CIDR Metagenomics Hub

CIDR

None

Table of contents

1. Metagenomics network hub 3

1.1 Network Sites 3

2. Lab Resources 5

2.1 Panmetagenomics protocol 5

3. Bioinformatics 6

3.1 Bioinformatics for clinical metagenomics 6

3.2 Setting up CIDR Metagenomics bioinformatics workflow 9

3.3 Running the metagenomics workflow 11

3.4 Bioinformatics - Organism query 15

3.5 Setting up CIDR Metagenomics bioinformatics workflow - alternative deployment 22

4. Analysis 24

4.1 Service evaluation report SOP 24

4.2 Network validation outline 25

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 Network Sites

1.1.1 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.

1.1.2 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.

1.1.3 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 [Unified metagenomic method for rapid detection of bacteria, fungi and viruses in clinical samples | Research Square which is currently going through journal review process. Work conducted during development and evaluation of metagenomics protocols are published in Baldan R et al J Infect. 2021 83:167. Charalampous T et al Genome Medicine 2021 13:182 Charalampous T et al Am J Resp Crit Care Med. 2024 209:164-174.

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.

3. Bioinformatics

3.1 Bioinformatics for clinical metagenomics

3.1.1 Overview

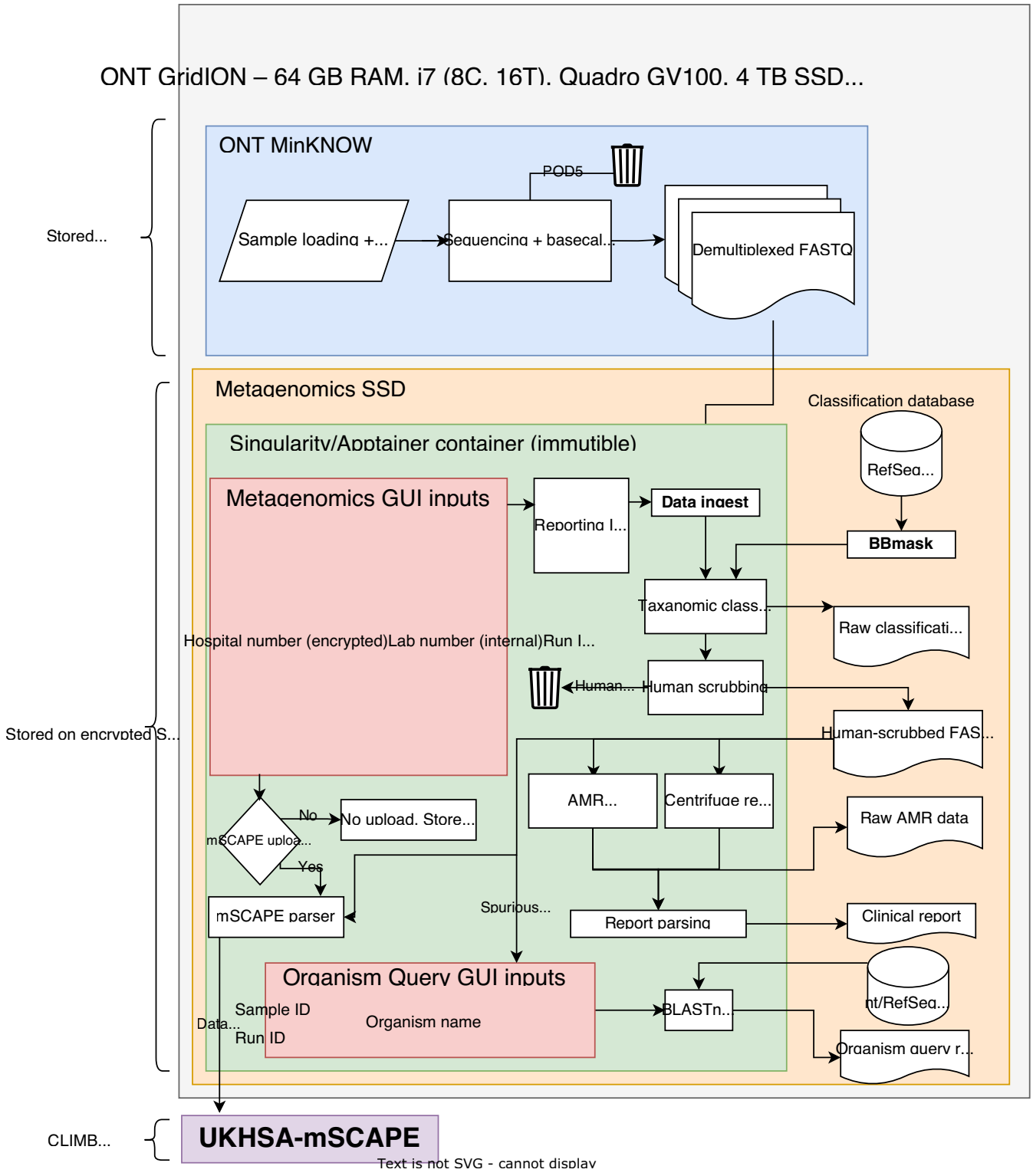
The principal output of the CIDR Metagenomics workflow is a PDF report listing organisms with detectable nucleic acids (RNA/DNA) and some additional information on AMR associated sequence data. The solution packages two applications - CIDR Metagenomics Workflow and Organism Query alongside a few scripts to help manage and analyse outputs. The Metagenomics Workflow runs ontop of MinKNOW, analysing sequencing data in real time producing 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 ~20 minutes. The Organism Query report is designed to provide the user with appropriate information to scrutinise a significant taxanomic classification.

