EDGE COVID-19 documentation

Los Alamos National Laboratory

v1.0.2

Overview

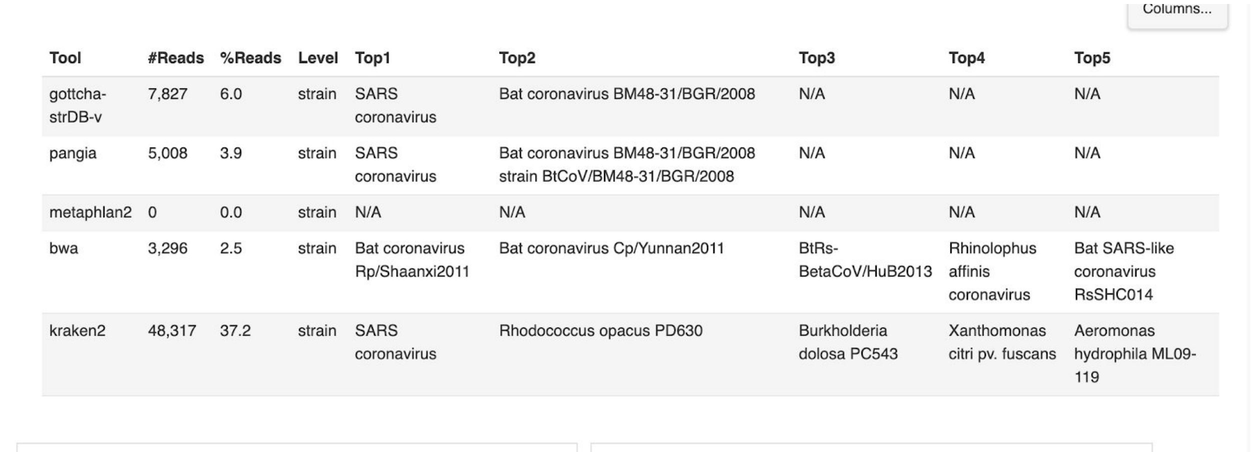
EDGE COVID-19 is a tailored bioinformatics platform based on the more flexible and fully open-source EDGE Bioinformatics software (Li et al. 2017). This mini-version consists of a user-friendly GUI that drives standardized workflows for genome reference-based 'assembly' and preliminary analysis of Illumina 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/or 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),
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. SRR11177792) and ONT (e.g. SRR11300652) 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).*

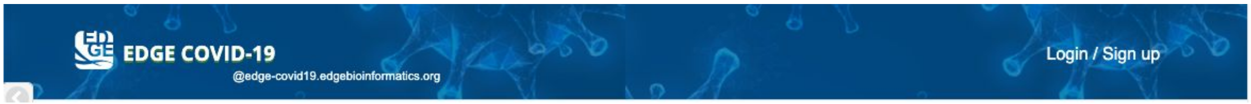


Tool	#Reads	%Reads	Level	Top1	Top2	Top3	Top4	Top5
gottcha-strDB-v	7,827	6.0	strain	SARS coronavirus	Bat coronavirus BM48-31/BGR/2008	N/A	N/A	N/A
pangia	5,008	3.9	strain	SARS coronavirus	Bat coronavirus BM48-31/BGR/2008 strain BtCoV/BM48-31/BGR/2008	N/A	N/A	N/A
metaphlan2	0	0.0	strain	N/A	N/A	N/A	N/A	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 EDGECOVID19:

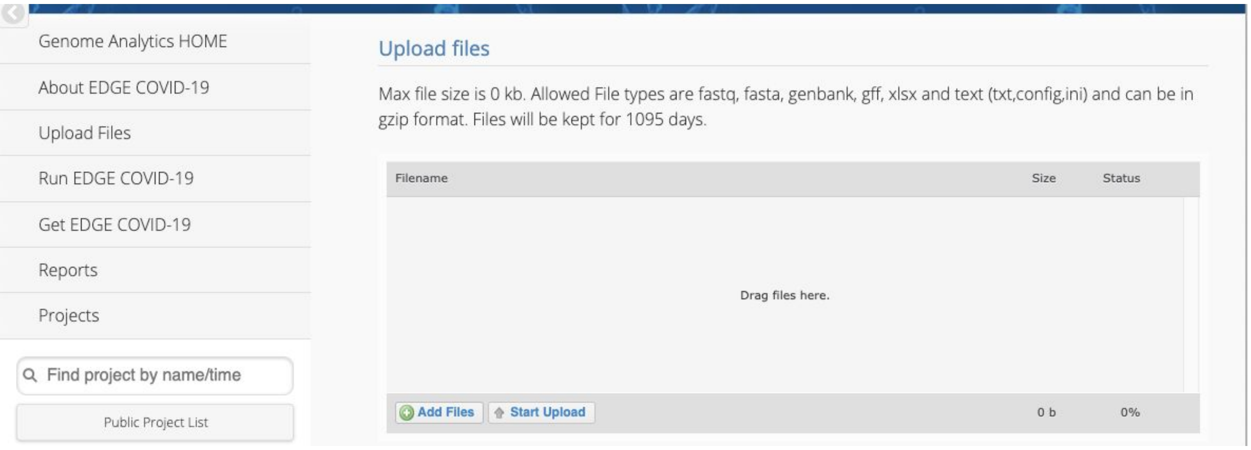
Step 1: Create an account

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” in all subsequent visits to provide your 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.

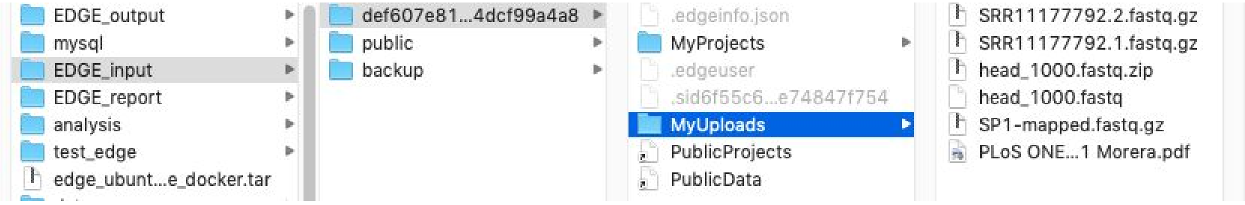


Upload files

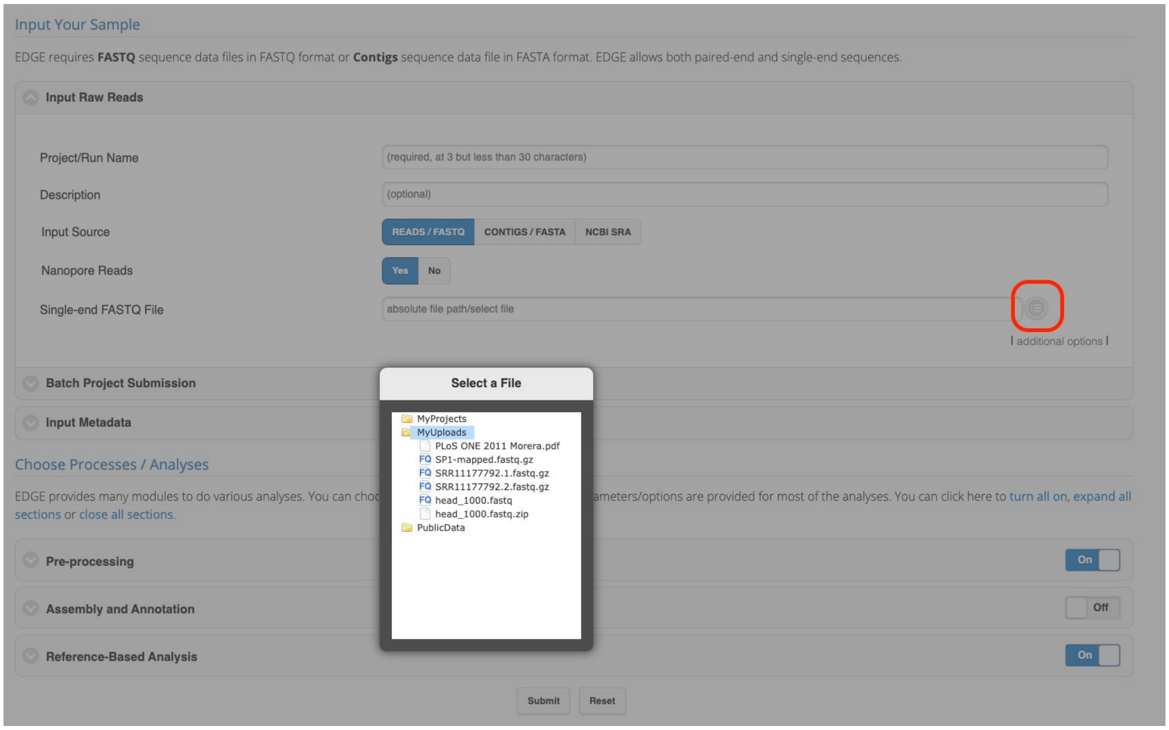
Max file size is 0 kb. Allowed File types are fastq, fasta, genbank, gff, xlsx and text (txt,config,ini) and can be in gzip format. Files will be kept for 1095 days.

