



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



CENTRE FOR
PATHOGEN
GENOMICS



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

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



CENTRE FOR
PATHOGEN
GENOMICS



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

- **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



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

- **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



CENTRE FOR
PATHOGEN
GENOMICS



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

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

Assembly metrics



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

< Medical Detectives
Patient 010 (CSF) ▾

Consensus Genome Pipeline v3.1.0 ▾ | processed 2 years ago | [SAMPLE DETAILS](#)

[Share](#) [Download All](#) [?](#)

Metagenomic Antimicrobial Resistance (Deprecated) **Consensus Genome**

Mapped to: Chikungunya virus [▼](#)

[Learn more about consensus genomes >](#)

Is my consensus genome complete? ⓘ

Assembly metrics

Taxon	Mapped Reads	GC Content	SNPs	%id	Informative Nucleotides	% Genome Called	Missing Bases	Ambiguous Bases
Chikungunya virus	20816	50.2%	0	100%	11772	99.7%	34	2

How good is the coverage? ⓘ

Coverage stats

Reference NCBI Entry	MK468620.1	Reference Length	11808	Coverage Depth	255.8x	Coverage Breadth	100.0%
----------------------	------------	------------------	-------	----------------	--------	------------------	--------

Coverage Plot

Coverage (SymLog)

Reference Accession

MK468620.1 - Chikungunya virus isolate CHRF_0012_07-11-2017, complete genome

Organism-specific considerations



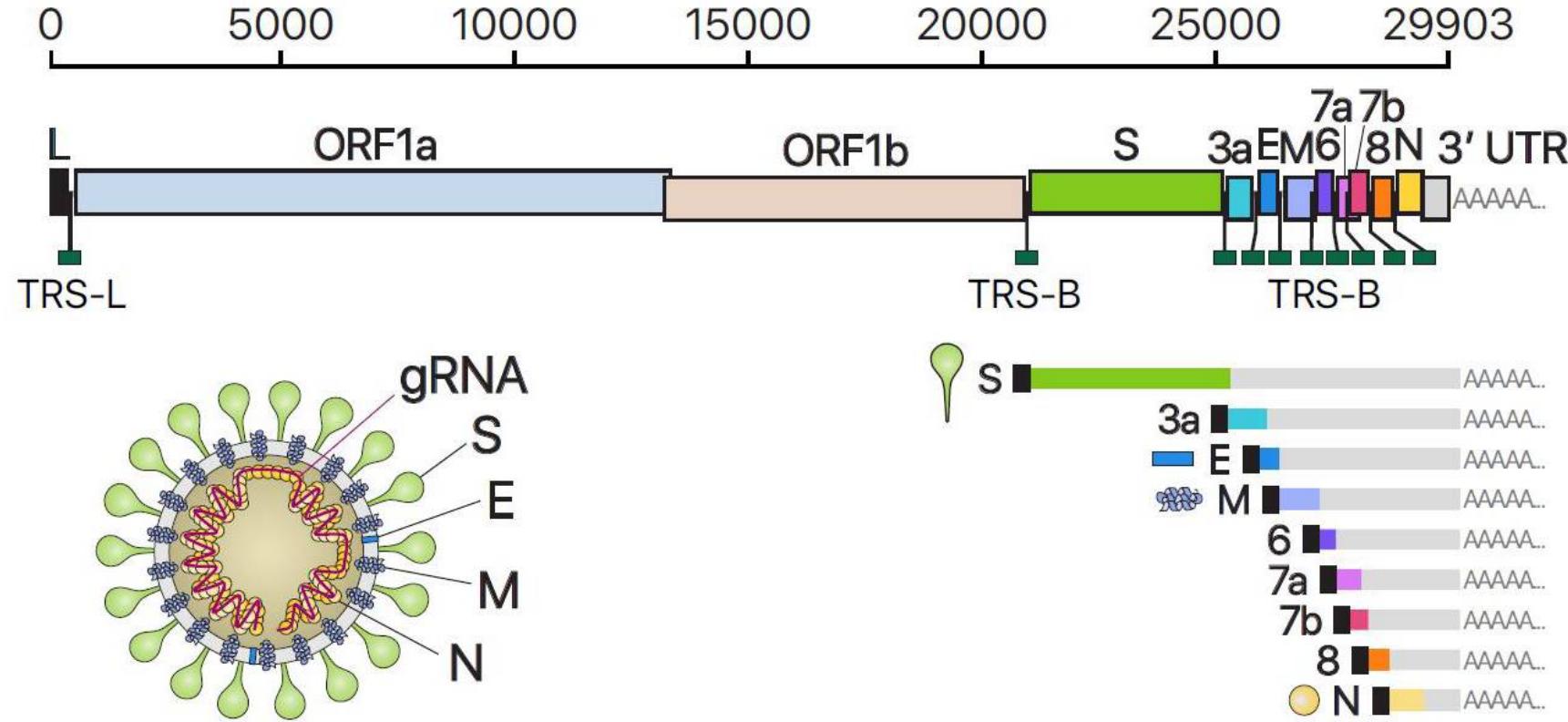
A joint venture between The University of Melbourne and The Royal Melbourne Hospital

- **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.

SARS-CoV-2 Genome



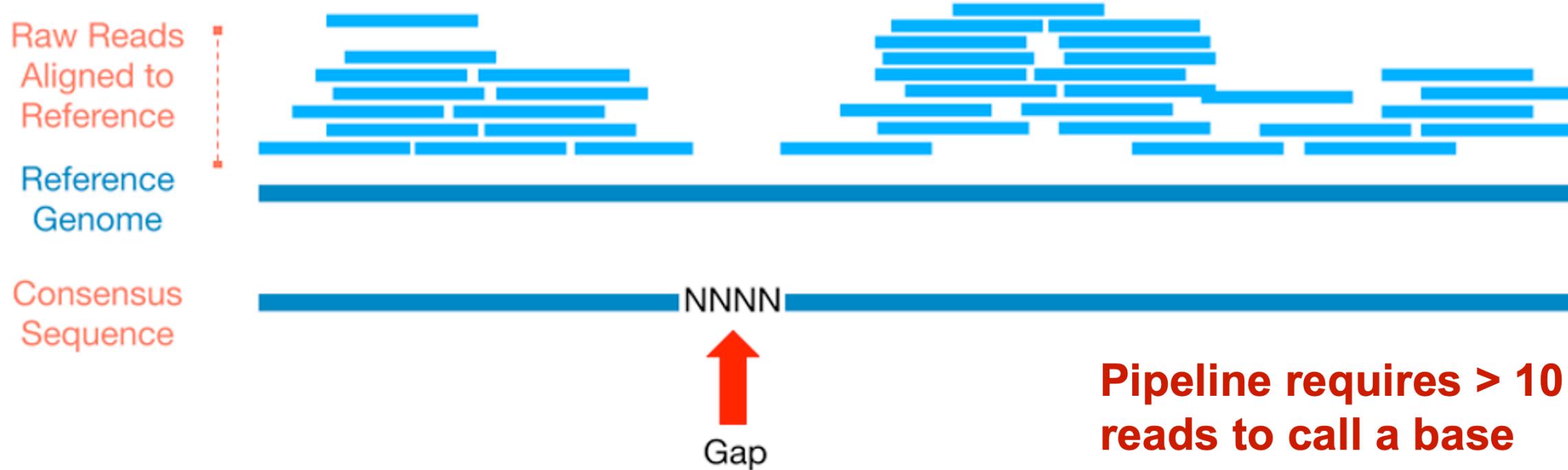
A joint venture between The University of Melbourne and The Royal Melbourne Hospital



