Open Access DNA, RNA and Amino Acid Sequences: The Consequences and Solutions for the International Regulation of Access and Benefit Sharing

Charles Lawson and Michelle Rourke*

This article addresses how open access to DNA, RNA and amino acid sequences might be reconciled with the benefit-sharing obligations under the United Nations' Convention on Biological Diversity and its Nagoya Protocol, the Food and Agriculture Organization of the United Nations' International Treaty on Plant Genetic Resources for Food and Agriculture, and the World Health Organization's Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits. Tracing the evolution of open access databases, the article posits models for reconciling open access and benefit sharing; the article concludes, however, that none of the proposed solutions – monitoring and tracing, the contract model, and the copyright and database right model – provides a perfect solution. Each model does, however, suggest that open access to these sequences might be at least partially reconciled with benefit sharing.

INTRODUCTION

Science in recent centuries has been founded on the open (and open access)[1] publication of scientific concepts and the data and information[2] as evidence that supports an understanding of those concepts.[3] This open publication has been important both to verify and confirm the concepts and to build on and develop them further "by standing on the shoulder of giants".[4] In modern neoliberal times this sequential building is best conceived as a market:

> [T]he coordinating functions of the market are but a special case of coordination by mutual adjustment. In the case of science, adjustment takes place by taking note of the published results of other scientists; while in the case of the market, mutual adjustment is mediated by a system of prices broadcasting current exchange relations, which make supply meet demand.[5]

* Charles Lawson, Professor, Australian Centre for Intellectual Property in Agriculture, Griffith Law School, Griffith University, Gold Coast, Queensland, Australia; Michelle Rourke, PhD Candidate, Griffith Law School, Griffith University, Gold Coast, Queensland, Australia; Lieutenant, Australian Army Malaria Institute, Gallipoli Barracks, Enoggera, Queensland, Australia. The opinions expressed here are those of the authors and do not necessarily reflect those of the Australian Defence Force.

Correspondence to: c.lawson@griffith.edu.au.

[1] The term "open" and "open access" are used in this article in the sense that the scientific concepts, the data and the information can be accessed without limitation other than being able to access the forums where the outputs are available, predominately books, journals, conference proceedings and so on, and various other repositories such as libraries, book shops, the internet and so on. The term "open access" covers the circumstances where access is free of charge and without restrictions (gratis) and *not* free of charge with some limits on how the accessed materials might be used (such as Creative Commons licenses) (libre). There are a number of regulatory incidents about open access: see, eg *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities* (2003) <https://openaccess.mpg.de/Berlin-Declaration>; *Bethesda Statement on Open Access Publishing* (2003) <http://legacy.earlham.edu/~peters/fos/bethesda.htm>; *Declaration of the Budapest Open Access Initiative* (2002) <http://www.budapestopenaccessinitiative.org>.

[2] "Data" might be considered the quantitative or qualitative values and "information" might be considered the meaning that is apparent from the data when considered alone or in combinations and within a particular context. See L Bygrave, "Information Concepts in Law: Generic Dreams and Definitional Daylight" (2015) 35 *Oxford Journal of Legal Studies* 91.

[3] See Working Group, *Science as an Open Enterprise* (Royal Society Science Policy Centre 02/12, The Royal Society, 2012) 13.

[4] H Turnbull (ed), *The Correspondence of Isaac Newton* (Cambridge University Press, 1959) Vol 1, 416.

[5] M Polanyi, "The Republic of Science: Its Political and Economic Theory" (2000) 38(1) *Minerva* 1, 4.

Within this market, however, there are competing claims for open access (through sharing, collaborating and disseminating through publication) and contrary claims for maintaining secrecy in order to protect possible commercial activities and possible intellectual property claims.[6] The evolution of open access to DNA, RNA and amino acid sequence data and information (generally through internet-accessible databases setting out DNA, RNA and amino acid sequences) provides an interesting insight into this balancing of open access against maintaining secrecy and other restrictions. The concern of this article is the development of an international regulatory framework to enable (or facilitate) access to genetic resources (comprising DNA, RNA, amino acids and other functional units of heredity)[7] and benefit sharing from using those resources (termed access and benefit sharing or ABS),[8] and the challenge posed by open access DNA, RNA and amino acid sequence data and information to the operation of these ABS arrangements. In short, this is a conflict between two regulatory models: open access that allows the benefits to dissipate into a broader public; and ABS that captures some of those benefits for a narrow and defined public.

