease” and press continue.



1. Select the position of your flow cell on MinKNOW. If you have not already run a flow cell check, you will have to run that now by selecting “Check flow cell”. After it is complete you can press continue.

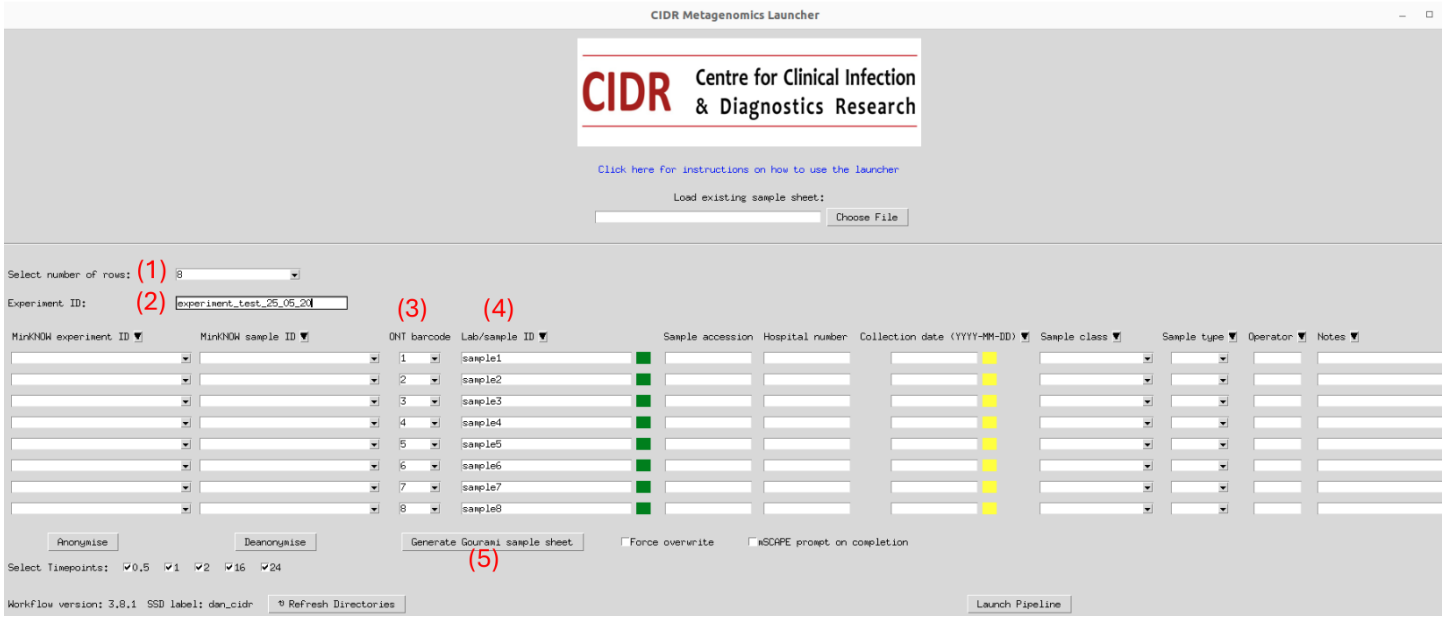


Tip

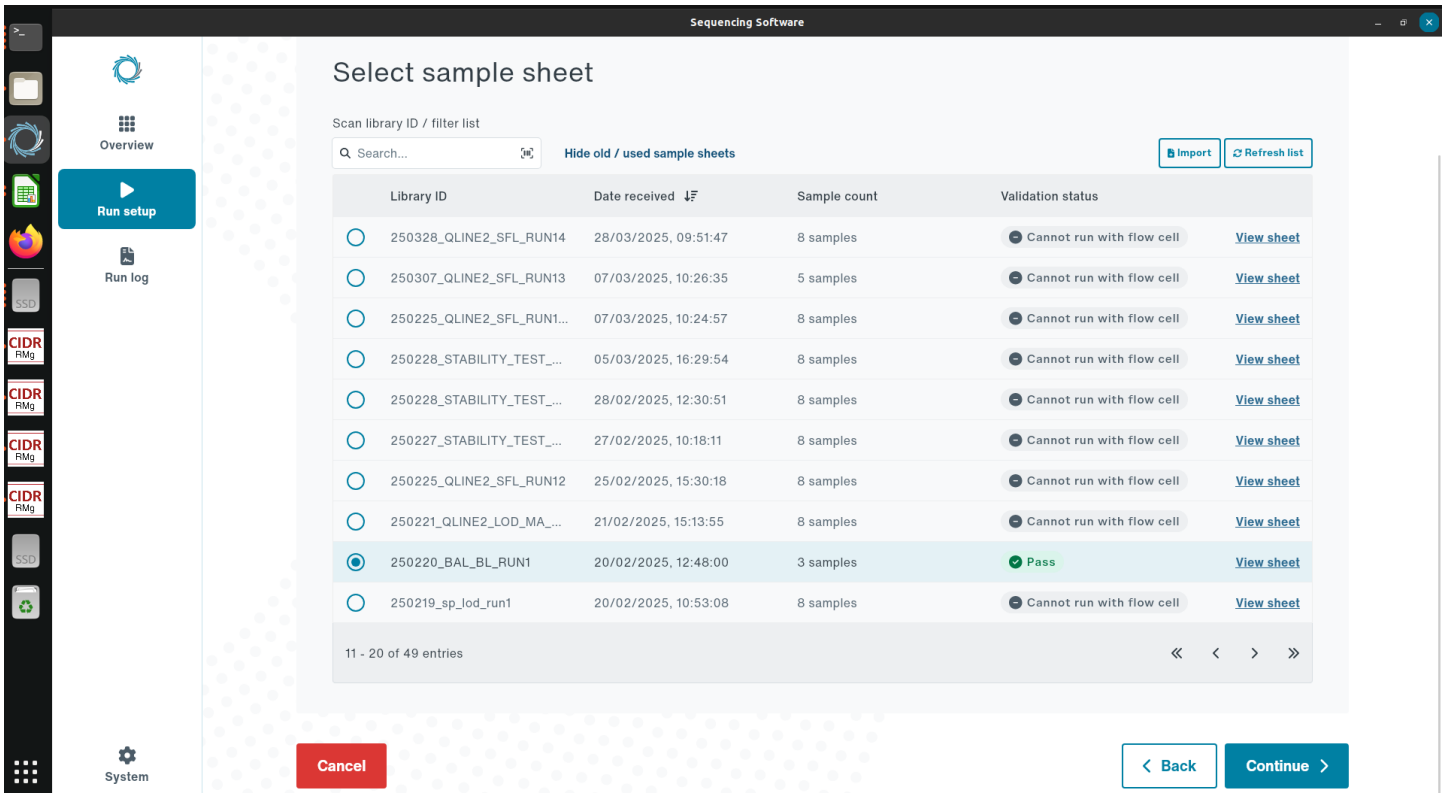
The sample sheet step is only necessary if you are using the 'Gourami' software.

