29 September 2014; updated August 2018

Annex 1

Questionnaire for databases

Responses provided by the Swiss Institute of Bioinformatics

Background

Check here if no change from 2014 ☒

1. Name of the database

OpenFluDB

2. When was your database established?

March 2009

3. Who hosts and manages the database?

Swiss Institute of Bioinformatics

Organization and functioning of the database

Check here if no change from 2014 ☒

4. Please describe the sequencing data (e.g. source material, type, etc.) stored in your databases.

Partial and complete genomic sequences of all influenza segments, types A and B, as well as their translations into the proteins.

5. Does your database also store/provide access to associated data (e.g. epidemiological data)? If so, please provide a description.

OpenFluDB is isolate-centric, rather than sequence-centric database. Each virus isolate can be associated with the name of the institution providing the sample, the name of the laboratory that sequenced it and the name of the institution that submitted the data. General information about an isolate includes type, subtype, lineage, passage history, host, and collection date and place. Several clinical data including host age, sex or vaccination status and epidemiological information including *in vivo*-tested antiviral resistance can also be attributed to a virus strain.

6. How many influenza genetic sequences are stored in your database?

515 thousands sequences in 120 thousands isolates

7. How many sequences from influenza viruses with human pandemic potential[1] are stored in your database?

8. What information/annotation is provided about the sequences?

Information	**Yes**	**No**
Submitting institution/Originating laboratory	yes	
Source material	yes	
Country of origin of the data	yes	
Country of origin of the source material	yes	
Date of submission	yes	
Other:		

9. Briefly explain the process to upload data in your database.

Users can deposit data either as a single isolate together with its sequences using a simple web form or a group of isolates by providing a properly formatted Microsoft Excel file together with the related sequences in a FASTA file. Users can populate OpenFluDB via two mechanisms: single isolate upload or batch upload. In addition, a daily automatic procedure imports isolates from GenBank. Uploaded data are checked for quality and consistency. The minimal criterion for each sequence metadata is presence of host species, year and country of the sample collection

10. Briefly explain how data is accessed by users and the general public.

Interaction via OpenFluDB web interface allows user to efficiently retrieve a set of isolates and related sequences according to a comprehensive set of criteria. The basic 'browse' form comprises several multiple select menus to restrict the search on virus type, subtype, lineage, host and sample collection geographical location. To further restrict the search criteria, additional filters, e.g. sample collection date, submission date, minimal sequence length, isolate name, OFL_ISL_ID, passage history, lineage, OFLID, DDBJ/EMBL/GenBank accession number, sequence submitter laboratory, etc can be applied. An estimation of the number of isolates and sequences returned by a query is updated dynamically and displayed when filters are set. The search results can be then submitted to several analysis tools like sequence similarity search and multiple sequence alignment (MSA), or mapped on geographical and sequence similarity maps (SSMs). Isolate records can be exported in Microsoft Excel format, and the nucleotide and protein sequences in FASTA format.

---

[1] The PIP Framework defines 'influenza viruses with human pandemic potential' as any wild-type influenza virus that has been found to infect humans and that has a haemagglutinin antigen that is distinct from those in seasonal influenza viruses so as to indicate that the virus has potential to be associated with pandemic spread within human populations with reference to the International Health Regulations (2005) for defining characteristics. (see PIP Framework Section 4.2)

11. From which countries do most sequences in your database originate?

Top-10:

239845	**USA**
59154	**China**
16856	**Canada**
13294	**Australia**
12400	**United Kingdom**
11897	**Japan**
10913	**Hong Kong**
10708	**Singapore**
10148	**Viet Nam**
9063	**New Zealand**

12. To the best of your knowledge, on average, how quickly after sequencing are sequences uploaded to your database?

On the next week a sequence appears in GenBank

13. How many sequences are uploaded into and downloaded from your database on a monthly basis?

Around 4 thousands sequences are uploaded each month

14. Who are the principal users of your database?

Users	**Yes**	**No**
General Public	Yes	
Academic Institutions	yes	
Researchers	yes	
Industry	Yes	
Unidentified	yes	

Policy

Check here if no change from 2014 ☒

15. Does your database have an access policy?

Data Access policy	**No**

16. If so, does your data access policy cover the following:

Data Access and Use	**Yes**	**No**
Access to the database (registration, identification)	yes	
Use of data for commercial purposes		No
Further sharing of downloaded data with a third-party		No
Uploading of downloaded data to another database		No
Acknowledgment of originating laboratory/country		No
Collaboration with originating laboratory/country		No

Intellectual property rights or other restrictions on the data		No
Suspension/Termination of access to the database		No

Other/Comments:

OpenFluDB is an open access database. Browsing data does not require any registration. Free and uncensored registration is required for user identification to upload sequences. All deposited data are open access. OpenFluDB provides for users a convenient interface to push their data to GenBank.