

MAY 01, 2023

SARS-CoV-2 consensus genome reconstruction, quality control, and lineage analysis

Benjamin
Schwessinger¹

¹Australian National University



Benjamin Schwessinger
Australian National University

OPEN ACCESS

Protocol Citation: Benjamin Schwessinger 2023. SARS-CoV-2 consensus genome reconstruction, quality control, and lineage analysis. **protocols.io** <https://protocols.io/view/sars-cov-2-consensus-genome-reconstruction-quality-ctipwkdn>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

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

Created: Apr 30, 2023

Last Modified: May 01, 2023

PROTOCOL integer ID:
81199

ABSTRACT

This protocols is part of the ANU Biosecurity mini-research project #2 "An SARS-CoV2 incursion scenario: Genomics, phylogenetics, and incursions." This mini-research project is modeled on the yearly Quality Assurance Program of The Royal College of Pathologists of Australia (RCPAQAP), we take part in together with ACT Pathology.

This research project is split into two major parts, identical to how the official RCPAQAP is run every year.

Part #1 is focusing on the 'wet-lab' by sequencing SARS-COV2 from real world RNA samples provided by ACT Pathology especially for our ANU biosecurity course (Thank YOU!). Here you will amplify and sequence five (5) RNA samples per research group. You will assess the SARS-COV2 genome sequences for their lineage assignments using online programs, put sequences into a global context, estimate the collection date based on genetic information, and describe mutations in the spike protein.

Part #2 is focusing on the 'dry-lab' by investigating a hypothetical incursion scenario in the so-called city Fantastica. You will combine genomic surveillance of SARS-COV2 with case interview data to trace the spread into of SARS-COV2 in the community and into high risk settings. We will provide you with real publicly available SARS-COV2 genome and fantasized case interviews. You will put these two together to trace the spread and suggest potential improvements in containment strategies with a focus on high risk settings.

This protocol describes the analysis component of Part #1. The metrics you are suppose to report for each of your samples are mostly borrowed from the official SARS-CoV-2 QAP. Don't worry if not all of these mean something to you at the moment as we will explain them again during the prac. In case all/most of your

samples have < 50% genome coverage please also include the analysis of MakeUp for points 1 to 7 and TimeMakeUp for point 8. You can access the MakeUp data [here](#) (ANU only). Make sure to read the README file so you understand what each item relates to.

The metrics you have to report for each of our samples (or MakeUps) include the following.

1. Consensus genome coverage.
2. Average read depth; You might want to include detailed read depth plots here as well.
3. Pangolin Lineage.
4. NextClade Lineage.
5. Base pair differences relative to the original SARS-CoV-2 genome.
6. Amino acid replacements and deletions in the S (spike) protein sequence.
7. Evaluation if your and/or the MakeUp samples would make the QC cut-off 90% genome coverage and other metrics that you deem important for QC. Would you 'flag' any of your samples as standing out e.g. being negative control?
8. Approximate sampling date of your sequences for month and year.

You must report the versions of all tools used in your report and the day the analysis was performed. This is extremely important for reporting as lineage naming and such change VERY frequently during the pandemic and any outbreak.

This protocol is applicable for week 9.

The following links might be useful for your report:

The original publication that describes the sequencing protocol is here:

<https://academic.oup.com/biomethods/article/5/1/bpaa014/5873518?login=true>

Original sources describing the consensus reconstruction from raw reads are here:

<https://artic.network/ncov-2019>, <https://artic.network/ncov-2019/ncov2019-bioinformatics-sop.html>, <https://labs.epi2me.io/>,
<https://www.nature.com/articles/s41467-020-20075-6>

Other websites and resources used in the protocol: <https://igv.org/>,

<https://clades.nextstrain.org/>, <https://pangolin.cog-uk.io/>,

<https://genome.ucsc.edu/cgi-bin/hgPhyloPlace>

GUIDELINES

You must have read, understood, and follow the health and safety instructions provided in the "Overview Mini-Research Project #2 BIOL3106/6106" provided on Wattle (ANU learning portal).

You must have signed and returned one copy of the "Student Safety Declaration Form For Practical Class Work" before starting any laboratory work.

You must have read and understood the Hazard Sheets (Risk assessment) of all chemicals listed below in the "Safety Warnings" section. These Hazard Sheets are provided on Wattle as part of the "Overview Mini-Research Project #2 BIOL3106/6106" document.

SAFETY WARNINGS



This protocol does not require any hazardous substances or infectious agents. However, maintain a proper posture while working on your computer.

BEFORE START INSTRUCTIONS

You must study the protocol carefully before you start. If anything is unclear post questions directly here on protocols.io.

Section I: Consensus genome reconstruction using Epi2Me ...

- 1 This section aims to reconstruct each sample's consensus genome based on the publication here <https://academic.oup.com/biomet/article/5/1/bpaa014/5873518?login=true> and an adaptation of the ARTIC Sars-CoV-2 analytics workflow <https://artic.network/ncov-2019>

The principle steps are:

1. Read filtering on quality and length.
2. Aligning reads to the reference.
3. Normalizing to 200x per base read depth on average.
4. Trimming of primer sequences.
5. Variant calling for both pools independently.
6. Combining variant calls of both pools.
7. Generation of sample's consensus genome reconstruction.
8. Generation of QC statistics like genome coverage, average read coverage etc.
9. Additional analysis like lineage calling, variant quantification, variant impact, etc.

We will run all of this within Epi2Me till step 8. The input for the Epi2Me analysis will be the "fastq_pass" folder of the MinKNOW basecalling output. This folder contains one "barcodeXX"

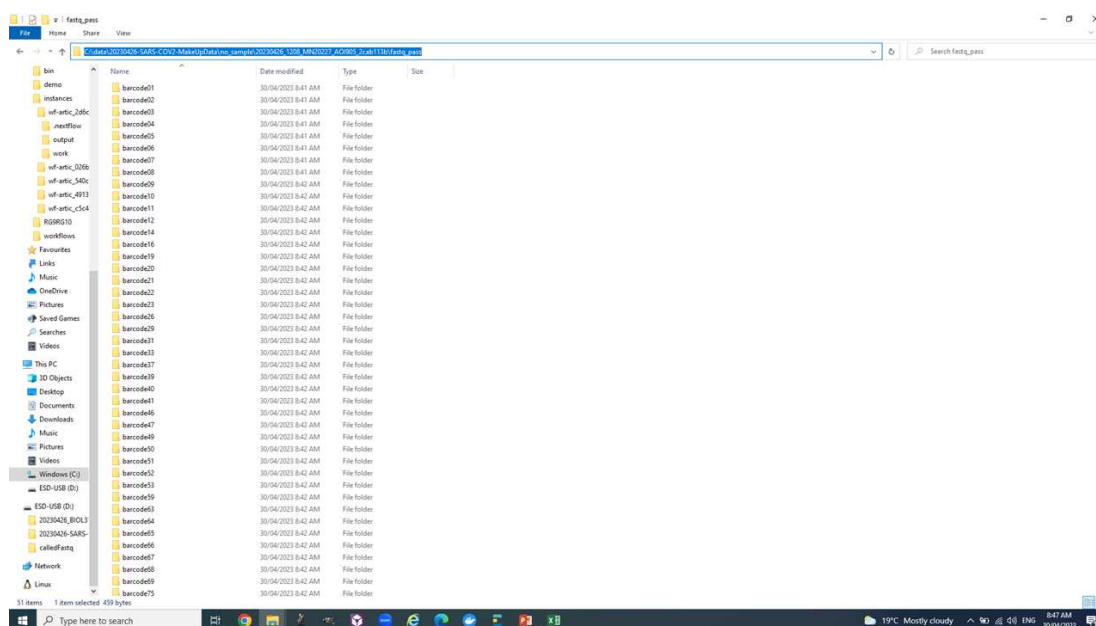
folder for each barcode identified by the basecalling software.

Overall, the Epi2Me "wf-artic" workflow takes two inputs.

1. The basecalled fastq output of MinKNOW (or equivalent e.g. guppy).
2. A sample sheet in csv (comma separated values) format with the columns "barcode", "alias", "type".

You might also want to execute this section on the MakeUp dataset if your data was of poor quality.

- 2 Your basecalled and demultiplexed, by barcode, fastq data should be in the 'fastq_pass' folder be in a similar to this C:\data\20230426-SARS-COV2-
MakeUpData\no_sample\20230426_1208_MN20227_AOI905_2cab113b\fastq_pass.

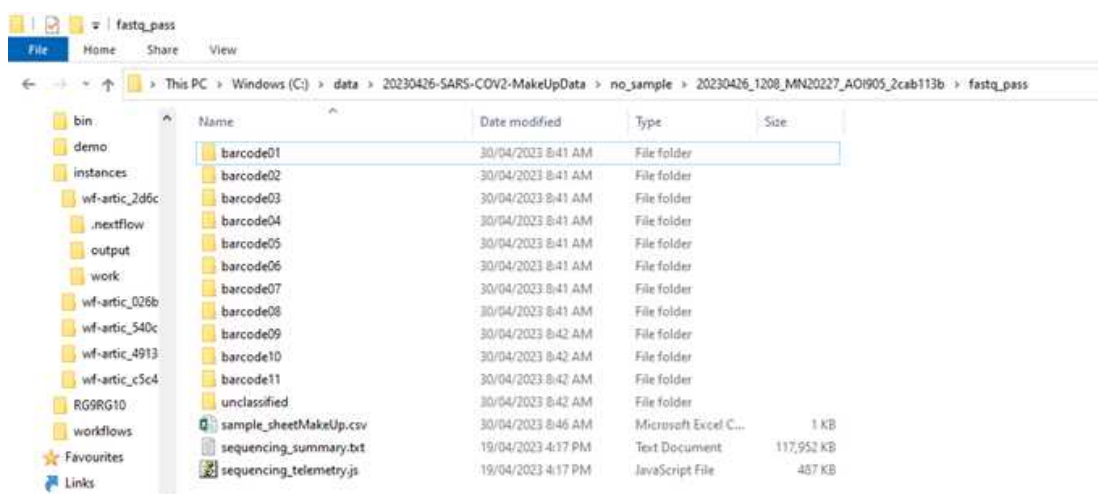


fastq_pass folder with all the demultiplexed barcodeXX folders containing fastq files for each barcode

- 3 We need to update the sample sheet to reflect your specific run and samples. This sheet should be located in the same folder your 'fastq_pass' barcode folders. It should look similar to the one below. (Done for you.)

