WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

World Health Organization
Philippines

KDCA
Korea Disease Control and Prevention Agency

Quality Control on Consensus Assembly

John Michael Egana
Core Facility for Bioinformatics
Philippine Genome Center
University of the Philippines

17 April 2024

CENTRE FOR PATHOGEN GENOMICS

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Consensus Assembly

The output of the process of generating **a single representative sequence** from a mapped set of reads to a reference sequence. In other cases, from a set of similar sequences that are also multiple sequence aligned.

Challenges in Consensus Assembly

- **Genetic variability**: Viruses often exhibit high mutation rates and genetic diversity.
- **Sequence errors**: Errors introduced during sequencing, such as base calling errors, insertions, deletions, sequence repeats, and chimeric sequences.
- **Mixed infections**: Presence of multiple viral variants within the same sample.
- **Laboratory contamination:** Unintended contamination due to human error in the laboratory.
- **Computational artifacts:** Bioinformatic tools used to analyze data and create a final genome can sometimes introduce errors.

Methods for Quality Control

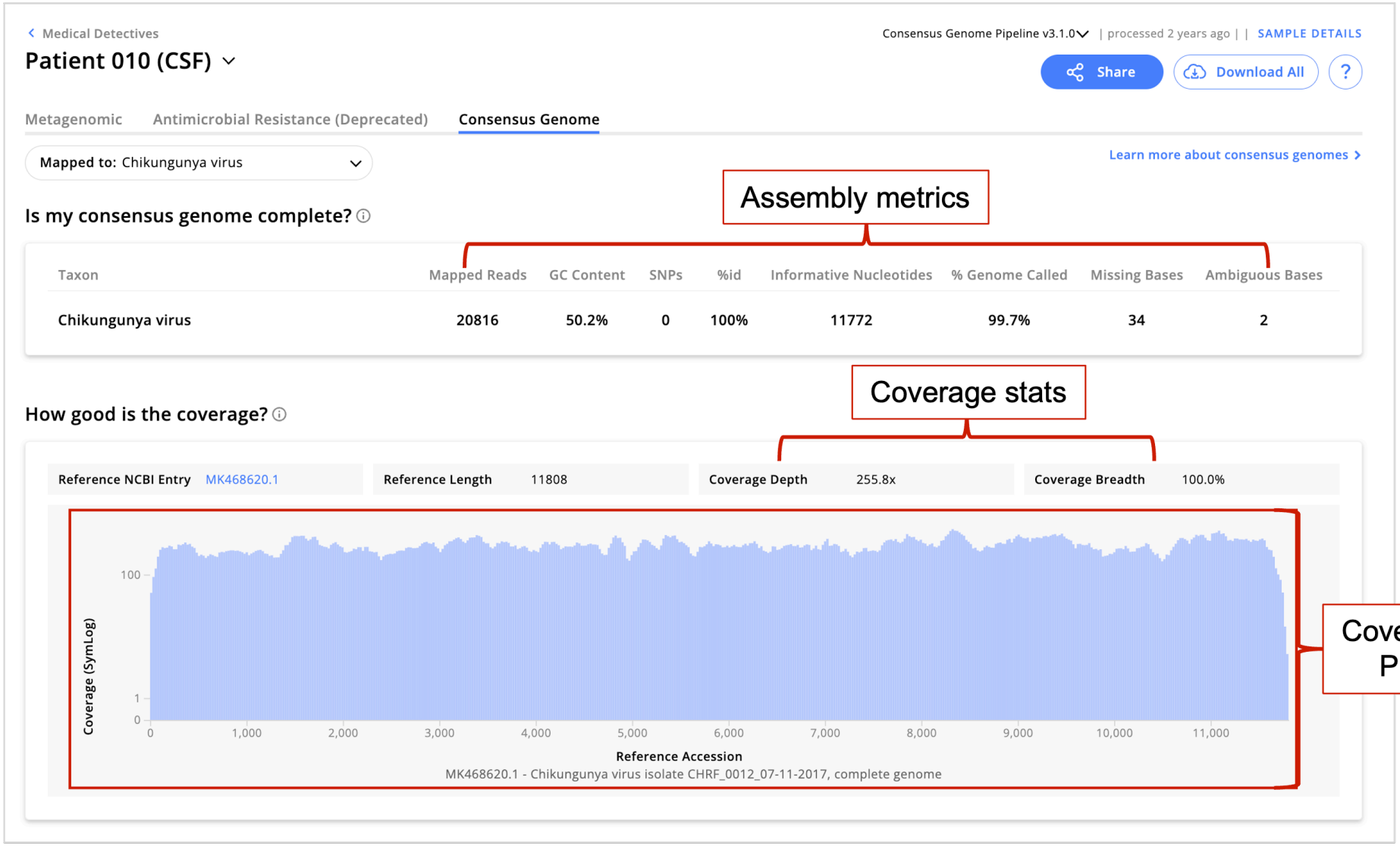
- **Pre-assembly**
  - **Read Filtering**: Exclude low-quality reads and adapter sequences to improve assembly accuracy.
  - **Error Correction**: Utilize tools for error correction, such as k-mer-based correction or consensus-based correction algorithms, e.g. minimum frequency threshold, minimum depth to call consensus
- **Alignment or read mapping**
  - **Alignment-based Filtering**: Align reads and/or consensus sequences to a reference genome to identify and remove erroneous reads or artifacts or gaps
  - **Variant Calling**: Identify and filter out variants that are likely sequencing errors rather than true genetic variation.
  - **Depth of Coverage Analysis**: Assess the depth of coverage across the genome to identify regions with low coverage or potential errors.
- **Cross-Validation**: Validate the consensus sequence with independent sequencing data or experimental methods like PCR and Sanger sequencing.
- **Recombination Detection:** Detect potential product of recombination.

Assembly metrics

- Coverage plot
- % Genome Called
- SNPs
- Informative bases
- Mapped reads
- GC content

https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

Assembly metrics



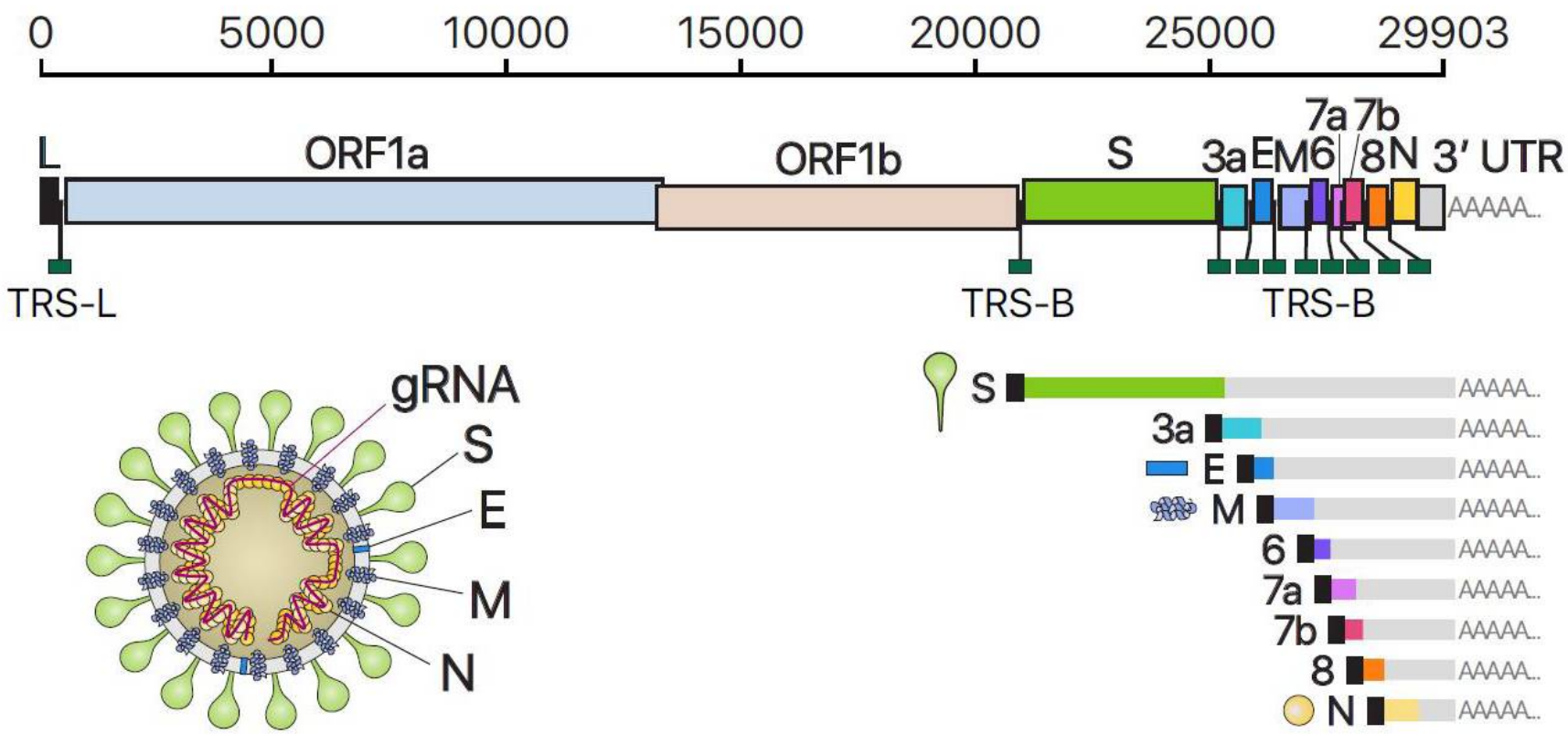
https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

Organism-specific considerations

- **Genome size:** How many base pairs is it?
- **Genome organization:** How many open reading frames (ORFs)? What is their orientation?
- **Repeat regions:** Are there known repeat regions? What are their positions? If reads don't span this region (i.e., region covered by single reads) the assembly or consensus sequence over these regions should not be trusted.
- **Low or high GC content areas:** Are there genomic regions with low or high GC content? Is important to inspect the assembly over these regions because they are prone to have sequencing bias or errors.

https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

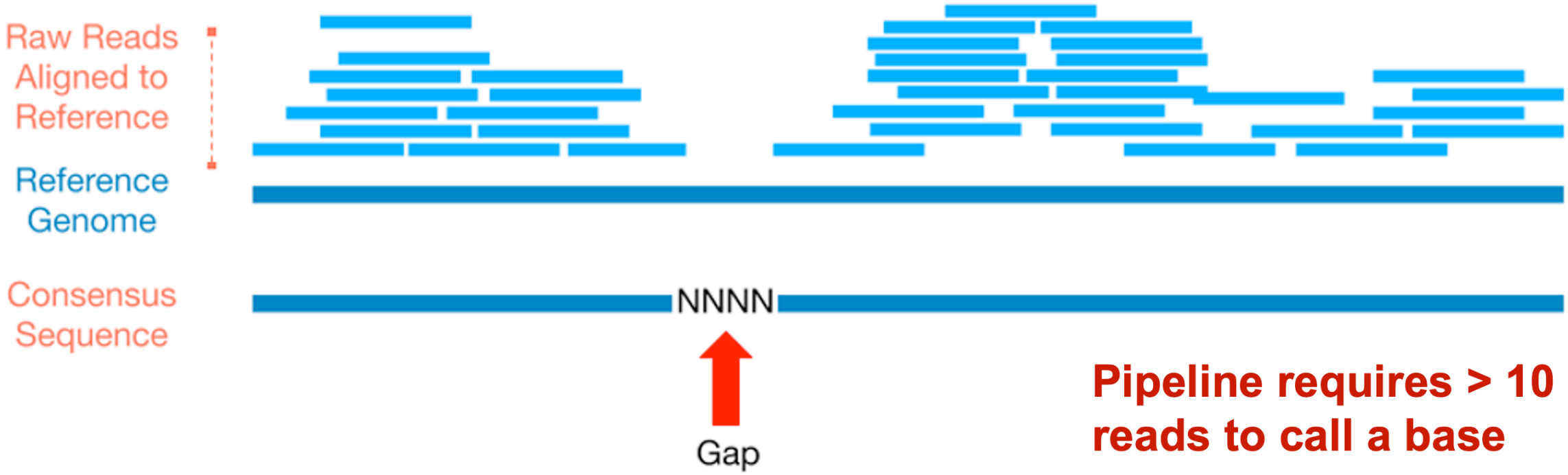
SARS-CoV-2 Genome



https://www.epigentek.com/catalog/insights-into-the-sars-cov-genome-transcriptome-and-epitranscriptome-n-40.html

Checking read alignments

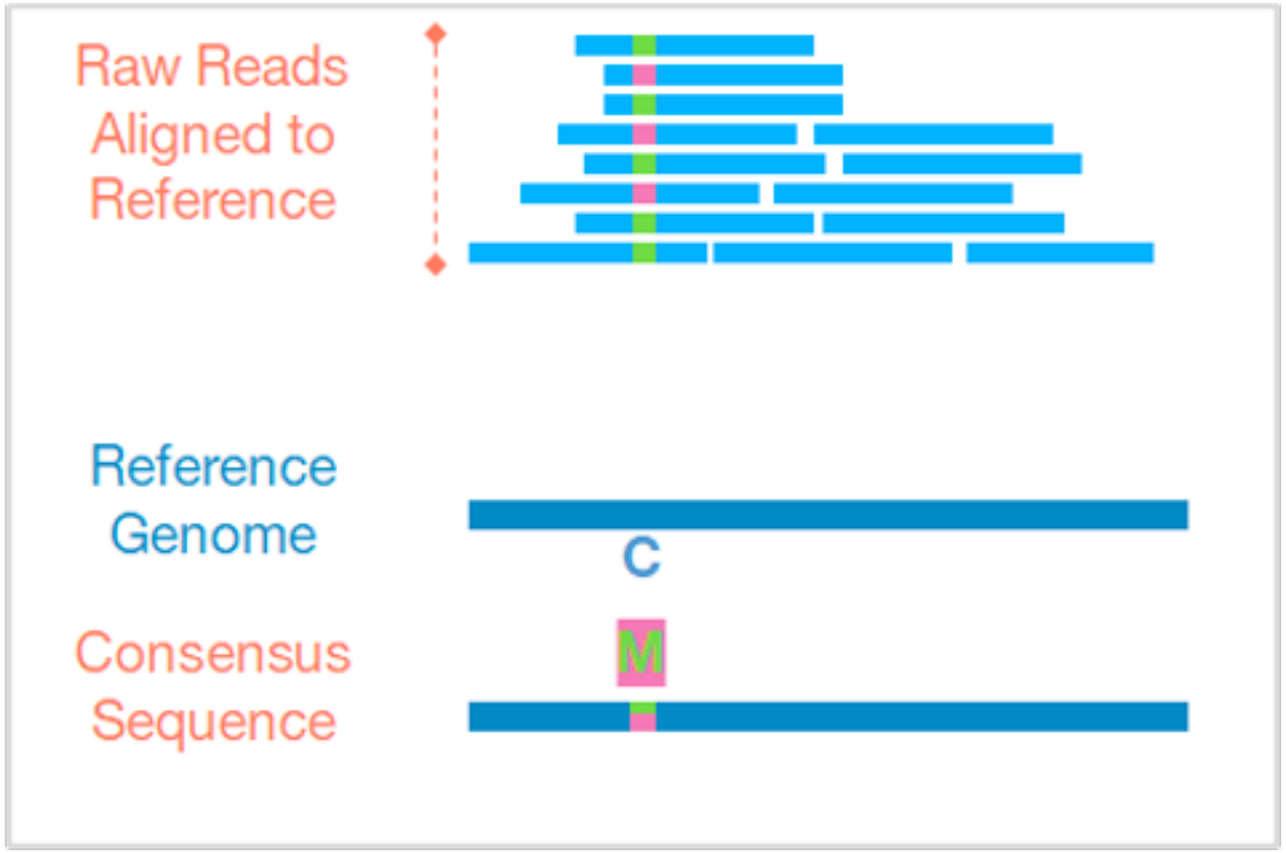
- Gaps or low coverage of consensus genome



https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

Checking read alignments

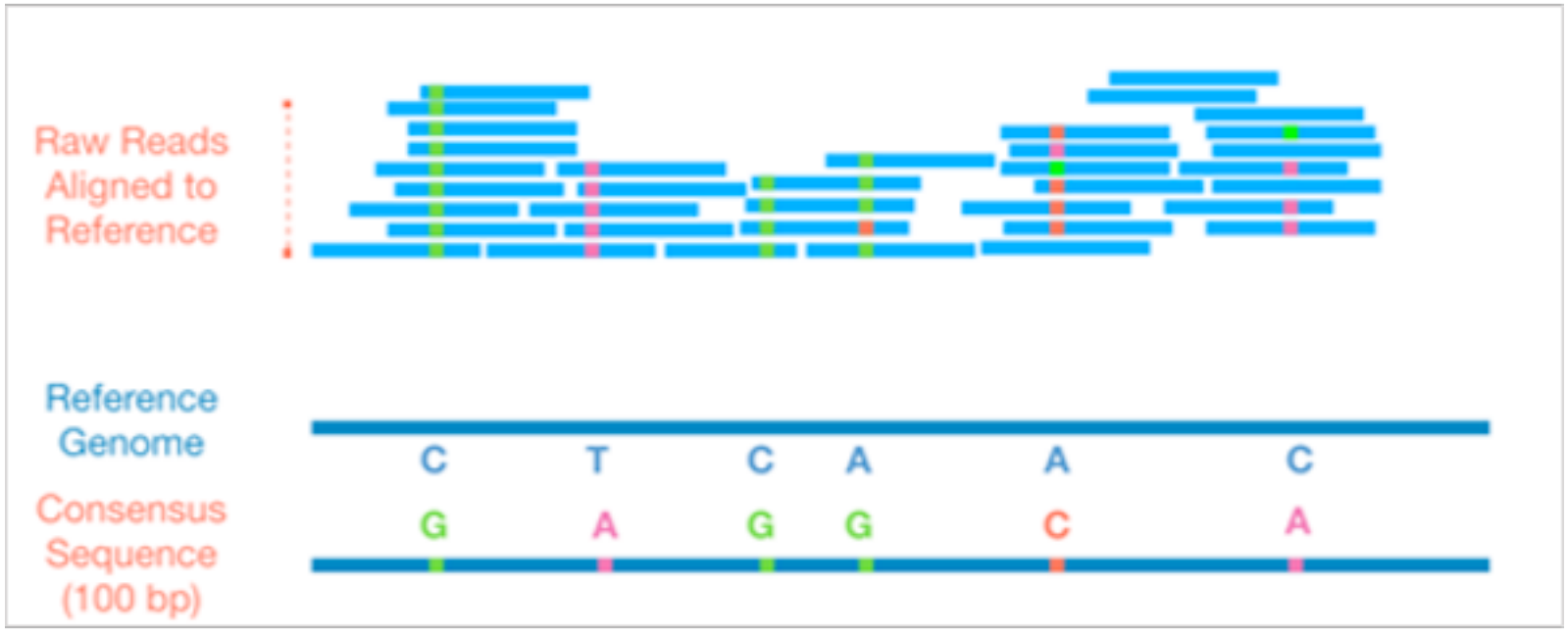
- Ambiguous bases



https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

Checking read alignments

- Single nucleotide polymorphisms (SNPs)



https://chanzuckerberg.zendesk.com/hc/en-us/articles/13619776085780-Consensus-Genome-Quality-Checks

