

Day 4: MPXV phylogenetics & interpretation

MPXV phylogenetics tutorial: [Áine O'Toole](#)

Quick analysis using NextClade, QC

Background data compilation with Pathoplexus

Running squirrel with EPI2ME for QC and APOBEC3 reconstruction

Learning Objectives	1
Primer on MPXV genomic epidemiology	2
MPXV diversity & genomics	2
APOBEC3 family of cytosine deaminases	2
Virus evolution during an outbreak	3
What is a FASTA file?	4
How to read a tree	4
What can MPXV genomes tell us?	5
Tutorial: Analysis of MPXV genomes	6
Running NextClade	6
Compiling a background dataset	10
Exercise: Repeat on newly generated data	15
Introduction to squirrel	16
Installing squirrel using EPI2ME	16
Running squirrel for sequence QC	19
Understanding the squirrel QC output	24
Running squirrel using EPI2ME for APOBEC3 reconstruction	25
References	30
Useful links	30

Learning Objectives

- Understanding of MPXV genomic epidemiology
- Understanding of host enzyme APOBEC3, what it means for MPXV
- How to run NextClade on new data & interpret results
- How to collect a background dataset to include in analysis
- How to install squirrel in EPI2ME
- How to run squirrel in QC mode
- How to run APOBEC3 reconstruction with squirrel
- How to analyse and interpret an outbreak of MPXV

Primer on MPXV genomic epidemiology

MPXV diversity & genomics

MPXV is a double stranded DNA virus of the family Poxviridae that causes mpox. It has a large, complex genome (~200 kb) and a large gene repertoire. There are two major clades of MPXV, Clades I and II, and recently each has been subdivided into Clades Ia and Ib, and IIa and IIb.

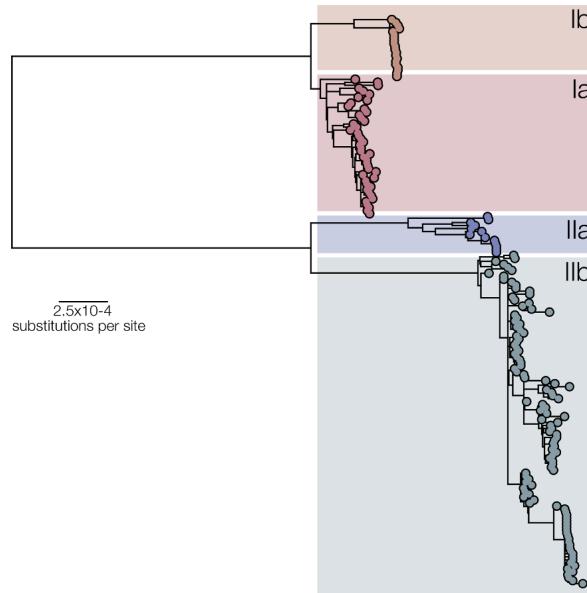


Figure 1. The known genetic diversity of MPXV is divided into two distinct clades (I and II), and further subdivided into Clades Ia and Ib, and Clades IIa and IIb (Wawina-Bokalanga et al 2025). Historically, cases of Clade I were reported primarily in Central Africa and cases of Clade II were majoritively reported in West Africa. In recent years, cases of Clade II have been reported across the globe.

The genomes themselves have many repetitive and low-complexity regions that can be challenging to both sequence and analyse. The analytic challenges can be overcome with automated pipelines such as NextClade and squirrel, and alignments and phylogenies can be produced easily without need for the command line. The major challenge that remains for robust genomic epidemiology of MPXV is in knowing how to correctly interpret these results.

MPXV has existed for most of its evolutionary history in a rodent reservoir, and recent spillover events have led to persistence in the human population. In humans, MPXV evolves at a faster rate due to the action of the human enzyme APOBEC3.

APOBEC3 family of cytosine deaminases

APOBEC3 is a family of host enzymes that can bind single stranded DNA and deaminate cytosine bases to uracil. Due to DNA synthesis, this change gets propagated back onto the DNA as a T and what we observe as a result is a C->T mutation. This can then be observed either on the positive strand (as a C->T mutation), or on the negative strand (as a G->A mutation). The function of APOBEC3 enzymes is varied, but has been well studied for its

antiviral function against HIV. APOBEC3F, which is the human enzyme thought to be acting in MPXV infections, has a preference for binding and deaminating TC dinucleotides, so the observed mutations in the MPXV population are TC->TT and GA->AA dinucleotide mutations.

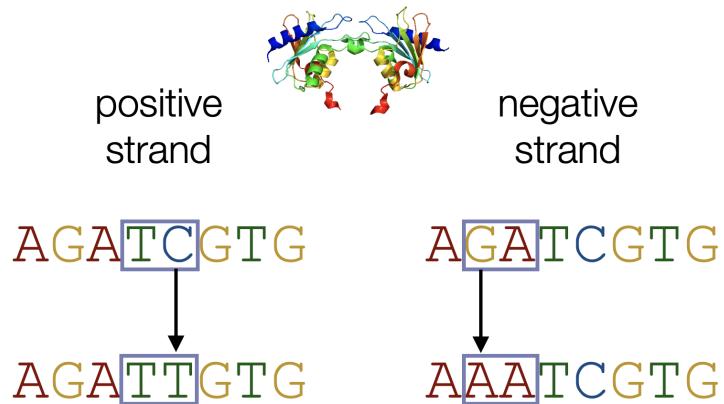


Figure 2. APOBEC3 is a family of host cytidine deaminases with a track-record of anti-viral function. Humans and other primates have 7 different APOBEC3 genes, whereas rodents only have a single APOBEC3 (Conticello 2008). Human APOBEC3F is thought to be responsible for the enrichment of TC->TT and GA->AA mutations that are characteristic of sustained transmission of MPXV in the human population (Suspène et al 2023). This signature is not observed when the virus circulates in the rodent population (O'Toole et al 2023).

These mutations are primarily antiviral in function, and are not indicative of MPXV adaptation to humans. To support this, the mutations observed are the residual least deleterious mutations that remain after natural selection has removed those with high fitness-costs to the virus. This means that there is not a constant molecular clock across the MPXV phylogeny, and branch lengths in the rodent population and in the human population need to be interpreted differently. Mutations that are not consistent with APOBEC3 editing will occur at a much lower rate, expecting one couple of years in the MPXV population on average. Genomes with many unique non-APOBEC3 mutations are likely a result of contamination, sequencing issues or bioinformatic error.

Virus evolution during an outbreak

In a virus outbreak, new cases arise as the virus is transmitted from person to person. Often, the number of sampled cases is less than the true number of cases that have occurred in the outbreak (Figure 3A). As the virus transmits, mutations can occur that can be the result of errors during virus replication or as a result of human host defense (e.g. APOBEC3 editing). Not every transmission event gives rise to a mutation and how often a mutation is observed is influenced by a variety of factors, for example the nature of the virus (e.g. RNA, DNA) and the time between observed cases (serial interval) (Drummond et al 2003). Over time mutations accumulate in the virus population, however very deleterious will be selected out through natural selection and likely not observed. The virus genome sequences of sampled cases can allow us to reconstruct the relationships between cases (Figure 3B).

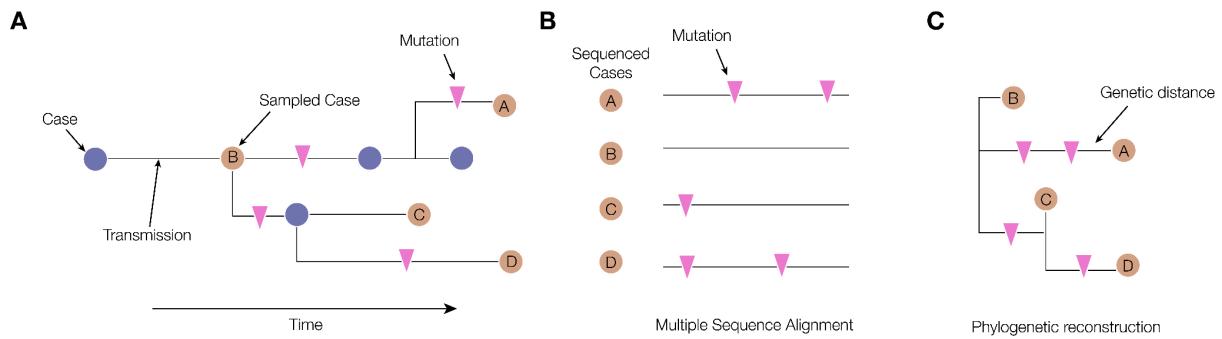


Figure 3. We can use the mutations that occur as the virus is transmitted from host to host to reconstruct the evolutionary history of the virus in the form of a phylogenetic tree. Not every transmission event will give rise to an observed mutation in the population.

What is a FASTA file?

FASTA is a standard format used to represent genetic sequence data. For MPXV, we represent our consensus genome sequences in FASTA format. A FASTA file can consist of one or more sequence records, represented by a header line and then the sequence itself. In the below example there are two sequence records.

```
>case001|MPXV|2024-12-03
TTACAGATCATTATTCACAAAATATTAACATATACGTTATTATATGATGTTAACGTGTAATTAA
TAAACATTATTTATGATGCAATTGCTGACAACCTAGATTGGCATAAGGATATTGATAAGCTCTA
CGAGAATATATTGTTGGACGTTATCGTTACGAAATAGTTGAGACATCAGAAAGAGGTTAATATT
TTTGAGACCATCGAAGAGAGAAAGAGAATAAAATATTGTTGAAACTTTTAA
AGACAANNNNNNNNNNNNNNNNNNNNNTAGTGATCATATCGTATCACATATTGAAACAG
>case002|MPXV|2024-12-06
TACAGATCATTATTCACAAAATATTAACATATACGTTATTATATGATGTTAACGTGTAATTAA
AACATTATTTATGATGCAATTGCTGACAACCTAGATTGGCATAAGGATATTGATAAGCTCTAC
GAGAATATATTGTTGGACGTTATCGTTACGAAATAGTTGAGACATCAGAAAGAGGTTACCATT
TTCAGATGAATAGAGTTATCGATTAGACACATGCTTGAGTTGTTGAATCGATGAGTGAAGT
ATCATCGGTTGCACCTTCAGATGCCGATCCGTCGACATACTGAATCCATCCTG
```

In a FASTA file, the header line always begins with a “>” character, followed by the sequence ID. The ID should not contain any whitespace (like space characters “ ”, or tab characters). It is good practice to include relevant metadata in the header, such as the date of sampling. The sequence follows on the next line and it can be represented in a single long line, or for readability is often represented across a number of lines. In this example the MPXV DNA sequence is shown as ACGT nucleotide characters and any unknown sites are represented as the ‘N’ ambiguity character.

How to read a tree

For a primer on “how to read a phylogenetic tree”, see the detailed guide on the ARTIC website at: <https://artic.network/how-to-read-a-tree.html>

Background datasets: *The need for context*

Without appropriate context, you are interpreting any newly generated sequence data with only part of the information. Sometimes no other relevant sequences are available and when interpreting the data we need to be aware of that. However, public databases can be very valuable resources for your analysis and for the interpretation of results. In this tutorial, you will learn how to access some background genome sequence data to include with your analysis.