1. Open the Metagenomics launcher application and enter:
    1. Number of samples (rows)
    2. Experiment ID
    3. 'ONT barcode' (for each sample in the library)
    4. 'Lab/Sample ID' (for each sample in the library)

After filling out the above fields, click **Generate Gourami sample sheet (5)**. This will create a sample sheet in the 'sample_sheet' directory in 'metagenomics'. The sample sheet will be named with the date/time and the text in the Experiment ID field.



1. Switching back to the Gourami sequencing software, select “Import” on the select sample sheet screen and import the sample sheet you have just created. Then make sure it is selected and the validation status says “Pass”, then press continue.



1. On the “Prime and load” screen ensure flow cell is correctly loaded as per SOP and press continue.
2. On the final “Review and start” screen, select “Run assay” to begin the run. Please wait all the sample barcodes to appear as found before starting the metagenomics pipeline.
3. Proceed to the Running the metagenomics workflow section to start the analysis.

Success!

After reads start to appear on the barcoding screen you can advance to Starting the metagenomics workflow.

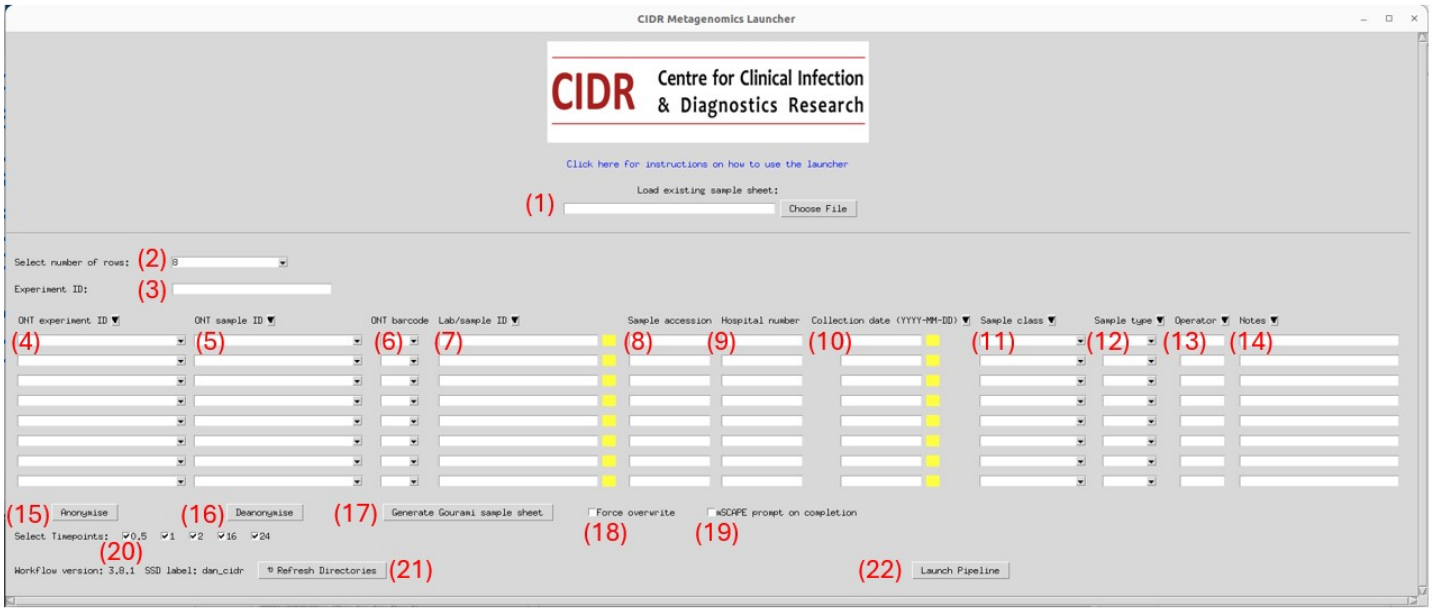
3.4 Running the CIDR metagenomics workflow

Before starting

1. The CIDR metagenomics workflow must only be executed during a sequencing experiment or after a sequencing experiment has completed. The pipeline must not be launched before a sequencing experiment has been commenced and has **started producing reads**.
2. Ensure the SSD is inserted in to one of the rear USB 3.1 ports, has been mounted and the encryption key has been entered successfully. Test the disk has been mounted by navigating to it in the Ubuntu file explorer.

3.4.1 Starting a run

1. Double click the **Metagenomics Launcher** icon on the GridION desktop, the CIDR Metagenomics Launcher should appear alongside a terminal window.



Known issues

The `'geocryptfs error not found...'` error can be ignored as it is not essential to the workflow.

For all clinical specimens, positive controls and negatives, a unique Lab/sample ID must be used, taking in to account across all previous runs. For positive controls, for example, add the date to each Lab/sample ID (POS_25_01_18) - you can not use only 'POS' as this will overwrite the previous run's control sample with the same name.

3.4.2 Feature descriptions:

Number	Field	Description
1	**Load existing sample sheet**	Load a pre-existing sample sheet. This will populate the fields below with the data from the TSV file.
2	**Number of samples**	The number of samples to be analysed. This will create the number of rows in the table below.
3	**Experiment ID**	Not to be confused with the ONT
4	**ONT experiment ID**	The exact name matching the experiment name on MinKNOW entered by the user when initiating a sequencing run. This is populated automatically from the `/data` directory.
5	**ONT sample ID**	The exact name matching the Sample name on MinKNOW entered by the user when initiating a sequencing run. This is populated automatically from the `/data/{experiment_id}/`
6	**ONT barcode**	The ONT library index/barcode used. Green colour indicates the barcode directory has been validated.
7	**Lab/Sample ID**	The unique lab accession number for the sample. This data is encrypted before transmission. **If repeating a sample, append with _n**
8	**Sample accession**	The lab's sample ID - identifying a specific patient specimen - ANONYMISED.
9	**Hospital number**	A value identifying the individual providing the sample - ANONYMISED.
10	**Collection date**	The date the specimen was collected. For positive and negative controls, this would be the day of library preparation.
11	**Sample Class**	The category of the sample loaded.
12	**Sample type**	The type of specimen.
13	**Operator**	Identifier for user operating the sequencer.
14	**Notes**	An open field for notes that will appear on all reports.
15	**Anonymise**	Anonymises the 'Sample accession' and 'Hospital number' values using an encryption cypher.
16	**Deanonymise**	Deanonymises the 'Sample accession' and 'Hospital number' values present in the launcher fields to their original values. The deanonymisation tool can be used to access previous runs.
17	**Generate Gourami sample sheet**	**Only for Q-line >=v1.1** Generates a Gourami compatible sample sheet for starting a sequencing experiment. The output can be found in the `./metagenomics/sample_Sheet/gourami` directory.
18	**Force overwrite**	Checking this box will move results and reports for all timepoints matching the 'Lab/sample ID' filed in the the launcher to the `./metagenomics/recycle_bin` directory and 'unlock' all directories. If you have aborted a run, or the terminal is reporting failures, try using this feature.
19	**mSCAPE prompt**	After the sequencing and analysis run has completed, open the mSCAPE uploader for user input. No data is uploaded without par-sample expressed authorisation.
20	**Select timepoints**	Select the timepoints you'd like to be generated. If you encounter errors generating a timepoint visit the FAQ section
21	**Refresh directories**	This button refreshes the contents of the MinKNOW experiment ID and MinKNOW sample ID columns. Useful if you have started the launcher before commencing the sequencing experiment.

