Oct 07, 2024

DDNS Data analysis and quality control checks

DOI
dx.doi.org/10.17504/protocols.io.5qpvok4jxl4o/v1

Alex Shaw[1], Joyce Akello[1], Catherine Troman[1], Aine OToole[2], c.ansley[2], Catherine Pratt[3], Erika Bujaki[4], Zoe Vance[2], rachel.colquhoun[2], Andrew Rambaut[2], Javier Martin[4], Nick Grassly[1]

[1]Imperial College London; [2]University of Edinburgh; [3]Biosurv International;
[4]Medicines and Healthcare products Regulatory Agency

Poliovirus Sequencing Co…

Joyce Akello
Imperial College London

DOI: dx.doi.org/10.17504/protocols.io.5qpvok4jxl4o/v1

Protocol Citation: Alex Shaw, Joyce Akello, Catherine Troman, Aine OToole, c.ansley, Catherine Pratt, Erika Bujaki, Zoe Vance, rachel.colquhoun, Andrew Rambaut, Javier Martin, Nick Grassly 2024. DDNS Data analysis and quality control checks . **protocols.io** https://dx.doi.org/10.17504/protocols.io.5qpvok4jxl4o/v1

Protocol status: Working
We use this protocol and it's working

Created: September 16, 2024

Last Modified: October 07, 2024

Protocol Integer ID: 107688

Keywords: Poliovirus detection, Nanopore sequencing, Direct detection, DDNS

Funders Acknowledgement:

Bill and Melinda Gates Foundation

Abstract

This standard operating procedure indicates how to perform data analysis, quality control checks, and data reporting for DDNS and provides guidance on best practice for ensuring relevant sequencing run data are recorded. The document is structured into quality control steps after a DDNS sequencing run.

Procedure

For DDNS results to be valid and suitable for reporting, sufficient sample metadata must be recorded and data integrity maintained throughout the planning stage of the experiment, during the experiment and after the experiment. Quality control checks have been included to ensure that the protocol has been performed correctly and that results are valid. Later comparison of the DDNS results with culture-based poliovirus detections can be facilitated by adding further metadata describing the timelines and results for processing of the same samples by cell-culture.

An overview of the procedure is shown in Figure 1.

It is the responsibility of the Lab senior scientist to designate staff to conduct the QC analysis report and to ensure that the personnel conducting the data analysis has provided complete and accurate data, and the report generated is correct prior to approval. Data must be reviewed and approved by the lab senior scientist before QC analysis review with the technical team (PSC member). Designated, technical team lead then submit the QC’d data to the GSL for review/approval. Once approved, the GSL share the QC'd data to the program.

A log should be kept of the sequencing runs that are performed, indicating whether all quality control checks have been completed or whether some remain pending, and confirming that all reportable sequences have been reported. All VP1 poliovirus sequences generated by DDNS (even in cases of sample or run QC fail) should be collected into a laboratory sequence QC database folder to aid in the identification of contamination. This addition can be performed by annotating the vp1_sequence.fasta from a run with the run name (e.g. "vp1_sequences_Run21.fasta") and copying it into the database folder.

We recommend that if you are using this protocol for stool surveillance, do not add RNA from cell-culture isolates to the same run. The RNA from cell-culture will be hugely more concentrated resulting in the sequencing data skewing towards these samples rather than the stool samples.