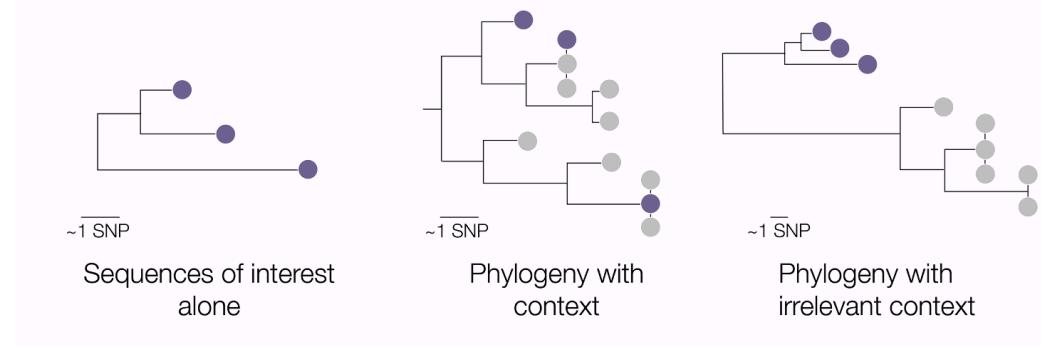


Figure 4. If all that is used to construct the phylogeny is your sequences of interest alone, it appears that all sequences are linked. However, when relevant genomes are included for additional context this may not be the case. Choosing appropriate sequences for context can be tricky, and you want to include as many as possible without hindering your ability to analyse the data efficiently. Choosing irrelevant contextual data will not help with interpretation.

What can MPXV genomes tell us?

Some of the first questions that are necessary to answer are:

1. What clades/ lineages are present in the data
2. Whether the observed genomes are part of a known human outbreak or whether they represent a novel spillover event
3. If part of an outbreak, which outbreak are they a part of
4. Are the cases linked
5. What are the most related genomes in public data, possible source attribution

Tutorial: Analysis of MPXV genomes

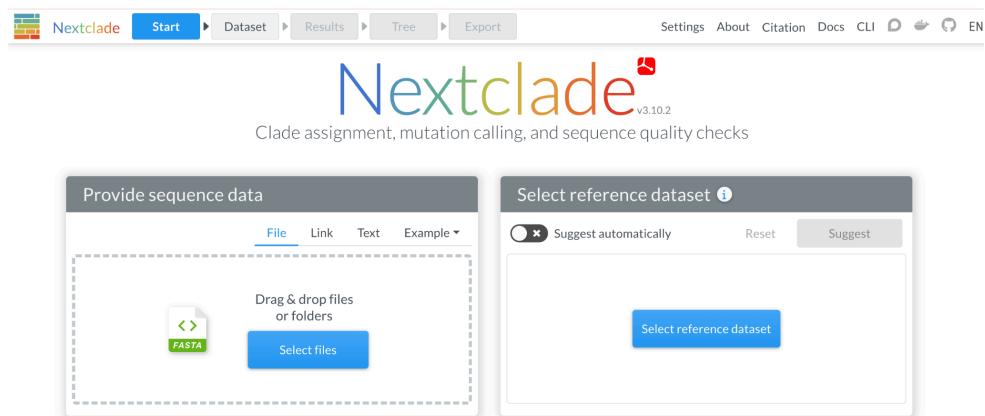
As part of this tutorial, you will become familiar with running nextclade and squirrel (through the EPI2ME interface). You can run the example outbreak dataset through the tutorial and also the newly generated sequence data to test your ability to apply these new skills to new data.

Download the example data here: https://aineotoole.co.uk/initial_cases.fasta

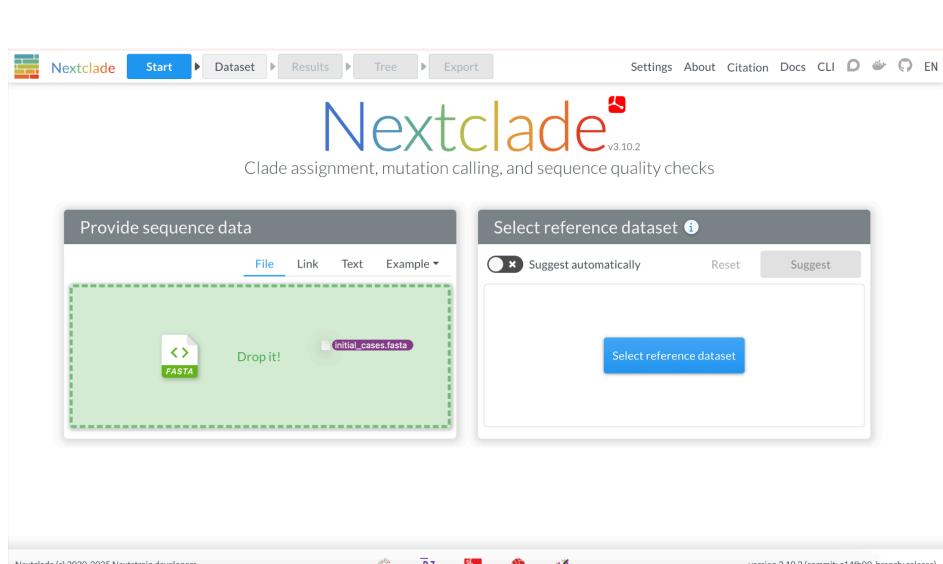
You will have the new consensus genomes from the field bioinformatics tutorial

Running NextClade

Citing NextClade: Aksamentov, I., Roemer, C., Hodcroft, E. B., & Neher, R. A., (2021). Nextclade: clade assignment, mutation calling and quality control for viral genomes. Journal of Open Source Software, 6(67), 3773, <https://doi.org/10.21105/joss.03773>



1. Navigate to the NextClade website at <https://clades.nextstrain.org/>.



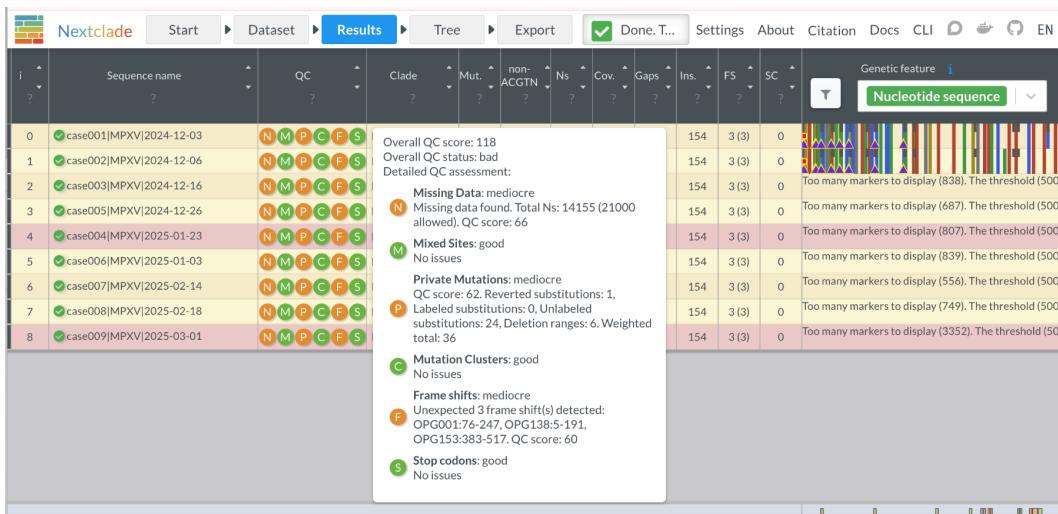
2. Drag and drop the initial_cases.fasta file you've just downloaded into the sequence data box, or select the file in the file browser.

The screenshot shows the NextClade web application. In the top navigation bar, the 'Start' button is highlighted. Below it, the main title 'NextClade' is displayed with the subtitle 'v3.10.2' and the tagline 'Clade assignment, mutation calling, and sequence quality checks'. The 'Add more sequence data' section on the left has tabs for 'File', 'Link', 'Text', and 'Example'. It includes a 'Drag & drop files or folders' area with a FASTA icon, a 'Select files' button, and a 'Sequence data you've added' list containing 'initial_cases.fasta (1.77 MB)'. The 'Selected reference dataset' section on the right has a 'Suggest automatically' toggle (ON), a 'Reset' button, and a 'Re-suggest' button. It shows a card for 'Mpxox virus (Clade I)' with a 100% official reference from Zaire_1979-005 (DQ011155.1) updated at 2024-11-19 14:18:53 (UTC). Buttons for 'Open tree' and 'Load example' are shown below the card. A note at the bottom indicates 'Multiple matching datasets.' with a link to change the reference dataset.

- In the reference dataset section, toggle the “suggest automatically” switch ON. Select “Suggest” to allow the appropriate reference to be selected. NextClade believes this file contains MPXV Clade I data. Select “Run” in the bottom right to run the pipeline.

The screenshot shows the 'Results' page of NextClade. The top navigation bar has 'Dataset' and 'Tree' buttons. The main area displays a table with columns: i, Sequence name, QC, Clade, Mut., non-ACGTN, Ns, Cov., Gaps, Ins., FS, SC, and ? (empty). The table rows represent individual sequences, each with a green checkmark and a unique ID (0-8). The 'Sequence name' column lists 'case001|MPXV|2024-12-03' through 'case009|MPXV|2025-03-01'. The 'Clade' column consistently shows 'Ib'. The 'non-ACGTN' column shows values like 113, 113, 112, etc. The 'Ns' column shows values like 0, 0, 0, etc. The 'Cov.' column shows values like 13573, 13543, 14202, etc. The 'Gaps' column shows values like 274, 274, 274, etc. The 'Ins.' column shows values like 154, 154, 154, etc. The 'FS' column shows values like 3 (3), 3 (3), 3 (3), etc. The 'SC' column shows values like 0, 0, 0, etc. To the right of the table, there are buttons for 'Genetic feature' (set to 'Nucleotide sequence'), 'Relative to' (set to 'Reference'), and a large 'T' icon. A note at the bottom right says 'Too many markers to display (838). The threshold (500) can be increased in "Settings".'

- When the analysis is complete you should see a results page similar to the above. NextClade has a variety of checks that flag problematic features in genomes, such as low coverage (lots of Ns) and unique mutations in a given sequence.