The generalised international ABS regulatory scheme covering *all* genetic resources (except human genetic resources)[9] is set out under the United Nations' *Convention on Biological Diversity* (CBD)[10] and its *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity* (Nagoya Protocol).[11] There are then ABS-specific compatible schemes now directed to some agricultural plants under the Food and Agriculture Organization of the United Nations' *International Treaty on Plant Genetic Resources for Food and Agriculture* (Plant Treaty)[12] and human pandemic influenza virus under the World Health Organization's (WHO) *International Health Regulations* and its *Pandemic Influenza Preparedness Framework for the Sharing of Influenza Viruses and Access to Vaccines and Other Benefits* (PIP Framework).[13] There are also schemes developing under the United Nations' *Law*

[6] See, eg P David, "The Economic Logic of "Open Science" and the Balance Between Private Property Rights and the Public Domain in Scientific Data and Information: A Primer" in J Esanu and P Uhlir (eds), *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium* (Basic Books, 2003) 19-34. See also G Church et al, "Public Access to Genome-wide Data: Five Views on Balancing Research with Privacy and Protection" (2009) 5(1) *PLoS Genetics* e1000665; P Uhlir and P Schröder, "Open Data for Global Science" (2007) 6 *Data Science Journal* OD36.

[7] See *Convention on Biological Diversity* [1993] ATS 32, Art 2 (CBD) defines "genetic resources" as "genetic material of actual or potential value" and "genetic materials" as "any material of plant, animal, microbial or other origin containing functional units of heredity". In the context of the CBD and the other conservation agreements, the subject of "genetic resources" is probably better characterised as broadly anything that comprises all living and fossil organism material unless it is expressly excluded, and will include biochemicals, excludes ex-situ holdings acquired before 29 December 1993, human genetic materials, and marine living (genetic and biochemical) resources beyond national jurisdiction: see Conference of the Parties to the Convention on Biological Diversity, *Access to Genetic Resources and Benefit-sharing: Legislation, Administrative and Policy Information*, UNEP/CBD/COP/2/13 (1995) [49]-[65].

[8] For an overview of these schemes, see C Lawson, *Regulating Genetic Resources: Access and Benefit-sharing in International Law* (Edward Elgar, 2012).

[9] See Conference of the Parties to the Convention on Biological Diversity, n 7, 15-18. See also Conference of the Parties to the Convention on Biological Diversity, *Report of the Sixth Meeting of the Conference of the Parties to the Convention on Biological Diversity*, UNEP/CBD/COP/6/20 (2002) 60-62 and 253-269 (Bonn Guidelines, cl 19).

[10] CBD, n 7.

[11] Conference of the Parties to the Convention on Biological Diversity, *Report of the Tenth Meeting of the Conference of the Parties to the Convention on Biological Diversity*, UNEP/CBD/COP/10/27 (2010) [103] and Annex (Decision X/1, Annex 1 (Nagoya Protocol) 89-109). There are presently 196 contracting parties to the CBD and 64 parties to the Nagoya Protocol: see <https://www.cbd.int/information/parties.shtml>.

[12] *International Treaty on Plant Genetic Resources for Food and Agriculture* [2006] ATS 10 (Plant Treaty). There are presently 136 contracting parties: see <http://www.planttreaty.org/list_of_countries>.

[13] World Health Organization, *Pandemic Influenza Preparedness: Sharing of Influenza Viruses and Access to Vaccines and Other Benefits*, A64/8 (2011) Attachment 2 (PIP Framework). This was adopted by the member states of the World Health Organization: World Health Organization, *Pandemic Influenza Preparedness: Sharing of Influenza Viruses and Access to Vaccines and Other Benefits*, Sixty-fourth World Health Assembly, WHA64.5 (2011) [1]. There are presently 194 member states: see <http://www.who.int/countries/en>.

of the Sea Convention,[14] and for various classes of organisms like microorganisms, livestock and forestry, although these are in the very early stages of negotiation.[15] The founding principle of each of these regulatory forms is recognition that nation states have sovereignty over the genetic resources within their jurisdictions and authority to determine the terms and conditions for accessing their genetic resources.[16] Each of the existing schemes then enables (or facilitates) access to genetic resources in exchange for benefit sharing, either through an individually negotiated contract between a resource holder and bioprospector under the CBD and Nagoya Protocol,[17] or as a Standard Material Transfer Agreement (SMTA) with fixed terms and conditions under the Plant Treaty[18] and PIP Framework.[19] In addition to the SMTA, the PIP Framework also provides for an annual contribution from influenza vaccine, diagnostic and pharmaceutical manufacturers using the framework according to a formula determined by the WHO.[20] Notably, national laws implementing these schemes are either still to be implemented[21] or only cover some genetic resources within a jurisdiction. So, for example, in Australia the Commonwealth scheme only applies to genetic resources collected in "Commonwealth areas",[22] and the State and Territory schemes (where they exist) only apply to some State or Territory lands,[23] with the remaining lands in Commonwealth, State and Territory jurisdictions (predominately privately owned lands) subject to no formal ABS obligations.[24] These distinctions are important for the following analyses because the benefits flowing from enabling (or facilitating) access to genetic resources have a range of potential beneficiaries, including the governmental holder with sovereignty (the Commonwealth, States and Territories in Australia), the provider of the resources under a SMTA, and the private landholders where there is no formal ABS regulation. It is only the likely beneficiaries under the governmental laws (the Commonwealth, States and Territories in Australia) implementing the CBD and Nagoya Protocol, Plant Treaty and PIP Framework regulations and SMTAs that are considered in this article.