GISAID Quality Control

CENTRE FOR PATHOGEN GENOMICS · Doherty Institute · The University of Melbourne · The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Virus name	Passage de	Accession ID	Collection da	Submission D		Length	Host	Location	Originating
hCoV-19/Philippines/PH-PGC-132210/2022	Original	EPI_ISL_17389138	2022-09-26	2023-04-04	ML	29,867	Human	Asia / Philippines	DAVAO INT
hCoV-19/Philippines/PH-PGC-132213/2022	Original	EPI_ISL_17389137	2022-09-26	2023-04-04					
hCoV-19/Philippines/PH-PGC-131761/2022	Original	EPI_ISL_17389136	2022-09-26	2023-04-04	ML	29,838	Human	Asia / Philippines	KIDAPAWA
hCoV-19/Philippines/PH-PGC-132196/2022	Original	EPI_ISL_17389135	2022-09-26	2023-04-04	ML	29,847	Human	Asia / Philippines	OTHERS -
hCoV-19/Philippines/PH-PGC-131762/2022	Original	EPI_ISL_17389134	2022-09-26	2023-04-04	ML	29,838	Human	Asia / Philippines	KIDAPAWA
hCoV-19/Philippines/PH-PGC-132215/2022	Original	EPI_ISL_17389133	2022-09-26	2023-04-04	ML	29,878	Human	Asia / Philippines	DAVAO INT
hCoV-19/Philippines/PH-PGC-131727/2022	Original	EPI_ISL_17389132	2022-09-25	2023-04-04	ML	29,880	Human	Asia / Philippines	ST. ELIZAB
hCoV-19/Philippines/PH-PGC-131723/2022	Original	EPI_ISL_17389131	2022-09-25	2023-04-04	ML	29,882	Human	Asia / Philippines	ST. ELIZAB
hCoV-19/Philippines/PH-PGC-131728/2022	Original	EPI_ISL_17389130	2022-09-25	2023-04-04	ML	29,854	Human	Asia / Philippines	ST. ELIZAB
hCoV-19/Philippines/PH-PGC-131726/2022	Original	EPI_ISL_17389129	2022-09-25	2023-04-04	ML	29,877	Human	Asia / Philippines	ST. ELIZAB
hCoV-19/Philippines/PH-PGC-131725/2022	Original	EPI_ISL_17389128	2022-09-25	2023-04-04	ML	29,846	Human	Asia / Philippines	ST. ELIZAB
hCoV-19/Philippines/PH-PGC-131680/2022	Original	EPI_ISL_17389127	2022-09-24	2023-04-04	ML	29,866	Human	Asia / Philippines	DR. ARTUF
hCoV-19/Philippines/PH-PGC-131682/2022	Original	EPI_ISL_17389126	2022-09-24	2023-04-04	⚠	29,249	Human	Asia / Philippines	DR. ARTUF

This submission requires further investigation! It appears to contain markers o
multiple lineages from both Delta and Omicron variants

GISAID Quality Control

CENTRE FOR PATHOGEN GENOMICS · Doherty Institute · The University of Melbourne · The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Virus name	Passage de	Accession ID	Collection da	Submission D	Length	Host	Location	Originating
hCoV-19/Philippines/PH-PGC-66977/2021	Original	EPI_ISL_5547858	2021-08-02	2021-10-26	29,847	Human	Asia / Philippines	Teresita Jal
hCoV-19/Philippines/PH-PGC-68464/2021	Original	EPI_ISL_5547879	2021-08-07	2021-10-26				
hCoV-19/Philippines/PH-PGC-68345/2021	Original	EPI_ISL_5547908	2021-08-06	2021-10-26				
hCoV-19/Philippines/PH-PGC-64107/2021	Original	EPI_ISL_5547982	2021-07-23	2021-10-26	29,857	Human	Asia / Philippines	Zamboanga
hCoV-19/Philippines/PH-PGC-69269/2021	Original	EPI_ISL_5548279	2021-08-09	2021-10-26	29,855	Human	Asia / Philippines	Philippine R
hCoV-19/Philippines/PH-PGC-68358/2021	Original	EPI_ISL_5548686	2021-08-06	2021-10-26	29,819	Human	Asia / Philippines	Dr. Jose N.
hCoV-19/Philippines/PH-PGC-66688/2021	Original	EPI_ISL_5548835	2021-07-27	2021-10-26	29,869	Human	Asia / Philippines	Qualimed H
hCoV-19/Philippines/PH-PGC-67607/2021	Original	EPI_ISL_5549643	2021-08-02	2021-10-26	29,853	Human	Asia / Philippines	Northern Mi
hCoV-19/Philippines/PH-PGC-66494/2021	Original	EPI_ISL_5550705	2021-07-26	2021-10-26	29,846	Human	Asia / Philippines	Western Vis
hCoV-19/Philippines/PH-PGC-67057/2021	Original	EPI_ISL_5550827	2021-07-29	2021-10-26	29,873	Human	Asia / Philippines	Vicente Sot
hCoV-19/Philippines/PH-PGC-66927/2021	Original	EPI_ISL_5551291	2021-08-01	2021-10-26	29,872	Human	Asia / Philippines	Corazon Lo
hCoV-19/Philippines/PH-PGC-66921/2021	Original	EPI_ISL_5551331	2021-08-03	2021-10-26	29,849	Human	Asia / Philippines	Corazon Lo
hCoV-19/Philippines/PH-PGC-66908/2021	Original	EPI_ISL_5551345	2021-07-31	2021-10-26	29,852	Human	Asia / Philippines	Corazon Lo
hCoV-19/Philippines/PH-PGC-66945/2021	Original	EPI_ISL_5551379	2021-08-04	2021-10-26	29,886	Human	Asia / Philippines	Corazon Lo

Stretches of NNNs (1.45% of overall sequence). Gap of 18 nucleotides when compared to the reference WIV04 sequence.

Total: 18,804 viruses

<< < 1 2 3 4 5 > >>

Charts | EPI_SET | Select | Analysis | Download

Nextclade Quality Control

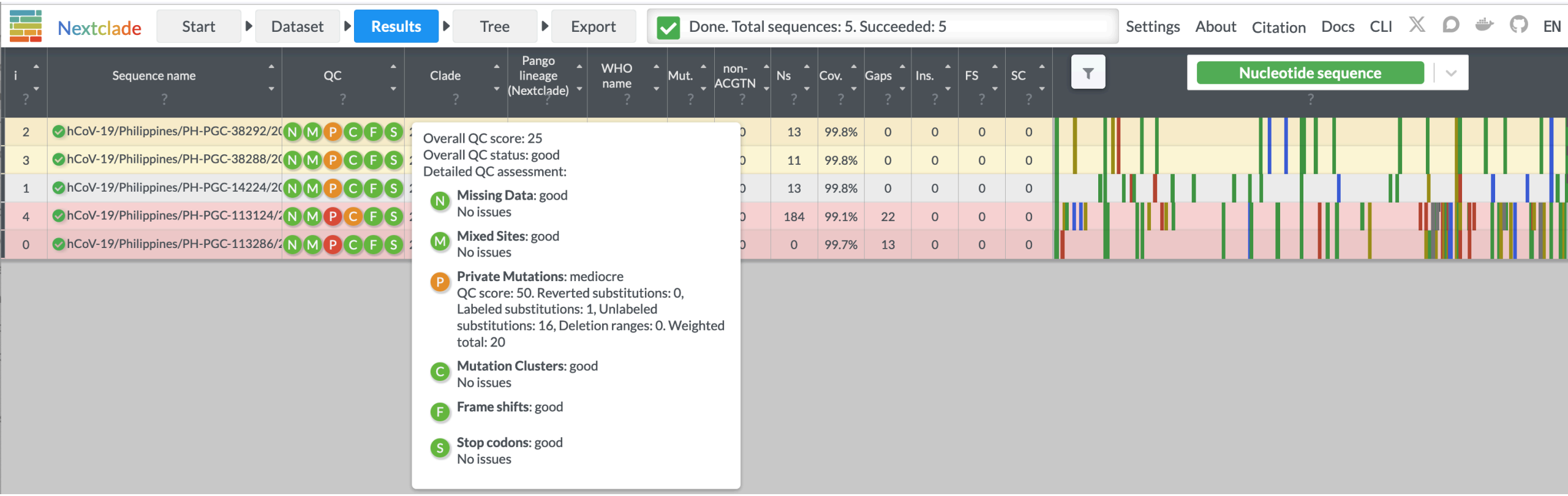
CENTRE FOR PATHOGEN GENOMICS · Doherty Institute · The University of Melbourne · The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Nextclade | Start | Dataset | **Results** | Tree | Export | Done. Total sequences: 5. Succeeded: 5 | Settings | About | Citation | Docs | CLI | EN

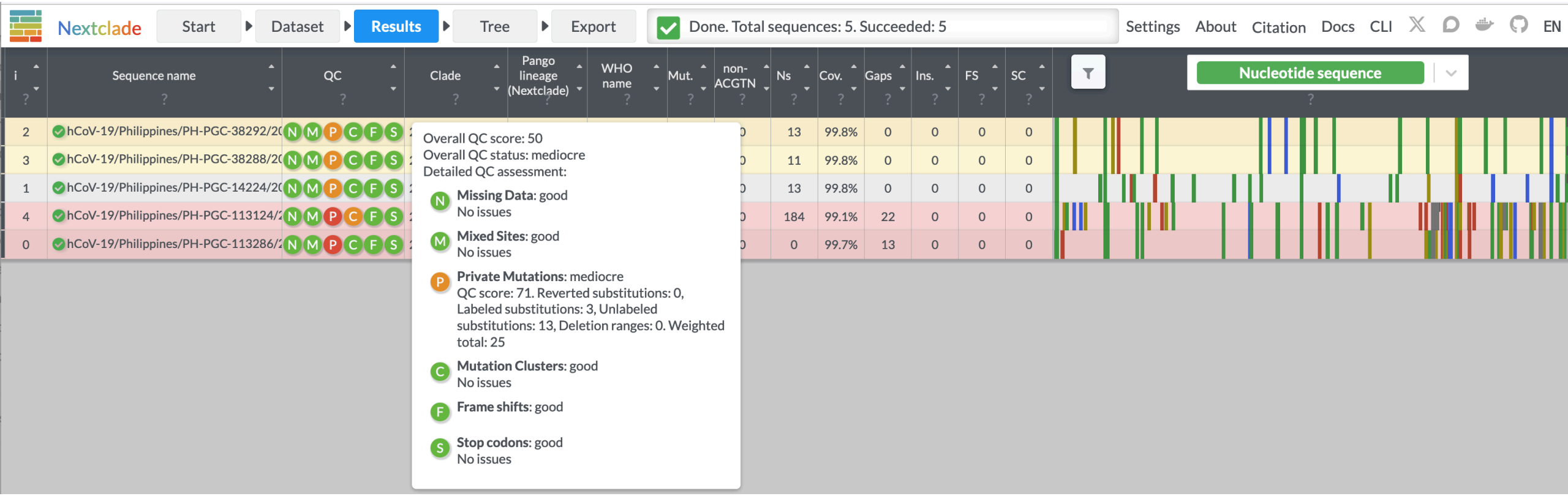
i	Sequence name	QC	Clade	Pango lineage (Nextclade)	WHO name	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC	Nucleotide sequence
2	hCoV-19/Philippines/PH-PGC-38292/20	N M P C F S	20B	B.1.1.28		25	0	13	99.8%	0	0	0	0	
3	hCoV-19/Philippines/PH-PGC-38288/20	N M P C F S	20B	B.1.1.28		25	0	11	99.8%	0	0	0	0	
1	hCoV-19/Philippines/PH-PGC-14224/20	N M P C F S	20B	B.1.1.263		26	0	13	99.8%	0	0	0	0	
4	hCoV-19/Philippines/PH-PGC-113124/2	N M P C F S	21I	B.1.617.2	Delta	70	0	184	99.1%	22	0	0	0	
0	hCoV-19/Philippines/PH-PGC-113286/2	N M P C F S	21I	B.1.617.2	Delta	52	0	0	99.7%	13	0	0	0	

Nextclade Quality Control



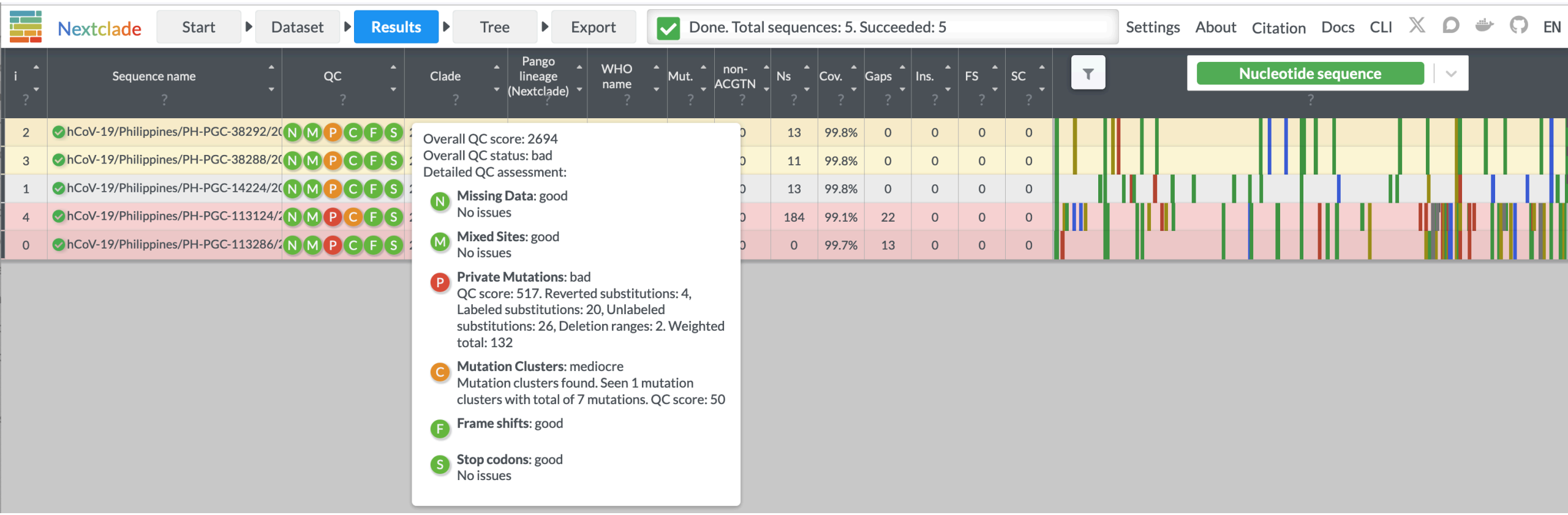
Example of a good 😍 overall QC status

Nextclade Quality Control



Example of a mediocre 🤷 overall QC status

Nextclade Quality Control



Example of a bad 👎 overall QC status

Nextclade Quality Scores

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

- **Individual Scores**
  - **Missing data (N):** If your sequence misses more than 3000 sites (N characters), it will be flagged as bad
  - **Mixed sites (M):** Ambiguous nucleotides are often indicative of contamination
  - **Private mutations (P):** As a by-product of the phylogenetic placement method of assigning lineages, Nextclade identifies the mutations, called “private mutations”, that differ between the query sequence and the nearest neighbor sequence.
  - **Mutation clusters (C)**: To be more sensitive for quality problems in a narrow area of a genome, the mutation cluster rule counts the number of private within all possible 100-nucleotide windows
  - **Stop codons (S)**: Premature stops
  - **Frame shifts (F)**: Wrong grouping of codons

- **Overall QC score**
  - multiple mildly concerning scores don’t result in a bad overall score, but a single bad score guarantees a bad overall score.
  - lower value means better quality, higher value means worse quality

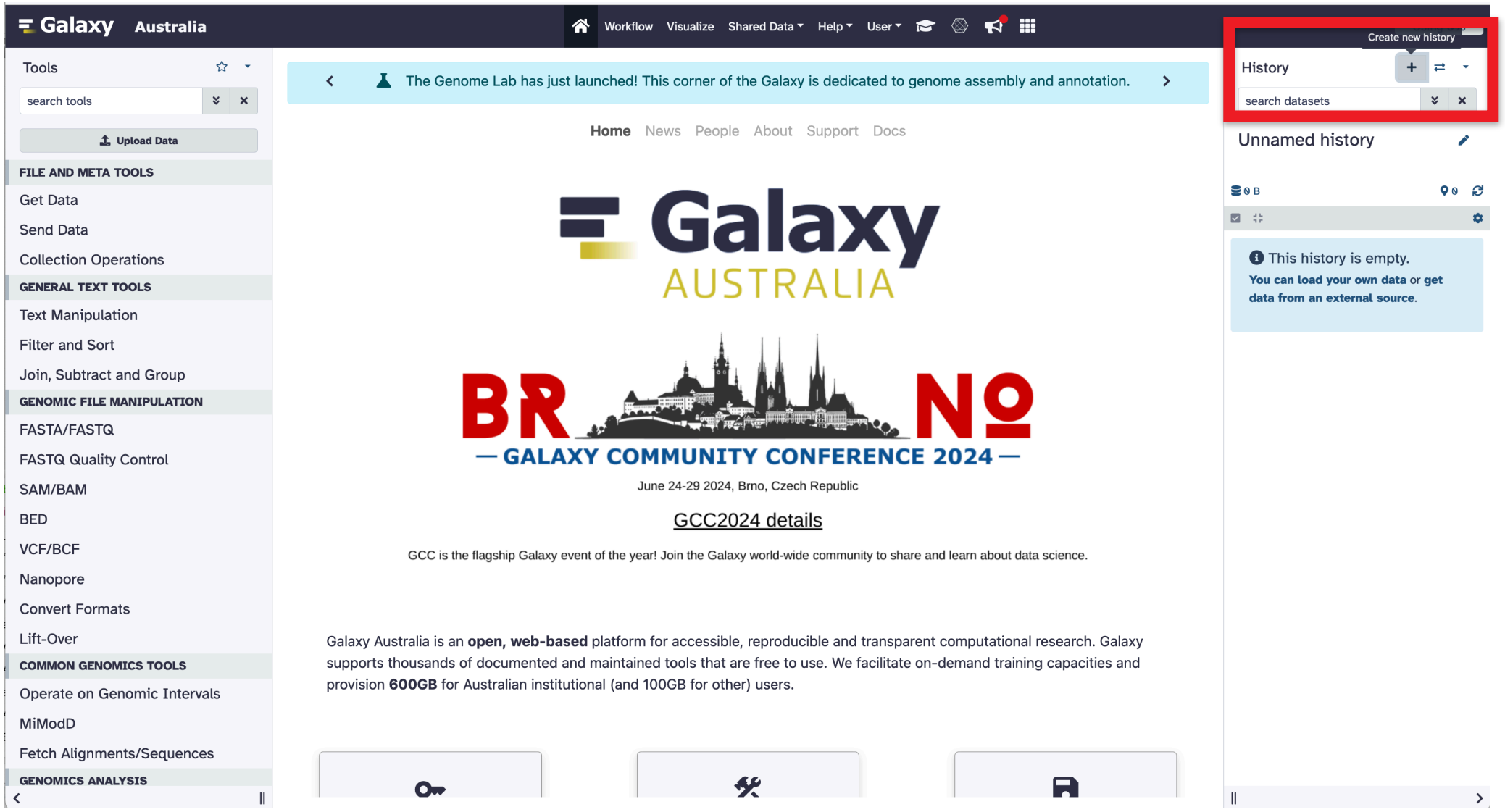
Hands-On

Objectives:

1. From a set of consensus assemblies of SARS-CoV-2 genome, we will trim nucleotide positions based on a gappyness threshold to minimize gaps and ambiguous bases

2. Assess each samples using Nextclade quality control to look at individual scores for each assembly metric

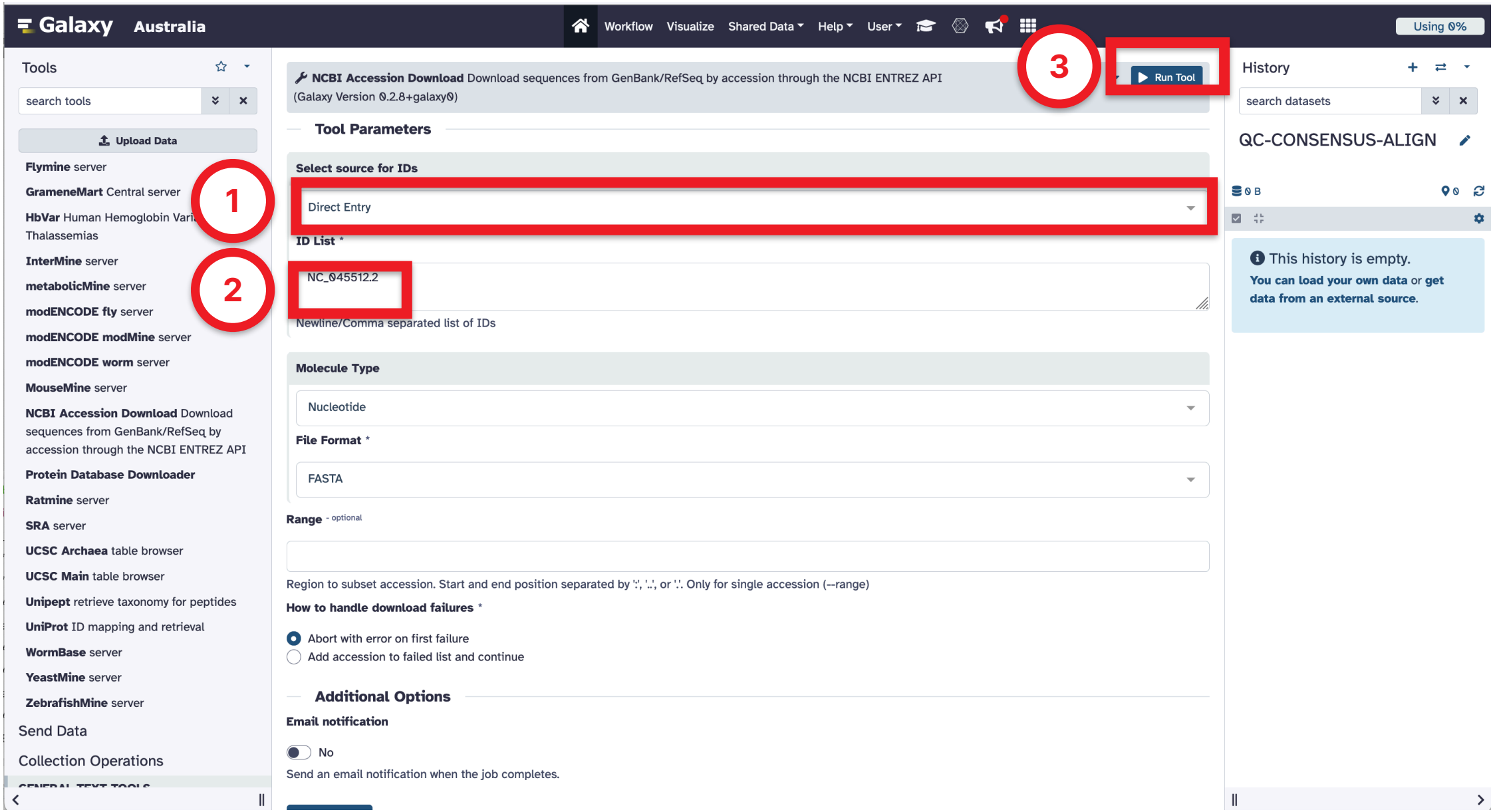
Create new history



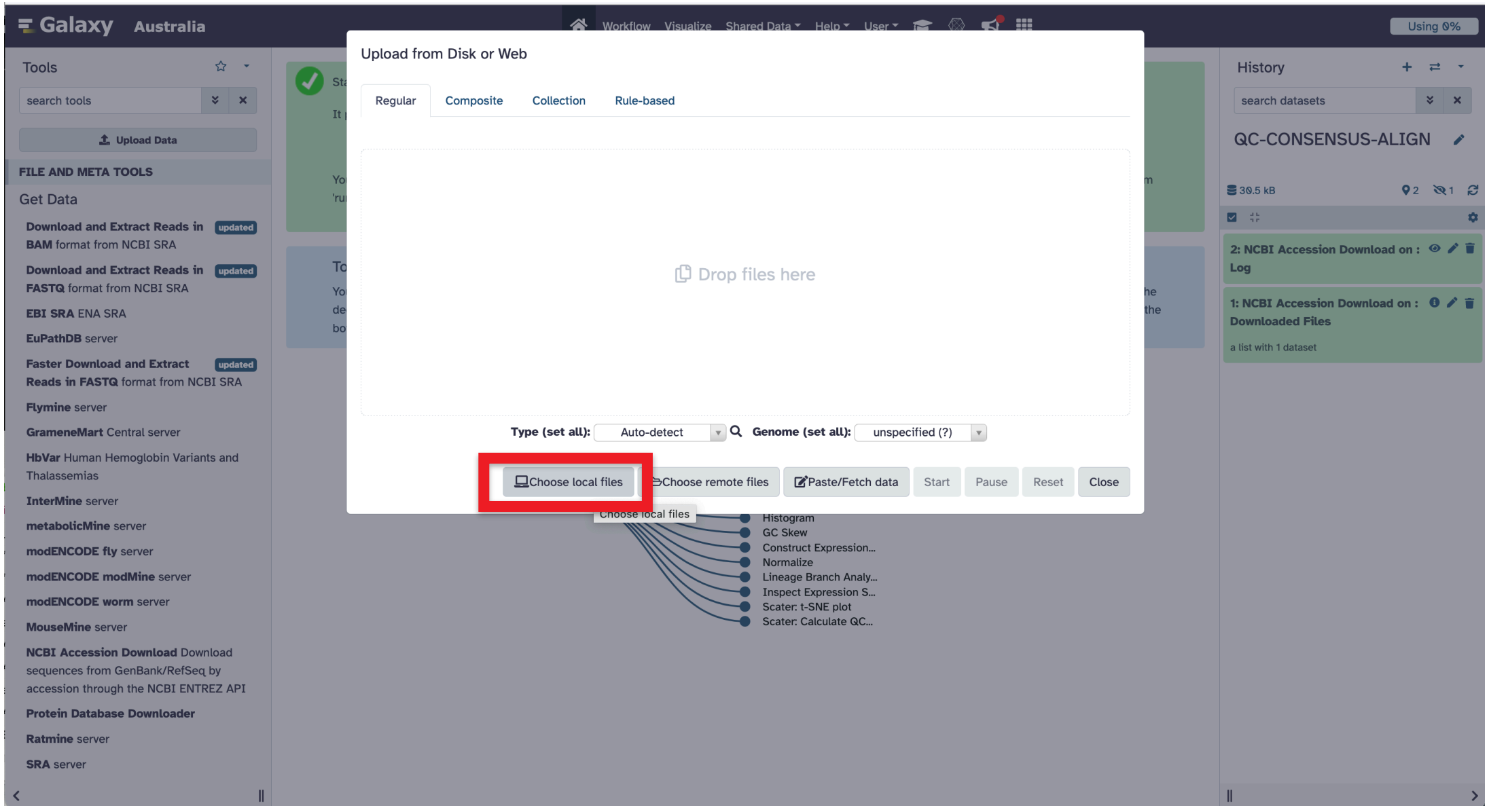
Name your history

NCBI accession download

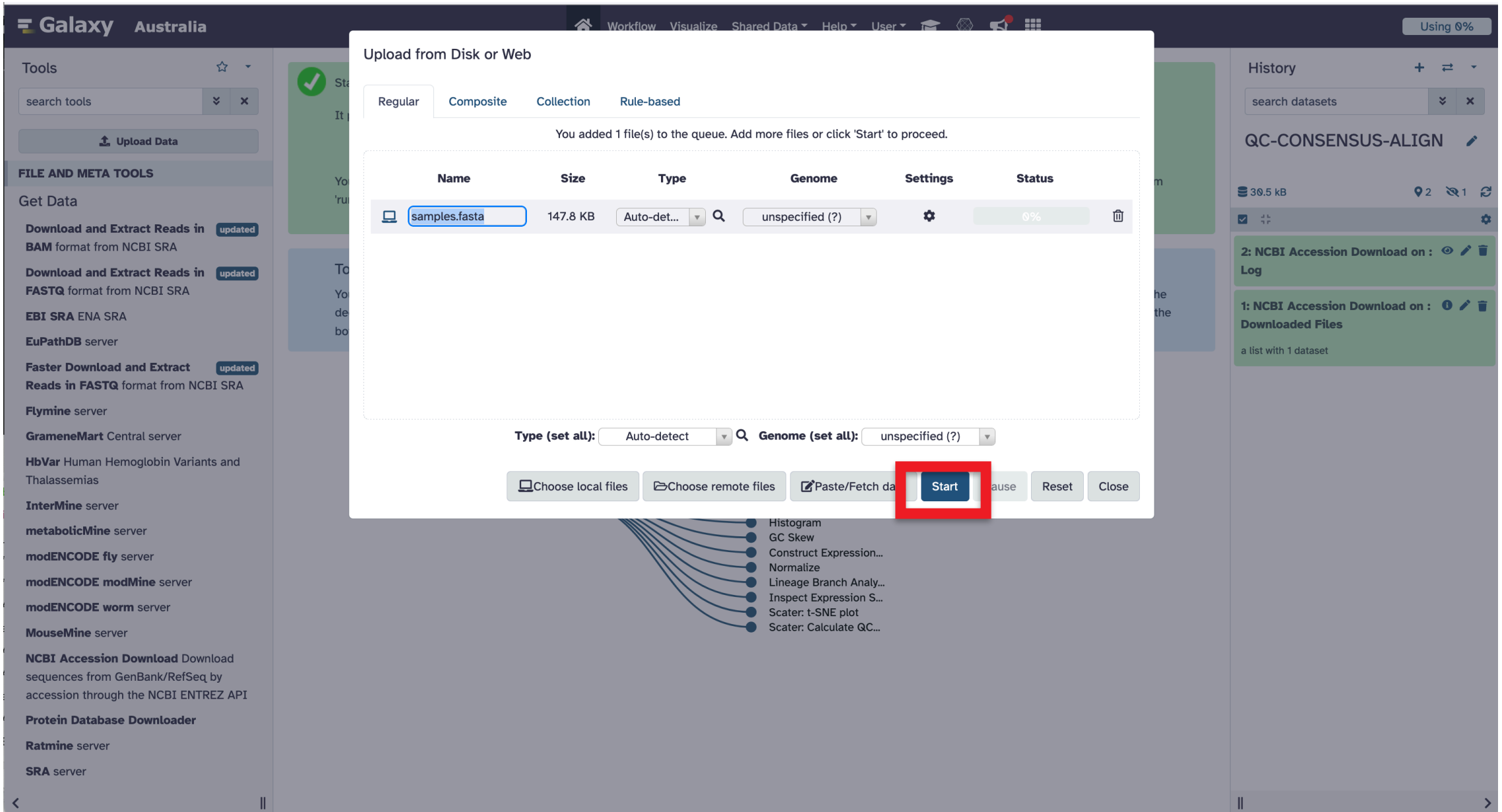
Add NC_045512.2 and run



Choose local files

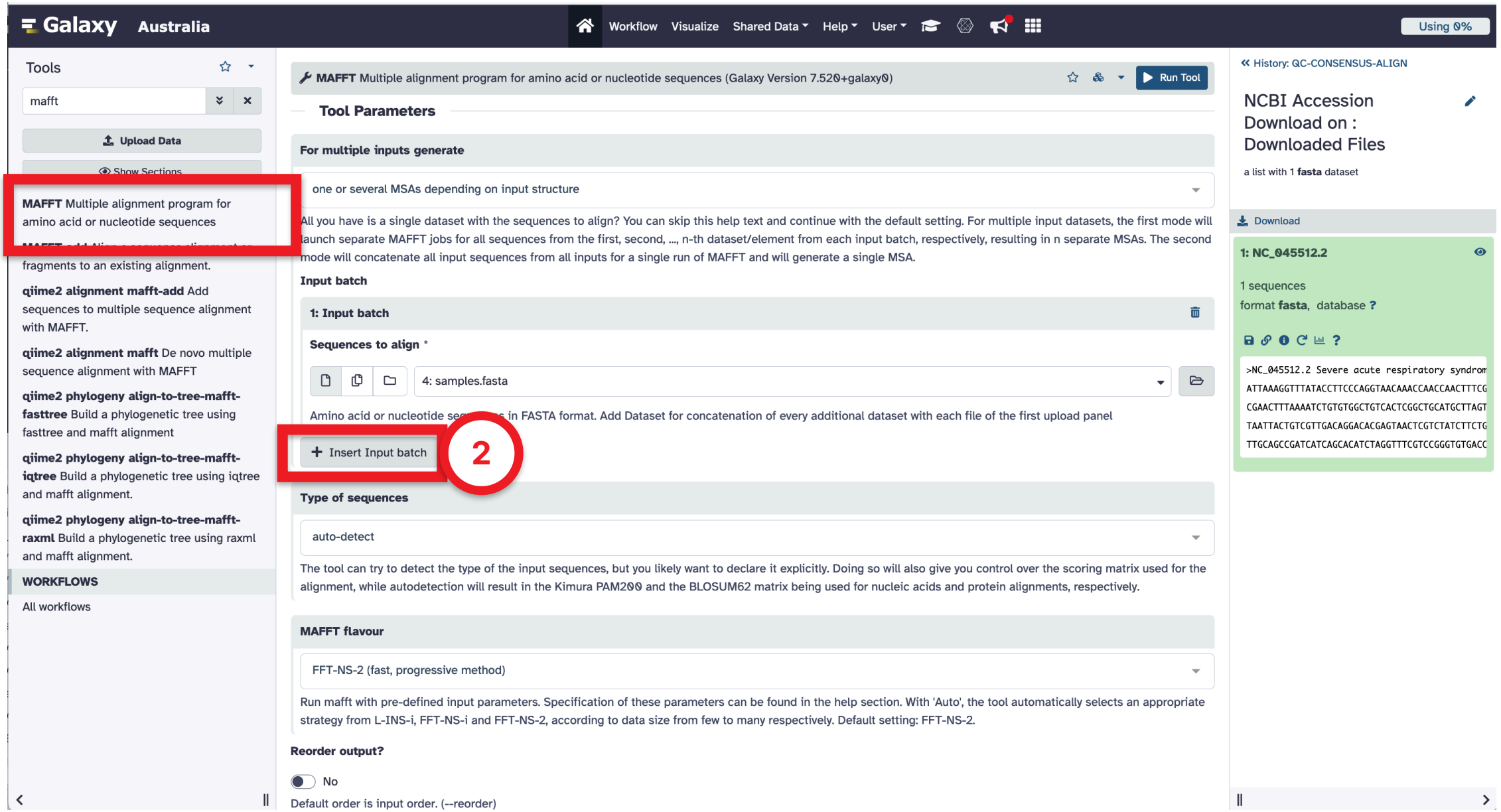


Select samples.fasta

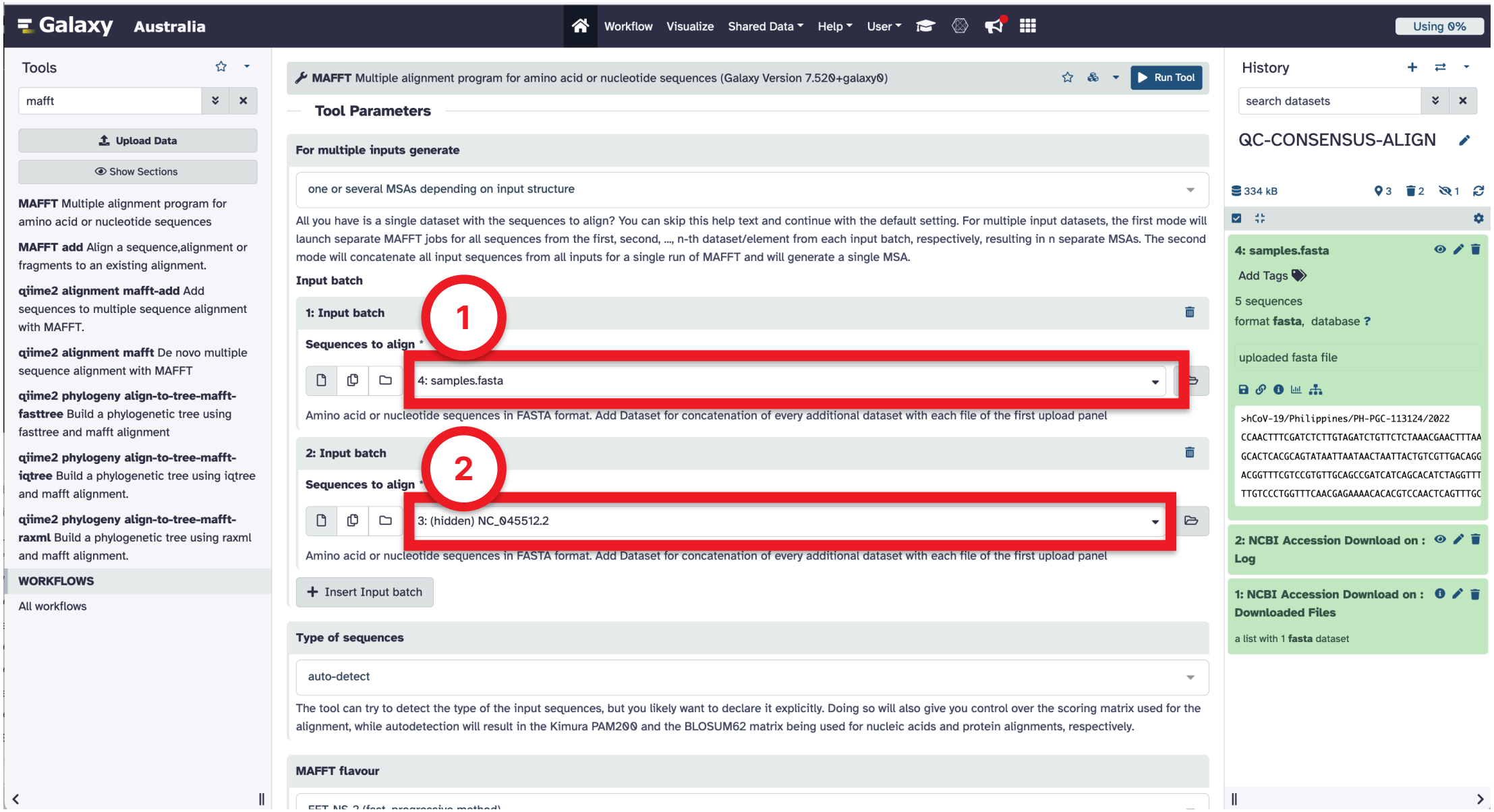


Check history for samples.fasta

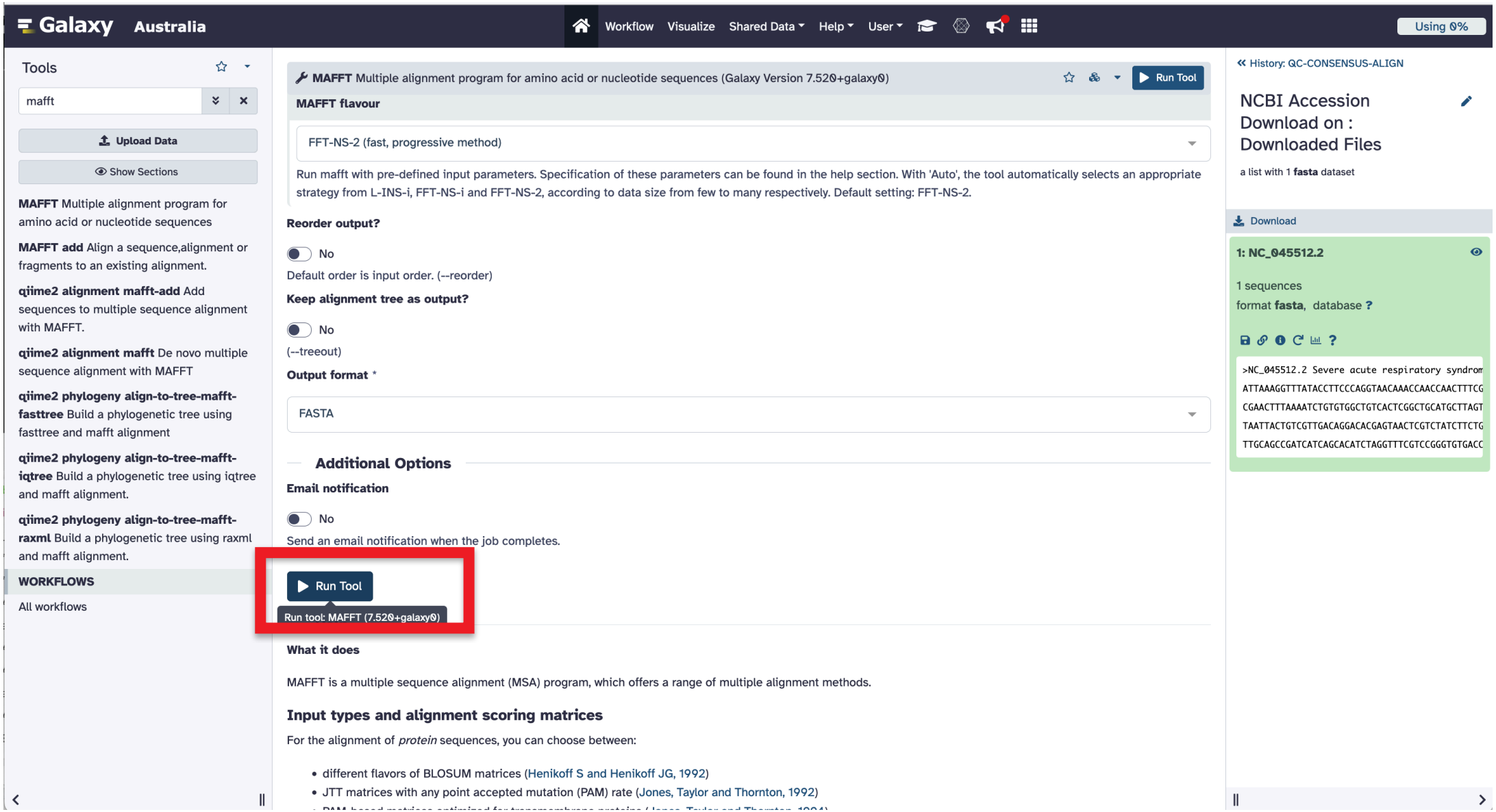
Search mafft in tools



Insert


Galaxy Australia
Workflow
Visualize
Shared Data
Help
User
Using 0%
Tools
mafft
Upload Data
Show Sections
MAFFT Multiple alignment program for amino acid or nucleotide sequences
MAFFT add Align a sequence,alignment or fragments to an existing alignment.
qiime2 alignment mafft-add Add sequences to multiple sequence alignment with MAFFT.
qiime2 alignment mafft De novo multiple sequence alignment with MAFFT
qiime2 phylogeny align-to-tree-mafft-fasttree Build a phylogenetic tree using fasttree and mafft alignment
qiime2 phylogeny align-to-tree-mafft-iqtree Build a phylogenetic tree using iqtree and mafft alignment.
qiime2 phylogeny align-to-tree-mafft-raxml Build a phylogenetic tree using raxml and mafft alignment.
WORKFLOWS
All workflows
MAFFT Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)
Run Tool
Tool Parameters
For multiple inputs generate
one or several MSAs depending on input structure
All you have is a single dataset with the sequences to align? You can skip this help text and continue with the default setting. For multiple input datasets, the first mode will launch separate MAFFT jobs for all sequences from the first, second, ..., n-th dataset/element from each input batch, respectively, resulting in n separate MSAs. The second mode will concatenate all input sequences from all inputs for a single run of MAFFT and will generate a single MSA.
Input batch
1: Input batch
1
Sequences to align *
4: samples.fasta
Amino acid or nucleotide sequences in FASTA format. Add Dataset for concatenation of every additional dataset with each file of the first upload panel
2: Input batch
2
Sequences to align *
3: (hidden) NC_045512.2
Amino acid or nucleotide sequences in FASTA format. Add Dataset for concatenation of every additional dataset with each file of the first upload panel
Insert Input batch
Type of sequences
auto-detect
The tool can try to detect the type of the input sequences, but you likely want to declare it explicitly. Doing so will also give you control over the scoring matrix used for the alignment, while autodetection will result in the Kimura PAM200 and the BLOSUM62 matrix being used for nucleic acids and protein alignments, respectively.
MAFFT flavour
History
search datasets
QC-CONSENSUS-ALIGN
334 kB
4: samples.fasta
Add Tags
5 sequences
format fasta, database ?
uploaded fasta file
>hCoV-19/Philippines/PH-PGC-113124/2022
CCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
GCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
ACGGTTTCGTCCGTGTTGCAGCCGATCATCAGCACATCTAGGTTT
TTGTCCCTGGTTTCAACGAGAAAACACACGTCCAACTCAGTTTGC
2: NCBI Accession Download on : Log
1: NCBI Accession Download on : Downloaded Files
a list with 1 fasta dataset

Run mafft


Galaxy Australia
Workflow
Visualize
Shared Data
Help
User
Using 0%
Tools
mafft
Upload Data
Show Sections
MAFFT Multiple alignment program for amino acid or nucleotide sequences
MAFFT add Align a sequence,alignment or fragments to an existing alignment.
qiime2 alignment mafft-add Add sequences to multiple sequence alignment with MAFFT.
qiime2 alignment mafft De novo multiple sequence alignment with MAFFT
qiime2 phylogeny align-to-tree-mafft-fasttree Build a phylogenetic tree using fasttree and mafft alignment