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

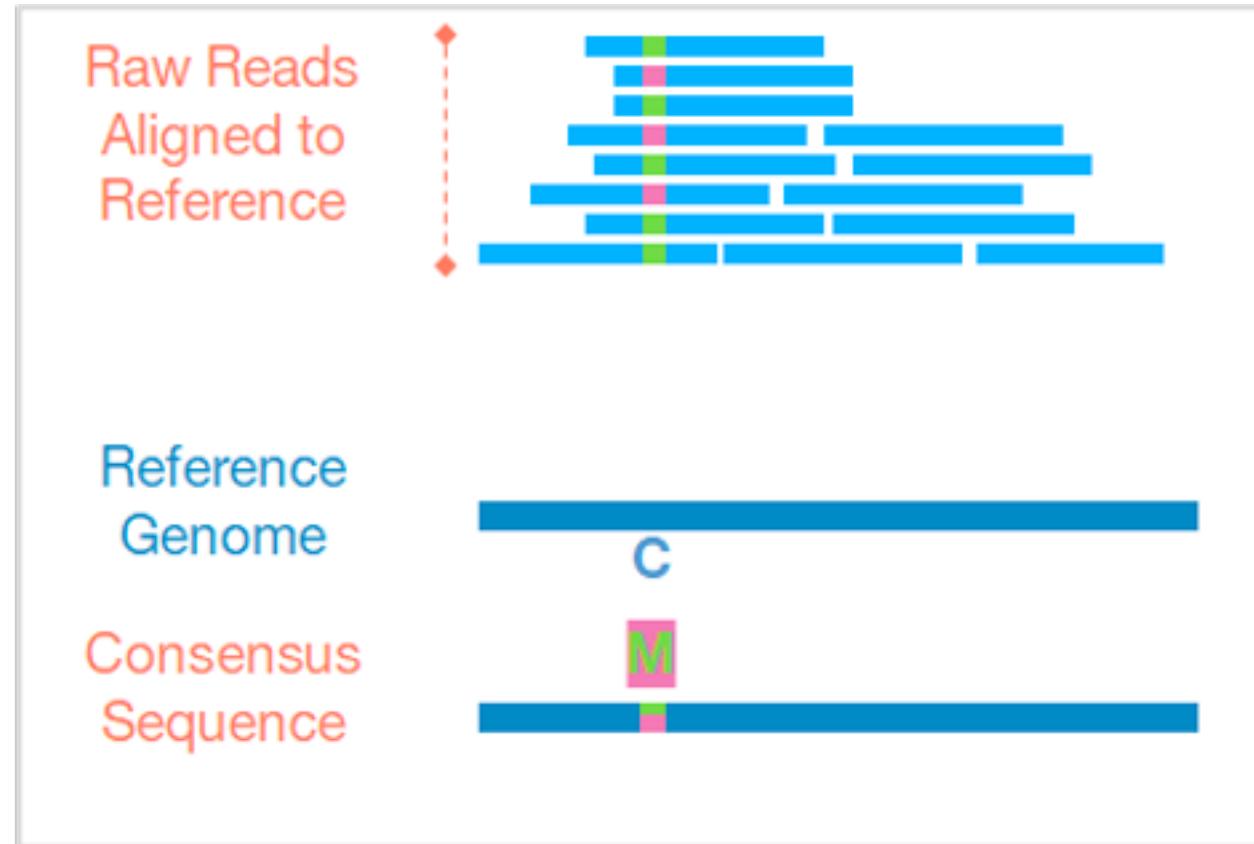
Checking read alignments

- Gaps or low coverage of consensus genome



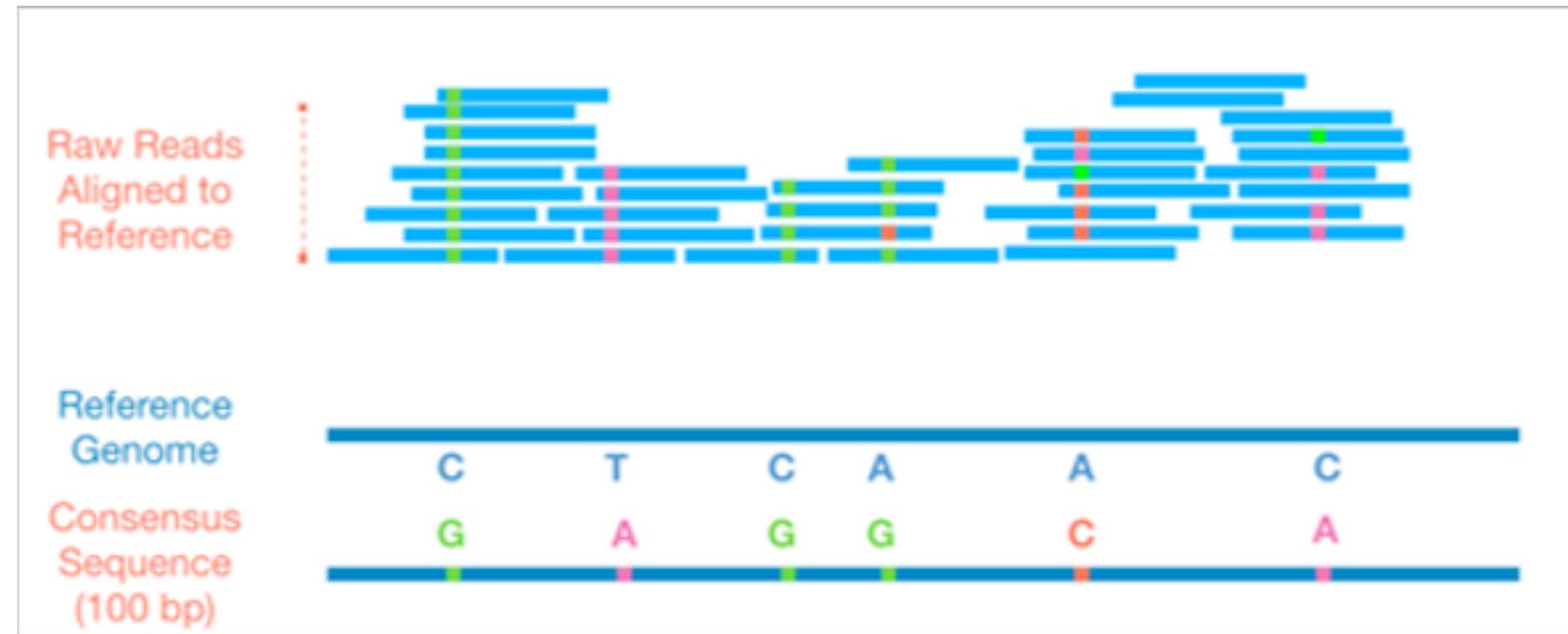
Checking read alignments

- Ambiguous bases



Checking read alignments

- Single nucleotide polymorphisms (SNPs)



GISAID Quality Control



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	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-132210/2022	Original	EPI_ISL_17389138	2022-09-26	2023-04-04		29,867	Human	Asia / Philippines	DAVAO INT	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-132213/2022	Original	EPI_ISL_17389137	2022-09-26	2023-04-04		This submission requires further investigation! It appears to contain markers of multiple lineages from both Delta and Omicron variants				
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131761/2022	Original	EPI_ISL_17389136	2022-09-26	2023-04-04		29,838	Human	Asia / Philippines	KIDAPAWA	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-132196/2022	Original	EPI_ISL_17389135	2022-09-26	2023-04-04		29,847	Human	Asia / Philippines	OTHERS - I	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131762/2022	Original	EPI_ISL_17389134	2022-09-26	2023-04-04		29,838	Human	Asia / Philippines	KIDAPAWA	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-132215/2022	Original	EPI_ISL_17389133	2022-09-26	2023-04-04		29,878	Human	Asia / Philippines	DAVAO INT	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131727/2022	Original	EPI_ISL_17389132	2022-09-25	2023-04-04		29,880	Human	Asia / Philippines	ST. ELIZAB	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131723/2022	Original	EPI_ISL_17389131	2022-09-25	2023-04-04		29,882	Human	Asia / Philippines	ST. ELIZAB	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131728/2022	Original	EPI_ISL_17389130	2022-09-25	2023-04-04		29,854	Human	Asia / Philippines	ST. ELIZAB	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131726/2022	Original	EPI_ISL_17389129	2022-09-25	2023-04-04		29,877	Human	Asia / Philippines	ST. ELIZAB	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131725/2022	Original	EPI_ISL_17389128	2022-09-25	2023-04-04		29,846	Human	Asia / Philippines	ST. ELIZAB	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131680/2022	Original	EPI_ISL_17389127	2022-09-24	2023-04-04		29,866	Human	Asia / Philippines	DR. ARTUF	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-131682/2022	Original	EPI_ISL_17389126	2022-09-24	2023-04-04		29,249	Human	Asia / Philippines	DR. ARTUF	

GISAID Quality Control



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

<input type="checkbox"/>	Virus name	Passage de	Accession ID	Collection da	Submission D		Length	Host	Location	Originating	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66977/2021	Original	EPI_ISL_5547858	2021-08-02	2021-10-26		29,847	Human	Asia / Philippines	Teresita Jal...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-68464/2021	Original	EPI_ISL_5547879	2021-08-07	2021-10-26		Stretches of NNNs (1.45% of overall sequence). Gap of 18 nucleotides when compared to the reference WIV04 sequence.				
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-68345/2021	Original	EPI_ISL_5547908	2021-08-06	2021-10-26		29,847	Human	Asia / Philippines	Dr. Jose N...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-64107/2021	Original	EPI_ISL_5547982	2021-07-23	2021-10-26		29,857	Human	Asia / Philippines	Zamboanga	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-69269/2021	Original	EPI_ISL_5548279	2021-08-09	2021-10-26		29,855	Human	Asia / Philippines	Philippine R...	
<input type="checkbox"/>	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...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66688/2021	Original	EPI_ISL_5548835	2021-07-27	2021-10-26		29,869	Human	Asia / Philippines	Qualimed H...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-67607/2021	Original	EPI_ISL_5549643	2021-08-02	2021-10-26		29,853	Human	Asia / Philippines	Northern Mi...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66494/2021	Original	EPI_ISL_5550705	2021-07-26	2021-10-26		29,846	Human	Asia / Philippines	Western Vis...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-67057/2021	Original	EPI_ISL_5550827	2021-07-29	2021-10-26		29,873	Human	Asia / Philippines	Vicente Sot...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66927/2021	Original	EPI_ISL_5551291	2021-08-01	2021-10-26		29,872	Human	Asia / Philippines	Corazon Lo...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66921/2021	Original	EPI_ISL_5551331	2021-08-03	2021-10-26		29,849	Human	Asia / Philippines	Corazon Lo...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66908/2021	Original	EPI_ISL_5551345	2021-07-31	2021-10-26		29,852	Human	Asia / Philippines	Corazon Lo...	
<input type="checkbox"/>	hCoV-19/Philippines/PH-PGC-66945/2021	Original	EPI_ISL_5551379	2021-08-04	2021-10-26		29,886	Human	Asia / Philippines	Corazon Lo...	