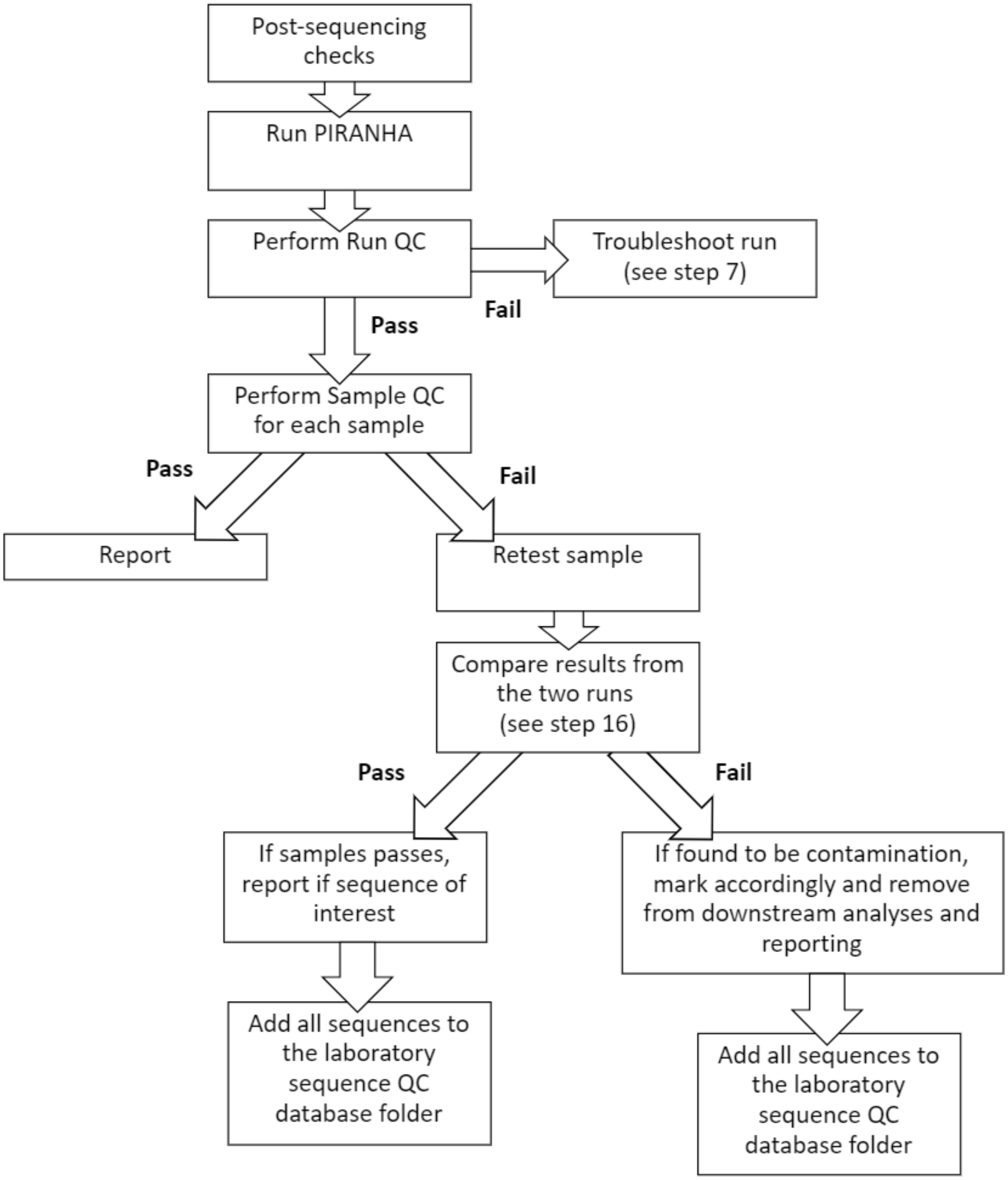

Post-sequencing checks
Run PIRANHA
Perform Run QC
Troubleshoot run (see step 7)
Fail
Pass
Perform Sample QC for each sample
Pass
Fail
Report
Retest sample
Compare results from the two runs (see step 16)
Pass
Fail
If samples passes, report if sequence of interest
If found to be contamination, mark accordingly and remove from downstream analyses and reporting
Add all sequences to the laboratory sequence QC database folder
Add all sequences to the laboratory sequence QC database folder

Figure 1. Overview of the data analysis and quality control procedure for DDNS. “Steps” refer to the procedure steps below.

Guidelines

All procedures to be performed by suitably trained members of staff.

Post Sequencing Run Checks

1 Perform the post sequencing run checks by confirming the following points manually
 a. Did the sequencing run complete its full run duration (check the MinKNOW run report).
 b. Was there no sudden reduction in pore numbers i.e. pores numbers did not fall beneath 400 (or 25% of total pores) in the first hour of the sequencing run (check the MinKNOW run report, see Figure 2).