Number	Field	Description
22	**Launch pipeline**	Launches metagenomics analysis, saving the sample sheet to the `./metagenomics/sample_sheets`.

1. Select the number of samples to be analysed from the dropdown (Feature 2).
2. Add an appropriate experiment ID to the field (Feature 3). This will label the stored sample sheet, useful for auditing. For Gourami users, this will set the ONT experiment ID.

Note

- Launching an analysis run will save all of the data in the launcher fields to a TSV file in the `metagenomics/sample_sheets` directory, with the date/time and the contents of the Experiment ID field. This can be reused if a repeat run is required - or quick edits need to be made to a set of samples without having to fill out the fields again.

1. Select the experiment and sample ID corresponding to the MinKNOW/Gourami run from the ONT Experiment ID and ONT sample ID dropdown menu (Feature 4). If your experiment is not there, click the refresh button (Feature 21).
2. Complete the remaining fields using the table above as a guide. The NHS service evaluation and mSCAPE protocols require that all fields are completed where appropriate.
3. Where appropriate, apply the anonymisation functions to the data using the 'Anonymise' button. This will pseudo-anonymise the 'Sample accession' and 'Hospital number' fields. Users can deanonymise a sample sheet by loading it in to the lanucher and selecting 'Deanonymise' or by using the Deanonymisation tool.
4. Click on `Launch pipeline` and follow the instructions to start the analysis.
5. After a minute, the terminal window accompanying the workflow launcher should start displaying log outputs from the workflow. See below for an example.
6. ~35 minutes after launching the sequencing experiment alongside the metagenomics workflow, the first reports will be available in `/media/grid/metagenomics/reports/{sample_name}/`. See below for a guide on how to access this.

Success!

The workflow will run for ~24 hours generating reports for 0.5, 1, 2, 16, 24 hour time-points. If you are using a previously sequenced dataset, reports will be generated as quickly as possible.

3.5 Bioinformatics - Organism query

The Organism Query tool is designed to help scrutinise taxonomic classification outputs from the CIDR metagenomics workflow. It uses a local (offline) version of NCBI BLASTn, with the full NCBI nt, RefSeq, and CIDR databases to produce a report, similar to that found on the NCBI BLAST website, providing the user with a second opinion on classifications.

3.5.1 Technical information

- The tool searches the classified reads for an organism indicated by the user. Selecting a random subset of 50 (max) reads assigned to that taxa.
- The centrifuge score is not used in this analysis, all reads matching a taxon are valid for selection
- Reads are extracted from the microbial FASTQ file stored in the workflow `results` folder and BLASTed against the prescribed database. The subsetted reads are saved as a FASTA file in the `./metagenomics/reports/{sample_id}/organism_query_XXX` directory.
- The results are parsed in to a HTML report with an interactive plot designed to help the user explore the different metrics of alignment.
- A full BLAST alignment report is available at the bottom of the HTML report.
- The report is stored in the `./metagenomics/reports/{sample_id}/organism_query_XXX` folder, with the other PDF reports from the metagenomics run.

3.5.2 Launching organism query

1. Load the relevant report for the run you'd like to query.

2. Click on the Organism Query launcher desktop icon.

3. Fill out the fields detailed below. Multiple queries can be run at once by selecting the `+` button. A separate report will be generated for each.

Parameter	Description
CIDR workflow Lab/sample ID	The Lab/Sample ID matching that of the report in question.
Workflow hour/ interval	The time-point corresponding to the dataset you'd like query.
Organism keyword	A keyword identifying the taxa to be queried eg. 'Aspergillus' (capturing all aspergillus spp.) or 'Bordetella parapertussis' for this species and all taxonomic children (eg strains)

1. Click on the `Launch script` button to start the query workflow. A Firefox browser window will appear after the workflow has finished. You can reopen the report from `` `reports/{sample_id}/organism_query_XXX `` on the metagenomics SSD.

