EDGE COVID-19 documentation

Los Alamos National Laboratory

v1.0.0

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 Nanopore 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 preliminary Single Nucleotide Polymorphism and Variant analyses, with some detail such as location and resulting coding differences if any.

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 only 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.

This light-weight version is available as a Docker container, able to run on any local hardware infrastructure or in the cloud. We have tested this Docker container on laptops and in the cloud, using Illumina (e.g. SRR11177792) and ONT (e.g. SRR11300652) datasets.

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

Requirements

1. Docker Engine version 19.03.2 or greater
2. Recommended minimum computational resource:
    - 8 GB memory
    - 4 CPUs
    - 20GB storage space for the image

How to install this image? A step by step guide:

Step 1: Install and run Docker

This step can be skipped if you have docker installed and opened in your system. If a docker instance is running in a MacOSX, dockers's icon ( ) will show up at the top bar of your screen.

If you do not have docker installed, See https://www.docker.com/products/docker-desktop to download and install a copy.

Step 2: Obtain the docker image

The image size is around 11.8GB. On a MacOSX, open Terminal and `cd` into the directory where you want to install the image. If you want to create a new folder, then first create that folder, and then pull the docker image:

```
$ mkdir EDGE-COVID19
$ cd EDGE-COVID19
$ docker pull bioedge/edge_ncov
```

This can take anywhere from 10-30 minutes depending on your internet speed.

Step 3: Setup necessary databases and folders

Download MySQL database for User Management. You can either directly download from Terminal using `wget` or `curl`, or just click the link and download it in the `EDGE-COVID19` folder.

```
$wget
https://edge-dl.lanl.gov/EDGE/docker/edge_ubuntu_init_mysql_EDGE_input_for_edge_docker.tgz
```

Unzip the download file

```
$tar -xvzf
edge_ubuntu_init_mysql_EDGE_input_for_edge_docker.tgz
```

Create Output and Report directories

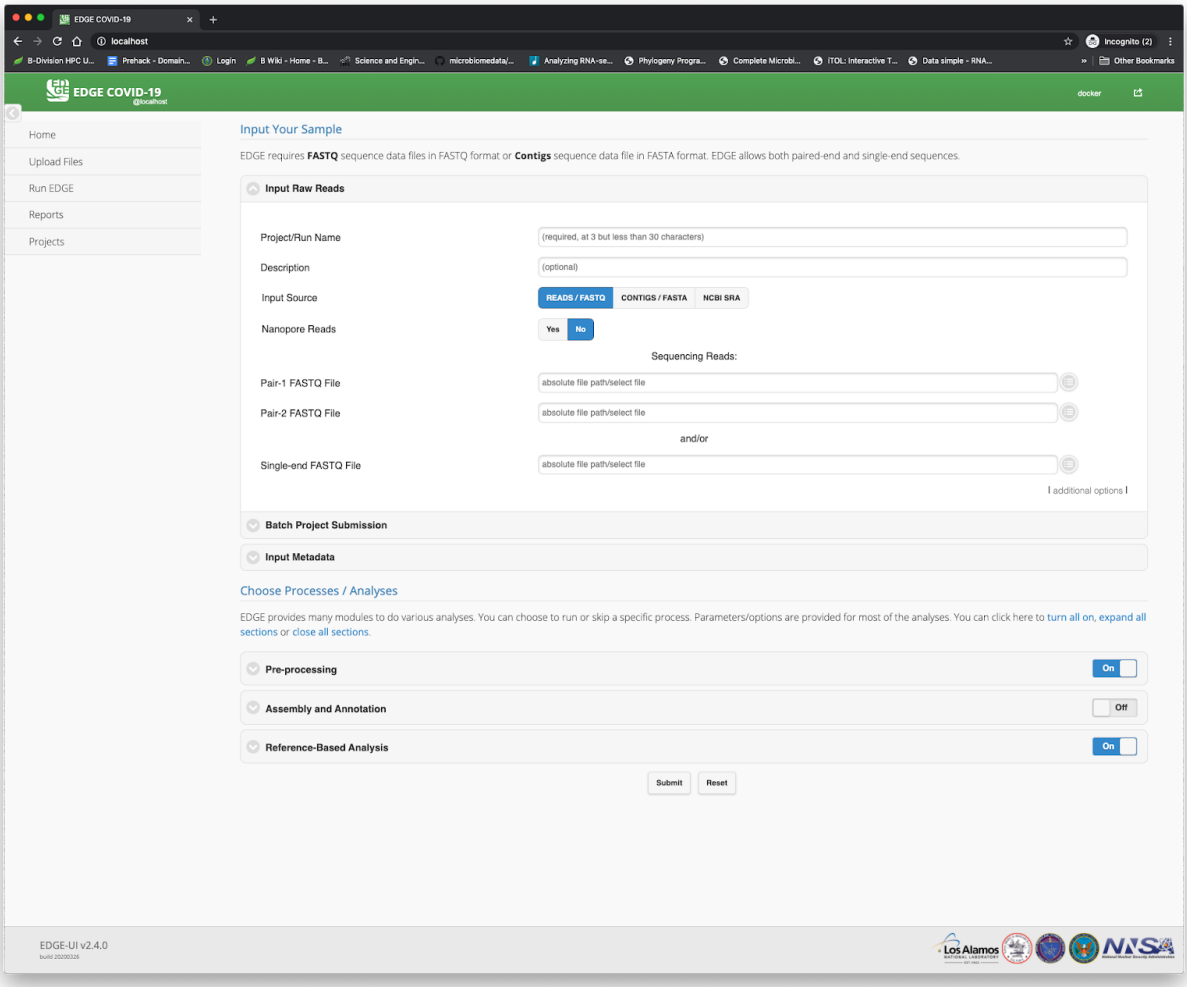
```
$ mkdir -p EDGE_output  EDGE_report EDGE_input
```