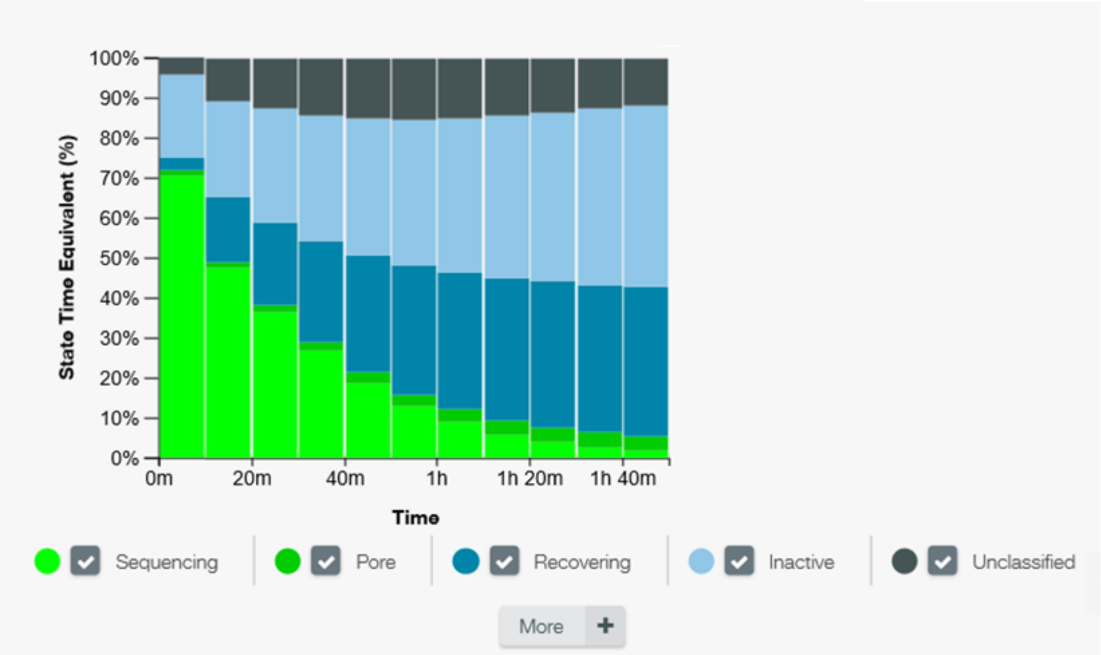


Figure 2*. The MinKNOW run report showing a failed sequencing run where the percentage of available pores (sum of bright green "Sequencing" and darker green "Pore" bars) falls below 25 % of total pores within the first hour of the sequencing run.*

If all answers are yes, then continue to running PiranhaGUI for sequence analysis. If any answers are no, then complete the troubleshooting steps;

2 Depending on the reason for failure:
a. Run has not reached its full duration:
 i. Restart the sequencing run. Check that there are still >500 pores available for sequencing (this will be reported when the run is restarted).
 ii. If there are insufficient pores available, repeat the library pooling and sequencing library preparation and load into a different flow cell.
b. Sudden reduction in pore numbers during the first hour:
 i. Repeat the sample pooling and library preparation and load into a different flow cell if the run doe snot pass QC checks after analysis with PIRANHA

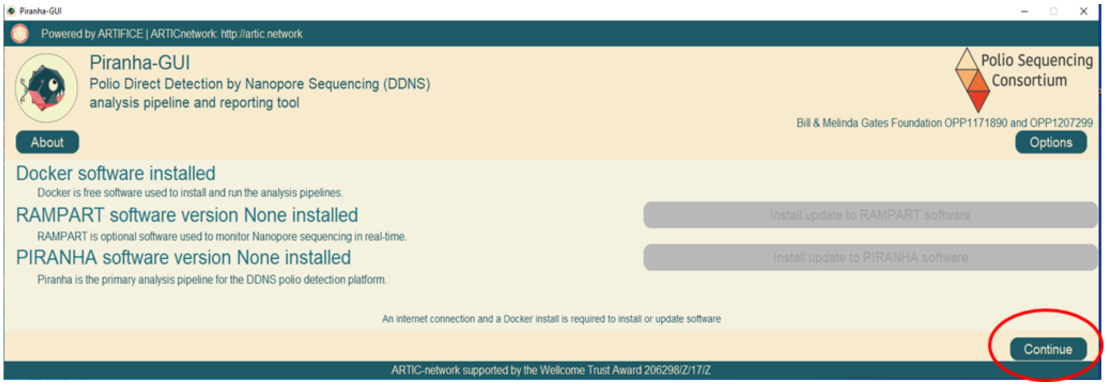
Running PiranhaGUI for DDNS Sequence Analysis

3 Run the PIRANHA analytical software once the sequencing run is complete. For full details on PiranhaGUI installation and usage visit https://github.com/polio-nanopore/piranhaGUI.

3.1 Open **docker** (click “Update to latest” in the top left if Docker Desktop needs to be updated)

3.2 Open **PiranhaGUI** (click “Install update to PIRANHA software” if piranha needs to be updated)

3.3 The PiranhaGUI window will pop up as shown below. Click "Continue" to begin setting up your run analysis.


Piranha-GUI
About
Options
Install update to RAMPART software
Install update to PIRANHA software
Continue

Powered by ARTIFICE | ARTICnetwork: http://artic.network

Piranha-GUI

Polio Direct Detection by Nanopore Sequencing (DDNS) analysis pipeline and reporting tool

Polio Sequencing Consortium

Bill & Melinda Gates Foundation OPP1171890 and OPP1207299

Docker software installed

Docker is free software used to install and run the analysis pipelines.

RAMPART software version None installed

RAMPART is optional software used to monitor Nanopore sequencing in real-time.