qiime2 phylogeny align-to-tree-mafft-iqtree Build a phylogenetic tree using iqtree and mafft alignment.
qiime2 phylogeny align-to-tree-mafft-raxml Build a phylogenetic tree using raxml and mafft alignment.
WORKFLOWS
All workflows
MAFFT Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)
Run Tool
MAFFT flavour
FFT-NS-2 (fast, progressive method)
Run mafft with pre-defined input parameters. Specification of these parameters can be found in the help section. With 'Auto', the tool automatically selects an appropriate strategy from L-INS-i, FFT-NS-i and FFT-NS-2, according to data size from few to many respectively. Default setting: FFT-NS-2.
Reorder output?
No
Default order is input order. (--reorder)
Keep alignment tree as output?
No
(--treeout)
Output format *
FASTA
Additional Options
Email notification
No
Send an email notification when the job completes.
Run Tool
Run tool: MAFFT (7.520+galaxy0)
What it does
MAFFT is a multiple sequence alignment (MSA) program, which offers a range of multiple alignment methods.
Input types and alignment scoring matrices
For the alignment of protein sequences, you can choose between:
different flavors of BLOSUM matrices (Henikoff S and Henikoff JG, 1992)
JTT matrices with any point accepted mutation (PAM) rate (Jones, Taylor and Thornton, 1992)
History: QC-CONSENSUS-ALIGN
NCBI Accession Download on : Downloaded Files
a list with 1 fasta dataset
Download
1: NC_045512.2
1 sequences
format fasta, database ?
>NC_045512.2 Severe acute respiratory syndrom
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCG
CGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATGCTTAGT
TAATTACTGTCGTTGACAGGACACGAGTAACTCGTCTATCTTCTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACC

Check mafft job and download MSA

Galaxy Australia | Workflow | Visualize | Shared Data | Help | User | Using 0%

Tools

mafft

Upload Data

Show Sections

MAFFT Multiple alignment program for amino acid or nucleotide sequences

MAFFT add Align a sequence,alignment or fragments to an existing alignment.

qiime2 alignment mafft-add Add sequences to multiple sequence alignment with MAFFT.

qiime2 alignment mafft De novo multiple sequence alignment with MAFFT

qiime2 phylogeny align-to-tree-mafft-fasttree Build a phylogenetic tree using fasttree and mafft alignment

qiime2 phylogeny align-to-tree-mafft-iqtree Build a phylogenetic tree using iqtree and mafft alignment.

qiime2 phylogeny align-to-tree-mafft-raxml Build a phylogenetic tree using raxml and mafft alignment.

WORKFLOWS

All workflows

Started tool **MAFFT** and successfully added 1 job to the queue.

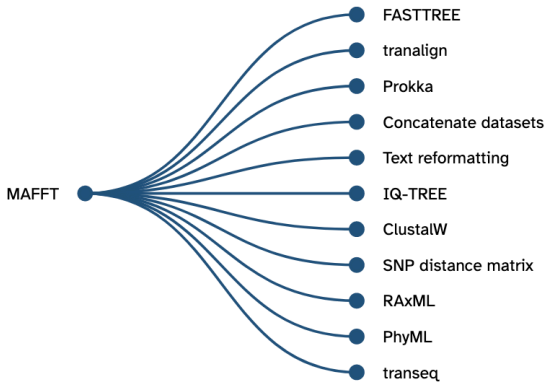
It produces this output:

- **7: MAFFT on data 3 and data 4**

You can check the status of queued jobs and view the resulting data by refreshing the History panel. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Tool recommendation

You have used rbc_mafft tool. For further analysis, you could try using the following/recommended tools. The recommended tools are shown in the decreasing order of their scores predicted using machine learning analysis on workflows. Therefore, tools at the top may be more useful than the ones at the bottom. Please click on one of the following/recommended tools to open its definition.



History

search datasets

QC-CONSENSUS-ALIGN

517 kB | 4 | 2 | 1

7: MAFFT on data 3 and data 4

Add Tags

6 sequences

format **fasta**, database **?**

nthread = 8
nthreadpair = 8
nthreadtb = 8

```
>hCoV-19/Philippines/PH-PGC-113124/2022
------------------------------------ccaactttcg
gttctctaaacgaactttaaaatctgtgtggctgtcactcggctg
cacgcagtataattaataactaattactgtcgttgacaggacacg
ttctgcaggctgcttacggtttcgtccgtgttgcagccgatcatc
```

4: samples.fasta

Add Tags

5 sequences

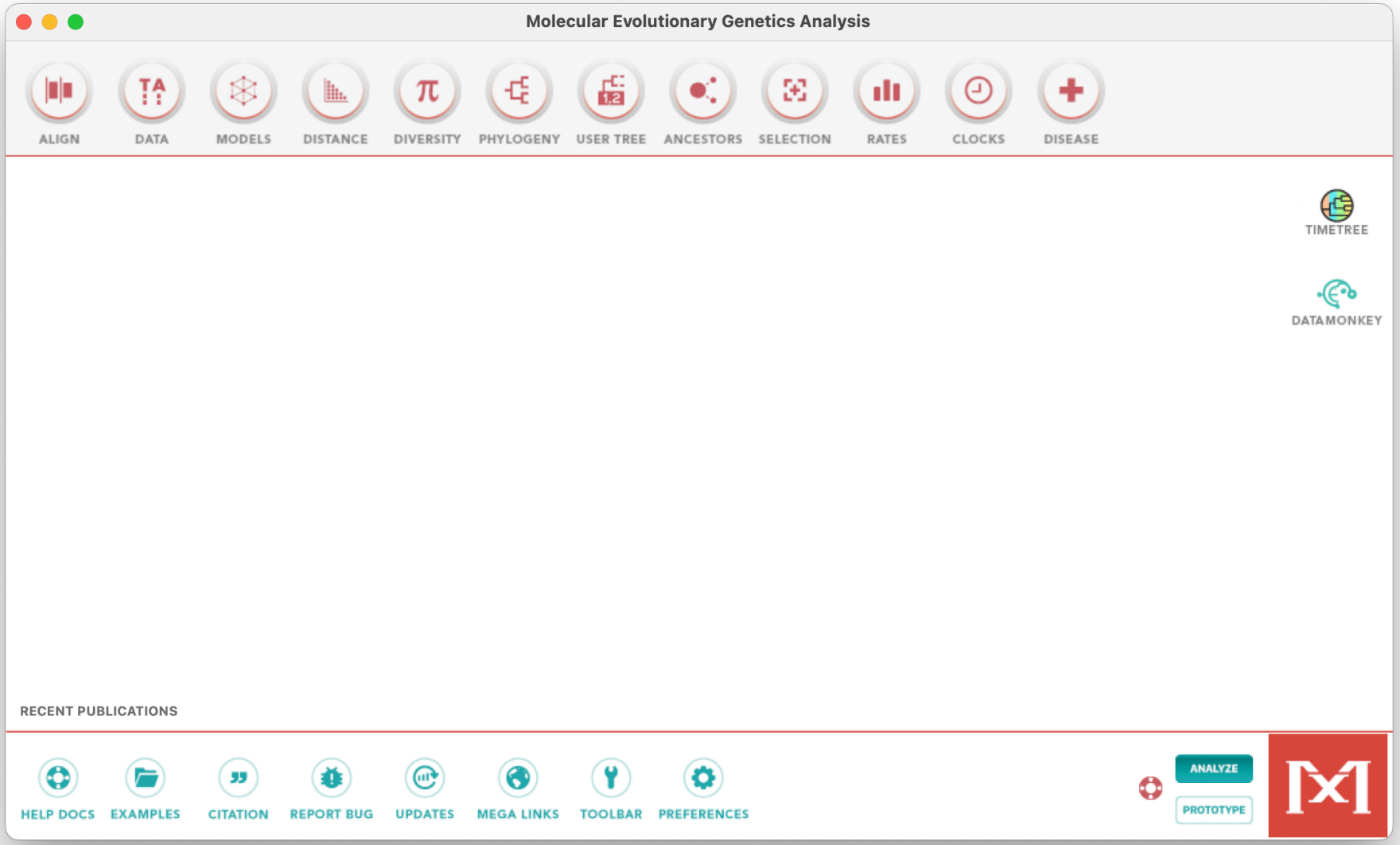
format **fasta**, database **?**

uploaded fasta file

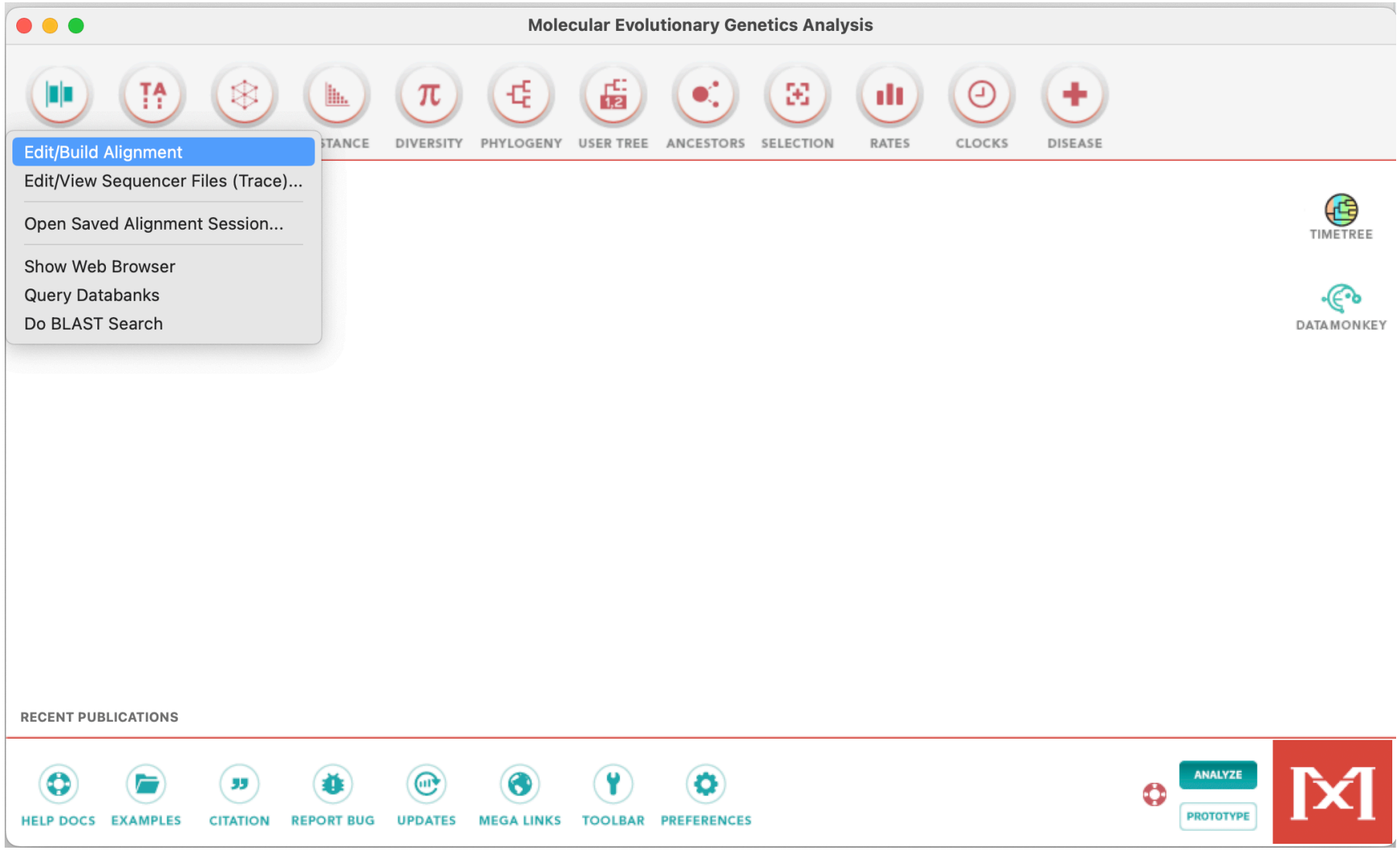
```
>hCoV-19/Philippines/PH-PGC-113124/2022
CCAACTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAA
GCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
```