	A	B	C	D	E
1	barcode	alias	type		
2	barcode01	MakeUp_sample_1	test_sample		
3	barcode02	MakeUp_sample_2	test_sample		
4	barcode03	MakeUp_sample_3	test_sample		
5	barcode04	MakeUp_sample_4	test_sample		
6	barcode05	MakeUp_sample_5	test_sample		
7	barcode06	MakeUp_sample_6	test_sample		
8	barcode07	MakeUp_sample_7	test_sample		
9	barcode08	MakeUp_sample_8	test_sample		
10	barcode09	MakeUp_sample_9	test_sample		
11	barcode10	MakeUp_sample_10	test_sample		
12	barcode11	MakeUp_sample_NegControl	test_sample		
13					

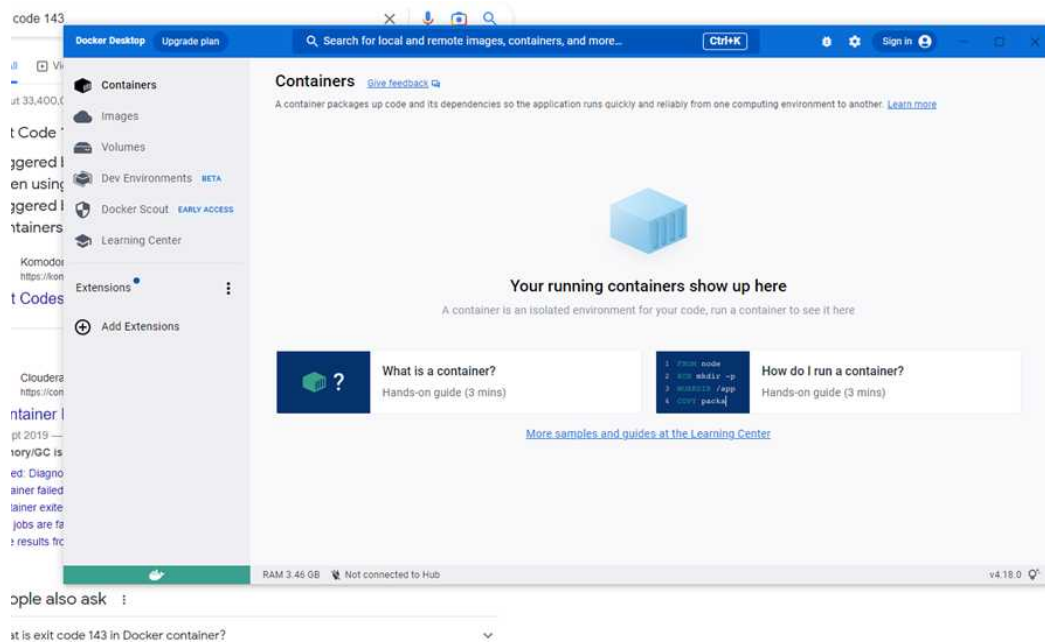
- 4 Before we start with the workflow we need to make a backup (e.g. copy of the 'fastq_pass' folder on the desktop for now) and then delete all the barcode folders we have not specified in the sample sheet. We also need to delete all folders of barcodes we have used but have not received any data.



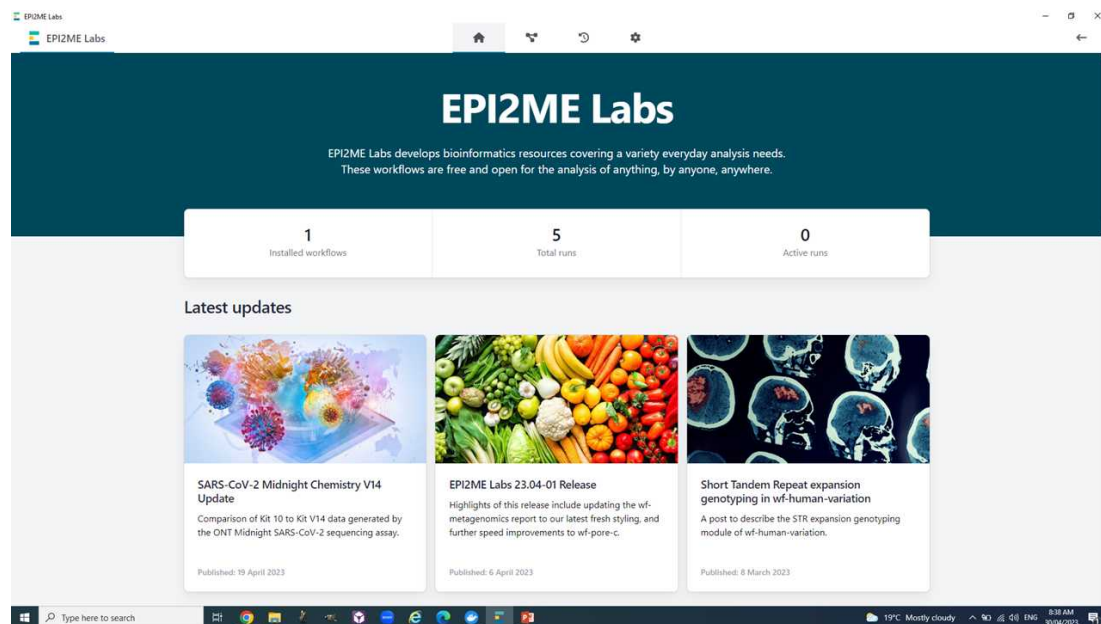
Fastq_pass folder after clean-up and with sample sheet.

- 5 Now we are ready to set-up and run the Epi2Me workflow for artic analysis.

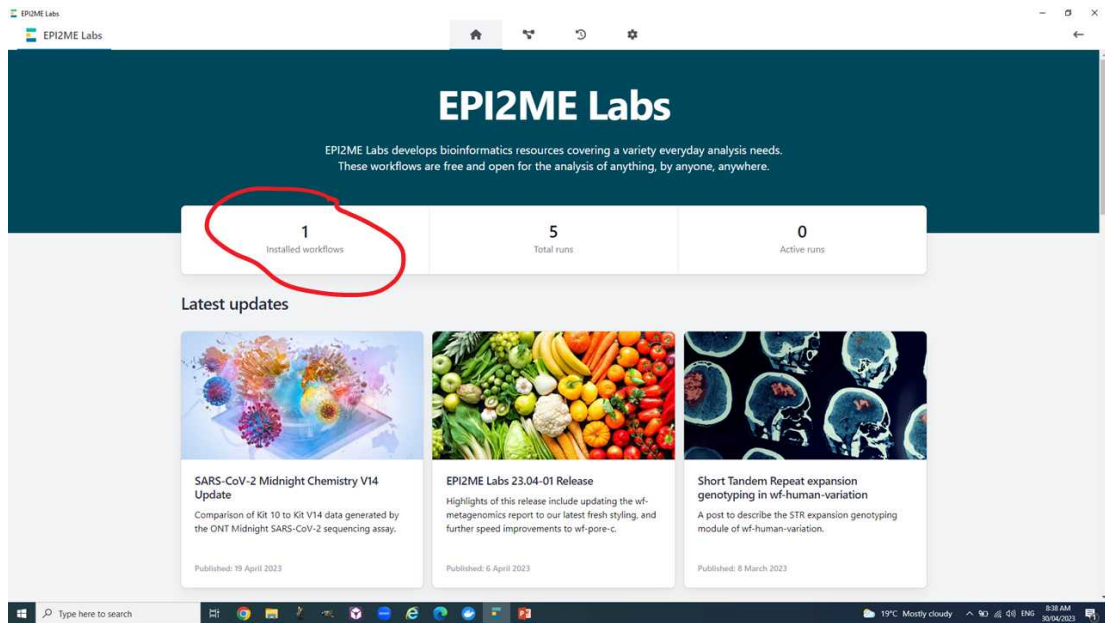
6 Start the "Docker App" and make sure it is running.



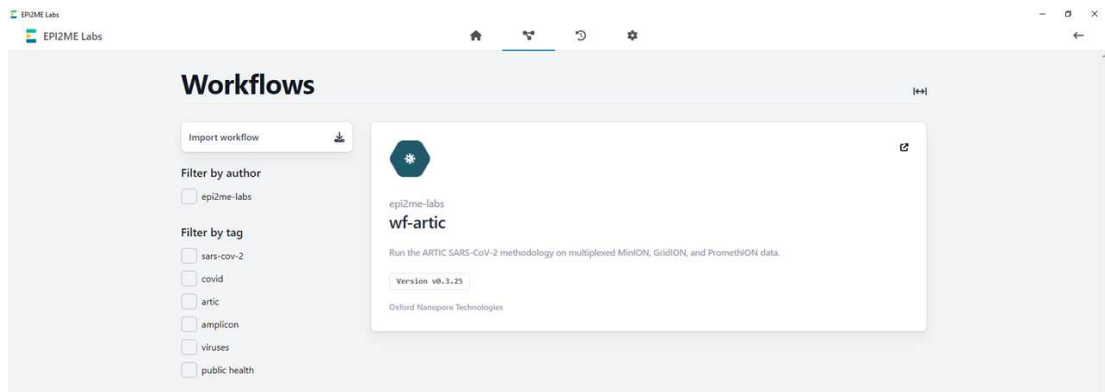
7 Start EPI2Me.



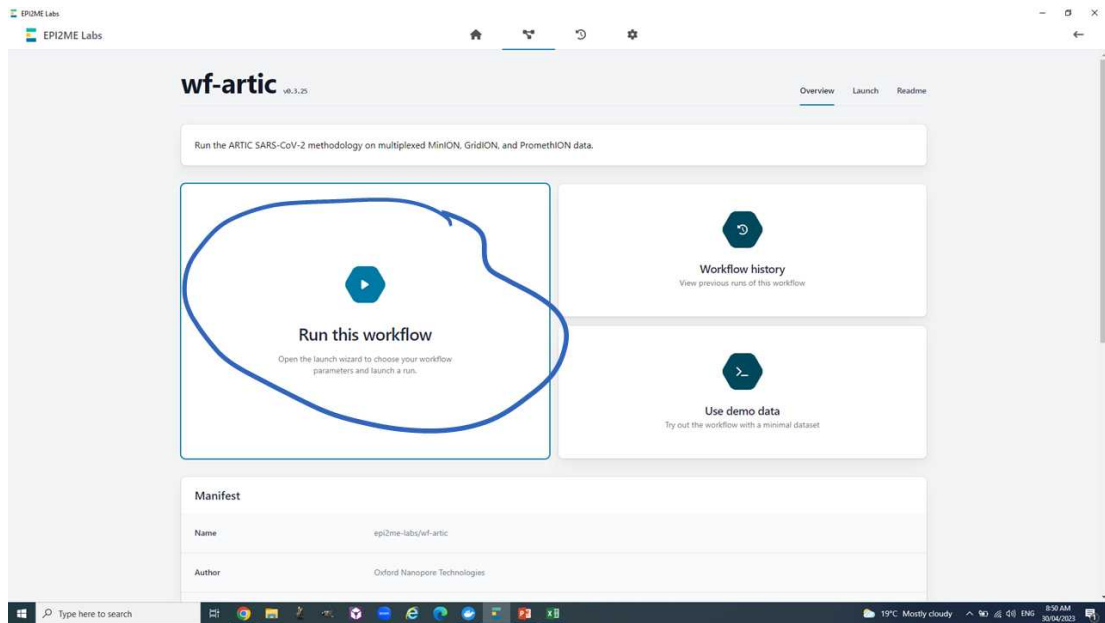
8 Click the "Installed workflows" bottom.



