

Jul 09, 2020

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

 Forked from [SARS-CoV-2 NCBI submission protocol: SRA, BioSample, and BioProject](#)

 In 2 collections

DOI

<https://dx.doi.org/10.17504/protocols.io.bhwdj7a6>

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

¹Quadram Institute Bioscience; ²US Food and Drug Administration; ³University of British Columbia;

⁴Centers for Disease Control and Prevention

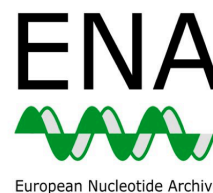
Coronavirus Method De...

PHA4GE



Nabil-Fareed Alikhan

University of Oxford



Create & collaborate more with a free account

Edit and publish protocols, collaborate in communities, share insights through comments, and track progress with run records.

Create free account

OPEN  ACCESS



DOI: <https://dx.doi.org/10.17504/protocols.io.bhwdj7a6>

Protocol Citation: Nabil-Fareed Alikhan, Ruth 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.bhwdj7a6>

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

Protocol status: Working

We use this protocol and it's working

Created: June 25, 2020

Last Modified: November 10, 2021

Protocol Integer ID: 38565

Keywords: metadata, INSDC, ERC000033, ENA, EBI, SARS-Cov2, COVID-19, biosamples submit sar, new ebi submission environment, raw read submission to ebi, metadata to biosample, pha4ge guidance for sar, bioproject this protocol, sample metadata to biosample, submit sar, existing bioproject, established bioproject, new bioproject, new ebi, raw data submission, pha4ge guidance, bioproject, ebi, first data submission, biosample, included protocol, sample metadata, metadata, raw read submission, european nucleotide archive,

Disclaimer

Please note that this protocol is public domain, which supersedes the CC-BY license default used by protocols.io.



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.

Preamble and other documentation

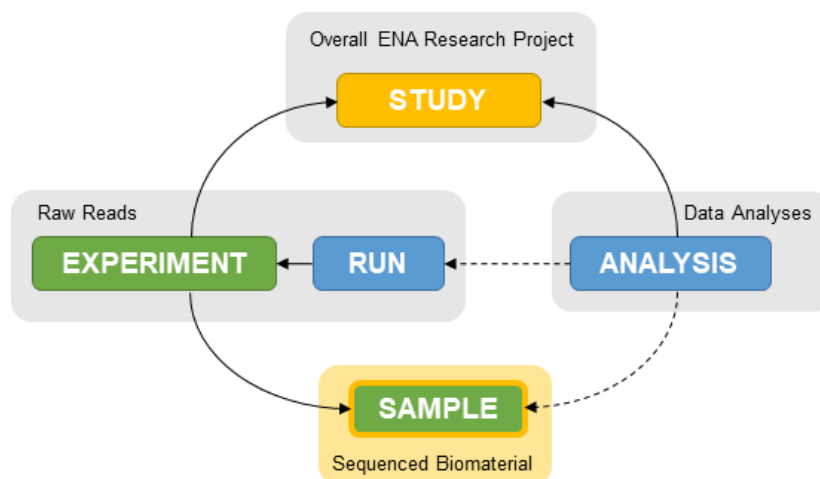
1 How data is structured in ENA

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

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

Note

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> 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/>

If you generate these XML files you can test them out here:

<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/AC87DEE56E44C12B5F731895CC821F65>

The protocol will take you through how to create a submission account, and how to set up Projects, Samples and Runs. There is a seperate 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>
- Production service URL: <https://www.ebi.ac.uk/ena/submit/sra>

3 **Important resources**

- ENA Documentation: https://ena-browser-docs.readthedocs.io/en/latest/help_and_guides/sars-cov-2-submissions.html
- ENA Submission site: <https://www.ebi.ac.uk/ena/submit/sra/#studies>
- COVID-19 data portal <https://www.covid19dataportal.org/>

Note

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

Creating a submission account

5m

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>

5m

Note

This will account should be shared with all members of your submission team. This does mean sharing the username and password.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support.

Registration form for a new submission (Webin) account.

You can now log in into the submission service .

