CoV-GLUE User Guide

Version 0.1.8, 11th May 2020, Dr. Josh Singer (josh.singer@glasgow.ac.uk)
MRC-University of Glasgow Centre for Virus Research

Introduction

CoV-GLUE is a publicly-accessible web application for the interpretation and analysis of hCoV-19 virus (also known as SARS-CoV-2) genome sequences, located at http://cov-glue.cvr.gla.uk . This document is aimed at users of the web application, typically researchers working on the virus, public health laboratory staff or diagnostic laboratory staff.

CoV-GLUE is developed within and funded by the COVID-19 Genomics UK Consortium. CoV-GLUE is based on the GLUE software framework. GLUE and CoV-GLUE are developed and maintained by the MRC-University of Glasgow Centre for Virus Research. CoV-GLUE is enabled by data from GISAID, see Elbe et al., 2017.

Functionality overview

CoV-GLUE provides two broad areas of functionality. Firstly it allows users to browse via the web a database of amino acid replacements and coding region indels that have been observed in sequences from the pandemic. Secondly, it allows users to analyse their own hCoV-19 sequences by submitting them to the web application to receive an interactive report.

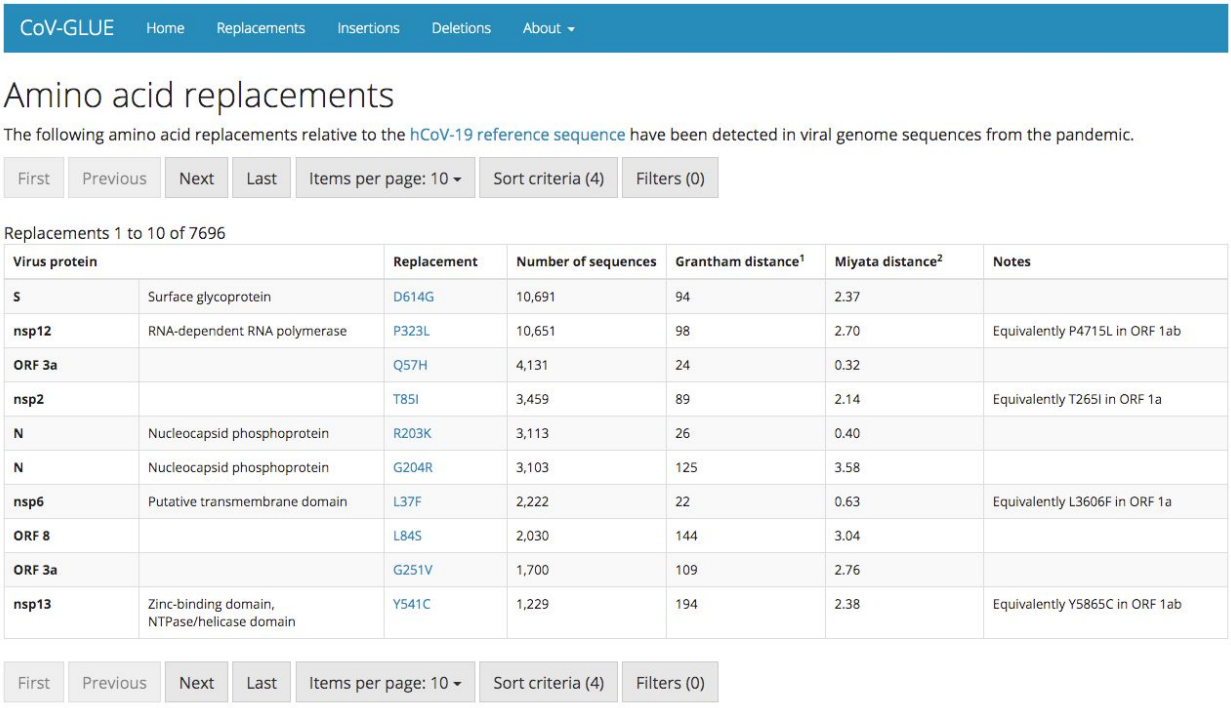
Changes within the virus genome are defined in CoV-GLUE relative to Wuhan-Hu-1 (GenBank accession MN908947.3). This was one of the first strains to be sequenced and is close to being ancestral to all pandemic strains. CoV-GLUE does not claim that Wuhan-Hu-1 is ancestral, rather its sequence is used as a reference.