Total: 18,804 viruses

<< < 1 2 3 4 5 > >>

Charts EPI_SET Select Analysis Download

Nextclade Quality Control



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



The screenshot shows the Nextclade interface with the following details:

- Header:** Nextclade, Start, Dataset, Results (highlighted in blue), Tree, Export, a green checkmark icon followed by "Done. Total sequences: 5. Succeeded: 5", Settings, About, Citation, Docs, CLI, and language links (EN).
- Table Headers:** i, Sequence name, QC, Clade, Pango lineage (Nextclade), WHO name, Mut., non-ACGTN, Ns, Cov., Gaps, Ins., FS, SC.
- Table Rows:** Each row represents a sequence entry with a green checkmark icon, the sequence name, QC status, clade, Pango lineage, WHO name, and various metrics (Mut., non-ACGTN, Ns, Cov., Gaps, Ins., FS, SC) followed by a barcode visualization.
- Filter:** A filter icon is located above the table.
- Panel:** A panel titled "Nucleotide sequence" is open on the right side of the interface.

Nextclade Quality Control



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 X D S Q EN

i	Sequence name	QC	Clade	Pango lineage (Nextclade)	WHO name	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC	Filter	Nucleotide sequence
2	✓ hCoV-19/Philippines/PH-PGC-38292/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
3	✓ hCoV-19/Philippines/PH-PGC-38288/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
1	✓ hCoV-19/Philippines/PH-PGC-14224/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
4	✓ hCoV-19/Philippines/PH-PGC-113124/20	N M P C F S	?	?	?	?	?	?	99.1%	22	0	0	0	?	?
0	✓ hCoV-19/Philippines/PH-PGC-113286/20	N M P C F S	?	?	?	?	?	?	99.7%	13	0	0	0	?	?

Overall QC score: 25
Overall QC status: good
Detailed QC assessment:

- N** Missing Data: good
No issues
- M** Mixed Sites: good
No issues
- P** Private Mutations: mediocre
QC score: 50. Reverted substitutions: 0, Labeled substitutions: 1, Unlabeled substitutions: 16, Deletion ranges: 0. Weighted total: 20
- C** Mutation Clusters: good
No issues
- F** Frame shifts: good
- S** Stop codons: good
No issues

Example of a good 😍 overall QC status

Nextclade Quality Control



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 X D S Q EN

i	Sequence name	QC	Clade	Pango lineage (Nextclade)	WHO name	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC	Filter	Nucleotide sequence
2	✓ hCoV-19/Philippines/PH-PGC-38292/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
3	✓ hCoV-19/Philippines/PH-PGC-38288/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
1	✓ hCoV-19/Philippines/PH-PGC-14224/20	N M P C F S	?	?	?	?	?	?	99.8%	0	0	0	0	?	?
4	✓ hCoV-19/Philippines/PH-PGC-113124/20	N M P C F S	?	?	?	?	?	?	99.1%	22	0	0	0	?	?
0	✓ hCoV-19/Philippines/PH-PGC-113286/20	N M P C F S	?	?	?	?	?	?	99.7%	13	0	0	0	?	?

Overall QC score: 50
Overall QC status: mediocre
Detailed QC assessment:

- N** Missing Data: good
No issues
- M** Mixed Sites: good
No issues
- P** Private Mutations: mediocre
QC score: 71. Reverted substitutions: 0, Labeled substitutions: 3, Unlabeled substitutions: 13, Deletion ranges: 0. Weighted total: 25
- C** Mutation Clusters: good
No issues
- F** Frame shifts: good
- S** Stop codons: good
No issues

Example of a mediocre 🤷 overall QC status

Nextclade Quality Control



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

Sequence name QC Clade Pango lineage WHO name Mut. non-ACGTN Ns Cov. Gaps Ins. FS SC

N M P C F S

Overall QC score: 2694
Overall QC status: bad
Detailed QC assessment:

- N** Missing Data: good
No issues
- M** Mixed Sites: good
No issues
- P** Private Mutations: bad
QC score: 517. Reverted substitutions: 4, Labeled substitutions: 20, Unlabeled substitutions: 26, Deletion ranges: 2. Weighted total: 132
- C** Mutation Clusters: mediocre
Mutation clusters found. Seen 1 mutation clusters with total of 7 mutations. QC score: 50
- F** Frame shifts: good
- S** Stop codons: good
No issues

Nucleotide sequence

The screenshot shows the Nextclade software interface. At the top, there's a navigation bar with tabs for 'Start', 'Dataset', 'Results' (which is currently selected), 'Tree', 'Export', and a success message 'Done. Total sequences: 5. Succeeded: 5'. Below the navigation is a table with columns for index (i), sequence name, QC status, clade, Pango lineage, WHO name, mutations, non-ACGTN, Ns, coverage, gaps, insertions, frame shifts, and stop codons. A modal window is open over the table, focusing on the first sequence (index 2). The modal provides a detailed 'Detailed QC assessment' for this sequence, listing various metrics and their status (e.g., 'Missing Data: good', 'Private Mutations: bad'). The background table shows other sequences with similar headers and some colored icons in the QC column.

Example of a bad 🤢 overall QC status

Nextclade Quality Scores



CENTRE FOR
PATHOGEN
GENOMICS



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



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

The Genome Lab has just launched! This corner of the Galaxy is dedicated to genome assembly and annotation.

Home News People About Support Docs

Galaxy AUSTRALIA

BR NO — GALAXY COMMUNITY CONFERENCE 2024 —

June 24-29 2024, Brno, Czech Republic

[GCC2024 details](#)

GCC is the flagship Galaxy event of the year! Join the Galaxy world-wide community to share and learn about data science.

Galaxy Australia is an **open, web-based** platform for accessible, reproducible and transparent computational research. Galaxy supports thousands of documented and maintained tools that are free to use. We facilitate on-demand training capacities and provision **600GB** for Australian institutional (and 100GB for other) users.

Workflow Visualize Shared Data Help User

Tools search tools Upload Data

FILE AND META TOOLS

- Get Data
- Send Data
- Collection Operations

GENERAL TEXT TOOLS

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group

GENOMIC FILE MANIPULATION

- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM

- BED
- VCF/BCF
- Nanopore
- Convert Formats
- Lift-Over

COLUMN GENOMICS TOOLS

- Operate on Genomic Intervals
- MiModD
- Fetch Alignments/Sequences

GENOMICS ANALYSIS

History + search datasets

Unnamed history

This history is empty.
You can load your own data or get data from an external source.

Name your history



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

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Tools

search tools

Upload Data

FILE AND META TOOLS

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

Lift-Over

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

MiModD

Fetch Alignments/Sequences

GENOMICS ANALYSIS

The Genome Lab has just launched! This corner of the Galaxy is dedicated to genome assembly and annotation.

Home News People About Support Docs

Galaxy AUSTRALIA

BR NO — GALAXY COMMUNITY CONFERENCE 2024 —

June 24-29 2024, Brno, Czech Republic

GCC2024 details

GCC is the flagship Galaxy event of the year! Join the Galaxy world-wide community to share and learn about data science.

Galaxy Australia is an **open, web-based** platform for accessible, reproducible and transparent computational research. Galaxy supports thousands of documented and maintained tools that are free to use. We facilitate on-demand training capacities and provision **600GB** for Australian institutional (and 100GB for other) users.

History

search datasets

Unnamed history

QC-CONSENSUS-ALIGN

Add Tags

Save Cancel

0 B

This history is empty.
You can load your own data or get data from an external source.

NCBI accession download



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

Galaxy Australia

Workflow Visualize Shared Data Help User Home News People About Support Docs

The Genome Lab has just launched! This corner of the Galaxy is dedicated to genome assembly and annotation.

History Using 0%

search datasets

QC-CONSENSUS-ALIGN

0 B

This history is empty.
You can load your own data or get data from an external source.

BRNO — GALAXY COMMUNITY CONFERENCE 2024 —

June 24-29 2024, Brno, Czech Republic

GCC2024 details

GCC is the flagship Galaxy event of the year! Join the Galaxy world-wide community to share and learn about data science.

Galaxy Australia is an **open, web-based** platform for accessible, reproducible and transparent computational research. Galaxy supports thousands of documented and maintained tools that are free to use. We facilitate on-demand training capacities and provision **600GB** for Australian institutional (and 100GB for other) users.

NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

Protein Database Downloader

Ratmine server

SRA server

UCSC Archaea table browser

UCSC Main table browser

Unipept retrieve taxonomy for peptides

UniProt ID mapping and retrieval

WormBase server

YeastMine server

ZebrafishMine server

Send Data

Collection Operations

GENERAL TEXT TOOLS

https://usegalaxy.org.au/tool_runner?tool_id=toolshed.g2.bx.psu.edu/repos/iuc/ncbi_acc_download/ncbi_acc_download/0.2.8+galaxy0

Add NC_045512.2 and run

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

Galaxy Australia

Workflow Visualize Shared Data Help User User History Using 0%

NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API (Galaxy Version 0.2.8+galaxy0)

Tool Parameters

Select source for IDs

1 Direct Entry

2 ID List *

NC_045512.2

Newline/Comma separated list of IDs

Molecule Type

Nucleotide

File Format *

FASTA

Range - optional

Region to subset accession. Start and end position separated by ':', '..' or '>'. Only for single accession (--range)

How to handle download failures *

Abort with error on first failure

Add accession to failed list and continue

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

History

search datasets

QC-CONSENSUS-ALIGN

0 B

This history is empty.
You can load your own data or get data from an external source.