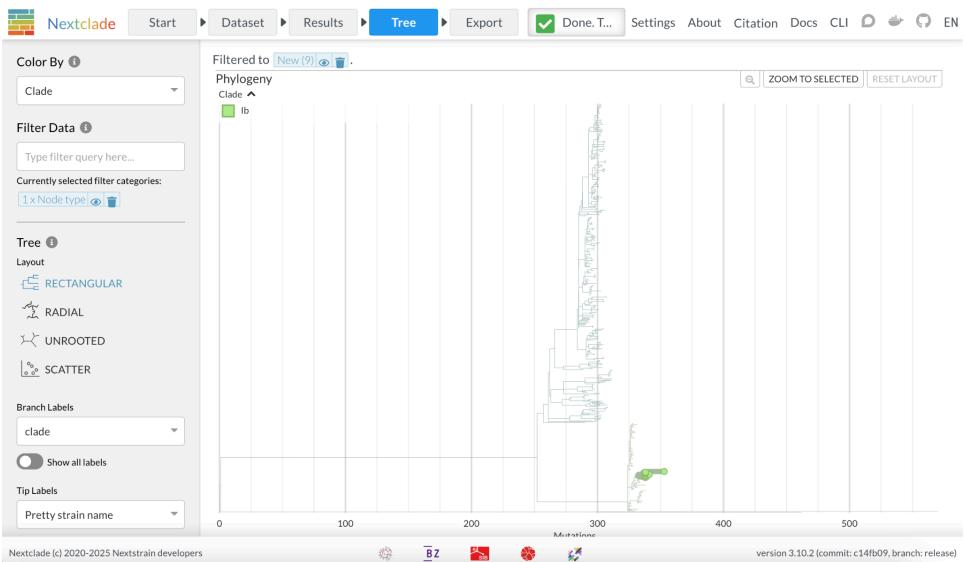
5. Hover over case004 to see why NextClade has flagged it. You can see it has lots of Ns, lots of Private mutations (mutations only seen in that sequence) and some unexpected frame shifts. A frameshift is caused by an insertion or deletion mutation that causes the *frame* of a coding sequence to change or *shift*. This is flagged here because frameshift mutations are rare as they disrupt gene function. Similarly, hover over case009 and see that it has more Ns than the others.

Question:

All sequences contain the flagged frameshifts, what does this say about your data?

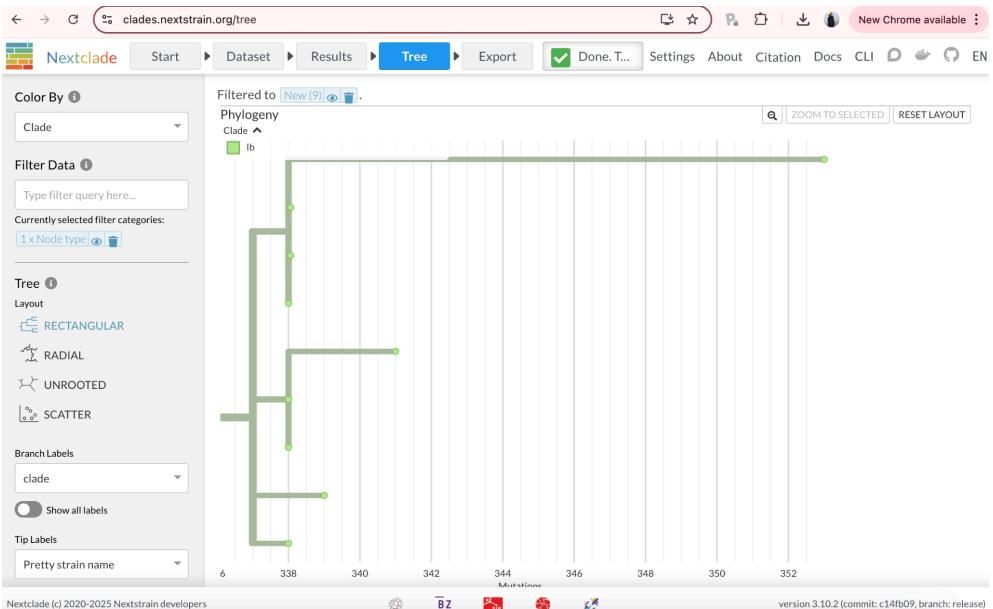
Question:

Are there any sequences in this dataset you think may have issues that could impact your interpretation of the results?



6. Navigate to the 'Tree' panel by clicking the 'Tree' button on the top of the window. You can see where the data lies in the diversity of Clade I.

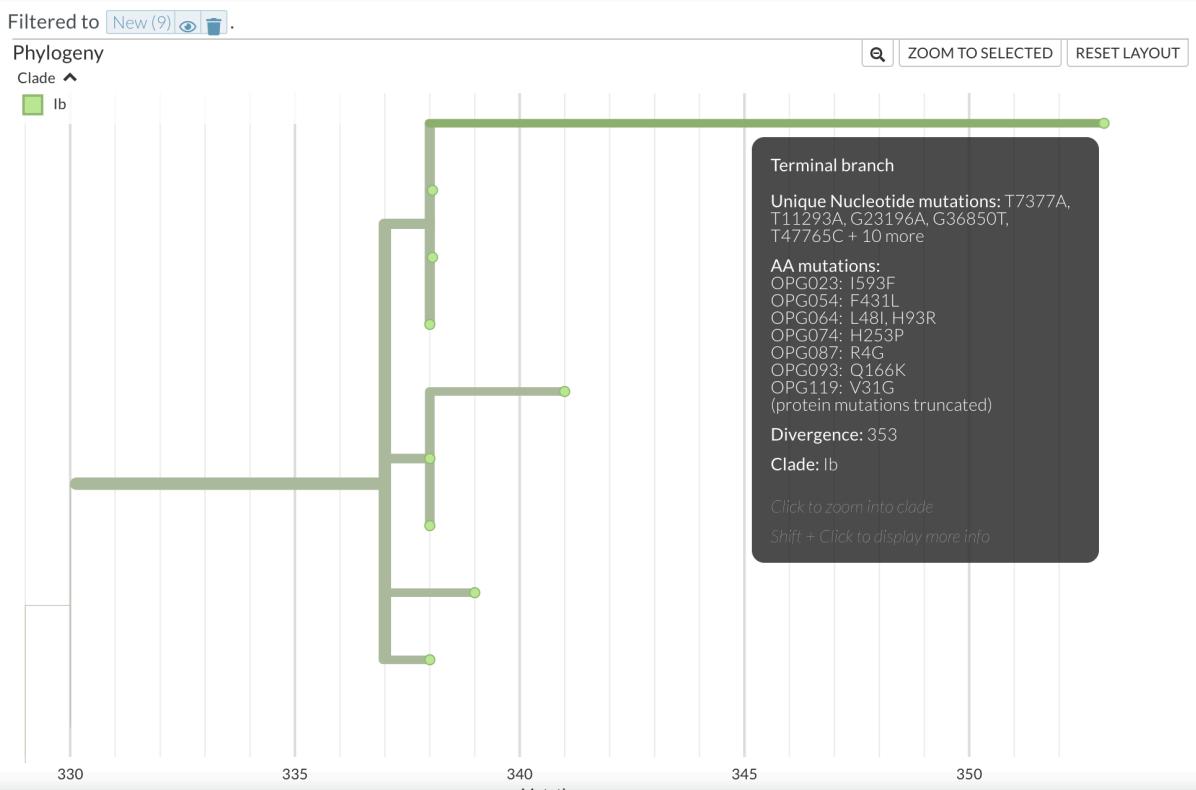
Question: What subclade is the data? Does the data cluster together?



7. Click on your sequences in the tree to zoom in to the data.

Question:

What can you say about the sequences?
How much diversity is there?
Do you think the cases are linked?



8. Hover over the branch leading to case004 to look at the SNPs present in that sequence only.

Question:

How many unique mutations are there?

Based on what you know about the rate of MPXV evolution, do you think these mutations are real biological variation?

Why/ why not?

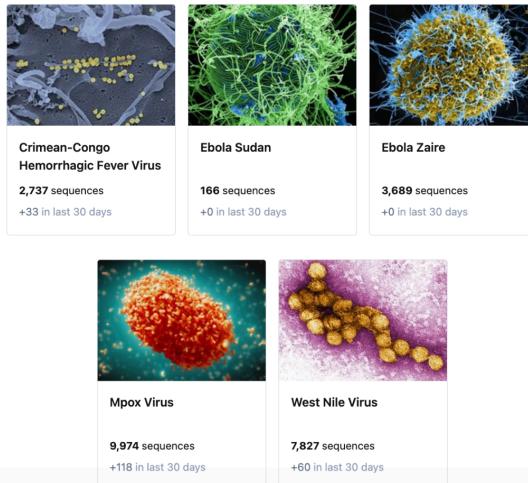
What other information might help us interpret this data?

Compiling a background dataset

Next, we will gather a background dataset. There are a number of public repositories for data that we will hear more about later in the workshop. Today, we will use Pathoplexus to search for and download some relevant background data.

Welcome to Pathoplex!

Pathoplex is a new, open-source database dedicated to the efficient sharing of human viral pathogen genomic data, fostering global collaboration and public health response.



1. Navigate to pathoplex.org. Scroll down and have a read through the homepage to learn more about Pathoplex.

Search

[Add Search Fields](#) [Reset](#) [Help](#)

Accession

Data use terms

Clade

Collection subdivision level 1

Collection country

Length

From

To

Lineage

Search returned 9,974 sequences

[Customize columns](#)
[Download all entries](#)

ACCESSION VERSION	CLADE	LINEAGE	COLLECTION DATE ▾	COLLECTION COUNTRY	COLLECTION SUBDIVISION LEVEL 1	AUTHORS	AUTHOR AFFILIATIONS	LENGTH
PP_0014AN4.1	IIb	E.1	2025-02-07	Portugal		Sobral, Daniel; ...	Instituto Nacio...	197,223
PP_0014AM6.1	IIb	E.1	2025-02-07	Portugal		Sobral, Daniel; ...	Instituto Nacio...	197,223
PP_00147XP.1	IIb	E.1	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,627
PP_00147YM.1	IIb	F.4	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,619
PP_00145MB.1	IIb	C.1	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,625
PP_001454C.1	IIb	E.1	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,627
PP_00145N9.1	IIb	E.1	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,621
PP_00145P7.1	IIb	F.4	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,630
PP_00147SZ.1	IIb	E.1	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,627
PP_00147R1.1	IIb	F.4	2025-02	Germany		Brinkmann, A.; ...	Robert Koch Ins...	190,619

2. Click in to Mpox Virus to see the data available for MPXV.

Search returned 287 sequences								
							Customize columns	Download all entries
ACCESSION VERSION	CLADE	LINEAGE	COLLECTION DATE	COLLECTION COUNTRY	COLLECTION SUBDIVISION LEVEL 1	AUTHORS	AUTHOR AFFILIATIONS	LENGTH
PP_00116S.1	Ib	None	2025-01-06	Thailand	Nitijanontakij,...	Bamrasnaradura ...	195,645	
PP_00111X1.1	Ib	None	2025-01	USA	GA	Gigante, C.; Fe...	CDC, DHCPP-PRB	195,536
PP_0010NBP.1	Ib	None	2024-11-28	United Kingdom	Everall, I.; Gr...	UKHSA, Research...	186,229	
PP_0013JPW.1	Ib	None	2024-11-07	Democratic Repu...	Kinshasa	Wawina-Bokalang...	195,184	
PP_0013JXE.1	Ib	None	2024-11-07	Democratic Repu...	Kinshasa	Wawina-Bokalang...	196,388	
PP_0013JCK.1	Ib	None	2024-11-05	Democratic Repu...	Kinshasa	Wawina-Bokalang...	195,170	
PP_0013J9R.1	Ib	None	2024-11-05	Democratic Repu...	Kinshasa	Wawina-Bokalang...	196,280	
PP_0013HRT.1	Ib	None	2024-11-04	Democratic Repu...	Kinshasa	Wawina-Bokalang...	195,160	
PP_0013GZC.1	Ib	None	2024-11-01	Democratic Repu...	Kinshasa	Wawina-Bokalang...	196,371	
PP_0010K8Y.1	Ib	None	2024-11	USA	CA	Gigante, C. M.,...	CDC, DHCPP-PRB	195,321

3. From the quick NextClade analysis, we know our data is Clade Ib. You can play around with the different filters on the left hand panel to specify only Clade Ib, and filter by length, date, country etc. When you're ready to download a dataset, click "Download all entries" in the top right corner.