Central to the CBD and Nagoya Protocol, Plant Treaty and PIP Framework is the transfer of physical samples of genetic materials between the parties to the contracts. For example, the Plant Treaty SMTA provides:

> The Plant Genetic Resources for Food and Agriculture specified in Annex 1 to this Agreement (hereinafter referred to as the "Material") … are hereby transferred from the Provider to the Recipient subject to the terms and conditions set out in this Agreement.[25]

[14] See United Nations General Assembly, *Letter Dated 13 February 2015 from the Co-Chairs of the Ad Hoc Open-ended Informal Working Group to the President of the General Assembly*, A/69/780 (2015).

[15] See Lawson, n 8, 241-246.

[16] See CBD, n 7, Arts 3 and 15; Plant Treaty, n 12, Art 10; PIP Framework, n 13, Art 1 (Principle PP11).

[17] See CBD, n 7, Art 15; Nagoya Protocol, n 11, Art 6.

[18] See Plant Treaty, n 12, Art 12.4.

[19] See PIP Framework, n 13, Art 5.4.

[20] See PIP Framework, n 13, Art 6.14.3. See also World Health Assembly, *Pandemic Influenza Preparedness: Sharing of Influenza Viruses and Access to Vaccines and Other Benefits* A67/36 Add 1 (2014) Annex ([7]-[14]).

[21] The CBD has 196 parties of which only 57 have implemented some type of law, measures or instruments to regulate ABS: JC Medaglia et al, *Overview of National and Regional Measures on Access and Benefit Sharing: Challenges and Opportunities in Implementing the Nagoya Protocol* (CISDL Biodiversity & Biosafety Law Research Programme, 2014) 9.

[22] *Environment Protection and Biodiversity Conservation Act 1999* (Cth) s 525; *Environment Protection and Biodiversity Conservation Regulations 2000* (Cth) reg 8A.02.

[23] "State land or Queensland waters": *Biodiscovery Act 2004* (Qld) ss 5, 10 and Sch; "Commonwealth areas" in the Territory: *Biological Resources Act 2006* (NT) s 9.

[24] On these privately owned lands the landholder determines the terms and conditions of access and benefit sharing according to existing land and other laws: see J Voumard, *Access to Biological Resources in Commonwealth Areas* (Commonwealth of Australia, 2000) 41-49.

[25] Interim Secretariat of the International Treaty on Plant Genetic Resources for Food and Agriculture, *First Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture*, IT/GB-1/06/Report (2006) Appendix G, Art 3.

Open access DNA, RNA and amino acid sequence data and information disrupt this exchange by making the sequence data and information available to the broader public *without* the binding terms and conditions of the contract. This also disrupts the benefit-sharing obligations because the benefit sharing is structured around the "materials" and not the data and information about those "materials". In short, open access is shifting the focus from the raw materials to the non-material values of these genetic resources – the "dematerialisation" of the use of genetic resources, being "the increasing trend for the information and knowledge content of genetic material to be extracted, processed and exchanged in its own right, detached from the physical exchange of the … genetic material".[26] As such, open access to DNA, RNA and amino acid sequences can undermine the ABS schemes set out in the CBD and Nagoya Protocol, Plant Treaty and PIP Framework.

This article traces the evolution of open access to DNA, RNA and amino acid sequences, showing that open access has been effectively achieved for many such sequences. However, some important limitations to open access exist, which, if addressed, could form the basis for resolving the apparent conflict between open access and fair and equitable benefit sharing. The article first discusses the early evolution of DNA sequence databases and the circumstances leading to the development of the GenBank sequence database at the National Center for Biotechnology Information (NCBI) in the United States, the EMBL-EBI database at the European Molecular Biology Laboratory's (EMBL) European Bioinformatics Institute (EBI) in Europe and the DNA Data Bank of Japan (DDBJ) in Japan. This analysis highlights the key decisions made about open access when these databases were conceived and the limitations that were imposed subsequently on those using the databases and their data and information. Next, the article looks at the Human Genome Project and the important influence this project had on promoting open access to DNA sequences. It also considers other recent developments, including the development of guidelines, aimed at addressing concerns such as personal privacy, while promoting broad disclosure and dissemination. The article then reviews the data and information obligations under the CBD and its Nagoya Protocol, the Plant Treaty and the PIP Framework. It concludes with a consideration about how open access to DNA, RNA and amino acid sequences might be reconciled with the benefit-sharing obligations under the CBD and Nagoya Protocol, Plant Treaty and PIP Framework.