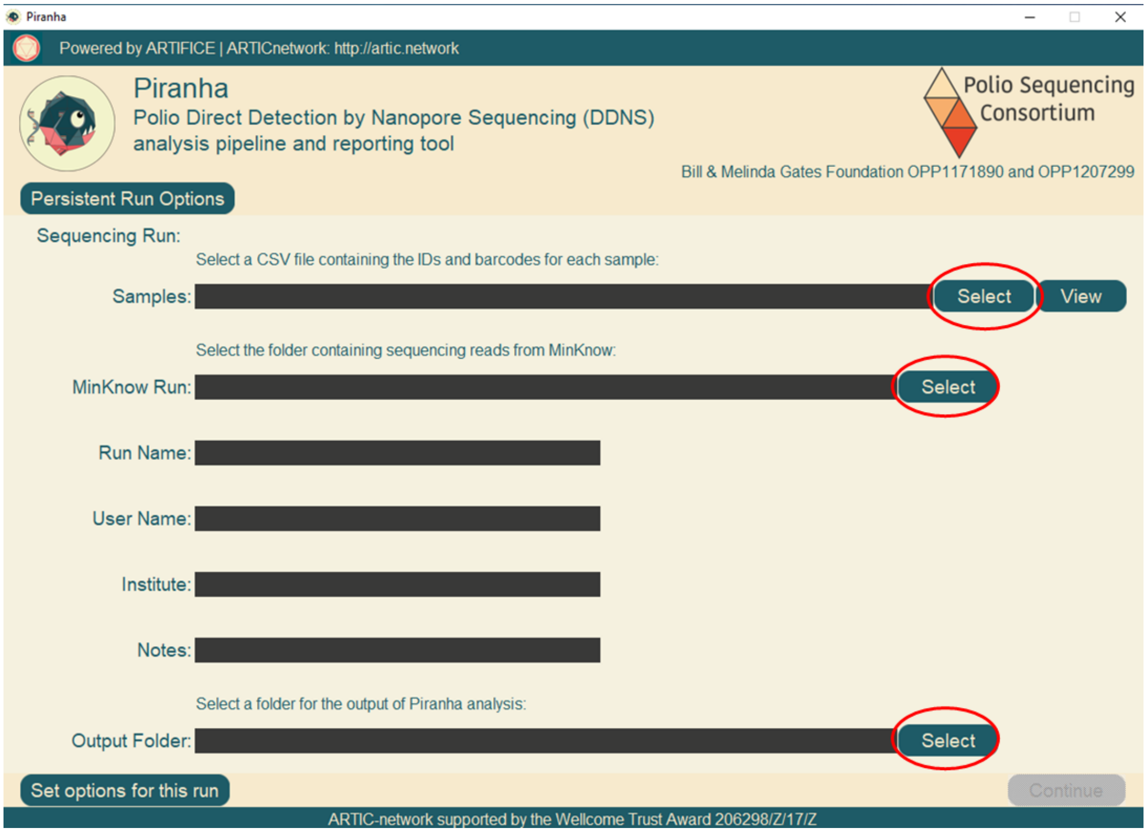
PIRANHA software version None installed

Piranha is the primary analysis pipeline for the DDNS polio detection platform.

An internet connection and a Docker install is required to install or update software

ARTIC-network supported by the Wellcome Trust Award 206298/Z/17/Z

3.4 The window to setup / input data for your run will pop up as shown below.



Piranha

Powered by ARTIFICE | ARTICnetwork: http://artic.network

Piranha
Polio Direct Detection by Nanopore Sequencing (DDNS)
analysis pipeline and reporting tool

Polio Sequencing Consortium

Bill & Melinda Gates Foundation OPP1171890 and OPP1207299

Sequencing Run:

Select a CSV file containing the IDs and barcodes for each sample:

Select the folder containing sequencing reads from MinKnow:

Select a folder for the output of Piranha analysis:

ARTIC-network supported by the Wellcome Trust Award 206298/Z/17/Z

Fill out the three fields circled red by clicking "Select" next to:

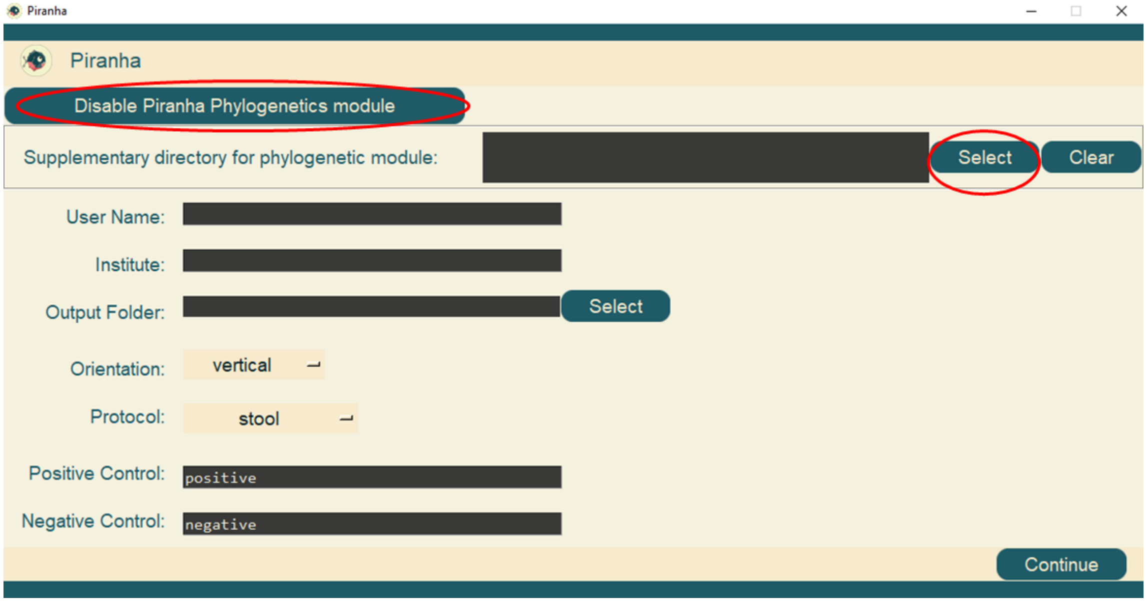
a. "**Samples**" to supply the location of your barcode.csv file which informs piranha which barcodes to analyse.

b. "**MinKnow Run**" to provide the location of your sequencing data (the demultiplexed fastq files). Usually this will be the fastq_pass folder.