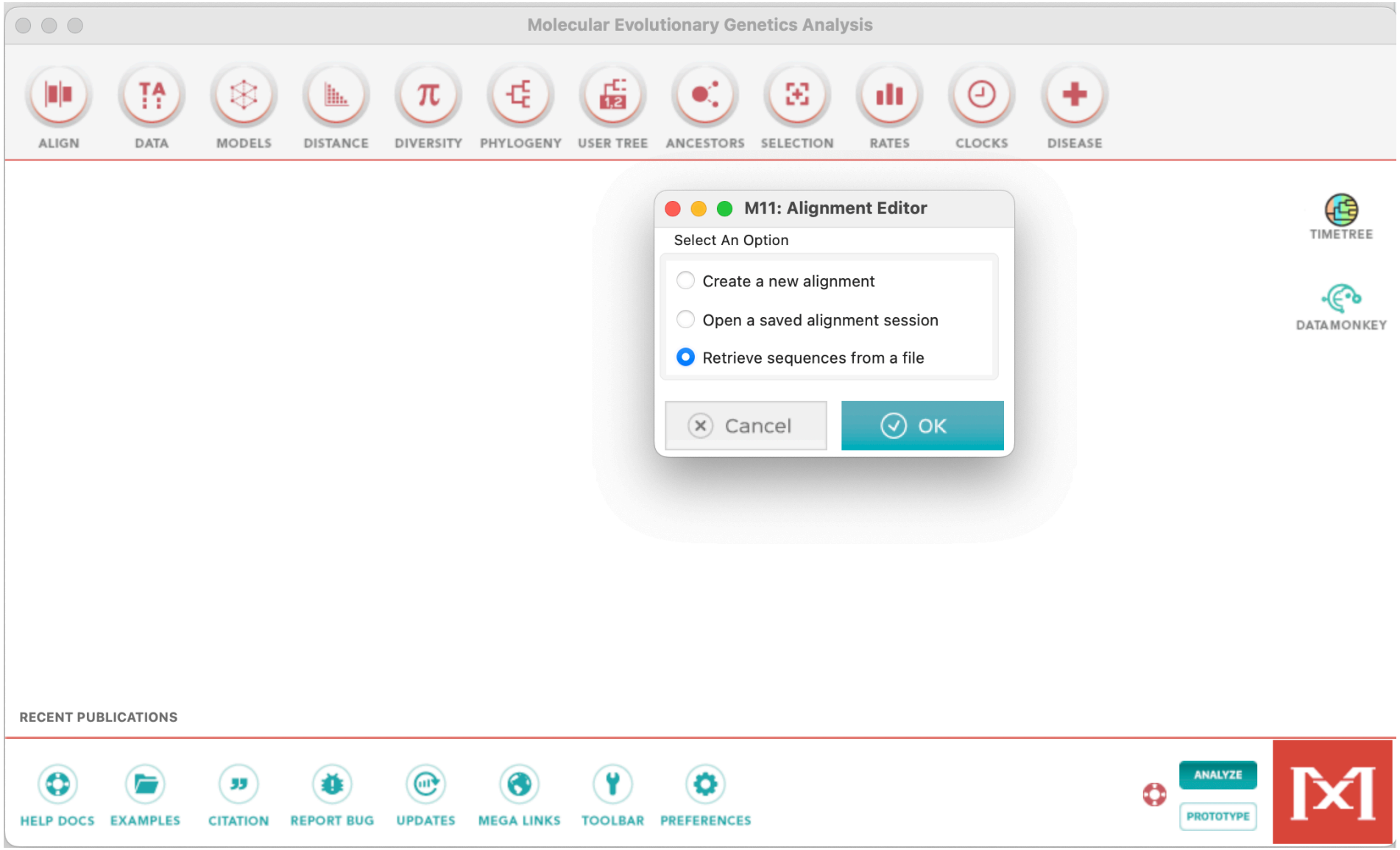
Open MEGA

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

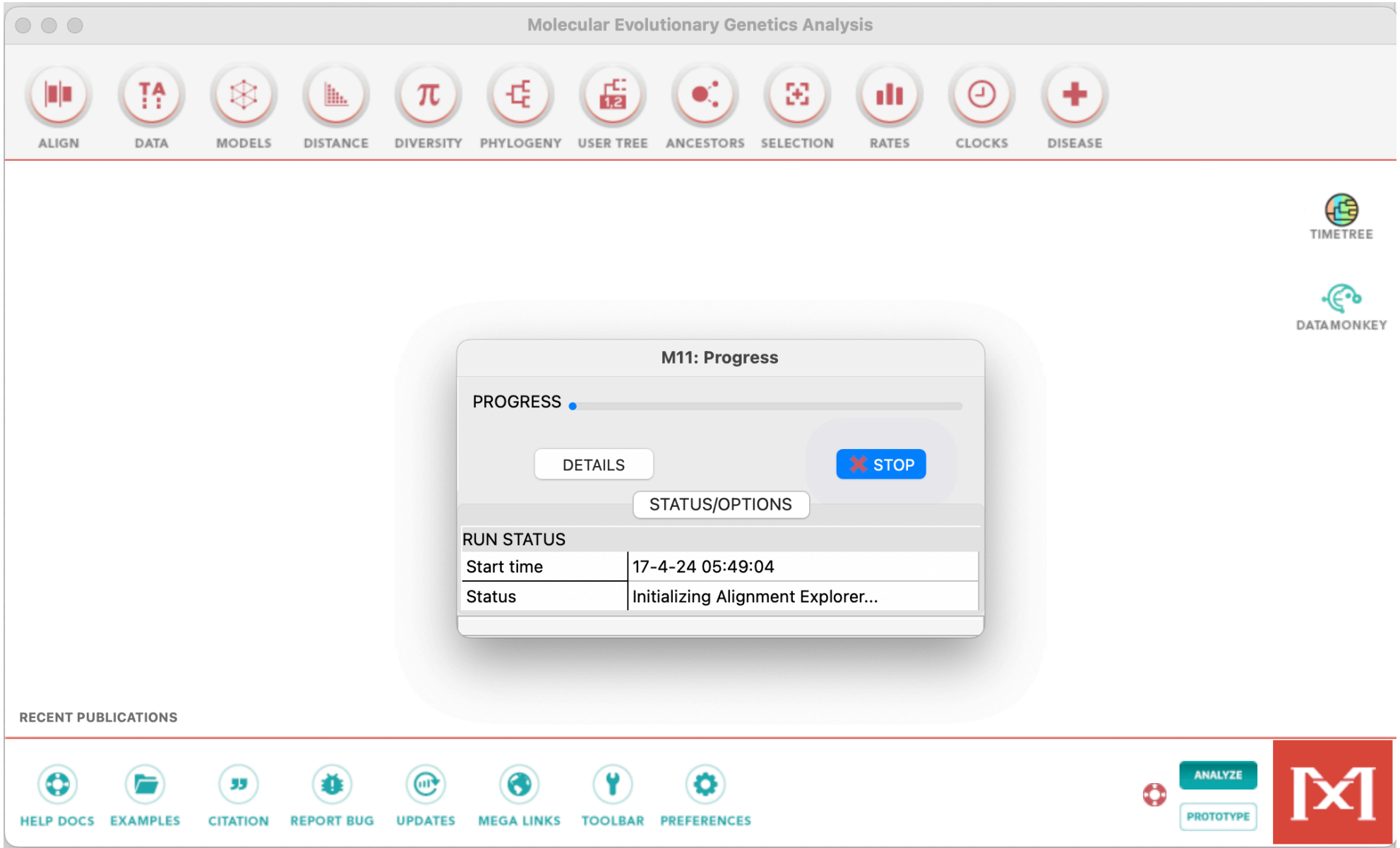


Align → Edit/Build Alignment

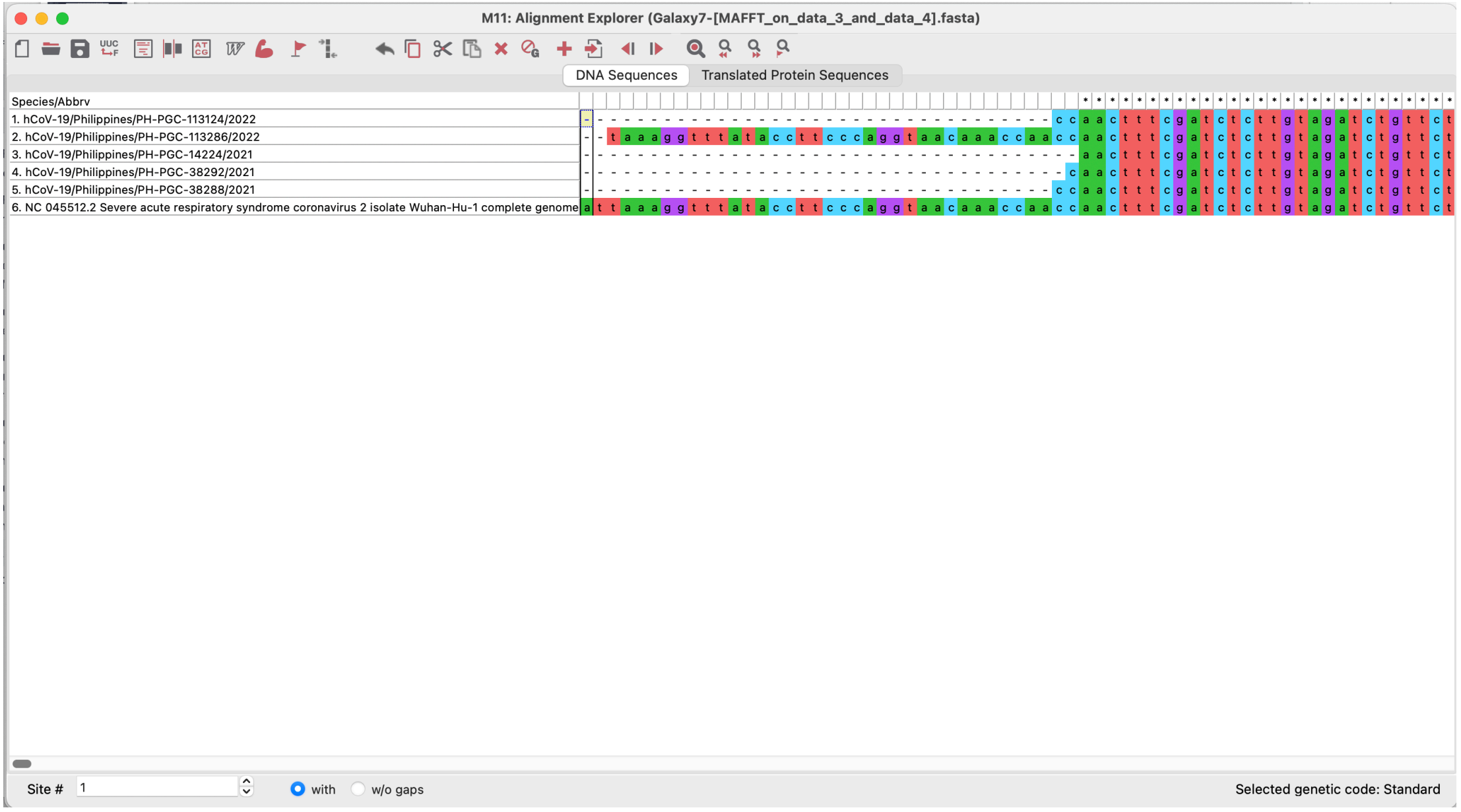



Molecular Evolutionary Genetics Analysis
ALIGN
DATA
MODELS
DISTANCE
DIVERSITY
PHYLOGENY
USER TREE
ANCESTORS
SELECTION
RATES
CLOCKS
DISEASE
M11: Alignment Editor
Select An Option
Create a new alignment
Open a saved alignment session
Retrieve sequences from a file
Cancel
OK
TIMETREE
DATAMONKEY
RECENT PUBLICATIONS
HELP DOCS
EXAMPLES
CITATION
REPORT BUG
UPDATES
MEGA LINKS
TOOLBAR
PREFERENCES
ANALYZE
PROTOTYPE