GENERAL TEXT TOOLS

Choose local files

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

The screenshot shows the Galaxy Australia interface. A central modal window titled "Upload from Disk or Web" is open. It has tabs for "Regular", "Composite", "Collection", and "Rule-based". Below the tabs is a large dashed rectangular area with the placeholder text "Drop files here". At the bottom of the modal are buttons for "Type (set all): Auto-detect", "Q Genome (set all): unspecified (?)", "Choose local files" (which is highlighted with a red box), "Choose remote files", "Paste/Fetch data", "Start", "Pause", "Reset", and "Close". A dropdown menu is open under "Choose local files", listing several options: Histogram, GC Skew, Construct Expression..., Normalize, Lineage Branch Analy..., Inspect Expression S..., Scatter: t-SNE plot, and Scatter: Calculate QC... . The background shows the main Galaxy interface with tool panels on the left and a history panel on the right.

Select samples.fasta

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Upload from Disk or Web

Regular Composite Collection Rule-based

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
samples.fasta	147.8 KB	Auto-det...	unspecified (?)	0%	

Type (set all): Auto-detect Genome (set all): unspecified (?)

Choose local files Choose remote files Paste/Fetch data Start Pause Reset Close

Histogram
GC Skew
Construct Expression...
Normalize
Lineage Branch Analy...
Inspect Expression S...
Scater: t-SNE plot
Scater: Calculate QC...

History

search datasets

QC-CONSENSUS-ALIGN

30.5 kB

2: NCBI Accession Download on : Log

1: NCBI Accession Download on : Downloaded Files

a list with 1 dataset

FILE AND META TOOLS

Get Data

Download and Extract Reads in BAM format from NCBI SRA

Download and Extract Reads in FASTQ format from NCBI SRA

EBI SRA ENA SRA

EuPathDB server

Faster Download and Extract Reads in FASTQ format from NCBI SRA

Flymine server

GrameneMart Central server

HbVar Human Hemoglobin Variants and Thalassemias

InterMine server

metabolicMine server

modENCODE fly server

modENCODE modMine server

modENCODE worm server

MouseMine server

NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

Protein Database Downloader

Ratmine server

SRA server

Check history for samples.fasta

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

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Tools

search tools

Upload Data

FILE AND META TOOLS

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

Lift-Over

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

MiModD

Fetch Alignments/Sequences

GENOMICS ANALYSIS

The Genome Lab has just launched! This corner of the Galaxy is dedicated to genome assembly and annotation.

Home News People About Support Docs

Galaxy AUSTRALIA

Galaxy Australia Genome Lab: The new online workbench for easier data analysis

Access ready-made tools, workflows and tutorials for

Data preparation Genome annotation Genome assembly

genome.usegalaxy.org.au

Click to switch to Genome Lab

Australian BioCommons

History

search datasets

QC-CONSENSUS-ALIGN

182 kB

4: samples.fasta

2: NCBI Accession Download on : Log

1: NCBI Accession Download on : Downloaded Files

a list with 1 fasta dataset

History

search datasets

QC-CONSENSUS-ALIGN

182 kB

4: samples.fasta

2: NCBI Accession Download on : Log

1: NCBI Accession Download on : Downloaded Files

a list with 1 fasta dataset

Search mafft in tools

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

MAFFT Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)

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

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

+ 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

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)

« 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 ?

Severe acute respiratory syndrome

ATTAAAGGTTTACCTCCAGGTAAACAAACCAACTTCGCGAATTAAAATCTGTGGCTGTACTCGGCTGCATGCTTAGTAATTACTGTCGTTGACAGGACACGAGTAACTCGTATCTTCGTTGCAGCCGATCATCAGCACATCTAGGTTCGCCGGGTGACCG

Insert

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

Galaxy Australia

Workflow Visualize Shared Data Help User Run Tool Using 0%

Tools

mafft

Upload Data Show Sections

MAFFT Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)

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

Sequences to align *

4: samples.fasta

2: Input batch

Sequences to align *

3: (hidden) NC_045512.2

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

EET_NC_2 (soft progressive method)

History

search datasets

QC-CONSENSUS-ALIGN

334 kB 3 2 1

4: samples.fasta

Add Tags

5 sequences format fasta, database ?

uploaded fasta file

>hCoV-19/Philippines/PH-PGC-113124/2022

CCAACTTCGATCTTGTAGATCTGTTCTAACGACATTAA

GCACTCACGCGATAATTAAACTAATTACTGTCGTTGACAGG

ACGGTTTCGTCGTTGAGCCGATCATCAGCACATCTAGGTTT

TGTCCTGGTTCAACGAGAAAACACGTCCAATCAGTTGC

2: NCBI Accession Download on : Log

1: NCBI Accession Download on : Downloaded Files

a list with 1 fasta dataset

Run mafft



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

Galaxy Australia

Workflow Visualize Shared Data Help User Run Tool Using 0%

MAFFT Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)

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)

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 syndrome coronavirus 2

ATTAAGTTTATACTTCCAGGTAAACAAACCAACTTCGCGAACCTAAATCTGTGGCTCACTCGCTGCATGCTTAATTACTGCTGTTGACAGGACAGAGTAACTCGTCTATCTCTGTTGCAGCCGATCATCAGCACATCTAGGTTCTGCCGGTGTGACC

Check mafft job and download MSA



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

Galaxy Australia

Workflow Visualize Shared Data Help User 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

samples.fasta

Add Tags

5 sequences format fasta, database ?

uploaded fasta file

>hCoV-19/Philippines/PH-PGC-113124/2022

ccaaaccttcg
gttcttaacgaacttaaatctgtgtggctgactcgctg
cagcgtataattaaataacttaatctgtgtgacaggacacg
ttctgcaggcttacggttcgccgtgtgcagccgatcatc

>hCoV-19/Philippines/PH-PGC-113124/2022

CCAACTTCGATCTTGTAGATCTGTTCTAACGAACTTAA
GCACTCACGCGAGTATAATTAAACTAAATTACTGTCGTTGACAGG

Open MEGA



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

Molecular Evolutionary Genetics Analysis

The interface features a top navigation bar with twelve circular icons representing different analysis modules: ALIGN, DATA, MODELS, DISTANCE, DIVERSITY, PHYLOGENY, USER TREE, ANCESTORS, SELECTION, RATES, CLOCKS, and DISEASE. Below this is a large central workspace. To the right of the workspace are two external links: 'TIMETREE' with its logo and 'DATAMONKEY' with its logo. At the bottom, there is a 'RECENT PUBLICATIONS' section followed by a row of eight small icons for 'HELP DOCS', 'EXAMPLES', 'CITATION', 'REPORT BUG', 'UPDATES', 'MEGA LINKS', 'TOOLBAR', and 'PREFERENCES'. On the far right, there is a red button labeled 'ANALYZE' and a red square containing a white 'MxI' logo.

ALIGN DATA MODELS DISTANCE DIVERSITY PHYLOGENY USER TREE ANCESTORS SELECTION RATES CLOCKS DISEASE

TIMETREE

DATAMONKEY

RECENT PUBLICATIONS

HELP DOCS EXAMPLES CITATION REPORT BUG UPDATES MEGA LINKS TOOLBAR PREFERENCES

ANALYZE
PROTOTYPE

MxI

Align → Edit/Build Alignment



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

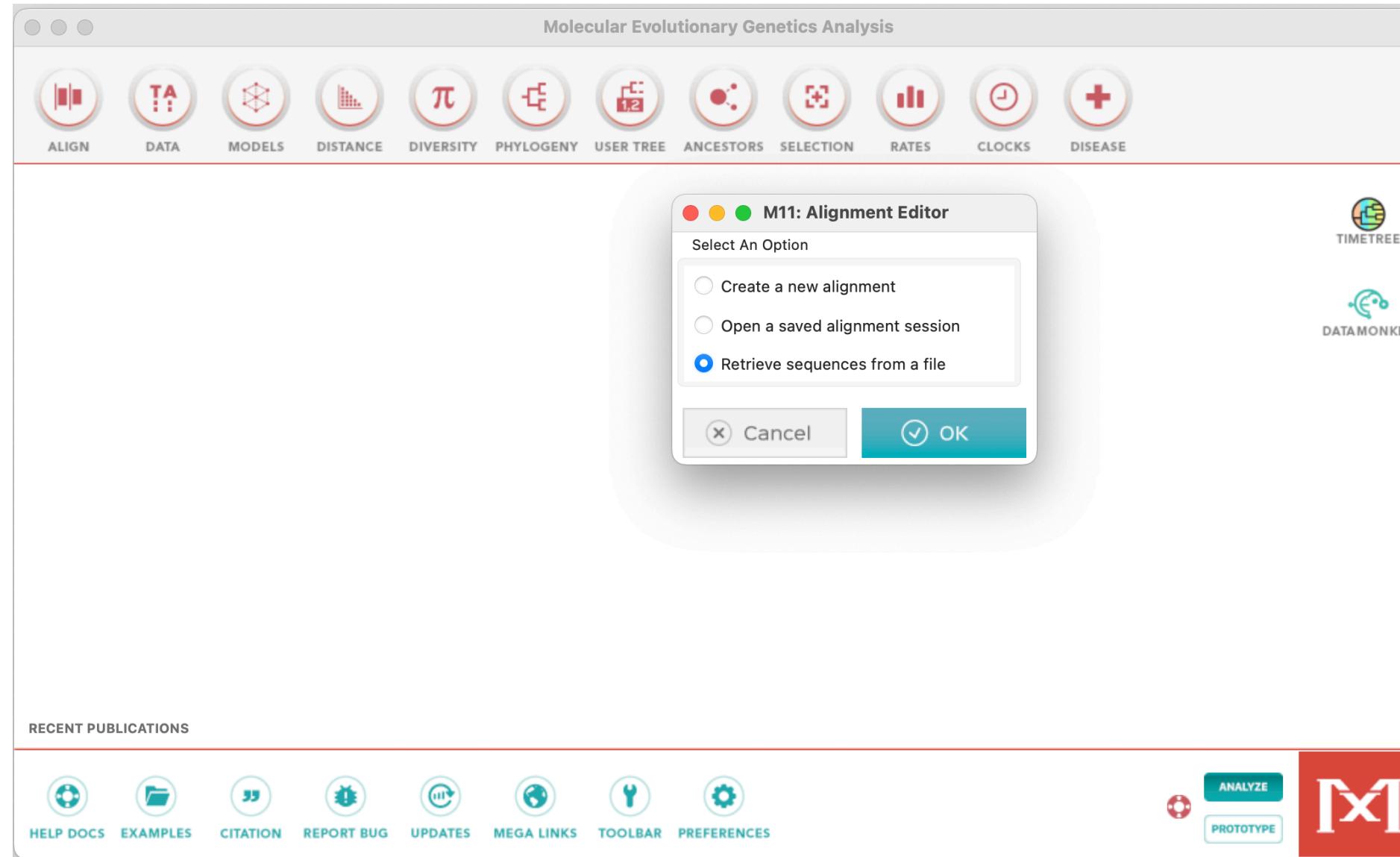
The screenshot shows the MEGA software interface. At the top, there is a toolbar with various icons: a flag, a DNA sequence, a network, a histogram, a pi symbol, a phylogenetic tree, a user tree, ancestors, selection, rates, clocks, and disease. Below the toolbar, a menu bar is visible with the title "Molecular Evolutionary Genetics Analysis". A dropdown menu for "Edit/Build Alignment" is open, containing the following options:

- Edit/View Sequencer Files (Trace)...
- Open Saved Alignment Session...
- Show Web Browser
- Query Databanks
- Do BLAST Search

On the right side of the interface, there are two additional tools shown as icons and names: "TIMETREE" and "DATAMONKEY".

At the bottom of the interface, there is a "RECENT PUBLICATIONS" section and a footer with links: HELP DOCS, EXAMPLES, CITATION, REPORT BUG, UPDATES, MEGA LINKS, TOOLBAR, PREFERENCES, ANALYZE, PROTOTYPE, and a large red "MEGA" logo.

Molecular Evolutionary Genetics Analysis



The MEGA software interface features a top navigation bar with twelve circular icons representing different analysis tools: ALIGN, DATA, MODELS, DISTANCE, DIVERSITY, PHYLOGENY, USER TREE, ANCESTORS, SELECTION, RATES, CLOCKS, and DISEASE. Below this is a large central workspace. A modal dialog box titled "M11: Alignment Editor" is displayed in the center, containing the text "Select An Option" and three radio button choices: "Create a new alignment", "Open a saved alignment session", and "Retrieve sequences from a file". The third option is selected. At the bottom of the dialog are "Cancel" and "OK" buttons. In the bottom right corner of the workspace, there is a red square icon with a white "x" and a white "I" symbol, labeled "ANALYZE" above "PROTOTYPE". The bottom navigation bar includes links for RECENT PUBLICATIONS, HELP DOCS, EXAMPLES, CITATION, REPORT BUG, UPDATES, MEGA LINKS, TOOLBAR, and PREFERENCES.

Open the downloaded MSA



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

Molecular Evolutionary Genetics Analysis

The screenshot shows the MEGA software interface. At the top, there is a toolbar with various icons and labels: ALIGN, DATA, MODELS, DISTANCE, DIVERSITY, PHYLOGENY, USER TREE, ANCESTORS, SELECTION, RATES, CLOCKS, and DISEASE. Below the toolbar, there is a large central area with a semi-transparent circular overlay containing a progress dialog. The dialog has a title "M11: Progress" and a "PROGRESS" bar. It includes two buttons: "DETAILS" and a blue "STOP" button with a red cross icon. Below these are tabs for "STATUS/OPTIONS" and "RUN STATUS". The "RUN STATUS" section displays "Start time" as "17-4-24 05:49:04" and "Status" as "Initializing Alignment Explorer...". At the bottom of the interface, there is a "RECENT PUBLICATIONS" section and a footer with links: HELP DOCS, EXAMPLES, CITATION, REPORT BUG, UPDATES, MEGA LINKS, TOOLBAR, PREFERENCES, ANALYZE, PROTOTYPE, and a large red "MEGA" logo.

ALIGN DATA MODELS DISTANCE DIVERSITY PHYLOGENY USER TREE ANCESTORS SELECTION RATES CLOCKS DISEASE

TIMETREE

DATA MONKEY

M11: Progress

PROGRESS

DETAILS STOP

STATUS/OPTIONS

RUN STATUS

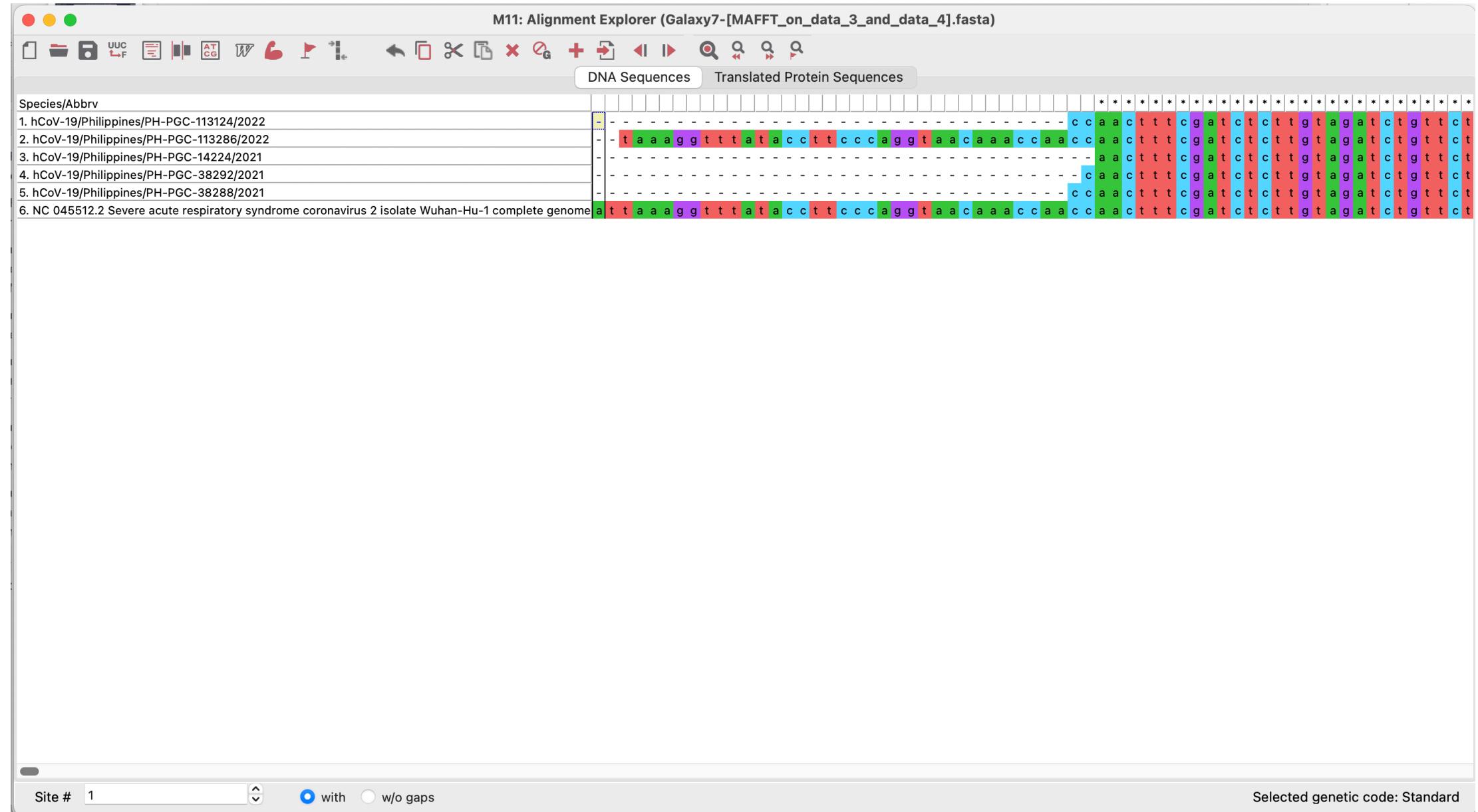
Start time 17-4-24 05:49:04

Status Initializing Alignment Explorer...

RECENT PUBLICATIONS

HELP DOCS EXAMPLES CITATION REPORT BUG UPDATES MEGA LINKS TOOLBAR PREFERENCES ANALYZE PROTOTYPE MEGA

Inspect alignment



Highlight leading sequences for all samples

M11: Alignment Explorer (Galaxy7-[MAFFT_on_data_3_and_data_4].fasta)

DNA Sequences Translated Protein Sequences

Species/Abbrv

- 1. hCoV-19/Philippines/PH-PGC-113124/2022
- 2. hCoV-19/Philippines/PH-PGC-113286/2022
- 3. hCoV-19/Philippines/PH-PGC-14224/2021
- 4. hCoV-19/Philippines/PH-PGC-38292/2021
- 5. hCoV-19/Philippines/PH-PGC-38288/2021
- 6. NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome

The screenshot shows a sequence alignment interface with six DNA sequences listed on the left. The sequences are aligned horizontally, with the first few bases of each sequence highlighted in yellow. To the right of the alignment, a translated protein sequence is shown, consisting of a series of amino acid codons represented by three-letter abbreviations. The background of the protein sequence area is color-coded in a repeating pattern of blue, red, green, and purple.