DNA, RNA AND AMINO ACID SEQUENCE DATABASES

The first systematic collection and assembly of amino acid sequences as a computerised collection of data was made publicly available through the printed *Atlas of Protein Sequence and Structure* by Margaret Dayhoff and others at the National Biomedical Research Foundation in 1965.[27] These atlases were widely distributed and readers were asked to contribute unpublished sequences in exchange for a free copy of a future atlas.[28] The editors made it clear they would not deal with questions of "history and priority" and asserted copyright limiting the republishing and redistribution of the materials.[29] From 1972, the database was made available on computer-readable magnetic tapes, and again subject to copyright claims and limited republishing.[30] The assertions of copyright, the delays in releasing computer readable magnetic tapes and the financial charges "put [the] project on the side of commercial ventures rather than publicly available resources".[31] In 1980, the nucleic acid sequence

[26] Secretariat of the International Treaty on Plant Genetic Resources for Food and Agriculture, *Report of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture*, IT/GB-5/13/Report (2013) Appendix I.2.

[27] M Dayhoff et al, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, 1965). See also B Strasser, "The Experimenter's Museum: GenBank, Natural History and the Moral Economies of Biomedicine" (2011) 102(1) *Isis* 60, 69-96; B Strasser, "Collecting, Comparing and Computing Sequences: The Making of Margaret Dayhoff's Atlas of Protein Sequence and Structure, 1954-1965" (2010) 43(4) *Journal of the History of Biology* 623, 624, 635-639.

[28] Strasser (2010), n 27, 644-645. See also M Dayhoff et al, "Banking DNA Sequences" (1980) 286(5771) *Nature* 326.

[29] Strasser (2011), n 27, 72, 84; Strasser (2010), n 27, 644.

[30] Strasser (2011), n 27, 72.

[31] Strasser (2011), n 27, 72. See also Strasser (2010), n 27, 644.

database was made available for "free",[32] although this was later changed to a subscription[33] subject to obtaining a password that involved signing an agreement not to redistribute the data.[34] The notice on accessing the database provided:

> Welcome to the NAS [Nucleic Acid Sequence] Reference Data System. You are licensed to use this data for your own research. As a licensee, you are legally obliged not to redistribute the data or otherwise make it available to any other party.[35]

By the late 1970s, however, it was apparent that DNA, RNA and amino acid sequence databases were central to the exploding molecular biology enterprise and that data organisation and analysis were limiting the progress of molecular biology.[36] The eventual outcome was the development of three independent databases of sequences in the United States (GenBank), Europe (EMBL-EBI), and Japan (DDBJ). Dayhoff's restricted database was sidelined by these developments and open access was established as the norm for DNA, RNA and amino acid sequences.

GenBank traces its origins to a workshop at Rockefeller University in 1979 to support a nucleic acid database and the development of appropriate analyses tools.[37] The concerns about establishing a centralised database were on the "moral tensions between different conceptions of credit attribution, data access, and knowledge ownership".[38] The result and outcome of the workshop, however, was only the recognition that there should be "a single computerised and non-proprietary database".[39] Another similar workshop followed in 1980 and recommended establishing at the National Institutes of Health (NIH) a nucleic acid sequence data bank that was computer-based and in "the public domain".[40] By late 1980, an ad hoc advisory committee met to draft the National Institute of General Medical Science (NIGMS) guidelines for a database, and in late 1981, proposals were publicly requested by the NIH that elicited three submissions: a Los Alamos and Bolt, Beranek and Newman Inc proposal (the BBN proposal); a Los Alamos and IntelliGenetics proposal (the IntelliGenetics proposal); and a proposal from Dayhoff at the National Biomedical Research Foundation (the Dayhoff proposal).[41] The BBN proposal and Dayhoff proposal were shortlisted,[42] and on 30 June 1982, the BBN proposal, later called "GenBank", was announced as the successful project.[43]

The content of these proposals highlighted the different ideas about limiting access to the database data and ultimately the choice in favour of open access. It was perhaps also significant that the BBN proposal "did not assert any proprietary interest whatsoever in any data" addressing the NIH's concern at the time[44] about future ownership of database information and copyright in the sequence data,[45] and government policy later reflected in the ideal "that, to the maximum extent possible, the products of

[32] See M Dayhoff et al, "Nucleic Acid Sequence Bank" (1980) 209(4462) *Science* 1182.

[33] Strasser (2011), n 27, 77.

[34] Strasser (2011), n 27, 76-77.

[35] Strasser (2011), n 27, 77.

[36] Strasser (2011), n 27, 66 and the references therein. See also Editorial, "Banking DNA Sequences" (1980) 285(5760) *Nature* 59; R Lewin, "Long-awaited Decision on DNA Database" (1982) 217(4562) *Science* 817.

[37] See Strasser (2011), n 27, 66; T Lenoir, "Shaping Biomedicine as an Information Science" in M Bowden, T Hahn and R Williams (eds), *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems* (ASIS Monograph Series, Information Today Inc, 1999) 35.

[38] Strasser (2011), n 27, 68.

