

# *EDGE COVID-19 documentation*

Los Alamos National Laboratory

Email us: [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov)

v1.0.16

<b>Overview</b>	<b>3</b>
<b>A step by step guide for running EC-19:</b>	<b>5</b>
Step 1: Create an account	5
Step 2: Upload your raw reads	5
Step 3: Run your sample	7
<b>EDGE COVID-19 output page</b>	<b>20</b>
<b>EDGE COVID-19 output files</b>	<b>23</b>
<b>ThirdParty Tools</b>	<b>27</b>
<b>FAQs</b>	<b>33</b>
For web based app:	33
How can I view alignments in a local viewer such as IGV?	33
<b>For local docker build:</b>	<b>33</b>
1. How to start/stop EDGE COVID-19 docker instance?	33
2. How to update EDGE COVID-19?	34
3. How long will it take to run my sample?	34
4. I am getting an error while pulling the image. What can I do?	34
6. IP address conflicts	35
7. What are the commands for checking status and error log?	35
8. How can I update the number of CPUs that EDGE COVID-19 uses?	36
<b>Contact:</b>	<b>36</b>
<b>Citation:</b>	<b>36</b>

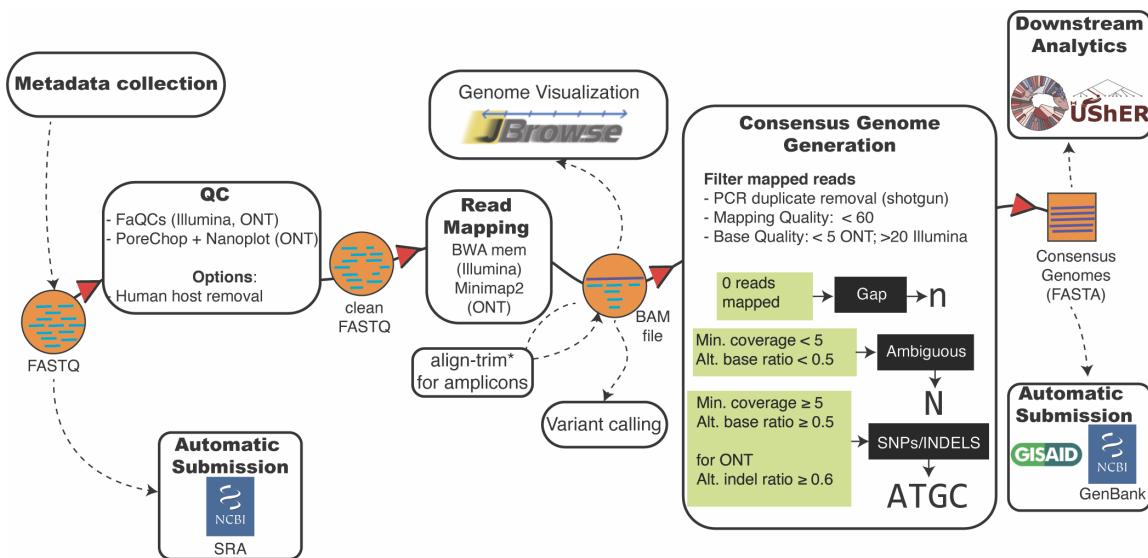
# Overview

[EDGE COVID-19](#) (EC-19) is a tailored bioinformatics platform based on the more flexible and fully open-source [EDGE Bioinformatics](#) software ([Li et al. 2017](#)). This mini-version consists of a user-friendly GUI that drives standardized workflows for genome reference-based ‘assembly’ and preliminary analysis of Illumina or Oxford Nanopore (ONT) or PacBio data for SARS-CoV-2 genome sequencing projects. **The result is a final SARS-CoV-2 genome ready for submission to [GISAID](#) and [GenBank](#).**

The default workflow in *EDGE COVID-19* includes (**Figure 1/Table 1**):

- 1) data quality control (QC) and filtering,**
- 2) alignment of reads** to the original (first available) reference genome ([NC\\_045512.2](#), we removed the PolyA tail from the 3' end (33 nt)),
- 3) creation of a consensus genome** sequence based on the read alignments, and
- 4) variant analyses**, with details on location (For eg. coding region. Synonymous changes, etc.)

The [EDGE COVID-19](#) platform can process Illumina, ONT data and PacBio, including ONT data from the [SARS-CoV-2 ARTIC network sequencing](#) protocols. Users can input/upload Illumina or ONT or PacBio FASTQ files (and/or download from [NCBI SRA](#)). For Illumina data, default analyses include read QC, read mapping to the reference, and variant analysis. For PacBio data, the read doesn't perform QC. For ONT data, the data must be demultiplexed prior to uploading; the samples will be processed individually. Also, the variant calling is not on by default for ONT. However, other functions (e.g. *de novo* assembly for whole genome data) are also available for these sequencing platforms. While command line execution is possible ([see here](#) and [here](#)), the GUI provides an easy data submission and results viewing platform, with the graphical and tabular views of variant/SNP data and a genome browser to view read coverage and location of SNPs or variants, as well as the reference annotations.



**Figure 1: Overview of EDGE COVID-19 workflow.** The workflow includes Quality Control (QC) and removal of low quality data in raw reads (FaQCs v2.09 (Lo and Chain, 2014), Porechop v0.2.3 (Wick, et al., 2017), mapping reads to a SARS-CoV-2 reference genome sequence using BWA v0.7.12 (Li and Durbin, 2009) (default for Illumina) or minimap2 v2.17(Li, 2018) (default for ONT), removing primer sequences (modified ver. of *align\_trim*) for genomes sequenced using amplicon-based enrichment strategies (like ARTIC (Tyson, et al., 2020), SWIFT, CDC-SC2 (Paden, et al., 2020), etc.), generation of consensus genomes, variant calling (using SAMtools v1.10 and BCFtools v1.10.2 (Li, 2011), lineage call using Pangolin v3.1.16, and phylogenetic placement using UsheR v0.5.1. Dotted line indicates optional steps.

**Table 1: EDGE COVID-19 Workflows summary**

	EC19-ONT	EC19-ILLUMINA	EC19-PacBio
QC	FaQCs -q 7 -min_L 350	FaQCs -q 20 -min_L 50	NA
Primer trimming	<i>align_trim*</i>	<i>align_trim*</i>	<i>align_trim*</i>
Mapping/Aligner	minimap2	BWA	minimap2
Mapping/Aligner args	--MD -La	mem	--MD -La -x map-pb
Filter clipped alignment length (SamClip <sup>\$</sup> )	150	50	150
Variant call/Consensus genomes tool	samtools mpileup	samtools mpileup	samtools mpileup
Variant call parameters	-q 60 -Q 5 -d 0 -A -B -a	-q 60 -Q 20 -d 0 -A -B -a	-q 60 -Q 5 -d 0 -A -B -a
Consensus <sup>&amp;</sup> parameters	--propThresh=0.5 --covThresh=5 --baseQual=5 --hetpropThresh=0.2 --filterHomopolymer --filterStrandBias	--propThresh=0.5 --covThresh=5 --baseQual=20 --hetpropThresh=0.2	--propThresh=0.5 --covThresh=5 --baseQual=5 --hetpropThresh=0.2 --filterHomopolymer
Consensus parameters Indel specific	--INDELpropThresh=0.6	--INDELpropThresh=0.5	--INDELpropThresh=0.6

\$ SamClip

<https://github.com/tseemann/samclip>

\* *align\_trim*

[https://github.com/LANL-Bioinformatics/EDGE/blob/SARS-CoV2/scripts/align\\_trim.py](https://github.com/LANL-Bioinformatics/EDGE/blob/SARS-CoV2/scripts/align_trim.py)

& consensus workflow

[https://gitlab.com/chienchi/reference-based\\_assembly](https://gitlab.com/chienchi/reference-based_assembly)

# for non-amplicon methods (no *align\_trim*), PCR deduplication is also performed by `samtools markdup -r -s`

We have tested these workflows using Illumina (e.g. [SRR11393704](#)) and ONT (e.g. [SRR11397722](#)) datasets; these projects (along with a few others) are made public on the [site](#). The workflow is also available as a Docker container (<https://hub.docker.com/r/bioedge/edge-covid19>), able to run on any local hardware infrastructure.

Note: For EDGE Bioinformatics users who would also like to use the phylogeny or read- and assembly-based taxonomy classification tools to identify all organisms that may be present within complex samples, we recommend using the original [EDGE Bioinformatics](#) platform which harbors several tools and associated (large) databases that enable such a search. *In initial tests of taxonomy classification of SARS-CoV-2 samples (with no SARS-CoV-2 genomes in any of the databases), we recover SARS coronavirus and Bat coronavirus as the nearest neighbors (See table below).*

Tool	#Reads	%Reads	Level	Top1	Top2	Top3	Top4	Top5	Columns...
gottcha-strDB-v	7,827	6.0	strain	SARS coronavirus	Bat coronavirus BM48-31/BGR/2008	N/A	N/A	N/A	
pangia	5,008	3.9	strain	SARS coronavirus	Bat coronavirus BM48-31/BGR/2008 strain BtCoV/BM48-31/BGR/2008	N/A	N/A	N/A	
metaphlan2	0	0.0	strain	N/A	N/A	N/A	N/A	N/A	
bwa	3,296	2.5	strain	Bat coronavirus Rp/Shaanxi2011	Bat coronavirus Cp/Yunnan2011	BtRs-BetaCoV/HuB2013	Rhinolophus affinis coronavirus	Bat SARS-like coronavirus RsSHC014	
kraken2	48,317	37.2	strain	SARS coronavirus	Rhodococcus opacus PD630	Burkholderia dolosa PC543	Xanthomonas citri pv. fuscans	Aeromonas hydrophila ML09-119	

## A step by step guide for running EC-19:

Visit <https://edge-covid19.edgebioinformatics.org/> and follow the steps below:

### Step 1: Create an account

You need to create an account. Click the “Sign up” link in the upper right corner of the page. After you have an account, you can click “Log in” for all subsequent visits and provide your user information. If you don’t want to create an account, the “GUEST” account is provided for anonymous login with data constraint to keep 24 hours only.