ACCESSION	LENGTH	EARLIEST RELEASE DATE
PP_00116S.1	195,645	2025-01-07
PP_00111X1.1	195,536	2025-01-23
PP_0010NBP.1	186,229	2024-12-13
PP_0013JPW.1	195,184	2025-02-04
PP_0013JXE.1	196,388	2025-02-04
PP_0013JCK.1	195,170	2025-02-04
PP_0013J9R.1	196,280	2025-02-04
PP_0013HRT.1	195,160	2025-02-04
PP_0013GZC.1	196,371	2025-02-04
PP_0010K8Y.1	195,321	2024-12-02
PP_0013JPW.1	196,348	2025-02-04
PP_0013JXE.1	196,398	2025-02-04
PP_0013JCK.1	196,408	2025-02-04
PP_0013J9R.1	196,370	2025-02-04

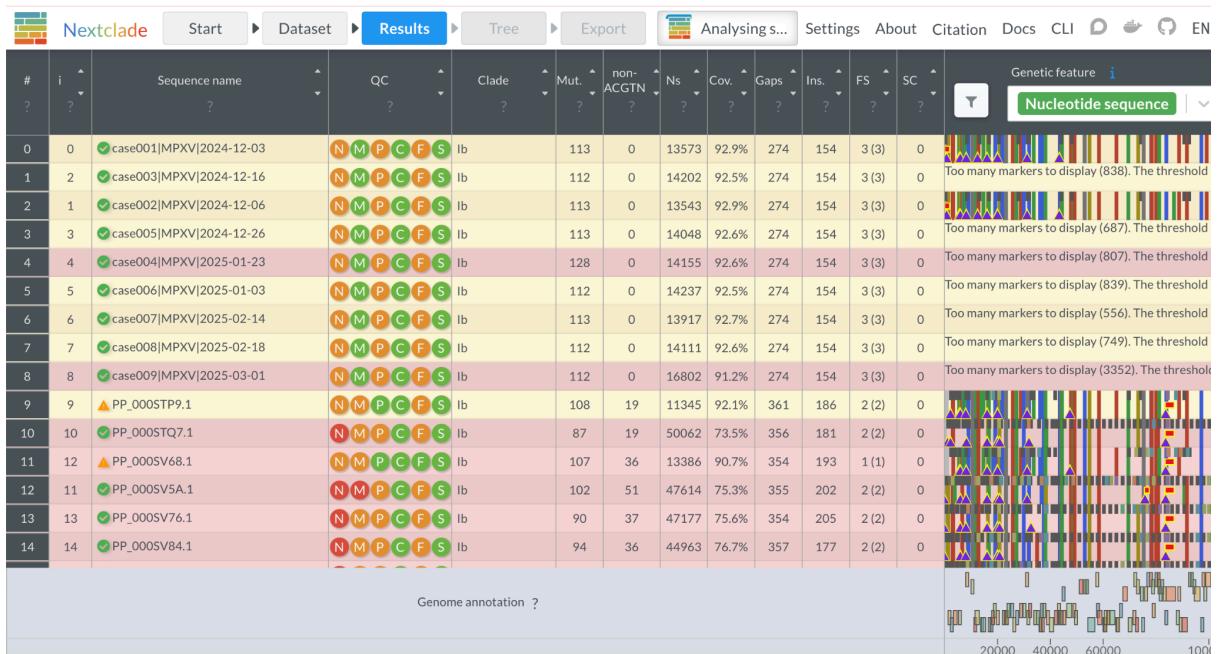
4. A dialogue box will pop up. You can decide whether you want to include restricted data or not. We will include it because our use of the data will just be for background purposes, which is allowed under the restricted data terms of use. To read more about the terms of use, see <https://pathplex.org/about/terms-of-use/restricted-data>. Select "raw nucleotide sequences" to download. The data can be downloaded without compressing, or with either Zstandard or Gzip compressions. We will download the sequences with no compression.

The screenshot shows the NextClade interface. At the top, there's a navigation bar with tabs: Start, Dataset, Results, Tree, Export, Settings, About, Citation, Docs, CLI, and a few others. Below the navigation bar is the NextClade logo with the text "v3.10.2". A sub-header reads "Clade assignment, mutation calling, and sequence quality checks".

The main area has two main panels:

- Add more sequence data:** This panel has tabs for File, Link, Text, and Example. It features a large dashed box for dragging and dropping files or folders, with a "Select files" button below it. Below this box is a list titled "Sequence data you've added" containing two items: "initial_cases.fasta (1.77 MB)" and "mpox_nuc_2025-03-09T0452.fasta (55.74 MB)".
- Selected reference dataset:** This panel shows a selected dataset: "Mpx virus (Clade I)" (official). It provides details: Reference: Zaire_1979-005 (DQ011155.1), Updated at: 2024-11-19 14:18:53 (UTC), Dataset name: nextstrain/mpox/clade-i. Buttons include "Suggest automatically", "Reset", "Suggest", "Open tree", "Load example", "Change reference dataset", and a green "Run" button.

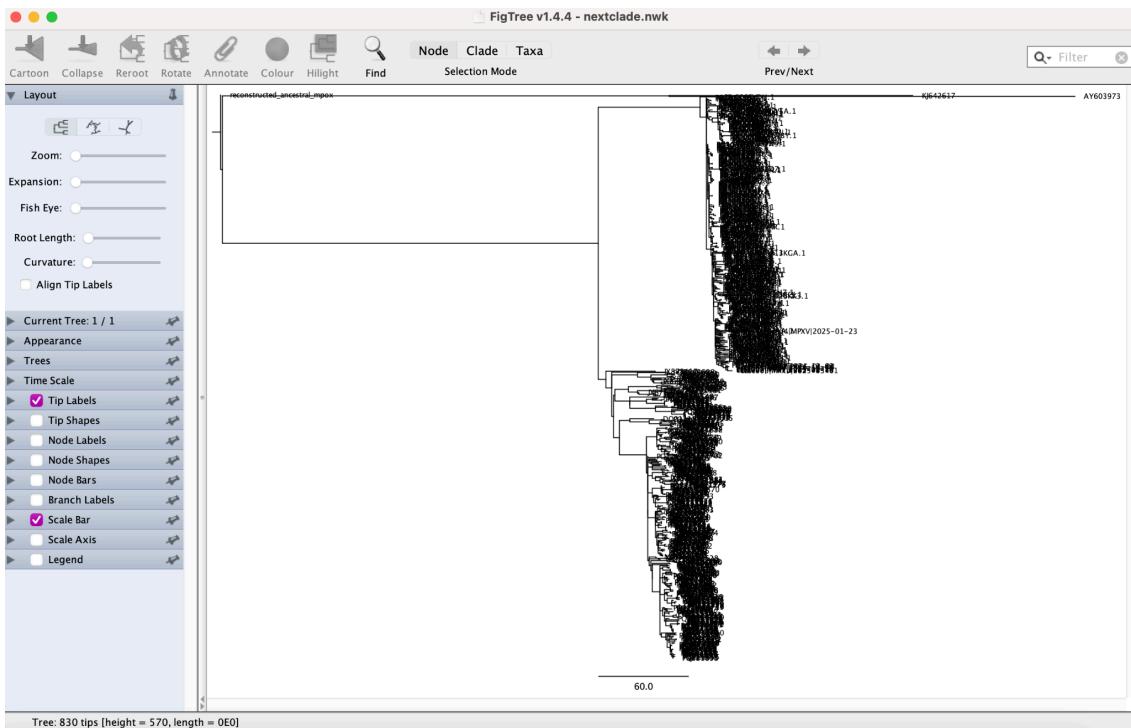
5. Return to the NextClade tab and navigate to the “Start” window by clicking the button in the top panel. Drag and drop the newly downloaded file with background sequences into the “Add more sequence data” box. You should see your additional file appear below in the “Sequence data you’ve added” panel.
6. Ensure the reference dataset is set to Mpxo virus (Clade I) and click “Run”.



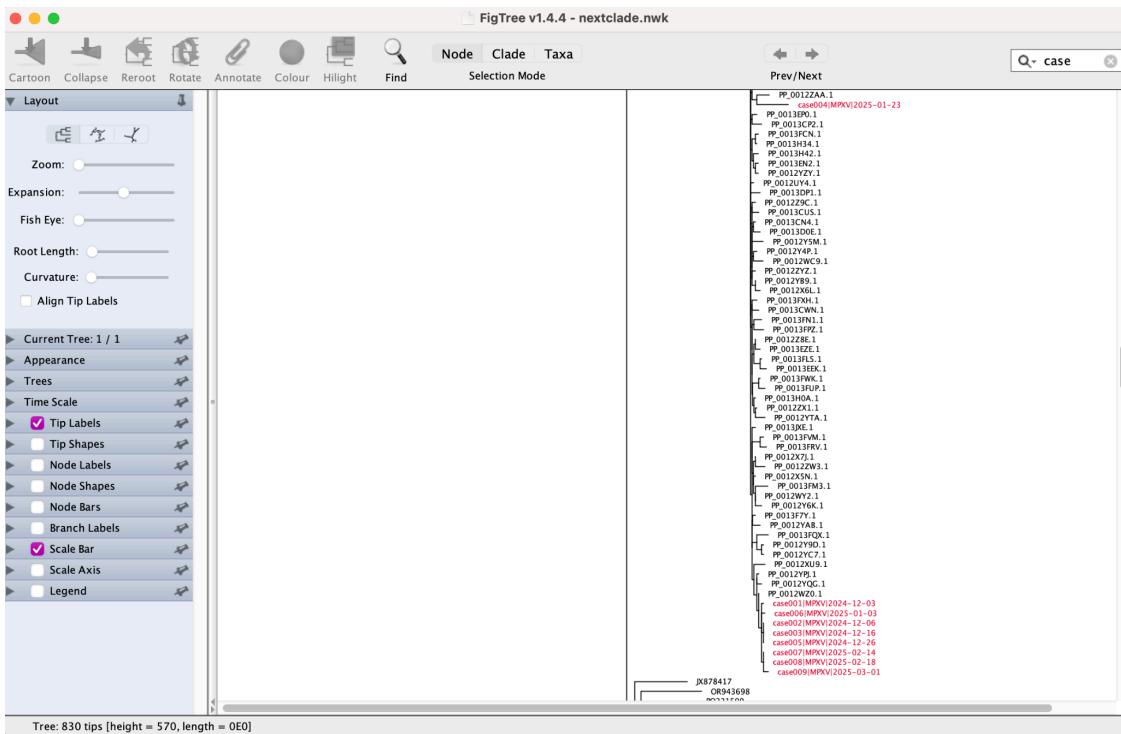
7. The “Results” window will show the case data alongside the background data. The analysis may take a little time as there are quite a few sequences to process. When the analysis has finished running, select “Tree” from the top panel to look at the tree that NextClade has produced.