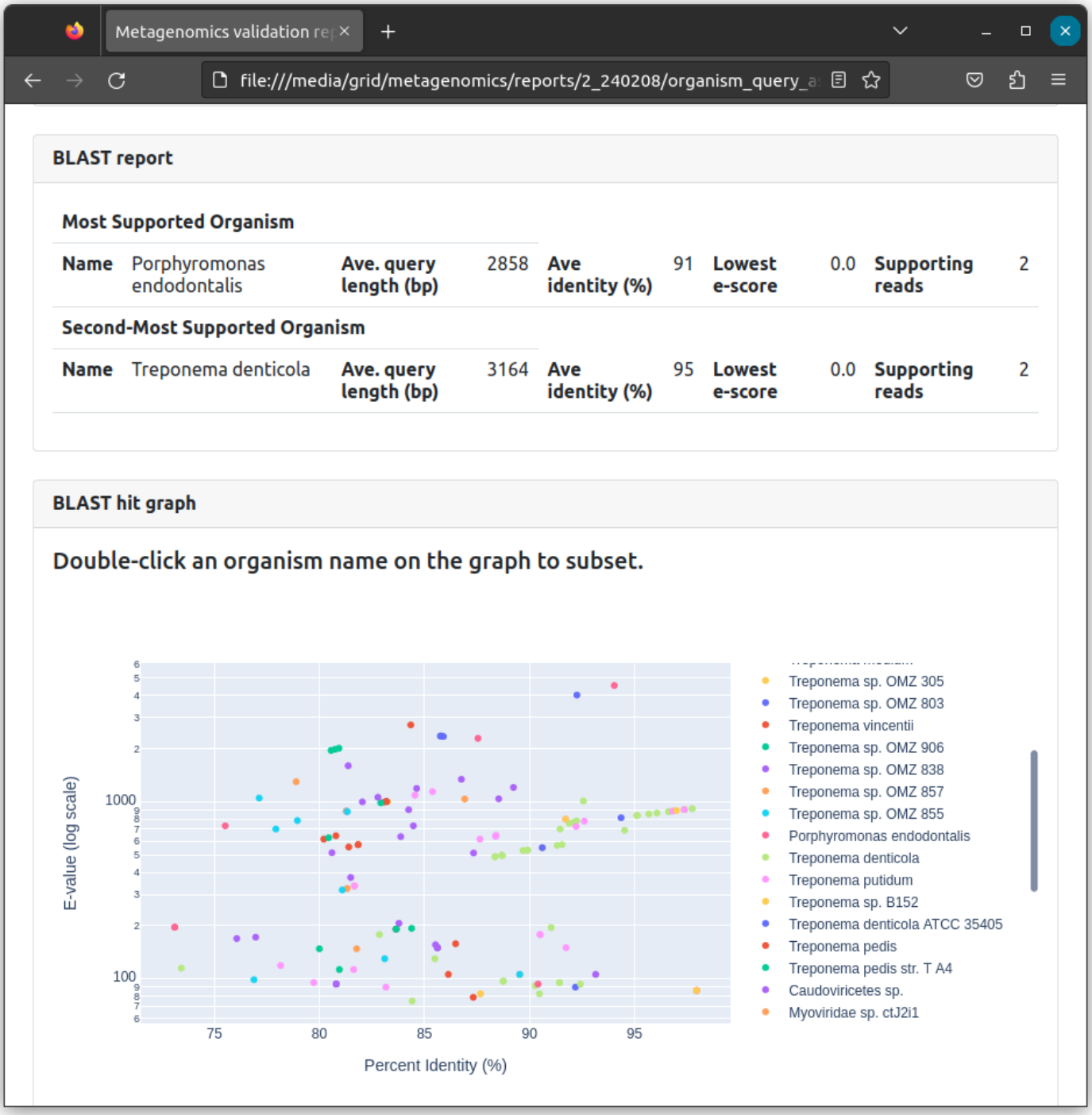
3.5.3 Interpreting results

Interpreting the results can be subjective. The interactive plot has been designed to help guide decision making. Using a combination of the alignment length (relative to the read length), the identity score and the e-score can be an informative approach.

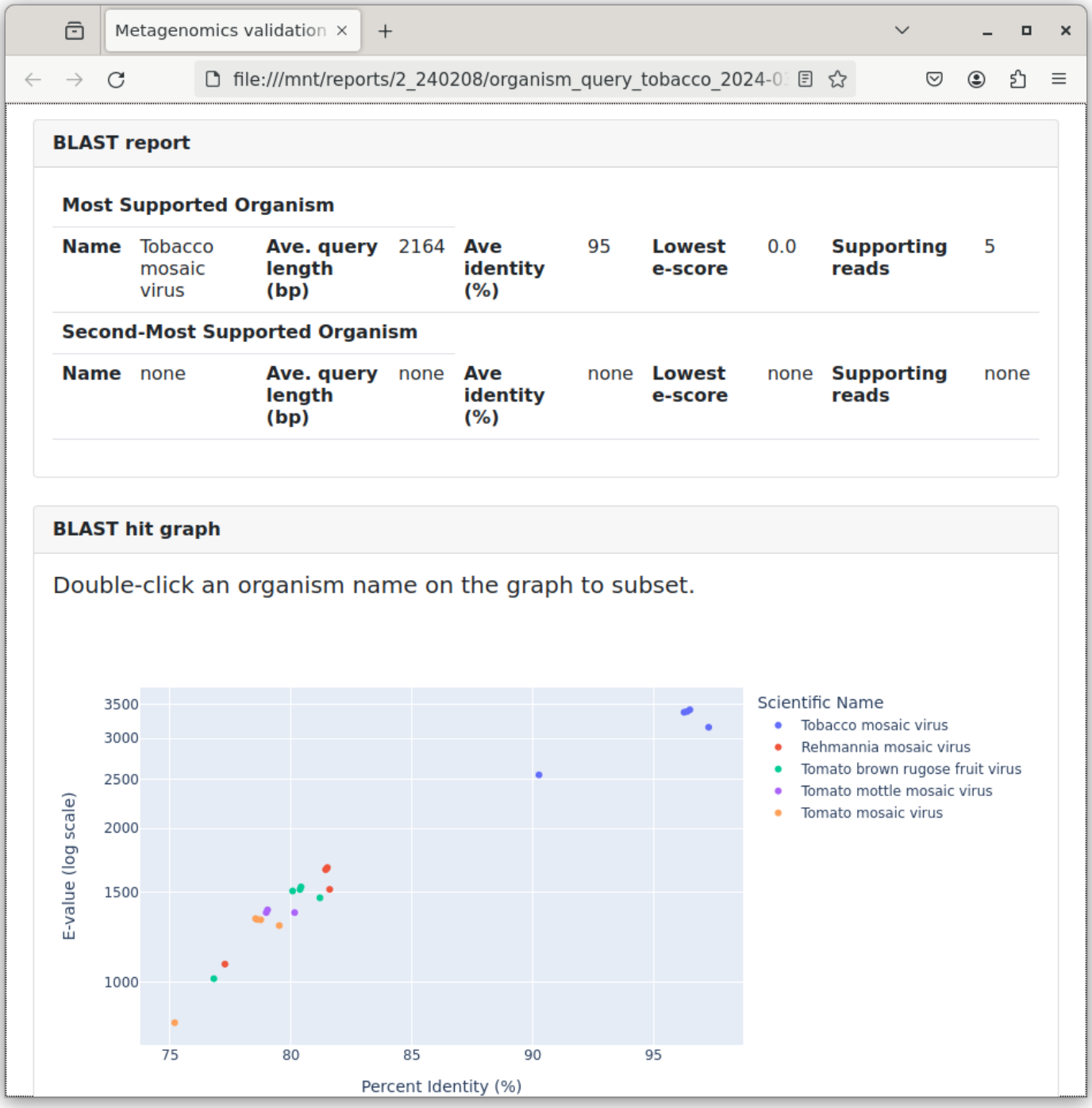
- Query Coverage: Percent of the query sequence length that is included in alignments against the sequence match.
- E-value: Indicates the number of hits or alignments that are expected to be seen by random chance with the same score or better. The lower the E-value, the more significant the alignment (the closer to 0, the better). E-value is the default metric used to sort the Descriptions table. Click here for a discussion of E-value thresholds.
- Percent Identity: Percent of nucleotides or amino acids that are identical between the aligned query and database sequences. A query sequence can share low percent identity with a sequence and still be a significant hit. It is essential to take the E-value into account and look for similarity between conserved regions (this will be more evident at the amino acid level).

Interactive plot

In a situation where there is either no confident match or no convergence across the read subset on an organism, or a alignments indicating a number of similar, closely related organisms, the plot appears disordered in a single cluster.