**Login to EDGE**

<input type="text"/> Email Address
<input type="password"/> Password
<b>Submit</b>
<input type="checkbox"/> Remember my email

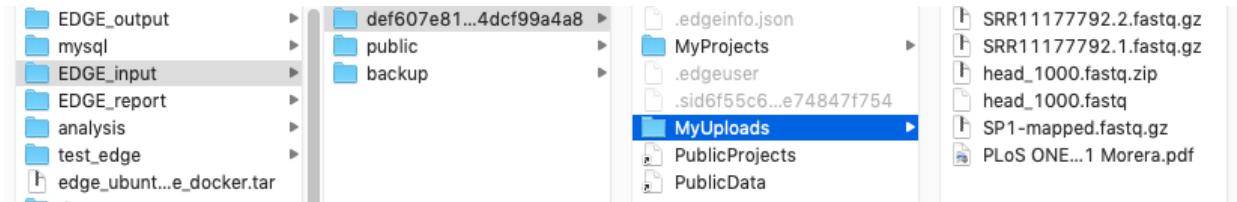
Forgot your password? [Reset it here!](#)  
 New to EDGE? [Sign up now!](#)  
 Continue as GUEST: [GO!](#)

## Step 2: Upload your raw reads

After you have logged in, you can click on “Upload Files” in the left menu. Drag and drop your data files into the window provided. Click “Start Upload” when you have added the files you need. The files will be put in a folder called MyUploads. The maximum file size is 5 gb and the total user upload space is 25 gb. The data will be stored for 180 days (It will be adjusted depends on the system resource and users will be notified) and user can clean up their uploaded data in the user icon popup menu.

### **For a local installation:**

The easiest way to upload your data is to put your data files in the upload folder, which can be found within the `EDGE_input` folder that you created when installing *EDGE COVID-19* docker [[see here](#)]. Within `EDGE_input`, there will be a folder with a long string of characters as the name and within that folder, there will be a folder called `MyUploads` where you can put your raw reads. This folder can then be seen from the web server (<http://localhost/>) by clicking on the button next to boxes where you input your FASTQ files.



The **MyUploads** folder can be seen from the web server by clicking on the button to the right of the box(es) where you input your FASTQ files. (See figure below.) Click the file(s) you want to analyze.

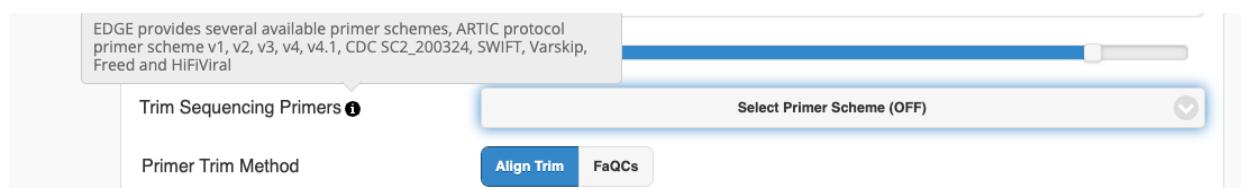
## Step 3: Run your sample

If you are familiar with the EDGE Bioinformatics environment then you can skip this step and jump right into analyzing your data. Even if you are not familiar, you may still be able to skip this step, as EDGE Bioinformatics has a relatively intuitive design to instinctively get to your analyses right away. However, for completeness, here is a short description that will get you started. For a more detailed description, you can also visit our documentation site for EDGE Bioinformatics at <https://edge.readthedocs.io>.

In the *EDGE COVID-19* web server,

1. Type in a unique **Project/Run Name** with no spaces, but use underscores and/or dashes, if needed.
2. Write in a short **Description**. Spaces are allowed.

3. In the **Input Source** section, select **READS/FASTQ** for analyzing your own raw reads or **NCBI SRA** if you want to analyze COVID-19 samples deposited in SRA.
4. Select **Platform** with **Nanopore** if your sample was generated using Nanopore; select **illumina** or **PacBio** if sample was from the corresponding platform.
5. Input your raw reads by clicking on the button to the right of the input box (highlighted with a red box in the figure above) and then within the GUI navigate to your **MyUploads** folder where you have added your raw reads in Step 2.
6. You can skip the `Batch Project Submission`, if you are only processing one sample. A detailed instructions on using the batch mode can be found [here](#).
7. In the **Input Metadata** section, you can fill in the metadata so that you have all needed information when you are ready to submit genomes to NCBI or GISAID.
8. **Pre-processing** (Data QC) is turned **ON** by default and uses [FaQCs](#). This includes trimming low quality regions of reads and filtering reads that either fail a quality threshold or minimum length. If you wish to change parameters, you can expand the module by clicking on it and modify as desired. The default parameters are as follows:
  - Trim Quality Level: 20 (Illumina), 7(Nanopore)
  - Minimum Read Length: 50 (Illumina), 350(Nanopore)
  - "N" Base Cutoff: 10
  - Low Complexity Filter: 0.85
9. **Trimming primers from samples sequenced using multiplex amplicon approach.** We provide two options to trim primers if a multiplex amplicon approach such as [ARTIC \(v1-4\)](#), [CDC protocols](#), [SWIFT protocols](#), [Freed \(Midnight\) protocols](#), [HiFiViral\(Pacbio\)](#) and [Varskip](#) were used. Default approach is to use the *align\_trim*, that soft clips primer region from the alignment file (BAM) based on the position of primers in the reference genome. Another approach is to use FaQC, which trims the regions from reads that match with primer sequences.



10. For samples with whole genome sequencing (WGS) data that are interested in *de novo* assembly, you can also turn on `Assembly and Annotation`. Currently we provide [IDBA\\_UD v1.1.1](#), [SPAdes v3.13.0](#), [MEGAHIT v1.2.9](#), [UniCycler v0.4.8](#), [wtbdbg2 v2.5](#), [flye v2.8](#) and [miniasm v0.3](#) as options for assemblies.
11. **Reference-Based SARS-CoV-2 Genome Analysis** is turned **ON** by default. For Illumina data, [BWA mem](#) is used as the default aligner, which is then automatically followed by generation of a consensus sequence and variant calling. For ONT/PacBio data, [minimap2](#) is the default aligner which is also automatically followed by generation of a consensus

sequence, but not variant calling. We currently turn **OFF** variant calling for ONT data as it takes well over 24 hours on this platform. However, you can change any of these parameters by expanding the module and selecting the desired changes. Additionally, to avoid partial short alignment, [samclip](#) script with *max clip* length 50 for Illumina and 150 for ONT/PacBio is also applied.

*Variant calling:* EDGE COVID-19 uses [bcftools mpileup](#) command to convert the aligned BAM file into genomic positions and call genotypes, reduce the list of sites to those found to be variants by passing this file into [bcftools call](#) command. The variant calls are filtered further by [vcfutils.pl](#) of SAMtools with following criteria:

- Minimum Root Mean Square (RMS) mapping quality for SNPs [10];
- minimum read depth [5];
- maximum read depth [1000];
- minimum number of alternate bases [3];
- minimum ratio of alternate bases [0.3];
- SNP within INT bp around a gap to be filtered [3];
- ‘window size for filtering adjacent gaps [10];
- min P-value for end distance bias [1e-15];
- maximum fraction of reads supporting an indel [0.5];

*The [consensus workflow](#):* For samples sequenced using non-amplicon methods, PCR deduplication is also performed. Various parameters are defaulted including a minimum of 5x depth coverage of support or variant siter coverage per base (otherwise the consensus will be “N”), base quality (<20 for Illumina and <5 for ONT), alternate base Threshold (0.5 to support an alternative for the consensus to be changed), indels Threshold (to support an INDEL for the consensus to be changed, 0.5 for illumina and 0.6 for amplicon-based ONT), and minimum mapping quality of 60. The strand bias and homopolymer were checked and filtered for samples from ONT..

12. Click the **submit** button at the bottom of the page to start your job

13. The status of all of your projects can be viewed by clicking the **Projects** tab on the left menu and then clicking on **My Project List**. A detailed description can also be found [here](#).

The screenshot shows the 'My Project List' section of the EDGE COVID-19 interface. At the top, there's a search bar labeled 'Find project by name/time' and a button labeled 'My Project List'. Below these are four completed projects listed in a table:

Submission Time (MDT)	Project Name	Status
2020-04-23 01:48:41	SRR11494509_v3	Complete
2020-04-23 01:37:44	SRR11494509_v2	Complete
2020-04-22 22:06:30	SRR11494509_v1	Complete

If a project is highlighted in **green** it means the project has finished; if **orange**, **red**, or **grey** (not shown) it means the project is running, cancelled/failed, or not yet started, respectively. You can access the details and outputs of each project by clicking on the project in the list. Once selected, the project results page displays a summary of the run and general statistics of what tools and modules were activated and their runtime and status, as well as links to log files that provide detailed information on the command lines and parameters used for each tool executed. The rest of the page is divided by the modules that were originally selected for analysis. When selecting a project which has not finished running, some of the completed results may still be viewed on the page, however graphics and links to interactive features will not be present, as the rendering of figures is performed only in the last step.

14. Tree Placement of consensus genome(s) by [UShER](#) (Ultrafast Sample placement on Existing Tree): you can select the consensus genomes from the project list to do tree placement and output to a new tab with UShER result (link to UCSC).