[39] Strasser (2011), n 27, 68.

[40] Strasser (2011), n 27, 80.

[41] Strasser (2011), n 27, 81; Lewin, n 36, 818.

[42] Strasser (2011), n 27, 81.

[43] Strasser (2011), n 27, 89; Lenoir, n 37, 35; T Smith, "The History of the Genetic Sequence Databases" (1990) 6(4) *Genomics* 701, 704; E Jordan and C Carrico, "DNA Database" (1982) 218(4568) *Science* 108, 108.

[44] Notably at this time there was a keen focus on the potential of commercialising research to benefit the national economy: see, eg R Teitelman, *Gene Dreams: Wall Street, Academia and the Rise of Biotechnology* (Basic, 1989).

fundamental research remain unrestricted".[46] It is notable that this followed soon after the United States Supreme Court's decision in *Diamond v Chakrabarty* that patentable subject matter was very broadly considered and included genetically modified organisms,[47] and the then recently passed *Bayh-Dole Act*[48] and *Stevenson-Wydler Act*[49] that promoted the private sector engagement in public sector research and development. On these measures, the NIH was concerned about the Dayhoff proposal's proprietary arrangements that were "not reassuring" when considering a public database.[50] Another significant factor in favour of the BBN proposal was its access to already established computer networks that would make it more accessible than its rival proposal.[51] The result was to choose a database with a model of open access and to expressly reject a model that favoured limiting access and the formal recognition of proprietary arrangements over data. The GenBank database formally started operations in late 1982.[52]

In contrast to these developments, the European EMBL-EBI accepted open access without hesitation and from its earliest conception.[53] A short time after the Rockefeller University meeting in 1979 initiating the GenBank developments, a meeting under EMBL sponsorship in Schönau Germany titled "EMBL Workshop on Computing and DNA Sequences" was convened.[54] The outcomes of this meeting were very clear – there should be a centralised and computerised sequence database that was freely available.[55] While initially expected to be a co-ordinated effort with those leading the GenBank initiative, the delays and funding uncertainty of GenBank led to the EMBL going ahead with its own bank, the Nucleotide Sequence Data Library, in late 1981/early 1982, some months before GenBank.[56] The EMBL-EBI was established with three primary goals consistent with open access:

- To make freely available a reliable and comprehensive collection of the published nucleic acid sequence data.
- To encourage standardisation and free exchange of data in the international molecular biology community.

[45] Strasser (2011), n 27, 86-87. See also B Strasser, "GenBank – Natural History in the 21st Century" (2008) 322(5901) *Science* 537, 538; A Rai and R Eisenberg, "Bayh-Dole Reform and the Progress of Biomedicine" (2003) 66(1-2) *Law and Contemporary Problems* 289.

[46] President of the United States, *National Policy on the Transfer of Scientific, Technical and Engineering Information*, National Security Decision Directive No 189 (21 September 1985) <http://fas.org/irp/offdocs/nsdd/nsdd-189.htm>. See also National Academy of Sciences, National Academy of Engineering and Institute of Medicine, *Scientific Communication and National Security* (National Academy Press, 1982).

[47] *Diamond v Chakrabarty*, 447 US 303, 309 (1980) deciding that "anything under the sun that is made by man" was patentable, including genetically modified organisms.

[48] *Bayh-Dole Act of 1980*, 35 USC §§ 200-212. See also D Mowery et al, "The Growth of Patenting and Licensing by US Universities: An Assessment of the Effects of the Bayh-Dole Act of 1980" (2001) 30(1) *Research Policy* 99.

[49] *Stevenson-Wydler Act of 1980*, 15 USC §§ 3701-3714. See also J Bagur and A Guissinger, "Technology Transfer Legislation: An Overview" (1987) 12(1) *Journal of Technology Transfer* 51.

[50] Strasser (2011), n 27, 87.

[51] Strasser (2011), n 27, 87-89.

[52] Jordan and Carrico, n 43, 108. See also R Lewin, "National Networks for Molecular Biologists" (1984) 233(4643) *Science* 1379, 1379. See also C Burks et al, "The GenBank Nucleic Acid Sequence Database" (1985) 1(4) *Bioinformatics* 225.

[53] For an overview of the historical developments, see M García-Sancho, *Biology, Computing, and the History of Molecular Sequencing: From Proteins to DNA, 1945-2000* (Palgrave Macmillam, 2012) 92-95. See also G Hamm and K Stüber, "The European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library" (1982) 1 *Nucleotide Sequence Data Library* 2.

[54] See M García-Sancho, "From Metaphor to Practices: The Introduction of Information Engineers into the First DNA Sequence Database" (2011) 33(1) *History and Philosophy of the Life Sciences* 71, 74-76; Smith, n 43, 703. See also Edward A Feigenbaum Papers, *EMBL Workshop on Computing and DNA Sequences*: Stanford University Libraries, Call No SC0340, Accession 2005-101, Box 51, Folder 10, EAF Printed Correspondence May 1980 – March 1981 <https://saltworks.stanford.edu/catalog/druid:wp136kv1744>; Hamm and Stüber, n 53.

