Jul 09, 2020

SARS-CoV2 EBI assembly submission protocol

Nabil-Fareed Alikhan[1], Emma Griffiths[2], Ruth Timme[3], Duncan MacCannell[4]

[1]Quadram Institute Bioscience; [2]University of British Columbia; [3]US Food and Drug Administration; [4]Centers for Disease Control and Prevention

1 Works for me dx.doi.org/10.17504/protocols.io.bhwqj7dw

Coronavirus Method Development Community PHA4GE

Nabil-Fareed Alikhan
Quadram Institute Bioscience

ABSTRACT

PURPOSE:
This protocol covers the steps for submitting a SARS-CoV-2 assembly to ENA

For new submitters, there's quite a bit of groundwork that needs to be established before a laboratory can start its first data submission. We recommend that one person in the laboratory take a few days to get everything set up in advance of when you expect to do your first data submission.

Two protocols cover the PHA4GE guidance for SARS-CoV-2 submission to ENA (Raw sequence data, metadata, and assemblies)

Complete in order (1 then 2):

1. SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject

- Step-by-step instructions for establishing a new Webin laboratory submission account and for creating and linking a new BioProject to an existing umbrella effort.
- Submit SARS-CoV-2 raw data to ENA (European Nucleotide Archive) and metadata.

2. SARS-CoV-2 EBI assembly submission protocol (included protocol)

Required: established BioProject and BioSamples

- Submit SARS-CoV-2 assemblies to ENA linking to existing BioProject, BioSamples, and raw data.

DOI

dx.doi.org/10.17504/protocols.io.bhwqj7dw

PROTOCOL CITATION

Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell 2020. SARS-CoV2 EBI assembly submission protocol . **protocols.io** dx.doi.org/10.17504/protocols.io.bhwqj7dw

KEYWORDS

metadata, INSDC, ERC000033, ENA, EBI, SARS-Cov2, COVID-19

CREATED

Jun 25, 2020

LAST MODIFIED

Jul 09, 2020

PROTOCOL INTEGER ID

38576

1 The Webin-CLI program is described as the only way to upload assembled sequences. This includes consensus sequences of SARS-COV2. Generally, you should follow the guidance here: https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html

The process, briefly, is as follows:

1. Download the Webin-cli tool. It is a Java program, so you will need the Java runtime enviroment installed as well.
2. Create manifest files, one for each assembly,
3. Submit the manifest file, and the associated assembly data via the Webin tool.

To begin SARSCoV2 consensus sequences will need to be submitted as a chromosome assembly, see https://ena-docs.readthedocs.io/en/latest/submit/assembly/genome.html#chromosome-assembly

Assemblies can only be submitted using Webin-CLI, using-context genome. During the process, you must define metadata in the manifest file(s). Please specify 'COVID-19 outbreak' as the 'ASSEMBLY_TYPE'.

2 The assembly submission system is set up to point to an existing sample record. You must have already created your Project/Study and registered your samples to proceed with the assembly submission.

For each record you will need 3 different files:

1. A manifest file that details the sample the assembly should be associated, and some other metadata.
2. The assembly seqeunce itself, this should be a FASTA file compressed with gz.
3. The chromsome file which details the order of the sequences in the FASTA file. This again, should be compressed with gz.

The manifest file could look something like this:

```
STUDY ERP123456
SAMPLE ERS123456
RUN_REF ERR123456
FASTA SARSCOV_Seq_Example.fasta.gz
NAME SARSCOV_Seq_Example
ASSEMBLY_TYPE COVID-19 outbreak
PROGRAM       ARTIC-ivar
PLATFORM Illumina
COVERAGE 1000
CHROMOSOME_LIST Illumina_NORW-EA35E.chrom.gz
```

Note here, that the STUDY, SAMPLE and RUN/EXPERIMENT must be specified, which means you should have already created these records and you should have the accession numbers to populate these fields.

See the Webin documentation for more information https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html

3 **Creating the supporting files**

The supporting files include the consensus sequence and the chromsome file. Both must be compressed (gz format) to be submitted by the Webin-Cli program.

Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SARS-CoV2 EBI assembly submission protocolÃÂ . https://dx.doi.org/10.17504/protocols.io.bhwqj7dw

The chromsome file which details the order of the sequences in the FASTA file. Since the your COVID19 consensus sequence should be single contigious sequence, the file is very simple, with the name of the seqeunce (the FASTA header), being the first (and only sequence) [1].

```
SARSCOV_Seq_Example  1       Chromosome
```

The FASTA seqeunce is a standard FASTA format. The header should match the name given in the chromosome file.

```
>SARSCOV_Seq_Example
ATAGTCACATAGCAATCTTTATCACATAGCAATCTTTATCACATAGCAATCTTTATCACATAGCAATCTTTA..
.
```

You need one chromsome file, for each FASTA file.

See the Webin documentation for more information https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html

4 Submitting a consensus sequence

With all the supporting files ready you can submit them with the webin tool, example:

```
java -jar webin-cli-3.0.0.jar -context genome --manifest manifest_file.txt -inputDir myData
 -outputDir mydata_reports  -submit  -userName Webin-12345 -passwordFile mypassword
```


Citation: Nabil-Fareed Alikhan, Emma Griffiths, Ruth Timme, Duncan MacCannell (07/09/2020). SARS-CoV2 EBI assembly submission protocolÃÂ . https://dx.doi.org/10.17504/protocols.io.bhwqj7dw