The screenshot shows the 'Project List' interface with a modal dialog box overlaid. The dialog is titled 'Proceeding action' and asks: 'Do you want place following genome(s) into a larger SARS-CoV-2 tree using UShER?'. It lists three selected genomes: '2111\_008', '2111\_007', and 'SRR12349131'. There are 'Cancel' and 'Confirm' buttons at the bottom right of the dialog. The background table shows a list of projects with columns for Project Name, Status, Submission Time (MDT), Total Running Time, and Owner.

Project Name	Status	Submission Time (MDT)	Total Running Time	Owner
2111_010	Complete	2020-10-09 11:24:22	00:20:53	Andrew Bartlow
2111_009	Complete	2020-10-09 11:24:07	00:17:07	Andrew Bartlow
2111_008	Complete	2020-10-09 11:23:52	00:35:11	Andrew Bartlow
2111_007	Complete	2020-10-09 11:23:37	00:25:04	Andrew Bartlow
2111_005	Complete			Andrew Bartlow
2111_006	Complete			Andrew Bartlow
2111_004	Complete			Andrew Bartlow
2111_003	Complete			Andrew Bartlow
2111_002	Complete			Andrew Bartlow
SRR12349131	Complete			Hajnalka Daligault
cats_2111_001_test	Complete			Andrew Bartlow
SRR11514749	Complete			Andrew Bartlow
SRR11593395_no_primer_removal	Complete	2020-10-01 13:01:50	00:15:36	Hajnalka Daligault
SRR19111157	Complete	2020-10-01 11:12:03	00:18:47	Hajnalka Daligault

UShER: Ultrafast Sample placement on Existing tRee

[view in Genome Browser](#) | [view subtree 1 in Nextstrain](#) | [view subtree 2 in Nextstrain](#) | [view subtree 3 in Nextstrain](#)

Fasta Sequence	Size ?	#Ns ?	#Mixed ?	Bases aligned ?	Insertions ?	Deletions ?	#SNVs used for placement ?	#Masked SNVs ?	Neighboring sample in tree ?	Lineage of neighbor ?	#Imputed values for mixed bases ?	#Maximally parsimonious placements ?	Parsimony score ?	Subtree number ?
SRR11514749_consensus_hCoV_19_SouthKorea_S2_2020_EPI_ISL_485393_2020_03_30_2020_07_09_South	29828 ?	2461 ?	0 ?	27227 ?	1268 ?	1262 ?	5 ?	1 ?	India/DL-NCDC-3982/2020   EPI_ISL_436445   20-04-09	B.6	0	18 ?	0	1 (view in Nextstrain)
SRR11514749_consensus_NC_045512_2	29782 ?	313 ?	0 ?	29469 ?	0 ?	0 ?	6 ?	1 ?	India/DL-NCDC-3982/2020   EPI_ISL_436445   20-04-09	B.6	0	3 ?	0	2 (view in Nextstrain)
SRR12110157_consensus_hCoV_19_USA_WA_UW_1631_2020_EPI_ISL_477693_2020_03_20_2020_06_29_USA	29864 ?	2 ?	0 ?	29862 ?	0 ?	0 ?	5 ?	0 ?	USA/CT_9849/2020   EPI_ISL_42616   20-03-08	A.1	0	1 ?	0	3 (view in Nextstrain)
SRR12110157_consensus_NC_045512_2	29868 ?	1 ?	0 ?	29867 ?	0 ?	0 ?	5 ?	0 ?	USA/CT_9849/2020   EPI_ISL_42616   20-03-08	A.1	0	1 ?	0	3 (view in Nextstrain)

#### Subtree 1: Unrelated sample

SRR11514749\_consensus\_hCoV\_19\_SouthKorea\_S2\_2020\_EPI\_ISL\_485393\_2020\_03\_30\_2020\_07\_09\_South

Differences from the reference genome (NC\_045512.2): C6310A, C6312A, C19370T, C19524T, C23929T

Mutations along the path from the root of the phylogenetic tree to SRR11514749\_consensus\_hCoV\_19\_SouthKorea\_S2\_2020\_EPI\_ISL\_485393\_2020\_03\_30\_2020\_07\_09\_South: C13730T > C28311T > C6312A > C23929T > C19524T > C6310A

This placement is not the only parsimony-optimal placement in the tree; 17 other placements exist.

Nearest neighboring GISAID sequence already in phylogenetic tree: India/DL-NCDC-3982/2020/EPI\_ISL\_436445|20-04-09: lineage B.6

15. Before submitting the pipeline run, you can prepare your genome for submission to GISAID and NCBI by inputting the metadata. You can do this after the pipeline is run as well.

EDGE COVID-19

@edge-covid19.edgebioinformatics.org

Login / Sign up

Upload Files

Run EDGE COVID-19

Get EDGE COVID-19

Reports

Projects

Input Raw Reads

Batch Project Submission

Input Metadata

**Virus detail**

Virus name: hCoV-19/Country/Identifier/2020

Passage details/history: Example: Original, Vero

**Sample information**

Collection date: Example: 2020-03-27, 2020-03 (collection in March, day unknown), 2020 (month and day unknown)

Location: Continent/Country/Region

Host: Example: Human, Environment, Canine, *Manis javanica*, *Rhinolophus affinis*, unknown

Gender: Example: Male, Female, or unknown

Patient age: Example: 65, 7 months, or unknown

After entering the metadata, the genome can be submitted to GISAID and NCBI directly through the *EDGE COVID-19* platform. You can access this functionality by clicking on the green check mark in the Reference-based results just below “Ready to Submit”.

Reference-Based SARS-CoV-2 Genome Analysis

a. Reads Mapped to Reference(s)

i. Mapped Reads By bwa

Columns...

SARS-CoV2 Reference	Ref Length	Ref GC%	Mapped Reads	Mapped Reads%	Base Coverage	Avg Fold	Bam File
NC_045512.2	29,870	38.01%	119,708	99.04%	99.97%	725.92X	

Link to | Directory

1 out of 1 reference(s) is(are) covered by input reads.

ii. Consensus Genome Statistics

Columns...

SARS-CoV2 Reference	Consensus Length	Gaps	Ns/ns	5' Ns/ns	3' Ns/ns	SNVs	INDELs	Lineage	Consensus Genome	Ready to Submit
NC_045512.2	29,870	1	88	82	6	5	0	A.1		

Link to | Directory



A menu will appear on the right side of the screen. Click the Metadata Action -> Upload to GISAID and NCBI option at the bottom of the menu to submit consensus genomes.

The screenshot shows the EDGE COVID-19 project interface. On the left, the project summary for SRR11241255 is displayed, including submission details and analysis status. The 'Job Progress' sidebar on the right lists completed tasks like 'Download SRA' and 'Count Fastq'. The 'EDGE Server Usage' section shows CPU, MEM, and DISK usage. The 'Action' menu on the far right includes options like 'View live log', 'Force this project to rerun', and 'Upload to GISAID and NCBI'. A red arrow points to the 'Upload to GISAID and NCBI' option in the 'Metadata Action' submenu.

Analysis	Run	Status	Running
Download SRA	On	Complete	00:00:10
Count Fastq	Auto	Complete	00:00:01
Quality Trim and Filter	On	Complete	00:00:04
Reads Mapping To Reference	On	Complete	00:00:30
Variant Analysis	Auto	Complete	00:00:01
Generate JBrowse Tracks	On	Complete	00:00:09
HTML Report	On	Complete	00:00:16

Report/Info	Location
Input Reads	SRR11241255
Output Directory	SRR11241255
PDF Report	final_report.pdf
MetaData	metadata.txt
Process log	process.log
Error log	error.log
Direct access	link

**Job Progress**

**SRR11241255**

- Download SRA
- Count Fastq
- Quality Trim and Filter
- Reads Mapping To Reference
- Variant Analysis
- Generate JBrowse Tracks
- HTML Report

Last checked: 2021-04-06 23:41:07

**EDGE Server Usage**

CPU 0.0 %  
MEM 10.5 %  
DISK 76.0 %

**Action**

- View live log
- Force this project to rerun
- Reconfig project (BETA)
- Interrupt running project
- Delete entire project
- Empty project outputs
- Share project
- Make project public
- Rename Project

**Metadata Action**

- Update Metadata
- Upload to GISAID and NCBI
- Upload to NCBI SRA

Submit SRR11241255\_3 to GISAID

**Virus detail**

Virus name	hCoV-19/USA/LANL01/2020
Passage details/history	Original

**Sample information**

Collection date	2020-09-08
Location	North America/USA/Los Alamos
Host	Homo Sapiens
Gender	Male
Patient age	65
Patient status	Unknown
Sequencing technology	Illumina
Assembly method	EDGE-covid19: bwa 0.7.12-r1039, Consensus min coverage: 5X, min map quality: 60, Alternate Base > 50%, Indel > 50%.
Consensus Fasta	NC_048512.2 (99.9%, 728X), Ready to Submit

**Institute information**

Originating lab	New Mexico Department of Health Scientific Laboratory Division
Address	1101 Camino De Salud NE, Albuquerque, NM 87102
Submitting lab	LANL Bioscience
Address	PO 1663 MS888, Los Alamos, NM 87544
Authors	Migun Shakya, Chienchi Lo, Patrick Chan, Cheryl Gleasner, Alina Deshpande

**Submitter information**

Submitter	Chienchi Lo
GISAID ID	[ ]
GISAID Password	[ ]
NCBI ID	[ ]
NCBI Password	[ ]

Required metadata fields must be properly filled for submission to proceed. You will need to have a registered [GISAID account](#) and a [NCBI account](#).

By clicking the "Confirm" button, you hereby authorize EDGE-COVID19 to submit the consensus genomes and metadata to the GISAID and NCBI Genbank, and agree to remit the samples and related metadata to the public domain.