Step 4: Start *EDGE COVID-19* instance

Start the *EDGE COVID-19* by running the following command in your Terminal from the `EDGE-COVID19` folder.

```
$ docker run -d -v $PWD/mysql:/var/lib/mysql \
 -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_ncov bioedge/edge_ncov
```

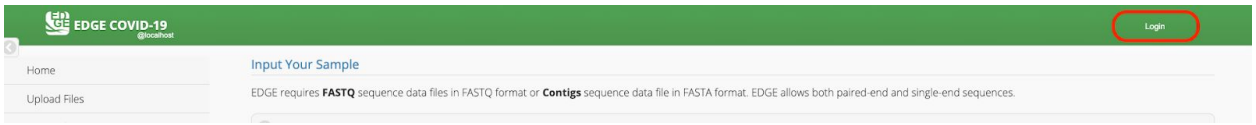
Wait a few minutes for the docker image to start the EDGE service; then open http://localhost/ in a browser (Firefox, Chrome, Safari) to start the *EDGE COVID-19*. The instance will keep running. You will see the following screen:



Step 5: Login information

If *EDGE COVID-19* is to be used by a single user then there is no need to create an account. You can log in directly using following credentials by clicking on the `Login` button on top right:

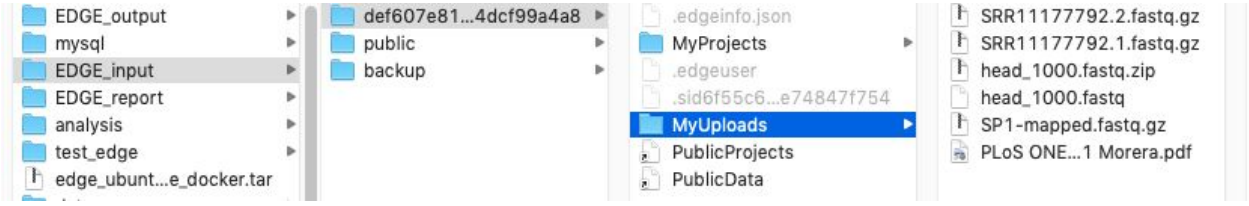
EDGE user: admin_docker@my.edge
EDGE password: admin_docker