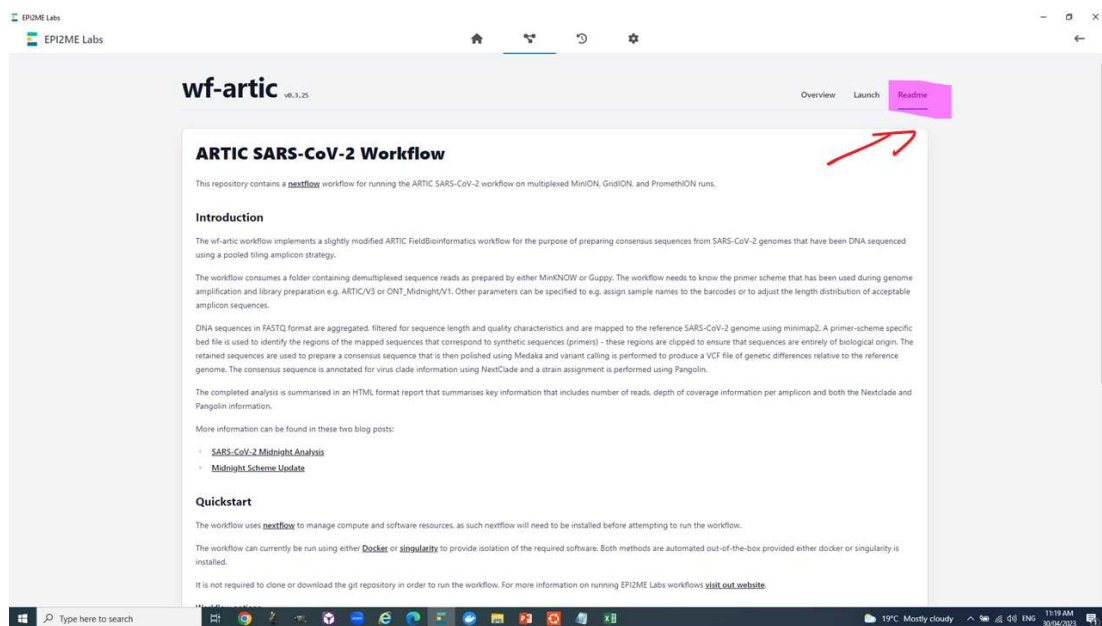
9 This should show "wf-artic" to be available.



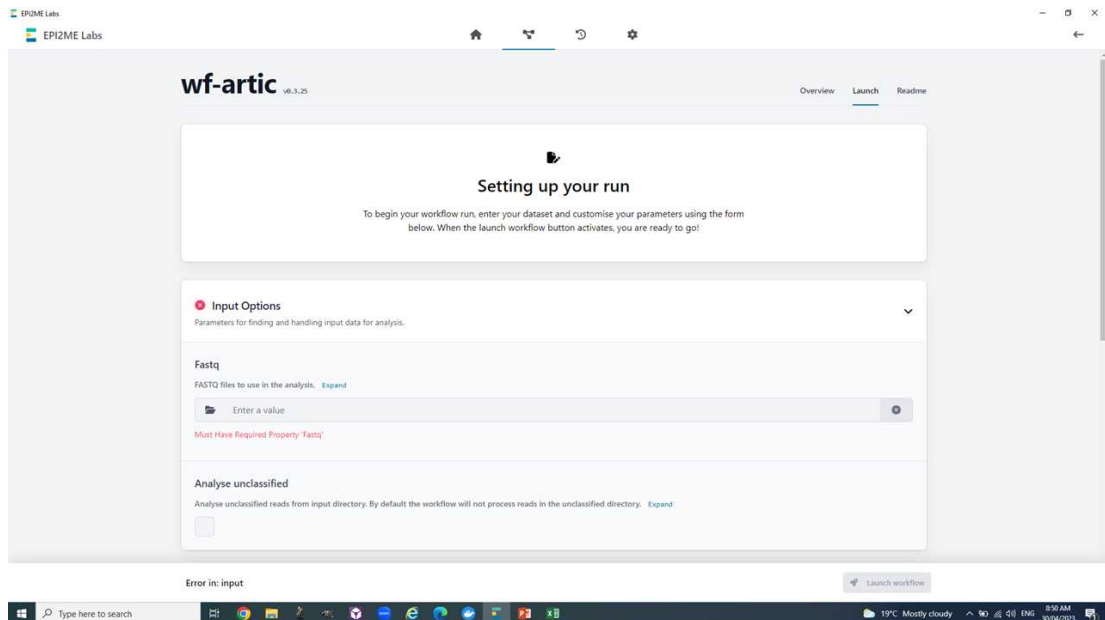
10 Now select the "wf-artic".



- 10.1 Site note. Also consider reading the "README" section to get more background for your report.

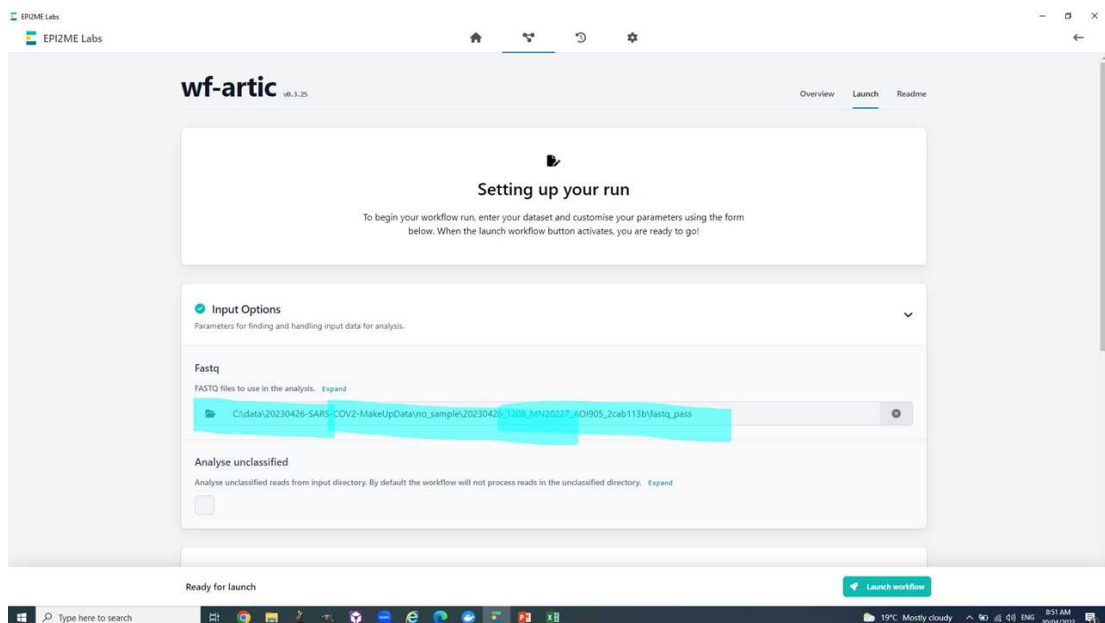


- 11 Now we are ready to set up our "wf-artic" specific to our samples and wet-lab protocols.



12 Tell the program where to find your 'fastq_pass' folder.

Similar to "C:\data\20230426-SARS-COV2-MakeUpData\no_sample\20230426_1208_MN20227_AOI905_2cab113b\fastq_pass".



13 Now set up all the other fields.

EPI2ME Labs

Primer Scheme Selection

Reference genome related files and options required for the workflow.

Scheme name

Primer scheme name. [Expand](#)

SARS-CoV-2

Scheme version

Primer scheme version. [Expand](#)

Midnight-ONT/V3

Custom scheme

Path to a custom scheme. [Expand](#)

Enter a value

Sample Options

Parameters that relate to samples such as sample sheets and sample names.

Reporting Options

Options controlling report outputs.

Ready for launch

Launch workflow

Type here to search

19°C Mostly cloudy 8:51 AM 30/04/2023

EPI2ME Labs

Path to a custom scheme. [Expand](#)

Enter a value

Sample Options

Parameters that relate to samples such as sample sheets and sample names.

Sample sheet

A CSV file used to map barcodes to sample aliases. The sample sheet can be provided when the input data is a directory containing sub-directories with FASTQ files. [Expand](#)

C:\data\20230426-SARS-COV2-MakeUpData\sample_20230426_1208_MN20227_AQ1905_2cab11701\fastq_pass/sample_sheetMakeUp.csv

Sample

A single sample name for non-multiplexed data. Permissible if passing a single .fastq(.gz) file or directory of .fastq(.gz) files.

Enter a value

Reporting Options

Options controlling report outputs.

Advanced Options

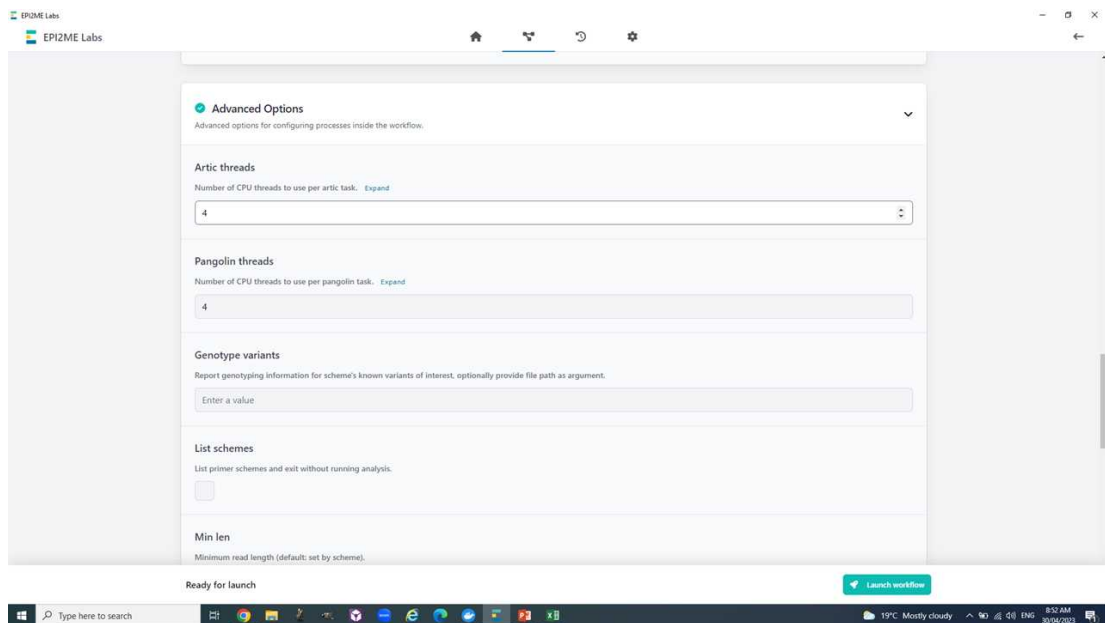
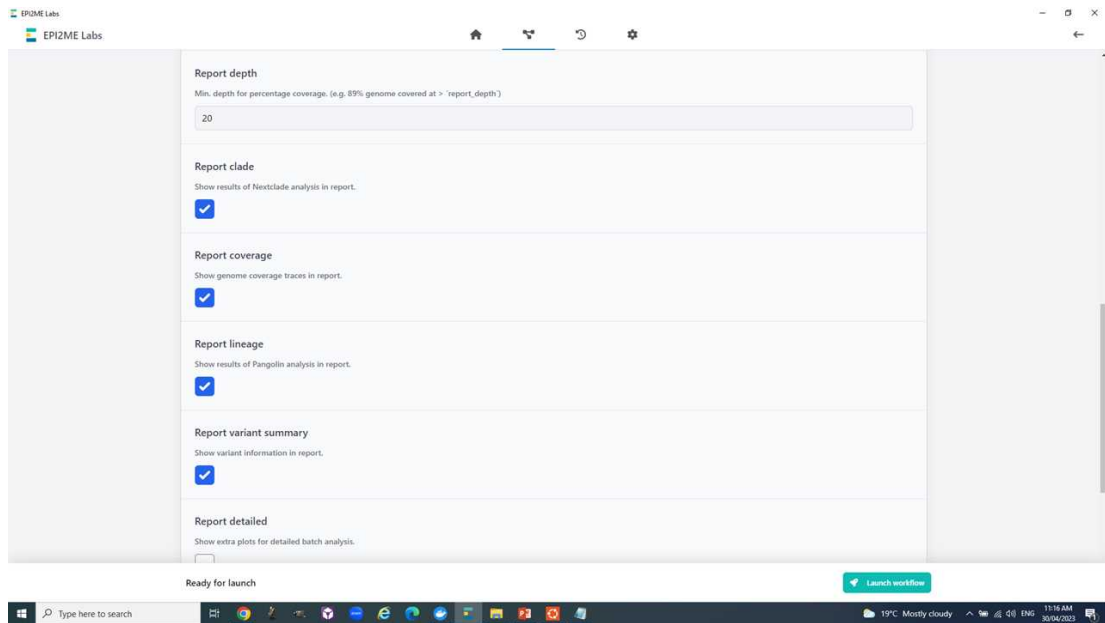
Advanced options for configuring processes inside the workflow.