For a local installation:

The easiest way is to put your data in the upload folder, which can be found within the `EDGE_input` folder that you created when installing edge-covid19 docker [see here] . Within `EDGE_input`, there will be a folder with a really long 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.


Input Your Sample

EDGE requires **FASTQ** sequence data files in FASTQ format or **Contigs** sequence data file in FASTA format. EDGE allows both paired-end and single-end sequences.

EDGE provides many modules to do various analyses. You can choo ... ameters/options are provided for most of the analyses. You can click here to turn all on, expand all sections or close all sections.

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 at https://edge.readthedocs.io.

In the *EDGE COVID-19* web server,

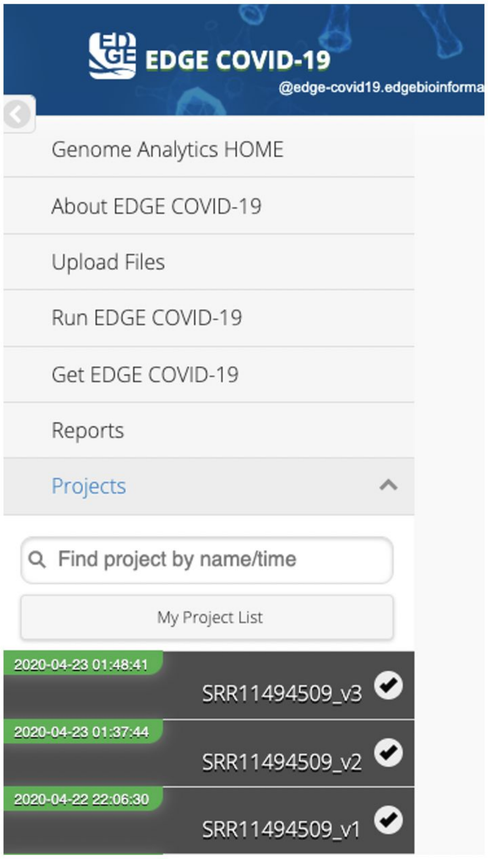
1. Type in a unique **Project/Run Name** with no spaces- use underscores, if needed.
2. Write in a short **Description-** spaces are allowed.
3. In the **Input Source** section, select **READS/FASTQ** for analyzing your own raw reads or **NCBI SRA** if you want to analyze COVID-19 samples deposited in 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 can have all the information that you need when you are ready to submit genomes to NCBI or GISAID.
8. **Pre-processing** (Data QC) is turned **ON** by default. 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 you wish. The default parameters are as follows:
   - Trim Quality Level: 20 (Illumina), 7(Nanopore)
   - Minimum Read Length: 50 (Illumina), 400(Nanopore)
   - "N" Base Cutoff: 10
   - Low Complexity Filter: 0.85
   - Trim ARTIC Protocol Primers: V3