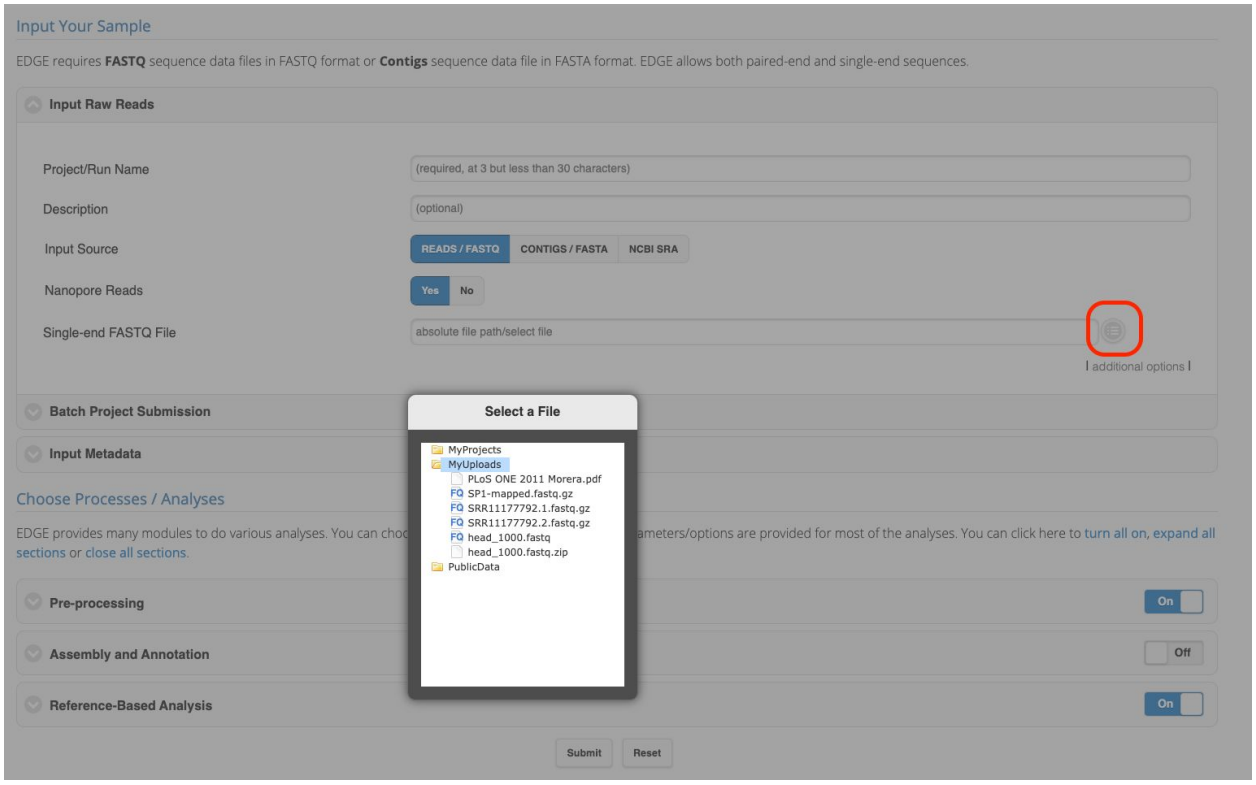


If there is going to be multiple users for this instance of *EDGE COVID-19,* one can create an account clicking the **Login** button and selecting **Create an account.**

Step 6: Upload your raw reads

The easiest way is to put your data in the upload folder, which can be found within the `EDGE_input` folder that you created in Step 3. 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.