Ready for launch

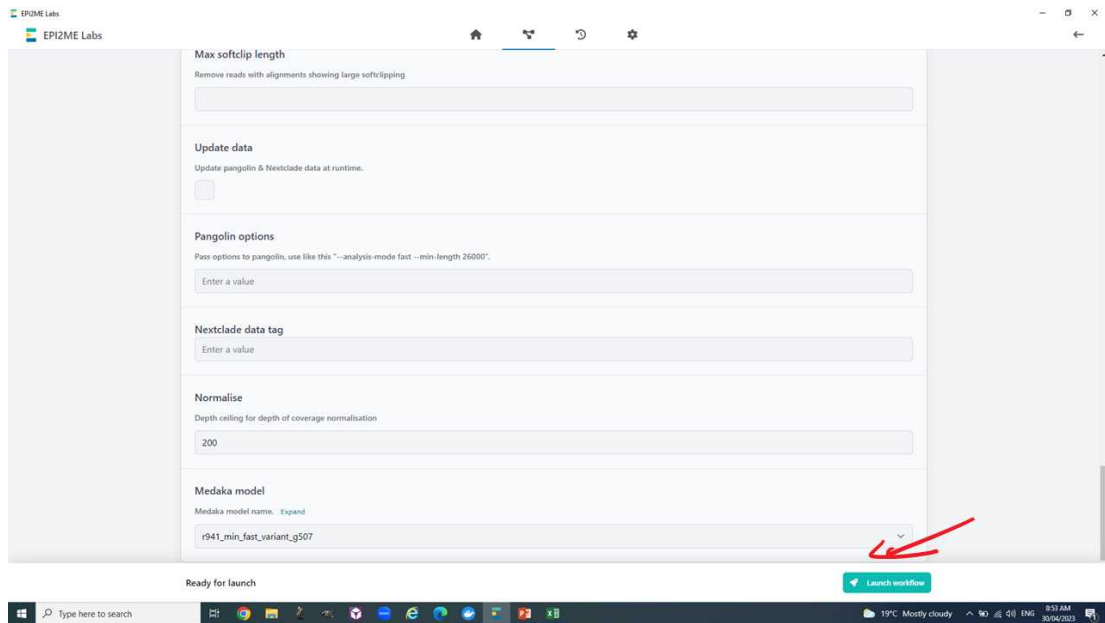
Launch workflow

Type here to search

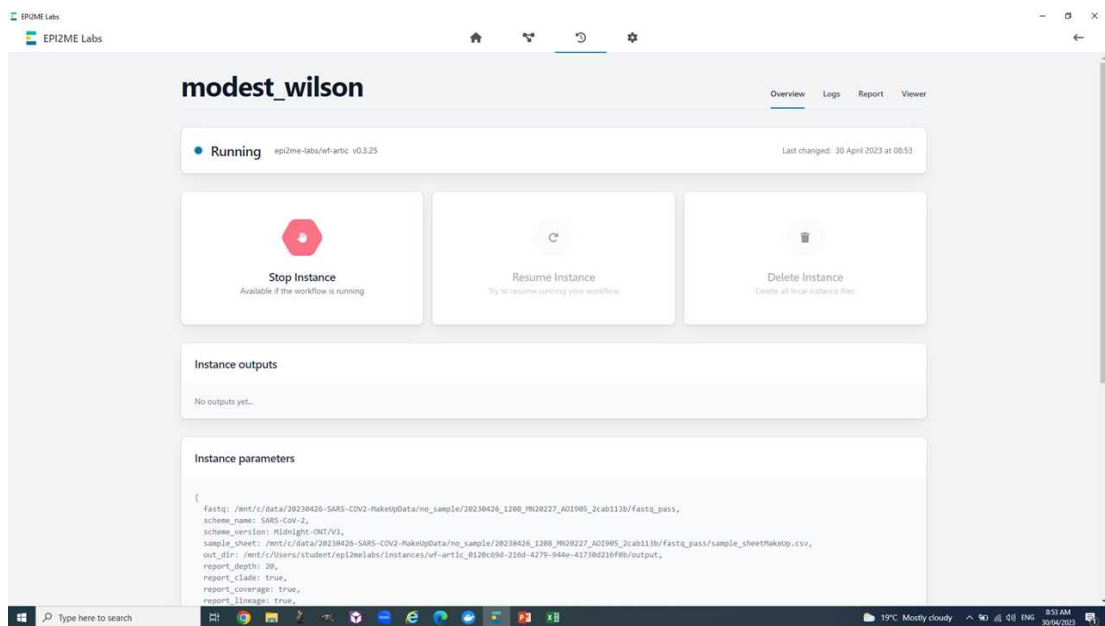
19°C Mostly cloudy 8:53 AM 30/04/2023



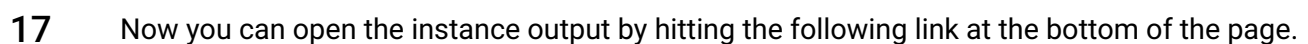
14 Hit the "Launch Workflow" button.

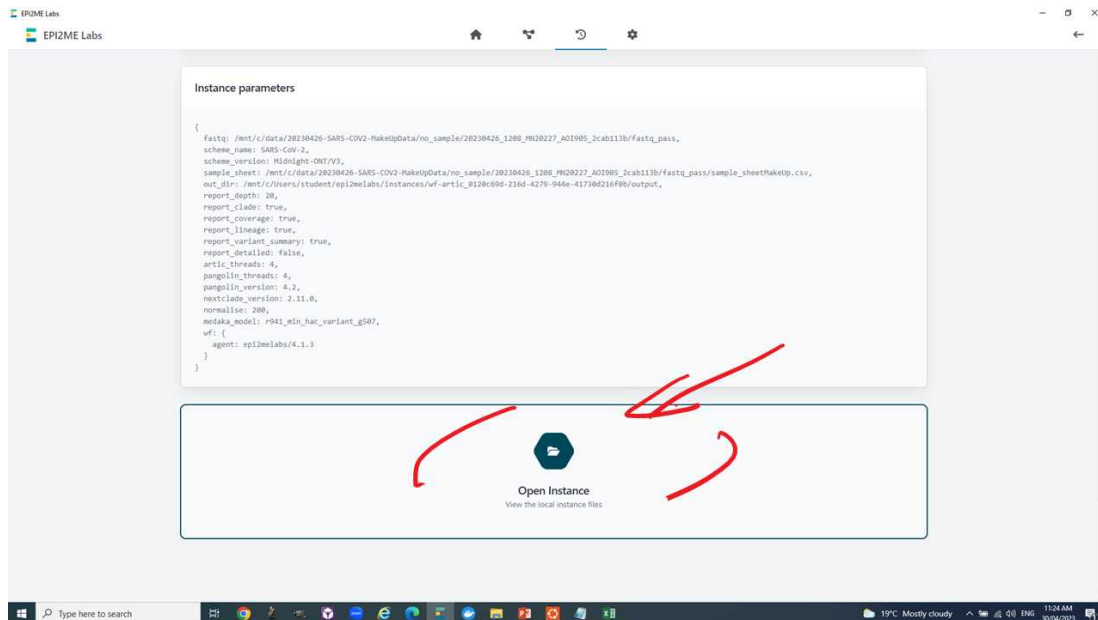


15 Now all should be running smoothly.

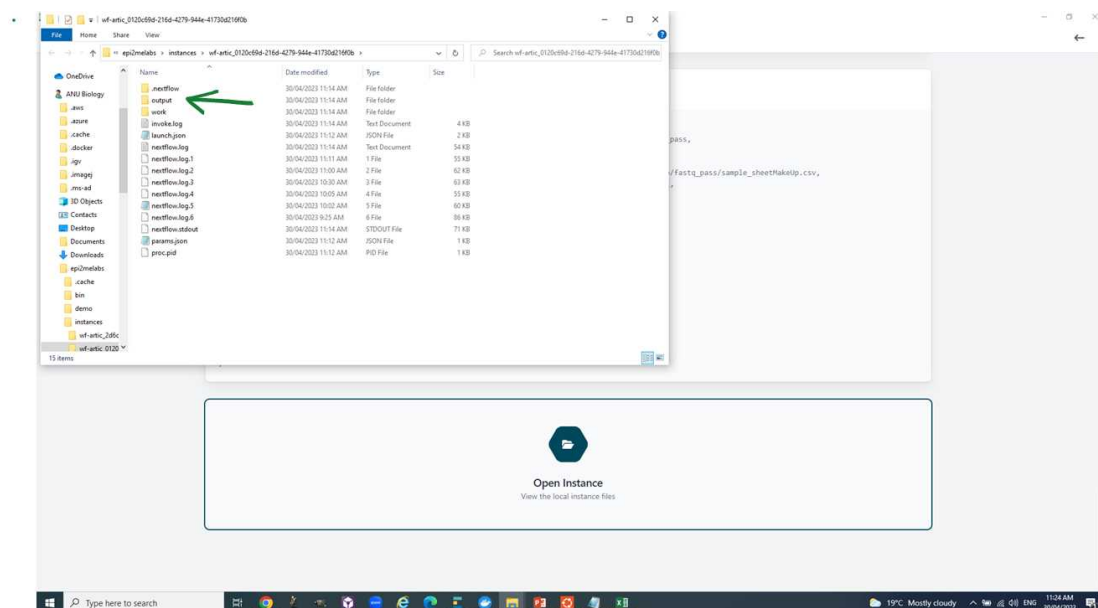


... and this after 2-5 minutes.