[55] Strasser (2011), n 27, 75. See also Editorial, n 36.

[56] Lenoir, n 37, 35; Smith, n 43, 704; Lewin, n 52, 1379; Lewin, n 36, 817.

- To serve as a European focus for efforts devoted toward computing and information services in molecular biology.[57]

The Japanese DDBJ was established in 1986 following discussions among its molecular biology and biophysics community.[58] The DDBJ was intended as a collaboration with EMBL-EBI and GenBank and, as such, the data within the database was made freely available following the model adopted by EMBL-EBI and GenBank.[59] The DDBJ was essentially a follow-on by the Japanese government to make sure its researchers stayed engaged with the global research community, and so adopted the existing open access standards and arrangements in place for EMBL-EBI and GenBank.

Following the original intention of a co-ordinated effort among researchers, both GenBank and EMBL-EBI agreed to share data.[60] Then, in 1988, GenBank, EMBL-EBI and the DDBJ developed a formal collaboration known as the International Nucleotide Sequence Database Collaboration (INSDC).[61] The collaboration agreed to provide free and unrestricted permanent access to all archived data, including raw data, assembly and alignment information, and functional annotated assembled sequences.[62] Each collaborator would therefore collect direct submissions, use a common format for data elements within a unit record, only update the records submitted to the individual collaborator, and distribute copies of all of the submitted sequences to the other collaborators.[63] Hence each collaborator would have a complete copy of all sequences, including those submitted to the other collaborators. The INSDC policy provides:

1. The [INSDC] has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilizing published scientific literature.
2. The [INSDC] will not attach statements to records that restrict access to the data, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.
3. All database records submitted to the [INSDC] will remain permanently accessible as part of the scientific record. Corrections of errors and update of the records by authors are welcome and erroneous records may be removed from the next database release, but all will remain permanently accessible by accession number.
4. Submitters are advised that the information displayed on the Web sites maintained by the [INSDC] is fully disclosed to the public. It is the responsibility of the submitters to ascertain that they have the right to submit the data.
5. Beyond limited editorial control and some internal integrity checks (for example, proper use of [INSDC] formats and translation of coding regions specified in CDS entries are verified), the quality and accuracy of the record are the responsibility of the submitting author, not of the database. The databases will work with submitters and users of the database to achieve the best quality resource possible.[64]

[57] G Hamm and G Cameron, "The EMBL Data Library" (1986) 14(1) *Nucleic Acids Research* 5, 5.

[58] Y Tateno and T Gojobori, "DNA Data Bank of Japan in the Age of Information Biology" (1997) 25(1) *Nucleic Acids Research* 14, 14. See also Deborah Shapley, "Japan Plans DNA Database" (1982) 300(5893) *Nature* 569.

[59] Tateno and Gojobori, n 58, 14. See also S Miyazawa, "DNA Data Bank of Japan: Present Status and Future Plans" (1990) 7 *Computers and DNA* 47.

[60] See Lewin, n 52, 1379. See also R Walgate, "Europe Leads on Sequence" (1982) 296(5858) *Nature* 596.

[61] See G Cochrane et al, "The International Nucleotide Sequence Database Collaboration" (2011) 39(Database issue) *Nucleic Acids Research* D15. See, eg Y Kodama et al, "The Sequence Read Archive: Explosive Growth of Sequencing Data" (2012) 40(Database issue) *Nucleic Acids Research* D54.

[62] See Cochrane et al, n 61, D15.

[63] See C Burks et al, "GenBank" (1992) 20(Supp) *Nucleic Acids Research* 2065.

[64] S Brunak et al, "Nucleotide Sequence Database Policies" (2002) 298(5597) *Science* 1333.

The INSDC asserts that this means: "INSDC databases are data hosts and not data owners."[65] The effect of the INSDC policy, however, was that all the data deposited in GenBank, EMBL-EBI and the DDBJ was immediately and freely available and might be used for any purpose.[66] This remained a contentious issue among researchers with some sequence data not being made available to GenBank, EMBL-EBI and the DDBJ as researchers sought to preserve their options to publish.[67] Most of these sequences were eventually released to GenBank without restrictions (and so also EMBL-EBI and the DDBJ).[68] The solution was about credit and not proprietary dealings with the data, and involved finding a balance between encouraging submissions while protecting confidentiality until the researchers doing the sequencing had an opportunity to gain credit (discussed further below).