8. Navigate to the “Export” window by clicking “Export” in the top panel. Scroll down until you see the `nextclade.nwk` download. This is the tree file.

Click the download button () to download the file to view in FigTree.



9. Open the newick (nwk) tree file in FigTree. Use the search bar in the top right to search for “case”. The sequences of interest will be highlighted. Select “Colour” to change the colour of the labels to make the sequences of interest clearer to see. You can expand the tree by sliding the “Expansion” dial in the left hand “Layout panel”.



10. Scroll to see your sequences of interest.

Question:

Do they cluster together?

What does this phylogeny tell us about the cases?

What do we not know?

Exercise: Repeat on newly generated data

Repeat these steps with the newly generated data to investigate the Clades and quality of the sequences you have generated during the workshop.

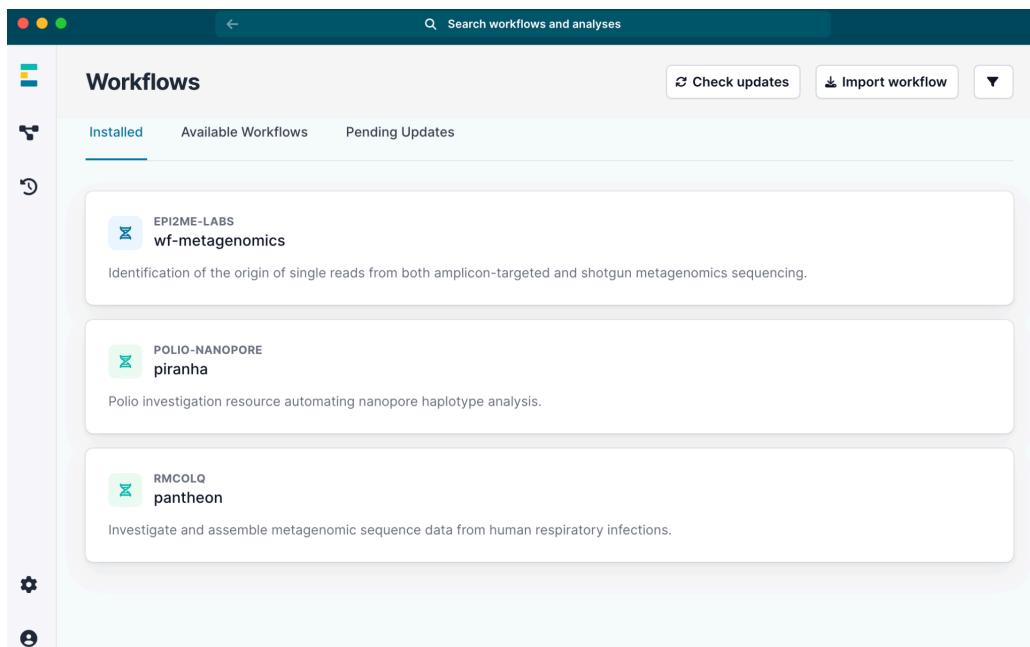
Introduction to squirrel

If the cases you're dealing with are part of known, sustained human outbreaks. It may not be necessary to take your analysis any further beyond what NextClade already provides. However, if it is unclear whether the data you're looking at is part of a known human outbreak, or if you are interested in APOBEC3 reconstruction for the data, squirrel can run this analysis.

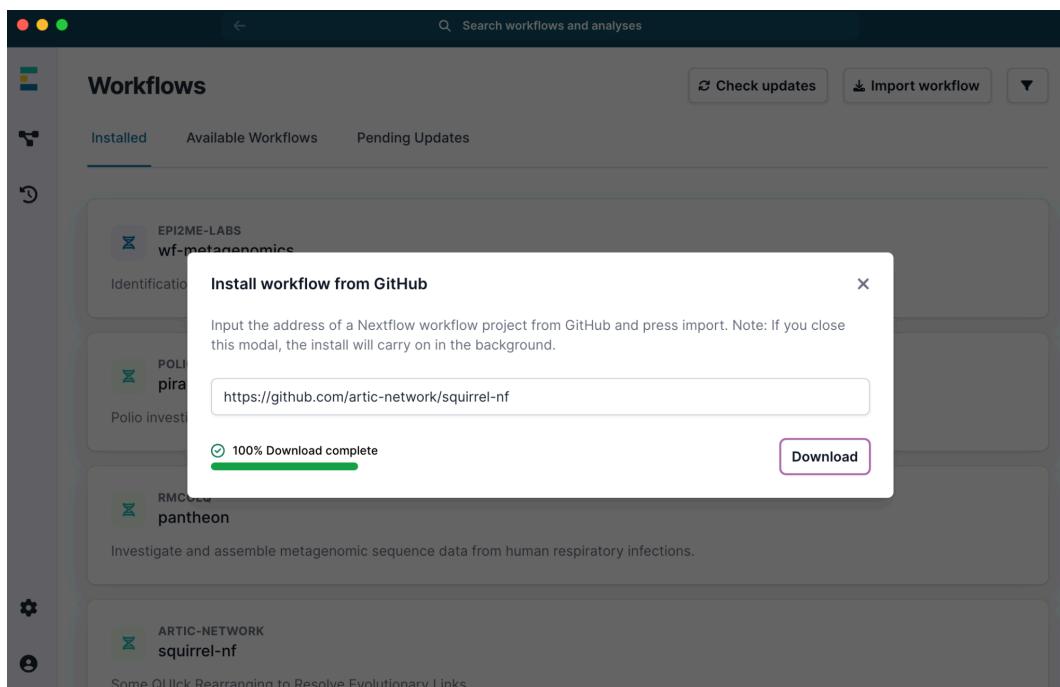
In [QC mode](#), squirrel can flag potential issues in the MPXV sequences that have been provided for alignment (e.g. SNPs near tracts of N, clusters of unique SNPs, reversions to reference alleles and convergent mutations) and outputs these in a mask file for investigation. We suggest you use this information to examine the alignment and pay close attention to the regions flagged. Squirrel can then accept this file with suggested masks and apply it to the sequences before doing phylogenetics.

Enrichment of APOBEC3-mutations in the MPXV population are a signature of sustained human-to-human transmission. Identifying APOBEC3-like mutations in MPXV genomes from samples in a new outbreak can be a piece of evidence to support sustained human transmission of mpox. Squirrel can run an [APOBEC3-reconstruction](#) and map these mutations onto the phylogeny.

Installing squirrel using EPI2ME



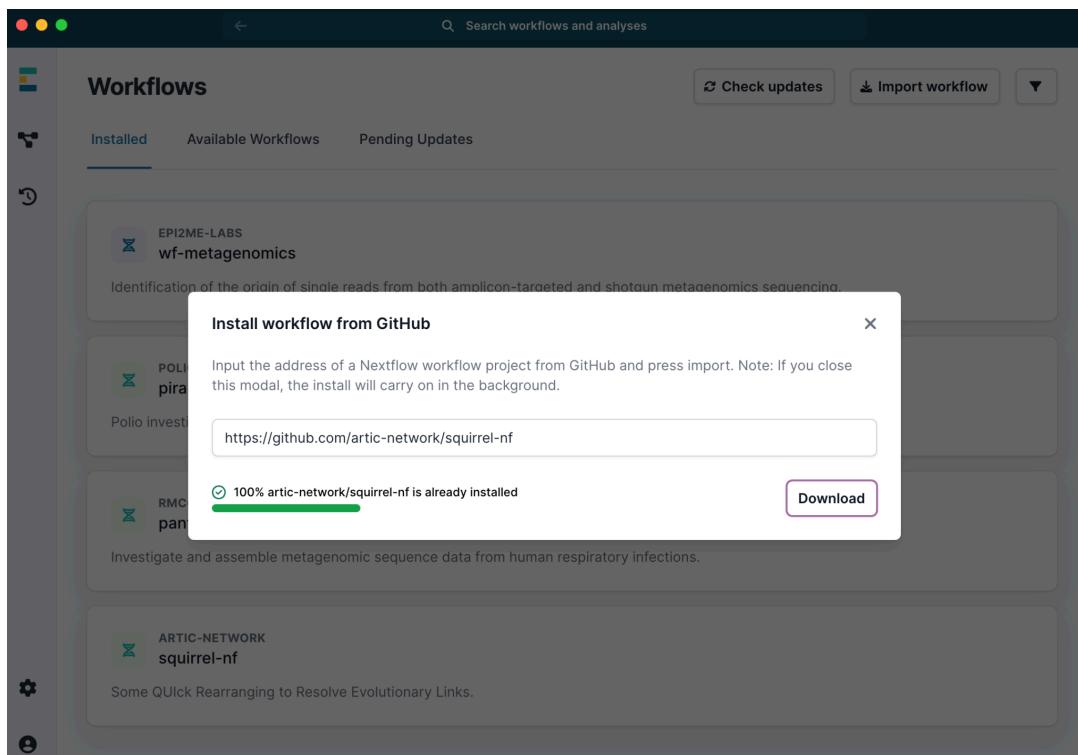
1. Open EPI2ME window and view workflows. If you haven't previously installed the squirrel workflow, it will not appear in the list of installed workflows. Click on the "Import workflow" button in the top right corner.



2. This will open a dialogue window as shown. Copy and paste the address to the squirrel workflow into the url box and click “Download”.

The address is:

<https://github.com/artic-network/squirrel-nf>



3. When the download has completed, click the top right X to close the dialogue box.

The screenshot shows the Nextflow Workflows interface. At the top, there's a search bar with the placeholder "Search workflows and analyses". Below the search bar, there are two buttons: "Check updates" and "Import workflow". A dropdown menu icon is also present. On the left side, there are several icons: a gear for settings, a magnifying glass for search, a circular arrow for refresh, and a downward arrow for more options. The main area is titled "Workflows" and contains four listed workflows:

- EPI2ME-LABS wf-metagenomics**
Identification of the origin of single reads from both amplicon-targeted and shotgun metagenomics sequencing.
- POLIO-NANOPORE piranha**
Polio investigation resource automating nanopore haplotype analysis.
- RMCOLQ pantheon**
Investigate and assemble metagenomic sequence data from human respiratory infections.
- ARTIC-NETWORK squirrel-nf**
Some QUILick Rearranging to Resolve Evolutionary Links.

4. You will return to the installed workflows page, where you can now see squirrel available as an installed workflow under the label `squirrel-nf. You are now ready to run squirrel.