If your submission is successful, you should receive an email from GISAID and NCBI for assigned accession numbers or further instructions.

Note: this feature is in beta format, and GISAID and NCBI can change the submission process at any time; if you run into any trouble, you can contact us at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov).

## 16. Raw Reads submit to NCBI SRA:

In the same menu on the right side of the screen, users can submit the raw reads fastq to NCBI SRA. Click the Metadata Action -> Upload to NCBI SRA option at the bottom of the menu to submit.

The screenshot shows the project details for SRR11241255. The 'Job Progress' section lists completed steps: Download SRA, Count Fastq, Quality Trim and Filter, Reads Mapping To Reference, Variant Analysis, Generate JBrowse Tracks, and HTML Report. The 'EDGE Server Usage' section shows CPU at 0.0%, MEM at 10.5%, and DISK at 76.0%. The 'Action' section includes options like View live log, Force this project to rerun, Reconfig project (BETA), Interrupt running project, Delete entire project, Empty project outputs, Share project, Make project public, Rename Project, and two options under 'Metadata Action': Update Metadata and Upload to NCBI SRA. A red arrow points to the 'Upload to NCBI SRA' button.

SRR11241255

Project Summary

Description: -

Submission Time: 2020 Sep 23 20:42:10 (MDT)

Number of CPUs: 10

Project Status: Complete

Total Analysis Run Time: 00:01:15

Last Run Time: 00:01:15

Owner: chienchi@lanl.gov

General

Analysis	Run	Status	Running
Download SRA	On	Complete	00:00:10
Count Fastq	Auto	Complete	00:00:01
Quality Trim and Filter	On	Complete	00:00:04
Reads Mapping To Reference	On	Complete	00:00:30
Variant Analysis	Auto	Complete	00:00:01
Generate JBrowse Tracks	On	Complete	00:00:09
HTML Report	On	Complete	00:00:16

Report/Info	Location
Input Reads	SRR11241255
Output Directory	SRR11241255
PDF Report	final_report.pdf
MetaData	metadata.txt
Process log	process.log
Error log	error.log
Direct access	link

SRA Sample Metadata

Job Progress

SR11241255	
Download SRA	✓
Count Fastq	✓
Quality Trim and Filter	✓
Reads Mapping To Reference	✓
Variant Analysis	✓
Generate JBrowse Tracks	✓
HTML Report	✓

Last checked: 2021-04-06 23:41:07

EDGE Server Usage

CPU	0.0 %
MEM	10.5 %
DISK	76.0 %

Action

- View live log
- Force this project to rerun
- Reconfig project (BETA)
- Interrupt running project
- Delete entire project
- Empty project outputs
- Share project
- Make project public
- Rename Project

Metadata Action

- Update Metadata
- Upload to NCBI SRA
- Upload to GISAID and NCBI

### Submit SRR13361443 to NCBI SRA

The screenshot shows a user interface for submitting a sample to the NCBI SRA. At the top, it says "Submit SRR13361443 to NCBI SRA". Below this, there's a section titled "BioProject" with a "BioProject" icon. It includes two buttons: "Use Registered BioProject" (with "Yes" selected) and "Existing BioProject" (with "PRJNA714680" entered). There are also sections for "BioSample", "Experiment", and "Additional Information", each with a corresponding icon.

Required metadata fields must be properly filled for submission to proceed.

By clicking the "Confirm" button, you hereby authorize EDGE-COVID19 to submit the samples and metadata to the NCBI SRA, and agree to remit the samples and related metadata to the public domain.

If your SRA submission is successful, you should receive an email with the subject *Submission ownership transfer*. After the ownership transfer, you can view the submission process at the [Submission Portal](#). You may need to log in with the NCBI credentials for the account you used in the submission metadata.

Note: this feature is in beta format, if you run into any trouble, you can contact us at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov).

#### 17. Batch submit (consensus genomes):

You can access this functionality by clicking on **My Project List**. Select on projects you would like to do the batch submission and then click on the upper-arrow action button at top of the table.

Project Name	Status	Submission Time (MDT)	Total Running Time	Type	Owner
<a href="#">ERR4868644</a>	Complete	2021-03-24 14:07:47	00:10:16	private	Chienchi Lo
<a href="#">SRR13361443</a>	Complete	2021-03-03 17:02:16	00:02:30	private	Chienchi Lo
<a href="#">SRR13361443_o</a>	Complete	2021-03-03 16:29:16	00:11:41	private	Chienchi Lo
<a href="#">SRR11547279</a>	Complete	2021-03-03 12:50:05	00:04:58	private	Chienchi Lo
<a href="#">SRR13530301</a>	Complete	2021-01-28 03:37:19	00:03:38	private	Chienchi Lo
<a href="#">SRR11445485</a>	Complete	2020-12-18 08:50:26	00:08:10	private	Chienchi Lo
<a href="#">ERR4206007</a>	Complete	2020-12-18 08:02:27	01:52:57	private	Chienchi Lo
<a href="#">SRR11494688</a>	Complete	2020-10-29 10:01:58	00:03:29	private	Chienchi Lo

The action button will bring up the selected projects metadata table for users to fill in. (can scroll to the right to see other metadata.)

### Selected Project Metadata

Type directly in the form below OR [upload a tab-delimited text file.](#)

Project Name *	Virus Name	Passage Details	Collection Date	Location	Host	Gender	Patient Age	Patient Status	Sequencing Technology	Consensus Fasta	Submit
<input type="checkbox"/> <a href="#">ERR4868644</a>	hCoV-19/USA/LANL01/2020	Original	2021-02-03	North America/USA/	Homo S.	Female	56	Unknown	Illumina	NC_046512_2 (99.71%, 823X), Ready to Submit	<input type="button" value=""/>
<input type="checkbox"/> <a href="#">SRR13361443</a>	hCoV-19/USA/LANL02/2020	Original	2021-02-16	North America/USA/1	Homo S.	Other	56	Unknown	Illumina	NC_046512_2 (99.98%, 2458X), Ready to Submit	<input type="button" value=""/>
<input type="checkbox"/> <a href="#">SRR13361443_o</a>						Select Gender		Select Status		Select consensus genome	<input type="button" value=""/>

Showing 1 to 3 of 3 entries

Don't see X scroll bar? On Mac, please try [this](#).

### Additional Information

Existing BioProject <a href="#">?</a>	[Optional] PRJNAXXXX, ex:PRJNA714680
Release date	Example: 2021-04-20

### Institute Information

Originating lab <a href="#">?</a>	New Mexico Department of Health Scientific Laboratory Division
Address	1101 Camino De Salud NE, Albuquerque, NM 87102
Submitting lab <a href="#">?</a>	Los Alamos National Laboratory Bioscience Division
Address	PO 1663 MS888, Los Alamos, NM 87545
Authors <a href="#">?</a>	Chien-Chi Lo, Migan Shakya, Cheryl Gleasner, Kim McMurry, Alina Deshpande, Twila Kunde, Joseph Hicks, Michael Edwards, Patrick Chain

### Submitter Information [?](#)

Submitter	Chienchi Lo
GISAID ID	chienchi
GISAID Password	
NCBI ID	andy4748
NCBI Password	

Required metadata fields must be properly filled for submission to proceed. You will need to have a registered [GISAID account](#) and a [NCBI account](#).

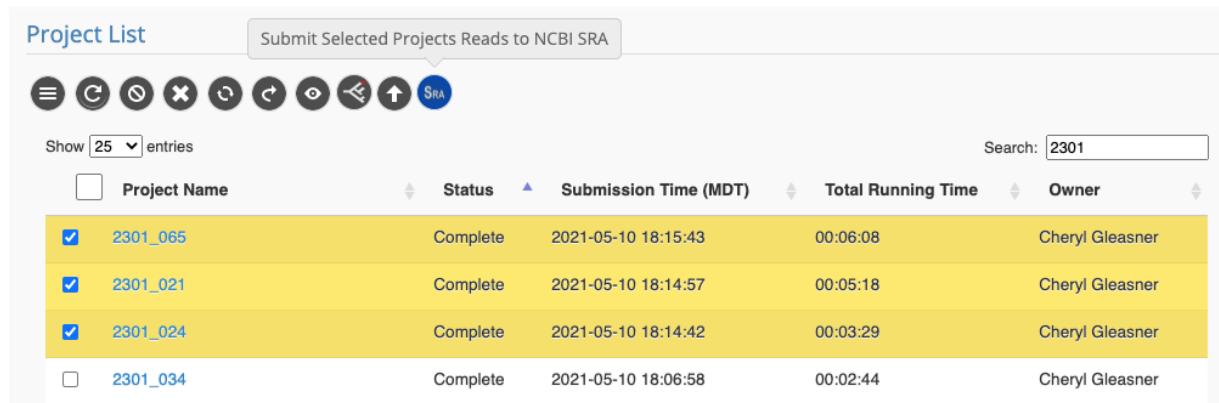
By clicking the "Confirm" button, you hereby authorize EDGE-COVID19 to submit the consensus genomes and metadata to the GISAID and NCBI Genbank, and agree to remit the samples and related metadata to the public domain.

If your submission is successful, you should receive an email from GISAID and NCBI for assigned accession numbers or further instructions.

Note: this feature is in beta format, and GISAID/NCBI can change the submission process at any time; if you run into any trouble, you can contact us at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov).

#### 18. Batch submit (NCBI SRA):

You can access this functionality by clicking on **My Project List**. Select on projects you would like to do the batch submission and then click on the right-most (SRA) action button at top of the table.