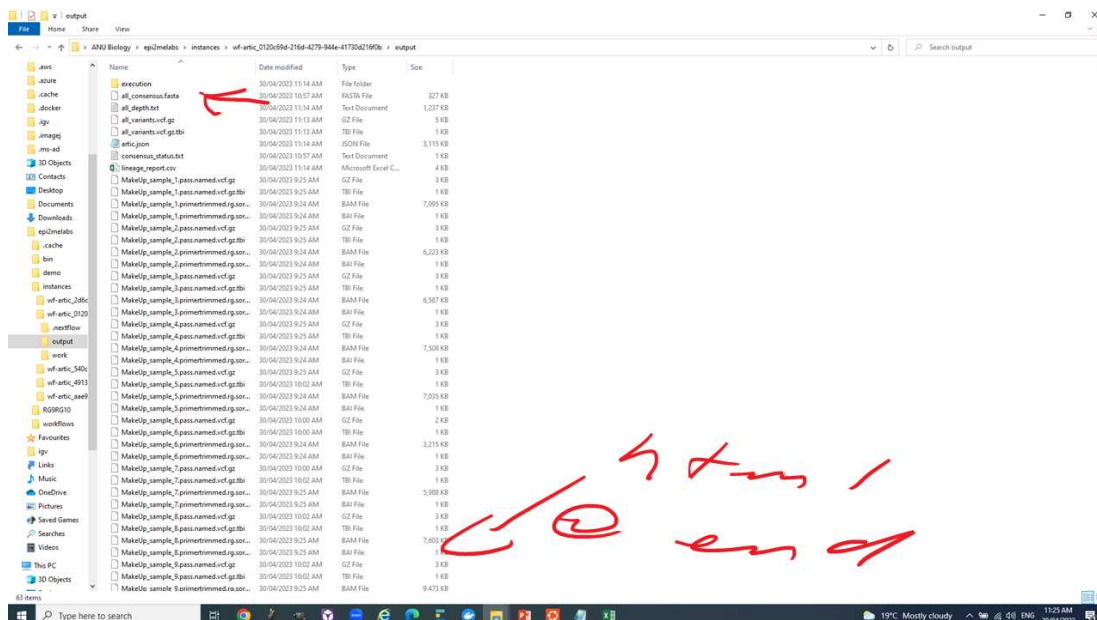




- 18 This will open the folder of your specific workflow instance that also contains your "output" folder.



- 19 This "output" folder contains LOADS of data. The most important is the "all_consensus.fasta" and the "wf-artic-report.html". We will also make use of some of the other files.



Now rename both files with unique and sensible identifiers for your group and share it with all the RG members.

20 The "all_consensus.fasta" contains your consensus genomes for each sample and the "wf-artic-report.html" looks like this. We will go over both in more detail during the prac.



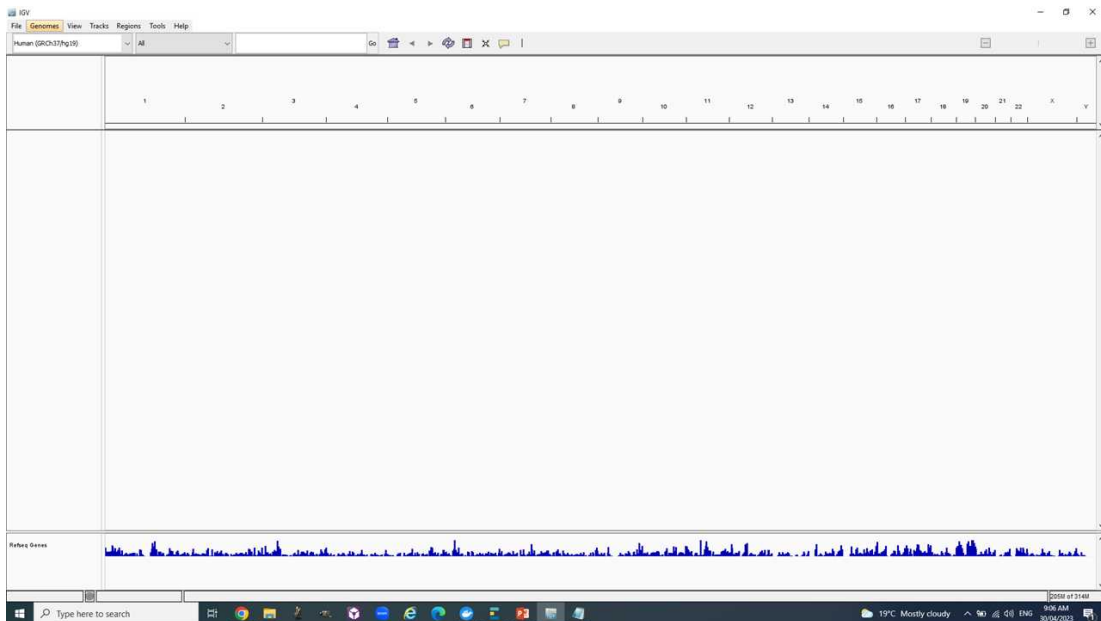
Snap shot of html summary report

21 We will use the "all_consensus.fasta" for most downstream analyses. It contains the consensus fasta sequences of your genomes.

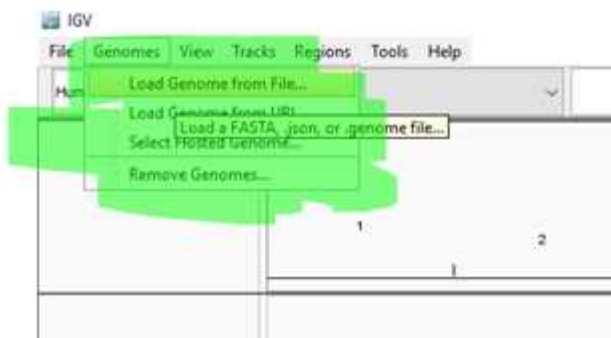
Second II: Visualization whole genome alignment and princ...

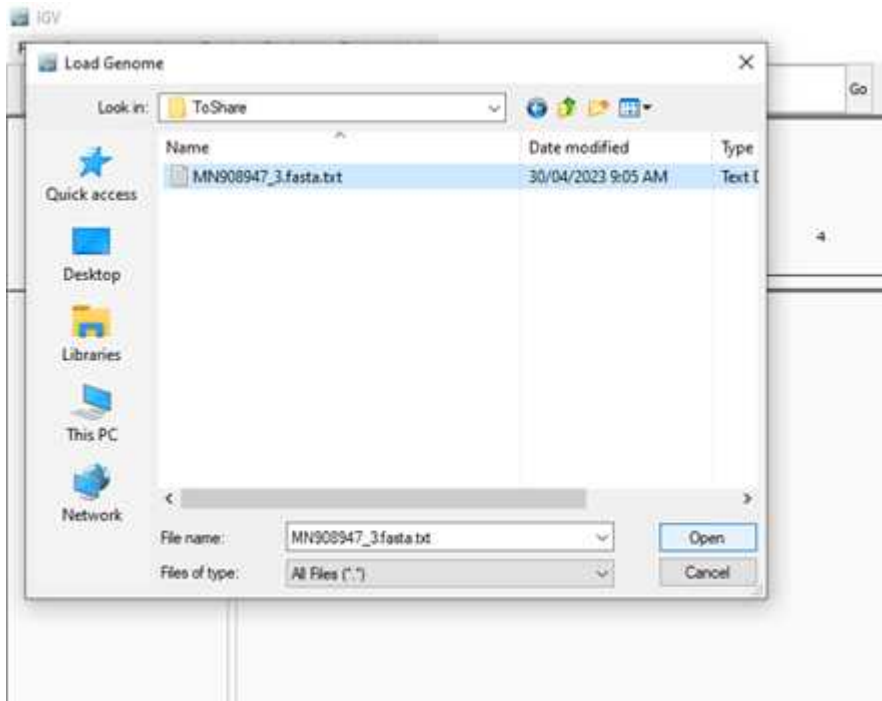
22 You will use IGV to look at one of your read sets mapped against the reference genome to get a better understanding of the "raw" data.

23 Open IGV.



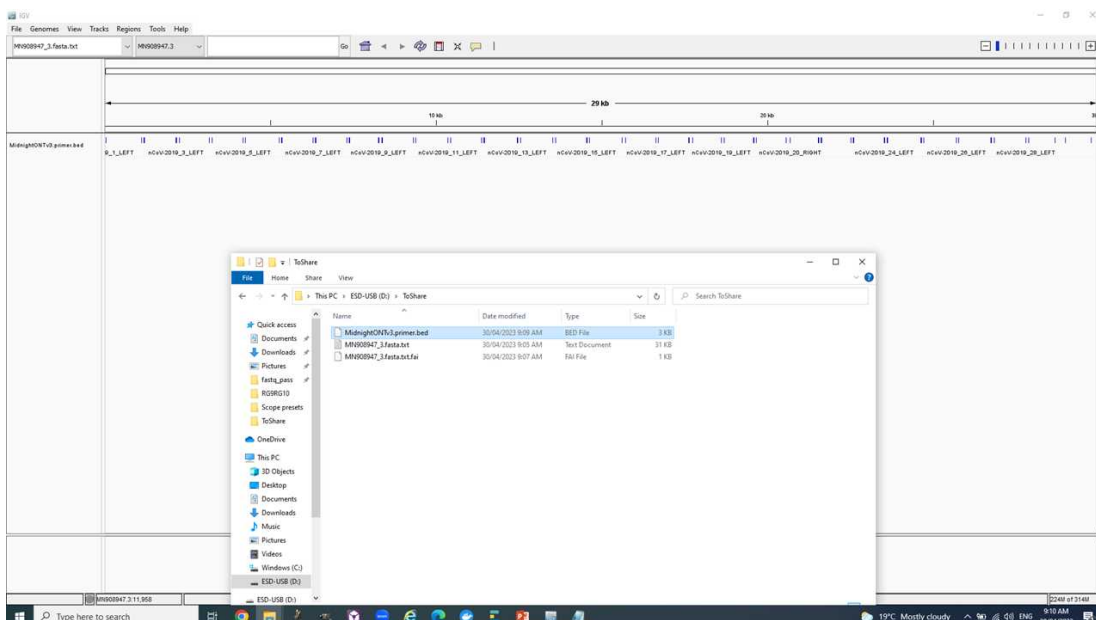
24 Select "Genomes" > "Load Genomes from File.." and add the "MN908947_3.fasta" reference shared [here](#) (ANU only).



