EDGE COVID-19 documentation

Los Alamos National Laboratory

Email us: edge-covid19@lanl.gov
v1.0.9

Table of Contents **2**

Overview **3**

A step by step guide to running EDGE COVID-19: **4**

Step 1: Create an account 4

Step 2: Upload your raw reads 4

Step 3: Run your sample 6

EDGE COVID-19 output page **11**

FAQs **14**

For web based app: 14

How can I view alignments in a local viewer such as IGV? 14

For local docker build: **14**

1. How to start/stop EDGE COVID-19 docker instance? 14
2. How to update EDGE COVID-19? 14
3. How long will it take to run my sample? 15
4. I am getting an error while pulling the image. What can i do? 15
6. IP address conflicts 16
7. What are the commands for checking status and error log? 16
8. How can I update the number of CPUs that EDGE COVID-19 uses? 16

Contact: **17**

Citation: **17**

Overview

EDGE COVID-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) 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:
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) a Single Nucleotide Polymorphism and Variant analyses, with detail such as location and resulting coding differences.

The *EDGE COVID-19* platform can accommodate Illumina or ONT data, including ONT data from the SARS-CoV-2 ARTIC network sequencing protocols. Users can input/upload Illumina or Nanopore sequencing FASTQ files (and/or download from NCBI SRA). For Illumina data, default analyses include read QC, read mapping to the reference, and SNP/variant analysis. For ONT data, the data must be demultiplexed prior to uploading; the samples will be processed individually. The SNP/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 both 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.

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. This light-weight version is also available as a Docker container, 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).

Columns...

Tool	#Reads	%Reads	Level	Top1	Top2	Top3	Top4	Top5
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 to running *EDGE COVID-19*:

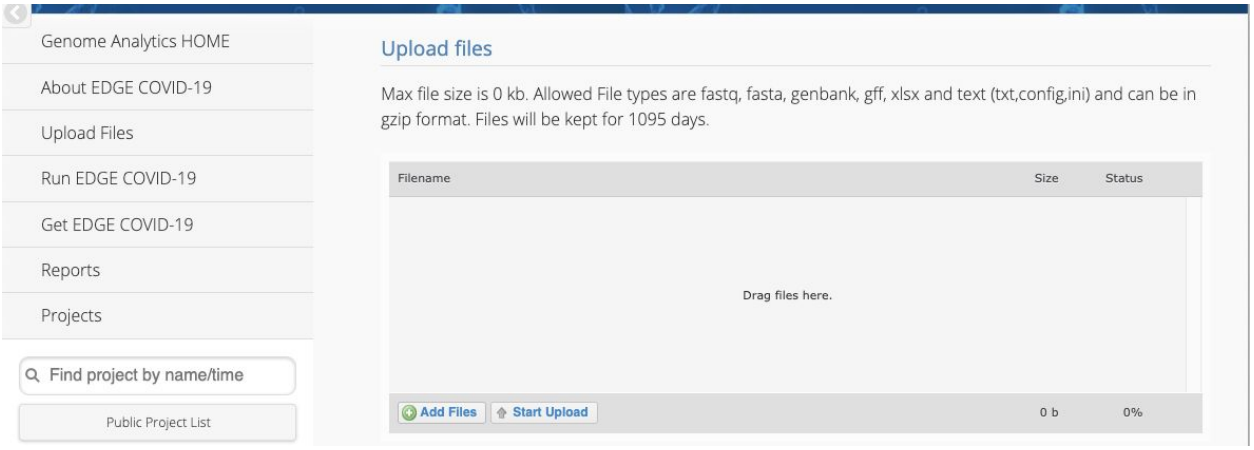
If you want to run the analyses in the cloud visit https://edge-covid19.edgebioinformatics.org/ and follow the steps below:

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.

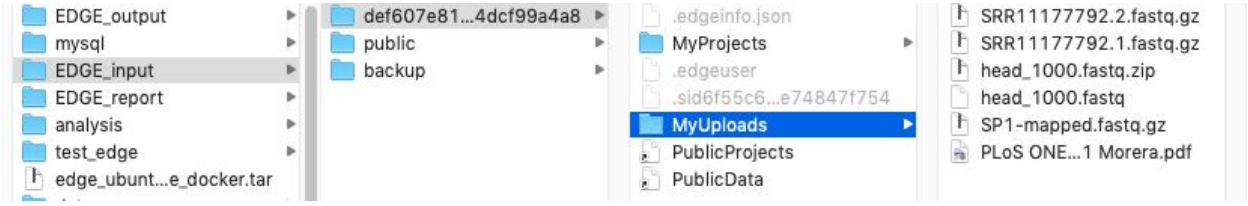
Step 2: Upload your raw reads

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.