Project List					
Submit Selected Projects Reads to NCBI SRA					
Show 25 entries Search: 2301					
<input type="checkbox"/>	Project Name	Status	Submission Time (MDT)	Total Running Time	Owner
<input checked="" type="checkbox"/>	2301_065	Complete	2021-05-10 18:15:43	00:06:08	Cheryl Gleasner
<input checked="" type="checkbox"/>	2301_021	Complete	2021-05-10 18:14:57	00:05:18	Cheryl Gleasner
<input checked="" type="checkbox"/>	2301_024	Complete	2021-05-10 18:14:42	00:03:29	Cheryl Gleasner
<input type="checkbox"/>	2301_034	Complete	2021-05-10 18:06:58	00:02:44	Cheryl Gleasner

The action button will bring up the selected projects metadata tables for users to fill in. (can scroll to the right to see other metadata.)

## Selected Projects Metadata

### BioProject

Use Registered BioProject

Existing BioProject [i](#)

PRJNA714680

### Biosamples

Type directly in the form below OR [upload a tab-delimited text file](#).

Show 25  entries

Search:

<input type="checkbox"/>	Project Name	Sample name	Isolate	Isolate source	Location	Passage Details	Collection Date
<input type="checkbox"/>	2301_021	2301_021	SARS-CoV-2/Homo sapiens/USA/NM-LANL-2		USA: New Mexico	Original	2021-02-08
<input type="checkbox"/>	2301_024	2301_024	SARS-CoV-2/Homo sapiens/USA/NM-LANL-2		USA: New Mexico	Original	2021-02-08
<input type="checkbox"/>	2301_065	2301_065	SARS-CoV-2/Homo sapiens/USA/NM-LANL-2		USA: New Mexico	Original	2021-02-19

Showing 1 to 3 of 3 entries

Previous  Next

Don't see X scroll bar!! On Mac, please try [this](#).

### Experiments

Type directly in the form below OR [upload a tab-delimited text file](#).

Show 25  entries

Search:

Project Name	Title	Design	Library Selection	Library Strategy	Library Layout	Library S
2301_021		SWIFT primer shceme V2 amplicon	Select libselection	Select libstrategy	Select liblayout	Select lib
2301_024		SWIFT primer shceme V2 amplicon	Select libselection	Select libstrategy	Select liblayout	Select lib
2301_065		SWIFT primer shceme V2 amplicon	Select libselection	Select libstrategy	Select liblayout	Select lib

Showing 1 to 3 of 3 entries

Previous  Next

### Additional information

Contact Email

cdgle@lanl.gov

Release date

2021-05-20

Questions? Please contact us at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov).

Required metadata fields must be properly filled for submission to proceed.

By clicking the "Confirm" button, you hereby authorize EDGE-COVID19 to submit the samples and metadata to the NCBI SRA, and agree to remit the samples and related metadata to the public domain.

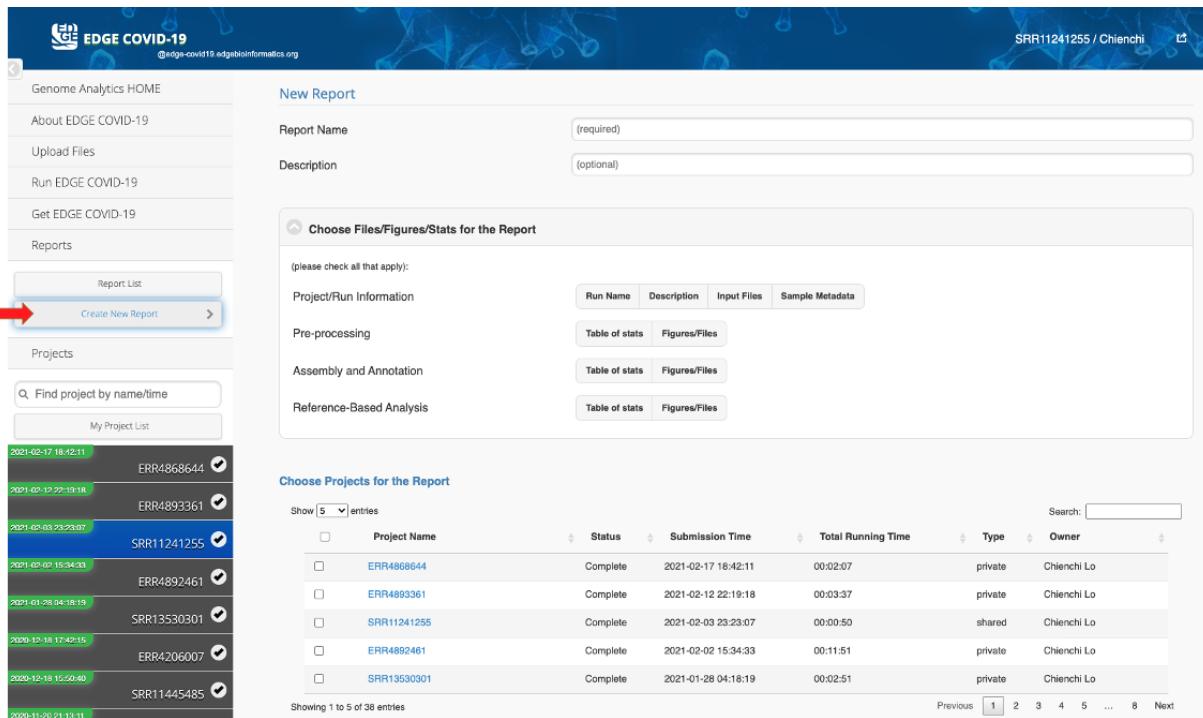
If your SRA submission is successful, you should receive an email with the subject *Submission ownership transfer*. After the ownership transfer, you can view the submission

process at the [Submission Portal](#). You may need to log in with the NCBI credentials for the account you used in the submission metadata.

Note: this feature is in beta format, and NCBI can change the submission process at any time; if you run into any trouble, you can contact us at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov).

## 19. Generate a report that contains comparison among multiple projects.

This feature can be accessed by clicking the “Reports” button on the left side of EC-19 page.



The screenshot shows the 'New Report' section of the EDGE COVID-19 genome analytics interface. On the left, there's a sidebar with links like 'Genome Analytics HOME', 'About EDGE COVID-19', 'Upload Files', 'Run EDGE COVID-19', 'Get EDGE COVID-19', and 'Reports'. Under 'Reports', a red arrow points to the 'Create New Report' button. The main area has sections for 'Choose Files/Figures/Stats for the Report' (with tabs for Run Name, Description, Input Files, and Sample Metadata) and 'Choose Projects for the Report' (listing 5 entries: ERR4868644, ERR4893361, SRR11241255, ERR4892461, and SRR13530301). A table below shows 38 entries with columns for Project Name, Status, Submission Time, Total Running Time, Type, and Owner (Chienchi Lo).

Project Name	Status	Submission Time	Total Running Time	Type	Owner
ERR4868644	Complete	2021-02-17 16:42:11	00:02:07	private	Chienchi Lo
ERR4893361	Complete	2021-02-12 22:19:18	00:03:37	private	Chienchi Lo
SRR11241255	Complete	2021-02-03 23:23:07	00:00:50	shared	Chienchi Lo
ERR4892461	Complete	2021-02-02 15:34:33	00:11:51	private	Chienchi Lo
SRR13530301	Complete	2021-01-28 04:18:19	00:02:51	private	Chienchi Lo
ERR4206007	Complete	2020-12-18 17:42:15			
SRR11445485	Complete	2020-12-18 15:50:40			
		2020-11-26 21:13:11			

Reference-Based Analysis

1. Reads Mapped to Reference(s)

**Mapped Reads**

Link to [ref\\_reads.refs.csv](#)

**Consensus Genome Statistics**

Link to [ref\\_reads\\_cns.csv](#)

**Variant Call**

Link to [ref\\_reads\\_cns.csv](#)

Project/Run Name	Reference	Length	GC%	Mapped Reads	Mapped Reads%	Base Coverage	Avg Fold	Fold std.
SRR13361443_o	NC_045512.2	29,870	38.01%	532,889	99.71%	99.95%	2439.56X	1187.11X
SRR11547279	NC_045512.2	29,870	38.01%	12,804	23.92%	78.35%	142.14X	205.06X
SRR13361443	NC_045512.2	29,870	38.01%	506,204	94.72%	99.94%	2131.33X	1116.63X

Project/Run Name	Reference	Consensus Length	GC%	GAP	Ns/ns	5' Ns/ns	3' Ns/ns	SNPs	INDELs	LINEAGE	Ready to Submit
SRR13361443_o	NC_045512.2	29,851	37.95%	3	32	17	11	29	12	B.1.1.7	
SRR11547279	NC_045512.2	29,787	20.87%	148	12,919	1,616	202	33	60	B	
SRR13361443	NC_045512.2	29,851	37.91%	3	52	38	14	29	4	B.1.1.7	

## EDGE COVID-19 output page

For a more detailed description of the EDGE bioinformatics output page, please refer to our full files list in the next section. Each selected module will be displayed as a subsection, and detailed results may be found in each section. The Pre-processing section, for example, will have details on various statistics from all reads both before and after quality trimming and filtering. If assembly/annotation is selected, this module's output will include the assembled contigs as a Fasta file in addition to assembly metrics and annotation files.