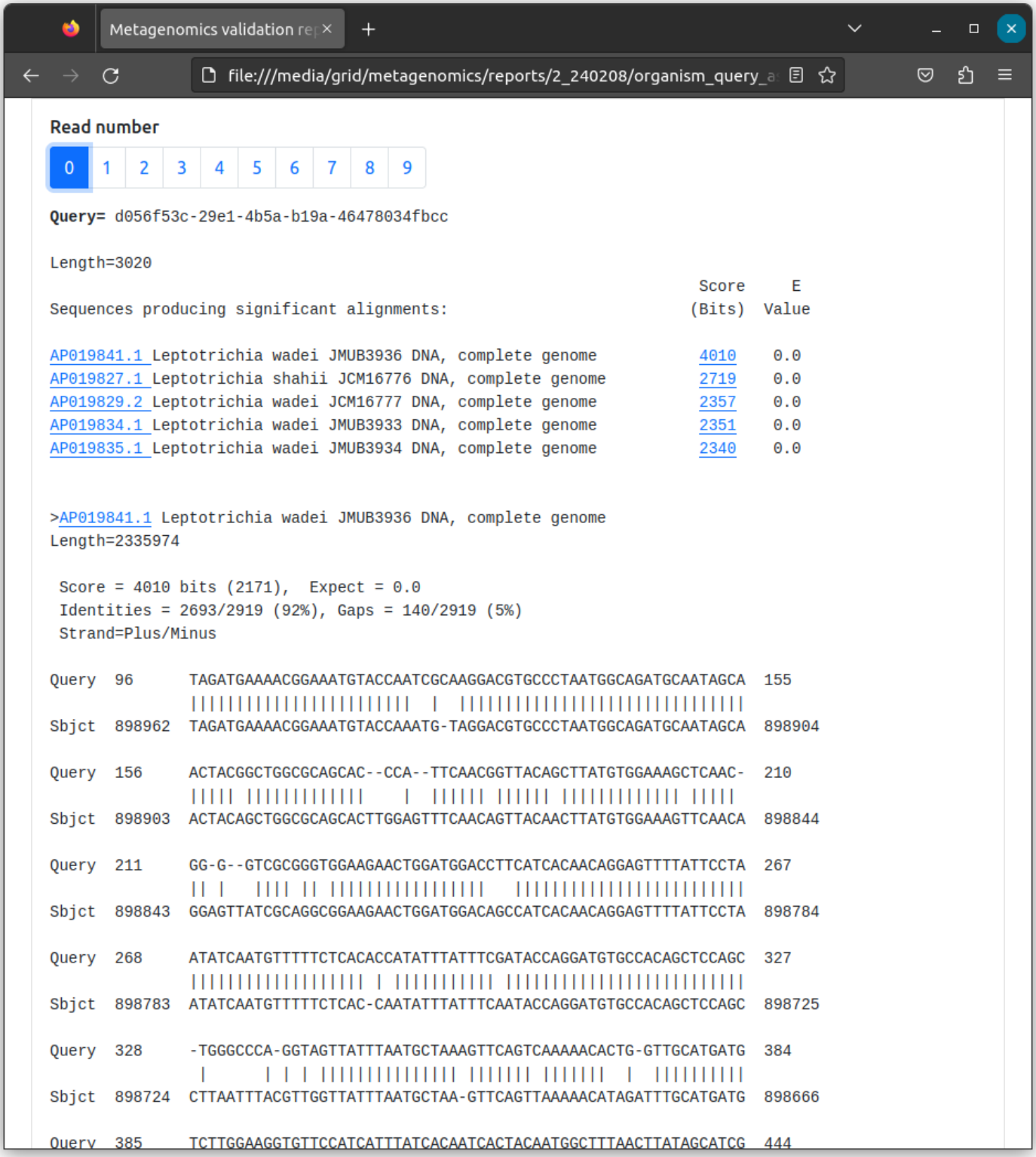
Situations where a convergent set of alignments are indicated results in a plot with two clusters, usually with one clear homogenous taxa. In this case, the Tobacco Mosaic virus.



BLAST alignments

Scrolling to the bottom of the report is a conventional BLAST alignment. We recommend looking at each read, the **Length** (query readlength), the **Score** and **Identities** values.

- If only a small alignment (80 bp) has been made from a long 2000 bp read, this is not likely to be a robust estimation.
- Check the alignment visualisation for long repeats, regions of low complexity etc. These regions often confound BLAST and taxonomic classification tools.


Metagenomics validation re
file:///media/grid/metagenomics/reports/2_240208/organism_query_a
Read number
0 1 2 3 4 5 6 7 8 9
Query= d056f53c-29e1-4b5a-b19a-46478034fbcc
Length=3020
Score E
Sequences producing significant alignments: (Bits) Value
AP019841.1 Leptotrichia wadei JMUB3936 DNA, complete genome 4010 0.0
AP019827.1 Leptotrichia shahii JCM16776 DNA, complete genome 2719 0.0
AP019829.2 Leptotrichia wadei JCM16777 DNA, complete genome 2357 0.0
AP019834.1 Leptotrichia wadei JMUB3933 DNA, complete genome 2351 0.0
AP019835.1 Leptotrichia wadei JMUB3934 DNA, complete genome 2340 0.0
>AP019841.1 Leptotrichia wadei JMUB3936 DNA, complete genome
Length=2335974
Score = 4010 bits (2171), Expect = 0.0
Identities = 2693/2919 (92%), Gaps = 140/2919 (5%)
Strand=Plus/Minus
Query 96 TAGATGAAAACGGAAATGTACCAATCGCAAGGACGTGCCCTAATGGCAGATGCAATAGCA 155
Sbjct 898962 TAGATGAAAACGGAAATGTACCAAATG-TAGGACGTGCCCTAATGGCAGATGCAATAGCA 898904
Query 156 ACTACGGCTGGCGCAGCAC--CCA--TTCAACGGTTACAGCTTATGTGGAAAGCTCAAC- 210
Sbjct 898903 ACTACAGCTGGCGCAGCACTTGGAGTTTCAACAGTTACAACTTATGTGGAAAGTTCAACA 898844
Query 211 GG-G--GTCGCGGGTGGAAGAACTGGATGGACCTTCATCACAACAGGAGTTTTATTCCTA 267
Sbjct 898843 GGAGTTATCGCAGGCGGAAGAACTGGATGGACAGCCATCACAACAGGAGTTTTATTCCTA 898784
Query 268 ATATCAATGTTTTTCTCACACCATATTTATTTCGATACCAGGATGTGCCACAGCTCCAGC 327
Sbjct 898783 ATATCAATGTTTTTCTCAC-CAATATTTATTTCAATACCAGGATGTGCCACAGCTCCAGC 898725
Query 328 -TGGGCCCA-GGTAGTTATTTAATGCTAAAGTTCAGTCAAAAACACTG-GTTGCATGATG 384
Sbjct 898724 CTTAATTTACGTTGGTTATTTAATGCTAA-GTTCAGTTAAAAACATAGATTTGCATGATG 898666
Query 385 TCTTGGAAGGTGTTCCATCATTTATCACAATCACTACAATGGCTTTAACTTATAGCATCG 444

3.6 mSCAPE upload tool

3.6.1 Introduction

With on-boarding to the mSCAPE programme, users are encouraged to upload samples to mSCAPE using the mSCAPE Launcher. This tool incorporates sample metadata inputted when using the Metagenomics Launcher, with some additional user inputs. The tool then packages human-scrubbed FASTQ data, producing metadata outputs in the required format for an mSCAPE submission.