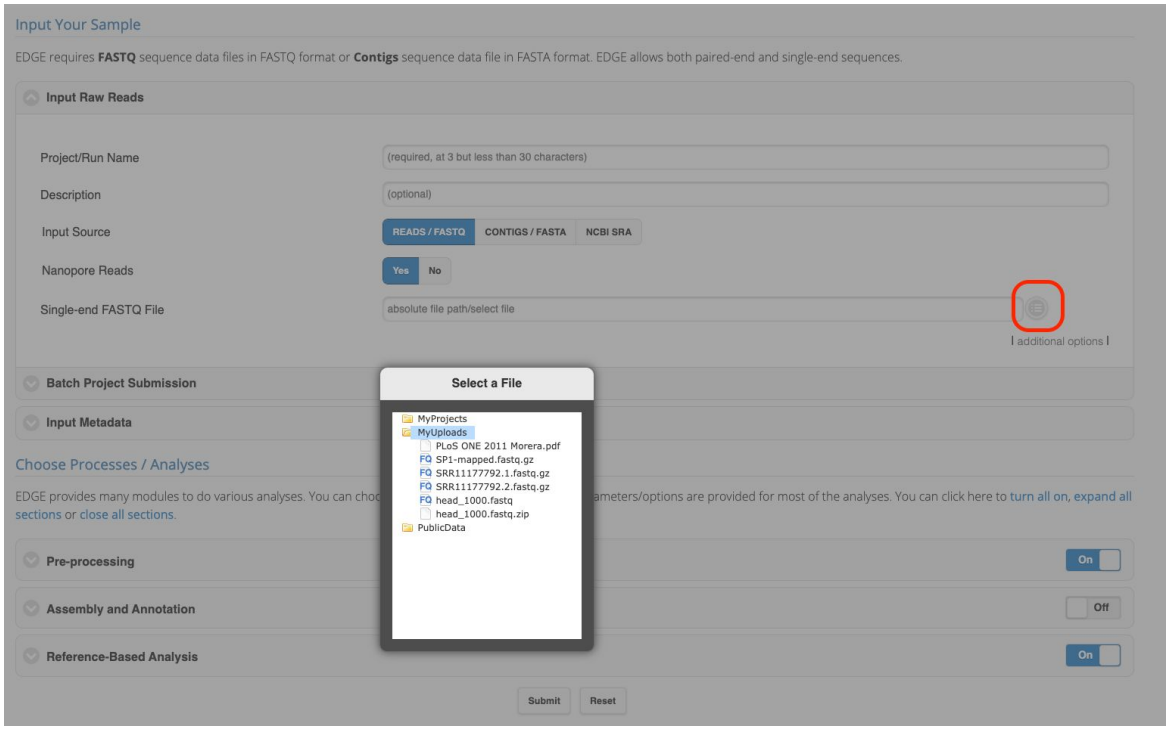


For a local installation:

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 dases, 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 **Yes** in **Nanopore Reads** if your sample was generated using Nanopore; select **No** if it's Illumina data.
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 (v1-3) and CDC protocols 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 v1.1.3, UniCycler v0.4.8, wtdbg2 v2.5, 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 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 is also applied.

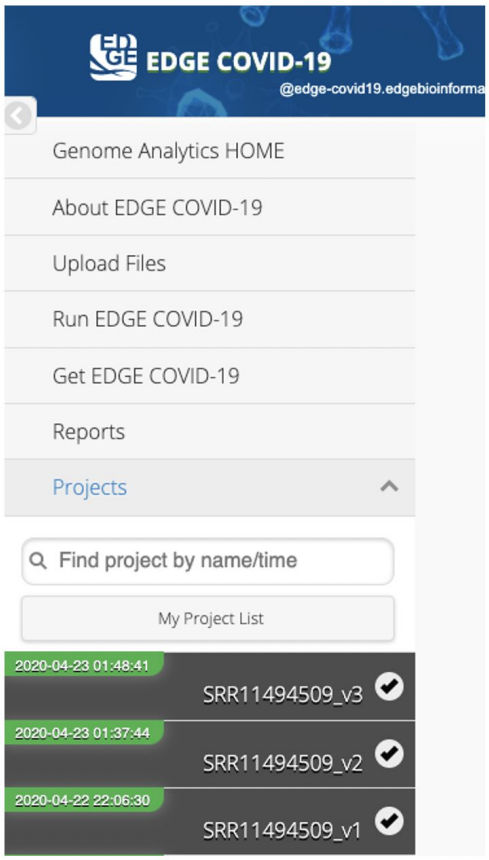
    Variant calling: *EDGE COVID-19* uses **bcftools** *mpileup* command to convert the aligned BAM file into genomic positions and call genotypes, reduce the list of sites to those found to be variants by passing this file into **bcftools** *call* command. The variant calls are filtered further by vcfutils.pl of SAMtools with following criteria:
    - Minimum Root Mean Square (RMS) mapping quality for SNPs [10];

- minimum read depth [5];
- maximum read depth [300];
- 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 [0.0001];
- maximum fraction of reads supporting an indel [0.5];

The consensus workflow: *EDGE COVID-19* uses a maximum of 8000X depth coverage reads for computational efficiency. 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 <7 for ONT), Base Alignment Quality for Illumina, 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 0.8 for ONT), and minimum mapping quality of 60.