The replacement and indel databases are created by regularly analysing genomic data from GISAID, see Shu et al., 2017. Some sequences are excluded; see the relevant section below for details. A global reference phylogenetic tree which captures the main lineages of the pandemic is used for some analysis, there is also a section below containing details of this tree.

Amino acid replacements database

CoV-GLUE analyses changes within viral proteins which arise as a consequence of changes in the virus genome. If you are unfamiliar with the hCoV-19 open reading frames and viral proteins, read the short explainer towards the end of this document. Amino acid replacements are denoted by 4 parameters: the viral protein, the amino acid present in the Wuhan-Hu-1 reference sequence, the number of the codon within Wuhan-Hu-1 and the replacement amino acid. For example, a sequence X may contain amino acid replacement S:D614G. S refers to the Spike protein. Wuhan-Hu-1 contains an Aspartic Acid (D) translated from codon 614 of the Spike coding region, in contrast sequence X contains a Glycine (G) translated from the homologous codon. Since non-structural proteins are products of a polyprotein, a change such as nsp12 replacement P323L is equivalent to ORF 1ab polyprotein replacement P4715L.

The replacements database may be accessed by clicking "Replacements" in the menu bar at the top of the application. The page displays 10 replacements by default, this can be changed using the "Items per page" button. You can browse through more pages by clicking "First", "Previous", "Next" or "Last".

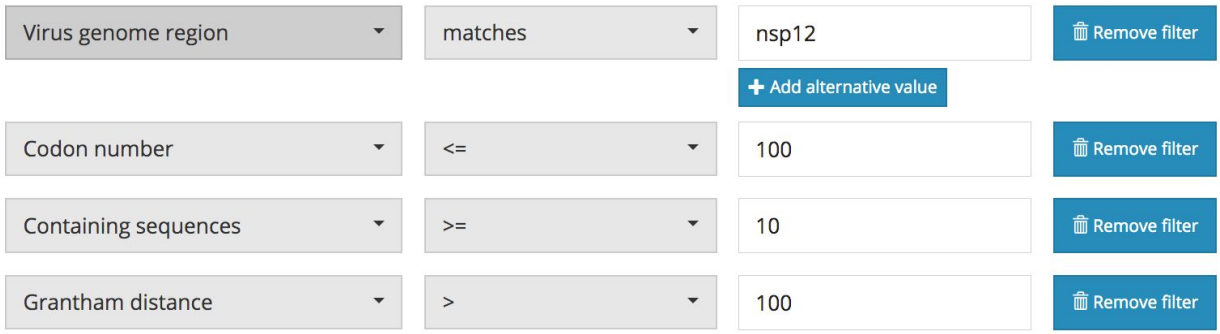


The number of sequences which contain each replacement is given, this is also the default sort criterion, so that the most frequent replacement is displayed at the top of the table.

Each replacement is also categorised by two standard metrics from the literature: Grantham distance and Miyata distance. These aim to provide a quick summary of the magnitude of the change in terms of protein structure and/or function. Grantham distance is based on

composition, polarity and molecular volume, with 100 being average, see Grantham, 1974. Miyata distance is based on volume and polarity, see Miyata et al., 1979.

The view of the replacements database can be filtered by clicking the "Filters" button. This brings up a dialog box. User defined filters may be applied, an example is given below.



Here we have filtered the view to only include replacements in the first 100 codons of the region coding for the nsp12 viral protein. Each replacement must have been observed in at least 10 sequences and must have a Grantham distance at least 100. You can also change the ordering of the table using the "Sort criteria" button, for example ordering by descending Miyata distance.

Replacement pages

The "Replacement" column of the table contains links, clicking one will take you to a page which provides more details and analysis for that specific replacement. The example below is for nsp12 replacement D63Y. The replacement page has three alternative tabs. The first tab "Containing sequences" provides a table of pandemic sequences in which the replacement was detected. Similar to other tables in the web application, you can page through it, and apply custom filters and sort criteria. For example you can filter the sequences by country.

nsp12 replacement D63Y

nsp12: RNA-dependent RNA polymerase
Equivalently D4455Y in ORF 1ab
Nucleotide position 13627 in the hCoV-19 reference sequence