Site # 36 with w/o gaps Selected genetic code: Standard

Trim the selected block

M11: Alignment Explorer (Galaxy7-[MAFFT_on_data_3_and_data_4].fasta)

Species/Abbrv

- 1. hCoV-19/Philippines/PH-PGC-113124/2022
- 2. hCoV-19/Philippines/PH-PGC-113286/2022
- 3. hCoV-19/Philippines/PH-PGC-14224/2021
- 4. hCoV-19/Philippines/PH-PGC-38292/2021
- 5. hCoV-19/Philippines/PH-PGC-38288/2021
- 6. NC 045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1 complete genome

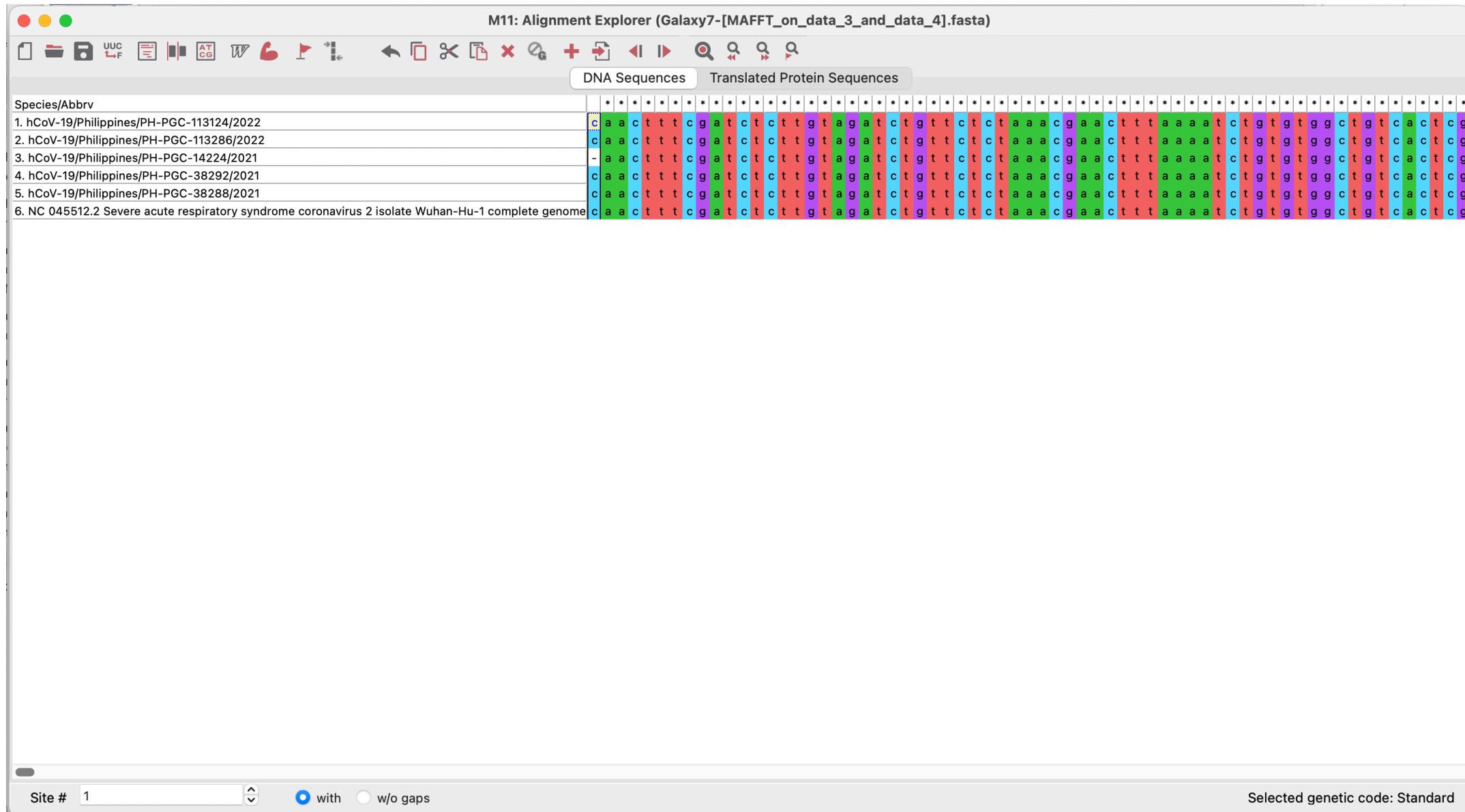
Cut the selected block (Meta+X) Sequences Translated Protein Sequences

1

2

Site # 36 with w/o gaps Selected genetic code: Standard

Inspect the sequence, again



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

1

2

3

4

The screenshot shows the Galaxy web interface for the ClipKIT tool. A red circle labeled '1' highlights the 'clipkit' tool entry in the 'Tools' dropdown menu. A red circle labeled '2' highlights the 'Alignment file' input field, which contains the value '7: MAFFT on data 3 and data 4'. A red circle labeled '3' highlights the 'gappyness' input field, which has the value '0.2' entered. A red circle labeled '4' highlights the 'Run Tool' button at the bottom of the tool configuration panel.

Tool Parameters

Alignment file *
7: MAFFT on data 3 and data 4

Select trimming mode from the list *
gappy

Select trimmed alignment output format from the list *
FASTA format

gappyness - optional
0.2
Specify gappyness threshold (between 0 and 1). Default: 0.9

create complement of the trimmed alignment
 No

Additional Options

Email notification
 No
Send an email notification when the job completes.

Modes

smart-gap: dynamic determination of gaps threshold
gappy: trim sites that are greater than the gaps threshold
kpic: keeps parsimony informative and constant sites
kpic-smart-gap: a combination of kpic- and smart-gap-based trimming
kpic-gappy: a combination of kpic- and gappy-based trimming
kpi: keep only parsimony informative sites
kpi-smart-gap: a combination of kpi- and smart-gap-based trimming
kpi-gappy: a combination of kpi- and gappy-based trimming

Gaps

Positions with gappyness greater than threshold will be trimmed. Must be between 0 and 1. (Default: 0.9). This argument is ignored when using the kpi and kpic modes of trimming as well as an iteration of trimming that uses smart-gap.

History

QC-CONSENSUS-ALIGN

7: MAFFT on data 3 and data 4

format fasta, database ?

nthread = 8
nthreadpair = 8
nthreaddb = 8

>hCoV-19/Philippines/PH-PGC-113124/2022

ccaaactttcg
gttctctaacgaactaaaaatctgtgtggctgtactcggctg
cacgcgatataatataactaaattactgtcggtgacaggacacg
ttctgcaggctgtttacggttcgttgcggatcatc

4: samples.fasta

format fasta, database ?

uploaded fasta file

>hCoV-19/Philippines/PH-PGC-113124/2022
CCAACTTCGATCTTGAGATCTGTTCTAAACGAACCTTAA
GCACGCAGCTATAATTAACAATTACTGTGTTGACAGG

Visualize alignment



Doherty
Institute



 The Royal
Melbourne
Hospital

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

The screenshot shows the Galaxy Australia web interface. The top navigation bar includes links for Home, Workflow, Visualize, Shared Data, Help, User, and various system icons. A user icon indicates 'Using 0%'.

Tools section:

- clipkit

Buttons: Upload Data, Show Sections.

ClipKIT. Alignment trimming software for phylogenetics.

WORKFLOWS section:

- All workflows

A red circle labeled '2' highlights a box containing the following text:

You can display your dataset with the following links:
1. [display with IGV \(local \)](#)

or select a visualization from below.

search visualizations

Editor: Manually edit text

Multiple Sequence Alignment: The MSA viewer is a modular, reusable component to visualize large MSAs interactively on the web.

A red circle labeled '1' highlights the first step of a workflow visualization:

6.93 MB 6 16 1

23: Clipkit log. Add Tags 2 sequences format fasta, database ?

22: Trimmed alignment. Add Tags 6 sequences format fasta, database ?

| Arguments |

1: MAFFT on data 3 and data 4 Add Tags 6 sequences format fasta, database ?

nthread = 8
ntheadpair = 8
ntheadtb = 8

7: MAFFT on data 3 and data 4 Add Tags 6 sequences format fasta, database ?

nthread = 8
ntheadpair = 8
ntheadtb = 8

Check for trimmed sequences



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

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Tools

clipkit

Upload Data

Show Sections

ClipKIT. Alignment trimming software for phylogenetics.

WORKFLOWS

All workflows

Import Sorting Filter Selection Vis.elements Color scheme Extras Export Help

ID Label . 2 . 4 . 6 . 8 . 10 . 12 . 14 . 16 . 18 . 20 . 22 . 24 . 26 . 28 . 30 . 32 . 34 . 36 . 38 . 40 . 42 . 44 . 46 . 48 . 50 . 52 . 54 . 56 . 58 .
1 hCoV-19/Philippines C A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T
2 hCoV-19/Philippines C A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T
3 hCoV-19/Philippines - A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T
4 hCoV-19/Philippines C A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T
5 hCoV-19/Philippines C A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T
6 NC_045512.2 C A A C T T T C G A T C T C T T G T A G A T C T G T T C T C T A A A C G A A C T T T A A A A T C T G T G T G G C T G T

History

search datasets

QC-CONSENSUS-ALIGN

6.93 MB 6 16 1

23: Clipkit log.

22: Trimmed alignment.