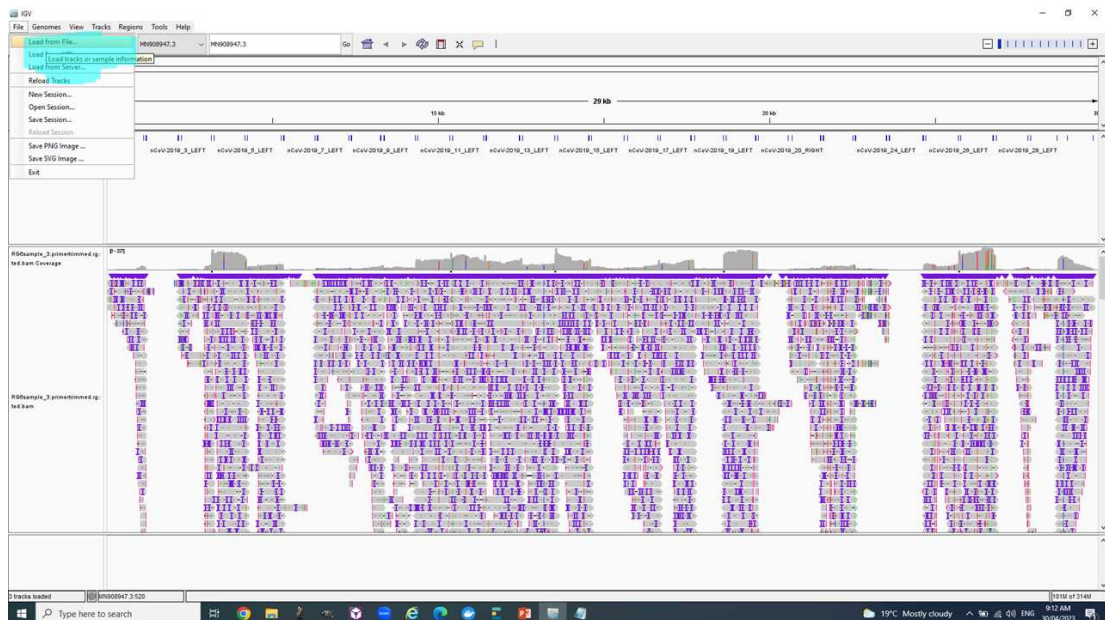


The file can end in ".txt" or better ".fasta"

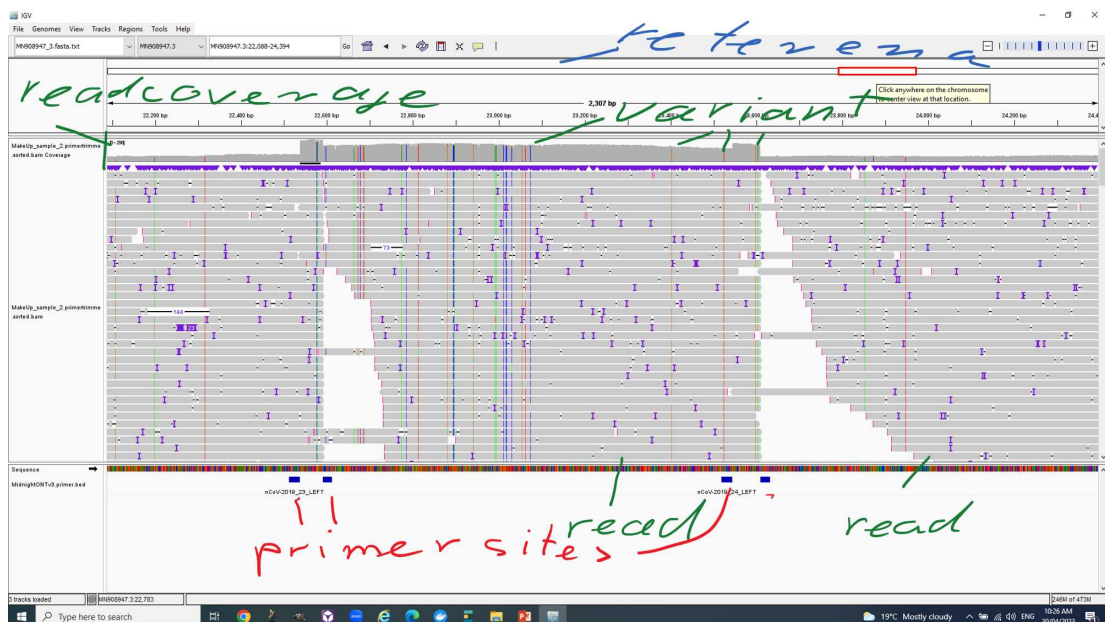
- 25 Drag and drop the "MidnightONTv3.primer.bed" into IGV to illustrate where your primer sequences bind.



- 26 Load one of the "XXX.primertrimmed.rg.sorted.bam" files from the "output" folder of Epi2Me [here](#) (ANU only) or from your own dataset.



27 Now you see the alignments of all the reads as pileup against the reference, the read coverage and variants.

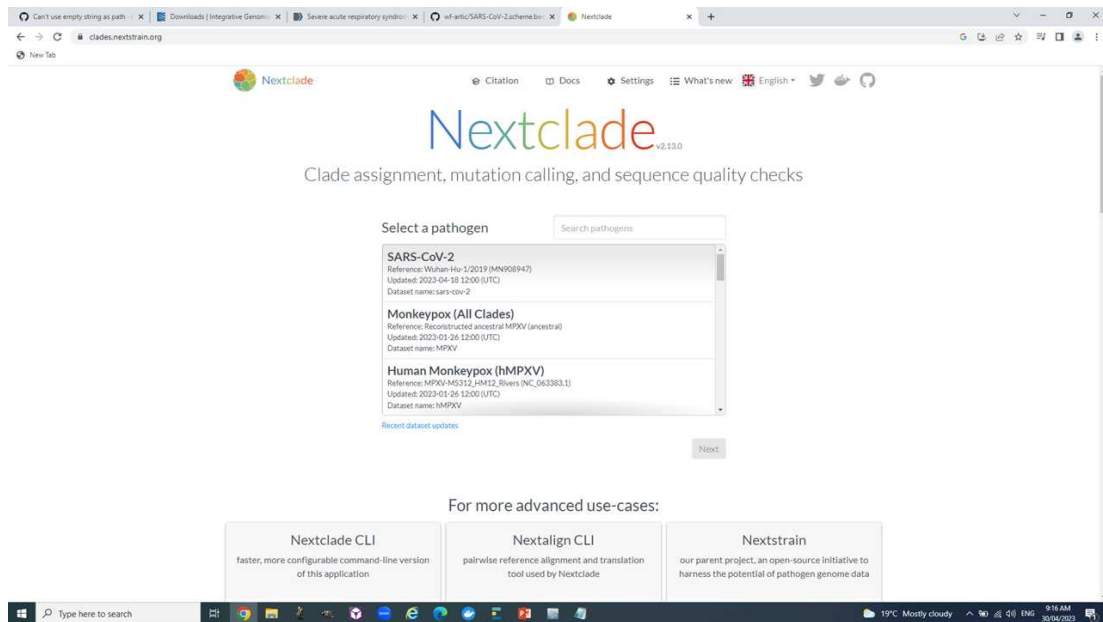


You can zoom into certain regions e.g. the S-gene 21563-25384 to see how the reads are different to the reference via the different colored bars in the grey area labelled with variants in the picture above. There is loads more to explore. Ask if you have any questions.

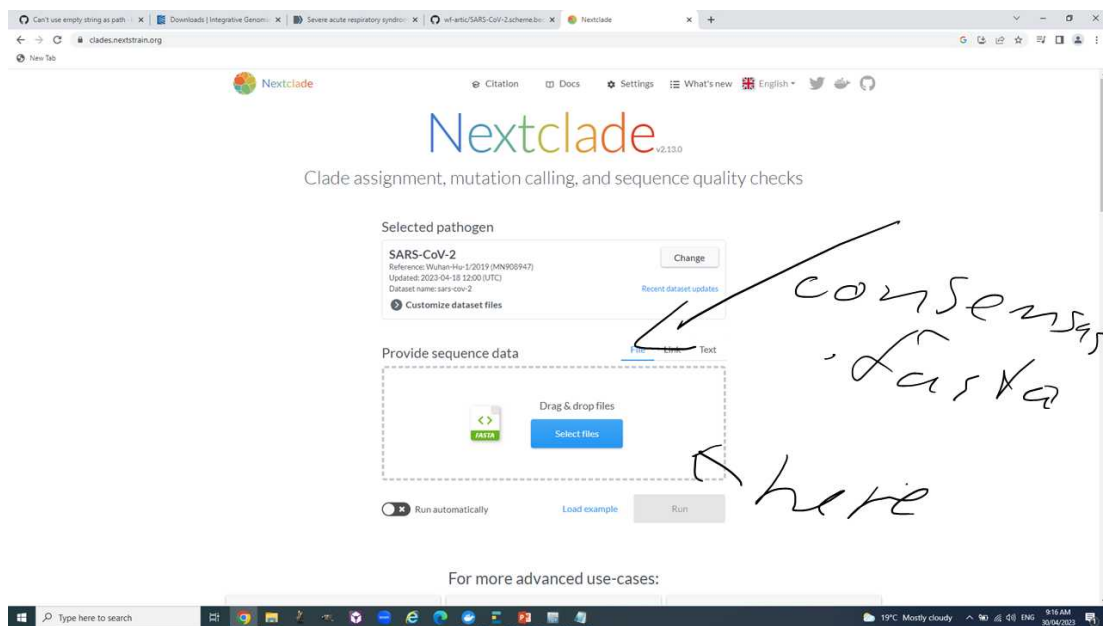
Section III: Nextclade for lineage calling, QC measures, am...