9. For those with wgs data and interested in *de novo* assembly, you can also turn on `**Assembly and Annotation**`. Currently we provide IDBA_UD, SPAdes, MEGAHIT, UniCycler, wtdbg2, and miniasm as options for assembly.
10. **Reference-Based SARS-CoV-2 Genome Analysis** is turned **ON** by default. For Illumina data, **BWA mem** is the default aligner for mapping reads to the reference genome; this is automatically followed by variant call and generation of a consensus sequence. For ONT data, **minimap2** is the default aligner and is followed by the generation of a consensus. However, variant call based on reads can take well over 24 hours on this platform especially for higher fold coverage of ONT reads. We currently recommend leaving the variant analysis **OFF** for ONT data. You can change any of these parameters by expanding the module and selecting the desired changes.
    - For variant call, 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 Variants calls were filtered further by vcfutils.pl of SAMtools with following criteria:
      (1) minimum RMS mapping quality for SNPs [10];
      (2) minimum read depth [5];
      (3) maximum read depth [300];
      (4) minimum number of alternate bases [3];
      (5) minimum ratio of alternate bases [0.3];
      (6) SNP within INT bp around a gap to be filtered [3];
      (7) window size for filtering adjacent gaps [10];
      (8) min P-value for end distance bias [0.0001];
      (9) maximum fraction of reads supporting an indel [0.5];
    - The consensus workflow uses maximum 300x depth coverage reads for computational efficiency with enough statistically significant, PCR deduplication and check various parameters by default including the minimum 5x depth coverage (otherwise the consensus will be “N”), alternate base Threshold at 0.5 (support an alternative for the consensus to be changed), Indels Threshold at 0.5 (support an INDELs for the consensus to be changed), minimum mapping quality 60, minimum base quality 20 for

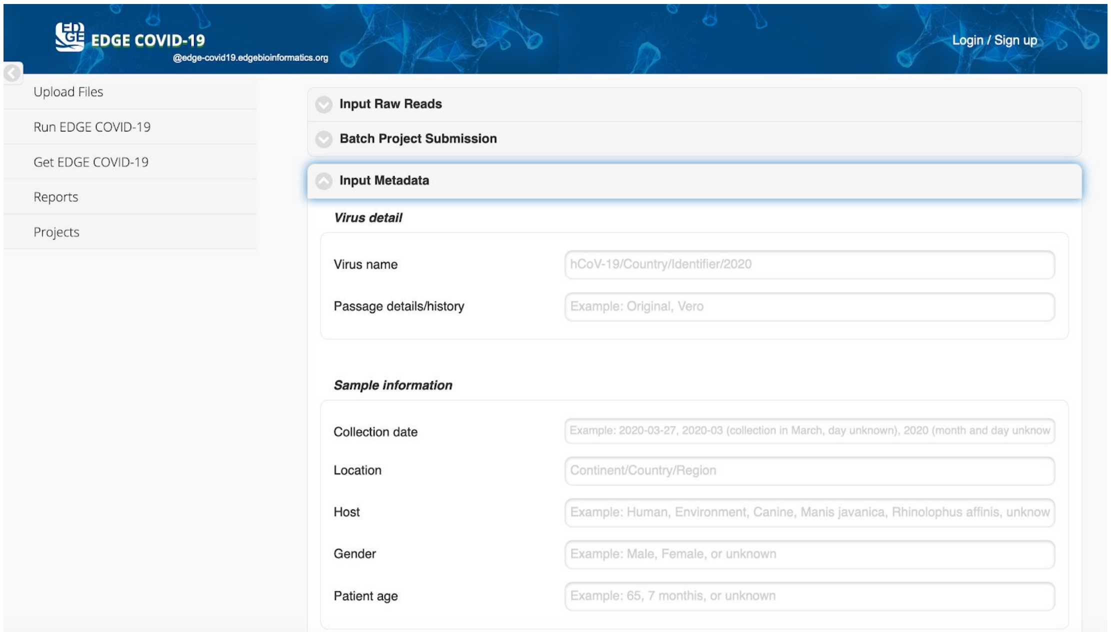
illumina data and minimum base quality 5 for ONT data. Click the **submit** button at the bottom of the page to start your job.