Add Tags

6 sequences format fasta, database ?

| Arguments |

hCoV-19/Philippines/PH-PGC-113124/2022
caacttcgatcttttagatctgttctaaacgaaacttaaa
actcggctgcatttttagtcactcacgcagataatataact
caggacacgagaactcgatctatcttcgcaggctgttacgg
ccgatcatcgacatcttagttttgtccgggtgtgaccgaaagg

7: MAFFT on data 3 and data 4

Add Tags

6 sequences format fasta, database ?

nthread = 8
nthreadpair = 8
nthreadtb = 8

Use nextclade to get indiv QC scores

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Tools

next

Upload Data

Show Sections

1 **Nextclade** Viral genome clade assignment, mutation calling, and sequence quality checks (Galaxy Version 2.7.0+galaxy0)

Tool Parameters

2 FASTA file with input sequences *

4: samples.fasta

(--input-fasta)

Organism *

3 SARS-CoV-2

Version of database to use

Download latest available database version from web

Output options - optional

Tabular format report

Output reports and optionally tree

Include header line in output file

No

Use advanced options

No

Additional Options

Email notification

No

5 Run Tool

History

deleted:false visible:any

QC-CONSENSUS-ALIGN

6.94 MB

23: Clipkit log.

22: Trimmed alignment.

Add Tags

6 sequences

format fasta, database ?

| Arguments |

>hCoV-19/Philippines/PH-PGC-113124/2022

caacttcgatctttagatctgtttctaaacgaactttaaa

actcggtgcgtttagtgactcagcgtataatataact

cggacacgagaatcgttatcttcgtcaggctgttaccgtt

ccgatcatcgcacatctaggtttgtccgggtgtaccgaaagg

7: MAFFT on data 3 and data 4

Add Tags

6 sequences

format fasta, database ?

nthread = 8

nthreadpair = 8

nthreadtb = 8

Check Nextclade results



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

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Tools

next

Upload Data

Show Sections

Nextclade Viral genome clade assignment, mutation calling, and sequence quality checks (Galaxy Version 2.7.0+galaxy0)

Tool Parameters

FASTA file with input sequences *

4: samples.fasta

(--input-fasta)

Organism *

SARS-CoV-2

Version of database to use

Download latest available database version from web

Output options - optional

Tabular format report

Output reports and optionally tree

Include header line in output file

No

Use advanced options

No

Additional Options

Email notification

No

Send an email notification when the job completes.

Run Tool

History

deleted:false visible:any

QC-CONSENSUS-ALIGN

6.94 MB

7 16 1

24: Nextclade on data 4 (TSV report)

Add Tags

5 lines

format tabular, database ?

1. seqName 2. clade

hCoV-19/Philippines/PH-PGC-113124/2022	211
hCoV-19/Philippines/PH-PGC-113286/2022	21A
hCoV-19/Philippines/PH-PGC-14224/2021	20B
hCoV-19/Philippines/PH-PGC-38292/2021	20B
hCoV-19/Philippines/PH-PGC-38288/2021	20B

23: Clipkit log.

22: Trimmed alignment.

Add Tags

6 sequences

format fasta, database ?

>hCoV-19/Philippines/PH-PGC-113124/2022

View tabular format using your favorite spreadsheet viewer



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



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

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 <code>N</code>)	non-negative integer	238
totalNonACGTNs	Total number of detected ambiguous nucleotides (nucleotide characters that are not <code>A</code> , <code>C</code> , <code>G</code> , <code>T</code> , <code>N</code>)	non-negative integer	2
totalUnknownAa	Total number of unknown aminoacids (aminoacid character <code>x</code>)	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



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



Create workflow



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

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

Workflow name

Workflow constructed from history 'QC-CONSENSUS-ALIGN'

Create Workflow **Check all** **Uncheck all**

Tool

NCBI Accession Download
 Include "NCBI Accession Download" in workflow

Data Fetch
This tool cannot be used in workflows

MAFFT
 Include "MAFFT" in workflow

ClipKIT. Alignment trimming software for phylogenetics.
 Include "ClipKIT. Alignment trimming software for phylogenetics." in workflow

Nextclade
 Include "Nextclade" in workflow

History items created

1 NCBI Accession Download on : Downloaded Files

2 NCBI Accession Download on : Log

4 samples.fasta
 Treat as input dataset samples.fasta

7 MAFFT on data 3 and data 4

22 Trimmed alignment.

23 Clipkit log.

24 Nextclade on data 4 (TSV report)

History

deleted:false visible:any

QC-CONSENSUS-ALIGN

6.94 MB 7 16 1

24: Nextclade on data 4 (TSV report)
Add Tags
5 lines
format tabular, database ?

1.seqName 2.clade
hCoV-19/Philippines/PH-PGC-113124/2022 21I
hCoV-19/Philippines/PH-PGC-113286/2022 21A
hCoV-19/Philippines/PH-PGC-14224/2021 20B
hCoV-19/Philippines/PH-PGC-38292/2021 20B
hCoV-19/Philippines/PH-PGC-38288/2021 20B

23: Clipkit log.

22: Trimmed alignment.
Add Tags
6 sequences
format fasta, database ?

| Arguments |

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



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

Galaxy Australia

Workflow Visualize Shared Data Help User

Using 0%

Workflow constructed from history 'QC-CONSENSUS-ALIGN'

MAFFT
Multiple alignment program for amino acid or nucleotide sequences (Galaxy Version 7.520+galaxy0)

Input batch

1: Input batch

Sequences to align *

Data input 'inputs' (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

Sequences to align *

Data input 'inputs' (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

+ 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

Tools

search tools

Inputs

FILE AND META TOOLS

Get Data

Send Data

Collection Operations

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

VCF/BCF

Nanopore

Convert Formats

Lift-Over

COMMON GENOMICS TOOLS

Operate on Genomic Intervals

MiModD

Fetch Alianments/Sequences

samples.fasta

output (input)

NCBI Accession Download

output (input)

error_log (txt)

MAFFT

Input batch 1 > Sequences to align

Input batch 2 > Sequences to align *

Input batch 3 > Sequences to align *

outputAlignment (fasta, clustal, maf, xmfa, stockholm, phylip)

ClipKIT

Alignment trimming software for phylogenetics.

Alignment file

trimmed_output (fasta, clustal, maf, xmfa, stockholm, phylip)

log_output (txt)

Nextclade

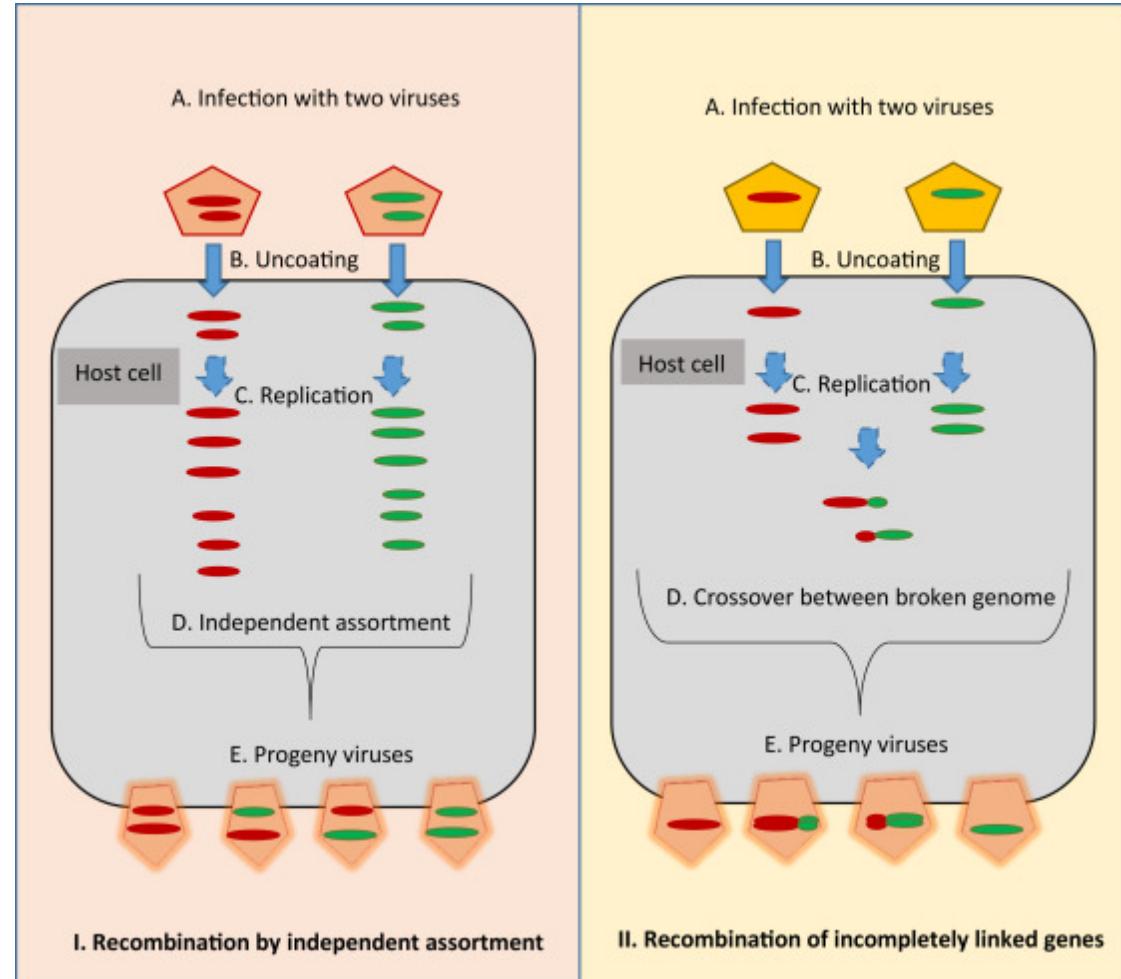
FASTA file with input sequences

report_tsv (tabular)

