# Open Access and Data Sharing of Nucleotide Sequence Data

MASANORI ARITA

SPECIAL COLLECTION: MULTIDISCIPLINARY DATA ACTIVITIES BRIDGING THE RESEARCH COMMUNITY AND SOCIETY

ESSAY

ubiquity press

## ABSTRACT

Open access, free access, and the public domain are different concepts. The International Nucleotide Sequence Database Collaboration (INSDC) permanently guarantees free and unrestricted access to nucleotide sequence data for all researchers, irrespective of nationality or affiliation. However, recent virus information is primarily distributed via the restricted-access repository known as the Global Initiative on Sharing Avian Flu Data (GISAID) supported by the World Health Organization. As compensation for the restriction, GISAID needs to meet its initial goal of benefit-sharing among countries and to curb ongoing vaccine diplomacy campaigns.

CORRESPONDING AUTHOR:
**Masanori Arita**
Bioinformation & DDBJ Center, National Institute of Genetics, Yata 1111, Mishima Shizuoka 411–8540, Japan; RIKEN Center for Sustainable Resource Science, 1-7-22 Tsurumi, Yokohama, Kanagawa 230-0045, Japan
*arita@nig.ac.jp*

KEYWORDS:
Nucleotide Sequence data (NSD); International Nucleotide Sequence Database Collaboration (INSDC); Global Initiative on Sharing Avian Flu Data (GISAID); Convention on Biological Diversity (CBD); Nagoya Protocol

TO CITE THIS ARTICLE:
Arita, M. 2021. Open Access and Data Sharing of Nucleotide Sequence Data. *Data Science Journal*, 20: 28, pp. 1–5. DOI: *https://doi.org/10.5334/dsj-2021-028*

# 1. INTERNATIONAL NUCLEOTIDE SEQUENCE DATABASE COLLABORATION

## HISTORY

The collaboration among nucleotide sequence databases began in 1982 and involved the data library at the European Molecular Biology Laboratory (EMBL, Heidelberg, Germany) and GenBank at the Los Alamos Science Laboratory, now called the Los Alamos National Laboratory (LANL, New Mexico, USA). At that time, sequence data were derived from publications and input manually. The manual work was resource-intensive, and both institutes agreed to exchange input results and to provide a free copying service for sequence data on magnetic tapes for all sectors (Arita et al., 2021). In 1987, the DNA Data Bank of Japan (DDBJ, Mishima, Japan) joined the collaboration as the third node in Asia (Fukuda et al., 2021).

This collaboration was later called the International Nucleotide Sequence Database Collaboration (INSDC – *http://www.insdc.org*) (Brunak et al., 2002). After severe competition between Celera Genomics Corp. led by Craig Venter and the international Human Genome Project (HGP) consortium, the policy of the INSDC was established by its advisory board in 2002, which guaranteed free and unrestricted access to all data records without use-restrictions, licensing requirements, or fees for the distribution or use by any party including commercial sectors. The success of the INSDC policy begat the open-access movement of academic journals (2003, Budapest Open Access Initiative) and subsequent waves of open science. The EMBL data library is now called the European Nucleotide Archive (ENA); it is located at the European Bioinformatics Institute (EMBL-EBI, Hinxton, UK). GenBank is maintained at the National Center for Biotechnology Information (NCBI, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA). The services of the INSDC are summarized in ***Table 1***.

| | ANNOTATED SEQUENCES | NGS READS | PROJECT METADATA | SAMPLE INFORMATION | FUNCTIONAL GENOMICS | HUMAN GENOMES |
|---|---|---|---|---|---|---|
| **DDBJ** | DDBJ (1987) | DRA (2009) | BioProject (2011) | BioSample (2013) | GEA (2018) | JGA (2013) |
| **EMBL-EBI** | European Nucleotide Archive (ENA) 1982 | | | | Array Express (2001)/ Expression Atlas | EGA (2008) |
| **NCBI** | GenBank (1982) | SRA (2008) | BioProject (2011) | BioSample (2011) | GEO (2001) | dbGaP (2007) |

**Table 1** The databases of the three INSDC nodes and the year of their inception. The four columns on the left indicate INSDC resources and the two columns on the right are associated repositories that are not formally part of the INSDC exchange. The table is reproduced from (Arita et al., 2021).