11. 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.

12. Once the pipeline is finished and you have a consensus genome, you can prepare your genome to submit to GISAID by inputting the metadata.



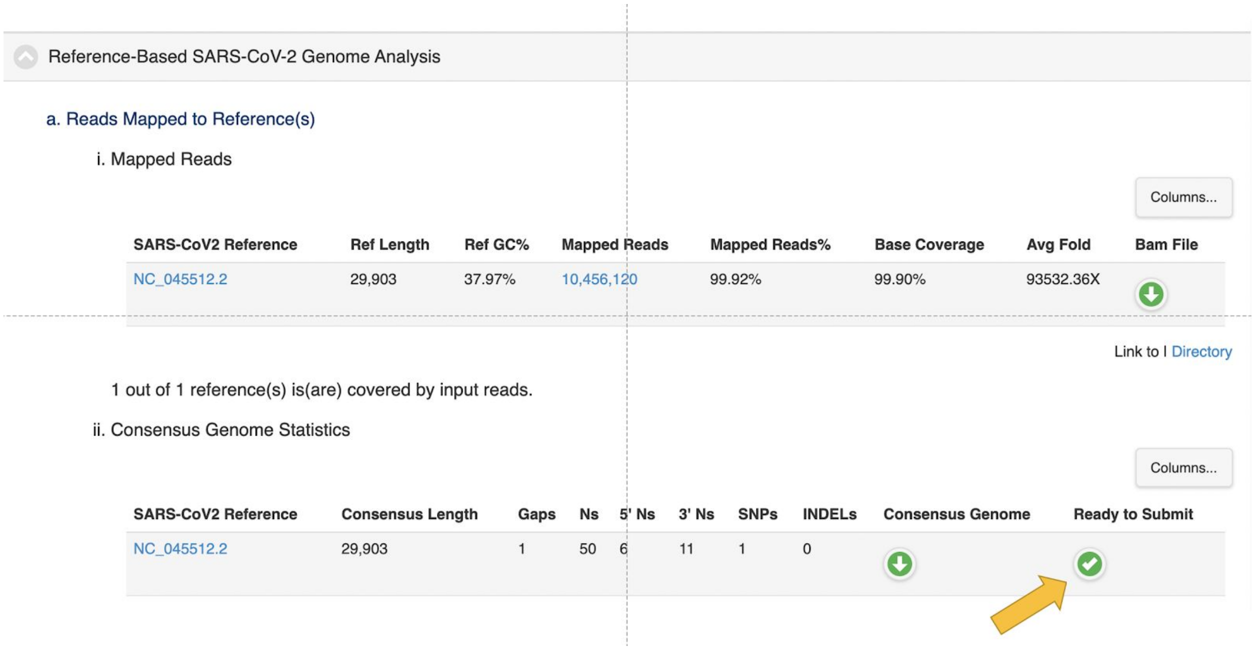
Virus detail

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 unknow
Location	Continent/Country/Region
Host	Example: Human, Environment, Canine, Manis javanica, Rhinolophus affinis, unknow
Gender	Example: Male, Female, or unknown
Patient age	Example: 65, 7 monthis, or unknown

After entering the metadata, you can even submit the genome directly through the platform. You can access this function 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

SARS-CoV2 Reference	Ref Length	Ref GC%	Mapped Reads	Mapped Reads%	Base Coverage	Avg Fold	Bam File
NC_045512.2	29,903	37.97%	10,456,120	99.92%	99.90%	93532.36X	⬇