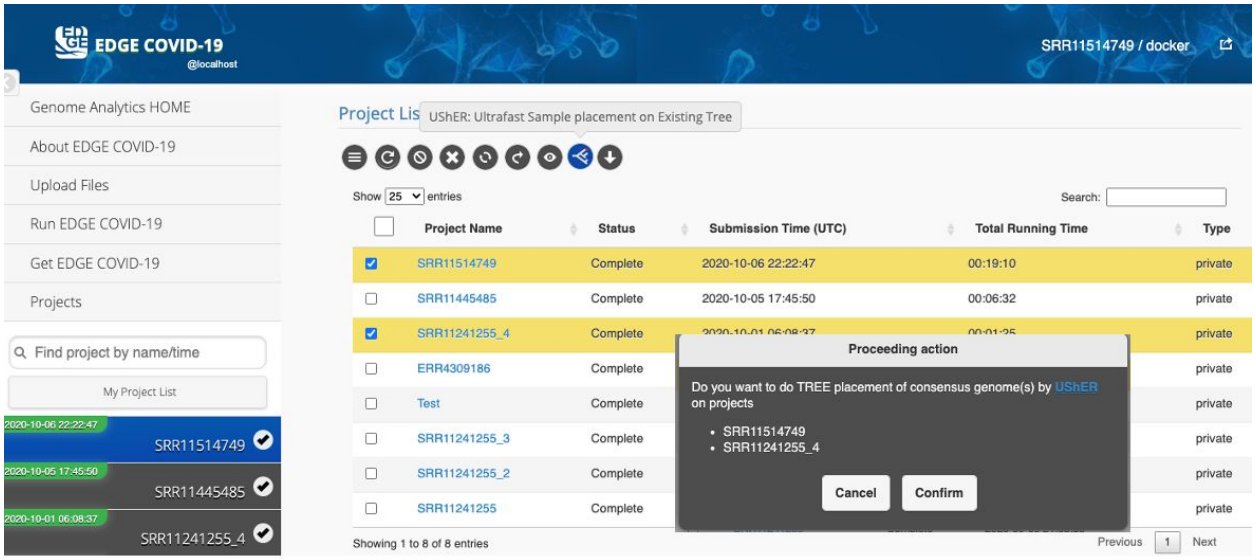
12. Click the **submit** button at the bottom of the page to start your job

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.



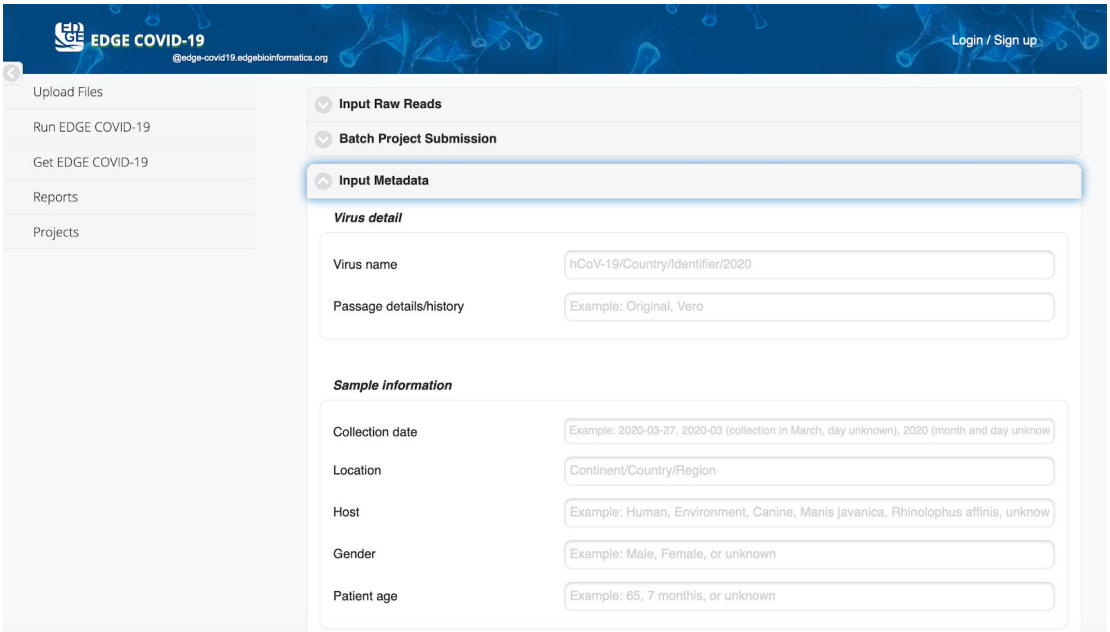
If a project is highlighted in **green** it means the project has finished; if **orange**, **red**, or **grey** (not shown) it means the project is running, cancelled/failed, or not yet started, respectively. You can access the details and outputs of each project by clicking on the project in the list. 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).