A key feature of the open access success of GenBank, EMBL-EBI and DDBJ has been the requirement by journals that sequence data be lodged as a condition of publication.[69] Without this obligation, the task of independently finding and recording sequences would have been too difficult.[70] Notably this was an element of the BBN proposal for GenBank and followed the model set by EMBL-EBI.[71] Over time this has been embraced by journals[72] and granting bodies,[73] so that now it is "mainstream dogma" that journals require an INSDC accession number to publish research dealing with sequences.[74] This was perhaps inevitable because lodging sequences with the databases provided a way for journals to avoid "page upon page of their publications with virtually unreadable sequences".[75] Interestingly, however, this "specialised" form of publishing has been framed as "electronic data publishing" that is "designed both to compliment and to support printed publications".[76] This has also allowed GenBank, EMBL-EBI and DDBJ to design a system that required the sequence data be submitted to them in a convenient form (and this has been regularly updated).[77] The result is now "a highly structured, network-based communication channel through which scientists can present their experimental results … alongside the standard journal publication process without being dependant on journals as a source of data".[78]

The unresolved problem, recalling that this was a problem at the foundation of GenBank,[79] was finding a way for rapid and timely submission without compromising the sequencers' ability to gain credit and attribution for their work doing the sequencing.[80] This has now been resolved by allowing limited confidentiality according to the following GenBank policy (with a similar policy also at

[65] Cochrane et al, n 61, D16.

[66] See L Roberts, "A Tussle Over the Rules for DNA Data Sharing" (2002) 298(5597) *Science* 1312, 1312.

[67] See Roberts, n 66.

[68] Roberts, n 66, 1312.

[69] See T Cech et al, "Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences" (2003) 132(1) *Plant Physiology* 19.

[70] See R Lewin, "DNA Databases are Swamped" (1986) 232(4758) *Science* 1599. See also M Cinkosky et al, "Electronic Data Publishing and GenBank" (1991) 252(5010) *Science* 1273, 1273.

[71] Strasser (2011), n 27, 82-83. See also Strasser (2008), n 45, 538; Walgate, n 60.

[72] See C Burks et al, "GenBank Status Report" (1987) 235(4786) *Science* 267, 267-268.

[73] See J Cassatt and J Peterson, "GenBank Information" (1987) 238(4831) *Science* 1215.

[74] Cochrane et al, n 61, D16.

[75] See Lewin, n 70.

[76] Cinkosky et al, n 70, 1273.

[77] See, eg D Benson et al, "GenBank" (2013) 41(Database issue) *Nucleic Acids Research* D36, D36 (Submission portal).

[78] Cinkosky et al, n 70, 1274.

[79] See Strasser (2011), n 27, 68.

[80] See Roberts, n 66.

EMBL-EBI and DDBJ):[81]

> Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work. GenBank will, upon request, withhold release of new submissions for a specified period of time. However, if a paper citing the sequence or accession number is published prior to the specified date, your sequence will be released upon publication. In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data. As soon as it is available, please send the full publication data – all authors, title, journal, volume, pages and date – to the following address: update@ncbi.nlm.nih.gov.[82]

Another significant limitation on open access has been to implement privacy measures for human sequences.[83] GenBank now applies the policy (with a similar policy also at EMBL-EBI and DDBJ):[84]

> If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. It is our assumption that you have received any necessary informed consent authorizations that your organizations require prior to submitting your sequences.[85]

And most importantly, there is a policy limiting the database managers' responsibilities for assessing the ownership and conditions of use.[86] For example, NCBI, which maintains GenBank, provides the following notice to those submitting data to GenBank:

> 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.[87]

The NCBI also sets out on the GenBank site:

> This site contains resources which incorporate material contributed or licensed by individuals, companies, or organizations that may be protected by US and foreign copyright laws … All persons

[81] See EMBL-EBI, *Terms of Use for EMBL-EBI Services* <http://www.ebi.ac.uk/about/terms-of-use> (providing "All scientific data will be made available by a time and release mechanism consistent with the data type (eg human data where access needs to be reviewed by a Data Access Committee, pre-publication embargoed for a specific time period)"); DDBJ, *Principle of "Hold-Until-Published" Data Release* <http://www.ddbj.nig.ac.jp/sub/hold_date-e.html> (providing "In principle we release 'Hold-Until-Published' data when one of the following three conditions is met: (1) The submitter requests to release the data. (2) The accession number has published and it has been confirmed. (3) A specified hold-date has come").

[82] GenBank, *How to Submit Data to GenBank* (2013) <http://www.ncbi.nlm.nih.gov/genbank/submit>.

[83] See, eg J McEwen et al, "Evolving Approaches to the Ethical Management of Genomic Data" (2013) 29(6) *Trends in Genetics* 375; C Heeney et al, "Assessing the Privacy Risks of Data Sharing in Genomics" (2011) 14 *Public Health Genomics* 17.

[84] See EMBL-EBI, n 81; DDBJ, *Data Submission of Human Subjects Research* <http://www.ddbj.nig.ac.jp/sub/human-e.html> (providing "For all data from human subjects researches submitted to DDBJ, it is submitter's responsibility to ensure that the privacy of participant (human subject) is protected in accordance with all applicable laws, regulations and policies of submitter's institute. In principle, make sure to remove any direct personal identifiers of human subjects from your submissions").