In the **EDGE COVID-19** version of **Reference-based SARS-CoV-2 genome analysis**, an overview of the statistics and reference genome coverage is presented, including fold coverage (in graphical form along the length of the reference genome), as well as number of SNPs and gaps discovered (including those at the 5' and 3' ends of the reference genome). The [Pangolin \(v3\) Lineage](#) assignment also reported with hyperlink to [outbreak.info](#) for detailed information. A warning icon will be shown if it is a [Variant of Concern \(VOC\)](#) or [Variant of Interest \(VOI\)](#). If any

INDELs cause the frameshift or SNP changes result in early stop codon in the CDS region, a warning icon will be shown too. You can directly download the consensus genome by clicking on the download icon. In our report, we also provide a quality check of the consensus genome by providing a green check mark if the resulting consensus genome is longer than 25kb, has coverage depth greater than 10X, and less than 5% of the genome is Ns. More data such as reference genome, BAM file, etc. can be accessed via the **Directory** link which allows access to all output files (e.g., there is an output file detailing the genomic location of SNPs or variant nucleotides, their prevalence within reads covering that position, any changes in translated amino acid composition, etc.). (See below)

#### a. Reads Mapped to Reference(s)

##### i. Mapped Reads By bwa

SARS-CoV2 Reference	Ref Length	Ref GC%	Mapped Reads	Mapped Reads%	Base Coverage	Avg Fold	Bam File	Columns...
NC_045512.2	29,870	38.01%	532,900	99.71%	99.95%	2458.68X		

[Link to I](#) [Directory](#)

1 out of 1 reference(s) is(are) covered by input reads.

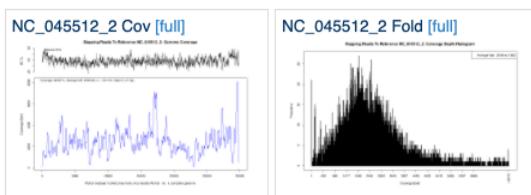
##### ii. Consensus Genome Statistics

SARS-CoV2 Reference	Consensus Length	Gaps	Ns/ns	5' Ns/ns	3' Ns/ns	SNVs	INDELs	Lineage	Consensus Genome	Ready to Submit	Columns...
NC_045512.2	29,851	3	49	38	11	29	4	B.1.1.7			

[Link to I](#) [Directory](#)

##### iii. Variant Call

SARS-CoV2 Reference	Variants	INDELs	Columns...
NC_045512.2	29	1	



Show the results in

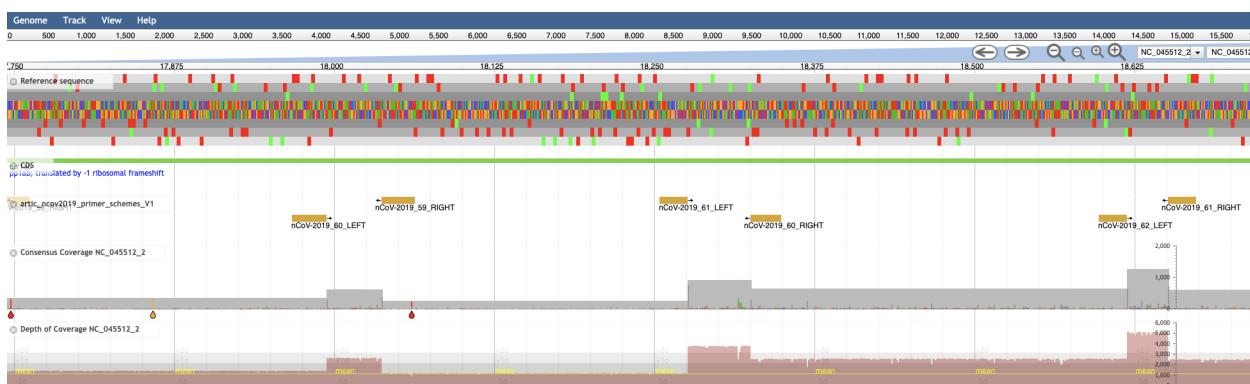
[Link to I](#) [All Plots PDF](#) [SNV Report](#) [INDELs Report](#) [Gap Table](#) [Directory](#)

Table											
	Chromosome	SNP_position	Ref_codon	Sub_codon	aa_Ref	aa_Sub	Synonymous	Product	CDS_start	CDS_end	CDS_stran
NC_045512_2	241	C	T					Intergenic region			
NC_045512_2	913	TCC	TCT	S	S	Yes	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	2110	AAC	AAT	N	N	Yes	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	2836	TGC	TGT	C	C	Yes	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	3037	TTC	TTT	F	F	Yes	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	3267	ACT	ATT	T	I	T1001I	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	6954	ATA	ACA	I	T	I2230T	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	7984	GAT	GAC	D	D	Yes	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	10319	CTT	TTT	L	F	L3352F	GU280_gp01:orf1a polyprotein	266	13483	+	
NC_045512_2	14120	CCA	CTA	P	L	P218L	GU280_gp01:orf1ab polyprotein	13468	21555	+	

Showing 1 to 10 of 31 entries

Previous 1 2 3 4 Next

For scientists wishing to examine the details underlying the statistics, a JBrowse link is also provided (right below the graphics), which will open another browser window and allows interactive examination of the reference genome alignment results including annotations, locations of SNPs or variants, and read alignments (See below).



If the primer **align\_trim** option has been used, users can also access the amplicon coverage plot in the output **Directory** and clicking the *readsToRef\_NC\_045512\_2\_amplicon\_coverage.html* will open the graphics in a new browser window (See below).

a. Reads Mapped to Reference(s)

i. Mapped Reads

SARS-CoV2 Reference	Ref Length	Base Coverage	Avg Fold	Bam File
NC_045512.2	29,870	99.70%	1249.44X	<a href="#">Download</a>

1 out of 1 reference(s) is(are) covered by input

ii. Consensus Genome Statistics

SARS-CoV2 Reference	Consensus Length	DELS	Consensus Genome	Ready to Submit
NC_045512.2	29,870	<a href="#">Download</a>	<a href="#">Link to Directory</a>	<a href="#">Download</a> <a href="#">Checkmark</a>

[Link to Directory](#)

NC\_045512\_2 Cov [full] [Mapping Results To Reference NC\\_045512\\_2\\_Coverage.html](#)

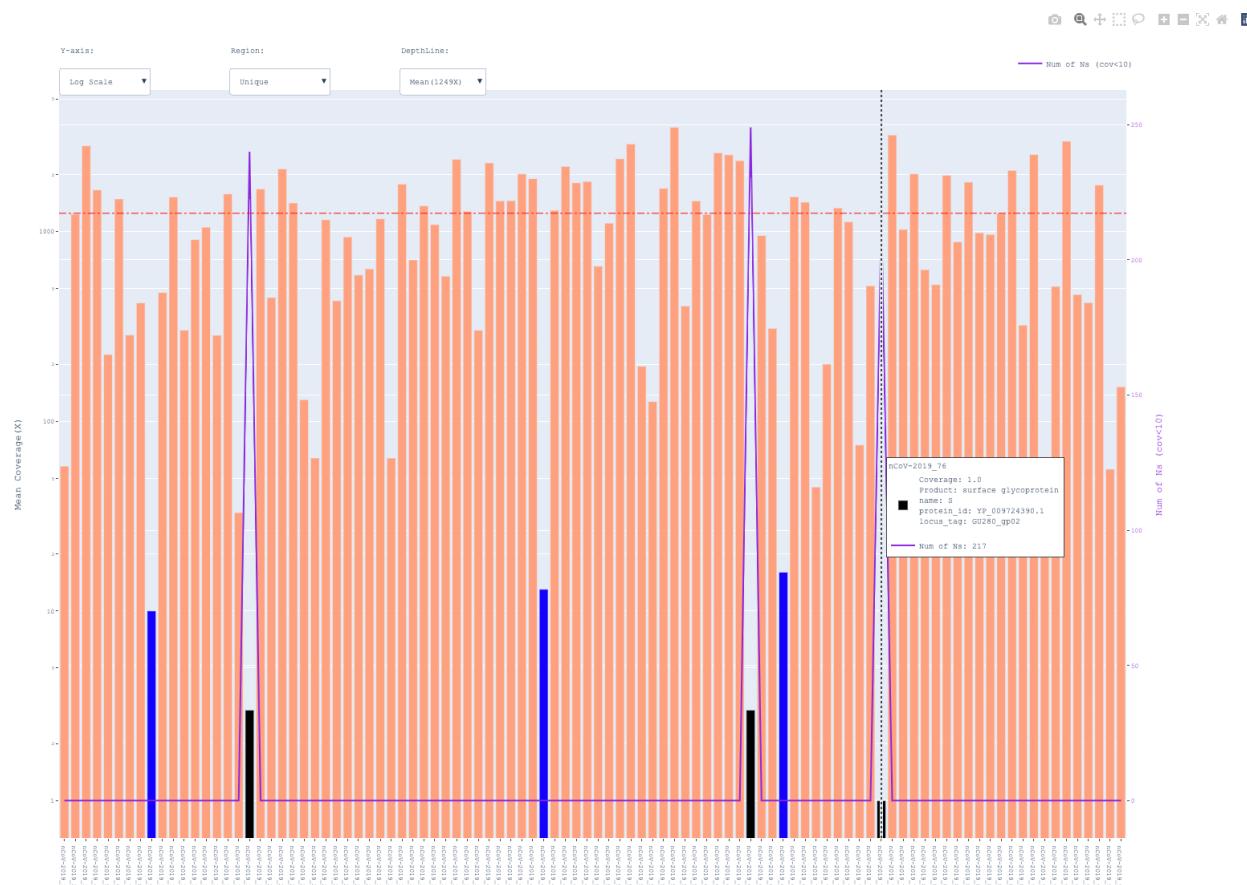
NC\_045512\_2 Cov [full] [Coverage Depth Histogram](#)

Y-axis: Log Scale Unique Depthline: Mean(1249X)

Base Coverage: 99.70% Avg Fold: 1249.44X Bam File: NC\_045512\_2\_consensus.fasta

File Selection Dialog:

- NC\_045512\_2\_consensus.fasta
- NC\_045512\_2\_consensus.fasta.comp
- NC\_045512\_2\_consensus.gaps
- NC\_045512\_2\_consensus.gaps\_report.txt
- consensus.log
- getConsensus.finished
- mapping.log
- readsToRef.alnstats.txt
- readsToRef.gaps
- readsToRef.stats.pdf
- readsToRef\_NC\_045512\_2.coverage
- readsToRef\_NC\_045512\_2.gap.coords
- readsToRef\_NC\_045512\_2\_amplicon\_coverage.html
- readsToRef\_plots.pdf
- runReadsToGenome.finished
- variantAnalysis.finished
- variantAnalysis.log



# EDGE COVID-19 output files

\* **Bold Files** are files with easy to download buttons in the project result page of EDGE GUI. For advanced users, other files are accessible using the GUI project file browser (**Directory**) in the result page. The directory is in a grey background.