Running squirrel for sequence QC

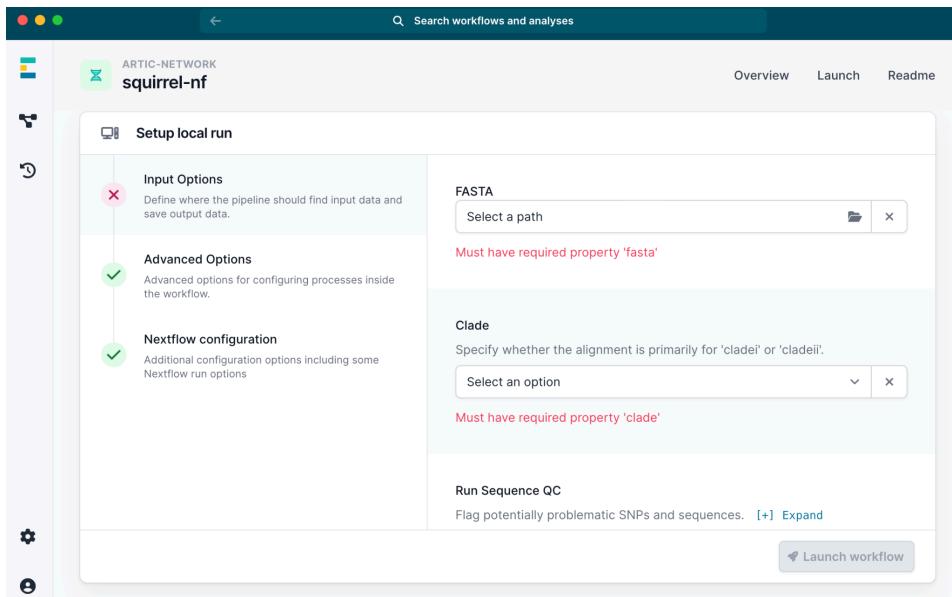
If squirrel is run in QC mode, it will flag sites that it believes may need to be masked from the alignment and produce a csv mask file summarising the sites. This file can then be provided to squirrel to redo the alignment with additional masking.

The screenshot shows the ARTIC-NETWORK interface with the 'squirrel-nf' workflow selected. The top navigation bar includes 'Overview', 'Launch', and 'Readme'. The workflow details section contains a description: 'Some QUICK Rearranging to Resolve Evolutionary Links.', a link to the GitHub repository (<https://github.com/aineniamh/squirrel/>), and the author's name, Aine O'Toole. It also indicates the workflow is 'Installed locally' and shows the current version as 1.0.12. Below this, there are four main actions: 'Run this workflow', 'Update workflow', 'Switch revision', and 'Delete workflow'. Each action has a corresponding icon and a brief description.

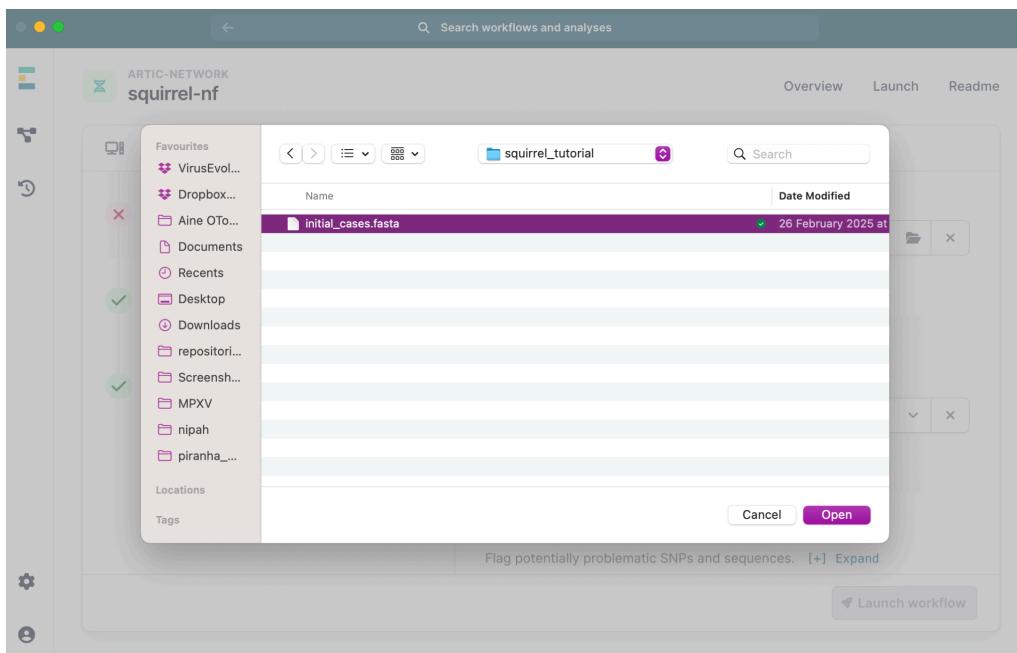
1. Click the squirrel workflow. You will be directed to this page, which has options to run, update and delete the workflow, as well as changing revision versions. We will run the workflow by clicking “Run this workflow”.

The screenshot shows the 'Launch' wizard for the squirrel-nf workflow. The first step, 'Select environment', is displayed. A single option, 'Run on your computer', is selected and highlighted with a blue border. At the bottom right of this step, there is a 'Continue' button.

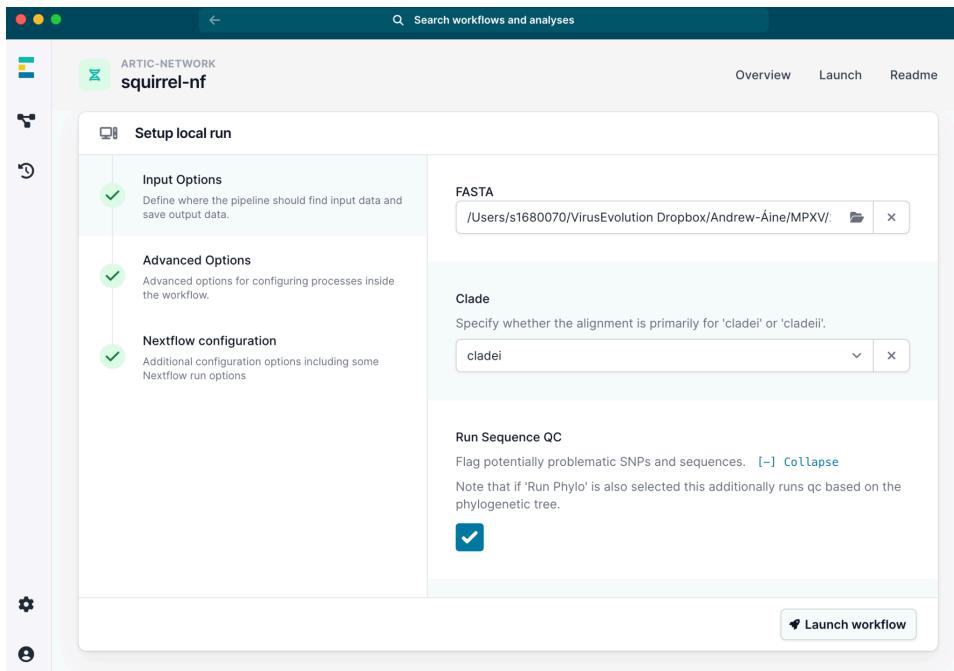
2. Select “Run on your computer” by clicking inside the box, and then click “Continue” on the bottom right corner.



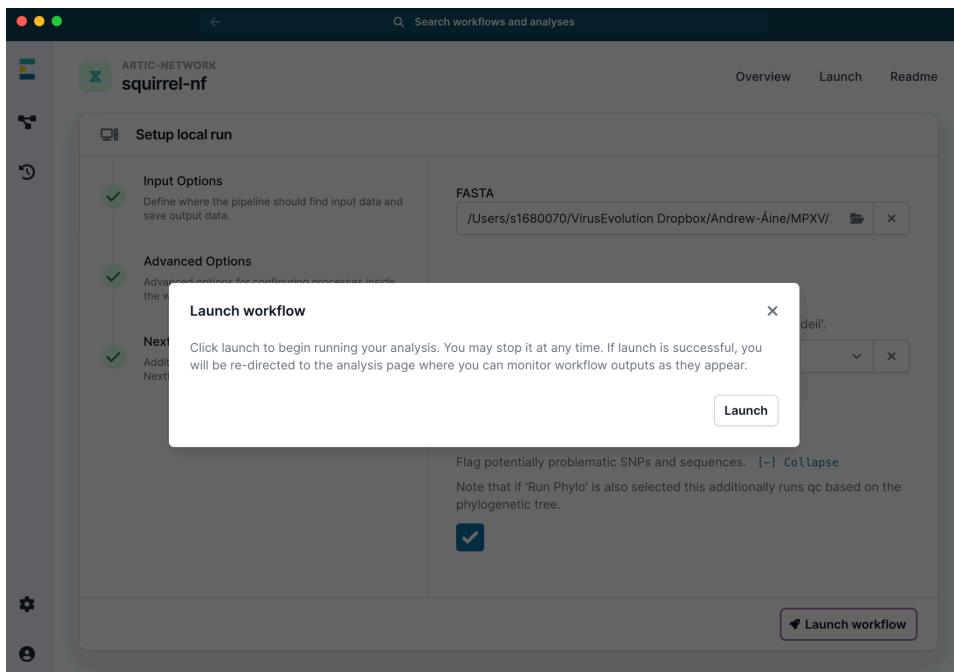
- This will bring up the setup screen as shown above. Along the left hand panel, there are “input options”, “advanced options” and “nextflow configuration”. The “advanced options” and “nextflow configuration” are already completed (green ticks) so we don’t need to change those settings. We do need to change some input options. This first pass we are going to run in `Sequence QC` mode. Select a FASTA file to input into squirrel by selecting the folder icon.



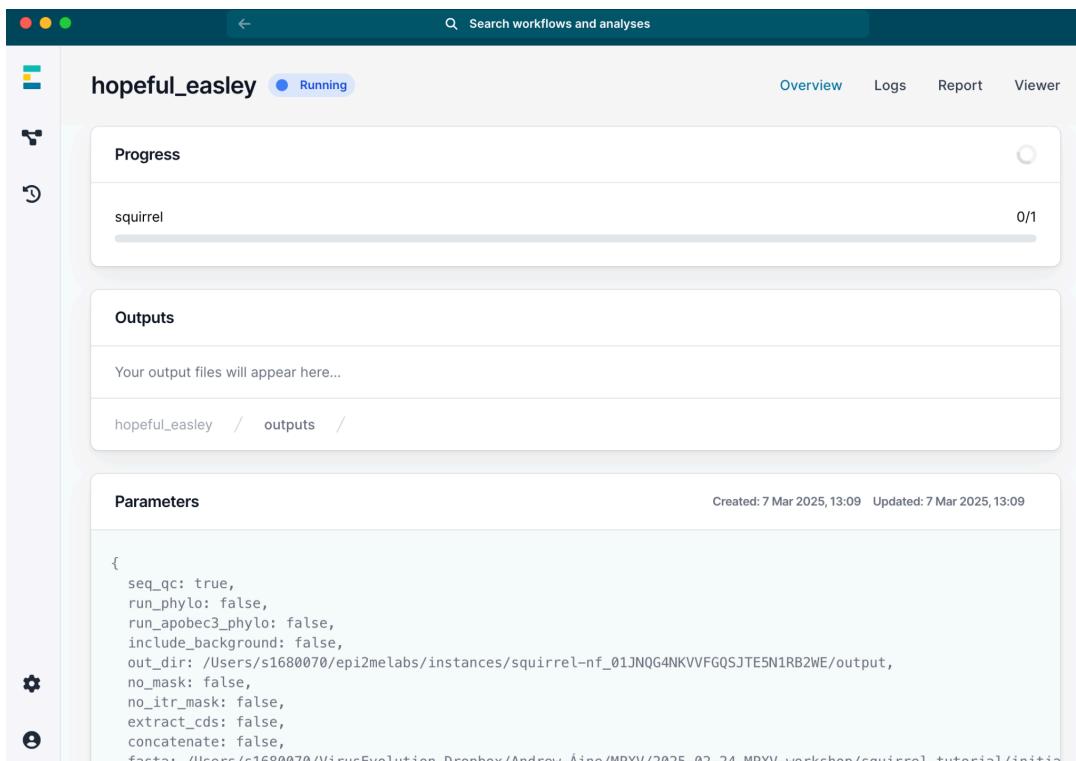
- Navigate in the file browser to the file you wish to run through squirrel. In this case we will run “initial_cases.fasta”, which has some simulated Clade Ib genome sequences.



