NCBI will no longer make taxonomy identifiers for individual influenza strains on January 15, 2018

Authors: Hatcher, Eneida L¹; Bao, Yiming²; Amedeo, Paolo³; Blinkova, Olga¹; Cochrane, Guy⁴; Fedorova, Nadia B³; Gruner, William E⁵; Leihe, Detlef D¹; Nakamura, Yasukazu⁶; Ostapchuk, Yuri¹; Palanigobu, Vasuki¹; Sanders, Robert¹; Schoch, Conrad¹; Smith, Catherine⁷; Wentworth, David E⁷; Yankie, Linda¹; Zhdanov, Sergey A; Karsch-Mizrachi, Ilene¹; Brister, J. Rodney¹.

1. Information Engineering Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health. Bethesda, MD, USA.
3. J. Craig Venter Institute. Rockville, MD, USA.
4. European Molecular Biology Laboratory, European Bioinformatics Institute, Welcome Trust Genome Campus. Hinxton, Cambridge, UK.
5. United States Air Force Materiel Command, School of Aerospace Medicine. Wright-Patterson AFB, OH, USA.
7. National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention. Atlanta, GA, USA.

*Co-first authors

Corresponding Author: Eneida Hatcher, Eneida.Hatcher@NIH.gov
Abstract

Currently the National Center of Biotechnology Information (NCBI) assigns individual taxonomy identifiers to each distinct influenza virus isolate submitted to GenBank. To support this practice, individual flu isolates must be manually added to the NCBI taxonomy database and unique taxonomy identifiers generated. This added layer of manual processing is unique to influenza virus and prevents automatization of the flu sequence submission process. Here we outline a new NCBI policy that normalizes Influenza virus taxonomy processing but maintains features supported by the previous approach. This change will reduce the amount of manual handling necessary for flu submissions and pave the way for increased automation of the submissions process. While this automation may disrupt some historic practices, it will better align influenza virus data processing with other viruses and ultimately lower the submission burden on data providers.

Introduction

GenBank is a member of the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane et al. 2016) - data repositories dedicated to providing public access to biological sequence data. Viral taxonomy within INSDC databases follows the guidelines provided by the International Committee on the Taxonomy of Viruses (ICTV). The scope of the ICTV mandate extends from species to higher level taxa, and no subspecific taxa are maintained by the ICTV (Adams et al. 2017).

All viral sequences submitted to GenBank and other INSDC repositories are assigned to a species. Sequences from characterized viruses are assigned to their pre-existing species. Sequences from novel viruses are assigned to newly created, unclassified species. Typically, subspecific taxonomic ranks are not created at the time of submission, though some formally unranked subspecific taxa are made during post-submission taxonomic revisions. Creation of new viral taxa within the NCBI taxonomy database - whether families, species, or subspecific ranks - requires manual validation and database operations.

There are currently more than 550,000 Influenzavirus A, B, and C nucleotide sequences in GenBank - nearly twenty percent of the entire viral nucleotide sequence content of this database (see Table 1). These sequences represent a coordinated effort by the international scientific community to share critical public health data (Bao et al. 2008), and it is imperative that GenBank provides efficient data distribution pathways to support this and similar efforts. Given the number of influenza virus sequences generated by the scientific community, efficient distribution to GenBank can only be sustained through increased automation of the submissions process.
Table 1. Number of influenza virus nucleotide sequences submitted by year.

<table>
<thead>
<tr>
<th>Year</th>
<th>Sequences submitted</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>50,383</td>
</tr>
<tr>
<td>2015</td>
<td>59,300</td>
</tr>
<tr>
<td>2016</td>
<td>84,015</td>
</tr>
<tr>
<td>Total*</td>
<td>551,664</td>
</tr>
</tbody>
</table>

*Total includes all sequences submitted before October 2017

Problem

Each newly submitted influenza isolate is given a unique strain name, which must be manually validated and added to the NCBI taxonomy database. Almost 14,000 individual influenza strain names were added to the NCBI taxonomy database in 2016 (see Table 2). With more than 120,000 strain names total, this manual step has become a significant impediment to influenza virus sequence submission automation that impacts both traditional GenBank submissions and large-scale submissions.

Table 2. Number of influenza virus strain names created by year.

<table>
<thead>
<tr>
<th>Species</th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
<th>Total*</th>
<th>Number of strains which include complete genome sets*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza A virus</td>
<td>8,567</td>
<td>12,600</td>
<td>11,120</td>
<td>103,927</td>
<td>37,283</td>
</tr>
<tr>
<td>Influenza B virus</td>
<td>1,301</td>
<td>3,874</td>
<td>2,761</td>
<td>16,394</td>
<td>6,066</td>
</tr>
<tr>
<td>Influenza C virus</td>
<td>9</td>
<td>36</td>
<td>48</td>
<td>272</td>
<td>26</td>
</tr>
<tr>
<td>Influenza D virus</td>
<td></td>
<td></td>
<td>45</td>
<td></td>
<td>25</td>
</tr>
</tbody>
</table>

*Includes all sequences submitted before October 2017
The burden of manual taxonomy operations falls not only on GenBank staff, but submitters also face delays and/or must complete extra steps to prepare submissions. For large scale submitters, this means submitting lists of proposed strain names and waiting for their approval before being able to submit sequence data. As it stands, this situation is at odds with both the need for GenBank to introduce new, more efficient submissions pathways and the desire of data providers to reduce submissions burden and timelines.