Containing sequences | Phylogenetic tree | Replacement analysis

The following pandemic sequences contain nsp12 replacement D63Y

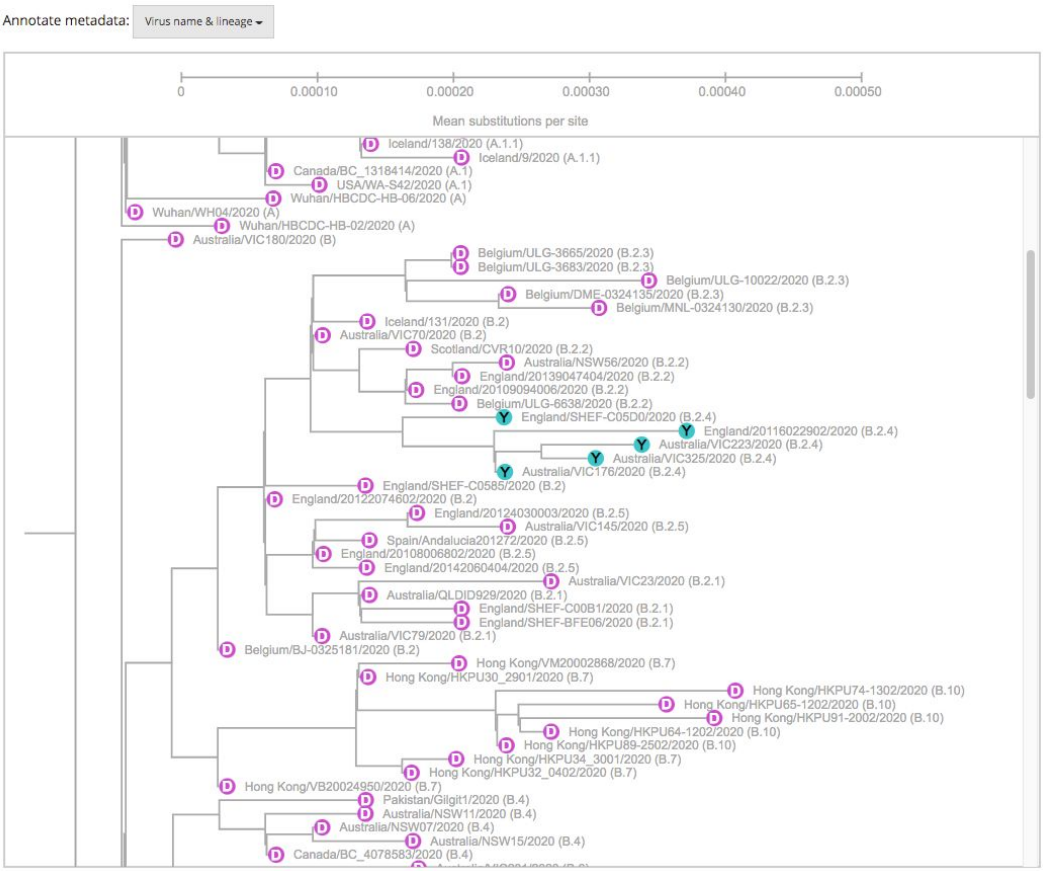
First | Previous | Next | Last | Items per page: 10 | Sort criteria (1) | Filters (0)

Sequences 1 to 10 of 59

Virus name	GISAID ID	Lineage	Country	Location	Collection date
Taiwan/TSGH-17/2020	EPI_ISL_436102	B.2.4 (99.28%)	Taiwan (TWN)	Asia / Taiwan / New Taipei City	19-MAR-2020
England/NOTT-10F2DD/2020	EPI_ISL_432988	B.2.4 (99.50%)	United Kingdom (GBR)	Europe / United Kingdom / England	1-APR-2020
Wales/PHWC-27935/2020	EPI_ISL_432270	B.2.4 (99.49%)	United Kingdom (GBR)	Europe / United Kingdom / Wales	1-APR-2020
Australia/VIC966/2020	EPI_ISL_430653	B.2.4 (99.73%)	Australia (AUS)	Oceania / Australia / Victoria	27-MAR-2020
Australia/VIC968/2020	EPI_ISL_430651	B.2.4 (99.73%)	Australia (AUS)	Oceania / Australia / Victoria	27-MAR-2020
Uruguay/UY-10/2020	EPI_ISL_429257	B.2.4 (99.99%)	Uruguay (URY)	South America / Uruguay / Montevideo	19-MAR-2020
Australia/NSW175/2020	EPI_ISL_427687	B.2.4 (99.99%)	Australia (AUS)	Oceania / Australia / New South Wales / Sydney	18-MAR-2020
Australia/NSW174/2020	EPI_ISL_427668	B.2.4 (99.99%)	Australia (AUS)	Oceania / Australia / New South Wales / Sydney	21-MAR-2020
Australia/VIC887/2020	EPI_ISL_427140	B.2.4 (99.19%)	Australia (AUS)	Oceania / Australia / Victoria	6-APR-2020
Australia/VIC808/2020	EPI_ISL_427082	B.2.4 (99.92%)	Australia (AUS)	Oceania / Australia / Victoria	28-MAR-2020