> **Note**
>
> PiranhaGUI is configured to detect and access FASTQ read data within the specified directory and subdirectories that correspond to each barcode in the sequencing run with FASTQ sequencing read files

c. "**Output Folder**" to specifiy where PIRANHA will place its' report files.

3.5 Click on "**Persistent Run Options**" to allow you to provide the supplementary sequences for the analysis, enable the "Piranha Phylogenetic module" or check that the "Piranha Phylogenetics module" is enabled, and also to set your own default options.



Supplementary directory for phylogenetic module:

User Name:

Institute:

Output Folder:

Orientation: vertical

Protocol: stool

Positive Control: positive

Negative Control: negative

a. Under "**Supplementary directory for phylogenetic module**" select the location of your laboratory sequence database. This contains all the vp1_sequences.fasta files from previous DDNS runs.

b. Under “**Orientation**”, select the orientation of the barcodes in your barcode primer plate. Horizontal has well A1 barcode01, A2 barcode02, A3 barcode03. Vertical has well A1 barcode01, B1 barcode02, C1 barcode03 and so on.

c. Under "**Protocol**", select the sample type. Default "stool". Options "environmental" or "stool"

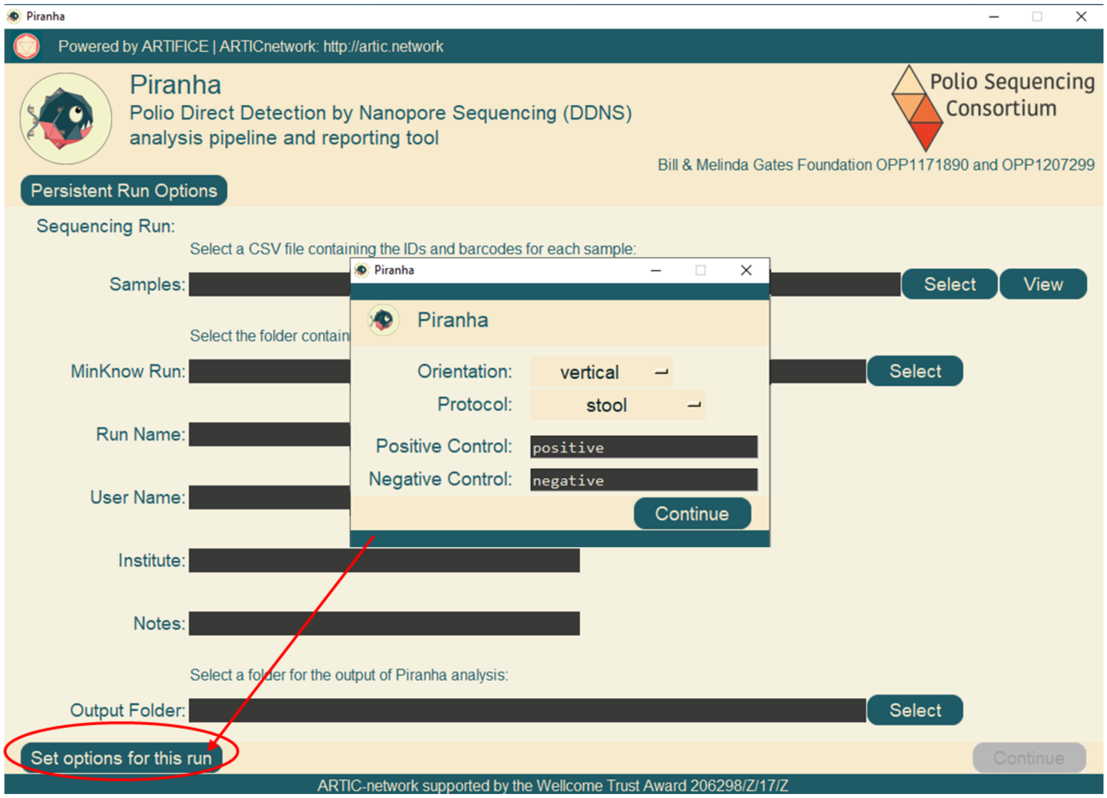
d. Enter or change the names of the positive and negative controls if different from the default. if you have multiple positive and negative controls, set them by separating the names with a comma. e.g "positiveEx1, positiveEX2" or "negativeEX, negativePCR1, negativePCR2"

e. Click "**Continue**" to confirm any changes.

Note

All the options selected under persistent run options are permanently set and will remain every time you launch Piranha.
With the button "Set options for this run" you can also set options for the analysis but these will not apply to future runs.