28 We will use Nextclade <https://clades.nextstrain.org/> to get some quality assessment metrics,

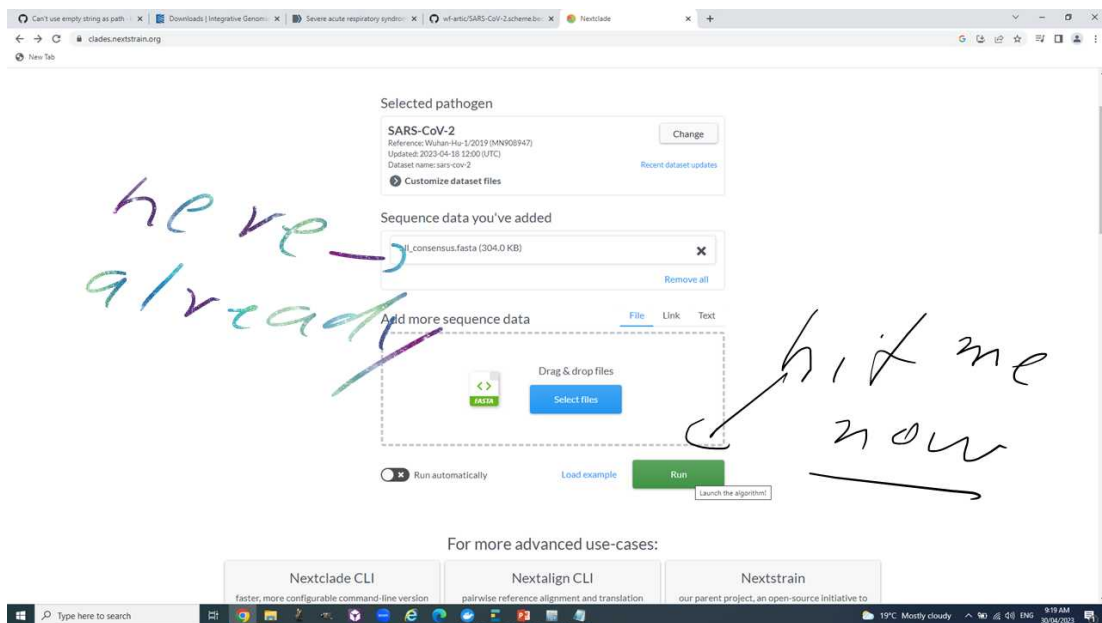
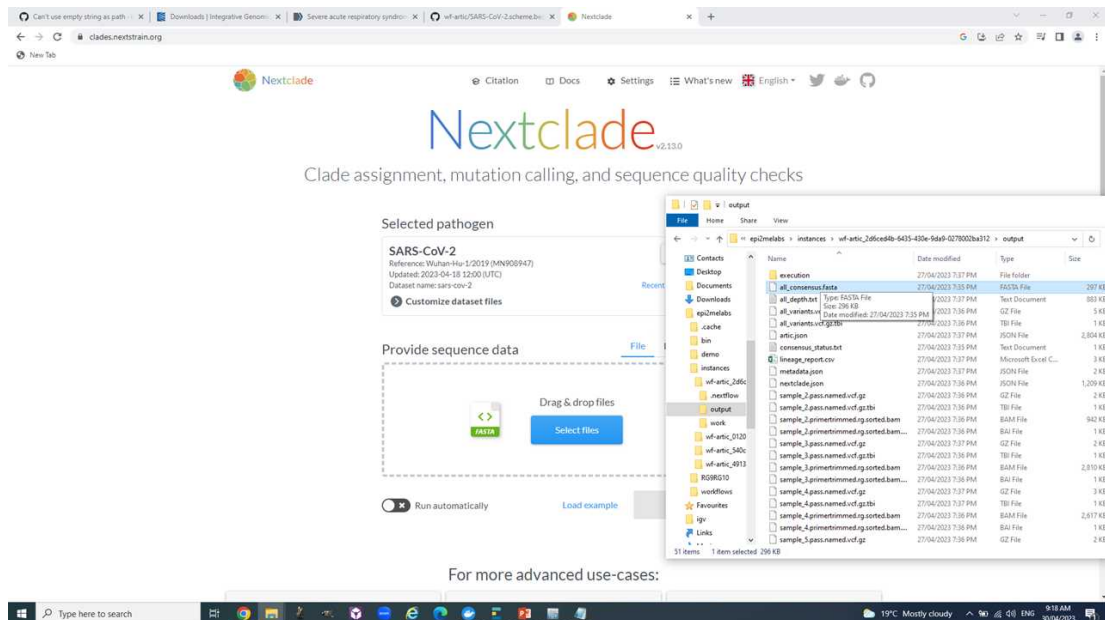
lineages calling, number of mutations, and much more.



- 29 Open the webpage <https://clades.nextstrain.org/>, select SARS-CoV-2 and drag and drop your consensus sequences into it.

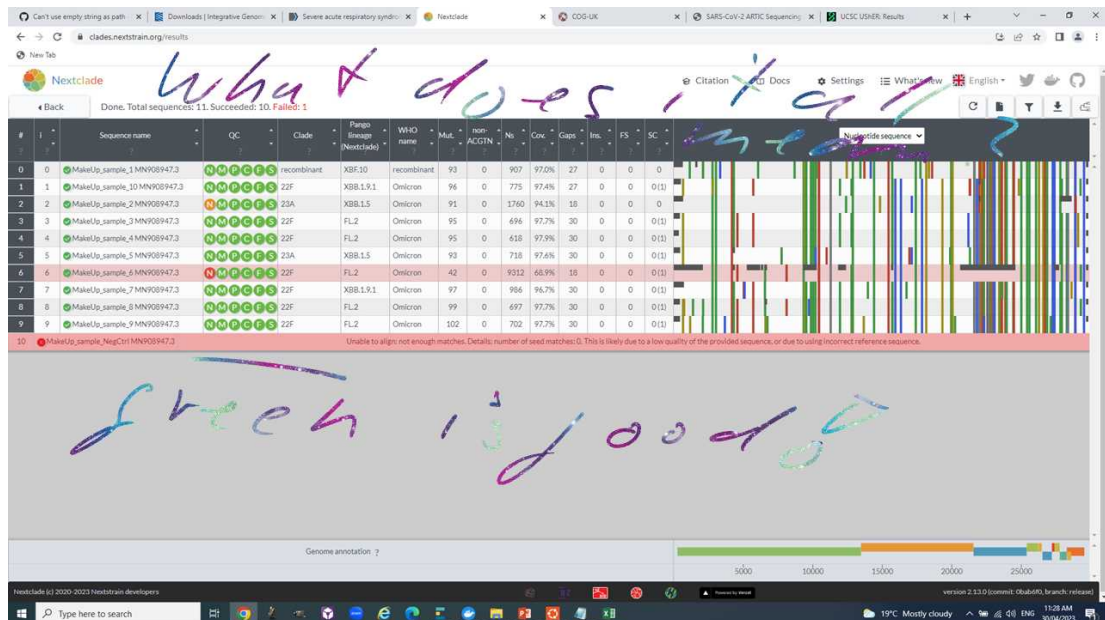


Pretty please!

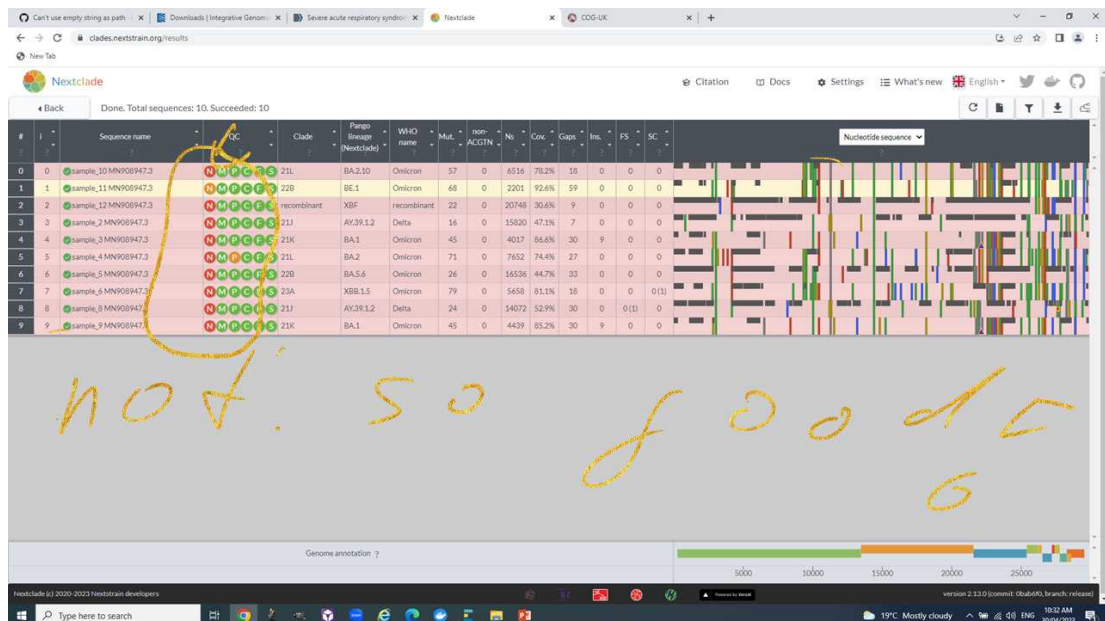


Thanks!

30 This is an example of a good Midnight consensus sequence set.

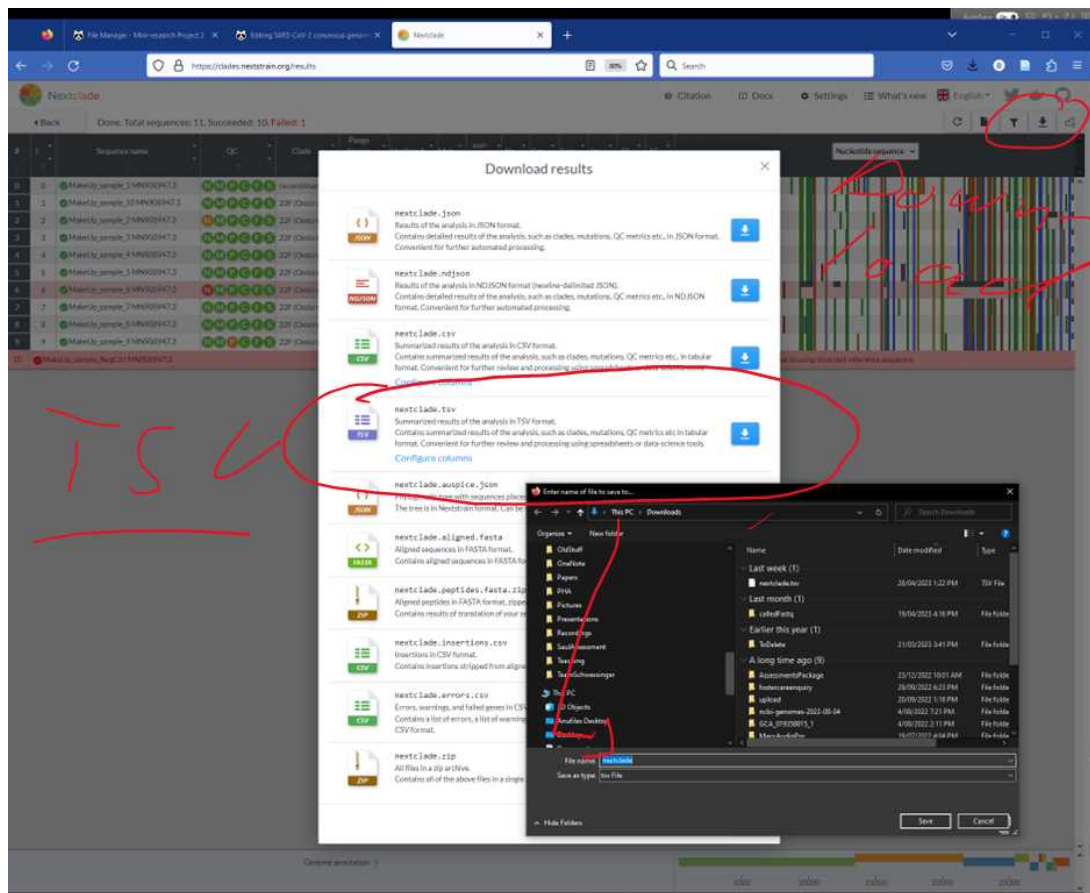


31 This is an example of a not so good Midnight consensus sequence set.



- 32 Explore the results page to get to the following:
1. Pango lineage assignment.
 2. Genome coverage.
 3. Number of mutations relative to reference.
 4. Number of Ns.
 5. Gaps.
 6. Changes to the amino acid sequence of the S protein.

- 33 Make sure to download your samples via the Download dialog in the "TSV" (tab separated values) format. You can import this into Excel later.