First | Previous | Next | Last | Items per page: 10 | Sort criteria (1) | Filters (0)

The second tab "Phylogenetic tree" allows you to view the relationships between the replacement and the main lineages of the pandemic. This tab displays CoV-GLUE's global reference tree with colour annotations representing the amino acid at the replacement codon for each taxon. Below we can see that nsp12 replacement D63Y is associated with the taxa of lineage B.2.4. By default each tree taxon is annotated with the virus name and lineage. You can change what text is displayed by clicking the drop down next to "Annotate metadata".



The third tab "Replacement analysis" gives the Grantham and Miyata distances for the replacement along with a third classification based on Hanada et al., 2006.

Indels database

While single nucleotide changes are the most common form of mutation, virus genome replication may occasionally result in deleted sections of nucleotides or, even more rarely, inserted sections. These insertions and deletions are known as indels. CoV-GLUE provides access to a database of indels that have been observed in pandemic sequences. You can access these by clicking on "Insertions" or "Deletions" in the application menu bar. Screenshots of these tables are given below. The view of these tables may be sorted and filtered as with the replacements table.

Insertions 1 to 5 of 5

Virus protein		Inserted nucleotides	Codon-aligned?	Inserted amino acids	Number of sequences	Notes
nsp6	Putative transmembrane domain	11074-TTT-11075	Yes	34-F-35	23	Equivalently ORF 1a 3603-F-3604
nsp12	RNA-dependent RNA polymerase	14605-TCCTTA-14606	No	N/A	1	
nsp16	2'-O-ribose methyltransferase	21384-TTC-21385	Yes	242-F-243	1	Equivalently ORF 1ab 7040-F-7041
S	Surface glycoprotein	22304-CCCACCAGA-22305	No	N/A	1	
S	Surface glycoprotein	22353-TTA-22354	No	N/A	1	

First Previous Next Last Items per page: 10 Sort criteria (2) Filters (0)

Deletions 1 to 10 of 69

Virus protein		Deleted nucleotides	Codon-aligned?	Deleted codons	Number of sequences	Notes
nsp2		1605-1607	No	N/A	405	
nsp1	Leader protein	686-694	Yes	141-143	36	Equivalently ORF 1a codons 141-143
nsp1	Leader protein	515-520	Yes	84-85	12	Equivalently ORF 1a codons 84-85
S	Surface glycoprotein	21991-21993	No	N/A	7	
nsp1	Leader protein	508-522	No	N/A	6	
nsp1	Leader protein	518-520	Yes	85	5	Equivalently ORF 1a codons 85
ORF 8		28090-28095	No	N/A	5	
nsp1	Leader protein	510-518	No	N/A	4	
nsp1	Leader protein	516-518	No	N/A	4	
nsp1	Leader protein	669-671	No	N/A	4	

First Previous Next Last Items per page: 10 Sort criteria (2) Filters (0)

CoV-GLUE only detects deletions strictly within coding regions. A 382nt deletion which spans ORF 7b and ORF 8 has been confirmed in sequences from Singapore, but such deletions are not yet included in CoV-GLUE's database. Similarly, indels in the untranslated regions of the genome are not included.

Indels in CoV-GLUE are primarily referred to by their nucleotide coordinates. These are based on Wuhan-Hu-1. So, if a sequence X contains deletion 1605-1607, this means that the nucleotides in X that are homologous to nucleotides 1605,1606 and 1607 in Wuhan-Hu-1 are absent from X. Similarly, if a sequence Y contains insertion 21384-TTC-21385, this means that Y contains the nucleotides TTC inserted between the nucleotides which are homologous to nucleotides 21384 and 21385 in Wuhan-Hu-1.