mSCAPE

Users can opt in to mSCAPE on an per-experiment basis for an automatic upload of sequencing data to mSCAPE. The mSCAPE tool collates and encrypts data directly from the metagenomics workflow outputs ready for transmission.

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, taxanomic classification, AMR identification, which is then consolidated in to a PDF reports which are saved in the `/media/grid/metagenomics/results/` directory. The diagram below illustrates the technical structure of the workflow:



Taxanomic classification

At its core, the pipeline leverages a taxanomic classification tool called Centrifuge. There a numerous alternatives, the most common being Kraken. We chose centrifuge namely because of its smaller memory footprint and existing deployment in ONT's WIMP. Each read in the raw data is aligned against an index, and is

assigned a confidence score and a taxonomy. Our index is an optimised database of curated eukrayotic, prokaryotic, and viral reference sequences formed primarily from NCBI RefSeq and FDA-ARGOS databases. This is provided on the SSD sent to each site.

The index was built from the following sequences - this map file contains the accession number for each sequence in the database and the accompanying taxa ID. This file is required to assemble the Centrifuge index alongside the sequences in FASTA format. Before building the index, the sequences were masked using BBmask. This tool is applied primarily to prevent false-positive matches in highly-conserved or low-complexity regions of genomes.

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 ./
```

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.

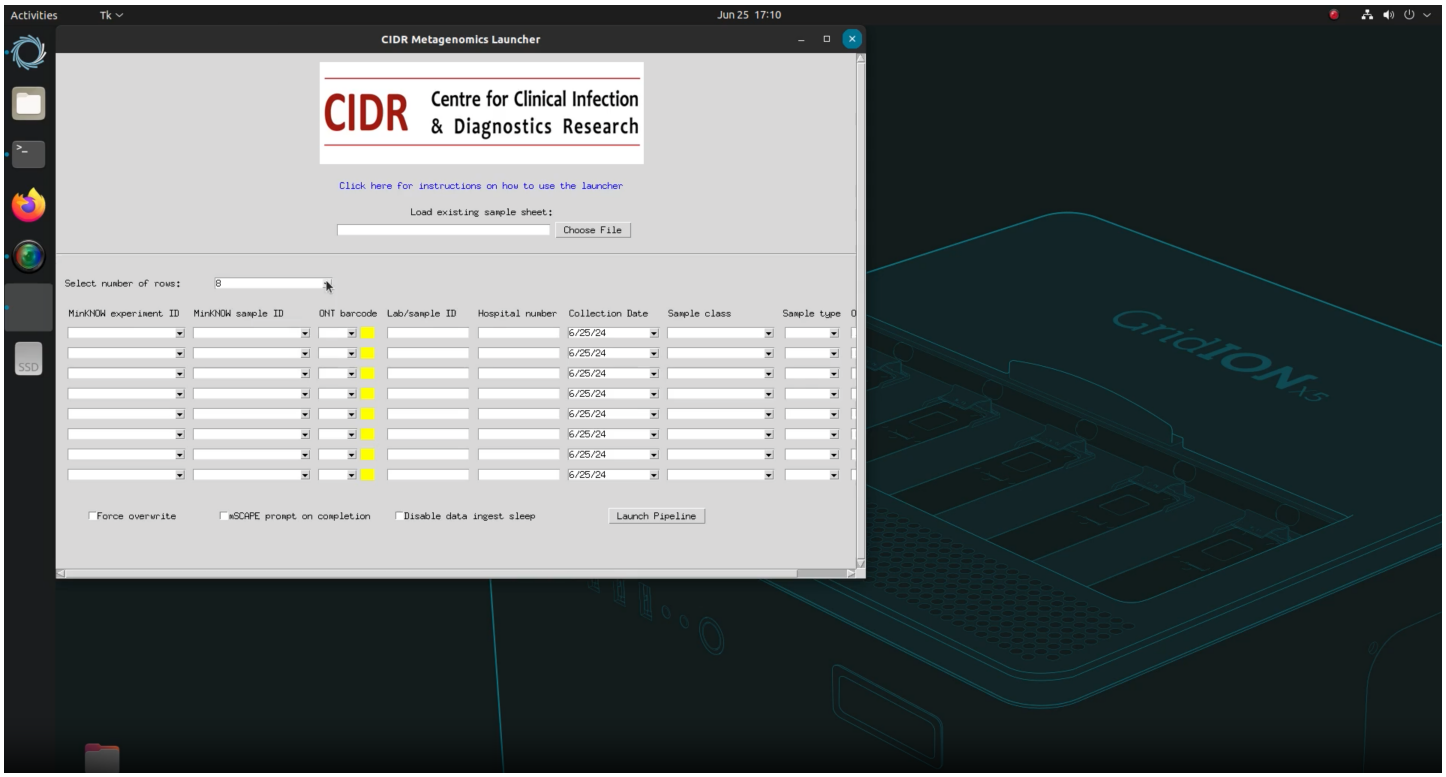
3.3 Running the metagenomics workflow

Before starting

1. The CIDR metagenomics workflow must be started during a sequencing experiment or after a sequencing experiment has completed. The pipeline must not be activated before a sequencing experiment has started in MinKNOW and has **started producing reads** (See MinKNOW setup - Lab Protocol).
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.

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.

1. Select the number of samples to be analysed from the dropdown.
2. You can choose to initiate the launcher using **one** of the below methods:

    • Fill out the fields on the form for each sample to be analysed.

    • Loading a pre-existing TSV - see example.

Field descriptions:

Field	Description
MinKNOW 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.
KinKNOW 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}/` directory.
ONT barcode	The ONT library index/barcode used. Green colour indicates the barcode directory has been validated.
Lab/Sample ID	The unique lab accession number for the sample. This data is encrypted before transmission.
Hospital number	Hospital number corresponding to the sample. Can be anonymised. This is encrypted before transmission.
Collection date	Collection data of the sample.
Sample Class	The class of sample loaded.