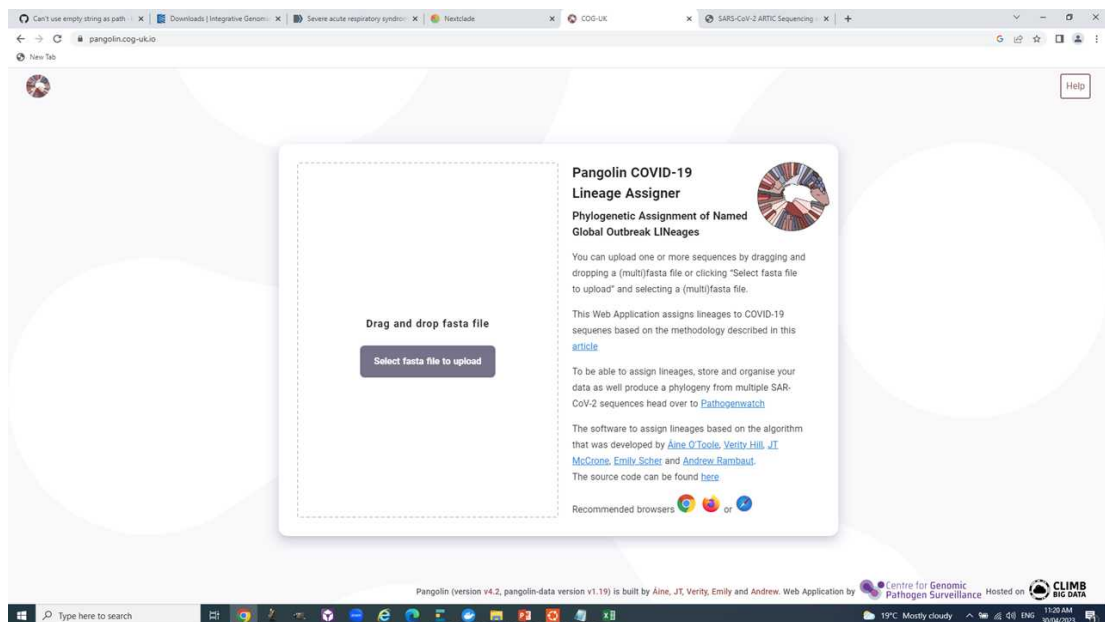


Ask question during the prac for anything that is unclear.

Section IV: Pangolin lineage assignment

- 34 Let's assign the latest lineage designation with Pangolin.

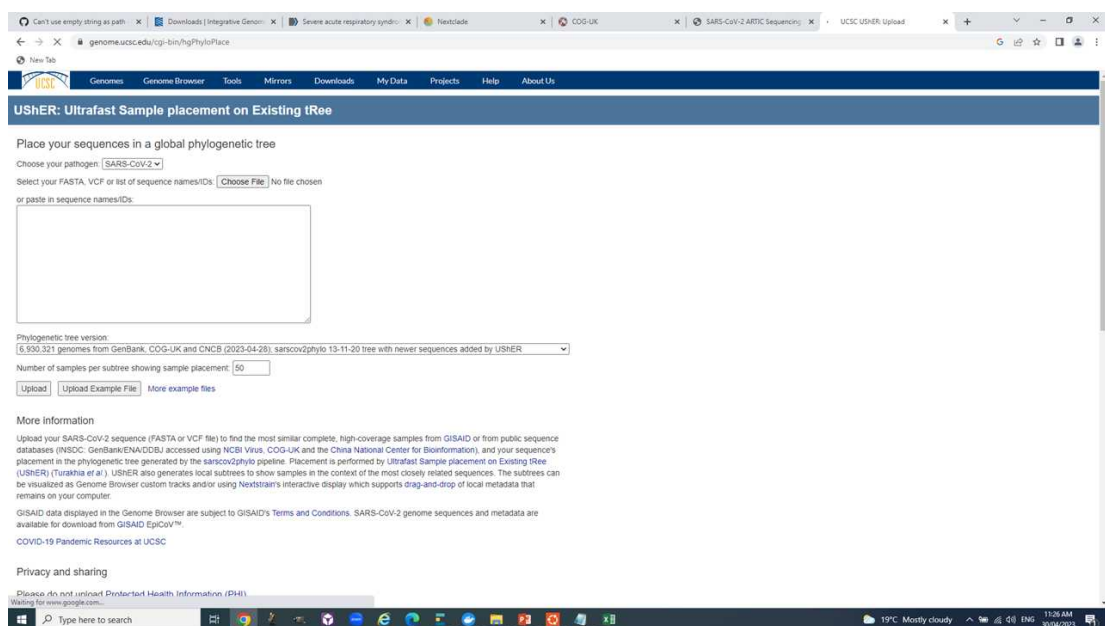
Navigate to <https://pangolin.cog-uk.io/>.



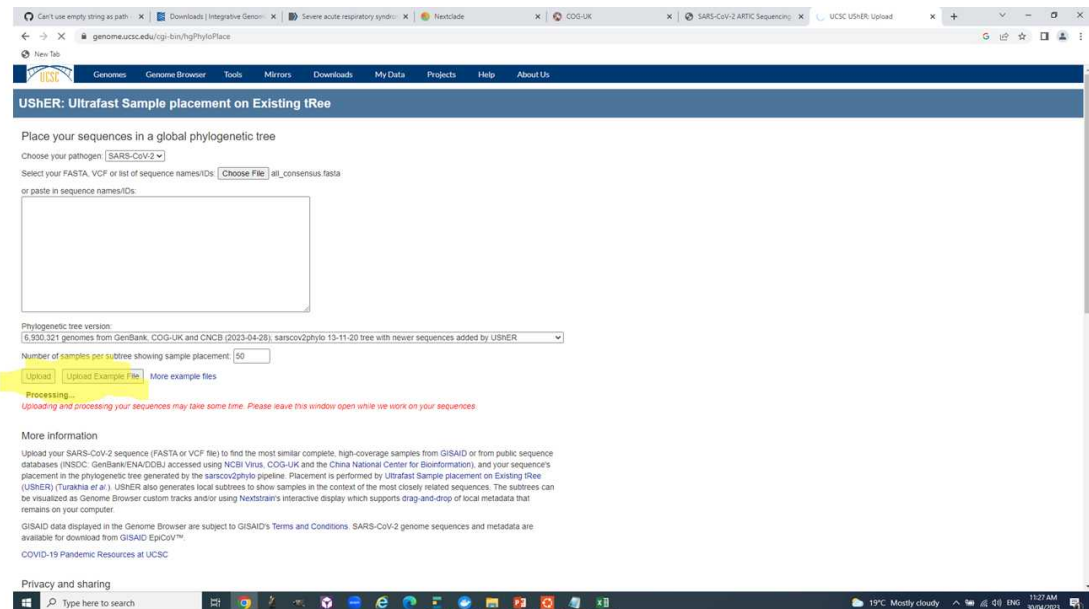
Drag and drop your sequences, start analysis, and hope for the best.

Step V: Use Usher to place your samples into a global conte..

35 Navigate to <https://genome.ucsc.edu/cgi-bin/hgPhyloPlace>.



36 Upload your sequences



37 Explore the webpage for results including:

1. Neighboring sample in Tree.
2. That ^^ samples deposition date.
3. Subtree of each of your sequences.

Sequence MakeUp_sample_NegCtrl has too few bases (0 excluding 29503 Ns at beginning and 0 Ns at end), must have at least 10000; skipping

view in Genome Browser | view downsampld global tree in Nextstrain

If you have metadata you wish to display, click a View subtree in Nextstrain button, and then you can drag on a CSV file to add it to the tree view.

Note: The Nextstrain subtree views, and Download files below, are temporary files and will expire within two days. Please download the Nextstrain subtree JSON files if you want to view them again in the future. The JSON files can be drag-dropped onto <https://auspice.usf>

Downloads: | Global phylogenetic tree with your sequences | TSV summary of sequences and placements | TSV summary of Spike mutations | ZIP file of subtree JSON and Newick files

Fasta Sequence	Size (T)	#Ns (T)	#Mixed (T)	Bases aligned (T)	Inserted bases (T)	Deleted bases (T)	#SNVs used for placement (T)	#Masked SNVs (T)	Nextstrain clade (T)	Pango lineage (T)	Neighboring sample in tree (T)	Lineage of neighbor (T)	#Imputed values for mixed bases (T)	#Maximally parsimonious placements (T)	Parsimony score (T)	Subtree number (T)
MakeUp_sample_1	28970 (T)	1 (T)	0	28969 (T)	0	27 (T)	92 (T)	1 (T)	22D (Omicron)	XBF	BHR/750600659/2023 OQ749542.1 2023-02-17	BA.2	0	3	13	3 (view in Nextstrain)
MakeUp_sample_10	29104 (T)	3 (T)	0	29101 (T)	0	27 (T)	85 (T)	1 (T)	22F (Omicron)	XBB.1.9.1	BHR/740906780/2023 OQ439604.1 2023-02-02	BA.2	0	1	3	4 (view in Nextstrain)
MakeUp_sample_2	28955 (T)	850 (T)	0	28125 (T)	0	18 (T)	90 (T)	1 (T)	22F (Omicron)	XBB.1.5	CHN/SH-NS204-2117/2023 C_AA009479.1 2023-04-04	BA.2	0	1	4	5 (view in Nextstrain)
MakeUp_sample_3	29177 (T)	0 (T)	0	29173 (T)	0	30 (T)	93 (T)	0	22F (Omicron)	FL.2	USA/CA-CDC-1C102656/2023 OQ601391.1 2023-02-28	BA.2	0	1	0	1 (view in Nextstrain)
MakeUp_sample_4	29255 (T)	0 (T)	0	29251 (T)	0	30 (T)	93 (T)	0	22F (Omicron)	FL.2	USA/CA-CDC-1C102656/2023 OQ601391.1 2023-02-28	BA.2	0	1	0	1 (view in Nextstrain)
MakeUp_sample_5	29156 (T)	1 (T)	0	29155 (T)	0	30 (T)	92 (T)	1 (T)	22F (Omicron)	XBB.1.5	USA/OH-CDC-QD-X47312602/2023 OQ610777.1 2023-02-27	BA.2	0	1	2	6 (view in Nextstrain)
MakeUp_sample_6	27536 (T)	8963 (T)	0	20573 (T)	0	18 (T)	42 (T)	0	22F (Omicron)	FL.2	BHR/221207666/2023 OQ439574.1 2023-02-06	BA.2	0	1	1	2 (view in Nextstrain)
MakeUp_sample_7	28887 (T)	0 (T)	0	28887 (T)	0	30 (T)	96 (T)	1 (T)	22F (Omicron)	FL.4	England/PHEC-YYEIG3W/2023 2023-04-14	BA.2	0	1	2	7 (view in Nextstrain)
MakeUp_sample_8	29177 (T)	1 (T)	0	29176 (T)	0	30 (T)	88 (T)	1 (T)	22F (Omicron)	FL.2	USA/CA-CDC-1C1038863/2023 OQ609386.1 2023-03-11	BA.2	0	1	3	2 (view in Nextstrain)
MakeUp_sample_9	29171 (T)	0 (T)	0	29171 (T)	0	30 (T)	190 (T)	2 (T)	22F (Omicron)	FL.2	USA/IL-CDC-1C1008907/2023 OQ457953.1 2023-02-08	BA.2	0	1	8	8 (view in Nextstrain)

Subtree 1: 2 related samples

view subtree 1 in Nextstrain

38 Ask questions in during the prac.