CoV-GLUE also considers the effect of indels on viral proteins. If the inserted string of nucleotides is of a length that is not a multiple of 3, this is considered a frameshifting indel. It may be a biologically implausible sequencing error. If not an error, it may invalidate prediction of amino acid replacements for the sequence. Consequently, the presence of a frameshifting indel causes a sequence to be excluded from the analysis.

Non-frameshifting indels may be codon-aligned or non-codon-aligned, indicated by columns in the insertion and deletion tables.

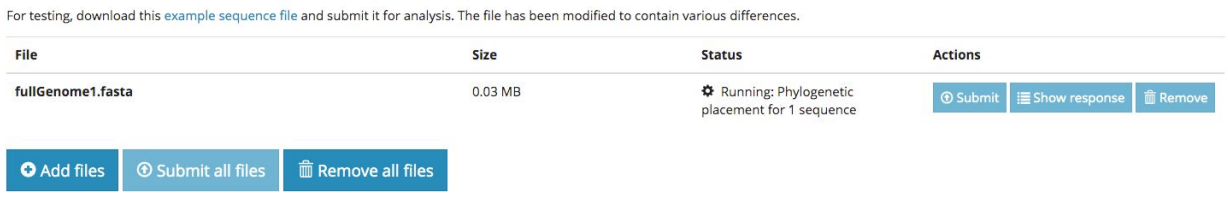
A codon-aligned insertion is between nucleotides at the end of one codon and the start of the next. Where insertions are codon-aligned, it is possible to refer to them in terms of the inserted amino acids and the numbers of adjacent codons, hence the codon-aligned nsp6 nucleotide insertion 11074-TTT-11075 may also be referred to as the nsp6 amino acid insertion 34-F-35.

A deletion is codon-aligned when the first and last deleted nucleotides lie at the start and end of codons. The codon-aligned nsp1 nucleotide deletion 686-694 is also the nsp1 amino acid deletion 141-143.

As in the replacements database, the insertion and deletion tables provide clickable links to pages for individual indels. These pages contain tabs listing the sequences containing the indel, and allowing visualisation of the indel on the global reference tree.

Analysis of user-submitted sequences

CoV-GLUE allows users to submit their own sequence data for rapid analysis. This is done by uploading a FASTA consensus sequence file. Click on "Home" in the application menu bar. Click "Add files" (see screenshot below) to add one or more FASTA files for upload, or alternatively drag them from your desktop to this part of the browser window. Each FASTA file may contain multiple sequences, but there is a limit of 50 sequences for each file. You can also use an example sequence file, hyperlinked from the same page.



Once a file has been added, click "Submit" in the "Actions" column. You will see a series of messages in the "Status" column for that file as the analysis of the file progresses. Once it displays "Complete", the "Show response" button will become enabled. Clicking this displays the analysis report for the file. This report will remain accessible until "Remove" is clicked or the browser window is closed. Note that submitted sequences are not stored by CoV-GLUE after the analysis is performed.

Phylogenetic classification

Phylogenetic classification is one key output of the analysis. If the sequence is not recognised as hCoV-19, this is indicated and no further analysis is performed. The lineage of the sequence

within the pandemic (see Rambaut, et al. 2020) is detected using a maximum-likelihood method. "Total LWR" is very approximately the confidence level that the sequence belongs to this lineage.

Analysis of sequence file 'fullGenome1.fasta'

Summary | Genome visualisation | Phylogenetic placement

Sequence	Classification			Primer/probe analysis			Differences from reference
	SARS-CoV-2?	Lineage	Total LWR	Diagnostics issues	Sequencing issues	Full report	
Sequence1	Yes	B	95.69%	0	1	View	Known amino acid replacement in ORF 3a: G251V (Tree) Novel amino acid replacement in M: R44M (Tree) Known codon-aligned insertion in nsp6: 34-F-35 (Tree) Novel codon-aligned insertion in S: 122-VK-123 (Tree) Known non-codon-aligned deletion in nsp2: nucleotides 1605-1607 (Tree) Novel frameshifting deletion in ORF 6: nucleotide 27361 (Tree) Novel codon-aligned deletion in N: codons 62-65 (Tree)

Differences from reference

The analysis reports coding region replacements and indels in the submitted sequences. If the replacement or indel has been observed in pandemic sequences, you can click "Known" to open the relevant page from the CoV-GLUE database. The tree link opens the "Phylogenetic placement" tab and annotates the selected change on the tree.