Alternatively, the button "**Set options for this run**" on the main run window can also be used to set options for the analysis but these will not apply to future runs.


Piranha
Powered by ARTIFICE | ARTICnetwork: http://artic.network
Polio Sequencing Consortium
Bill & Melinda Gates Foundation OPP1171890 and OPP1207299
Persistent Run Options
Select
View
Select
Select
Set options for this run
Continue
ARTIC-network supported by the Wellcome Trust Award 206298/Z/17/Z
Piranha
Continue

Piranha

Polio Direct Detection by Nanopore Sequencing (DDNS)
analysis pipeline and reporting tool

Sequencing Run:

Select a CSV file containing the IDs and barcodes for each sample:

Samples:

Select the folder contain

MinKnow Run:

Run Name:

User Name:

Institute:

Notes:

Select a folder for the output of Piranha analysis:

Output Folder:

Orientation: vertical

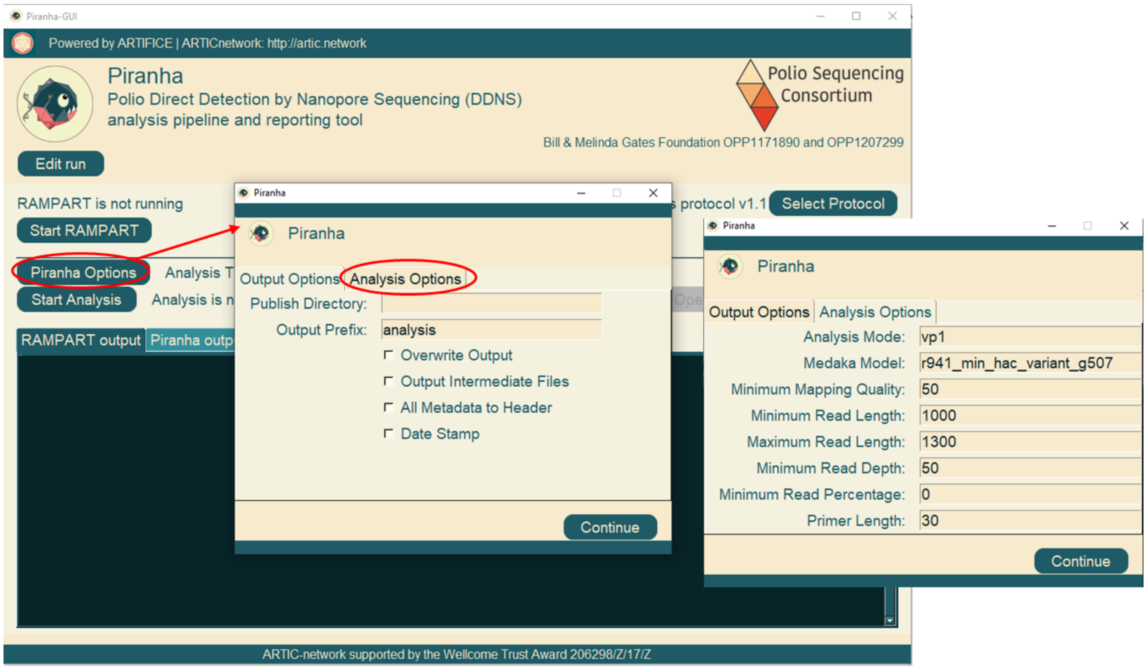
Protocol: stool

Positive Control: positive

Negative Control: negative

3.6 Before starting the analysis, click the "**Piranha Options**" button to set the "**Analysis options**". When running PIRANHA for stool samples set the following options:

a. Minimum read length: 1000
b. Maximum read length: 1300
c. Minimum depth: 50
d. Minimum read percentage: 0


Piranha-GUI
Powered by ARTIFICE | ARTICnetwork: http://artic.network
Piranha
Polio Direct Detection by Nanopore Sequencing (DDNS)
analysis pipeline and reporting tool
Polio Sequencing Consortium
Bill & Melinda Gates Foundation OPP1171890 and OPP1207299
Edit run
RAMPART is not running
Start RAMPART
Piranha Options
Analysis T
Start Analysis
Analysis is n
protocol v1.1
Select Protocol
RAMPART output
Piranha outp
Ope
ARTIC-network supported by the Wellcome Trust Award 206298/Z/17/Z
Continue
Continue

Piranha

Output Options | Analysis Options

Publish Directory:

Output Prefix: analysis

☐ Overwrite Output

☐ Output Intermediate Files

☐ All Metadata to Header

☐ Date Stamp

Piranha

Output Options | Analysis Options

Analysis Mode:	vp1
Medaka Model:	r941_min_hac_variant_g507
Minimum Mapping Quality:	50
Minimum Read Length:	1000
Maximum Read Length:	1300
Minimum Read Depth:	50
Minimum Read Percentage:	0
Primer Length:	30

3.7 Click "**Continue**" to return to the run window. For quicker analysis, select the maximum number of threads for the Piranha pipeline.