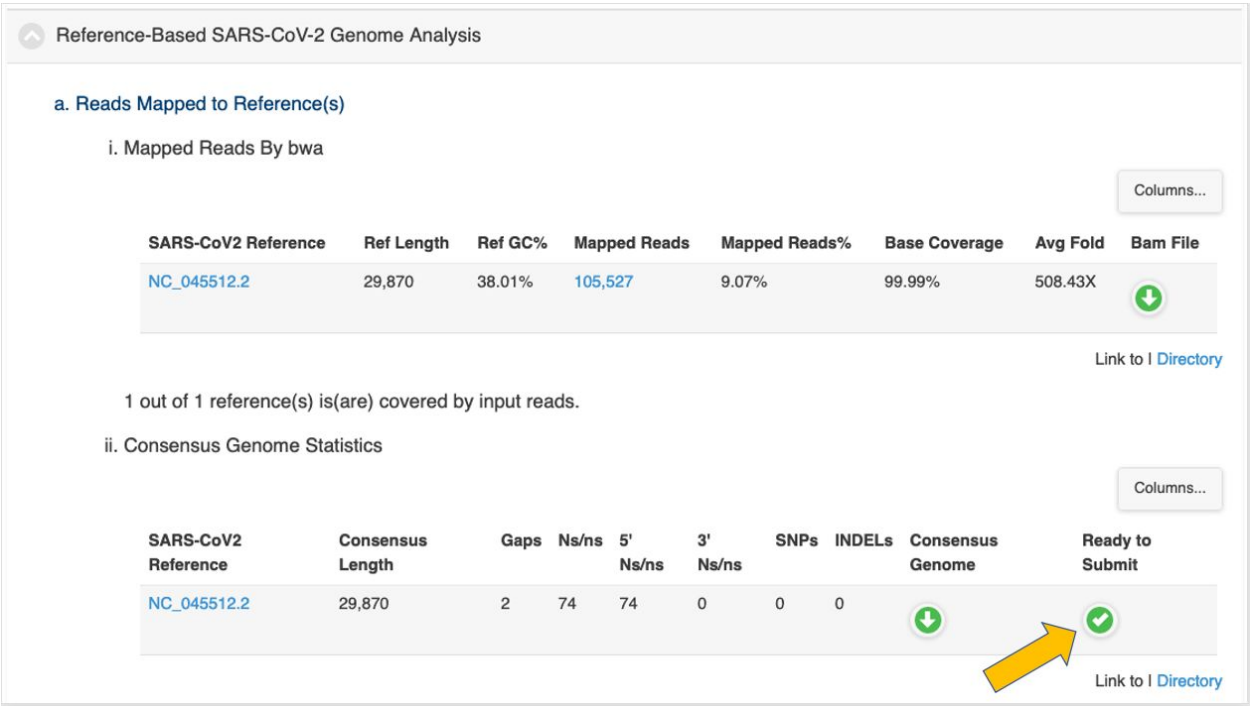


Fasta Sequence	Size (?)	#Ns (?)	#Mixed (?)	Bases aligned (?)	Insertions (?)	Deletions (?)	#SNVs used for placement (?)	#Masked SNVs (?)	Neighboring sample in tree (?)	Lineage of neighbor (?)	#Imputed values for mixed bases (?)	#Maximally parsimonious placements (?)	Parsimony score (?)	Subtree number (?)
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_426416 | 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_426416 | 20-03-08	A.1	0	1	0	3 (view in Nextstrain)

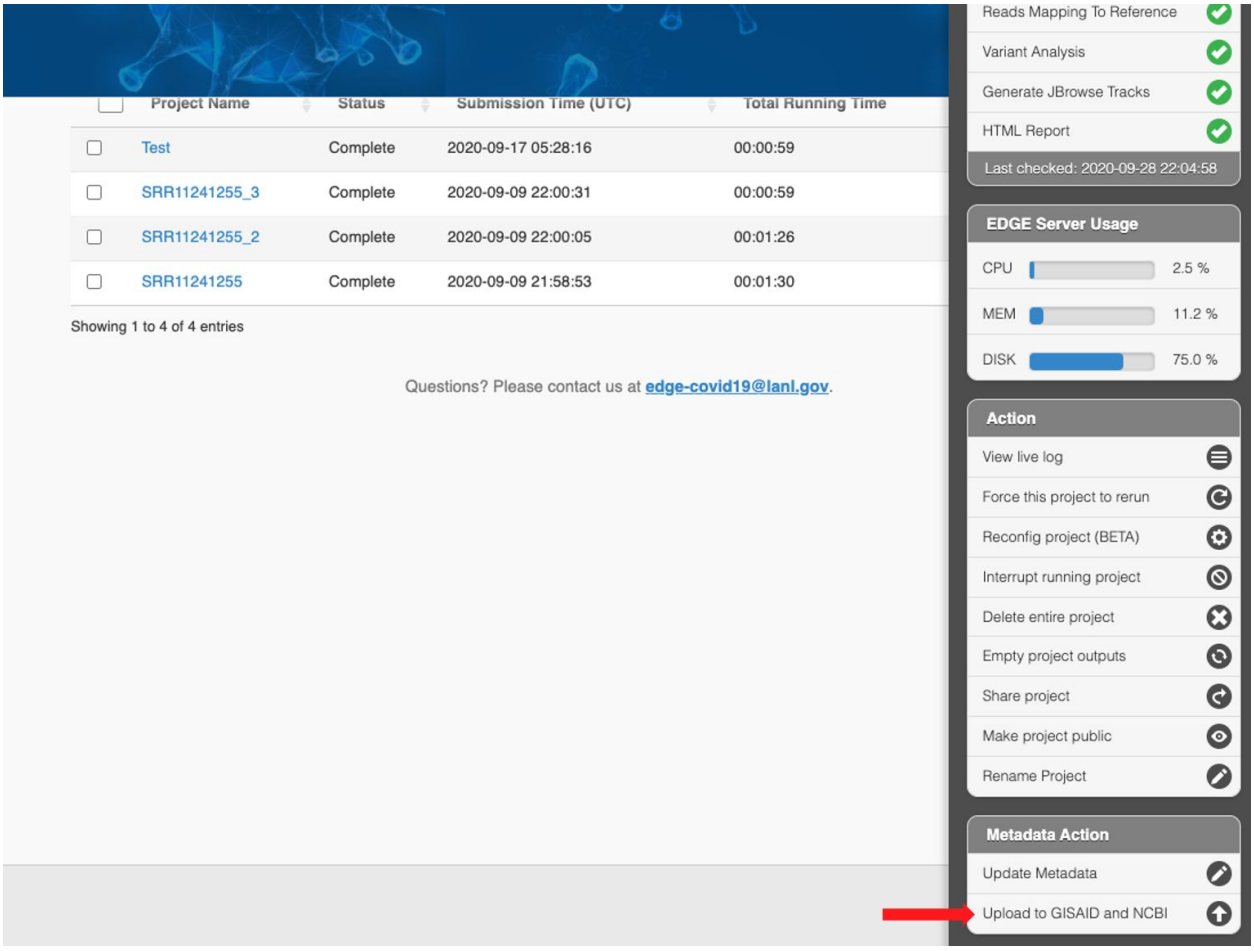
15. Before submitting the pipeline run, you can prepare your genome for submission to GISAID and NCBI by inputting the metadata. You can do this after the pipeline is run as well.