3.6.2 Installation

The mSCAPE on-boarding team will provide all of the credentials required to make a test submission. These credentials should be stored in `~/.aws/` on the host machine. Contact the bioinformatics lead at GSTT for activation of the mSCAPE Uploader.

3.6.3 Uploading samples

Following the successful completion of the CIDR Metagenomics Workflow, users can double click the 'mSCAPE Launcher' on the desktop. The tool reads the sample information from the sample sheet saved when a workflow run is started and parses it in to the correct format ready for upload.

Note

Using the 'filename suffix' field on the Metagenomics Launcher appends the sample sheet filename with a string of your choosing, making it easier to find during audits or mSCAPE uploads.

Check out the video at the bottom of the page for a visual guide on how to run the uploader

1. From the 'Dropdown Options' section, select the parameters appropriate for the samples to be uploaded. Some addition fields for mSCAPE are inferred from the sample sheet, the data in the dropdowns or the sequencing reads themselves. See the table below for more details.

Note

For more details on how to fill the metadata fields associate with the samples, for example 'Study description', see the additional information section at the bottom of the page.

Parameter	Description
StudySite	The RMg Network site the sequencing took place.
Extraction Method	the nucleic acid extraction methodology used.
Spike-in	The spike-in control used.
ISOCountry	Country/nation.
Sequencing Protocol	The methodology used for sample preparation.
Library Protocol	the sequencing kit used for library preparation.
Bioinformatics protocol	The version of the CIDR Metagenomics bioinformatics workflow.
Clinical or research	mSCAPE clinical or research.
Human scrubbing	mSCAPE informatic Human Scrubbing protocol used.
Study description	Code provided by mSCAPE team.

1. Load a sample sheet by clicking the 'Load sample sheet' button at the bottom of the interface. The file browser takes you to the 'sample sheet' directory. Find the date/time corresponding to your run (or the sheet identified by the filename suffix) and open it. With this, the Main Table section should be populated with the samples from the corresponding sequencing run.
2. From the 'FASTQ Selection', select the dataset and the timepoint to be uploaded.

Note

Sample which are not intended for upload can be checked in the 'Main Table' and removed by clicking the 'Delete Selected rows' button at the bottom of the interface.

1. Once you have selected all of the timepoints for upload, choose the 'Update DataFrame' button at the bottom of the interface.
2. Review the data in the 'Main Table' panel ensuring it is suitable for upload.
3. Select 'Upload to S3'. You can check the terminal window for outputs confirming successful upload. A status indicator is also present in the FASTQ selection window.

Video tutorial

Video demonstration of using mSCAPE Uploader features

3.6.4 Additional information

Study description

Once onboarded by the CLIMB team, the site will be provided with a predetermined [study_centre_id] and three [study_id's].

A study_centre_id is an abbreviation of the site name e.g. name of the NHS trust. Each site will have one ID.

A study_id will be used to identify if samples are from a particular research study, or an NHS residual sample. Every NHS trust will have one Study ID related to their clinical samples, and a separate verification Study ID. This is to clearly differentiate between their verification stage and clinical service. Additional study_ids can be added as desired to differentiate any research studies.

A list of the predetermined site and study codes can be found here: (Link to be added soon)

See Table 1 for example of a study_centre_id, study_id and their purposes for GSTT.

Table 1. Example of a study centre ID and study ID:

study_centre_id	study_id	Purpose
GSTT	GSTT-CLI-01	Clinical service samples
GSTT	GSTT-VER-01	Verification
GSTT	GSTT- RES-01	Research – outside of network of excellence protocol

3.7 Summary report generator

3.7.1 Purpose

We have found it useful to be able to generate summaries of multiple runs for downstream analysis. This takes the form of a spreadsheet or CSV, with rows corresponding to each sample and columns containing information derived from the sample sheet and the metagenomic analyses across timepoints. The tool features an end-to-end GUI to select samples and build the sheet.

Important note

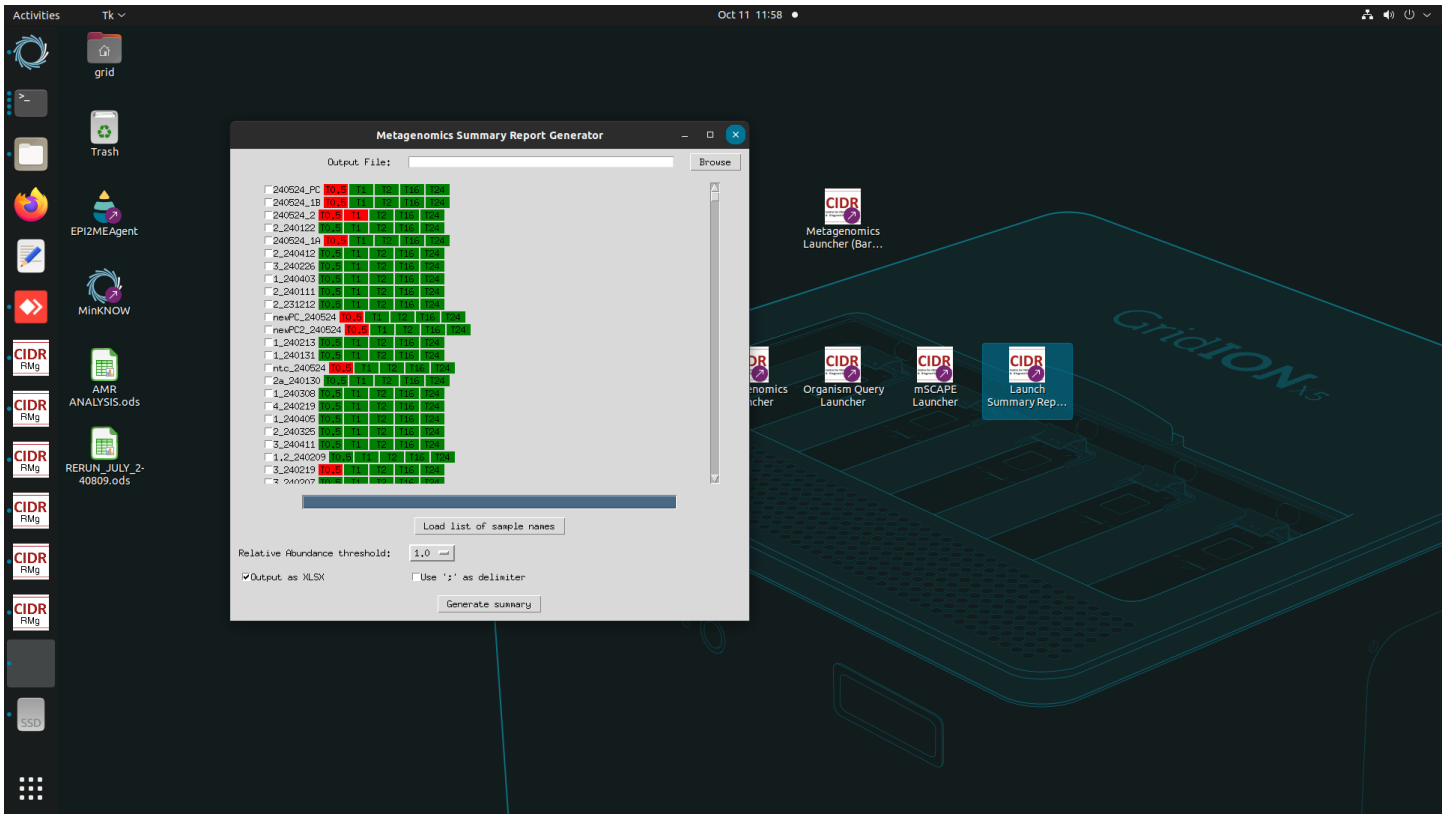
We ask that for the NHS RMg service evaluation, you set the Summary Report threshold cut-off to the lowest value (currently 0.1) so that all classifications are included.