Note

All settings are saved even after closing the GUI, so there is no need to reapply the changes every time.

3.8 Click "**Start Analysis**". A progress bar will show how much has been completed.

4 Verify that the Piranha run completed successfully. The last line before “###PIRANHA SOFTWARE FINISHED###” should say “**Generating: /data/run_data/output/piranha_output/report.html**”. If that is not the case, then Piranha may have encountered an error. Check and resolve the error before continuing.

5 After the PIRANHA analysis is complete, you can view the output by clicking "**Open Output**". The "piranha_output" folder will contain the following:

5.1 barcode_reports: This contains the HTML reports for each sample barcode with information on the mutations when compared to respective reference sequence for each poliovirus serotype. The snipit plot shows the percentage co-occurrence of SNPs called against reference and it can give an idea if mixed populations are present within the sample.

5.2 published_data folder: This contains a folder for each barcode and a file called vp1_sequences.fasta containing all consensus sequences for each samples' classified haplotype

5.3 detailed_run_report: This contains all sequencing results appended to your barcode.csv file. Any additional metadata can be added to the "detailed_run_report.csv" file. This final report is the definitive document containing all the data for the sequencing run and can be uploaded for storage and data rows shared when reporting detections.

5.4 report : This contains the sample summary information specific to the sequences generated for each sample.

6 PIRANHA will also create a html report, where it will have confirmed whether:
a. The positive control(s) yielded at least 500 sequences that have mapped to Coxsackievirus A20.
b. The negative control(s) yielded less than 50 reads mapped to poliovirus or NPEVs

Perform Run Quality Control Checks

7 Perform the RunQC by checking the following points:
a. Were there more than 500 reads for the positive control mapped to Coxsackievirus A20?
b. Were there less than 50 reads mapped for the negative control?

If all answers are Yes, then enter Pass in the column "RunQC" of the detailed_run_report.
If any answers are No, enter Fail in the "RunQC" column and add an explanation of the failure in the "comments" field, then complete the following troubleshooting steps:

7.1 Too few positive control reads:
a. Confirm that your earlier positive control check has passed QC checks. Repeat the library pooling and confirm the presence of your library after the cleanup steps using a Tapestation or a Qubit fluorometer.
b. Check that you are ligating the correct adaptor (LA) and are using the short fragment buffer (SFB) during library preparation and that none of your end-prep or ligation enzymes are expired

7.2 Too many negative control reads:
a. Confirm that your earlier negative control check has passed QC checks.
b. Rewash the flow cell with a DNAse wash and repeat the library pooling and sequencing run.

Sample Quality Control Checks

8 If EPIDs and metadata are only available after the sequencing run, add them to the detailed_run_report.

9 Generate a phylogenetic tree comparing your run to all prveious data by using one of the methods below:

a. Using the piranha phylogenetic tree module: Open the piranha html report in the output folder and confirm that the piranha report contains a phylogenetic tree and the table that shows the IDs of identical sequences.
b. Alternatively, use Geneious or Nextstrain to generate an alignment and phylogenetic tree of the new sequences generated by PIRANHA and your laboratory sequence QC database with the DDNS sequences from prior runs.

10 Perform individual sample QC checks for all samples and enter Pass/Fail in the column “SampleQC” of the detailed_run_report. Samples should be marked as “Fail” if:

a. A pair of samples with the same EPID (i.e. from the same case) are 3 or more nucleotides different from each other over VP1.
b. A sample is identical to another sample in the run that does not have the same EPID (i.e. they are from different cases), unless the sequences are both the same Sabin serotype with less than 1 mutations from the original vaccine.

11 Check for bleed-through between runs if the flow cell has been reused and for amplicon contamination; compare your current run to previous DDNS run to see if there is a sample with the same barcode that yielded a highly similar sequence (less than 3 nucleotides different over VP1). If a sample matches a sequence from a previous run in this manner, mark it as a “Fail”. Refer to your phylogenetic tree to identify similar sequences with matching barcodes.

Note

Identical sequences appearing at low read numbers over multiple samples could possibly indicate cross-contamination. Such samples should be analysed with care. If contamination is suspected, mark the sample as "Fail".

12 Any sample that are not marked as “Fail” can be marked “Pass” in column “SampleQC”, and the column “SampleQCChecksComplete” can be marked as “Yes” for the passed samples. It should be marked as “Fail” for the failed samples.

13 Samples that pass both SampleQC can be reported, and should have:
a. “SampleQCCheckComplete” marked as “Yes”.
b. “ToReport” should be marked as “Yes”.

14 Samples marked as fail in “SampleQC” should have:

a. “SampleQCCheckComplete” marked as “Fail”.
b. “ToReport” should be marked as “No”.
c. An explanation added in “QCComments”

Note

Note: These samples should be repeated on a later run (with column “IsQCRetest” flagged “Yes” in
that run). These samples should not be grouped together upon retesting and should use different barcodes.