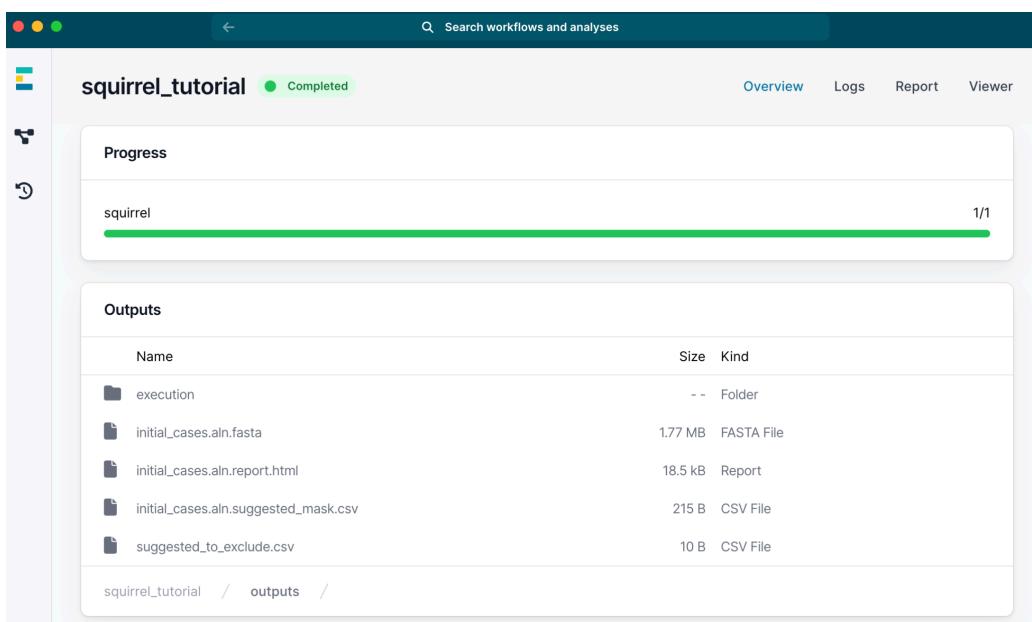
5. Select cladei for the clade and tick the box under “Run Sequence QC”, and then select “Launch workflow”.



6. A dialogue box will appear. To run the workflow, select “Launch”.



- EPI2ME will select a random name for the workflow run. In the above example it's "hopeful_easley". The name doesn't matter and only functions to create a unique run ID. The window shows the workflow is "Running" with the blue box.



- When the workflow has finished running, it will change to green and say "Completed" as shown above.

squirrel_tutorial • Completed

Overview Logs Report Viewer

initial_cases.aln.suggested_mask.csv 215 B CSV File
suggested_to_exclude.csv 10 B CSV File

squirrel_tutorial / outputs /

Parameters

Created: 7 Mar 2025, 13:09 Updated: 7 Mar 2025, 13:15

```
{
  seq_qc: true,
  run_phylo: false,
  run_apobec3_phylo: false,
  include_background: false,
  out_dir: "/Users/s1680070/epi2melabs/instances/squirrel-nf_01JNQ4NKVVFQ0SJTEN1RB2WE/output",
  no_mask: false,
  no_itr_mask: false,
  extract_cds: false,
  concatenate: false,
  fastaa: "/Users/s1680070/VirusEvolution Dropbox/Andrew-Aine/MPXV/2025-02-24_MPXV-workshop/squirrel_tutorial/initia
  clade: cladei
}
```

artic-network / squirrel-nf Version: 1.0.12 | 266ab8f

9. Scroll down to the “Parameters” section, which lists the workflow setup parameters. One of the pieces of information here describes the out_dir, which is the location that EPI2ME stores the output files from the workflow run. Navigate to this location in a file browser to find the output files from the squirrel analysis run. You can open the output folder by scrolling down and clicking “Open folder”.

To find the folder in your file browser manually follow the following instructions. There will be different ways to navigate to this location depending on your operating system.

For a windows machine, open up file explorer and on the left hand side panel you should be able to navigate to:

*This PC > Local Disk(C) > Users > *your_username* > epi2melabs > instances*

The most recent directory in this location will be the squirrel workflow that has just run. If you've previously run other workflows, such as artic_field_bioinformatics, they will also be shown here in this directory.

Name	Date Modified	Size	Kind
invoke.log	Today at 13:15	450 bytes	Log File
launch.json	Today at 13:09	1 KB	JSON
nextflow.log	Today at 13:15	10 KB	Log File
nextflow.stdout	Today at 13:09	209 bytes	Document
> output	Today at 13:15	--	Folder
params.json	Today at 13:09	405 bytes	JSON
progress.json	Today at 13:15	100 bytes	JSON

10. Within the most recent directory, you'll see a folder called "output", click into this to access the output of the EPI2ME squirrel run.

Understanding the squirrel QC output

Name	Size	Kind
> execution	--	Folder
initial_cases.aln.fasta	1.8 MB	TextEdit Document
initial_cases.aln.report.html	32 KB	HTML text
initial_cases.aln.suggested_mask.csv	215 bytes	comma-separated values
suggested_to_exclude.csv	10 bytes	comma-separated values

Contents of the squirrel QC output directory:

- initial_cases.aln.fasta
The file that ends in ` `.aln.fasta` is the aligned genome sequence data.
- initial_cases.aln.report.html
The html file is the file displayed within the EPI2ME report window when the run finishes
- initial_cases.aln.suggested_mask.csv
This suggested mask file is the file that we want out of our QC run. MPXV sequencing can be challenging, and the suggested mask file will flag any mutations in your genome that seem to be problematic/ a result of assembly or sequencing error.
- suggested_to_exclude.csv
This file contains a list of any sequences in the provided file that are very incomplete or problematic, so are suggested to be excluded from the alignment for phylogenetics. Whether to exclude the sequences is left to the user's discretion.

Double click the suggested mask file to open it:

Name,Minimum,Maximum,Length,present_in,note
65224,65224,65224,1,case009 MPXV 2025-03-01,N_adjacent
102759,102759,102759,1,case004 MPXV 2025-01-23,N_adjacent
105223,105223,105223,1,case001 MPXV 2024-12-03,N_adjacent

The file describes 3 sites with changes that squirrel believes are due to error because they occur immediately next to an N base (or block of Ns), which suggests they occur in an area of low read coverage.

When you next run squirrel, you can supply this file and the mutations that are likely due to error can be masked out.

Running squirrel using EPI2ME for APOBEC3 reconstruction

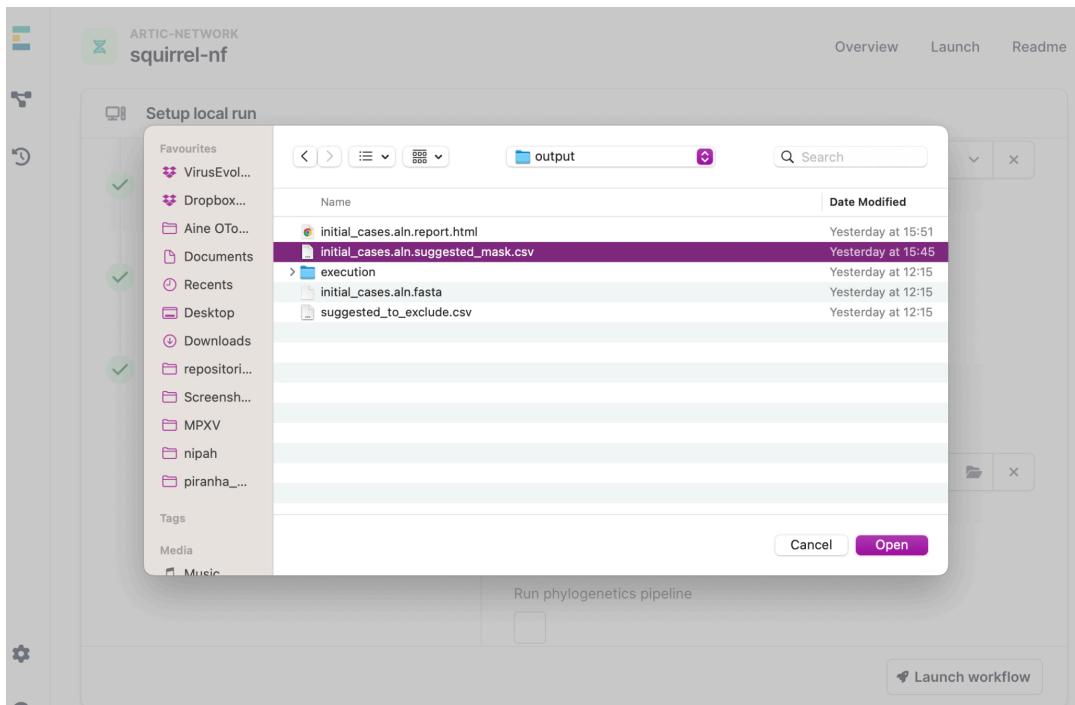
We will now run APOBEC3 reconstruction and phylogenetics through squirrel, using the suggested_mask file produced in the QC analysis.

The screenshot shows the squirrel-nf workflow page. At the top, there's a header with the project name "ARTIC-NETWORK squirrel-nf" and links for "Overview", "Launch", and "Readme". Below the header, there's a brief description: "Some QUick Rearranging to Resolve Evolutionary Links.", a link to the GitHub repository (<https://github.com/aineniamh/squirrel/>), and the author's name, Aine O'Toole. It also indicates the workflow is "Installed locally" and shows the "Current version: 1.0.12". On the left, there's a sidebar with icons for "Run this workflow", "Update workflow", "Switch revision", and "Delete workflow". The main content area lists these four options with their descriptions and a right-pointing arrow for each.

