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

Forked from SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject

Nabil-Fareed Alikhan1, Ruth E Timme2, Emma Griffiths3, Duncan MacCannell4

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

Works for me dx.doi.org/10.17504/protocols.io.bhwj7a6

Coronavirus Method Development Community PHA4GE

Nabil-Fareed Alikhan
Quadram Institute Bioscience

ABSTRACT

This protocol provides the steps needed to establish a new EBI submission environment for your laboratory, including BioProject(s). Once established, this protocol covers raw read submission to EBI and sample metadata to BioSample.

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 EBI

Complete in order (1 then 2):

1. SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject (included protocol)
   - Step-by-step instructions for establishing a new EBI (Webin) submission account and for creating and linking a new BioProject to an existing umbrella effort.
   - SARS-CoV-2 raw data submission to ENA (European Nucleotide Archive) and metadata to BioSample.

2. SARS-CoV-2 EBI assembly submission protocol
   - Required established BioProject and BioSamples
   - Submit SARS-CoV-2 assemblies to EBI, linking to existing BioProject, BioSamples, and raw data.

DOI

dx.doi.org/10.17504/protocols.io.bhwj7a6

PROTOCOL CITATION

Nabil-Fareed Alikhan, Ruth E Timme, Emma Griffiths, Duncan MacCannell 2020. SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject. protocols.io

https://dx.doi.org/10.17504/protocols.io.bhwj7a6

FORK FROM

Forked from SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject, Ruth Timme

KEYWORDS

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

LICENSE

This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Preamble and other documentation

1 How data is structured in ENA

Data in ENA are structured in a hierarchy. Data is grouped under a "STUDY," which contains multiple "SAMPLES," which in turn have multiple "RUNS/EXPERIMENTS," and from these are derived "ANALYSES."

- **STUDY**: records overarching information about the organisation, list of authors, and an abstract about the study. This is also referred to as a PROJECT.
- **SAMPLES**: records the contextual information about how the sample was collected. E.g. Host, geographic location, collection dates.
- **EXPERIMENT/RUNS**: records the details about how the sequence was produced, e.g. sequencing platform and library preparation and points to the read data.
- **ANALYSES**: records and data derived from the sequence data, such as de novo genome assemblies.

Note that at each level, there can be multiple records pointing to the parent e.g. multiple sequence runs pointing to a single sample, or multiple genome assemblies pointing to a set of sequence data.

Relationship of the different data records in ENA

As data depends on another data records, you will need to submit them in a particular order. i.e. STUDY > SAMPLE > EXP/RUN > ANALYSIS.

There are many ways to submit data to ENA:
- **INTERACTIVE WEB (This protocol)**
This is done through a web browser. This is the easiest method to get started, but will become tedious with large submission (> 50 records).

**WEBIN CLI CLIENT**
This is a Java program you can download that will accept prepared plain text files (MANIFESTS). These manifests specify the same information you would enter in the interactive web client, but it is easier to generate programatically. The program also submits data (reads, assemblies) for you. This is also the ONLY way to submit assembled sequences. For any sequencing centre producing data *en masse* this would be the best option. See [https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html](https://ena-docs.readthedocs.io/en/latest/submit/general-guide/webin-cli.html) for more information.

**MANUAL XML SUBMISSION**
This requires creating XML documents with the same information you would submit via the interactive web option. These files can then be submitted through [https://www.ebi.ac.uk/ena/submit/webin/](https://www.ebi.ac.uk/ena/submit/webin/)
If you generate these XML files you can test them out here: [https://wwwdev.ebi.ac.uk/ena/submit/webin/](https://wwwdev.ebi.ac.uk/ena/submit/webin/)

**PROGRAMATIC XML SUBMISSION**
This requires the XML files again but the files can be submitted through an API.

In this protocol, we will use the INTERACTIVE WEB submission system, which is through a web browser. This should be the the easiest way to get started. As you get used to the system, you can try the more advanced methods for bulk submissions.

There is a detailed protocol for filling out the submission contextual data (metadata) here. [https://www.protocols.io/private/AC87DEE56E44C12B5F731895CC821F63](https://www.protocols.io/private/AC87DEE56E44C12B5F731895CC821F63)

The protocol will take you through how to create a submission account, and how to set up Projects, Samples and Runs. There is a separate protocol dealing with assemblies/consensus sequences.