5 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.

5m

Welcome to the Webin submission service.

You can use this service to submit [sequence reads](#), [targeted assembled and annotated sequences](#) and to register [studies \(projects\)](#) and [samples](#). To register umbrella studies, used to group together sequencing studies, please contact [ENA helpdesk](#).

[Help about submitting read data](#)

Form to update details about your organisation

6 Bookmark "Studies" at ENA: <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.

1m



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

7 Starting a new submission

2m

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

Note

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

Creating a new project

1w

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

1w

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.

9 Describe the study

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

Registering a study form

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



Congratulations! You successfully completed a TEST submission, please go to [this page](#) to perform your

Dear Colleague,

Thank you for your recent submission to the European Nucleotide Archive with your account 'Webin-55756'.

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: PRJEB39067

Your study unique name is: ena-STUDY-test-25-06-2020-12:58:02:313-6

Your study will be held private until '25-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. To do this see here: <http://www.ebi.ac.uk/ena/about/data-release-mechanism>.

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

Note that citations are added only on Study level. Therefore, once your study is published, please add the PubMed ID (if available).

Please DO NOT reply to this email. If you need further assistance, contact ENA helpdesk <https://www.ebi.ac.uk/ena/browse>

Accession Summary

Type Accession Unique Name

Study PRJEB39067 ena-STUDY-test-25-06-2020-12:58:02:313-6

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.

Search by:				Accession / Unique name: ERP122535		
Show: ☒ accession ☐ unique name						
Primary Accession	Secondary Accession	Title	Submission Date	Status	Release Date	Actions
PRJEB39067	ERP122535	test	25-Jun-2020	Confidential	25-Aug-2020	Edit
						Edit
						Edit

Uploading sequencing data

1w

10 Uploading sequencing data

It is recommended 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)

1w

11

Samples and metadata

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 SARSCov2 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/AC87DEE56E44C12B5F731895CC821F65>

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.

Start

>>

Study

>>

Sample

>>

Run

You can use this service to submit [sequence reads](#), [targeted assembled and annotated sequences](#) and to register [studies \(projects\)](#) and [samples](#). To register umbrella studies, use please contact [ENA helpdesk](#).

Please select the type of submission you would like to make:

☒ Submit sequence reads and experiments

We recommend that Fastq, BAM, and CRAM read files are submitted using [Webin-CLI](#).

When using this interface instead of [Webin-CLI](#), raw sequences must be [uploaded](#) in one of the supported [data formats](#) before they can be submitted. All data submitted in a si the same study. Data for different studies must be submitted in separate submissions. The study and the sequenced samples can be either pre-registered or registered durin that each individual study and sample should be registered only once. In addition, you will be asked to provide information about the sequencing libraries and instruments. Ple (ERP") when citing data submitted to ENA.

Read [here](#) for more information on how submit your raw reads to ENA.

☐ Register study (project)

☐ Taxonomy Check/Request

☐ Register samples

☐ Submit other assembled and annotated sequences [formerly EMBL-Bank]

The first step of your submission is to [upload data files](#). Data files can be uploaded using FTP or Aspera, or using the [Webin File Uploader](#). If you have already uploaded your data files into your Webin upload area please proceed directly to the next step. Please note that unsubmitted files that are older than 2 months will be deleted as explained in our [Fair Use Policy](#).

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

Home

New Submission

Studies

Samples

Runs

Analyses

Start

>>

Sample

>>

Finish

Start building your submission

We use checklists to help provide required information in a standard format.

You will be guided through the following steps:

- Selecting a checklist
- Selecting optional fields in addition to mandatory ones
- Entering your data directly into this application

Alternatively, after selecting the checklist and fields you will be able to download a template spreadsheet. You can then enter your data in the spreadsheet and upload it.

Select Checklist >

Upload a submission completed using a template spreadsheet

If you have downloaded and filled a template spreadsheet please upload it using the **Submit Completed Spreadsheet** button.

Please note that only spreadsheets in tab-delimited text format are supported (with either .tsv or .txt extensions). If you edited the spreadsheet in Microsoft Excel (or equivalent) please save the spreadsheet as Text (Tab delimited). To do this please see [these instructions](#).