15 When the report has been submitted, complete the “DateReported” for samples that were included.

Note

The following files are required when reporting data following QC
a. MinKNOW Run Report
b. Piranha output report
c. barcode_reports
c. QC detailed run report (contains summary of the QC sample results). Note: ensure to include any QC anomalies and corrective actions
d. Fasta files of positive samples

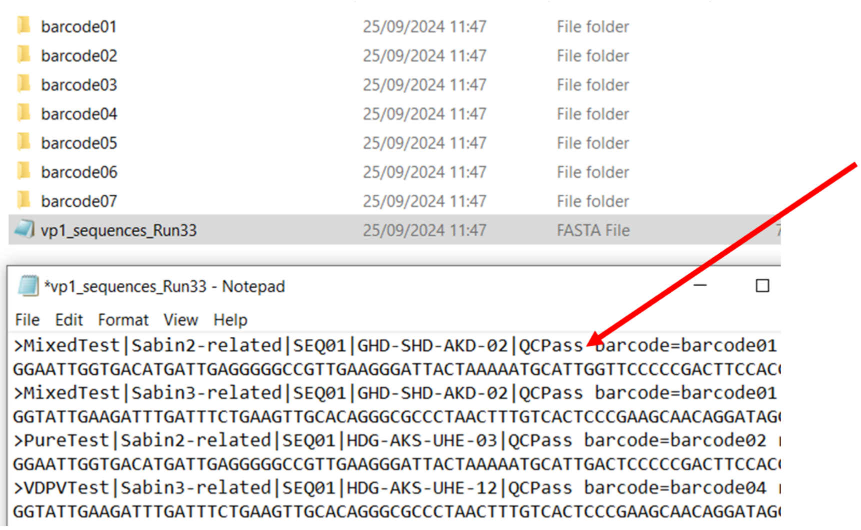
16 If on retesting:
a. Samples yield the same sequence as in the initial run, mark them on each run as “Pass”. If a pair of samples (same EPID) yield the the same sequences as the first run, but these differ by 3 or more nucleotides they can be marked as “Pass”.
b. Samples no longer yield sequences, mark them on each run as negative under “DDNSClassification” and add a note of“Likely contamination” in “QCComments” in the original run. These sequences can be removed from further analyses of the sequencing run and should not be submitted.

17 After retesting is complete, mark “SampleQCChecksComplete” in the original run as “Yes”

18 Any data that is not avilable before or during the run (e.g. Sanger sequencing results) can be added to the detailed_run_report.csv when available.

19 The filename of vp1_sequence.fasta should be appended to include the run name (e.g. "vp1_sequences.fasta" becomes "vp1_sequences_run21.fasta". Within the file sequences can be appended with "QC_Pass" if the sequence has passed the Run and Sample QC checks (as

shown below). The file should then be transferred to the laboratory sequence QC database folder



barcode01	25/09/2024 11:47	File folder
barcode02	25/09/2024 11:47	File folder
barcode03	25/09/2024 11:47	File folder
barcode04	25/09/2024 11:47	File folder
barcode05	25/09/2024 11:47	File folder
barcode06	25/09/2024 11:47	File folder
barcode07	25/09/2024 11:47	File folder
vp1_sequences_Run33	25/09/2024 11:47	FASTA File

```
>MixedTest|Sabin2-related|SEQ01|GHD-SHD-AKD-02|QCPass barcode=barcode01
GGAATTGGTGACATGATTGAGGGGGCCGTTGAAGGGATTACTAAAAATGCATTGGTTCCCCCGACTTCCAC
>MixedTest|Sabin3-related|SEQ01|GHD-SHD-AKD-02|QCPass barcode=barcode01
GGTATTGAAGATTTGATTTCTGAAGTTGCACAGGGCGCCCTAACTTTGTCACTCCCGAAGCAACAGGATAG
>PureTest|Sabin2-related|SEQ01|HDG-AKS-UHE-03|QCPass barcode=barcode02
GGAATTGGTGACATGATTGAGGGGGCCGTTGAAGGGATTACTAAAAATGCATTGACTCCCCCGACTTCCAC
>VDPVTest|Sabin3-related|SEQ01|HDG-AKS-UHE-12|QCPass barcode=barcode04
GGTATTGAAGATTTGATTTCTGAAGTTGCACAGGGCGCCCTAACTTTGTCACTCCCGAAGCAACAGGATAG
```

Sequences passing QC are annotated according to the barcode.csv

Protocol references

1. https://github.com/polio-nanopore/piranhaGUI.
2. https://github.com/polio-nanopore/piranha.
3. Áine O'Toole, Rachel Colquhoun, Corey Ansley, Catherine Troman, Daniel Maloney, Zoe Vance, Joyce Akello, Erika Bujaki, Manasi Majumdar, Adnan Khurshid, Yasir Arshad, Muhammad Masroor Alam, Javier Martin, Alexander G Shaw, Nicholas C Grassly, Andrew Rambaut, Automated detection and classification of polioviruses from nanopore sequencing reads using piranha, *Virus Evolution*, Volume 10, Issue 1, 2024, veae023, https://doi.org/10.1093/ve/veae023