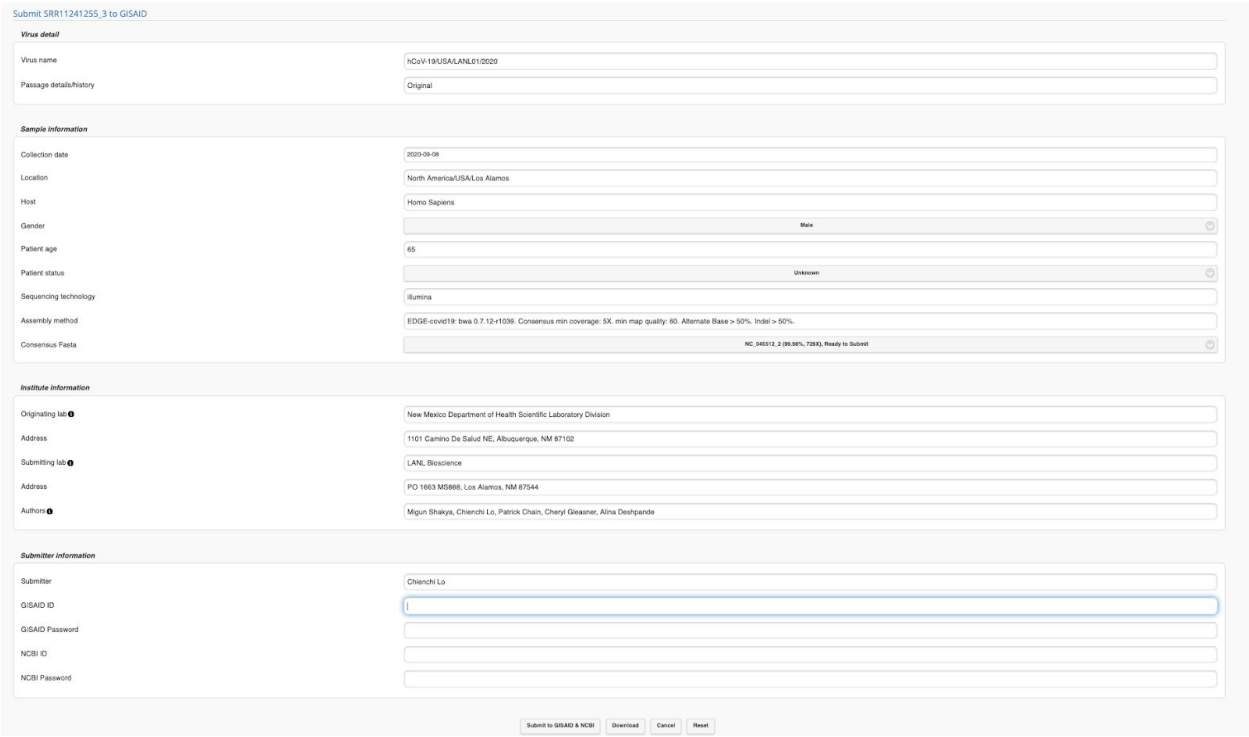


After entering the metadata, the genome can be submitted to GISAID and NCBI directly through the *EDGE COVID-19* platform. You can access this functionality by clicking on the green check mark in the Reference-based results just below “Ready to Submit”.



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.


Project Name
Status
Submission Time (UTC)
Total Running Time
Test
Complete
2020-09-17 05:28:16
00:00:59
SRR11241255_3
Complete
2020-09-09 22:00:31
00:00:59
SRR11241255_2
Complete
2020-09-09 22:00:05
00:01:26
SRR11241255
Complete
2020-09-09 21:58:53
00:01:30
Showing 1 to 4 of 4 entries
Questions? Please contact us at edge-covid19@lanl.gov.
Reads Mapping To Reference
Variant Analysis
Generate JBrowse Tracks
HTML Report
Last checked: 2020-09-28 22:04:58
EDGE Server Usage
CPU 2.5 %
MEM 11.2 %
DISK 75.0 %
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