Submit Completed Spreadsheet

Start

>>

Sample

>>

Finish

Please select the most appropriate checklist from the list below then click the **Next >>** button.

This group currently includes Genomic Standards Consortium (GSC) MixS sample checklists

Marine Checklists

This group currently includes Micro B3 and Tara Oceans sample checklists

Pathogens Checklists

This group currently includes several prokaryote and virus pathogen sample checklists

☐ ENA Global Microbial Identifier reporting standard checklist GMI_MDM-1.1
Minimum Data for Matching (MDM). A checklist for reporting metadata of pathogen samples for the Global Microbial Identifier (GMI) reporting system. More about GMI can be found here <http://www.g-m-i.org/>

☐ COMPARE-ECDC-EFSA pilot food-associated reporting standard
A checklist for reporting metadata of food-borne pathogen samples for the COMPARE-ECDC-EFSA reporting system.

☐ COMPARE-ECDC-EFSA pilot human-associated reporting standard
A checklist for reporting metadata of human-associated pathogen samples for the COMPARE-ECDC-EFSA reporting system.

☐ ENA parasite sample checklist
Minimum information about parasite samples. A checklist for reporting metadata of parasite samples associated with molecular data. This standard was developed by the COMPARE platform and can be used for submission of sample metadata derived from protozoan parasites (e.g. Cryptosporidium) and also multicellular eukaryotic parasites (e.g. Platyhelminthes and Nematoda).

☐ ENA prokaryotic pathogen minimal sample checklist
Minimum information required for a prokaryotic pathogen sample

☒ ENA virus pathogen reporting standard checklist
Minimum information about a virus pathogen. A checklist for reporting metadata of virus pathogen samples associated with genomic data. This minimum metadata standard was developed by the COMPARE platform for submission of virus surveillance and outbreak data (such as Ebola) as well as virus isolate information.

☐ ENA Influenza virus reporting standard checklist
Minimum information about an Influenza virus sample. A checklist for reporting metadata of Influenza virus samples associated with genomic data. This minimum metadata standard supports submission of avian, human and mammalian surveillance data as well as serology and virus isolate information (where available). The ENA Influenza sample checklist is based on standards in use at the Influenza Research Database.

Other Checklists

This group currently includes the ENA default sample checklist and a few project specific checklists

<< Previous

Next >>

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*

protocols.io | <https://dx.doi.org/10.17504/protocols.io.bhwdj7a6>

July 9, 2020

12/17

☐ Collection event information

☒ **collection date** - *recommended*
 The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated (i.e. all of these are valid ISO8601 compliant times: 2005-01-23T19:23:10+00:00; 2005-01-23T19:23:10; 2005-01-23; 2005-01; 2005).

☒ **geographic location (country and/or sea)** - *mandatory*
 The geographical origin of the sample as defined by the country or sea. Country or sea names should be chosen from the INSDC country list (<http://insdc.org/country.html>).

☐ **geographic location (latitude)** - *recommended*
 The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system

☐ **geographic location (longitude)** - *recommended*
 The geographical origin of the sample as defined by latitude and longitude. The values should be reported in decimal degrees and in WGS84 system

☒ **geographic location (region and locality)** - *recommended*

22 of 35 fields selected

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

[illegible]

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**

Note

- You may need to change "centuries" (host age units) to "years".
- The scientific name for SARSCov2 is "Severe acute respiratory syndrome coronavirus 2"
- The correct taxon id for SARSCov2 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

Start ✓ >> **Sample** >> Finish

Start building your submission

i We use checklists to help provide required information in a standard format. You will be guided through the following steps:

- Selecting a checklist
- Selecting optional fields in addition to mandatory ones
- Entering your data directly into this application

Alternatively, after selecting the checklist and fields you will be able to download a template spreadsheet. You can then enter your data in the spreadsheet and upload it.

[Select Checklist](#)

Upload a submission completed using a template spreadsheet

i If you have downloaded and filled a template spreadsheet please upload it using the **Submit Completed Spreadsheet** button.