Primer/probe analysis

Each sequence is also analysed against published primer/probe schemes for diagnostic and sequencing assays. The aim here is to detect sequence polymorphisms which may reduce the effectiveness of these assays. The full primer/probe analysis report may be accessed by clicking "View" in the "Full report" column on the "Summary" tab. An example report is shown below.

hCoV-19/SARS-CoV-2 Primer/probe analysis report

File path	fullGenome1.fasta
Query sequence ID	Sequence1
Report generation date	11/05/2020
GLUE engine version	1.1.94
CoV-GLUE project version	0.1.7
Reporting software author	Josh Singer <josh.singer@glasgow.ac.uk>

Overview

This bioinformatics procedure analyses an hCoV-19/SARS-CoV-2 genomic "query sequence" to compare it with a range of published primer/probe designs for sequencing and diagnostics assays.

Any mismatches, insertions or deletions detected could suggest that the sequenced viral strain may fail to bind to the specific primer/probe, and so the associated assay may be ineffective or suboptimal. However, please do bear in mind that the reported issues are based purely on a simple sequence comparison, so further investigation of primer and probe properties (e.g. melting temperature, secondary structure, complementarity) and laboratory investigation of assay efficacy is advised.

Where coverage / alignment issues are reported, this indicates N characters in the primer/probe region, or that the query sequence does not fully cover the primer/probe region, or that there has been a problem in the alignment step of the procedure.

Genome reference coordinates are based on the Wuhan-Hu-1 strain, GenBank accession MN908947.3.