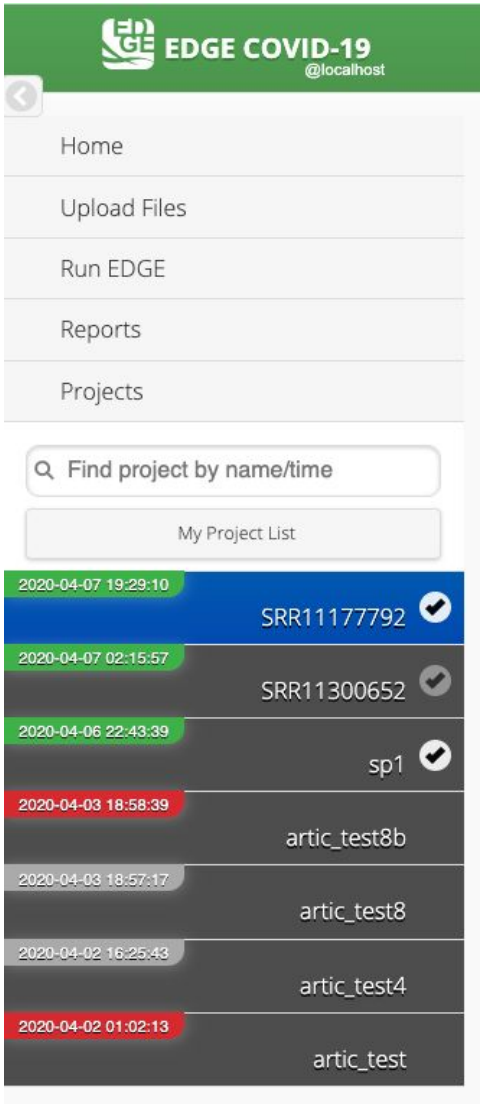
Step 7: Run your sample

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.