Please note that only spreadsheets in tab-delimited text format are supported (with either .tsv or .txt extensions). If you edited the spreadsheet in Microsoft Excel (or equivalent) please save the spreadsheet as Text (Tab delimited). To do this please see [these instructions](#).

[Submit Completed Spreadsheet](#)

<< Previous

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.

Start ✓ >> **Sample** >> Finish

i Please add samples to the submission. Multiple samples can be created by increasing the number by the add button

[Add](#) 1 samples

☒	NORW-EB469	i	x
☐	NORW-EBAEF	i	x
☐	NORW-EBAFE	i	x
☐	NORW-EBB28	i	x
☐	NORW-EBB45	i	x
☐	NORW-EBB70	i	x
☐	NORW-EBB8F	i	x
☐	NORW-EBD22	i	x
☐	NORW-EBD6E	i	x
☐	NORW-EBD8C	i	x

1-10 of 296

i Please submit by clicking the **Submit** button. Alternatively, download your data as a spreadsheet using the **Download Spreadsheet** button. Once you have filled the spreadsheet please restart the submission process and upload the spreadsheet using the **Upload Completed Spreadsheet** button.

[Download Spreadsheet](#)

[Restart Submission](#)

[Previous Sample](#) [Next Sample](#)

Basic Details

* Unique Name: NORW-EB469 **i** **x**

* Title: NORW-EB469 **i** **x**

Description: **i**

Organism Details

i If your organism is not found please go [here](#)

Search: **i**

* Tax Id: 2697049 **i** **x**

* Scientific Name: Severe acute respiratory syndrome coronavirus 2 **i** **x**

Common Name: **i** **x**

Collection event information

collection date: **i** **x**

* geographic location: **i** **x**

<< Previous [Submit](#)

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.

Navigation: < Previous Sample | Next Sample >

* Tax Id: 2697049

* Scientific Name: Severe acute respiratory syndrome coronavirus 2

Common Name:

Collection event information

collection date:

* geographic location (country and/or sea): United Kingdom

Tooltip: The date of sampling, either as an instance (single point in time) or interval. In case no exact time is available, the date/time can be right truncated i.e. all of these are valid ISO8601 compliant times: 2008-01-23T19:23:10+00:00; 2008-01-23T19:23:10; 2008-01-23; 2008-01; 2008.

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 submit.

Registering samples and submitting sequencing metadata (Part 2)

1w

12 Submitting run & experiment metadata

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

Please select the file format. If you have files of different types please submit them in separate submissions.

☐ CRAM

☒ BAM

One BAM file is submitted for each run.

☐ SFF

☐ One Fastq file (Single)

☐ Two Fastq files (Paired)

☐ Complete Genomics

☐ PacBio HDF5

☐ Oxford Nanopore

Mandatory fields are denoted by (*).

Download Template Spreadsheet | Upload Completed Spreadsheet | Download Spreadsheet

[Sample reference suggestions]	Sample reference (*)	Instrument Model (*)	Library Name	Library Source (*)	Library Selection (*)	Library Strategy (*)	Des
✖	test	test	Illumina HiSeq 40i	VIRAL RNA	PCR	AMPLICON	
+							

Note

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 informations 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! You successfully completed a TEST submission, please go to [this page](#) to perform your submission in PRODUCTION

Dear Colleague,

Thank you for your recent submission to the European Nucleotide Archive with your account 'Webin-55756'.

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

Please DO NOT reply to this email. If you need further assistance, contact ENA helpdesk <https://www.ebi.ac.uk/ena/browser/support>

Accession Summary

Type Accession Unique Name
Sample ERS4783409 (SAMEA7015592) test
Experiment ERX4244183 ena-EXPERIMENT-test-25-06-2020-15:19:22:337-1
Run ERR4292371 ena-RUN-test-25-06-2020-15:19:22:337-1

Kind regards,

European Nucleotide Archive (ENA)

13 Epilogue

Congratulations! If you've made it this far you've successfully uploaded SARSCov2 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/>