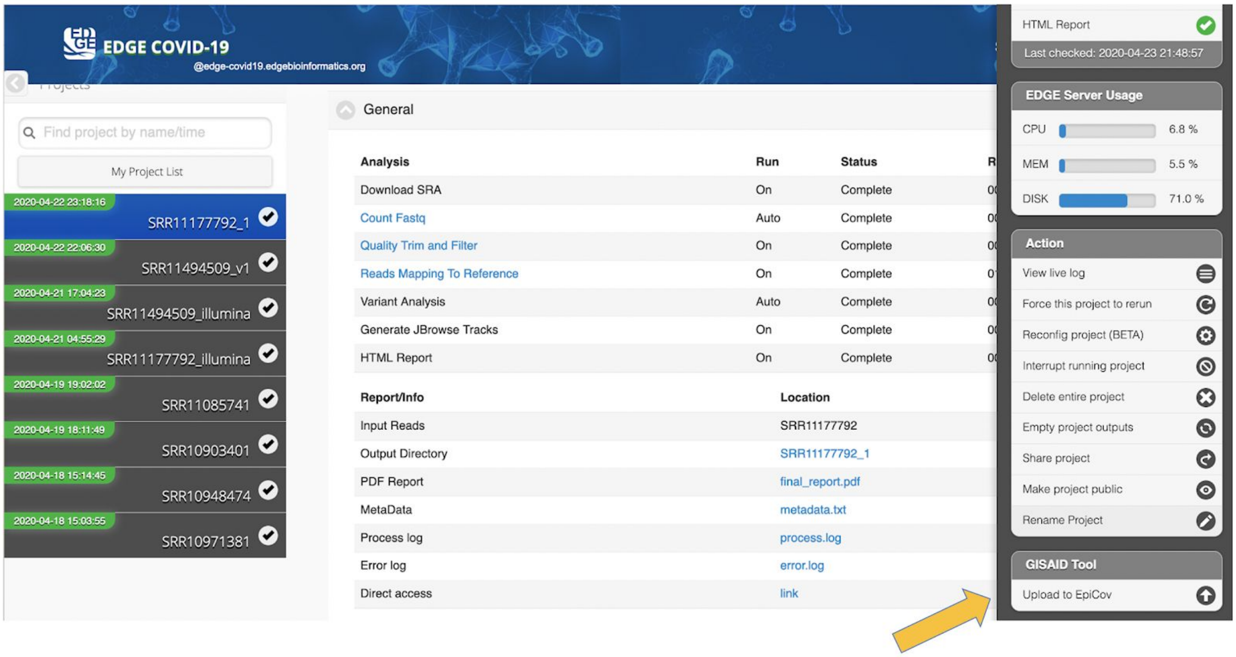
Link to I Directory

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

ii. Consensus Genome Statistics

SARS-CoV2 Reference	Consensus Length	Gaps	Ns	5' Ns	3' Ns	SNPs	INDELs	Consensus Genome	Ready to Submit
NC_045512.2	29,903	1	50	6	11	1	0	⬇	✓

A menu will appear on the right side of the screen. Click the bottom of the menu to submit.



2020-04-22 23:18:16	SRR11177792_1
2020-04-22 22:06:30	SRR11494509_v1
2020-04-21 17:04:23	SRR11494509_illumina
2020-04-21 04:55:29	SRR11177792_illumina
2020-04-19 19:02:02	SRR11085741
2020-04-19 18:11:49	SRR10903401
2020-04-18 15:14:45	SRR10948474
2020-04-18 15:03:55	SRR10971381

Analysis	Run	Status	R
Download SRA	On	Complete	00
Count Fastq	Auto	Complete	00
Quality Trim and Filter	On	Complete	00
Reads Mapping To Reference	On	Complete	01
Variant Analysis	Auto	Complete	00
Generate JBrowse Tracks	On	Complete	00
HTML Report	On	Complete	00