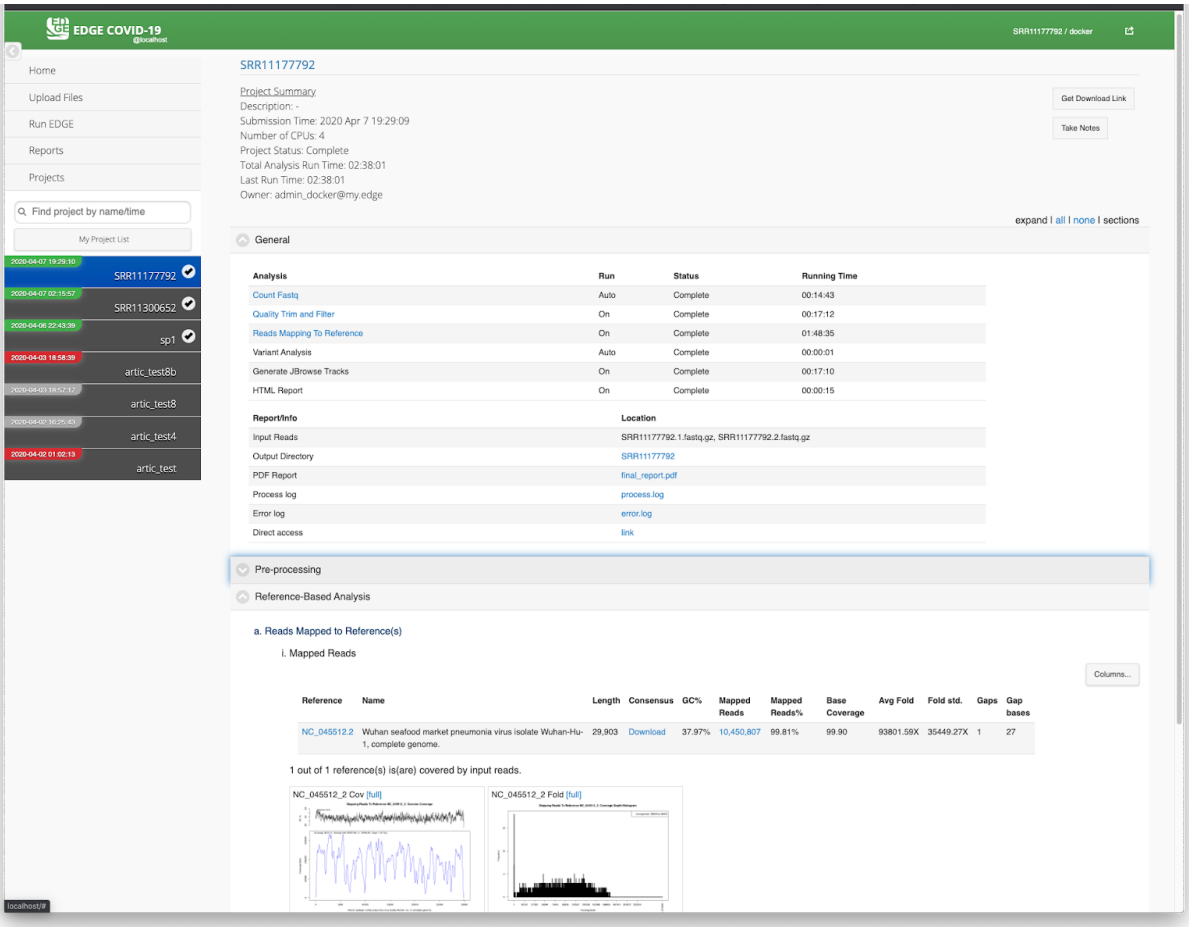
In the *EDGE COVID-19* web server,

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 next to 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 6.
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 trimming of reads that either fail a quality threshold. If you wish to change parameters, you can expand the module by clicking on it and modify as you wish.
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 analysis 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 analysis based on reads can take well over 24 hours on this platform especially for higher fold coverage of Nanopore 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.
11. Click the **submit** button at the bottom of the page to start your job.

12. 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.



A project with red highlight means the project failed, green means the project has finished, orange (not shown) means the project is running, and grey means the project has not yet started. You can access the details and outputs of each project by clicking on the project in the list. For example:



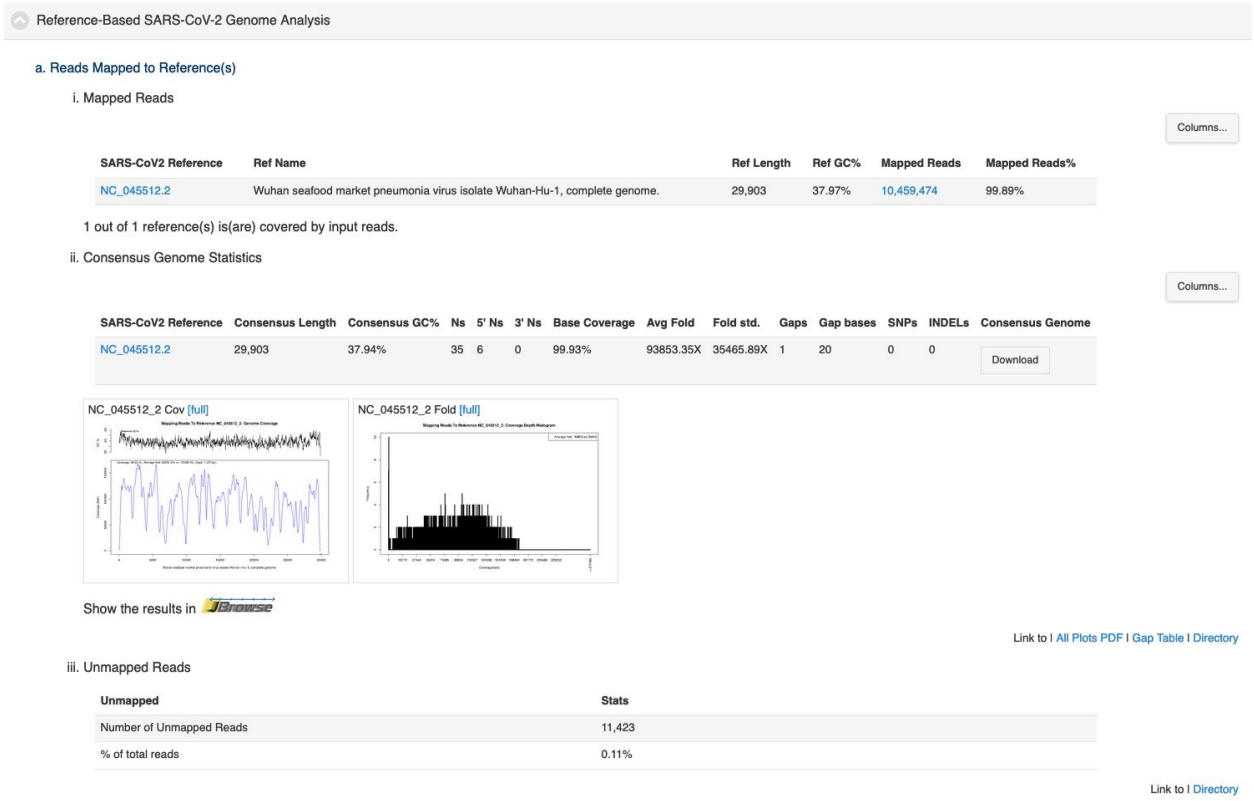
Once selected, the project results page displays a summary of the run and general statistics of what tools and modules were activated and their runtime and status, as well as links to log files that provide detailed information on the command lines and parameters used for each tool executed. The rest 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.

