Annex 1

Questionnaire for databases

Responses provided by the National Center for Biotechnology Information, National Institutes Of Health, United States Department of Health and Human Services

Background

Check here if no change from 2014 ☒

1. Name of the database

International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/)

2. When was your database established?

Early 1980s

3. Who hosts and manages the database?

Three international partners: DDBJ (NIG, Japan; https://www.ddbj.nig.ac.jp/) GenBank (NCBI, US; https://www.ncbi.nlm.nih.gov/genbank/) ENA (EMBL-EBI; https://www.ebi.ac.uk/ena)

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.

INSDC collates, preserves, integrates and presents globally comprehensive sequence and associated data, covering the spectrum from raw reads, through assemblies to functional annotation across all taxa. For influenza, INSDC includes raw sequence reads, assembled sequences corresponding to all segments of influenza and derived protein annotation, metadata (including but not limited to strain name, subtype, host, collection date, country, passage history, segment number). Content currently covers some 22 petabases, 1.1 billion assembled/annotated sequences and 1.8 million taxa.

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

INSDC hosts rich associated contextual information, such as sample information (serotype, geographical coordinates, collection details, host phenotypes, etc.), links to the scientific literature and information on experimental configuration. The nature of these contextual data vary across applications and user communities. Typically we work with expert communities to develop the appropriate standards and reporting structures for the information essential to the expert's domain.

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

Approximately 673,000 Influenza sequences with 628,000 in the specialized Influenza Virus Resource.

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

About 110,800 including Pandemic (H1N1) 2009 sequences; about 4,000 excluding Pandemic (H1N1) 2009 sequences.
*These numbers are based on non-seasonal influenza sequences that have human as host in IVR

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.

Various options, web and programmatic, are available to suit different user types and localities, typically including spreadsheet upload facilities for contextual data and fasta upload for sequence. Submission tools and services are made freely available to all and support and training is provided from the INSDC partner institutions in their use.

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

All data are made freely and openly discoverable and retrievable through search (e.g. sequence similarity, gps coordinates, serotype), web browse and download functions. Programmatic and batch download services are also provided.

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

[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)

Top 10 countries/regions where influenza viruses were collected: USA, China, Canada, Japan, United Kingdom, Australia, Hong Kong, India, South Korea, and Thailand

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

We can upload new sequences within a day of submission if the submitter does not request that we keep the sequences confidential.

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

We are unable to give a full response here as many secondary services mirror and replicate data that this tracking is not possible. At a single site, e.g. GenBank, average monthly upload for Influenza sequences in the past 12 months: 7,700 (range 5,700-14,000).

14. Who are the principal users of your database?

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

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)		
Use of data for commercial purposes		
Further sharing of downloaded data with a third-party		
Uploading of downloaded data to another database		
Acknowledgment of originating laboratory/country		
Collaboration with originating laboratory/country		
Intellectual property rights or other restrictions on the data		
Suspension/Termination of access to the database		
Other/Comments: The three INSDC host institutions impose no restrictions on users' access to data and make the following statements in relation to this: DDBJ: http://www.ddbj.nig.ac.jp/copyright-e.html NCBI: http://www.ncbi.nlm.nih.gov/About/disclaimer.html (see 'Molecular Database Availability')		

EMBL-EBI: http://www.ebi.ac.uk/about/terms-of-use (see item 9)