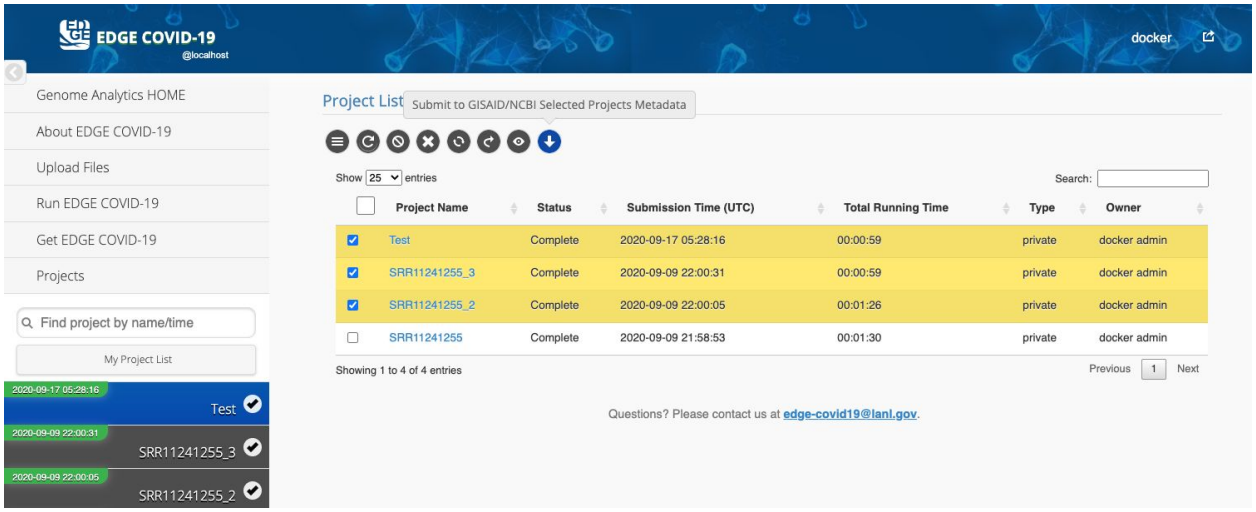


Required metadata fields must be properly filled for submission to proceed.

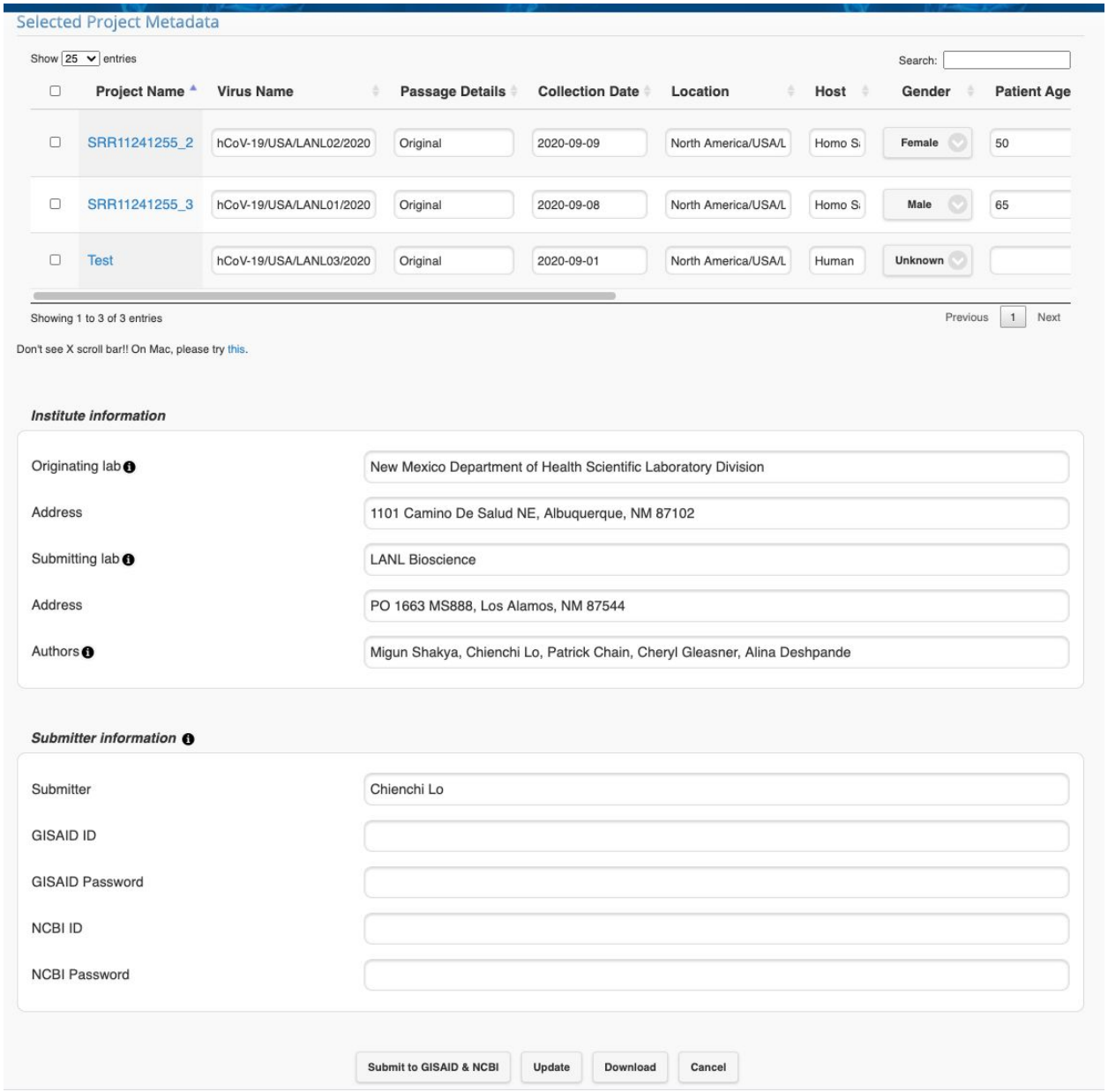
Note: this feature is in beta format, and GISAID and NCBI can change the submission process at any time; if you run into any trouble, you can contact us at edge-covid19@lanl.gov.

15. Batch submit:

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 action button at top of the table.



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.)



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

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.