A description of the fields featured in the spreadsheet is available here.

3.7.2 Instructions for use

Check out the video at the bottom of the page for an end-to-end demonstration

1. Double click the Launch Summary Report icon on the desktop. The window should appear with a loading bar. Please wait until all of the samples are sourced and loaded.



Note

In the above example, we have a list of samples, some of which have red indicators at specific time point. In both instances, it is likely that the reports were not generated because there were no reads present in the dataset. This is an especially frequent occurrence in NTC samples given that few reads should be detected in these samples.

1. The program reads all Metagenomics Workflow runs from the 'results' folder and populates a list. The list has a checkbox to include or exclude samples from the report, and a coloured box indicating the time point is present in the dataset. Select the samples for reporting using one of the two below methods.

   a. Select samples using the check boxes on the interface.

   b. Produce a simple list (newline delimited) of sample names, matching exactly (case sensitive) the Sample ID/Accession number used in the Metagenomics run. Save the list in the 'Sample Sheet' directory on the Metagenomics SSD. Select the 'Load list of sample names' button.
2. Choose whether you'd like to export as a spreadsheet (xlsx) or a CSV using the checkboxes at the bottom of the interface.

Note

Both report formats (xlsx/CSV) contain lists of taxa in single cells. In the xlsx, this is delimited by a newline (\n). In the CSV this is swapped for a semicolon ';' to avoid parsing errors.

1. Specify the output location by clicking the 'Browse' button in the 'Output File' section at the top of the interface. Fill out the 'Save As' prompt. **Please be sure add the '.xlsx' or '.csv' file extension to the filename if it not done automatically.**
2. Specify the 'Relative Abundance Threshold'. The default is 1.0%. this default parameter means that no organism < 1% relative abundance will will feature in the output report.

Important note

We ask that for the NHS RMg service evaluation, you set the Summary Report threshold cut-off to the lowest value (currently 0.1) so that all classifications are included.

1. Click 'Generate Summary'. The output will appear in the 'summary_report' directory.

Video tutorial - Manually selecting samples

Video tutorial - Providing a list of sample names

Output fields

Column	Description
Sample	The Lab/sample ID.
Experiment	The specific MinKNOW experiment.
SampleID	The specific MinKNOW sample.
Barcode	The barcode used in the library.
AnonymisedIdentifier	A de-identified hospital number.
CollectionDate	The date on which the sample was collected.
SampleClass	The classification of the sample, PC, NC, NTC, specimen.
SampleType	The sample site, eg. BAL, SPT, NPS.
Operator	The name or identifier of the individual who processed or sequenced the sample.
Notes	Any additional notes or remarks related to the sample or its processing.
Total reads XX hrs	The total number of sequencing reads generated after XX hours of sequencing.
Human reads XX hrs	The number of reads identified as being from human DNA after XX hours of sequencing.
Human reads (%) XX hrs	The percentage of total reads that are identified as human DNA after XX hours.
Total classified reads XX hrs	The number of reads that have been classified (assigned to an organism or category) after XX hours.
Sequencing N50 (bp) XX hrs	The N50 statistic for the reads generated after XX hours, indicating read length distribution.
Proportion >Q15 quality (%) XX hrs	The percentage of reads with a quality score greater than Q15 after XX hours.
Median read quality (PHRED score) XX hrs	The median PHRED quality score of the reads after XX hours, indicating overall data quality.
Total bases (bp) XX hrs	The total number of base pairs generated by the sequencing run after XX hours.
Organisms (excluding viruses) XX hrs	The list of organisms (excluding viruses) identified from the reads after XX hours.
Organisms (excluding viruses) read counts XX hrs	The read counts associated with organisms (excluding viruses) after XX hours.
Organism (excluding viruses) percentage abundance XX hrs	The percentage abundance of each organism (excluding viruses) in the sample after XX hours.
Viral organisms XX hrs	The list of viral organisms identified from the reads after XX hours.
Viral read counts XX hrs	The read counts associated with viral organisms after XX hours.

3.8 Setting up CIDR Metagenomics bioinformatics workflow - alternative deployment

Note

The tools and documentation described here and on the CIDR GitHub are not validated for diagnostic use and are for research and evaluation purposes only.

3.8.1 Overview

For collaborators outside of the Network, an alternative configuration can be provided. This will bypass the GUI allowing users to provide a `sample_sheet.csv` through a CLI. Organism query will not be available to headless users as this tools is heavily reliant on GUI I/O.

3.8.2 Install instructions

1. Decompress CIDR_metagenomics_vX.X.tar.gz:

```
tar -xvzf CIDR_metagenomics_vX.X.tar.gz
```

1. Install conda/mamba.
2. Build the appropriate environment for running the CIDR metagenomics containers.

```
wget https://raw.githubusercontent.com/GSTT-CIDR/metagenomics_container/main/conda/apptainer.yml conda env create -f apptainer.yml
```

1. Allocate a directory for MinKNOW data outputs. This will be mounted to the `/data` directory in the container in a later step.