## FREE ACCESS, OPEN ACCESS, AND THE PUBLIC DOMAIN

Free access, open access, and the public domain are different notions in their licensing terms.

Free access on the internet means "no fees" only. An often-associated statement "all rights reserved" indicates that the redistribution or commercial use of data may incur fees. To avoid complications arising from implicit rights with respect to free digital contents, "open access (OA)" was defined in 2001 as contents with clearly stated usage rights (Hagemann, 2012). Best known are Creative Commons (CC) licenses, providing multiple options at different restriction levels. Depending on the restriction levels, OA are grouped as either *gratis* or *libre* (Suber, 2008); the former means free as in free beer or free access, while the latter is free as in freedom of speech. Most OA journals adopt *libre* OA, but some adopt *gratis* OA, dictating that, for example, contents are not available for commercial sectors.

Lastly, the public domain indicates the absence or expiration of all rights; any type of use is allowed even without crediting the original data providers.

Nucleotide sequence data (NSD) from the INSDC are closest to the public domain. The INSDC does not demand data-submitters to abandon all rights; instead, it dictates the terms of *libre* data sharing and it does not declare ownership. The INSDC policy permanently guarantees free and unrestricted access to all data using unique identifiers (accession numbers) representing permanent aliases of the original information (Arita et al., 2021; Brunak et al., 2002). Via the accession numbers, the analysis, modification, and sharing of sequence data and of associated information (metadata) are permitted to all users permanently.

The cost of data collection and maintenance has been borne by the US-, European Union (EU)-, and Japanese governments for more than 30 years. The operational cost of the DDBJ

is 10 million USD each year; the total cost of the INSDC is much larger. The INSDC commits to treating all users fairly and unequivocally. Essentially all sequence analyses in biology and medicine depend on INSDC data.

## 2. CHALLENGE OF COUNTRIES AFFECTED BY AVIAN INFLUENZA

The fully public service of the INSDC has not always been welcomed. When the human infection of avian influenza turned into an epidemic in 2006, some countries including Indonesia were hesitant to deposit sequence data in the INSDC to avoid utilization by commercial enterprises in the Global-north without crediting the original data providers. Although the Global Initiative on Sharing Avian Flu Data (GISAID – *https://gisaid.org*) initially intended to accumulate influenza data in the INSDC, an alternative, login-based repository, the GISAID EpiFlu database, was started in 2008 (Bogner et al., 2006; Wikipedia, n.d.).

The hallmark of EpiFlu is its restriction of data reuse. Users, including the original data providers, are forbidden to re-distribute, or even display EpiFlu sequences in connection with any other database; the merging, sharing, or full publication of EpiFlu data are also prohibited (*https://www.gisaid.org/registration/terms-of-use/*). To reproduce any results based on the EpiFlu, researchers must start from the original, un-curated data. The same restriction applies to the database for SARS-CoV-2 genome called EpiCoV. For example, the NextStrain project (*https://nextstrain.org*) visually provides the time-course of pathogens' evolution based on EpiCoV data, but the research team is only allowed to provide contact information on the original sequence data by the GISAID administrators; they cannot share sequences used for their analyses. In essence, the GISAID is a *gratis* OA (free-to-access) database that severely restricts the sharing and reproduction of research results.

This restriction resulted in friction (Noorden, 2021), because the INSDC advocates open sharing of SARS-CoV-2 data without restriction (*https://covid19dataportal.org*). While the US and the EU countries register sequences in both the INSDC and GISAID, other countries, including Japan, deposit data only in GISAID. Consequently, far more SARS-CoV-2 sequences have been accumulated in GISAID. Its supporters attribute the rapid growth to the data restriction, such as the condition of crediting the original data providers especially from the Global-south (Maxmen, 2021).

In summary, there is a trade-off between the rate of global data accumulation and the freedom in data reuse. The overall effect on science is not immediately clear. The rapid accumulation of data may be more important for locating new variants of the virus or its pathogenicity. In the long term, however, access restriction is troublesome because the original data providers may become unreachable. Since re-distribution of the GISAID data is forbidden, reanalysis and validation become time-consuming.