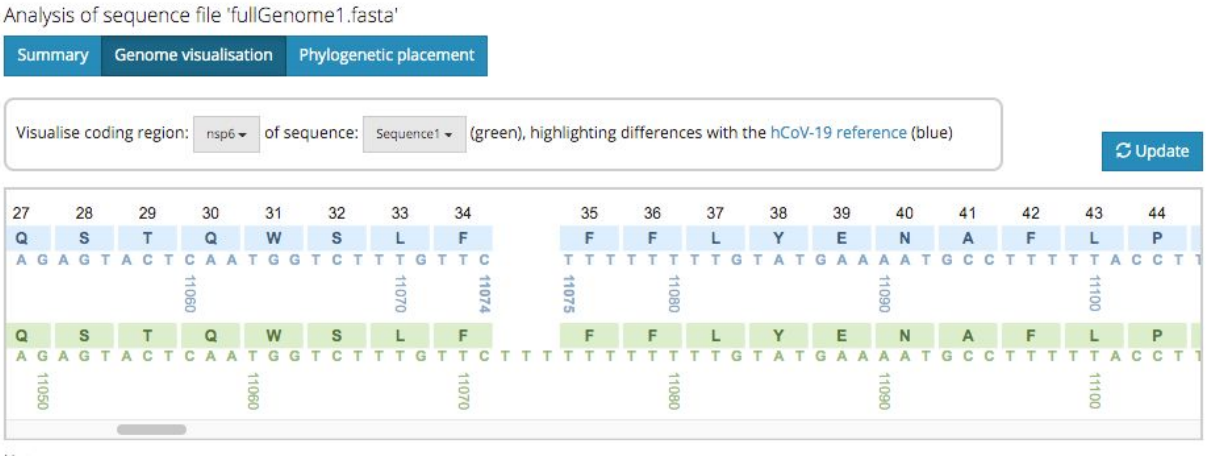
Assay reports

Publication	Assay	Purpose	Primer/probe	Primer/probe sequence	Location on reference	Query sequence issues
ARTIC Network	nCoV-2019 nanopore primers V3	Whole genome sequencing	nCoV-2019_93_RIGHT	CTCAACATGGCAAGGAAGACCT	29443-29464	1 deletion: 29457-29468
China CDC Primers and probes for detection 2019-nCoV	N	Amplification for diagnostics	No issues detected			
China CDC Primers and probes for detection 2019-nCoV	ORF1ab	Amplification for diagnostics	No issues detected			
Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (HKU)	HKU_N	Amplification for diagnostics	No issues detected			
Detection of 2019 novel coronavirus (2019-nCoV) in suspected human cases by RT-PCR (HKU)	HKU_ORF1b-nsp14	Amplification for diagnostics	No issues detected			
Diagnostic detection of Wuhan coronavirus 2019 by real-time RT-PCR – Charité, Berlin Germany	E_Sarbeco	Amplification for diagnostics	No issues detected			
Diagnostic detection of Wuhan coronavirus 2019 by real-time RT-PCR – Charité, Berlin Germany	RdRP_SARSr	Amplification for diagnostics	RdRP_SARSr-P1	CCAGGTGGWACRTCATCMGGTGATGC	15469-15494	2 mismatches: R15480C*, T15488A*
			RdRP_SARSr-R1	TATGCTAATAGTGTSTTTAACATYTG	15505-15530	1 mismatch: S15519T*
PCR and sequencing protocol for 2019-nCoV - Ministry of Public Health, Thailand	WH-NIC N	Amplification for diagnostics	No issues detected			
PCR and sequencing protocol for 2019-nCoV - National Institute of Infectious Diseases, Japan	NIID_2019-nCOV_N	Amplification for diagnostics	NIID_2019-nCOV_N_R2	GTTGACCTACACAGCTGCCA	29263-29282	1 mismatch: C29277G*
US CDC panel primer and probes – U.S. CDC, USA	2019-nCoV_N1	Amplification for diagnostics	No issues detected			
US CDC panel primer and probes – U.S. CDC, USA	2019-nCoV_N2	Amplification for diagnostics	No issues detected			
US CDC panel primer and probes – U.S. CDC, USA	2019-nCoV_N3	Amplification for diagnostics	No issues detected			
Wuhan coronavirus (2019-nCoV) real-time RT-PCR N gene 2020	Wuhan-TM2020	Amplification for diagnostics	No issues detected			

* This is a known issue relating to the primer/probe design rather than a sequence polymorphism.

Genome visualisation

The "Genome visualisation" tab of the analysis report allows you to visualise in detail the submitted sequence coding region nucleotides and protein translation (in green) relative to Wuhan-Hu-1 (in blue). The example below shows nsp6, with the codon-aligned insertion 34-F-35. In addition to the codon numbering, key nucleotide positions on both the submitted and reference sequences are also given. After changing selection of coding region or submitted sequence, click "Update" to update the view.