2 **Test your submission first!**
We highly recommend doing a run through of your submission through the TEST service first.
There are two interactive Webin submission services. One for test submissions and another for production submissions:

- Test service URL: [https://wwwdev.ebi.ac.uk/ena/submit/sra](https://wwwdev.ebi.ac.uk/ena/submit/sra)
- Production service URL: [https://www.ebi.ac.uk/ena/submit/sra](https://www.ebi.ac.uk/ena/submit/sra)

3 **Important resources**

- ENA Submission site: [https://www.ebi.ac.uk/ena/submit/sra/#studies](https://www.ebi.ac.uk/ena/submit/sra/#studies)
- COVID-19 data portal: [https://www.covid19dataportal.org/](https://www.covid19dataportal.org/)

If you have any queries or require assistance with your SARS-CoV2 submission please contact: virus-dataflow@ebi.ac.uk

4 **Create an Webin user account at ENA**
There is a form to fill in at [https://wwwdev.ebi.ac.uk/ena/submit/sra/#registration](https://wwwdev.ebi.ac.uk/ena/submit/sra/#registration)

This will account should be shared with all members of your submission team. This does mean sharing the
You can now log in into the submission service.

**Add additional contacts for your lab.**

Under *Home > My account details* you can add extra contacts for data submissions. These contacts will be notified if there are any major changes to data submissions and they will be listed as contacts on public data.

**Bookmark “Studies” at ENA:** [https://www.ebi.ac.uk/ena/submit/sra/#studies](https://www.ebi.ac.uk/ena/submit/sra/#studies). This is the page where you view and track all of your study submissions. You should bookmark this page.

Studies page at ENA is a good place to start your submissions.
Starting a new submission

You can start a new submission under "New Submission". From here you can do multiple different types of submissions. Remember that there is a particular order for data submissions i.e. STUDY > SAMPLE > EXP/RUN > ANALYSIS. So we will start with "Register STUDY"

Webin - New submission page

If you get stuck at any stage there is a "Restart Submission" link at the bottom

Identify or establish new BioProjects (Umbrella and/or Data BioProjects)

Umbrella projects/studies. If you are already part of a surveillance network, (e.g. SPHERES, COG-UK, CanCOGeN, etc) you should link your study to one of their established umbrella projects. For reference, here are some of the umbrella projects established for SARS-CoV-2 surveillance:

SPHERES (US): PRJNA615625
CanCOGeN (Canada), PRJNA623807

You will need to talk to your consortium about having your new study linked with their umbrella project.

Data BioProjects. Does your consortium have an established data BioProject for this effort? You should ask them to act as a broker and submit the data on your behalf.

Countries with single data projects (not exhaustive):
COG-UK (United Kingdom): PRJEB37886
Turkey: PRJNA636004
Switzerland: PRJEB38472
South Africa: PRJNA624358

Once you have confirmed that there are no existing projects that your data should be submitted under, you can safely create a new study for yourself.

Describe the study

Fill in the form with the details of the study. Fields marked with (*) are mandatory. You can edit this information later.

Citation: Nabil-Fareed Alikhan, Ruth E Timme, Emma Griffiths, Duncan MacCannell (07/09/2020). SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject. https://dx.doi.org/10.17504/protocols.io.bhwdj7as

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Registering a study form

Once you have submitted the form, you should see a receipt like this:

Dear Colleague,

Thank you for your recent submission to the European Nucleotide Archive with your account 'Webin-55758'.

We are pleased to advise you that the metadata has passed validation and the following accession numbers have been assigned:

- Your study accession number is: PRJEB239667
- Your study unique name is: ena-STUDY-text-25-06-2020-12:50:02:313-6
- Your study will be made public on '29 Aug 2020'.

Before the study has been made public, you can edit the release date if you wish to extend this date or release the study at an earlier date here: http://www.ebi.ac.uk/ena/about/data-release-mechanics/

Release is mandatory once the data is published in a journal.

Note that citations are added only at study level. Therefore, once your study is published, please add the PubMed ID (if available) to this email. If you need further assistance, contact EDA helpdesk https://www.ebi.ac.uk/ena/browse/accession-help

Accession Summary:

- Study accession number: PRJEB239667
- Study unique name: ena-STUDY-text-25-06-2020-12:50:02:313-6
- Kind regards,
- European Nucleotide Archive (ENA)

Your accession summary is also available as a downloadable .tsv spreadsheet. Download Accession Summary

Study registered success receipt

If you complete this on the production submission site, you can cite the study accession number in your publication.

You can see the new study when you go back the "Studies". You can also edit the details and release date here.