What results can I access via the EDGE GUI?

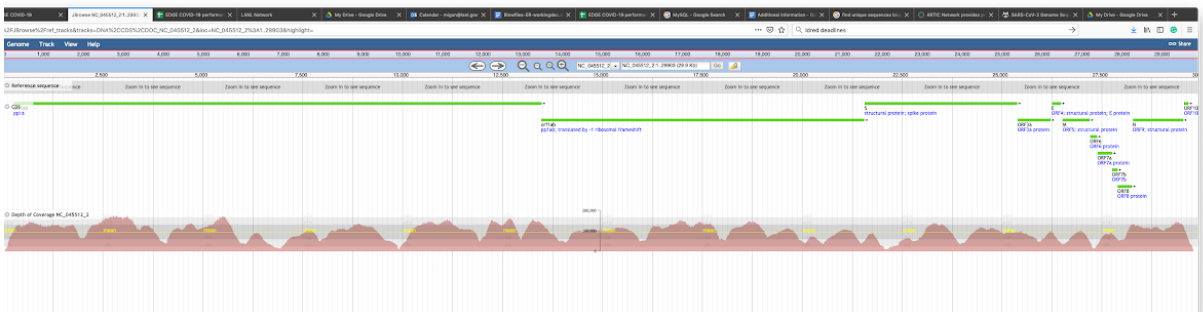
For a more detailed description of the output page, please refer to our full documentation here. Each selected module will be displayed as a subsection, and detailed results may be found in each section. The Pre-processing section, for example, will have details on various statistics of all reads both 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). These and more data can be accessed via some of the links at the bottom of the section, including 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)



For scientists wishing to examine the details underlying the statistics, a JBrowse link is provided 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).



How to start/stop *EDGE COVID-19*?

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_ncov
$ docker run -d -v $PWD/mysql:/var/lib/mysql \
    -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_ncov bioedge/edge_ncov
```

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

To stop the docker:

```
$ docker stop edge_ncov
```

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

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_ncov
```

After pulling the latest docker, then starting the image from terminal:

```
$ cd EDGE-COVID19
$ docker rm -v edge_ncov
$ docker run -d -v $PWD/mysql:/var/lib/mysql \
    -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_ncov bioedge/edge_ncov
```

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
10,493,168	430,923	Nanopore	Direct RNA sequencing (virion only), library preparation using SQK-RNA002,	4	0:12:11
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

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

Troubleshooting and Miscs

Error in pulling image

If you have issues pulling this image, you may increase the basesize when launching docker daemon or use a different [Storage Driver] (https://goo.gl/8YUeUA). See [issue] (https://github.com/moby/moby/issues/8971).

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.

Commands for checking status and error log

Check the MySQL status in container:

```
$ docker exec edge_ncov service mysql status
```

where "edge_ncov" 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_ncov service tomcat7 status
```

Check the Apache web server status and log:

```
$ docker exec edge_ncov service apache2 status
$ docker exec edge_ncov tail /var/log/apache2/error.log
$ docker exec edge_ncov tail /var/log/apache2/access.log
```

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