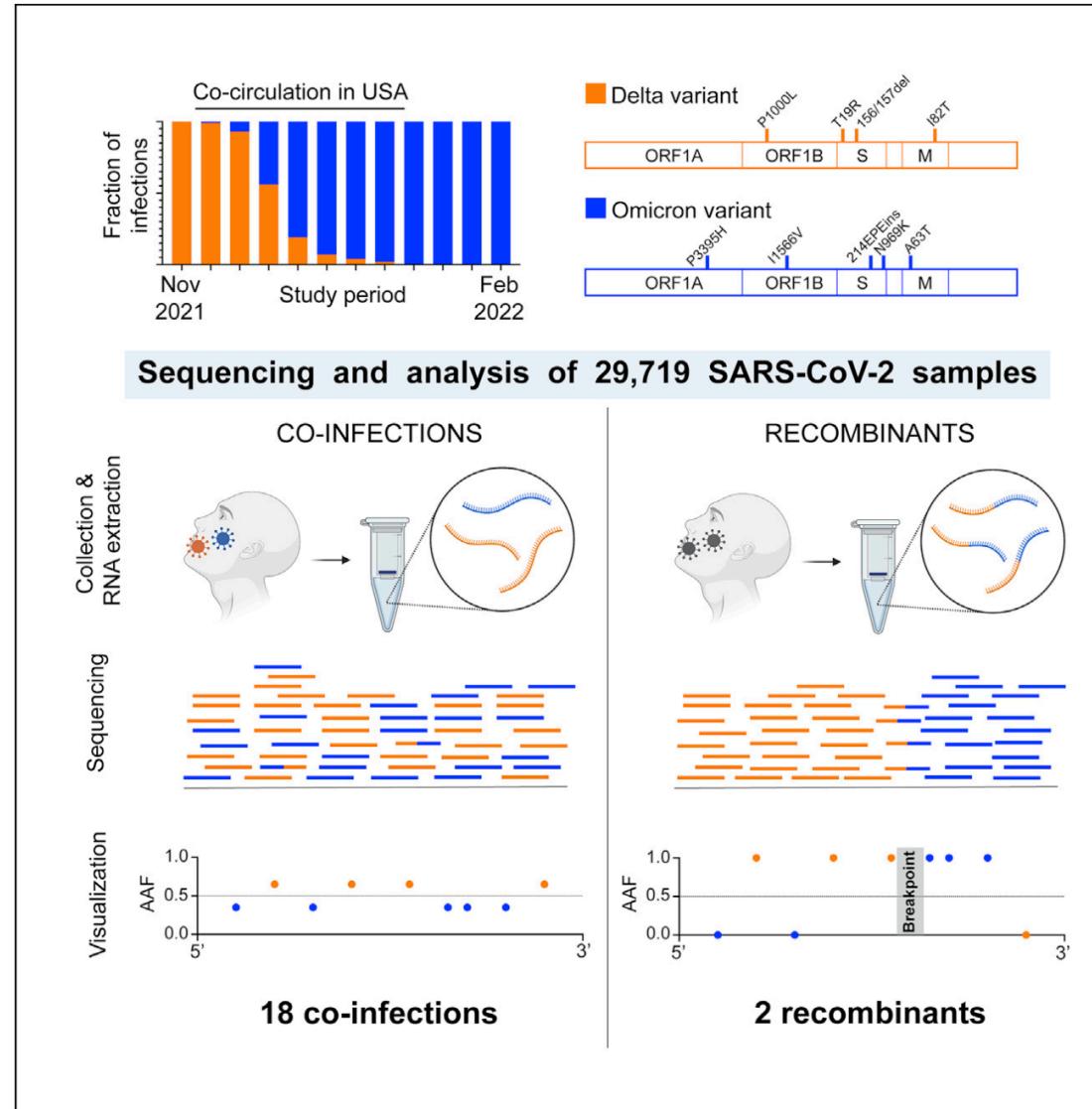
Workflow diagram showing the 'QC-CONSENSUS-ALIGN' history. The workflow starts with 'samples.fasta' (output of step 1) as input for 'MAFFT' (step 3). 'MAFFT' also receives inputs from 'NCBI Accession Download' (step 2) and 'ClipKIT' (step 5). 'NCBI Accession Download' receives 'samples.fasta' as input. 'ClipKIT' receives 'outputAlignment' from 'MAFFT' and 'FASTA file with input sequences' from 'Nextclade' (step 4). 'Nextclade' receives 'outputAlignment' from 'MAFFT' and 'FASTA file with input sequences' from 'ClipKIT'. The 'Type of sequences' dropdown is set to 'auto-detect'.

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.



Case Study: Recombination vs Co-infection



Case Study: Recombination vs Co-infection



A Potential recombinants between ['Delta / 21I', 'Omicron / BA.2 / 21L']:

coordinates

234455999990000112345567890112222222222233333333444555666666777777788888899

2267013153458801444580447441906922566667788999000004556894440450257782333678234888457

1479382884236929441181051616551802777887818199145670290541260686770655888350716888014

0100741444446198798409814603558707849685632725305353594480490940079789234827111123202

genes

ref

Delta / 21I

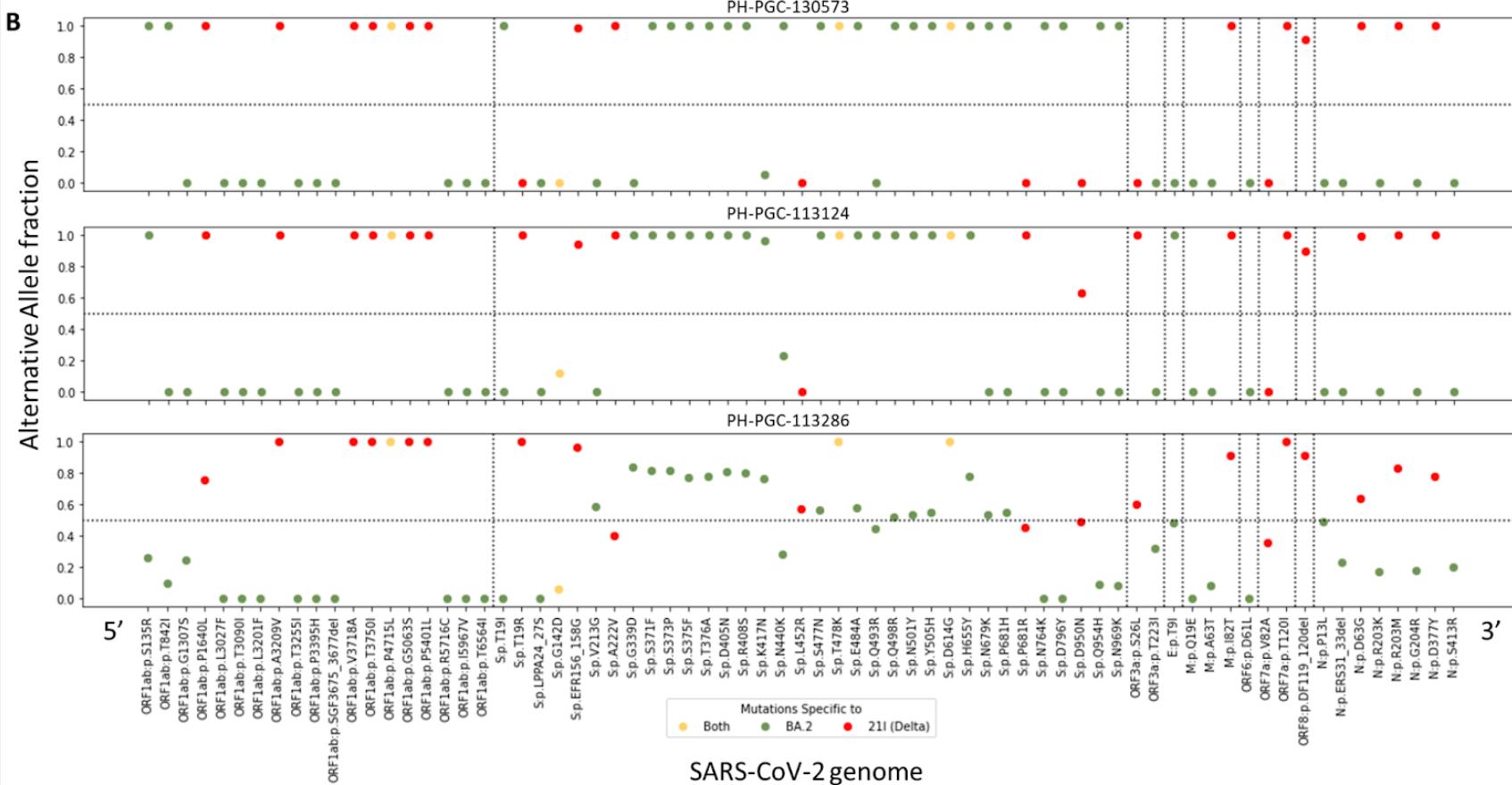
Omicron / BA.2 / 21L

consensus

PH - PGC - 113124 | 2022-03-20

PH-PGC-113286 | 2022-03-31

Made with Sc2rf - available at <https://github.com/lenaschimmel/sc2rf>



Tools for detecting recombination



CENTRE FOR
PATHOGEN
GENOMICS



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

- 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



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



The Royal
Melbourne
Hospital