Uploading sequencing data

10 Uploading sequencing data

It is recommend that read data is submitted via one of the Webin clients. The easiest method to get started with Webin File Uploader but there may be a better approach depending on your environment. There is extensive documentation here https://ena-docs.readthedocs.io/en/latest/submit/fileprep/upload.html

Registering samples and submitting sequencing metadata (Part 1)

11 Samples and metadata

Citation: Nabil-Fareed Alikhan, Ruth E Timme, Emma Griffiths, Duncan MacCannell (07/09/2020). SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject. https://dx.doi.org/10.17504/protocols.io.bhwdj7as

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
This is by far the most complicated step in the submissions process. This is where you must format your rich metadata according to the EBI data submission requirements. For SARS-CoV2 submission, they ask that the metadata comply with Checklist ERC000033: https://www.ebi.ac.uk/ena/browser/view/ERC000033.

There is a detailed protocol for filling out the sample sheet here.
https://www.protocols.io/private/AC87DEE55E4C4DF731895CC821F65

There is extended guidance available at the PHA4GE SARS-CoV-2 metadata specification: https://github.com/pha4ge/SARS-CoV-2-Contextual-Data-Specification

You should start formatting your metadata according these specifications before you begin.

11.1 Preparing the sample spreadsheet
Skip this step, if your already have a spreadsheet completed.

If your metadata is in order, starting from the new submissions page, select “Submit sequence reads and experiments” and click Next.

Specify a checklist for the metadata. In this case, it is the ENA VIRUS PATHOGEN REPORTING STANDARD CHECKLIST, under Select Checklist > Pathogens Checklist.

Citation: Nabil-Fareed Alikhan, Ruth E Timme, Emma Griffiths, Duncan MacCannell (07/09/2020). SARS-CoV-2 EBI submission protocol: ENA, BioSample, and BioProject. https://dx.doi.org/10.17504/protocols.io.bhwdj7as

This is an open access protocol distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Clicking Next at this point will show you the details of the various fields. You can check or uncheck different fields to add or remove them. Once this is complete, click Download template spreadsheet.

The download will be a tab separated table which you can open in your favourite spreadsheet program (Excel).

Check the PH4GE guidance for how to fill out this table with your metadata. There is also a separate protocol with informations about the various fields, see SOP for populating EBI submission templates.

- You may need to change "centuries" (host age units) to "years".
- The scientific name for SARS-CoV2 is "Severe acute respiratory syndrome coronavirus 2"
- The correct taxon id for SARS-CoV2 is "2697049"

Click Previous twice to go back to the first part of the submission process and click Submit complete.
spreadsheet to submit your completed spreadsheet (Continued on the next step). If there are any problems just restart the submission (click restart submission at the bottom).

11.2 Uploading the sample spreadsheet

From this page, click Submit completed spreadsheet, which will allow you to upload your prepared spreadsheet. If successful, you will see this dashboard with all the sample names on the left and the details of the selected sample on the right.

Any record with errors will have the red exclamation warning. The columns with invalid values will be shown in the right hand pane. Mouse over the blue (i) to see what values the field expects.

Once all the samples are green in the left hand pane, and everything looks OK, click submit. Any errors will appear below the form. If there are no errors, then there will be a final confirmation pop-up to...
Registering samples and submitting sequencing metadata (Part 2)

12 Submitting run & experiment metadata

This is a continuation of the previous section. By this point the raw sequencing data should be have been uploaded and the sample information should have been completed. Select a file format according to the directions.

You can only submit one type of file at once. If you have files of different types, these would need to be submitted through separate submissions.

Again, like the samples, download the template spreadsheet and fill in the information about your samples using your favourite spreadsheet program. You can then upload the sheet back, which will populate the table at the bottom. Review, correct any errors and then click Submit.

Check the PH4GE guidance for how to fill out this table with your metadata. There is also a separate protocol with information about the various fields, see SOP for populating EBI submission templates.

Errors on submission will be shown at the bottom.

If the submission is successful you will see a submission receipt. There is a downloadable table at the bottom that has all the accession codes, which can be included in your publication.
Congratulations! If you've made it this far you've successfully uploaded SARS-CoV2 data to ENA.

There are a few things to keep in mind:
- You can change the release date for your data under the studies page. 
  https://www.ebi.ac.uk/ena/submit/sra/#studies
- You can review the submitted data under the various tables.
- If you want to download the accession or check if you have missed any samples, it is actually easier to view them on the Webin submission portal here https://www.ebi.ac.uk/ena/submit/webin/