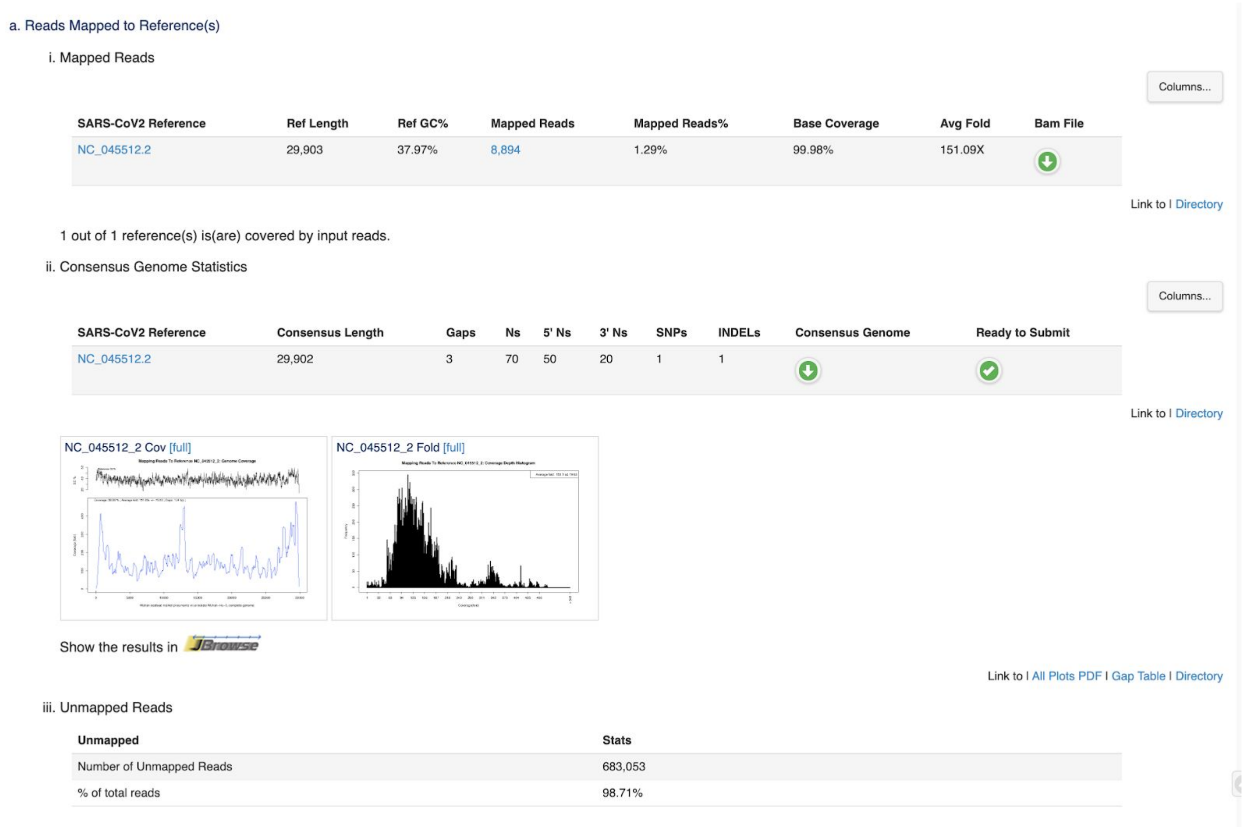
Report/Info	Location
Input Reads	SRR11177792
Output Directory	SRR11177792_1
PDF Report	final_report.pdf
MetaData	metadata.txt
Process log	process.log
Error log	error.log
Direct access	link

However, this feature is in beta format; if you run into any trouble you can contact us at edge-covid19@lanl.gov.

What results can I access via the EDGE GUI?

For a more detailed description of the 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 of all reads both prior to and after quality trimming and filtering. If assembly/annotation is selected, this module's output will include the assembled contigs as a Fasta file, 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 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 consensus genome by providing a green check mark if the resulting consensus genome is longer than 25kb, has depth coverage greater than 10X, and less than 5% of the genomes are 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, the prevalence within the reads covering that position, any changes in amino acid composition, etc.). (See below)



a. Reads Mapped to Reference(s)

i. Mapped Reads

Columns...