## 3. CHALLENGE BY THE CONVENTION ON BIOLOGICAL DIVERSITY

Another challenge to the INSDC derives from an international treaty, the Convention on Biological Diversity (CBD – *https://cbd.int*), which was signed by 150 countries in 1992 to promote globally sustainable development. The turning point was the Nagoya Protocol on Access and Benefit-Sharing (ABS). It took effect in 2014 to enforce benefit-sharing arising from the utilization of genetic resources. The Nagoya Protocol has changed the landscape of all biology research that uses plants and animals from foreign countries. To access any bio-resource and associated traditional knowledge, researchers must obtain Prior Informed Consent (PIC) from the foreign governments and exchange Mutually Agreed Terms (MAT) with the foreign provider.

At the 2018 CBD Conference of the Parties (COP14 – *https://www.cbd.int/cop/*), digital sequence information (DSI) on genetic resources was integrated within the scope of the CBD (decision 14/20). Not a few developing countries now demand benefit-sharing from the use of DSI, and the benefit-sharing is considered important for the setup of the post-2020 global biodiversity framework (decision 14/34). The post-2020 framework is called so, as the next phase of United Nations Decade on Biodiversity (*https://www.cbd.int/2011-2020/*); one of its aims is setting measurable targets to monitor and assess global efforts.

Accordingly, viable options for DSI have been discussed extensively by multiple parties such as the German WiLDSI project, the consultancy project of the European Commission, and the ICF Consulting Services of the UK (*https://www.cbd.int/article/dsi-webinar-series-2020*). In all discussions, the current INSDC policy is cited as an extreme example without any monetary consequences. More restrictive options have been proposed as pragmatic choices. It should be noted, however, that the INSDC has been supporting research, education, and commerce fairly for all. In addition, the US has not signed on to the CBD framework. The resolution effort for benefit-sharing falls more on the shoulders of the EU and Japan.

All CBD countries recognize the importance of (conditioned) open access to DSI. Discussions revolve around realizability of the accurate tracing of DSI utilization and the types of benefits from DSI. Then the benefits should accrue to the country from which the genetic resources are originally obtained. However, the correct attribution seems difficult as it is not possible to prove the true origin of any DSI. We know only a part of the global genetic ecosystem; sequences are circulating among organisms through many factors. Proving the uniqueness of a certain sequence would be already difficult, not to say of locating its geographical origin. Good examples are traveling animals in the open sea.

The INSDC is recording geographical and species origin of DSI. Using the BioSample repository (***Table 1***), the INSDC provides options to describe the geographical and species origin. The country qualifier of the DDBJ/ENA/GenBank also identifies the "locality of isolation of the sequenced organism indicated in terms of political names for nations, oceans or seas, followed by regions and localities" (*http://insdc.org*). This scheme was determined through the history of the INSDC to trace authors and original location of the sequences. The same scheme works well with respect to patented sequences. In the US, Europe, Japan, and Korea, patented sequences are tagged with accession numbers and publicized within the INSDC framework. Although all records including updates are permanently available, the extent of data utilization and the amount of acquired benefits cannot be traced fully in the current system. Consequently, a different scheme or protocol will be needed.

## 4. RESOLVING THE DSI CONTROVERSY

Historically, the GISAID databases (EpiFlu and EpiCoV) were created to satisfy the benefit-sharing demands of developing countries. For this reason, the system became login-based with severe restrictions on data reuse. Although the scope of the CBD excludes the human genome and pathogens, GISAID represents a model platform for assessing sharable benefits from viral DSI.

Viral DSI yields huge commercial benefits such as mRNA vaccines (Trefis Team, 2020). In late 2020, Pfizer and Moderna already had pre-orders for 1.3 billion and 800 million doses, respectively, and Moderna stock jumped more than 7 times in 2020. The financial benefits to these companies can be imagined.

The size and financial profit of commercial enterprises tend to follow a power-law distribution, i.e., majority of the total profits are harvested by a small group of companies. Developing countries that insisted on the current GISAID framework should assess the contribution of GISAID in terms of ABS, in addition to the size of accumulated sequence data.

