Annex 1

Questionnaire for databases

Responses provided by J. Craig Venter Institute

Background

Check here if no change from 2014 ☒

1. Name of the database

Influenza Research Database – IRD (www.fludb.org)

2. When was your database established?

2004 (as part of BioHealthBase), name changed to IRD in 2009

3. Who hosts and manages the database?

Team lead by Northrop Grumman and the J. Craig Venter Institute through a contract provided by the U.S. National Institute of Allergy and Infectious Diseases (NIAID)

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.

Nucleotide sequence data from GenBank
Protein sequence data from GenBank and UniProt
Structured and curated metadata about virus isolation location, date, host species, etc. from GenBank, UniProt and IRD
Extensive novel sequence annotations from IRD (e.g. sequence features, phenotypic sequence variations, H1 and H5 clade membership, antiviral drug resistance markers)
Sequence annotations indicating potential sequence artefacts in nucleotide sequences, identified by a detailed curation process using a combination of algorithms and expert review. Curation is applied to all stored sequences and is available to users for quality assessment of all sequences.

Additional epitope sequence annotations from the Immune Epitope Database (IEDB)

(N.B. No sequences from GISAID due to access restrictions.)

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

In addition, IRD supports a database of influenza surveillance records (human, non-human mammalian, and avian), including extensive metadata, from the Centers of Excellence for Influenza Research and Surveillance (CEIRS). Surveillance metadata includes source material, assessment of the presence/absence of influenza virus in the collected specimen, and geolocation and date of sample collection. Epidemiology-focused users can query surveillance and linked sequence data, and use IRD custom tools to

investigate geographic (state/province/nation) factors, strain relatedness, phylogeny, and sequence conservation. Users can search for strains by seasonality in summer, winter, northern and southern hemisphere data quadrants. The availability of such metadata in association with sequence and surveillance data can support epidemiological risk assessment-based studies such as outbreak curves, phylogeography analysis, geolocation heatmaps, and other public health risk factor assessments.

Outbreak strains can be rapidly compared to vaccine strains using the suite of analysis tools and related data (e.g. location of immune epitopes) provided in IRD to determine best fit and relevance of epitope, host, and sequence feature relationships between vaccine and pandemic sequences.

IRD provides data about H1 and H5 clade classification for existing sequence records in IRD based on the WHO H5 classification scheme and the USDA H1 classification scheme. IRD has also implemented H5 and H1 clade classification tools for user-supplied sequences. IRD is also collaborating with the USDA to develop a global H3 classification scheme and analysis tool.

In 2014, IRD added a new feature in which all sequences are now annotated with information about the presence of sequence variations from the CDC H5N1 Genetic Changes Inventory (http://www.cdc.gov/flu/pdf/avianflu/h5n1-inventory.pdf). These include genetic changes that have been demonstrated to affect influence virulence, replication efficiency, polymerase activity, receptor binding, host adaptation, transmission, and anti-viral drug susceptibility. Users can search for and download sequences with any of these genetic characteristics. A tool to screen user-supplied sequences for the presence of these genetic changes is also available in IRD. These data can be used for the development of risk assessment models based on these genetic markers of important viral phenotypic characteristics.

Finally, IRD contains transcriptomic, proteomic and lipidomic data examining host responses to viral infection from experiments involving 15 different influenza virus strains, including A/California/04/2009 and A/Viet Nam/1203/2004.

**** IRD focuses on data aggregation, integration and enhancement, and on support for a wide range of analysis and visualization tools.**

- Examples of data enhancements include the calculation of sequence variation score for selected groups of sequence records, the screening of every sequence record for the presence of sequence variations with known functional consequences (e.g. increased virulence, anti-viral drug resistance), and the annotation of H1 and H5 clade membership.

- Examples of data integration include the integration of sequence variation and immune epitope information with 3D protein structure data for visualization.

- Examples of analysis tools include: BLAST, short sequence search, multiple sequence alignment, phylogenetic analysis, clade classification, sequence variation calculations, statistical comparative sequence analysis, and PCR primer design.

**** IRD provides private workbenches where investigators can store their retrieved data, upload private sequence data, combine IRD sequence data with private sequence data**

for comparative analysis, and save analysis results for future use and controlled sharing with designated collaborators.

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

625,693 segment sequences and 1,007,235 protein sequences from 135,549 strains as of 02JUL2018
(N.B. No sequences from GISAID due to access restrictions.)

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

Subtype	counts*
H5N1	**2559**
H5N6	**24**
H7N1	**1**
H7N2	**8**
H7N3	**16**
H7N7	**60**
H7N9	**918**
H9N2	**163**
H10N7	**2**
H10N8	**48**
Total	**3799**

***Only human isolates; excludes all human H1N1, H1N2, H2N2, H3N2, and LAB strains**

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

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

Other:
*** Information contained in the Reference field Direct Submission type.**

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

IRD facilitates submission of new sequences to GenBank by providing a custom sequence annotation and formatting tool.
Sequences are retrieved from GenBank on a nightly basis and processed through custom annotation and quality assurance tests, including manual curation, to ensure a

[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)

high level of data accuracy and consistency before loading and integrating into the IRD database.
Direct submissions to the public portion of the IRD database are not supported.
Uploading of private sequence data is allowed in a user's personal workbench where the sequence can be annotated with the various custom annotation tools provided by IRD, and integrated with public sequence records from IRD for combined analysis. These private sequence records are never made available through the public IRD interface.

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

Open access public website using custom search interfaces
No restrictions

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

Top 10 countries from which sequence data is available:

Country	# of segment records*
USA	254216
China	58692
Canada	19863
Hong Kong	13345
Japan	12322
United Kingdom	11655
Viet Nam	10750
Australia	10490
Singapore	9391
Netherlands	9019

***As of 02JUL2018**

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

For some sequences, this information is impossible to determine since we have no way of knowing when a sequence was determined for many GenBank records.

However, we work very closely with the NIAID-funded Centers of Excellence for Influenza Research and Surveillance (CEIRS) and the Genomic Centers for Infectious Diseases (GCID) through which many of the full-length influenza genome sequences have been determined. Indeed, of the 38,655 full genome sequences available in IRD, over 22,132 were determined by the GCID at the J. Craig Venter Institute (JCVI). With regards to the timing of submission, both the CEIRS and GCID programs have a policy requiring submission of sequences into public databases (usually GenBank) within 45 days after the sequence has been completed.

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

For the 12-month period July 2017 - June 2018, a monthly average of 7431 segment sequences were uploaded into the IRD database. **For the 12-month period July 2017 - June 2018, a monthly average of 1997 segment sequences were downloaded from the IRD database.**

14. Who are the principal users of your database?

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

Policy

Check here if no change from 2014 ☐

15. Does your database have an access policy?

Data Access policy	**Yes X**	**No**

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

Data Access and Use	**Yes**	**No**
Access to the database (registration, identification)	X	
Use of data for commercial purposes	X	
Further sharing of downloaded data with a third-party	X	
Uploading of downloaded data to another database	X	
Acknowledgment of originating laboratory/country	X	
Collaboration with originating laboratory/country		X
Intellectual property rights or other restrictions on the data	X	
Suspension/Termination of access to the database		X

Other/Comments: **The IRD Data Access and Use Policy is copied below. The mission of the NIH-supported IRD is to provide the highest quality data and data analysis tools to all interested parties worldwide in order to promote research on influenza virus and the promotion of new solutions to fight this on-going serious public health threat by all possible means.**

Irrespective of the type of data access policy in place, no technology currently exists to control and/or restrict the type of uses of internet downloadable data. Furthermore, policies with options to suspend/terminate access to individuals or groups of individuals are ineffective, as those restrictions are very easily circumvented. Instead, such restrictions only serve to discourage the broadest distribution of data to the research/health community. Therefore, in order to ensure the appropriate acknowledgement of the originators of the data or specimens from which data are generated, it is ultimately the research/public health communities who have to develop and implement diligent practices for self-policing to ensure appropriate

acknowledgement of data sharing. Data use policies can be one important way to achieve and encourage this. A thorough peer-review process of scientific manuscripts prior to publication in scientific journals is another such mechanism. To fully achieve this, individual researchers have to be made aware of these issues.

Finally, most databases, including IRD, have no legal standing in protecting intellectual property (IP); instead data providers need to use legally binding mechanisms (patents, copyrights, trademarks, etc.) to protect their IP.

One option that IRD is currently evaluating that promises to provide a solution to data tracking is the use of the blockchain technology that has enabled the cryptocurrency revolution. Blockchain was invented to serve as a public transaction ledger for cryptocurrencies. A blockchain is a growing list of records, call blocks, in which each block contains a cryptographic hash of the previous block, a timestamp, and transaction data. By design, a blockchain is resistant to modification of the data. Once recorded, the data in any given block cannot be altered retroactively without alteration of all subsequent blocks, which requires consensus of the network majority. While this approach to data tracking may provide an ideal solution to the tracking of sequence data use, it is untested for this purpose at the present time.

IRD Data Access and Use Policy

** Please note that the use of any data and/or tools available in IRD for research and teaching purposes must be acknowledged by citation and acknowledgement. The preferred methods for attribution are to cite both* ***the original data providers AND the IRD resource*** *as follows:*

- ***If used as a bibliographic citation, we recommend citing****: Squires et al. (2012) Influenza research database: an integrated bioinformatics resource for influenza research and surveillance. Influenza and Other Respiratory Viruses DOI: 10.1111/j.1750-2659.2011.00331.x.*
- ***If used in acknowledgements or footnotes section we recommend the format****: "The datasets and tools used in this study were provided by Investigator X, Investigator Y and Investigator Z and were obtained through the Influenza Research Database (www.fludb.org), a federally funded project supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, Contract No. NIH N01AI 400041".*
- ***If primary data provider information is available in IRD*** *for a dataset used in your research, it is required to acknowledge them in addition to citing IRD. Please contact them separately if co-authorship is a relevant option.*

If you include an IRD citation or acknowledgment in an accepted manuscript or abstract, the IRD Team would greatly appreciate e-mail notification of the publication citation, solely for tracking IRD usage. Please send email notifications to: feedback@virusbrc.org.

Please note: *While this resource is freely available to researchers neither we as data providers nor our funding agencies are responsible for incorrect data or for incorrect interpretations of any results. Accuracy cannot be guaranteed for data that is predicted using computational tools and sequence similarity methods as they may not be experimentally validated. In some cases data can be incomplete as it may be part of an ongoing research project. To minimize errors, we strongly encourage users to contact the primary data provider(s) to ensure that any anomalies present in the primary data are adequately considered.*

17. Provide any other relevant additional information on the database