SARS-CoV2 Reference	Ref Length	Ref GC%	Mapped Reads	Mapped Reads%	Base Coverage	Avg Fold	Bam File
NC_045512.2	29,903	37.97%	8,894	1.29%	99.98%	151.09X	⬇

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	5' Ns	3' Ns	SNPs	INDELs	Consensus Genome	Ready to Submit
NC_045512.2	29,902	3	70	50	20	1	1	⬇	✓

Link to I Directory

NC_045512_2 Cov [full]

NC_045512_2 Fold [full]

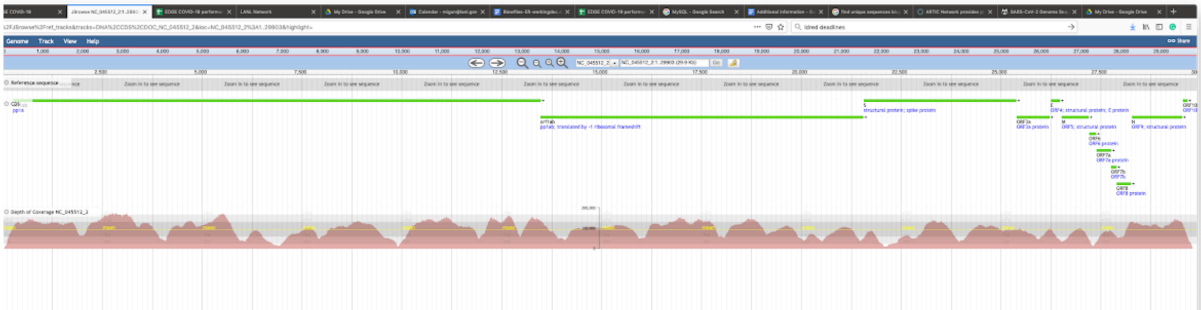
Show the results in JBrowse

Link to I All Plots PDF I Gap Table I Directory

iii. Unmapped Reads

Unmapped	Stats
Number of Unmapped Reads	683,053
% of total reads	98.71%

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).



FAQs

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

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

FAQs for local docker build:

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

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

```
$ 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
```

Wait a few minutes and go to http://localhost.

To stop the docker:

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

Note that the docker 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 from **Step 1**.

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

After pulling the latest docker, then starting 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?

We are interested in compiling runtime information from different SRA or other examples executed on laptops or any local/remote/cloud based servers:

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 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 environment, users need to set the /path/to/mysql directory into 0777 mode. Please try opening up the directory permission if you run into trouble.

```
 * Use data volume container (a.k.a: OS independent.)
  $ 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 docker0 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 user `docker run` it 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-covid19 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 (edgeui_tol_cpu-1)/max_num_jobs in integer CPUs (as you see on the GUI). These number can be changed by editing the sys.properties file in the edge-covid19 container:

```
    $ CPU_NUM=10  # the number to update
    $ #below is one line command
    $ docker exec edge-covid19 sed -ie
"s/edgeui_tol_cpu=[0-9]\+/edgeui_tol_cpu=${CPU_NUM}/"
/home/edge/edge/edge_ui/sys.properties
```

Refresh the Browser page to see the update on the GUI.

Contact:

You can view the discussions in the google group below and join the group to post questions or comments.
EDGE user's google group at https://groups.google.com/d/forum/edge-users

Email us: edge-covid19@lanl.gov

Citation

Po-E Li, Chien-Chi Lo, Joseph J. Anderson, Karen W. Davenport, Kimberly A. Bishop-Lilly, Yan Xu, Sanaa Ahmed, Shihai Feng, Vishwesh P. Mokashi, Patrick S.G. Chain; Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform, Nucleic Acids Research, Volume 45, Issue 1, 9 January 2017, Pages 67–80, https://doi.org/10.1093/nar/gkw1027