## 5. CONCLUSIONS

The WHO-recommended GISAID databases were derived from ABS discussions regarding DSI on the bird-flu. Because the SARS-CoV-2 sequences yield significant monetary benefits as mRNA vaccines, responsible sectors should assess the benefit-sharing scheme that is based on the access statistics of the GISAID databases. Although the US and its commercial sectors are not included in the CBD treaty, this exercise provides a good example on how ABS is handled by DSI databases including the INSDC.

DSI from the INSDC is far more complex and comprehensive than the GISAID information. The SARS-CoV-2 example offers a good initial step for the ABS discussion on DSI. OA options for SARS-CoV-2 information are important not only for human health but also for the CBD framework.

## ACKNOWLEDGEMENTS

The author is grateful to the INSDC members. DDBJ is supported by the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of Japan, the AMED-CREST (19gm1010006), JST-CREST (JPMJCR20H1), and the National Bioscience Database Center (NBDC). The author also thanks Ursula Petralia for editing the manuscript.

## COMPETING INTERESTS

The author serves as Head of DDBJ, a part of INSDC, since 2018.

## AUTHOR AFFILIATIONS

**Masanori Arita** *orcid.org/0000-0001-6706-0487*
Bioinformation & DDBJ Center, National Institute of Genetics, Yata 1111, Mishima Shizuoka 411–8540, Japan; RIKEN Center for Sustainable Resource Science, 1-7-22 Tsurumi, Yokohama, Kanagawa 230-0045, Japan

## REFERENCES

**Arita, M, Karsch-Mizrachi, I** and **Cochrane, G.** 2021. The international nucleotide sequence database collaboration. *Nucleic Acids Res*, 49(D1): D121–D124. DOI: *https://doi.org/10.1093/nar/gkaa967*

**Bogner, P, Capua, I, Lipman, DJ, Cox, NJ,** et al. 2006. A global initiative on sharing avian flu data. *Nature*, 442: 981. DOI: *https://doi.org/10.1038/442981a*

**Brunak, S, Danchin, A, Hattori, M, Nakamura, H, Shinozaki, K,** et al. 2002. Nucleotide sequence database policies. *Science*, 298(5597): 1333. DOI: *https://doi.org/10.1126/science.298.5597.1333b*

**Fukuda, A, Kodama, Y, Mashima, J, Fujisawa, T** and **Ogasawara, O.** 2021. DDBJ update: streamlining submission and access of human data. *Nucleic Acids Res*, 49(D1): D71–75. DOI: *https://doi.org/10.1093/nar/gkaa982*

**Hagemann, M.** 13 Feb 2012. Ten years on, researchers embrace open access. *Open Society Foundations (Voices)*. *https://www.opensocietyfoundations.org/voices/ten-years-on-researchers-embrace-open-access*.

**Maxmen, A.** 2021. Why some researchers oppose unrestricted sharing of coronavirus genome data. *Nature (news)*, 593: 176–177. DOI: *https://doi.org/10.1038/d41586-021-01194-6*

**Noorden, RV.** 2021. Scientists call for fully open sharing of coronavirus genome data. *Nature (news)*, 590: 195–196, DOI: *https://doi.org/10.1038/d41586-021-00305-7*

**Suber, P.** Aug 2, 2008. SPARC Open Access Newsletter issue #124. *https://legacy.earlham.edu/~peters/fos/newsletter/08-02-08.htm*.

**Trefis Team.** 16 Dec 2020. Pfizer and Moderna's vaccines could be more profitable than you think. *Forbes*. *https://www.forbes.com/sites/greatspeculations/2020/12/16/pfizer-and-modernas-vaccines-could-be-more-profitable-than-you-think*.

**Wikipedia.** GISAID. *https://en.wikipedia.org/wiki/GISAID*. Last accessed 30 Mar 2021.

TO CITE THIS ARTICLE:
Arita, M. 2021. Open Access and Data Sharing of Nucleotide Sequence Data. *Data Science Journal*, 20: 28, pp. 1–5. DOI: *https://doi.org/10.5334/dsj-2021-028*

Submitted: 31 March 2021
Accepted: 04 September 2021
Published: 15 September 2021



*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.