NCBI – GenBank Influenza Submission Using the Submission Portal Programmatic Interface, May, 2024

This is a description of how to use the influenza virus programmatic interface of the GenBank Submission Portal. This tool can be used for large and/or frequent submissions of Influenza A, B, or C virus sequences (updates to include Influenza D virus are coming soon).

Users should contact gb-admin@ncbi.nlm.nih.gov to discuss requirements for submission through the programmatic interface. An FTP upload directory will be created for each group to deliver their submission files. Submitters should also have a MyNCBI user account.

Each time a user submits, they should create a new submission folder under their FTP upload directory. Please make sure the submission folder has group read, write, and execute permissions (use Linux commands "ls -dl ..." to see permissions, and "chmod" to change them if needed).

Data Files

An archive file (.zip or tar.gz) containing the following should be uploaded to the appropriate submission folder. Each file must have a specific extension, as shown in parentheses below.

Description of the files and mandatory file extensions:

- **Sequence data (.fsa)** Nucleotide sequences in FASTA format. The sequence identifier must match the Sequence_ID used in the source information table. Spaces are not allowed in the sequence identifier.
- **Source information table (.src)** Tab-delimited text file which must include: sequence_ID, strain, collection-date, host, geo_loc_name (geographic location name where sample was collected), isolation-source, and serotype for Influenza A viruses. Other source metadata may also be included; see here for a list of all source modifiers: https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers . A source information tables is required unless the source information is provided in the FASTA file as part of the deflines.
- **Submission template (.sbt)** Text file with submitter names and organizations, as well as publications associated with or describing the sequence. Users can generate submission template files by entering their information at https://submit.ncbi.nlm.nih.gov/genbank/template/submission/, and saving a template that can be submitted with each of their submissions.
- **Structured comment (.cmt)** This is an OPTIONAL tab-delimited text file which can be used to provide additional metadata that does not fit standard source modifiers. If a structured comment file is included with the submission, a "comment" column cannot be included in the source information table. For more information, please see https://www.ncbi.nlm.nih.gov/genbank/structuredcomment/

In addition to the archive file, users must upload a submission form with the file name "submission.xml", which includes instructions for the submission pipeline.
http://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/submit/public-docs/common/

- The .zip or .tar.gz file name must match the "file_path" in the submission.xml file.
- The SPUID is a user-generated identification number. Before an accession number is assigned, the SPUID allows submitters to keep track of samples as they are being processed by NCBI and within their own LIMS if they choose. If there are problems during submission, the SPUID will be used to identify samples in communications, and it can be used to link accession numbers to other records like BioSample.
    - Here is an example of a SPUID line from a submission.xml file:
        - `<SPUID spuid_namespace="winter-grippe-surveillance">2018-07-20_09.102</SPUID>`
    - The value for "spuid_namespace" in the first part of the line (`winter-grippe-surveillance`) will stay the same for each submission from the submitter.

- The SPUID is the **second part** of the line (`2018-07-20_09.102`), and **must be unique** for each submission from the submitter.

After the data files and submission.xml are ready to be submitted, the submitter needs to upload an empty text file named "submit.ready" into the submission folder.

For example, a submission might look like this:
FTP upload directory (*This is an example and will not work. Please contact gb-admin@nlm.nih.gov to have an FTP upload directory generated for you or your group*):
`ftp://login@ftp-private.ncbi.nlm.nih.gov/`

Submission folder:
`ftp://login@ftp-private.ncbi.nlm.nih.gov/Production/20200420/`

Upload these files to folder 20200420:

1. .fsa, .sbt, .src, (optional .cmt) – all together in "**`flu.zip`**"
2. `submission.xml` – references **`flu.zip`** in the `file_path` line
3. `submit.ready`

`The submission folder must contain only the data.zip file, the submission.xml file and the submit.ready file. Any additional files or subfolders will cause the submission to fail.`

After submission

The Submission Portal software scans the upload directories several times per day. When it finds a new "submit.ready" file, it checks the associated archive file to make sure all of the required data files are present, and begins processing the files. The influenza virus annotation tool FLAN is included in the pipeline, so users do not need to provide annotation for their submissions. To see how your sequence will be annotated, please use the NCBI Influenza virus Annotation Tool at https://www.ncbi.nlm.nih.gov/genomes/FLU/annotation/.

In case of errors, Submission Portal provides diagnostic message with listing of failed actions and files and provides error descriptions.

Upon completion, Submission Portal creates submission report file in the submission folder. This submission report has the name "report.<N>.xml", where <N> stands for consecutive numbers 1, 2, etc. The first report file made by Submission Portal is always "report.1.xml" and, if there are more updates to correct errors in the submission, can create report files "report.2.xml," "report.3.xml," etc.

The report file contains the status for every action in the submission. See the complete list of statuses in https://www.ncbi.nlm.nih.gov/viewvc/v1/trunk/submit/public-docs/common/docs/UI-lessSubmissionProtocol.docx?view=log, download the most recent revision and go to Appendix A. If some actions have status Processed-error, those need to be corrected and resubmitted. Actions that have status Processed-ok cannot be resubmitted. Once all actions in the submission have status Processed-ok, no further updates to the folder are processed by Submission Portal and this folder could be removed to reclaim space.