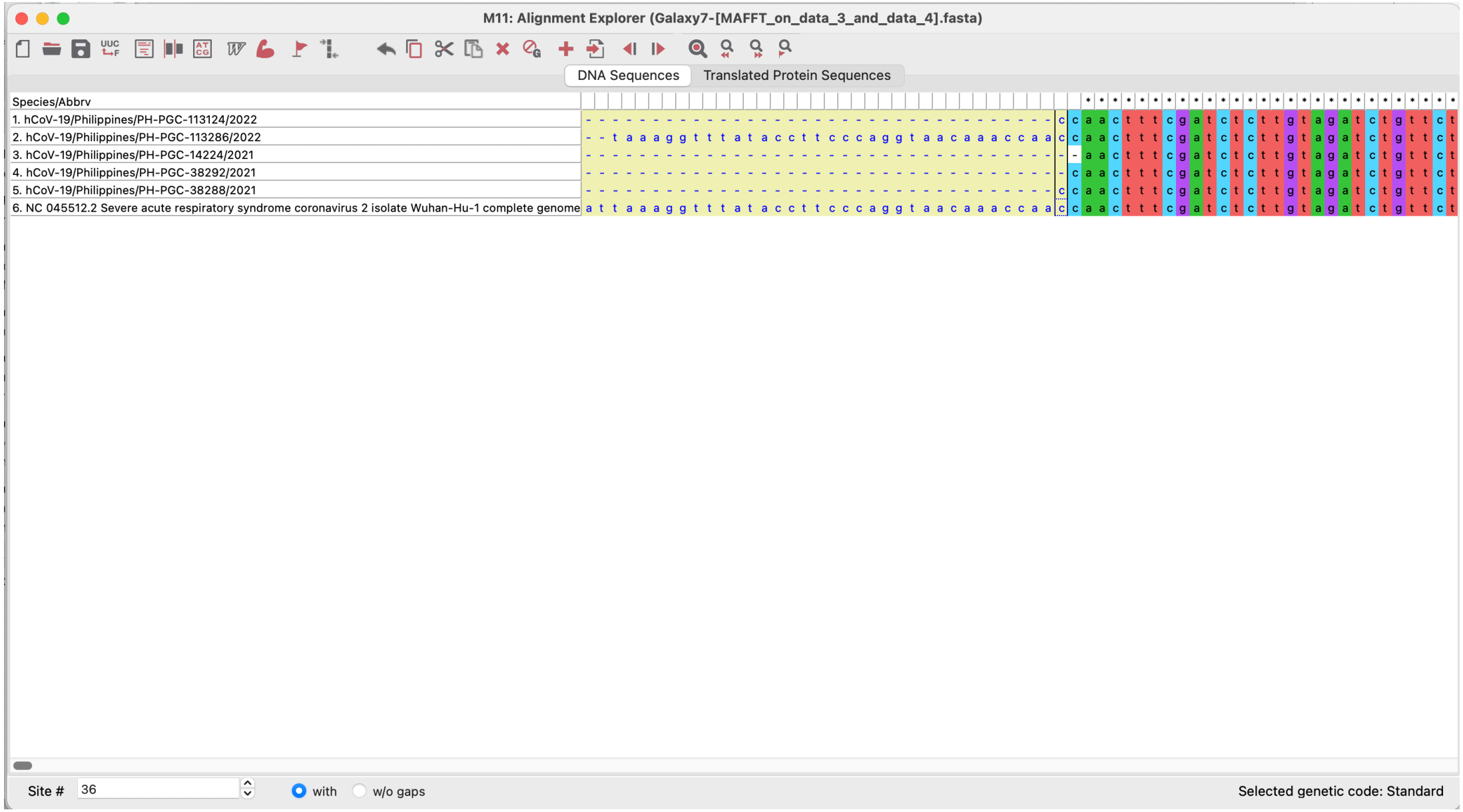
Open the downloaded MSA



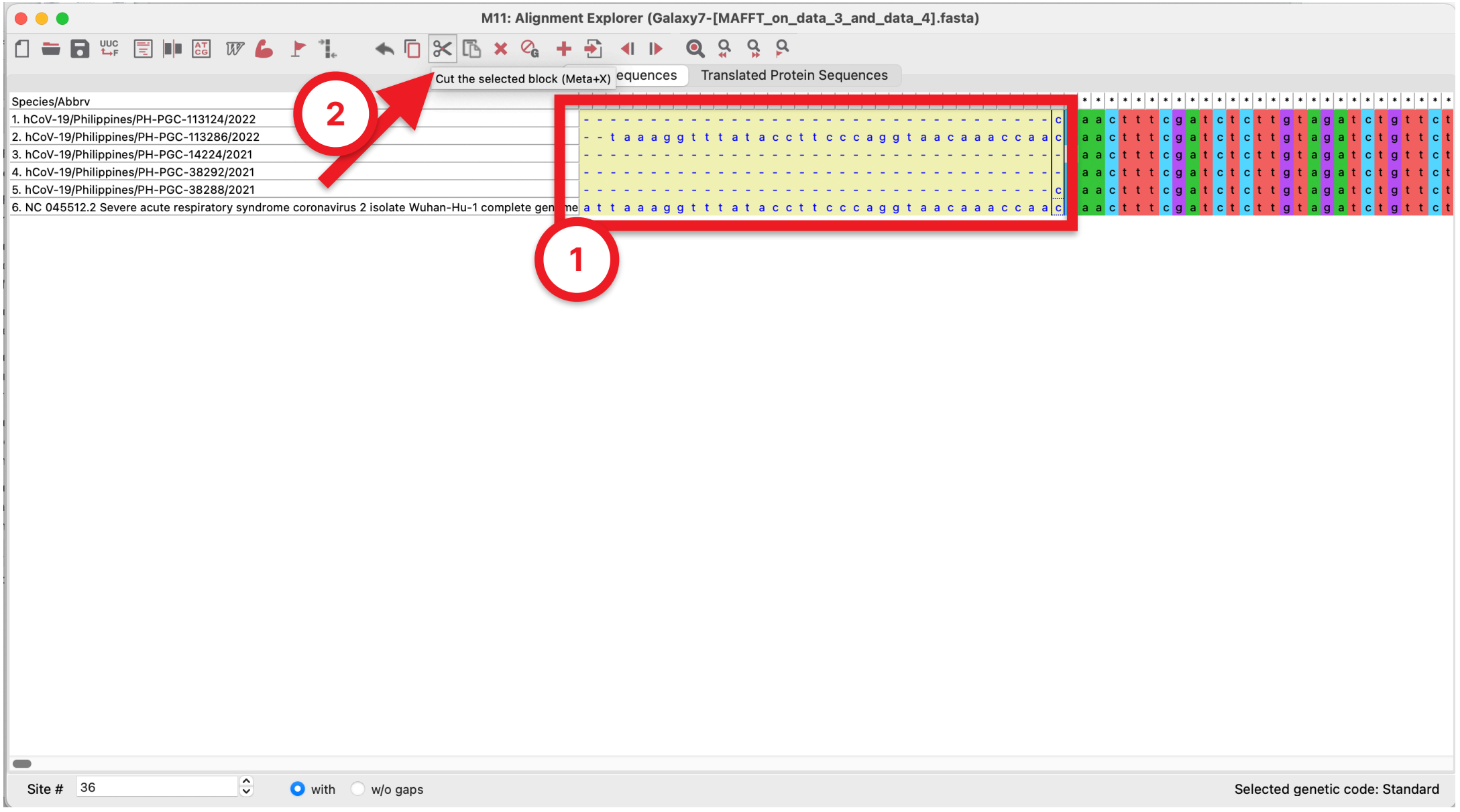
Inspect alignment



Highlight leading sequences for all samples

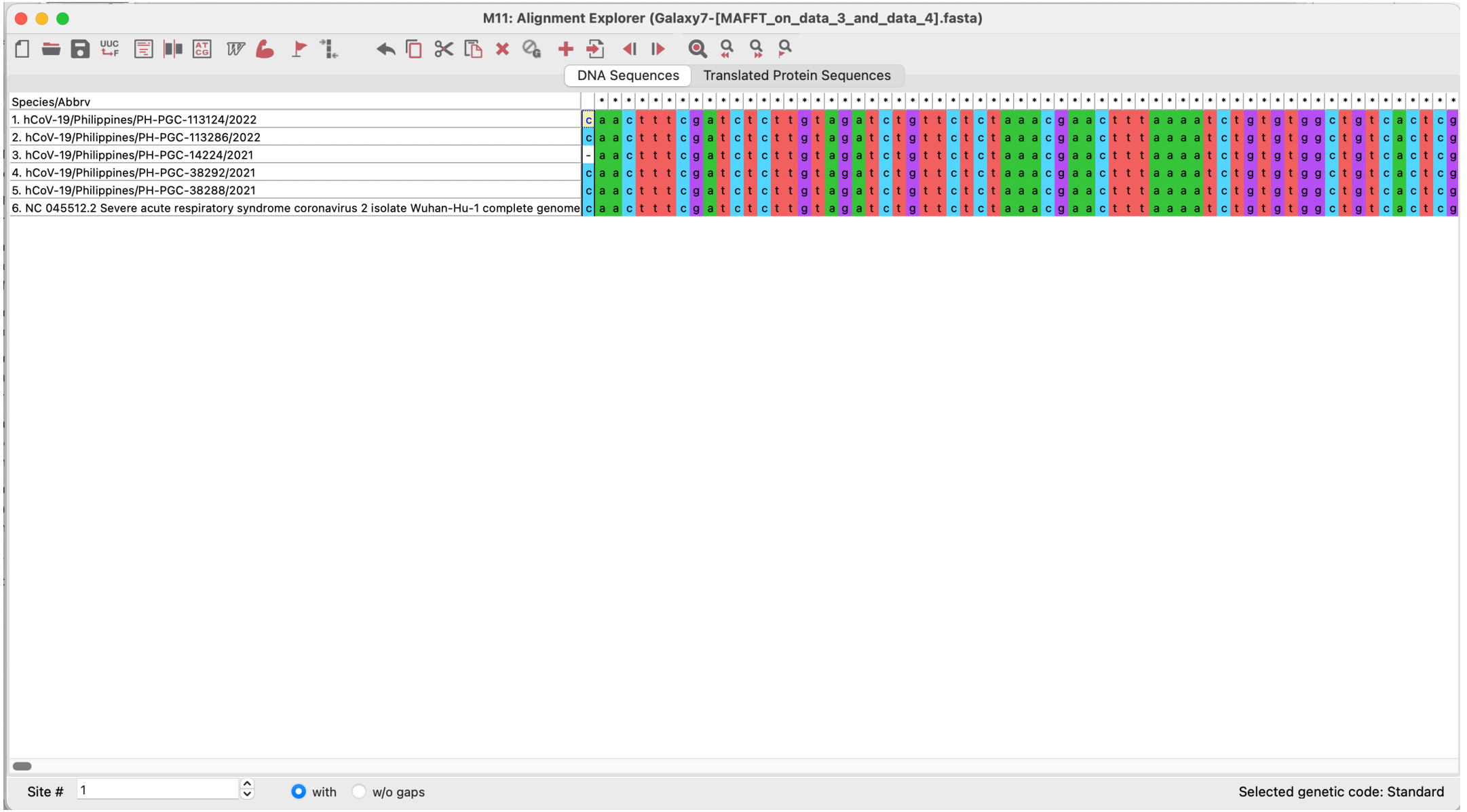


Trim the selected block

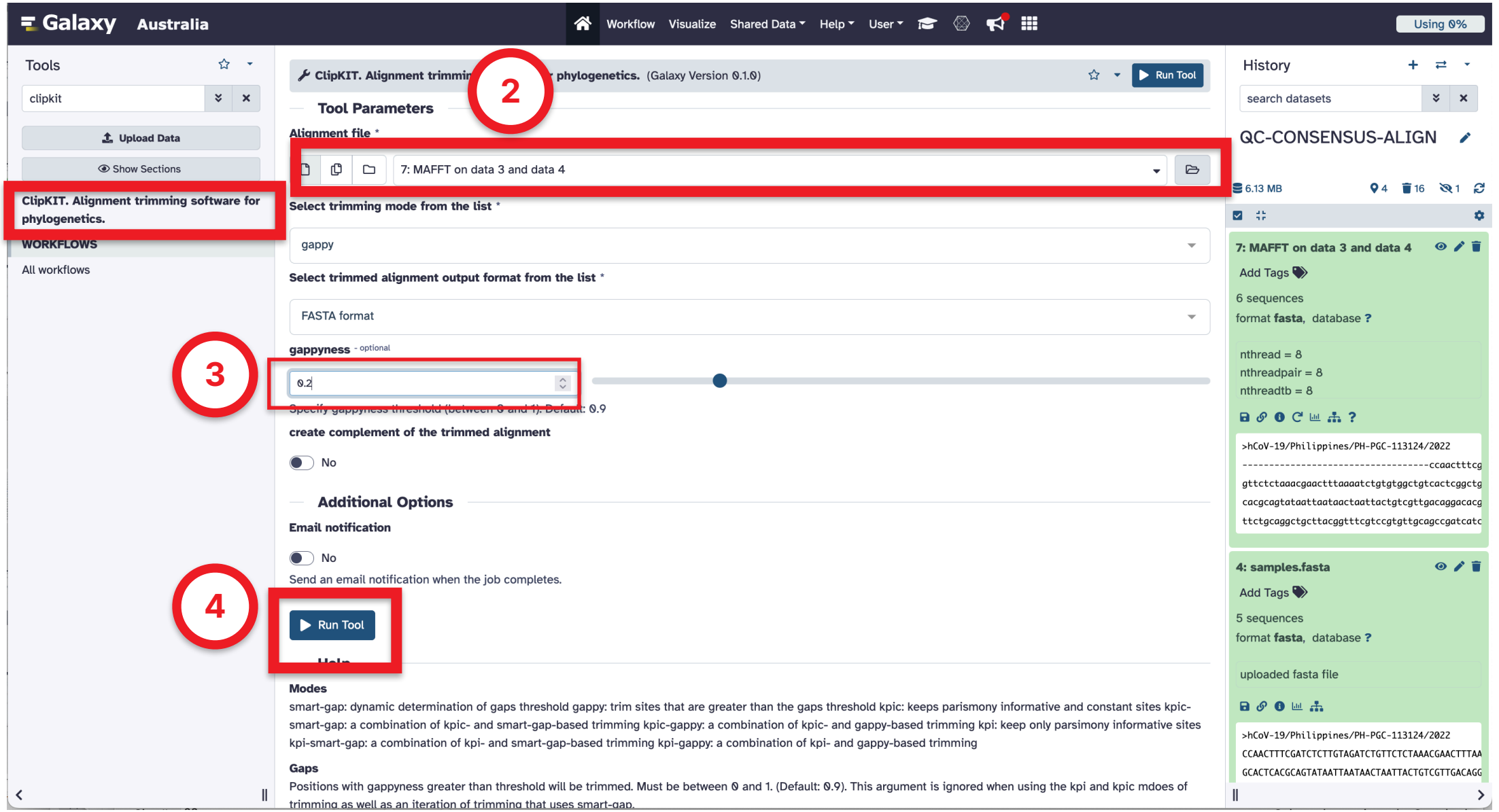


Inspect the sequence, again

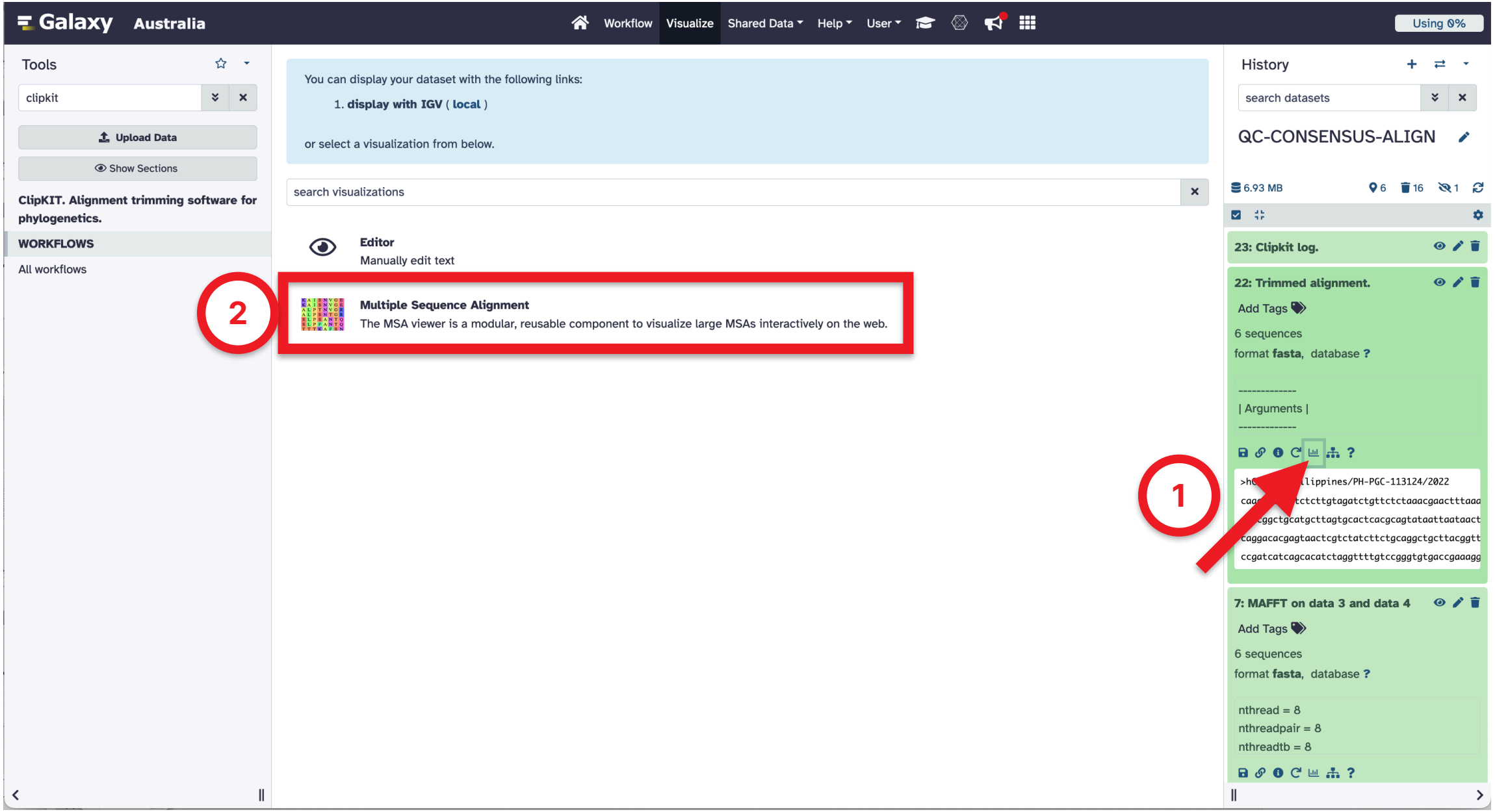
CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital



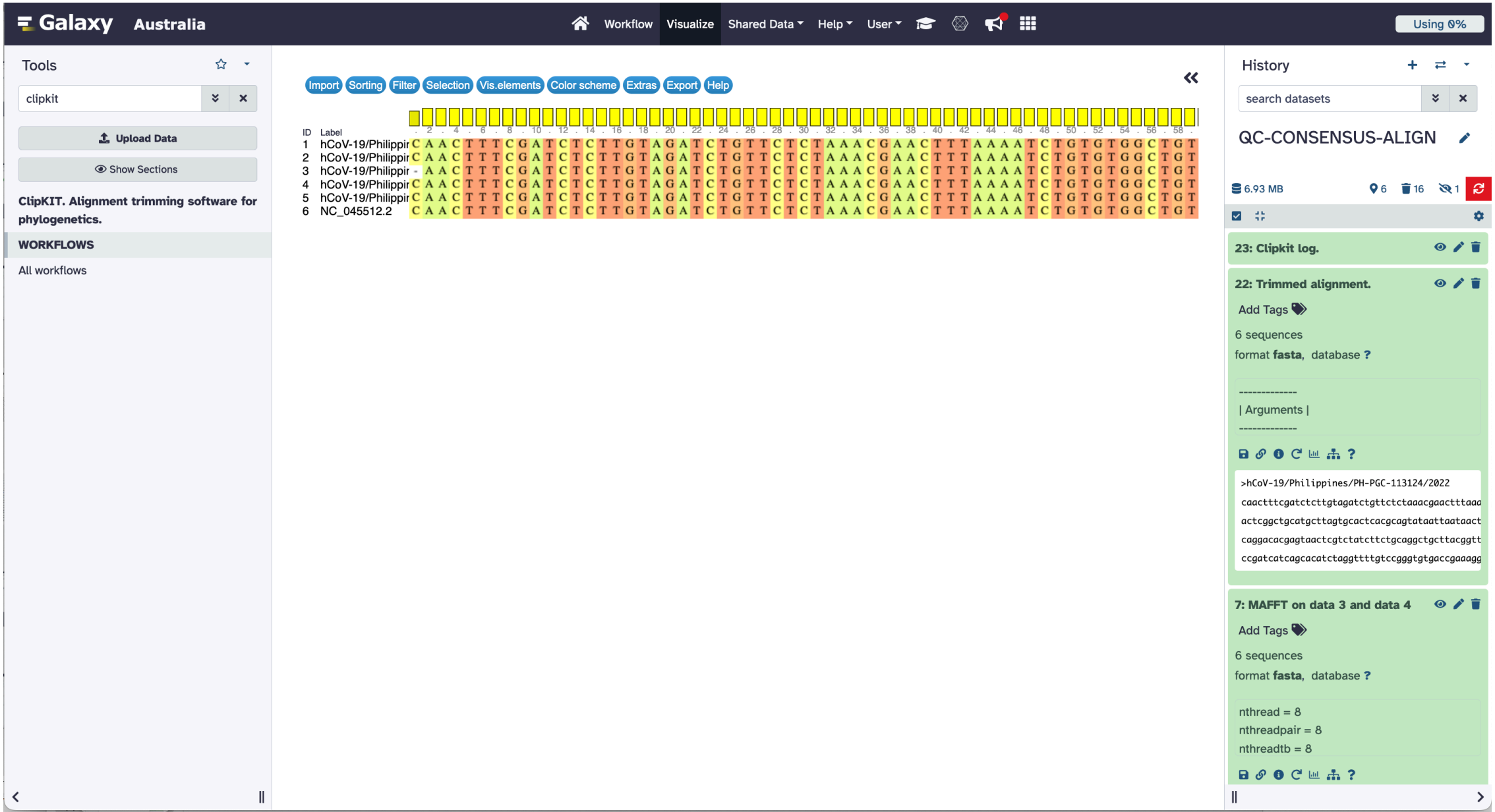
Now, we're gonna do it programmatically in Galaxy using ClipKIT