Project	File Descriptions
batch_input.json	batch input parameters
clusterJob.log	cluster submit ID and log
clusterSubmit.sh	cluster submit shell script
config.json	EDGE configuration in JSON
config.txt	EDGE configuration in text
error.log	project error log
final_report.pdf	result pdf
HTML_Report	result HTML directory
JBrowse	JBrowse tracks directory
metadata_gisaid_ncbi.txt	project sample metadata for gisaid and ncbi
metadata_run.txt	project run ID
process_current.log	project last run log
process.log	project overall process log
QcReads	QC processed directory
all.1.fastq	input forward reads
all.2.fastq	input reverse reads
fastqCount.txt	stats for fastq count
QC.1.trimmed.fastq	trimmed forward reads
QC.2.trimmed.fastq	trimmed reverse reads

- QC.log	QC log
- QC_qc_report.pdf	QC stats report pdf
- QC.stats.txt	QC stats text
- QC.unpaired.trimmed.fastq	trimmed single end/orphan reads
- ReadsBasedAnalysis	
└─ readsMappingToRef	
- AlignTrimMapping/	Align Trim reads fastq directory
- consensus.log	consensus workflow log
- Coverage_plots/	reference coverage and histogram png files directory
- GapVSReference.report.json	Gap analysis result json file
- GapVSReference.report.txt	report of All reference genes affected by gap regions
- mapping.log	reads mapping log
- NC_045512.2.alnstats.txt	reads mapping to NC_045512 stats
- NC_045512.2_consensus.changelog	consensus workflow nucleotide changes info
- NC_045512.2_consensus.fasta *	consensus fasta file
- NC_045512.2_consensus_w_ambiguous.fasta *	consensus fasta file with IUPAC code
- NC_045512.2_consensus.fasta.comp	consensus fasta nucleotide composition
- NC_045512.2_consensus_w_ambiguous.fasta.comp	consensus fasta with IUPAC code nucleotide composition

NC_045512.2_consensus.gaps	no reads mapped to NC_045512 regions
NC_045512.2_consensus.gaps_report.txt	report of NC_045512 genes affected by gap regions
NC_045512.2_consensus.Indels_report.txt	report of NC_045512 genes affected by INDELs
NC_045512.2_consensus_lineage.txt	report of Pangolin lineage assignment
NC_045512.2_consensus.SNPs_report.txt	report of NC_045512 genes affected by SNPs
NC_045512.2_consensus_w_ambiguous.SNPs_report.txt	report of NC_045512 genes affected by SNPs with IUPAC code
NC_045512.2.sort.bam *	reads mapping bam file
NC_045512.2.sort.bam.bai	reads mapping bam index file
NC_045512.2.vcf	variant call of reads mapping to NC_045512 result by bcftools
pangolin.log	Pangolin lineage run log
readsToRef.alnstats.txt	reads mapping to All reference stats
readsToRef.gaps	reads mapping to All reference gaps
readsToRef.Indels_report.txt	report of All reference genes affected by INDELs
readsToRef_NC_045512_2_amplicon_coverage.html	amplicon coverage interactive plot
readsToRef_NC_045512_2_amplicon_coverage.txt	amplicon coverage based on primer scheme bed file
readsToRef_NC_045512_2.coverage	NC_045512 genome coverage per base position

readsToRef_plots.pdf	coverage plots of reads mapping to all reference
readsToRef.SNPs_report.txt	report of All reference genes affected by SNPs
readsToRef.vcf	variant call of reads mapping to all reference result by bcftools
swift_primer_schemes_v2.bed	bed file for primer trimming and amplicon coverage plot
└ variantAnalysis.log	variants gene analysis
└ Reference	
NC_045512.2.fasta	input reference fasta
reference.fasta	all input reference fasta
reference.gbk	all input reference genbank
reference.gff	all input reference gff (convert from gbk)
└ ref_list.txt	the accession list of input reference
└ ReferenceBasedAnalysis	
└ readsMappingToRef	symlink directory from ReadsBasedAnalysis
└ UPLOAD	
sra_experiments.txt	metadata for SRA submission
└ sra_samples.txt	metadata for SRA submission

## ThirdParty Tools

- Alignment

- Bowtie 2
    - Citation: Langmead, B. and Salzberg, S.L. (2012) [Fast gapped-read alignment with Bowtie 2](#), *Nature methods*, 9, 357-359.
    - Site: <http://bowtie-bio.sourceforge.net/bowtie2/index.shtml>
    - Version: 2.4.1
    - License: GPLv3
  - BWA
    - Citation: Li, H. and Durbin, R. (2009) [Fast and accurate short read alignment with Burrows-Wheeler transform](#), *Bioinformatics*, 25, 1754-1760.
    - Site: <http://bio-bwa.sourceforge.net/>
    - Version: 0.7.12
    - License: GPLv3
  - minimap2
    - Citation: Li, H. (2018) [Minimap2: fast pairwise alignment for nucleotide sequences](#). *Bioinformatics*, 34:3094-3100.
    - Site: <https://github.com/lh3/minimap2>
    - Version: 2.17
    - License: MIT
  - Kallisto
    - Citation: Nicolas L Bray, et al. (2016) [Near-optimal probabilistic RNA-seq quantification](#), *Nature Biotechnology* 34, 525–527
    - Site: <https://pachterlab.github.io/kallisto/>
    - Version: 0.46.0
    - License: BSD 2-Clause
- Annotation
    - RATT
      - Citation: Otto, T.D., et al. (2011) [RATT: Rapid Annotation Transfer Tool](#), *Nucleic acids research*, 39, e57.
      - Site: <http://ratt.sourceforge.net/>
      - Version:
      - License: GPLv3

- Note: The original RATT program does not deal with reverse complement strain annotations transfer. We edited the source code to fix it.

- Assembly

- IDBA-UD
  - Citation: Peng, Y., et al. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *Bioinformatics*, 28, 1420-1428.
  - Site: [http://i.cs.hku.hk/~alse/hkubrg/projects/idba\\_ud/](http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/)
  - Version: 1.1.1
  - License: GPLv2
- SPAdes
  - Citation: Nurk, Bankevich et al. (2013) Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol.* 2013 Oct;20(10):714-37
  - Site: <http://bioinf.spbau.ru/spades>
  - Version: 3.13.0
  - License: GPLv2
- MEGAHIT
  - Citation: Li D. et al. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. 2015 May 15;31(10):1674-6
  - Site: <https://github.com/voutcn/megahit>
  - Version: 1.2.9
  - License: GPLv3
- LRASM: Long Read Assembler
  - Citation:
  - Site: [https://gitlab.com/chienchi/long\\_read\\_assembly](https://gitlab.com/chienchi/long_read_assembly)
  - Version: 0.1.0
  - License: GPLv3
- RACON

- Citation: Vaser R et al.(2017) [Fast and accurate de novo genome assembly from long uncorrected reads](#). *Genome Res.* 2017 May;27(5):737-746.
  - Site: <https://github.com/isovic/racon>
  - Version: 1.4.13
  - License: MIT
- Unicycler
  - Citation: Wick RR et al.(2017) [Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads](#). *PLoS Comput Biol.* 2017 Jun 8;13(6):e1005595.
  - Site: <https://github.com/rrwick/Unicycler>
  - Version: 0.4.8
  - License: GPLv3
- Lineage Assignment
  - Pangolin
    - Citation: Andrew Rambaut et al. (2020) [A dynamic nomenclature proposal for SARS-CoV-2 to assist genomic epidemiology](#). *Nat Microbiol.* 2020 Nov;5(11):1403-1407.
    - Site: <https://pangolin.cog-uk.io/>
    - Version: 3.1.20
    - License: GPLv3
- Reads Quality Control
  - FaQCs
    - Citation: Chienchi Lo, PatrickS.G. Chain (2014) [Rapid evaluation and Quality Control of Next Generation Sequencing Data with FaQCs](#). *BMC Bioinformatics.* 2014 Nov 19;15
    - Site: <https://github.com/LANL-Bioinformatics/FaQCs>
    - Version: 2.09
    - License: GPLv3
  - NanoPlot
    - Citation: De Coster W, et al.(2018) [NanoPack: visualizing and processing long read sequencing data](#), *Bioinformatics.* 2018 Mar 14.
    - Site: <https://github.com/wdecoster/NanoPlot>

- Version: 1.13.0
  - License: GPLv3
- Porechop
  - Citation: Wick RR, Judd LM, et al (2017). Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom.* 2017;3(10):e000132. Published 2017 Sep 14.  
doi:10.1099/mgen.0.000132
  - Site: <https://github.com/rrwick/Porechop>
  - Version: 0.2.3
  - License: GPLv3
- Align Trim
  - Site:  
[https://github.com/artic-network/fieldbioinformatics/blob/master/artic/align\\_trim.py](https://github.com/artic-network/fieldbioinformatics/blob/master/artic/align_trim.py)
  - Version:
  - License: MIT
  - Note: **The original Align Trim script does not deal with illumina reads and strandness. We edited the source code to work on it.**
- Utility
  - R
    - Citation: R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
    - Site: <http://www.r-project.org/>
    - Version: 3.6.3
    - License: GPLv2
  - GNU\_parallel
    - Citation: O. Tange (2011): GNU Parallel - The Command-Line Power Tool, ;login: The USENIX Magazine, February 2011:42-47
    - Site: <http://www.gnu.org/software/parallel/>
    - Version: 20190422
    - License: GPLv3
  - Seqtk