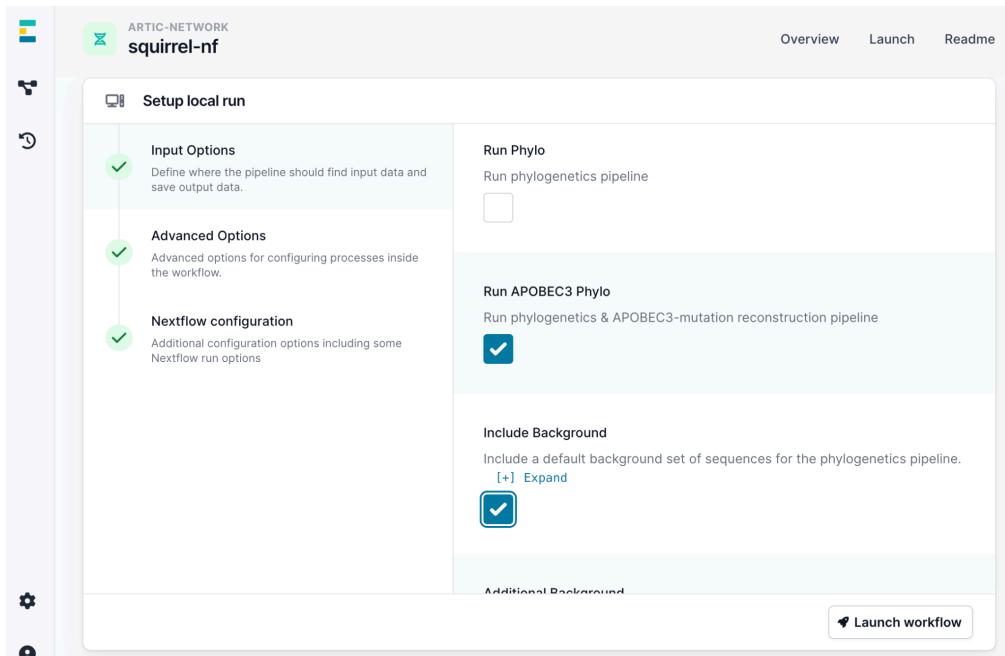
1. Navigate to the squirrel-nf workflow page by clicking the workflow panel button on the left hand side (), and then selecting squirrel-nf.

The screenshot shows the "Setup local run" configuration screen. It has a sidebar with icons for "Run this workflow", "Update workflow", "Switch revision", and "Delete workflow". The main area is titled "Setup local run" and contains three sections: "Input Options" (selected, with a note about defining input and output paths), "Advanced Options" (selected, with a note about configuring processes), and "Nextflow configuration" (selected, with a note about additional Nextflow run options). To the right, there are fields for "FASTA" (set to "/Users/s1680070/VirusEvolution Dropbox/Andrew-Áine/MPXV/"), "Clade" (set to "cladeib" in a dropdown menu), and "Run Sequence QC" (with a note about flagging problematic SNPs and sequences, and a checkbox that is unchecked). At the bottom, there's a "Launch workflow" button.

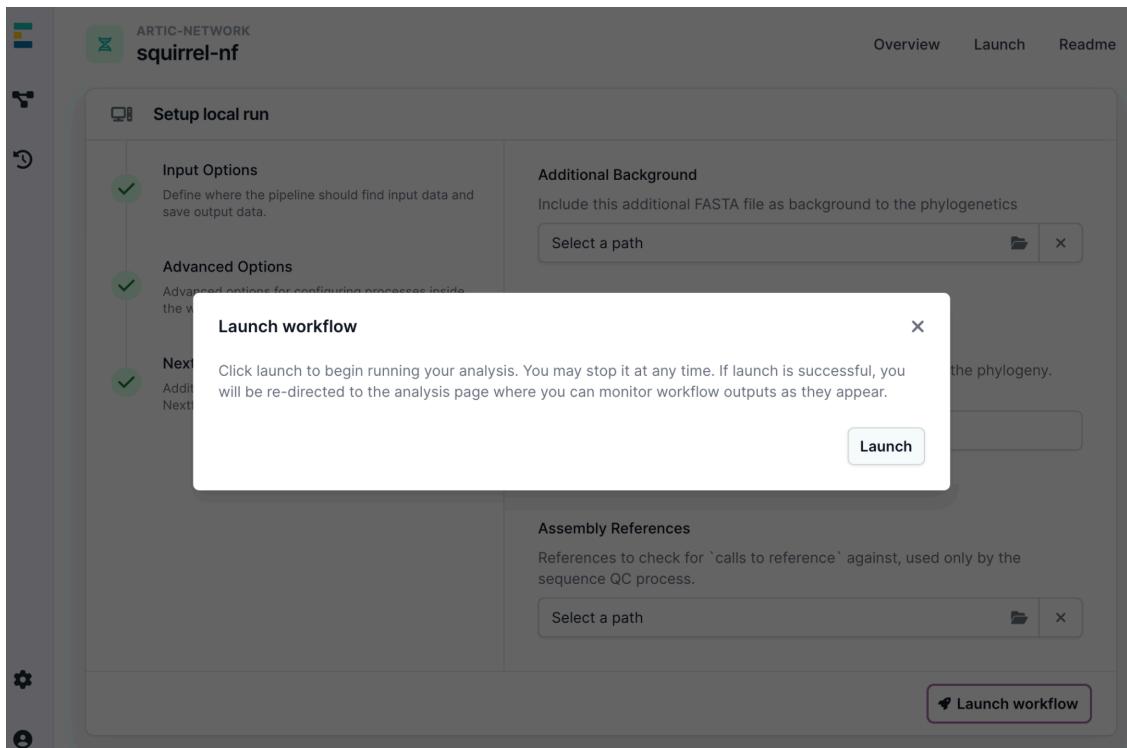
2. Run this workflow and click through as before until you get to the setup screen above. Select the initial_cases.fasta file as before, and you can select cladeib from the Clade dropdown menu. We do not need to run sequence QC this time, so you don't need to tick that box.



3. Scroll down to the “Additional Mask” file panel and navigate to the epi2melabs workflow output as you did previously. Select the suggested_mask.csv file.



4. Continue to scroll down the page and select “Run APOBEC3 Phylo” and “Include Background”. This will run the full phylogenetic analysis with APOBEC3 reconstruction and include some minimal background Clade Ib sequences in the analysis.



5. Click “Launch workflow” and then “Launch” when the dialogue box appears to run the workflow. You can rename the workflow as before by scrolling down and selecting “Rename analysis”.

squirrel_apobec3 ● Completed

Report files initial_cases.aln.report

squirrel | Some QQuick Reconstruction to Resolve Evolutionary Links

squirrel report 2025-03-08

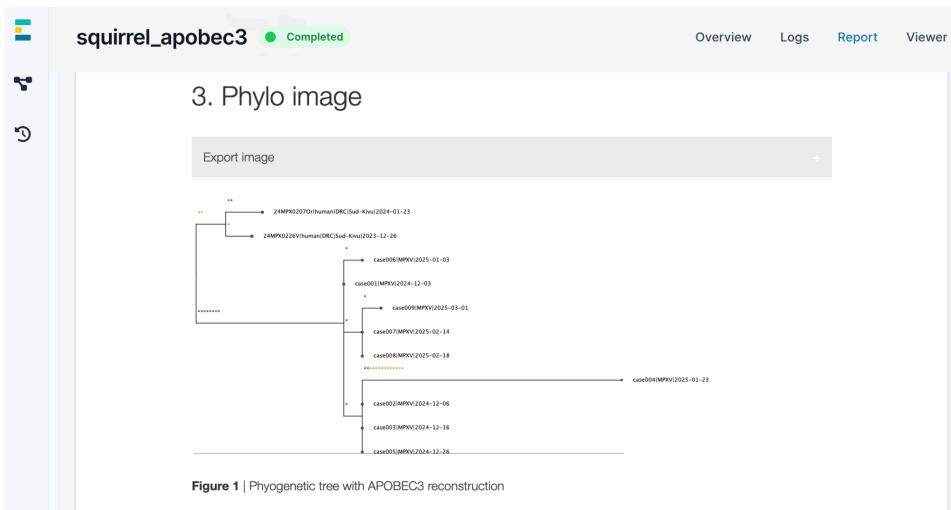
Best-practice phylogenetics for MPXV analysis

1. Alignment
Output alignment written to [initial_cases.aln.fasta](#)

2. Tree file
Output tree file written to [initial_cases.aln.tree](#)

3. Phylo image
[Export image](#)

6. When the workflow has finished running (the status says “Completed” in green at the top), you can navigate to the Report viewer, by clicking “Report” in the top right.



7. Scroll down to view the image of the phylogenetic tree with APOBEC3 reconstruction. Mutations are displayed along branches as dots, with red dots representing APOBEC3-type mutations that are characteristic of human-to-human transmission and yellow dots representing non-APOBEC3 type mutations. Locate the samples in the phylogenetic tree and identify the background sequences that have been included too.

Question:

What sort of SNPs are seen across the branches of the tree?

What can you say about case004?

Are all the sequences in the initial_cases.fasta file linked?

What might provide more information about whether the cases are related?

8. Return to the Overview window and scroll down until you see the option “Open folder”. Open the folder to explore the output files.

Name	Kind
> execution	Folder
initial_cases.aln.fasta	TextEdit Document
initial_cases.aln.report.html	HTML text
initial_cases.aln.tree	TextEdit Document
initial_cases.aln.tree.amino_acid.reconstruction.csv	comma-separated values
initial_cases.aln.tree.branch_snps.reconstruction.csv	comma-separated values
initial_cases.aln.tree.png	PNG image
initial_cases.aln.tree.state	Document
initial_cases.aln.tree.state_differences.csv	comma-separated values
initial_cases.aln.tree.svg	SVG Image

9. The output files from the squirrel and APOBEC3-reconstruction analysis are shown above.

- `initial_cases.aln.fasta`

The alignment file, with alignment scaffolded against a clade-specific reference. By default one of the ITR regions and a curated set of problematic regions is masked as Ns.

- `initial_cases.aln.tree`

The output maximum likelihood tree file from IQTREE2 with Node labels that correspond to the reconstruction Node labels. This tree can be viewed in various tree viewers, for example FigTree.

- `initial_cases.aln.tree.state` and `initial_cases.aln.tree.state_differences.csv`

The output ancestral state reconstruction file from IQTREE2 and the compiled list of unambiguously variable sites from squirrel.

- `initial_cases.aln.tree.branch_snps.reconstruction.csv`

A report of individual site changes mapped to specific branches and their dinucleotide context.

- `initial_cases.aln.tree.amino_acid.reconstruction.csv`

A report of each mutation that occurs across the phylogeny, their location, dinucleotide context, APOBEC3 status, which gene they're present in, codon position, amino acid change and a prediction of how extreme that amino acid change is with Grantham score.

- `initial_cases.aln.tree.png` and `initial_cases.aln.tree.svg`

Visualisation of reconstructed phylogeny showing whether mutations are consistent with APOBEC3 editing or not.

- `initial_cases.aln.report.html`. Summary report of analysis run.

Exercise:

Give squirrel a try with your newly generated MPXV genome sequence data. Squirrel analysis is clade specific, so we have provided you with a clade I file and clade II file.

References

Conticello, S.G. 2008. <https://doi.org/10.1186/gb-2008-9-6-229>

Drummond et al 2003

<https://www.sciencedirect.com/science/article/abs/pii/S0169534703002167>

O'Toole et al 2023 <https://www.science.org/doi/10.1126/science.adg8116>

Suspène et al 2023 <https://pubmed.ncbi.nlm.nih.gov/37224627/>

Wawina-Bokalanga et al 2025

<https://www.medrxiv.org/content/10.1101/2024.11.15.24317404v1>

Useful links

<https://clades.nextstrain.org/>

<https://github.com/aineniamh/squirrel>

<https://nanoporetech.com/products/analyse/epi2me>

<https://artic.network/how-to-read-a-tree.html>

<https://artic.network/mpxv/mpxv-phylogenetics-epi2me-sop.html>