Sample type	The methodology used to collect the sample.
Operator	Arbitrary identifier of the user operating the sequencer.

Note

- Option 1 will generate a sample sheet stored in the `metagenomics/sample_sheets` directory. 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.

- For full information on data encryption protocol visit the mSCAPE uploader page.

- Test

1. With the metadata form filled, select the run parameter check boxes.

Parameter	Description
Force overwrite	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.
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.
Disable data ingest sleep	**Not for real-time analysis!** Analyse all data immediately - do not wait for it to be generated by the sequencer.

Known issues

You should wait to launch the pipeline after the sequencer has reported producing reads in MinKNOW, the workflow will display errors in red if no reads have been found.

The NTC will exhibit the same 'error' behavior as no reads are present in the corresponding barcode folder. We are working on functionality to circumvent this.

You can stop the analysis or close the Launcher window at any point by typing `CTRL+C` while the terminal window is active and closing the launcher window.

1. Click on `Launch pipeline` and click `OK` to start analysis.

2. After a minute, the terminal window accompanying the workflow launcher should start displaying log outputs from the workflow. See below for an example.

1. ~40 minutes after launching the sequencing experiment alongside the metagenomics workflow, the first reports will be available in `/media/grid/metagenomics/reports/{sample_name}/{timepoint}`. See below for a guide on how to access this.

Success!

We have now run the CIDR metagenomics workflow. The workflow will run for ~24 hours generating PDF reports for 0.5, 1, 2, 16, 24 hour time-points.

3.4 Bioinformatics - Organism query

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

3.4.1 Technical information

- The tool searches the classified reads for an organism indicated by the user. Selecting a subset of 10 (max) reads assigned to that taxa.
- The reads are extracted from the FASTQ file stored in the workflow `results` folder and BLASTs them against the prescribed database.
- 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 `reports/{sample_id}/organism_query_XXX` folder, with the other PDF reports from the metagenomics run.
- In the report folder is the BLAST HTML report and the subsetted FASTQ and FASTA reads.

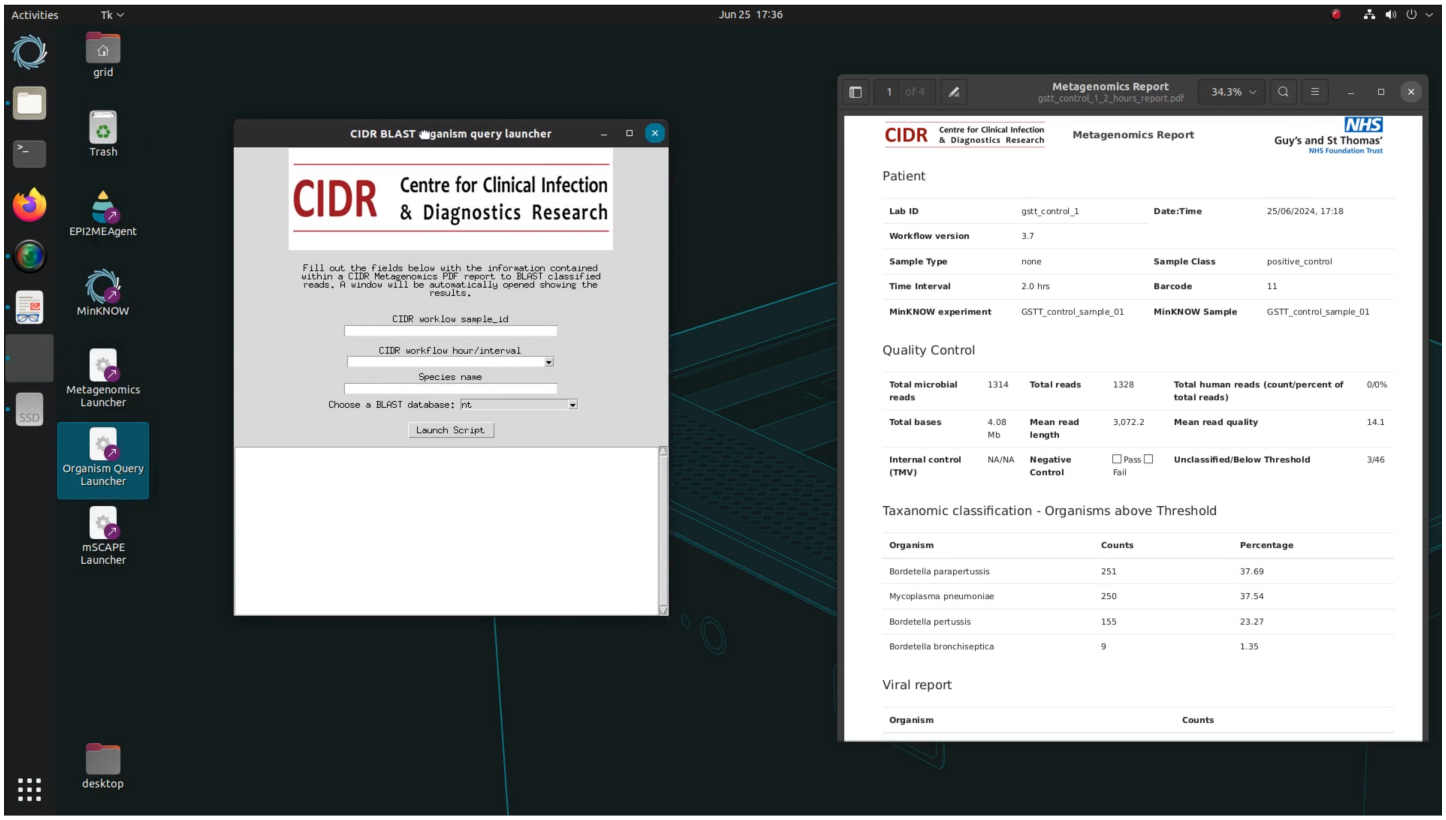
Info

The most thorough analysis is performed using the default nt database. This usually takes ~30 minutes to generate a report.

Known issues

The interactive plot is only available immediately after running the tool, when the original Launcher window is open. Loading the query report after closing will allow you to view the BLAST alignments but not the plot section, which will display an 'Internal Server Error'.