Note

The directory structure of data for ingest must be maintained as in standard MinKNOW outputs eg. **Example for control sample**
`[minknow_outputs_directory]/GSTT_control_sample_01/GSTT_control_sample_01/20240424_1408_X4_FAY88387_d3868a4f/fastq_pa`
Naming schema `[minknow_outputs_directory]/[experiemnt]/[sample_id]/[*]/fastq_pass/barcodeXX`

3.8.3 Install validation

1. Navigate to the root of the `CIDR_metagenomics_vX.X` directory.
2. Move `CIDR_metagenomics_vX.X/GSTT_control_sample_XX` to the allocated directory for MinKNOW data outputs (from Install instructions: Step 4).

3. activate the apptainer conda environment: `conda activate apptainer`
4. Initiate the run for analysing the control dataset:

```
apptainer exec --bind ./:/mnt --bind ./data:/data ./containers/cidr_metagenomics_v3.6.sif bash -c 'cd /workflow ; source /opt/conda/etc/profile.d/conda.sh
```

1. When the workflow has completed, inspect the `CIDR_metagenomics_vX.X/reports/CIDR_control_1` PDF report, it should match the CIDR validation report provided here.

Info

Variables to change in step 3
--bind ./:/mnt - Binding the workflow root directory to the container /mnt.
--bind ./data:/data - binding the allocated directory for MinKNOW data outputs to /data.
./containers/cidr_metagenomics_v3.6.sif - launching the metagenomics container.
for t in 0.5 1 2 16 24 - time-points for analysis.
--cores 20 - number of samples to be processed simultaneously - not the same as threads.
samples=/mnt/sample_sheets/CIDR_control_1.csv - the mounted path for the sample sheet - remember this is the relative mounted path, so `/mnt/sample_sheets` corresponds to `CIDR_metagenomics_vX.X/sample_sheets` on the host machine.

3.8.4 Implementation

1. Build a **sample sheet** copying the structure of the example in `CIDR_metagenomics_vX.X/sample_sheets`. Importantly, 'Experiment', 'SampleID' and 'Barcode' must be correct and correspond to the `[minknow_outputs_directory]/[experiemnt]/[sample_id]/[*]/fastq_pass/barcodeXX` scheme.
2. activate the apptainer conda environment: `conda activate apptainer`
3. Run the container, changing the flags explained in the validation step:

```
apptainer exec --bind ./:/mnt --bind ./data:/data ./containers/cidr_metagenomics_v3.6.sif bash -c 'cd /workflow ; source /opt/conda/etc/profile.d/conda.sh
```

1. PDF outputs should be found in `CIDR_metagenomics_vX.X/reports/` corresponding to each LabID in the **sample sheet** loaded.

4. Analysis

4.1 Service evaluation report SOP

Note

Please reference use of this method in any presentation or publication as:

Alcolea-Medina, A., Alder, C., Snell, L.B. et al. Unified metagenomic method for rapid detection of microorganisms in clinical samples. Commun Med 4, 135 (2024). https://doi.org/10.1038/s43856-024-00554-3

Important

Risk assessment for handling respiratory samples needs to be performed by each laboratory.

The tools and documentation described here and on the CIDR GitHub are not validated for diagnostic use and are for research and evaluation purposes only.

4.2 Network validation outline

Note

Please reference use of this method in any presentation or publication as:

Alcolea-Medina, A., Alder, C., Snell, L.B. et al. Unified metagenomic method for rapid detection of microorganisms in clinical samples. Commun Med 4, 135 (2024). https://doi.org/10.1038/s43856-024-00554-3

Important

Risk assessment for handling respiratory samples needs to be performed by each laboratory.

The tools and documentation described here and on the CIDR GitHub are not validated for diagnostic use and are for research and evaluation purposes only.

5. FAQ

5.1 FAQ

5.1.1 1) My 30 minute reports are persistently missing.

Presentation

During live analysis runs only, the 30-minute reports are not being produced. All other reports are coming out fine. The clock is also running an hour behind (during BST).

Solution

The GridION is not configured out of the factory for the UK timezone - an issue specifically during BST. this means the data ingest script timings fail and the 30 minute timepoint is missed. **Note - if other reporting timepoints are missing, this is not your solution.**

1. **Type the following in to a terminal window** `sudo timedatectl set-timezone Europe/London`
2. Check the time is correct. If not, manually update the time to be correct.

5.1.2 2) Some timepoint reports are missing from my run.

Presentation

Entire samples, or timepoint reports are sporadically missing from the report folder. In many cases, reports are not generated because no microbial reads are detected - this causes the pipeline to crash. We are working on improving this error reporting and softer crashes.

There are a number of causes for reports failing. Follow the flowchart to find the correct solution.

flowchart TB A["All reports before
 a specific time
 point are missing. eg.
 (~~0.5hr~~)(~~1hr~~)(~~2hr~~)(16hr)(24hr)"] --> B[In the first
 available report, is
 the read count very low?] B -->|Yes| C[The sequencing yield
 is low. Previous reports
 were not generated because
 no microbial reads were found.] B -->|No| D[See Action 1] E[A single report is
 missing where previous
 time points had
 reports generated.] --> F[Action 1] G[All reports are missing] --> H[Check MinKNOW barcode panel.
 Are there reads reporting for
 the barcode in question?] H -->|Yes| I[See Action 2] H -->|No| J[The sequencing run has
 failed. No reads are
 present to analyse.]

Actions

1. Delete the folders in `../metagenomics/results` and `../metagenomics/reports` corresponding to the specific sample with issues. Rerun the the pipeline by loading the saved sample sheet. You can do this by using the 'Load Sample Sheet' function on the metagenomics launcher - find the sample sheet corresponding to the date/time. Alternatively, you can re-enter the run information in to the launcher.
2. At the moment, no report is generated if the sample consists of human reads only. Run the following command in a terminal - it opens one of the intermediate classification files which we can use to confirm a human-only run. replace {sample_id} with the appropriate (identical) problematic sample IDand {timepoint} with an appropriate timepoint eg. 0.5_hours.

   The number 9606 in the terminal output corresponds to the Homo sapiens taxon. If only this is present, the read is human -only. If there are other numbers, try Action 1.

Note

Open the terminal py pressing CTRL-ALT-T

```
zcat -f /media/grid/metagenomics/results/{sample_id}/{timepoint}/centrifuge/centrifuge_raw.tsv* | awk -F'\t' '{print $3}' | sort -u
```

5.1.3 3) My summary report has returned a "....returned non-zero exit status 1" and is not generating spreadsheets.

Presentation

The summary report tool outputs an error message and does not create a summary report.

Actions

The tool should automatically add the correct file extension on to the filename provided by the user. If it has not, add either the '.xlsx' extension or the '.csv' extension depending on your needs.

5.1.4 CIDR RMg genome database