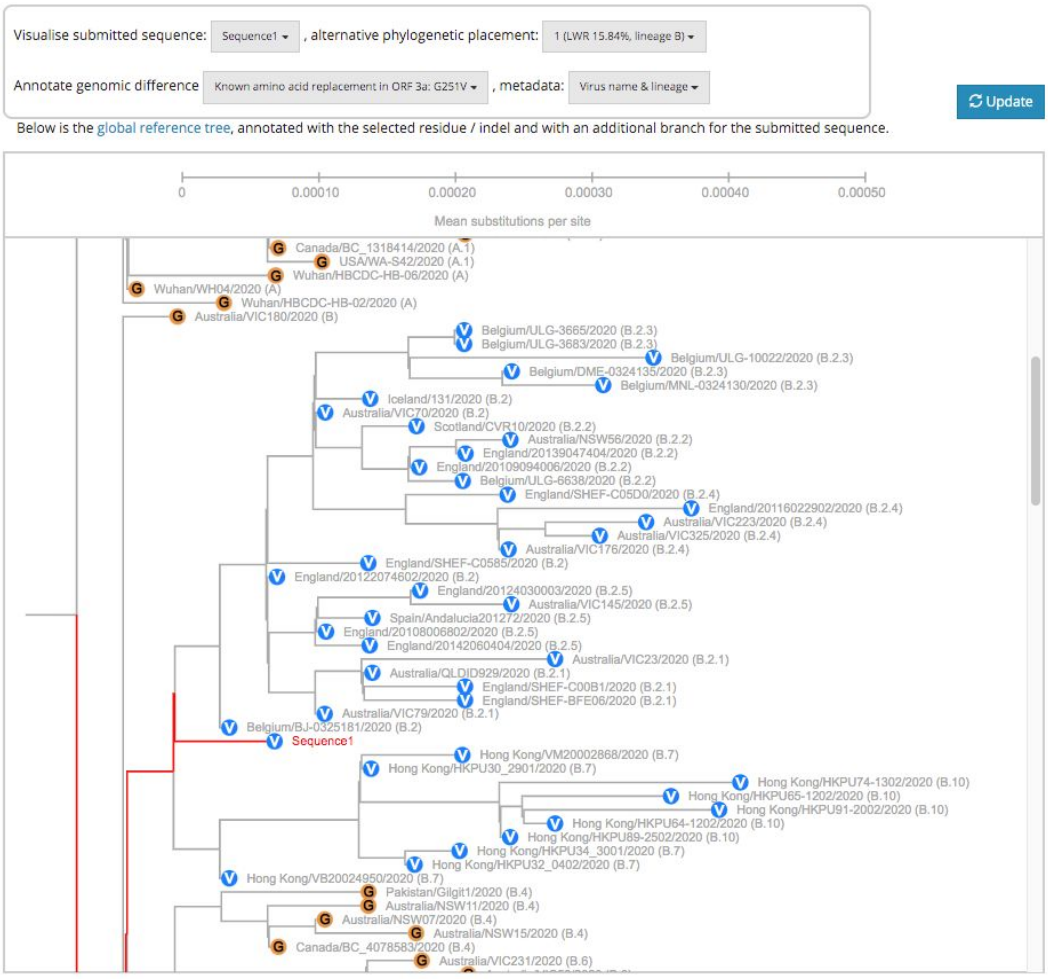
Phylogenetic placement

The "Phylogenetic placement" tab of the analysis report allows you to investigate in detail the placement of the submitted sequence within the global reference tree. The branch for the submitted sequence is picked out in red (see below).

The RAxML-EPA phylogenetic method often produces multiple alternative placements for the submitted sequence within the reference tree. You can switch between these using the drop-down next to "alternative phylogenetic placement". Each alternative placement has a relative confidence level (the "likelihood weight ratio") and a lineage. The overall assignment displayed on the "Summary" tab is the most specific lineage with a total likelihood weight ratio of at least 50%.

The tree can be annotated with any of the detected replacements or indels within the submitted sequence, or various metadata.

After changing any selection, click "Update" to update the view.



Excluded sequences

CoV-GLUE excludes the following GISAID sequences from its analysis:

- Any sequence of nucleotide length less than 29,000 bases or greater than 35,000 bases
- Any sequence with a non-human host, e.g. bat, pangolin
- Sequences from environmental samples
- Any sequences marked with a warning flag, as having quality issues on GISAID
- Any sequence with more than 10 unique (across the whole dataset) single nucleotide mutations relative to the hCoV-19 reference sequence.
- Any sequence which we detect has a frameshifting deletion or insertion relative to the hCoV-19 reference sequence.

Reference tree

CoV-GLUE uses a reference tree to represent the main lineages of the pandemic at the global level. The lineage definitions are taken from Rambaut, et al. 2020. They can be accessed via a GitHub repository https://github.com/hCoV-2019/lineages. CoV-GLUE is updated regularly based on changes in this repository.

The CoV-GLUE reference alignment and tree are computed as follows:

- A multiple sequence alignment of the lineage representatives is computed using MAFFT.
- The tree is computed using RAxML, 100 bootstraps, GTRGAMMAI substitution model, using the coding region part of the alignment.
- The rooting is based on an early split in the virus population, characterised by the L/S polymorphism at ORF 8 position 84, lineage A has S, lineage B has L.

Acknowledging GISAID data contributors

Publishing analysis requires acknowledgement of GISAID data contributors from Submitting and Originating laboratories, on which the research is based. Details of how to provide suitable acknowledgements are given on the "Data acknowledgement" page under the "About" option in the CoV-GLUE web application menu.

Open reading frames and viral proteins of hCoV-19

It is thought that the hCoV-19 genome has 12 open reading frames (ORFs) translated by cellular ribosomes: ORF 1a, ORF 1ab, S, ORF 3a, E, M, ORF 6, ORF 7a, ORF 7b, ORF 8, N and ORF 10. The two large ORFs, 1a and 1ab, share the same start point. At nucleotide location 13468, about 4400 codons to the 3' side of this start point, there is a -1 ribosomal slippage site, hence a certain percentage of the time this nucleotide is processed twice. If the slippage does not occur, a stop is encountered after only 5 codons, producing the ORF 1a polyprotein. If it does occur, a stop occurs after ~2700 codons, producing the longer ORF 1ab polyprotein.

There are then sites where polyproteins 1a and 1b are cleaved by viral proteases into the mature non-structural proteins (nsps), numbered nsp1 to nsp16. The nsps 1-10, (on the 5' side of the slippage site) are produced from both polyproteins 1a and 1ab, nsps 13-16 (on the 3' side) are produced only from 1ab. The nsp12 (RdRp) protein is produced only from 1ab and overlaps the slippage site. The short nsp11 protein is produced only from 1a.