EDGE COVID-19 output page

For a more detailed description of the EDGE bioinformatics output page, please refer to our full documentation here. 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). 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

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%	119,542	98.90%	99.54%	331.94X	

Link to I 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	SNPs	INDELs	Consensus Genome	Ready to Submit
NC_045512.2	29,870	43	591	82	34	5	0		

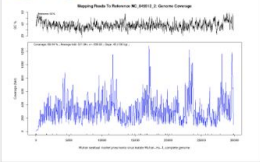
Link to I Directory

iii. Variant Call

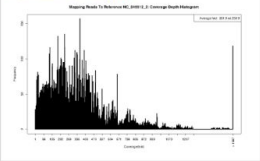
Columns...

SARS-CoV2 Reference	Variants	INDELs
NC_045512.2	5	0

NC_045512_2 Cov [full]



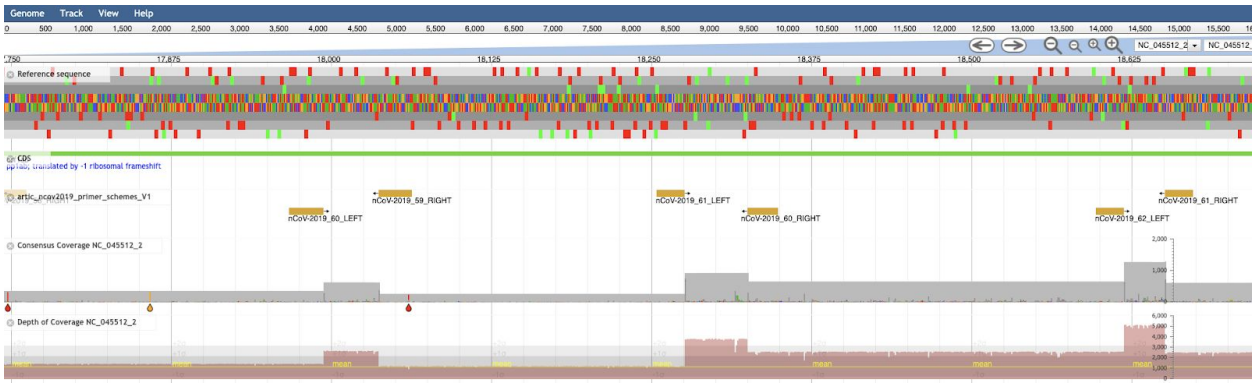
NC_045512_2 Fold [full]