- Citation: Heng Li <https://github.com/lh3/seqtk>
  - Site: <https://github.com/lh3/seqtk>
  - Version: 1.3
  - License: MIT
- sratoolkit
  - Citation:
  - Site: <https://github.com/ncbi/sra-tools>
  - Version: 2.9.6
  - License: Public Domain
- ea-utils
  - Citation: Erik Aronesty (2011) ea-utils : “Command-line tools for processing biological sequencing data”
  - Site: <https://code.google.com/archive/p/ea-utils/>
  - Version: 1.1.2-537
  - License: MIT License
- Anaconda3 (Python 3)
  - Citation:
  - Site: <https://anaconda.org>
  - Version: 2020.02
  - License: 3-clause BSD

## ● Variants Calling

- SAMtools
  - Citation: Li, H., et al. (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 25, 2078-2079.
  - Site: <http://www.htslib.org/>
  - Version: 1.10
  - License: MIT
- BCFtools
  - Citation: Heng Li (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics* (2011) 27(21) 2987-93.
  - Site: <http://www.htslib.org/>
  - Version: 1.10

- License: MIT
- Visualization
  - JBrowse
    - Citation: Skinner, M.E., et al. (2009) [JBrowse: a next-generation genome browser](#), *Genome research*, 19, 1630-1638.
    - Site: <http://jbrowse.org>
    - Version: 1.16.8
    - License: Artistic License 2.0/LGPLv.1
  - IGV.js
    - Citation: James T. Robinson, et al. (2020) [igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer \(IGV\)](#). *bioRxiv* 2020.05.03075499.
    - Site: <https://igv.org/>
    - Version: 2.10.4
    - License: MIT

## FAQs

For web based app:

1. How can I view alignments in a local viewer such as IGV?

You can download the BAM file using the green download button from the *Mapped Reads* section and the index file can be downloaded by clicking the hyperlinked **Directory** and then clicking on the index file (.bai).

2. Can edge-covid19 handle PacBio data?

Our QC tool FaQC cannot process the base quality values reported by PacBio Sequel as it reports all base qualities as PHRED 0. We recommend you run QC separately or turn **OFF** the **Preprocessing** module (FaQCs will throw an error due to Quality issue), select **YES** to **Nanopore Reads in Input Raw Reads** module, and add "-x map-pb" in the **Aligner Option** field in the additional options of **Reference-Based analysis module**.

For local docker build:

## 1. How to start/stop EDGE COVID-19 docker instance?

To start or restart *EDGE COVID-19*, you will need to run following command in your Terminal from the same directory:

```
$ cd EDGE-COVID19
$ docker rm -v edge-covid19
$ docker run -d --volumes-from mysql_data \
  -v $PWD/EDGE_output:/home/edge/EDGE_output \
  -v $PWD/EDGE_input:/home/edge/EDGE_input \
  -v $PWD/EDGE_report:/home/edge/EDGE_report \
  -p 80:80 -p 8080:8080 --name edge-covid19 bioedge/edge-covid19
```

Then wait a few minutes and go to <http://localhost> in your favorite browser.

To stop the docker run following command in your directory:

```
$ docker stop edge-covid19
```

Note that the docker container will keep running in the background until you restart your computer or specifically stop it using the above command.

## 2. How to update EDGE COVID-19?

To update the image to the latest version, you can pull the docker again in the original *EDGE-COVID19* folder used in **Step 1**.

```
$ docker pull bioedge/edge-covid19
```

After pulling the latest docker, start the image from terminal:

```
$ cd EDGE-COVID19
$ docker rm -v edge-covid19
$ docker run -d --volumes-from mysql_data \
  -v $PWD/EDGE_output:/home/edge/EDGE_output \
  -v $PWD/EDGE_input:/home/edge/EDGE_input \
  -v $PWD/EDGE_report:/home/edge/EDGE_report \
  -p 80:80 -p 8080:8080 --name edge-covid19 bioedge/edge-covid19
```

### 3. How long will it take to run my sample?

Using a Macbook Pro with 16GB RAM and 8 processors available:

dataset size (bases)	# Raw reads	Type of data (Nanopore/ Illumina)	Protocol	# of CPUs	Total Wall Clock Time
416,793,360	10,493,168	Illumina	Amplicons	4	2:38:01
594,064,863	1,382,016	Nanopore	ARTIC protocol	4	0:21:07

### 4. I am getting an error while pulling the image. What can I do?

If you have issues pulling this image, you may increase the basesize when launching docker daemon or use a different [Storage Driver](#). See similar [issue](#) here.

### 5. I am getting an error while trying to login on GUI: “Failed to login in. Please check server log for details”

In some linux environments, users need to set the /path/to/mysql directory into 0777 mode.

Please try opening the directory permissions if you run into trouble.

```
$ docker pull bioedge/edge_ubuntu_mysql  
$ docker create --name mysql_data --volume /var/lib/mysql  
bioedge/edge_ubuntu_mysql  
$ docker run -d --volumes-from mysql_data \  
-v $PWD/EDGE_output:/home/edge/EDGE_output \  
-v $PWD/EDGE_input:/home/edge/EDGE_input \  
-v $PWD/EDGE_report:/home/edge/EDGE_report \  
-p 80:80 -p 8080:8080 --name edge-covid19 bioedge/edge-covid19
```

### 6. IP address conflicts

Docker is hard coded to look for 172.17.0.1. If the IP address conflicts with the subnet of your WiFi, you may need to customize the docker bridge by editing the /etc/docker/daemon.json as described [here](#).

## 7. What are the commands for checking status and error log?

Check the MySQL status in container:

```
$ docker exec edge-covid19 service mysql status
```

where "edge-covid19" is the container name when using `docker run` with --name flag  
Check container status.

```
$ docker ps -a
```

Check user management system service status:

```
$ docker exec edge-covid19 service tomcat7 status
```

Check the Apache web server status and log:

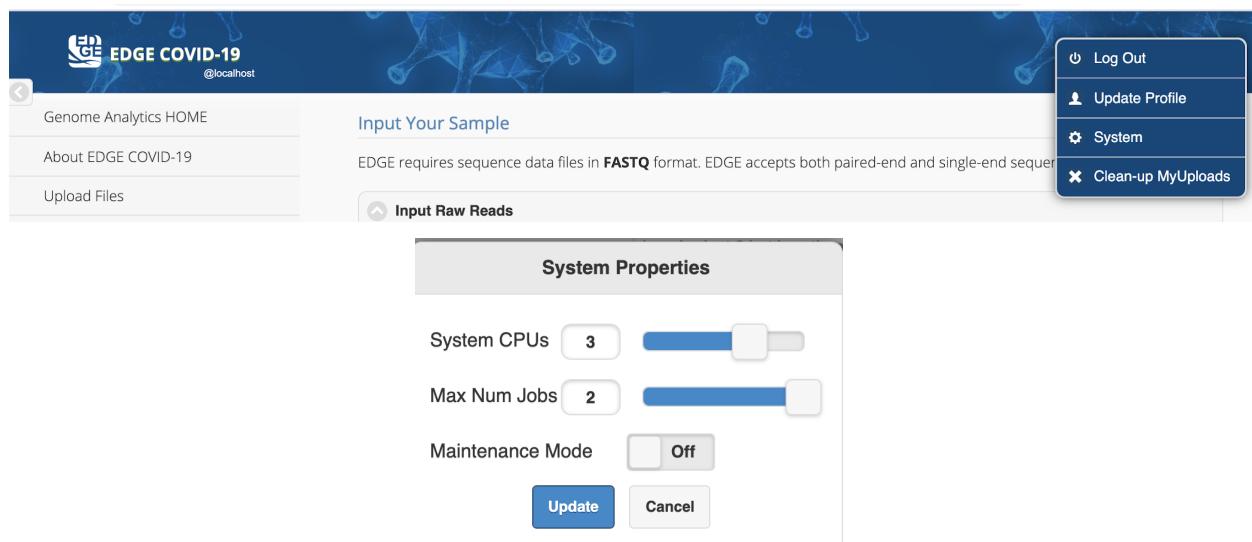
```
$ docker exec edge-covid19 service apache2 status
$ docker exec edge-covid19 tail /var/log/apache2/error.log
$ docker exec edge-covid19 tail /var/log/apache2/access.log
```

## 8. How can I update the number of CPUs that EDGE COVID-19 uses?

\*

The default number of CPUs available to EDGE inside the container is 4 and the maximum number of jobs can run simultaneously is 2.

Each job will use  $(\text{edge\_system\_cpu-1})/\text{max\_num\_jobs}$  in integer CPUs (as you see on the GUI). These numbers can be changed by login using an admin account and click on the user name to pop up the user menu where you can click the **System** button to open the system properties menu.



## Contact:

You can view the discussions in the google group below and join the group to post questions and/or comments.

EDGE user's google group at <https://groups.google.com/d/forum/edge-users>

You can also directly contact us through email at [edge-covid19@lanl.gov](mailto:edge-covid19@lanl.gov)

## Citation:

Lo, Chien-Chi, Migun Shakya, Karen Davenport, Mark C. Flynn, Jason D. Gans, Adán Myers y Gutiérrez, Bin Hu, Po-E Li, Elais Player Jackson, Yan Xu and Patrick S. G. Chain. "EDGE COVID-19: A Web Platform to generate submission-ready genomes for SARS-CoV-2 sequencing efforts." (2020).

<https://arxiv.org/abs/2006.08058>