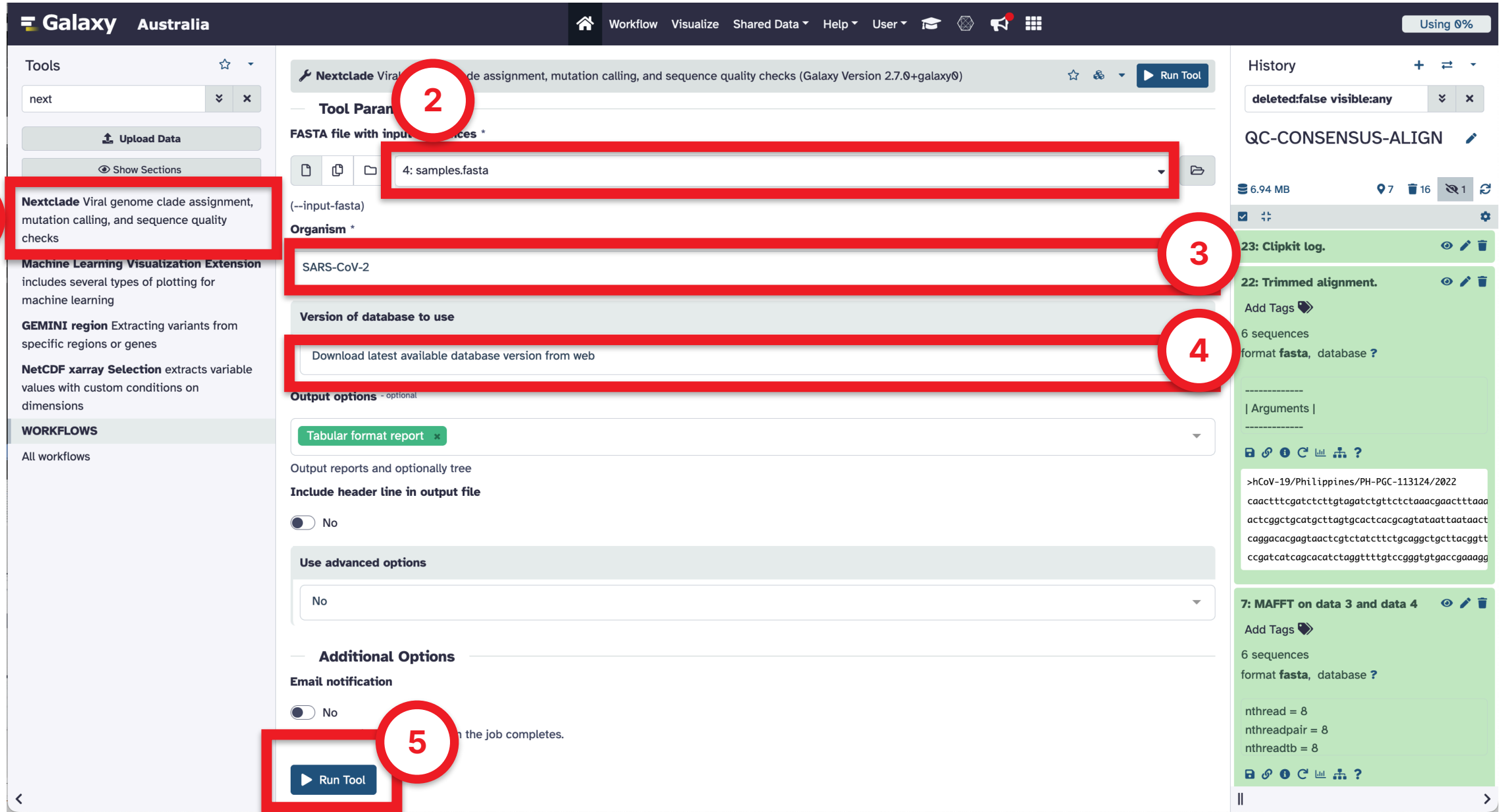
Visualize alignment



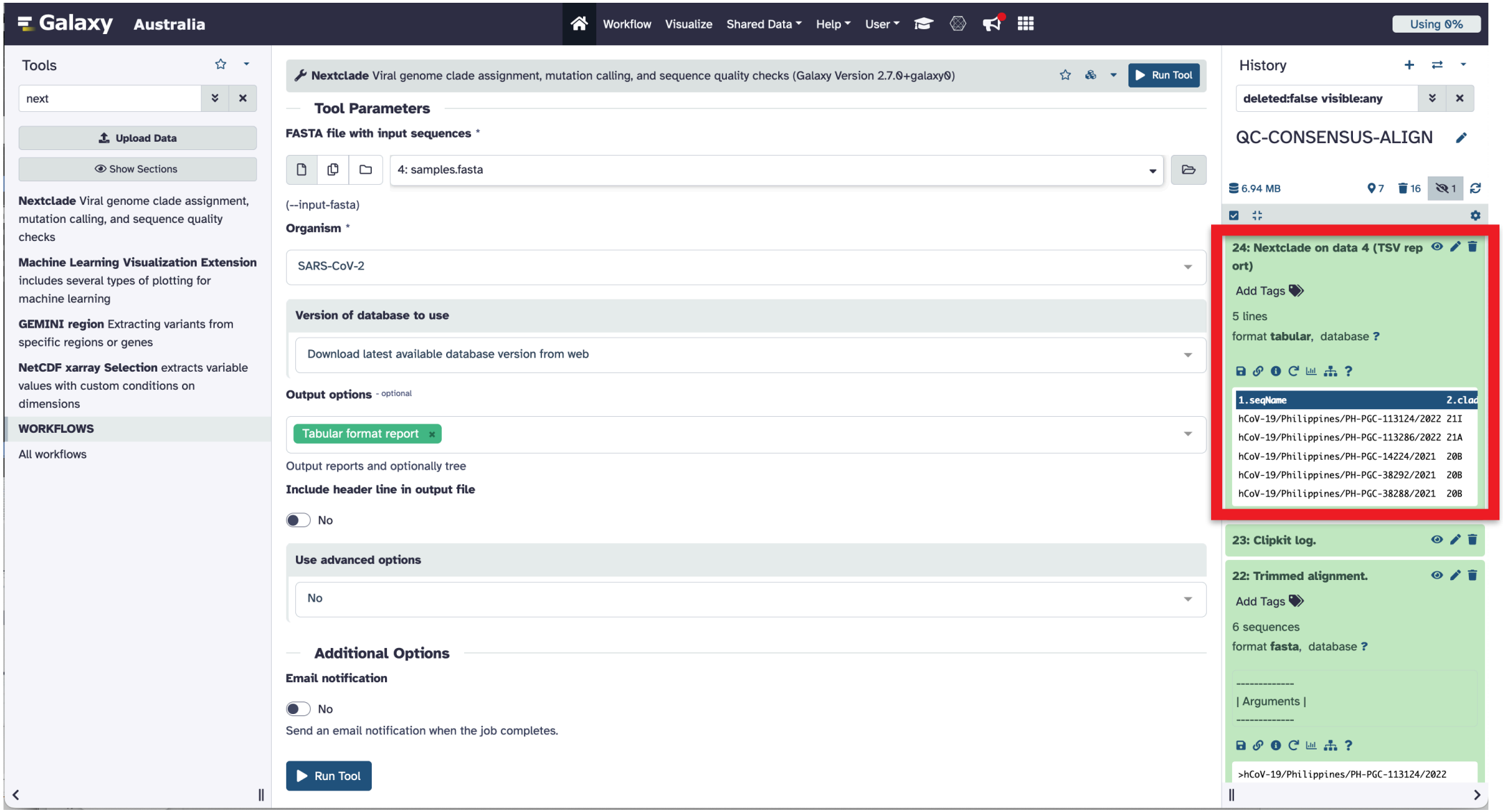
Check for trimmed sequences



Use nextclade to get indiv QC scores



Check Nextclade results



View tabular format using your favorite spreadsheet viewer

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

Python

1: Python

File Edit View Column Row Data Plot System Help | VisiData 3.0.2 | Alt+H for help menu

hCoV-19/Philippines/PH-PGC-113124/2022	21I	B.1.617.2	B.1.617.2	21I	Delta	21I (Delta)	2694.444444	bad	70	22	0	0	57	7	0	184	0	3	C241T,T670G,C835T,…	22029-22034,22289-…		T210…
hCoV-19/Philippines/PH-PGC-113286/2022	21A	B.1.617.2	B.1.617.2	21A	Delta	21A (Delta)	2139.062500	bad	52	13	0	0	46	4	0	0	0	2	G210T,C241T,G569A,…	22029-22034,28248-…		A219…
hCoV-19/Philippines/PH-PGC-14224/2021	20B	B.1.1.263	B.1.1.263	20B		20B	25	good	26	0	0	0	13	0	0	13	0	4	C241T,G2755T,C3037…			
hCoV-19/Philippines/PH-PGC-38292/2021	20B	B.1.1.28	B.1.1.28	20B		20B	50.173611	mediocre	25	0	0	0	12	0	0	13	0	4	C241T,A1291G,C3037…			
hCoV-19/Philippines/PH-PGC-38288/2021	20B	B.1.1.28	B.1.1.28	20B		20B	50.173611	mediocre	25	0	0	0	12	0	0	11	0	4	C241T,A1291G,C3037…			

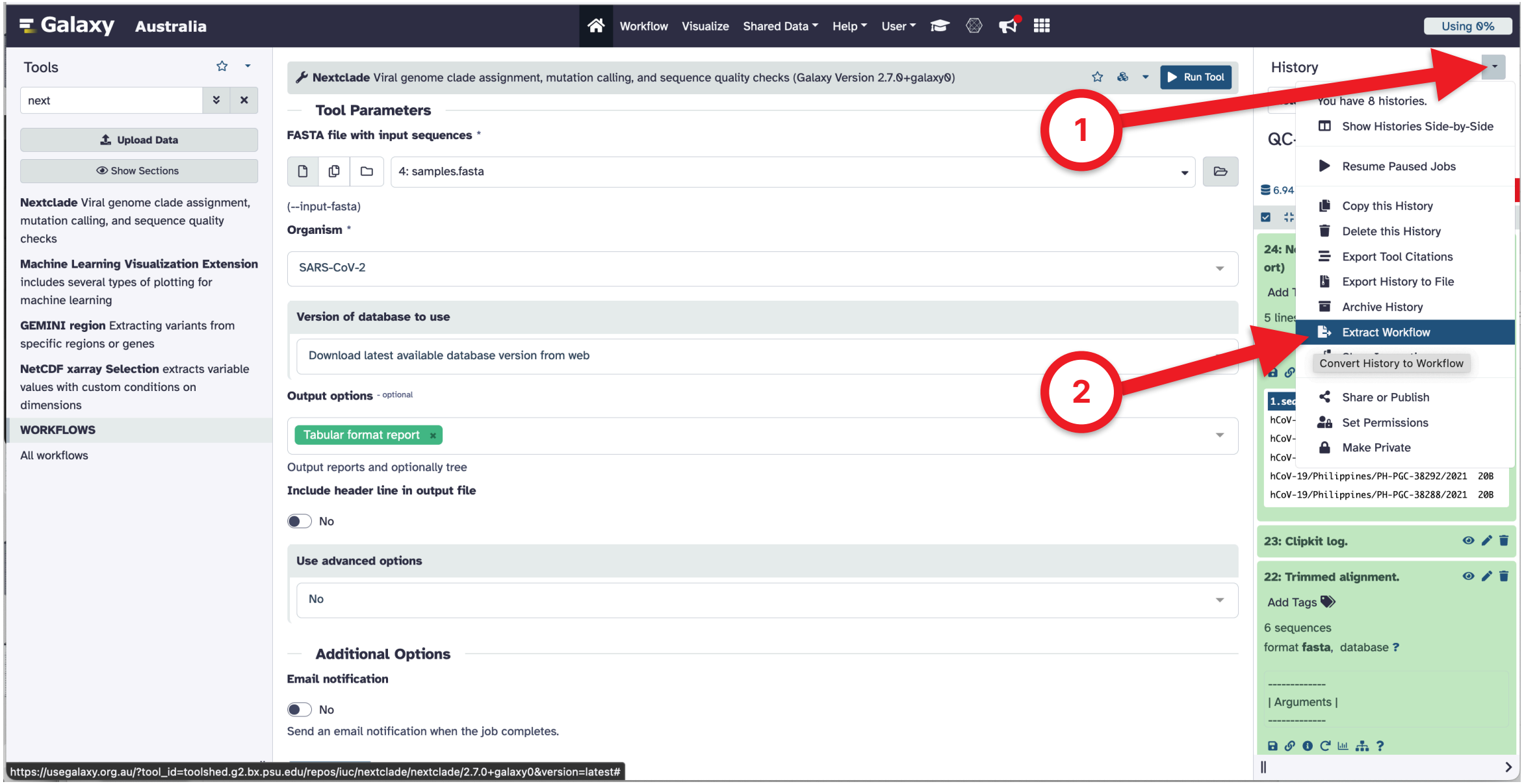
1› Galaxy24-[Nextclade_on_data_4_(TSV_report)]| Right go-right 5 rows

Nextclade columns

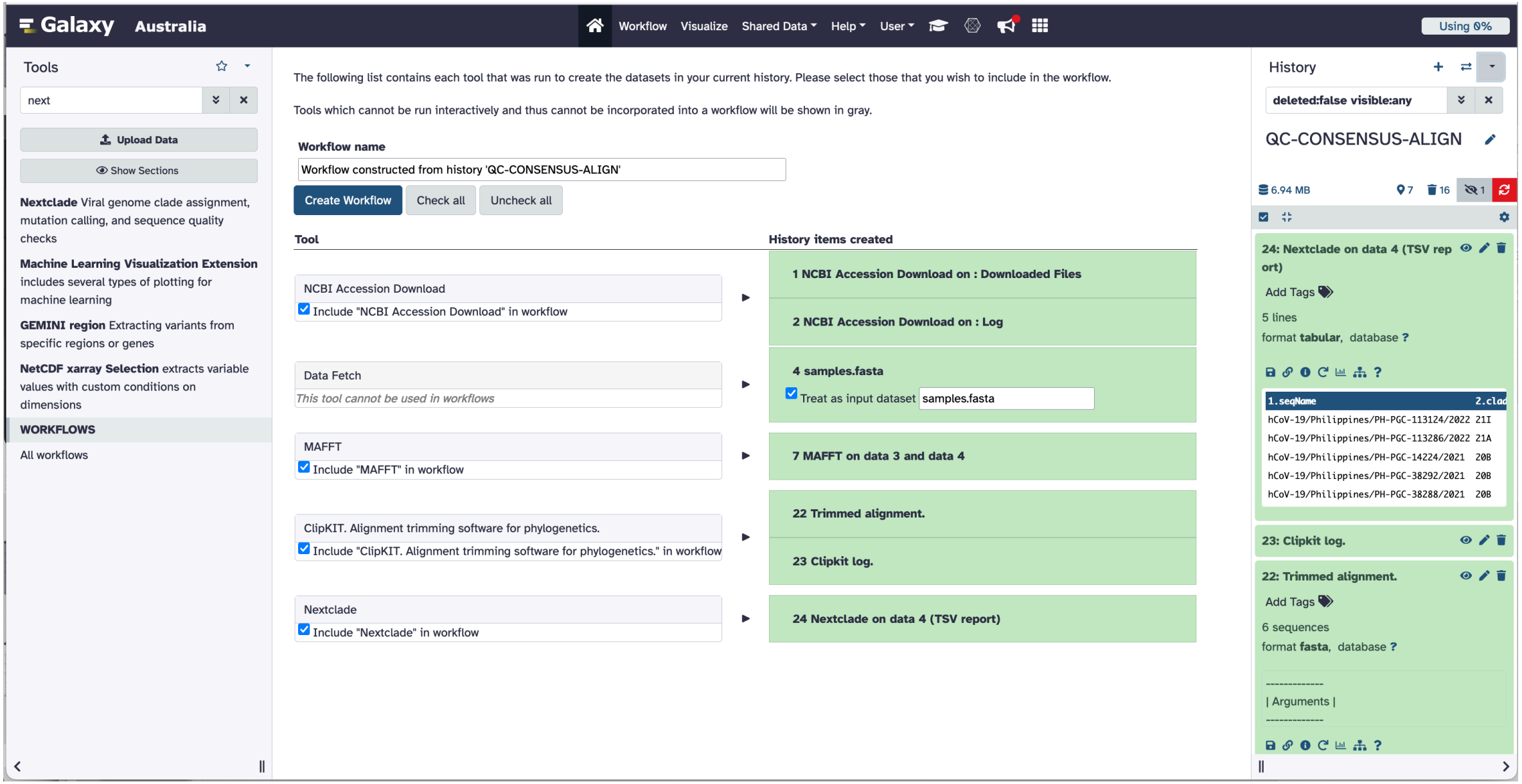
Every row in tabular output corresponds to 1 input sequence. The meaning of columns is described below:

Column name	Meaning	type	Example
index	Index (integer signifying location) of a corresponding record in the input fasta file(s)	non-negative integer	0
seqName	Name of the sequence (as provided in the input file)	string	hCoV-19/USA/SEARCH-4652-SAN/2020
clade	Assigned clade	string	20A
qc.overallScore	Overall quality control score	float	23.5
qc.overallStatus	Overall quality control status	string: `good|mediocre|bad`	mediocre
totalSubstitutions	Total number of detected nucleotide substitutions	non-negative integer	2
totalDeletions	Total number of deleted nucleotide bases	non-negative integer	15
totalInsertions	Total number of inserted nucleotide bases	non-negative integer	3
totalFrameShifts	Total number of detected frame shifts	non-negative integer	0
totalAminoacidSubstitutions	Total number of detected aminoacid substitutions	non-negative integer	1
totalAminoacidDeletions	Total number of deleted amino acid residues	non-negative integer	7
totalAminoacidInsertions	Total number of inserted amino acid residues	non-negative integer	8
totalMissing	Total number of detected missing nucleotides (nucleotide character `N`)	non-negative integer	238
totalNonACGTNs	Total number of detected ambiguous nucleotides (nucleotide characters that are not `A`, `C`, `G`, `T`, `N`)	non-negative integer	2
totalUnknownAa	Total number of unknown aminoacids (aminoacid character `X`)	non-negative integer	0
totalPcrPrimerChanges	Total number of nucleotide mutations detected in PCR primer regions	non-negative integer	0
substitutions	List of detected nucleotide substitutions	comma separated list of strings	C241T,C2061T,C11514T,G23012A
deletions	List of detected nucleotide deletion ranges	comma separated list of strings	201,28881-28882