Running queries simultaneously is currently not supported as the interactive plot function can run a single session. If two reports are opened, the incorrect plot may display on the report. This does not effect the full BLAST analysis, only the plot displayed at that time.



3.4.2 Launching organism query

1. Load the PDF report from the run you'd like to query a classification from.

2. Click on the Organism Query launcher icon on the GridION desktop.

3. Fill out the fields as indicated in the video below.

Parameter	Description
SampleID	The Lab/Sample ID assigned to the sample when launching the metagenomics workflow.
Workflow hour/interval	The timepoint, corresponding to the dataset you'd like query.
Species name	The name of the species to be queried eg. Aspergillus fumigatus

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.

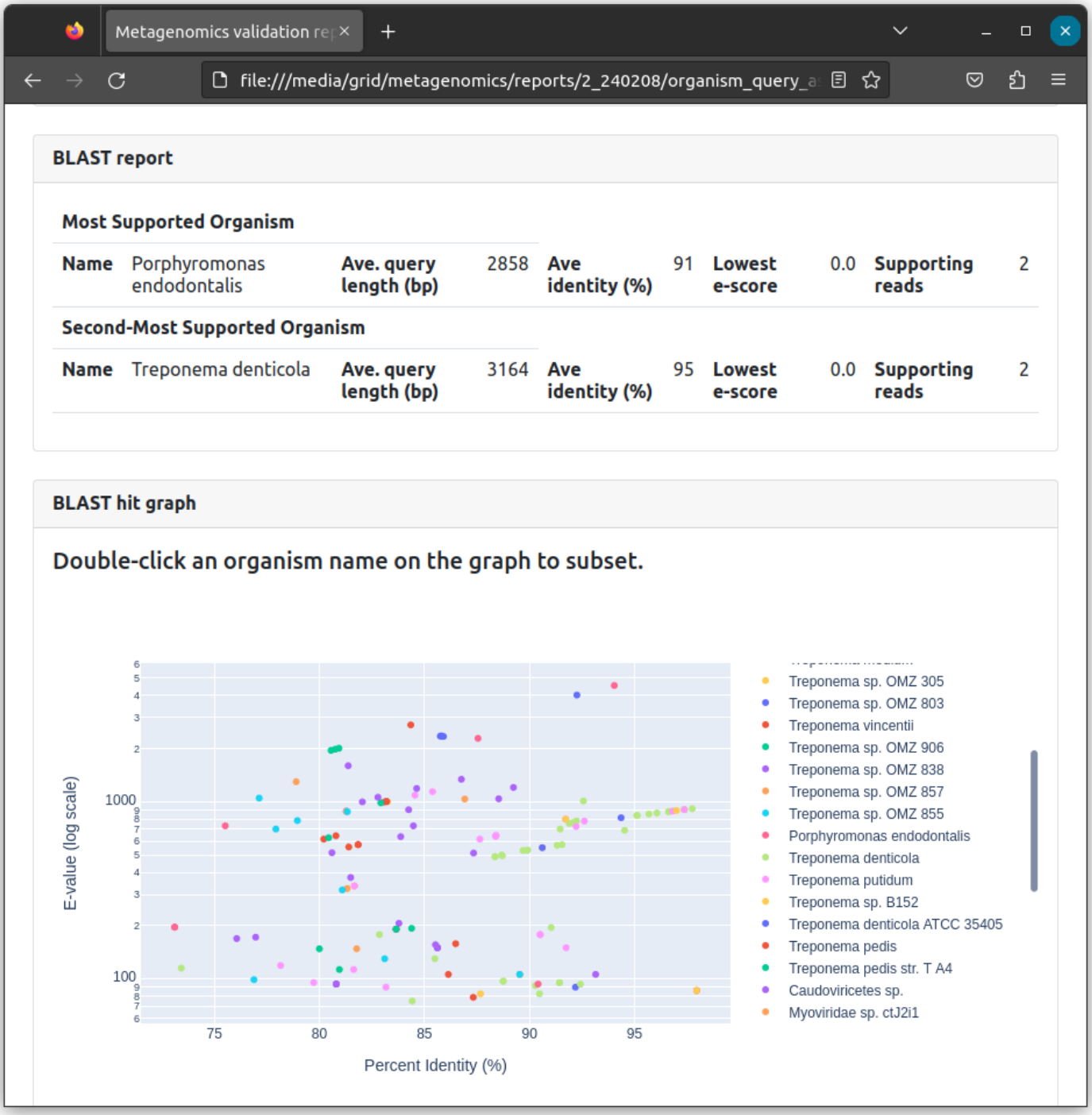
3.4.3 Interpreting results

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.

Metagenomics validation

file:///mnt/reports/2_240208/organism_query_tobacco_2024-0

BLAST report

Most Supported Organism

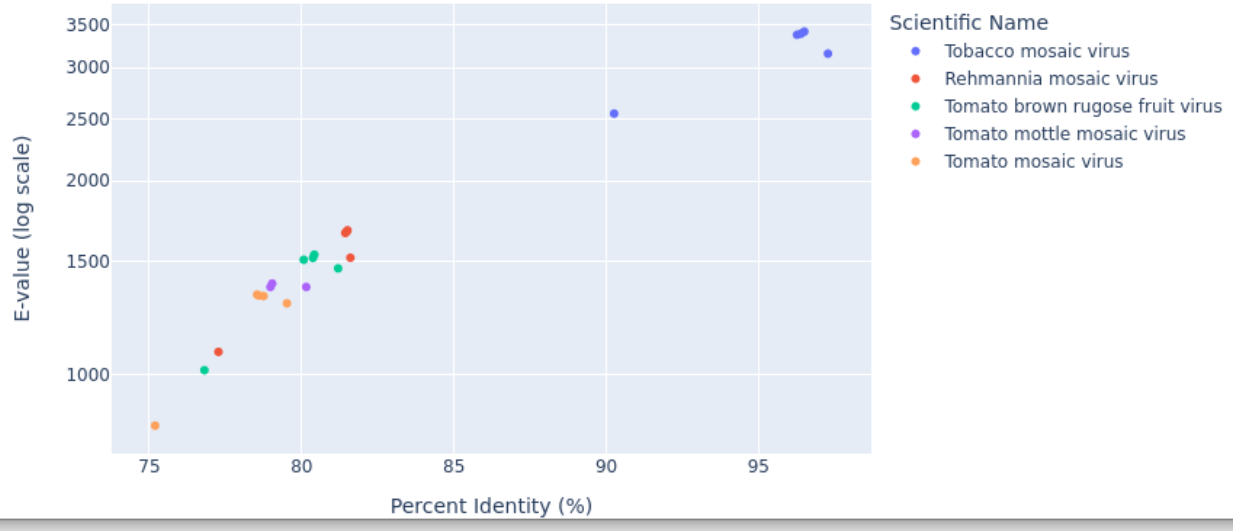
Name	Tobacco mosaic virus	**Ave. query length (bp)**	2164	**Ave identity (%)**	95	**Lowest e-score**	0.0	**Supporting reads**	5

Second-Most Supported Organism

Name	none	**Ave. query length (bp)**	none	**Ave identity (%)**	none	**Lowest e-score**	none	**Supporting reads**	none

BLAST hit graph

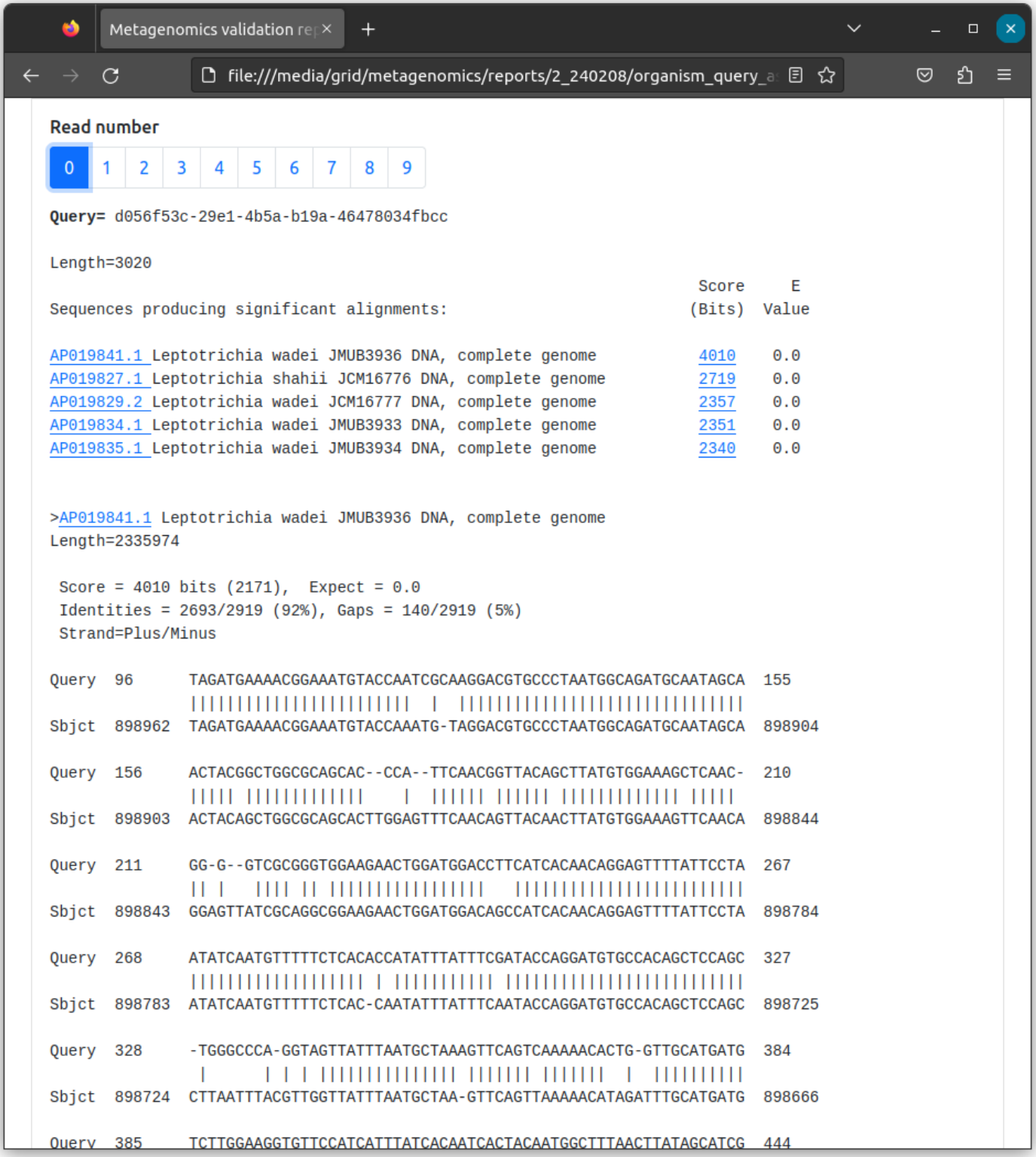
Double-click an organism name on the graph to subset.



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

3.5 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.

3.5.1 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.

3.5.2 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.ym
```

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_FAY8838
```

Naming schema `[minknow_outputs_directory]/[experiemnt]/[sample_id]/[*]/fastq_pass/barcodeXX`

3.5.3 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/con
```

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.

3.5.4 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/con
```

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 [Unified metagenomic method for rapid detection of bacteria, fungi and viruses in clinical samples | Research Square which is currently going through journal review process. Work conducted during development and evaluation of metagenomics protocols are published in Baldan R et al J Infect. 2021 83:167. Charalampous T et al Genome Medicine 2021 13:182 Charalampous T et al Am J Resp Crit Care Med. 2024 209:164-174.

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 [Unified metagenomic method for rapid detection of bacteria, fungi and viruses in clinical samples | Research Square which is currently going through journal review process. Work conducted during development and evaluation of metagenomics protocols are published in Baldan R et al J Infect. 2021 83:167. Charalampous T et al Genome Medicine 2021 13:182 Charalampous T et al Am J Resp Crit Care Med. 2024 209:164-174.

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.