**Proposed solution**

To facilitate more efficient submissions, NCBI will stop assigning strain level taxonomy designations to influenza virus sequences on January 15, 2018. From that point forward, influenza sequences will be assigned to the relevant species and will be associated with a species-level taxonomy identifier. The taxonomy identifiers shown in Table 3 will be automatically assigned to sequences as part of the GenBank submission process after the species of the sequence is verified by homology.

**Table 3.** Influenza virus species NCBI Taxonomy identification numbers.

<table>
<thead>
<tr>
<th>Species</th>
<th>Taxonomy identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Influenza A virus</td>
<td>11320</td>
</tr>
<tr>
<td>Influenza B virus</td>
<td>11520</td>
</tr>
<tr>
<td>Influenza C virus</td>
<td>11552</td>
</tr>
<tr>
<td>Influenza D virus</td>
<td>1511084</td>
</tr>
</tbody>
</table>

Once this change is made, both species and strain names will continue to be visible in the DEFINITION lines of GenBank records, BLAST results, NCBI’s ftp sites, and the NCBI Influenza Virus Resource. However, there will be changes to influenza GenBank records. The organism name will no longer include both species and strain information. Only the species name will be listed, and strain information will be included in the "/strain" field as illustrated in Fig. 1 (Please see Supp. Fig. 1 to see an example of a complete GenBank record).
Current format:

DEFINITION  Influenza A virus (A/alien/Mars/1/2033(H20N15)) segment 4, complete sequence.

SOURCE  ORGANISM  Influenza A virus (A/alien/Mars/1/2033(H20N15))

   source 1.1740
    /organism="Influenza A virus (A/alien/Mars/1/2033(H20N15))"
    /mol_type="viral cRNA"
    /strain= "A/alien/Mars/1/2033"
    /serotype="H20N15"
    /host="alien"
    /segment="4"

Proposed format after December 1, 2017:

DEFINITION  Influenza A virus (A/alien/Mars/1/2033(H20N15)) segment 4, complete sequence.

SOURCE  ORGANISM  Influenza A virus

   source 1.1740
    /organism="Influenza A virus"
    /mol_type="viral cRNA"
    /strain= "A/alien/Mars/1/2033"
    /serotype="H20N15"
    /host="alien"
    /segment="4"

Figure 1. Differences in influenza virus GenBank records after NCBI no longer makes strain-level organism names for flu sequences.

GenBank nucleotide (https://www.ncbi.nlm.nih.gov/nuccore/) and protein (https://www.ncbi.nlm.nih.gov/protein/) databases will continue to support searches based on components of strain names (e.g. A/New York/61A/2003), as will the NCBI Influenza Virus Resource (https://www.ncbi.nlm.nih.gov/genome/viruses/variation/flu/). This resource also supports sequence downloads that include user-defined DEFINITION lines derived from isolate descriptors and other sequence metadata.

New strain-level taxonomy names will not be created in NCBI's taxonomy database. For example, pages such as this one https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=311639 will not be made for individual influenza virus strains after January 15, 2018. All new submissions will point to species level pages such as this one for Influenza A virus - https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Info&id=11320.
Influenza virus strains will not be found by searching the NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) using new strain names. However, previously existing strain-level names will not be removed from the NCBI taxonomy database and will still be found using strain name searches. The NCBI Taxonomy pages for the Orthomyxoviridae family, Influenzavirus genera, influenza virus species, and legacy strain rank pages will include a highlighted banner to make users aware of the changes in policy.

The NCBI BioSample resource (Barrett et al. 2012) will continue to provide unique accessions for individual samples/isolates if requested. This accession will be linked to submitted sequences in SRA and GenBank. BioSample records also provide storage for highly detailed sample descriptors, providing a much richer biological context to sequences compared to the source descriptors in the GenBank record. More information about BioSample can be found here: https://www.ncbi.nlm.nih.gov/biosample.

**Submitting Influenza sequences to GenBank**

As of February 1, 2018, there will be two options to submit Influenza A, B, or C virus sequence data: an interactive web wizard and a programmatic interface. Both options will facilitate submissions by including an automated process for annotation, so users will not have to include annotation files. To see how your sequence will be annotated, please use the NCBI Influenzavirus Annotation Tool at https://www.ncbi.nlm.nih.gov/genomes/FLU/annotation/. We expect to add the capability to accept Influenza D virus sequences through these submission tools in early 2018.

The new interactive web wizard for submitting influenza sequences is at https://submit.ncbi.nlm.nih.gov/subs/genbank/. A description of the submission process and details on the required files are provided at https://submit.ncbi.nlm.nih.gov/genbank/help/ (see also Table 4 and the Supplemental text documents). Data providers will need to submit FASTA-formatted sequence files and a tab-separated table with source information - sequence_ID, isolate, collection-date, host, collection country/geographic origin, isolation-source, and serotype for Influenza virus A. We encourage submitters to include passage history in a “note” column if known, however it is not required at this time. Influenza strain names will be constructed from the source information table in the format “virus type/country/isolate/year(serotype),” although serotype will only be included for influenza A viruses.

A programmatic submission interface will be available for centers submitting large-scale surveillance data. Submitters are encouraged to contact NCBI at gb-admin@ncbi.nlm.nih.gov prior to submission to ensure that the information included in the submission is properly formatted. An archive file (.zip or tar.gz) containing the sequence(s) in FASTA format, source information table in tab-delimited format, a submission template file, and a submission form which includes information necessary for processing should be submitted. Each file must have a specific extension, as shown
in Table 4. Users can generate submission template files by entering their information at [https://submit.ncbi.nlm.nih.gov/genbank/template/submission/](https://submit.ncbi.nlm.nih.gov/genbank/template/submission/), and saving a template that can be submitted with each of their submissions. Details on the submission process through the programmatic interface will be provided soon.

Reassortant influenza strains play an important role in pandemics and vaccine research. Currently, reassortant viruses cannot be submitted through the web wizard, and users should contact gb-admin@ncbi.nlm.nih.gov if they would like to submit reassortant influenza viruses. Although the programmatic interface can be used to submit reassortants, we strongly encourage submitters to send an email first to ensure that their submission files are in the correct format. The strain name should follow the accepted format, e.g., Influenza A virus (A/reassortant/VG-552(Japan/5685/2016 x New York/5689/2016)(H3N2).

**Table 4.** Files required for submission of influenza virus sequences to NCBI.

<table>
<thead>
<tr>
<th>File</th>
<th>Requirement - Web wizard</th>
<th>Requirement - Programmatic Interface (mandatory file extension)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence data</td>
<td>Required</td>
<td>Required (.fsa)</td>
<td>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.</td>
</tr>
<tr>
<td>Source information table</td>
<td>Required</td>
<td>Required* (.src)</td>
<td>Tab-delimited table which must include: <strong>Wizard submissions:</strong> sequence_ID, isolate, collection-date, host, collection country, isolation-source, serotype. <strong>Programmatic submissions:</strong> sequence_ID, strain, collection-date, host, collection country, isolation-source, serotype and any other source metadata that fits the submission. See here for a list of all source modifiers: <a href="https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers">https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers</a></td>
</tr>
<tr>
<td>Submission template</td>
<td>N/A</td>
<td>Required (.sbt)</td>
<td>Submitter names and organizations, and publications associated with or describing the sequence. Can be generated at <a href="https://submit.ncbi.nlm.nih.gov/genbank/template/submission/">https://submit.ncbi.nlm.nih.gov/genbank/template/submission/</a></td>
</tr>
<tr>
<td>Submission form</td>
<td>N/A</td>
<td>Required (.xml)</td>
<td>Includes instructions for the submission pipeline.</td>
</tr>
<tr>
<td>Structured comment</td>
<td>N/A</td>
<td>Optional (.cmt)</td>
<td>Additional metadata that does not have designated fields in GenBank records. For more information, please see <a href="https://www.ncbi.nlm.nih.gov/genbank/structuredcomment/">https://www.ncbi.nlm.nih.gov/genbank/structuredcomment/</a></td>
</tr>
</tbody>
</table>

*Required unless all required source information is provided in the FASTA file as part of the deflines.*
After successful submission, an email will be sent to inform the submitter of the new accession numbers associated with the sequence IDs, the expected release date, and a preview of the records in GenBank format.

In order to facilitate this transition for providers and users of flu data, we will maintain a webpage with the changes outlined in this paper and include links to it from the Influenza virus resource at https://www.ncbi.nlm.nih.gov/genome/viruses/variation/flu/help-center/.

Summary

The change in GenBank influenza virus strain name processing will support improved submissions efficiency. While individual strain names will no longer be added to the NCBI taxonomy database after January 15, 2018, virus strain names will remain available and searchable through NCBI resources. GenBank staff will provide assistance to data providers to ensure a smooth transition to this new policy.

Funding

This research was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. This project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under Award Number U19AI110819.

References


Supplemental Figure 1. Example of a GenBank record for after new Influenza taxonomy policy
Supplemental Text Files

A. Submission template file
B. Source information table including strain for programmatic interface, .src
C. Source information table including isolate for web wizard, .txt
D. Sequence data
E. Submission form