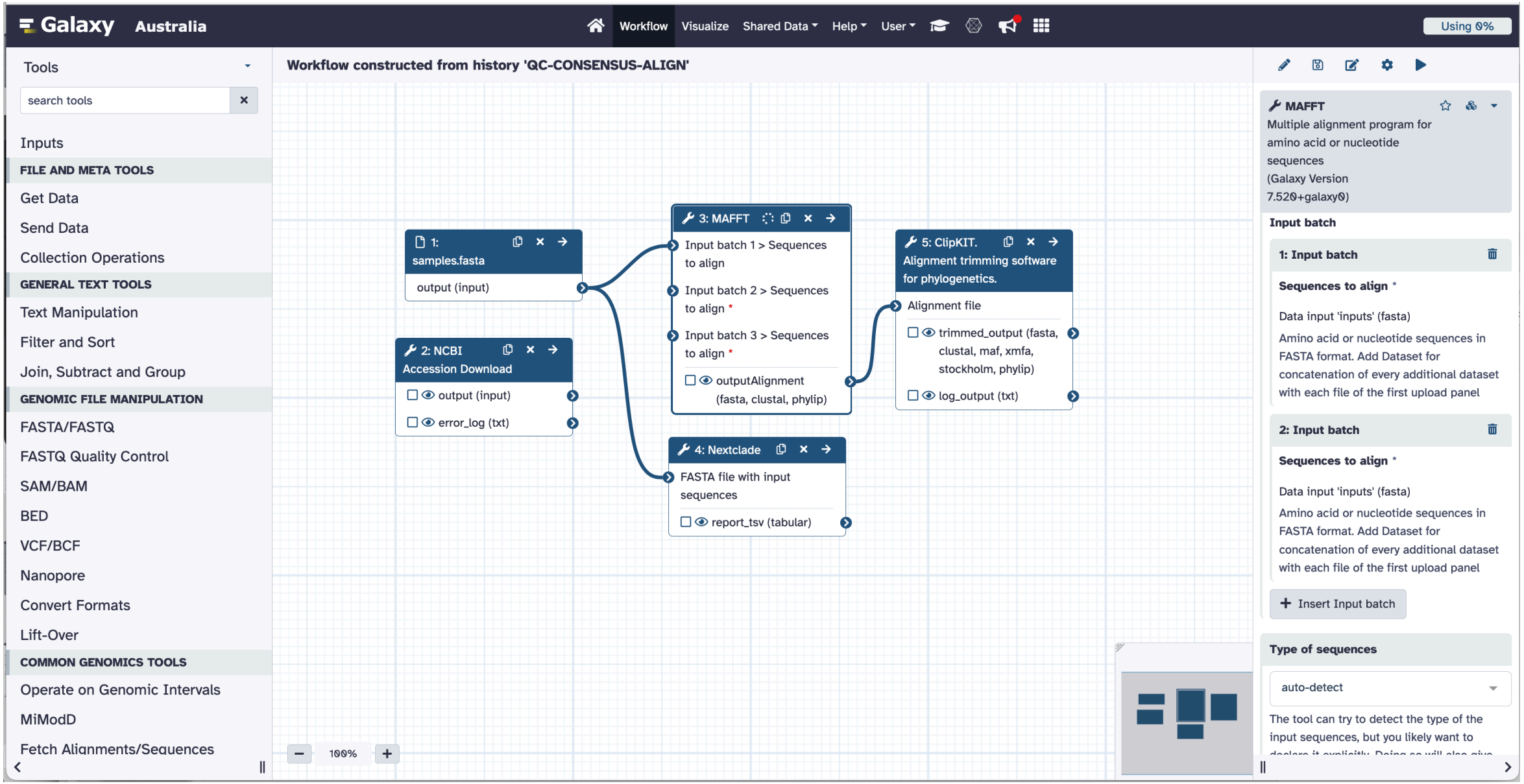
Extract workflow



Create workflow

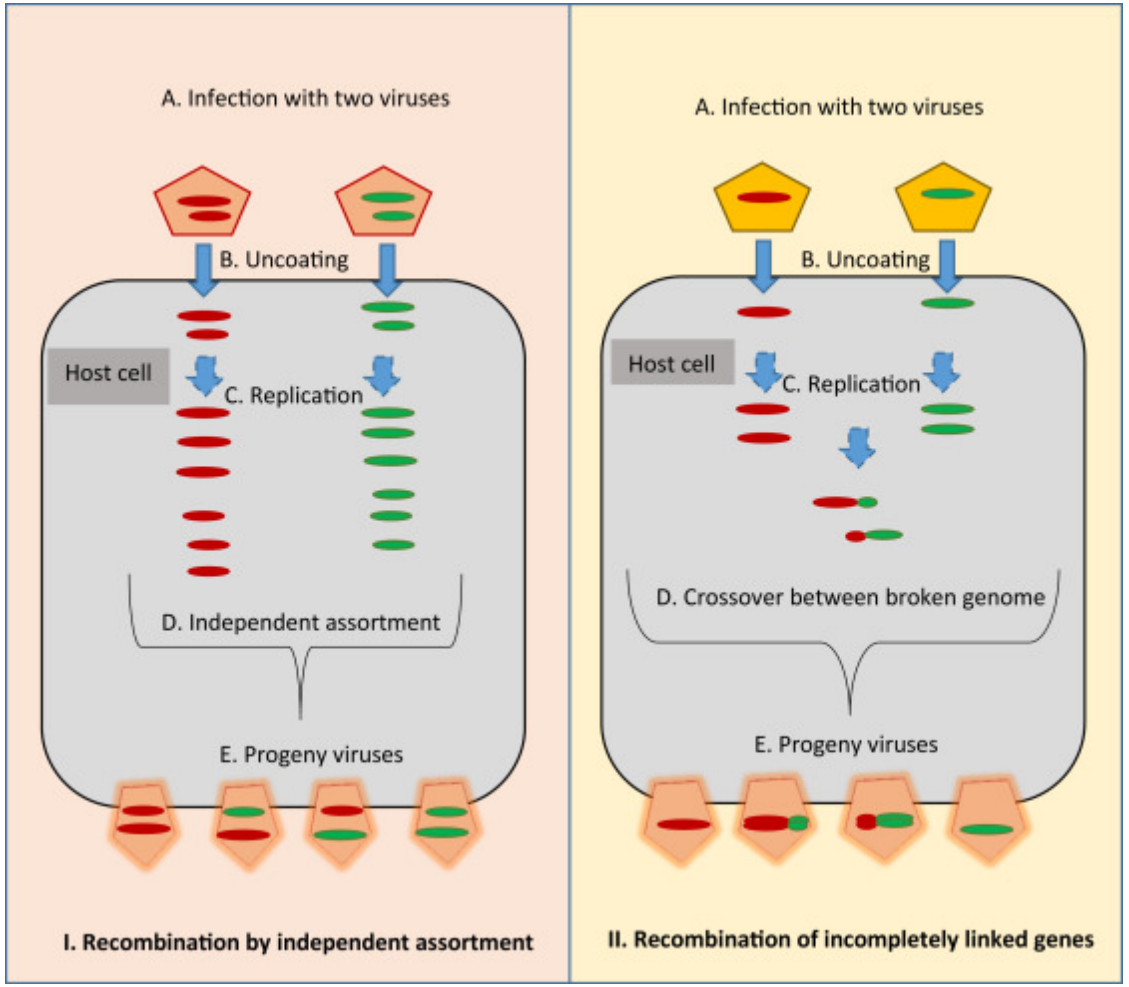


Edit/View Workflow; Use other set of samples for the workflow



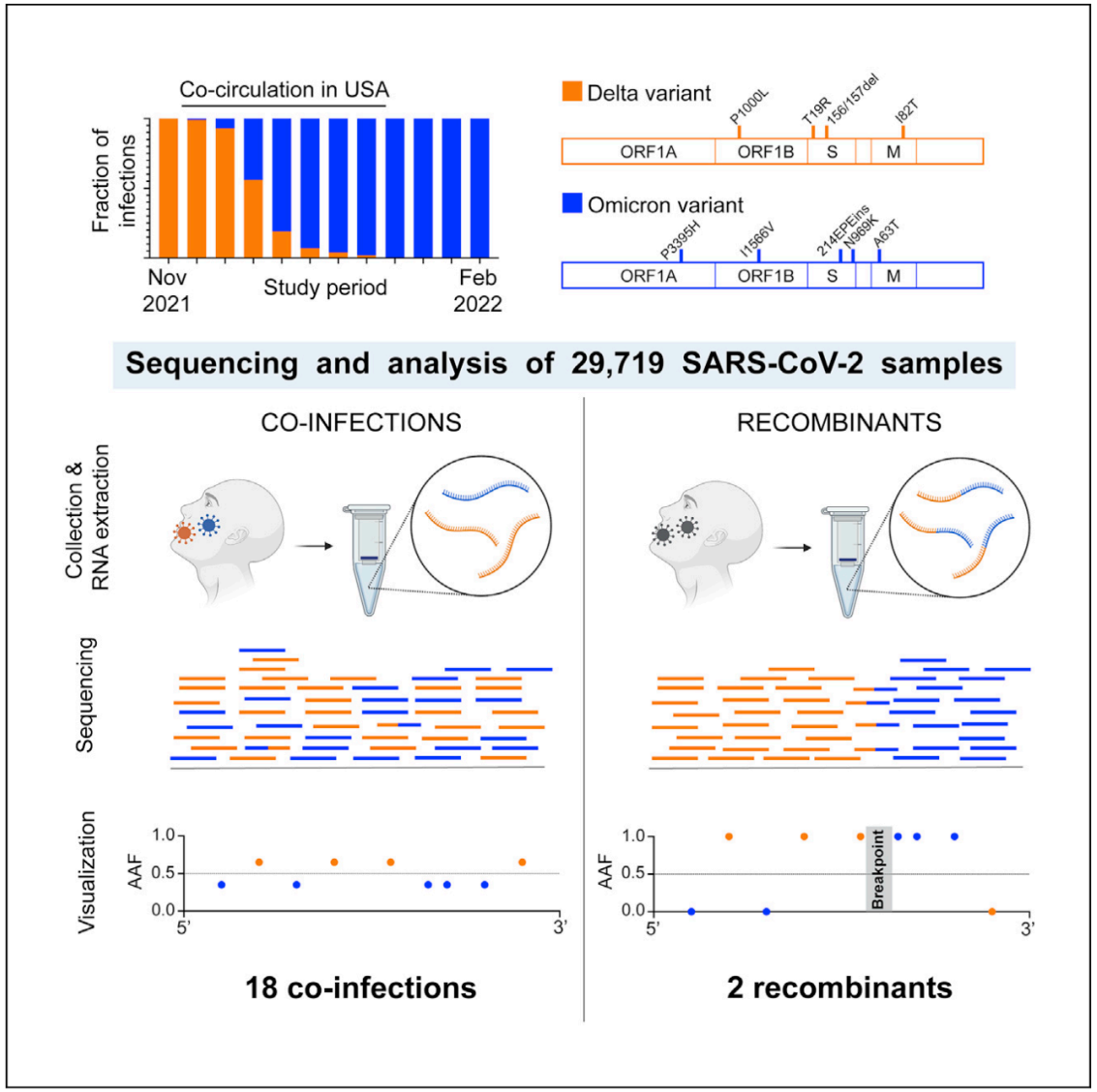
Recombination Detection

- Two viruses of different parent strains infect a host cell at the same time
- **Recombination by independent assortment**
  - occurs when viruses with segmented genomes exchange genetic material which are unlinked and assorted at random during replication
- **Recombination of incompletely linked genes**
  - Recombination may also happen between genes present on the same piece of nucleic acid. If recombination occurs between linked genes, the linkage is thought to be incomplete.



Dhara, A. K., & Nayak, A. K. (Eds.). (2023). Viral infections and antiviral therapies. Academic Press, an imprint of Elsevier.

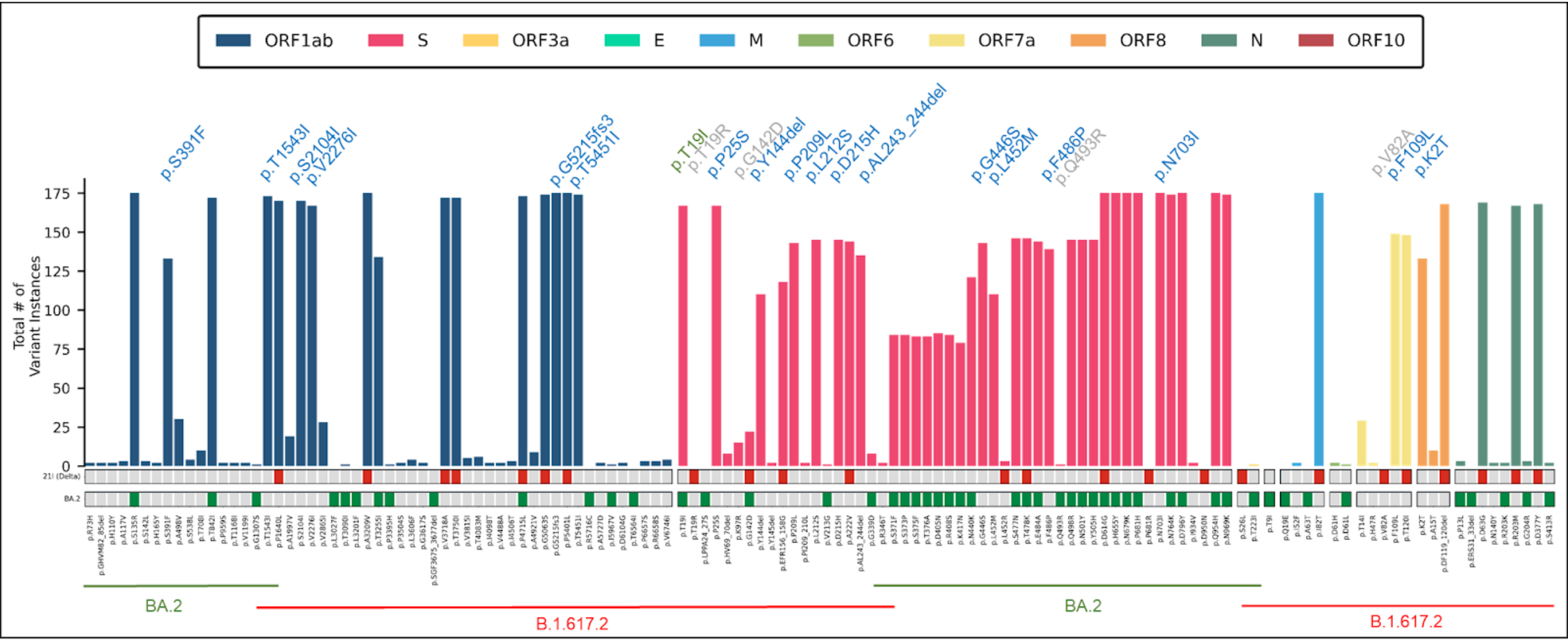
Case Study:
Recombination vs Co-infection



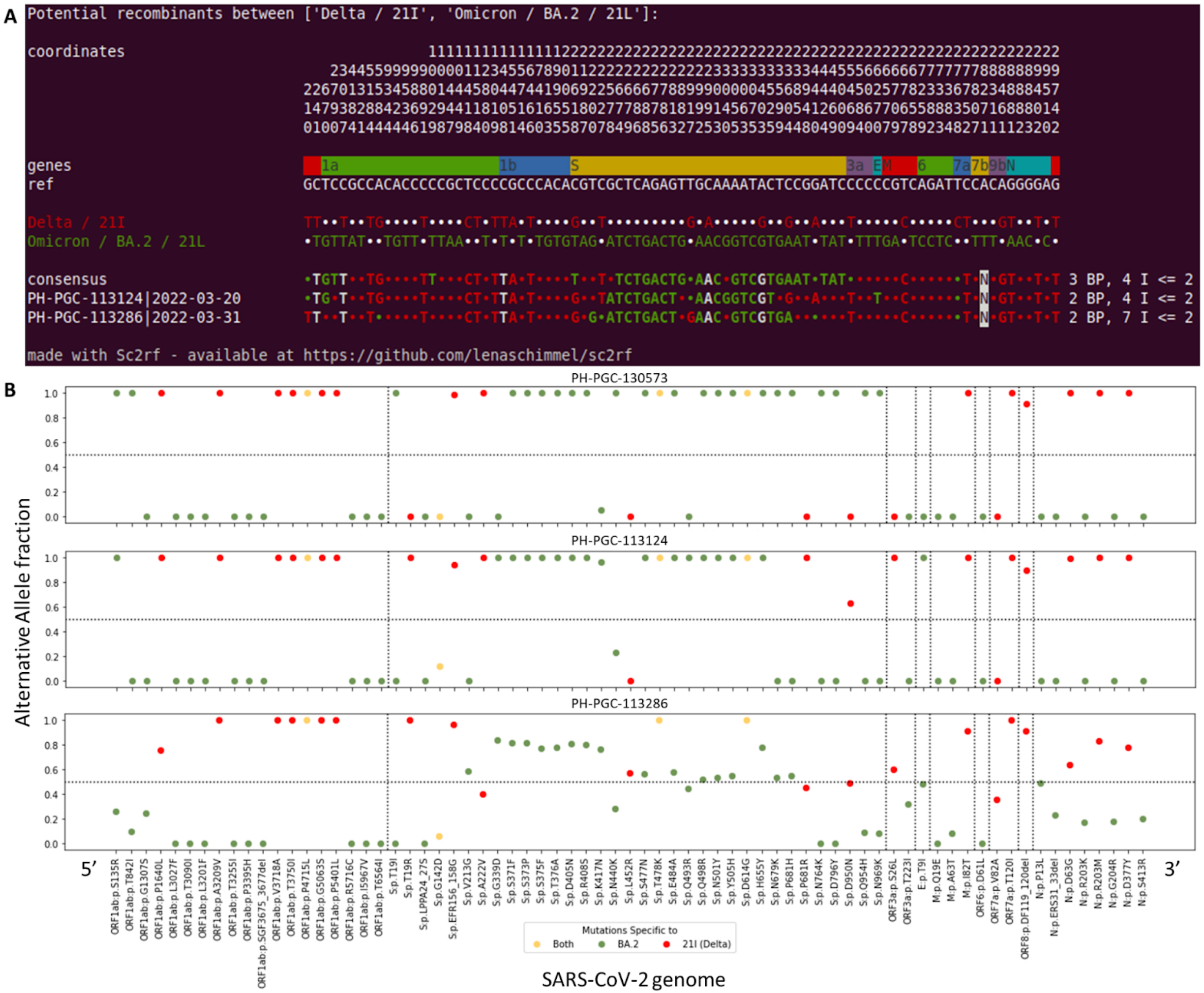
Bolze, A., Basler, T., White, S., Rossi, A. D., Wyman, D., Dai, H., Roychoudhury, P., Greninger, A. L., Hayashibara, K., Beatty, M., Shah, S., Stous, S., McCrone, J. T., Kil, E., Cassens, T., Tsan, K., Nguyen, J., Ramirez, J., Carter, S., ... Luo, S. (2022). Evidence for SARS-CoV-2 Delta and Omicron co-infections and recombination. Med, S2666634022004494. https://doi.org/10.1016/j.medj.2022.10.002

Case Study:
Recombination vs Co-infection

CENTRE FOR PATHOGEN GENOMICS
Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital
A joint venture between The University of Melbourne and The Royal Melbourne Hospital



Pangilinan, E. A. R., Egana, J. M. C., Mantaring, R. J. Q., Telles, A. J. E., Tablizo, F. A., Lapid, C. M., Yangzon, M. S. L., Endozo, J. J. S., Padilla, K. S. A. R., Nipales, J. E., Carandang, L. C. D. L., Enriquez, Z. M. R., Barot, T. A. U., Manlimos, R. A., Mangonon, K. N. P., Plantig, Ma. E. L., Araiza, S. M. M., Llames, J.-H. S., Punayan, K. P., … Saloma, C. P. (2023). Analysis of SARS-CoV-2 Recombinant Lineages XBC and XBC.1 in the Philippines and Evidence for Delta-Omicron Co-infection as a Potential Origin [Preprint]. Genetics. https://doi.org/10.1101/2023.04.12.534029


A
Potential recombinants between ['Delta / 21I', 'Omicron / BA.2 / 21L']:
coordinates
genes
1a 1b S 3a E M 6 7a 7b 9b N
ref
Delta / 21I
Omicron / BA.2 / 21L
consensus
PH-PGC-113124|2022-03-20
PH-PGC-113286|2022-03-31
3 BP, 4 I <= 2
2 BP, 4 I <= 2
2 BP, 7 I <= 2
made with Sc2rf - available at https://github.com/lenaschimmel/sc2rf
B
PH-PGC-130573
PH-PGC-113124
PH-PGC-113286
Alternative Allele fraction
0.0 0.2 0.4 0.6 0.8 1.0
5'
3'
ORF1ab:p.S135R
ORF1ab:p.T842I
ORF1ab:p.G1307S
ORF1ab:p.P1640L
ORF1ab:p.L3027F
ORF1ab:p.T3090I
ORF1ab:p.L3201F
ORF1ab:p.A3209V
ORF1ab:p.T3255I
ORF1ab:p.P3395H
ORF1ab:p.SGF3675_3677del
ORF1ab:p.V3718A
ORF1ab:p.T3750I
ORF1ab:p.P4715L
ORF1ab:p.G5063S
ORF1ab:p.P5401L
ORF1ab:p.R5716C
ORF1ab:p.I5967V
ORF1ab:p.T6564I
S:p.T19I
S:p.T19R
S:p.LPPA24_27S
S:p.G142D
S:p.EFR156_158G
S:p.V213G
S:p.A222V
S:p.G339D
S:p.S371F
S:p.S373P
S:p.S375F
S:p.T376A
S:p.D405N
S:p.R408S
S:p.K417N
S:p.N440K
S:p.L452R
S:p.S477N
S:p.T478K
S:p.E484A
S:p.Q493R
S:p.Q498R
S:p.N501Y
S:p.Y505H
S:p.D614G
S:p.H655Y
S:p.N679K
S:p.P681H
S:p.P681R
S:p.N764K
S:p.D796Y
S:p.D950N
S:p.Q954H
S:p.N969K
ORF3a:p.S26L
ORF3a:p.T223I
E:p.T9I
M:p.Q19E
M:p.A63T
M:p.I82T
ORF6:p.D61L
ORF7a:p.V82A
ORF7a:p.T120I
ORF8:p.DF119_120del
N:p.P13L
N:p.ERS31_33del
N:p.D63G
N:p.R203K
N:p.R203M
N:p.G204R
N:p.D377Y
N:p.S413R
Mutations Specific to
Both
BA.2
21I (Delta)
SARS-CoV-2 genome

Tools for detecting recombination

- 3SEQ
- Sc2rf (for SARS-CoV-2)
- RDP4 / RDP5
- OpenRDP
- GENECONV
- (many more)

World Health Organization
Philippines

CENTRE FOR PATHOGEN GENOMICS

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

THANK YOU

jcegana@up.edu.ph

KDCA
Korea Disease Control and Prevention Agency

Doherty Institute

THE UNIVERSITY OF MELBOURNE

The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital