FACT SHEET

Genetic sequence data and databases

Background
Genetic sequence data (GSD)
Organisms are built, and their functions are determined, by their genetic code. This code is contained in DNA molecules, which are found in human, animal and plant cells, as well as in microorganisms like bacteria and viruses. DNA has four components, or building blocks, called C (cytosine), G (guanine), A (adenine), or T (thymine). The nucleotides line up in an order particular to each organism, comprising the genetic code. The code can be "read" by the cell to produce RNA – which is also a type of genetic information, encoded by the nucleotides A (adenine), C (cytosine), G (guanine) and U (uracil) – and in turn, proteins that are responsible for all of the structure and function of a living organism.

The order of the nucleotides in DNA and RNA – that is, the sequence – is critical because genetic sequences contain information about the structure and specific properties of an organism. In the case of viruses, this includes characteristics such as pathogenicity, transmissibility and antiviral susceptibility. Laboratories can determine the genetic sequence of a particular organism, using sequencing technologies. The data generated through this process is called genetic sequence data (GSD), which is represented by listing the nucleotides in order (e.g., an RNA sequence might look like AGAAAUGAAUGGCUCUGUCAA).

GSD databases

GSD databases are databases that receive, host and provide access to GSD that has been submitted to them. Most databases also provide important additional, contextual information related to the sequences, to enrich the sequence data with information about the patient/animal/other source from which the sample was extracted. This additional information is known as “metadata”.

Principal databases that host influenza GSD
There are more than 1700 molecular biology databases, at least six of which contain influenza GSD. These are:
- GISAID Epiflu™: https://www.gisaid.org/
- Influenza Research Database: https://www.fludb.org/brc/home.spg?decorator=influenza
- Openflu database: http://openflu.vital-it.ch/browse.php
- The three databases are part of the International Nucleotide Sequence Database Collaboration (INSDC):
  - European Nucleotide Archive: https://www.ebi.ac.uk/ena
  - DNA Data Bank of Japan: http://www.ddbj.nig.ac.jp/

GSD and databases in the PIP Framework
Although GSD is not included in the definition of PIP Biological Materials, genetic sequences are defined in section 4.2: “‘Genetic sequences’ means the order of nucleotides found in a molecule of DNA or RNA. They contain the genetic information that determines the biological characteristics of an organism or a virus”. Additionally, the PIP Framework has several references to GSD and databases:
- 5.2.1 Genetic sequence data, and analyses arising from that data, relating to H5N1 and other influenza viruses with human pandemic potential should be shared in a rapid, timely and systematic manner with the originating laboratory and among WHO GISRS laboratories.
• 5.2.2 Recognizing that greater transparency and access concerning influenza virus genetic sequence data is important to public health and there is a movement towards the use of public-domain or public-access databases such as Genbank and GISAID respectively; and
• 5.2.3 Recognizing that in some instances the publication of genetic sequence data has been considered sensitive by the country providing the virus;
• 5.2.4 Member States request the Director-General to consult the Advisory Group on the best process for further discussion and resolution of issues relating to the handling of genetic sequence data from H5N1 and other influenza viruses with pandemic potential as part of the Pandemic Influenza Preparedness Framework.  

The PIP Framework also sets out certain expectations regarding GSD:
• Annex 5: WHO Collaborating Center Terms of Reference, section B.5: WHO CC “upload available haemagglutinin, neuraminidase and other gene sequences of A(H5) and other influenza viruses with pandemic potential to a publicly accessible database in a timely manner but no later than three months after sequencing is completed, unless otherwise instructed by the laboratory or country providing the clinical specimens and/or viruses (Guiding Principle 9)”.  
• Annex 4, Guiding Principle 9: “WHO GISRS laboratories will submit genetic sequences data to GISAID and Genbank or similar databases in a timely manner consistent with the Standard Material Transfer Agreement”.  

Main features of GSD databases
As elaborated in the table below, GSD databases vary according to several factors, including
• **Scope of data coverage**
• **Level of data curation and analysis**
• **Database access**
• **Terms and conditions of access to, and use of, data**
• **Intellectual property (IP) and other legal rights over the data**

Databases address these factors, depending on their purpose and the community of data providers and data users they serve. The table below provides an overview of the main approaches taken by existing databases.
<table>
<thead>
<tr>
<th>DATABASE FEATURES</th>
<th>Scope of data coverage</th>
<th>Data curation &amp; analysis</th>
<th>Database access</th>
<th>Terms and Conditions (TC) of access to and use of data</th>
<th>Intellectual property rights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comprehensive: Contains data from different organisms</td>
<td>Archival repository: Little curation, annotations and data analysis</td>
<td>Publicly-accessible with no registration: • Accessible to all; • Direct access, no registration/log in required</td>
<td>Database does not have TC • No terms and conditions imposed by database; • Free for all to use, subject to pre-existing rights (e.g. IP) or applicable domestic legislation</td>
<td>Database does not address IP rights or licenses for the use of data: • Publication of GSD in publicly-accessible database places data in public domain (see glossary below), except if there are pre-existing IP rights; • Data users will need to verify if there are any existing rights and restrictions on the use of data</td>
<td></td>
</tr>
<tr>
<td>Specialized: Focuses on specific organisms</td>
<td>Curated database: Curation, annotations and data analysis</td>
<td>Publicly-accessible with registration: • Accessible to all; • Requires registration; • User must log-in to access data</td>
<td>Database has TC • Some terms and conditions, but not enforceable (e.g. scientific etiquette); • Free for all to use, subject to pre-existing rights (e.g. IP) or applicable domestic legislation</td>
<td>Database contains provisions on IP – possible approaches include: • Publication of GSD in publicly-accessible database places data in public domain, except if there are pre-existing IP rights; • Data providers can be requested to grant the database a license to use the data and the database then grants licences to data users; • Future IP on technologies developed using data may be subject to terms and conditions; • Users may be asked not to impose any restriction on the data submitted to the database; • Data users may need to verify if there are any existing rights and restrictions on the use of data</td>
<td></td>
</tr>
<tr>
<td>Private access: Access to data is limited to an institution or a restricted group of users</td>
<td></td>
<td></td>
<td>Database has legally-binding TC • Legally-binding, enforceable terms and conditions on use of data (e.g acknowledgement or third-party transfers of data); • Free for all authorized users, subject to a Data Access Agreement (‘DAA’), pre-existing rights (e.g. IP) and/or applicable domestic legislation</td>
<td>Private access: Future IP rights are not affected by publication in the private database, which has confidentiality constraints on data access. Data must generally be transferred to a publicly-accessible database in order to obtain IP rights over technology developed using the data or to publish in scientific journal.</td>
<td></td>
</tr>
</tbody>
</table>
Glossary
This glossary provides definitions of terms frequently used in the context of data access and databases. It does not purport to be comprehensive as new terms come into use and definitions evolve. The definitions provided in the Glossary are commonly used but not necessarily universally agreed. Therefore, some terms may be used differently in different contexts.

<table>
<thead>
<tr>
<th>Data access agreement (also referred to as ‘database access agreement’ or ‘data use agreement’)</th>
<th>A <em>Data access agreement</em> is a contract between a database user (this includes data providers and data users) on the one hand and the database (custodian of the data) on the other. It generally contains terms and conditions for accessing and using data hosted by the database. These terms and conditions can include: conditions of access, use and re-use of data; further distribution of data; attribution and collaboration with the data provider; intellectual property rights, commercialization and licensing rights over data; suspension and termination of access.</th>
</tr>
</thead>
</table>
| Open access | The concept of ‘Open access’ arose in response to the growing demand for unrestricted, free access to scholarly research. According to the *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* and the *Bethesda Statement on Open Access Publishing*, open access contributions must satisfy two conditions. The Berlin Declaration states:
1. The author(s) and right holder(s) of such contributions grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship (community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use.
2. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, interoperability, and long-term archiving. |
| Open data | ‘Open data’ is a relatively recent term. The concept behind “open data” however predates the invention of the Internet and has arisen from the belief that knowledge is a common good and “that the enormous amount of information routinely collected by government entities should be available to all citizens.”

According to *Open Knowledge International*, ‘open data’ is data that can be “freely used, modified, and shared by anyone for any purpose”. For example, the European Data Portal, which gives access to “metadata of Public Sector Information available on public data portals across European countries”, states that: “[open data] should have no limitations that prevent it from being used in any particular way”, “[it] must be free to use, but this does not mean that it must be free to access”, and “once the user has the data, they are free to use, reuse and redistribute it – even commercially.” The World Bank states that open data has two dimension: 1. The data must be legally open, which means they must be placed in the public domain or under liberal terms of use with minimal restrictions. 2. The data must be technically open, which means they must be published in electronic formats that are machine readable and preferably non-proprietary, so that anyone can access and use the data using common, freely available software tools. Data must also be publicly available and accessible on a public server, without password or firewall restrictions.

| Open source | The concept of open source arose from the field of computer programming. Open source licenses allow a community of users to access source materials, tools and platforms under terms that are meant to encourage free flow of information, collaboration and innovation. Open source licenses usually require that no intellectual property rights be asserted on the materials, tools or platforms. Some licences also require that innovations arising from their use be shared with the community. |
**Public access**

In the context of scientific publication, public access refers to the publication of scientific results, notably scientific manuscripts, in repositories that are accessible to the public. This access is often at low or no cost and is generally subject to copyright (under the principle of Fair Use). For example, the US National Institutes of Health’s Public Access Policy requires that “all investigators funded by the NIH submit or have submitted for them to the National Library of Medicine’s PubMed Central an electronic version of their final peer-reviewed manuscripts upon acceptance for publication, to be made publicly available no later than 12 months after the official date of publication. Provided, That the NIH shall implement the public access policy in a manner consistent with copyright law.”

<table>
<thead>
<tr>
<th>Public access database</th>
<th>A public access database can be described as a publicly-accessible database “where access to the data is provided to a user after registration and explicit acceptance of a data access and use agreement. After registration, access requires using a log-in procedure”(^3); database with a data access agreement “extending a number of key “protections” and assurances to data contributors.”(^3); database with “identification of contributors and users.”(^3).</th>
</tr>
</thead>
</table>

**Public domain**

In the context of intellectual property rights, public domain is understood to mean “intangible materials that are not subject to exclusive IP rights and which are, therefore, freely available to be used or exploited by any person.”

<table>
<thead>
<tr>
<th>Public domain database</th>
<th>A public domain database can be described as a publicly-accessible database that allows “free and unencumbered access to digital sequence information”(^4); database that “do not require a data access or use agreement, nor registration or log-in”(^4); “anonymous access”(^4) database. While some sequences in “public domain databases” are in the public domain (free from exclusive IP rights), many are claimed as part of granted patent or published patent applications and are not in the public domain.(^4)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Publicly accessible</th>
<th>Accessible to all; not private, not limited to an institution and not restricted to a category or group of users.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Publicly accessible database</th>
<th>Database that is accessible to all, including the scientific community, policymakers and the general public. Publicly accessible databases may have access mechanisms, registration and authorization procedures or terms and conditions, but are not private, not limited to an institution and not restricted to a category or group of users.</th>
</tr>
</thead>
</table>

---

3. The INSDC website can be found at [http://www.insdc.org/](http://www.insdc.org/)
5. PIP Framework Annex 5 Terms of Reference for WHO Collaborating Centres, Core Term of Reference B 5
6. PIP Framework Annex 4, Guiding Principles for the development of Terms of Reference for current and potential future WHO global influenza surveillance and response system (GISRS) laboratories for H5N1 and other human pandemic influenza viruses, Principle 9
7. Zou et al., Biological Databases for Human Research, Genomics, Proteomics and Bioinformatics, available at [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411498/#b0005](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4411498/#b0005)
8. Ibid
10. Other legal rights over data will depend on international obligations, national legislation and other pre-existing contractual rights.
12. From Scoping Paper on approaches to seasonal influenza and genetic sequence data under the PIP Framework, p. 19: “In most jurisdictions, naturally-occurring influenza GSD would not be considered patentable subject matter. However, innovations from the development of influenza-related products could be protected if patentability requirements are met (these are generally: novelty; patentability of the subject-matter; and an inventive step). There are therefore a number of sequences in publicly-accessible databases that are claimed as part of patents.” However, “the early, open publication of the gene sequence of a

See for example: "The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank." (GenBank overview, GenBank Data Usage, available at https://www.ncbi.nlm.nih.gov/genbank/

13 See for endnote 11 above.

14 See endnote 11 above

15 Laid SA and Wynberg RP, A Fact-Finding and Scoping Study on Digital Sequence Information on Genetic Resources in the Context of the Convention on Biological Diversity and the Nagoya Protocol, par. 29, available at https://www.cbd.int/doc/c/e95a4fddd/4baee2e772be28edcd10358/dsi-ah teg-2018-01-03-en.pdf; See for example GISAID Database Access Agreement: "2. License Terms. You are hereby granted a non-exclusive, worldwide, royalty-free, non-transferable and revocable license to access and use the GISAID EpiFlu™ Database and Data solely in accordance with this Agreement in all its terms. Without limiting the foregoing, your access to and use of the GISAID EpiFlu™ Database and Data is subject to the following terms and conditions:

a. Use of Any Data Provided by You. You hereby grant (i) GISAID and (ii) all users and Data providers that have agreed to be bound by the GISAID EpiFlu™ Database Access Agreement and that continue to abide by its terms (collectively “Authorized Users”) a non-exclusive, worldwide, royalty-free, and irrevocable license to collect, store, reproduce, access, modify, display, distribute, coordinate, arrange, and otherwise use the Data submitted by You as contemplated by this Agreement. This Agreement does not transfer any other rights or ownership interests in the Data." (GISAID EpiFlu™ Database Access Agreement, Effective: March 16, 2011, available at https://www.gisaid.org/registration/terms-of-use/).

See also for example in the field human genomics, UK Biobank Material Transfer Agreement for data and/or samples, 20 August 2012, section 3, available at http://www.ukbiobank.ac.uk/wp-content/uploads/2012/09/Material-Transfer-Agreement.pdf.

16 Division G, Title II, Section 218 of PL 110-161 (Consolidated Appropriations Act, 2008), sec. 218

17 Ibid. GISAID Database Access Agreement, section 2, para. g.

18 Ibid.

19 See for example GISAID EpiFlu™ Database Access Agreement: "g. Intellectual Property. "You acknowledge and agree that: (i) It is your sole responsibility to obtain any additional authorization or license as may be necessary for your use of the Data." (GISAID EpiFlu™ Database Access Agreement, Effective: March 16, 2011, available at https://www.gisaid.org/registration/terms-of-use/)

20 Authorized users are generally those who have agreed to the DAA and have not had their rights suspended or terminated.

21 "Generally speaking, genetic information that is published without confidentiality constraints on access would be considered non-novel. This does not mean that the information must be available free of charge – genetic information available only on a paid-subscription database would still be considered sufficiently public to preclude novelty and hence patentability – provided those accessing the data were not bound by confidentiality obligations regarding the data." (see WIPO Patent issues related to influenza viruses and their genes: working paper. World Intellectual Property Organization. Geneva: WIPO; 2007, p.19, available at http://apps.who.int/medicinedocs/documents/s21417en/s21417en.pdf)


31 The Open Source Definition (1998) is a series of criteria established by the Open Source Initiative (OSI) for determining what is an open source license. The definition can be found at https://opensource.org/osd.


33 Division G, Title II, Section 218 of PL 110-161 (Consolidated Appropriations Act, 2008), sec. 218

34 As referred to in PIP Framework, section 5.2.


40 As referred to in PIP Framework, section 5.2.