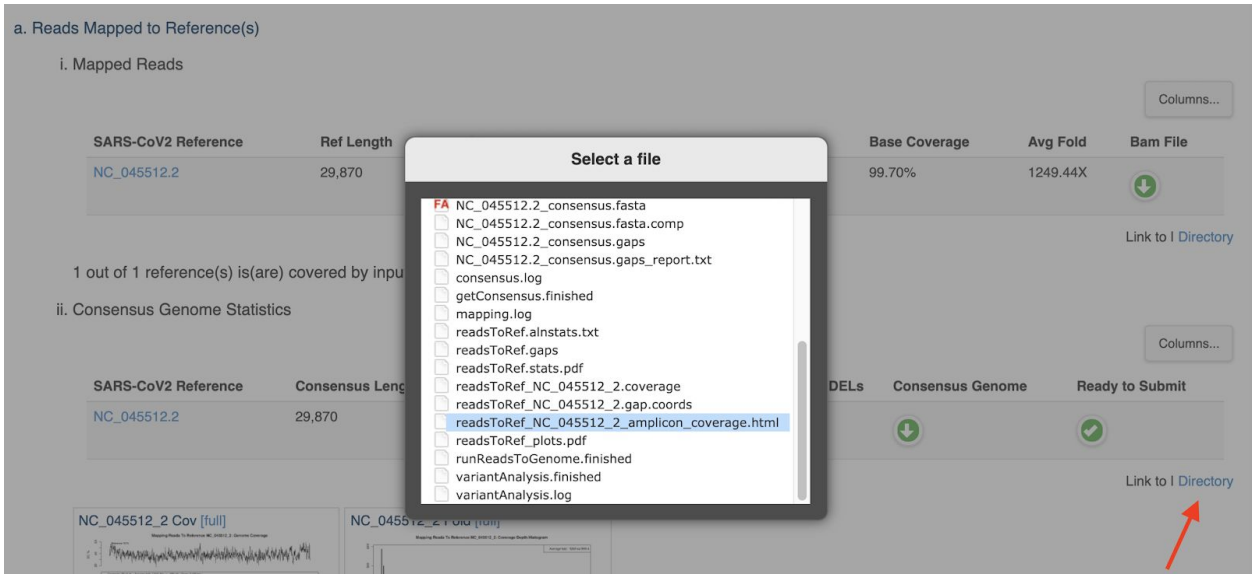
Show the results in JBrowse

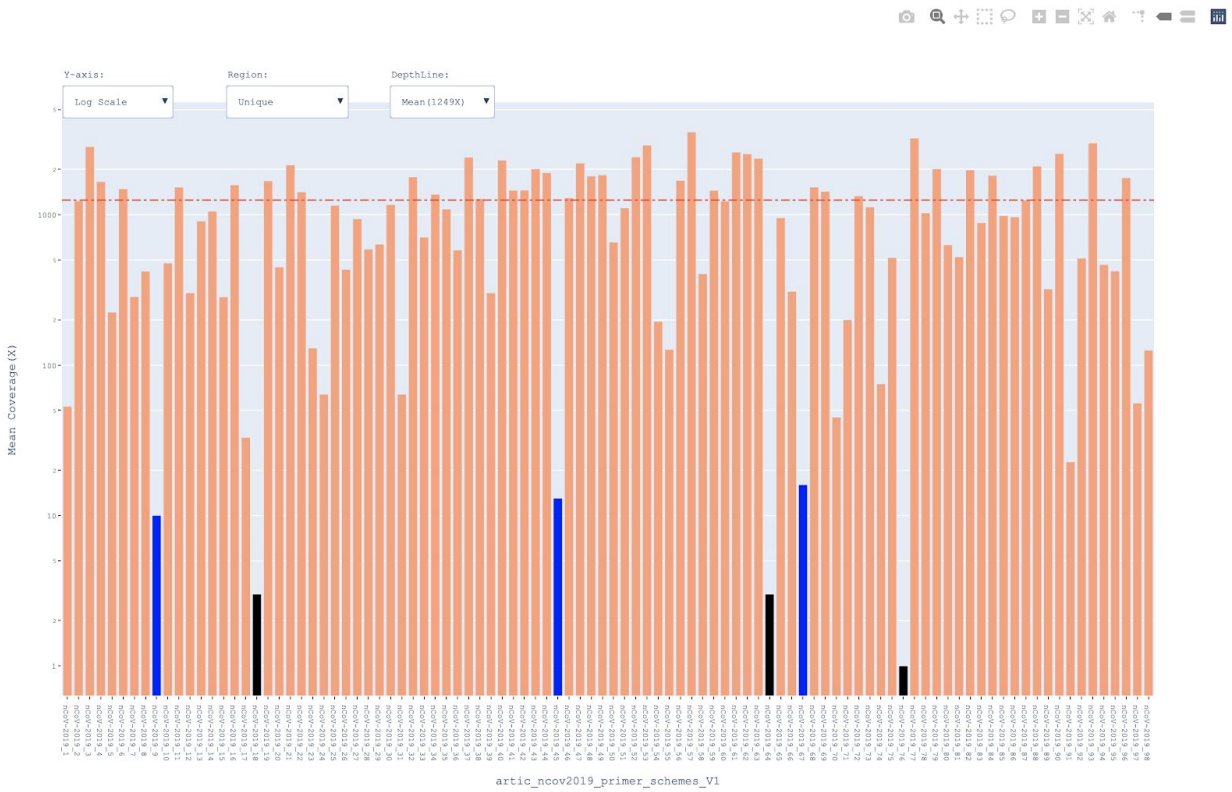
Link to I All Plots PDF I SNP Report I INDELs Report I Gap Table I Directory

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
Columns...
SARS-CoV2 Reference
Ref Length
Base Coverage
Avg Fold
Bam File
NC_045512.2
29,870
99.70%
1249.44X
Link to I Directory
1 out of 1 reference(s) is(are) covered by inpu
ii. Consensus Genome Statistics
Columns...
SARS-CoV2 Reference
Consensus Leng
DELs
Consensus Genome
Ready to Submit
NC_045512.2
29,870
Link to I Directory
NC_045512_2 Cov [full]
Select a file
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


Y-axis:
Log Scale
Region:
Unique
DepthLine:
Mean(1249X)
Mean Coverage(X)
artic_ncov2019_primer_schemes_V1

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
```

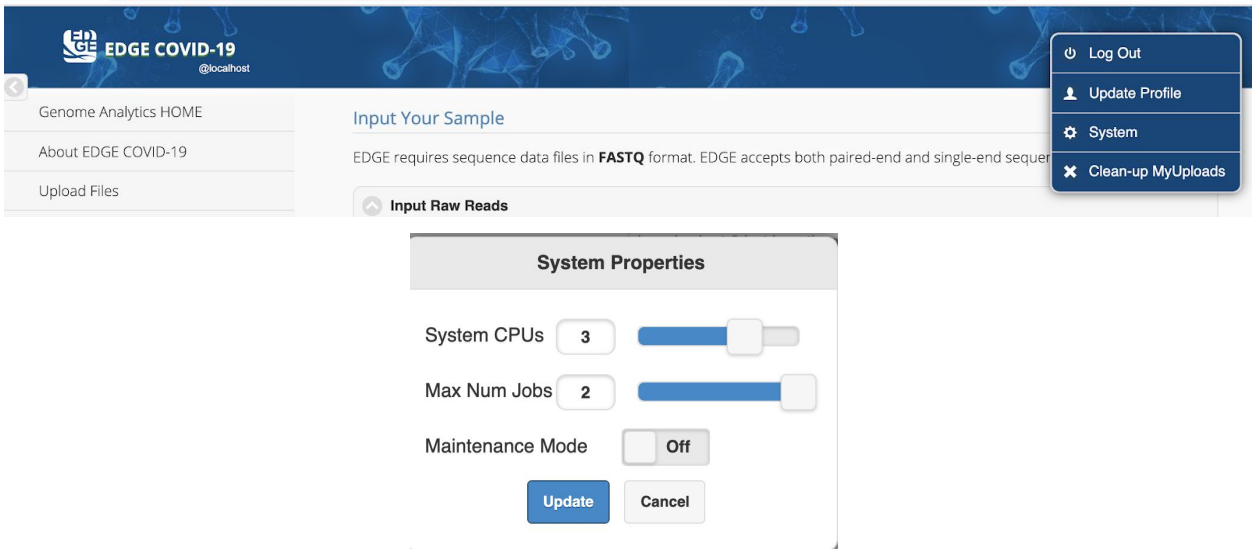
Check the Apache web server status and log:

```
$ 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 (*edge_system_cpu-1)/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

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