[85] GenBank, n 82.

[86] See NCBI, *GenBank Data Usage* <http://www.ncbi.nlm.nih.gov/genbank>. Noting the journal *Nucleic Acids Research* publishes an annual review of new and updated databases: see, eg M Galperin et al, "The 2015 Nucleic Acids Research Database Issue and Molecular Biology Database Collection" (2015) 43(Database issue) *Nucleic Acids Research* D1.

[87] NCBI, n 86. EMBL-EBI, n 81, similarly provides: "These Terms of Use reflect EMBL-EBI's commitment to OpenScience through its mission to provide freely available online services, databases and software relating to data contributed from life science experiments to the largest possible community. They impose no additional constraints on the use of the contributed data than those provided by the data owner … EMBL-EBI is not liable to you or third parties claiming through you, for any loss or damage … The original data may be subject to rights claimed by third parties, including but not limited to, patent, copyright, other intellectual property rights, biodiversity-related access and benefit-sharing rights. For the specific case of the [European Genome-phenome Archive] database and human data consented for biomedical research, these rights may be formalised in Data Access Agreements. It is the responsibility of users of EMBL-EBI services to ensure that their exploitation of the data does not infringe any of the rights of such third parties." DDBJ similarly provides: "Although DDBJ does not impose any control over the use of any part of the accumulated records, there have not been any copyright transfer from authors of the records upon submission. This is the reason why we, DDBJ, avoid making any definite statement that anybody may freely copy/modify/redistribute any part of the data set". See DDBJ, *Copyright and Limitation of Using DDBJ Data* <http://www.ddbj.nig.ac.jp/copyright-e.html>.

> reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder. Transmission or reproduction of protected items beyond that allowed by fair use as defined in the copyright laws requires the written permission of the copyright owners.[88]

GenBank, EMBL-EBI and DDBJ remain the main databases and they have increased the ranges of data they support.[89] The consequence of the various policies adopted by GenBank, EMBL-EBI and DDBJ has been to effectively limit the early ideal of open access because the data stored and accessible through the database may not be available for use without restrictions; and importantly for users, these restrictions may not be immediately apparent to those accessing the databases. Significantly, the Human Genome Project – probably because of its size and significance as a "big science" project in biology[90] – shaped many of these norms and demonstrated that open access is not necessarily fixed but open to some negotiation.

THE HUMAN GENOME PROJECT

The Human Genome Project sought to list the nucleotide sequence of an entire human genome and was the first "big science" project in biology.[91] The project traces its origins to somewhere around 1985 when planning began.[92] At that stage the entire genomes of the RNA virus bacteriophage MS2,[93] simian virus 40,[94] phage ϕX174,[95] bacteriophage T7,[96] and lambda phage[97] were already publicly available. These whole genome sequence achievements coincided with the then use of computer databases to both store the sequence information and as a data source for rudimentary ways of analysing that sequence data.[98] Since then sequencing technology has improved dramatically[99] so that now the availability of open access sequence data has become the norm with vast open access databases – the main ones being GenBank, EMBL-EBI and DDBJ, described above. The Human Genome Project's main participants envisioned the project as an open access venture – "free of any legal or commercial constraints":[100]

> A key issue for the Human Genome Project is how to promote and encourage the rapid sharing of materials and data that are produced, especially information that has not yet been published or may never be published in its entirety. Such sharing is essential for progress toward the goals of the program

[88] NCBI, *NCBI Website and Data Usage Policies and Disclaimers* <https://www.ncbi.nlm.nih.gov/home/about/policies.shtml>.

[89] See D Benson et al, "GenBank" (2011) 39(Database issue) *Nucleic Acids Research* D32; E Kaminuma et al, "DDBJ Progress Report" (2011) 39(Database issue) *Nucleic Acids Research* D22; Cochrane et al, n 61; R Leinonen et al, "The European Nucleotide Archive" (2011) 39(Database issue) *Nucleic Acids Research* D28.

[90] See F Collins et al, "The Human Genome Project: Lessons from Large-Scale Biology" (2003) 300(5617) *Science* 286.

[91] See Collins et al, n 90.

[92] See R Sinsheimer, "The Santa Cruz Workshop – May 1985" (1989) 5(4) *Genomics* 954.

[93] See W Fiers et al, "Complete Nucleotide Sequence of Bacteriophage MS2 RNA: Primary and Secondary Structure of the Replicase Gene" (1976) 260(5551) *Nature* 500.

[94] See V Reddy et al, "The Genome of Simian Virus 40" (1978) 200(4341) *Science* 494; W Fiers et al, "Complete Nucleotide Sequence of SV40 DNA" (1978) 273(5658) *Nature* 113.

[95] See F Sanger et al, "The Nucleotide Sequence of Bacteriophage ϕX174" (1978) 125(2) *Journal of Molecular Biology* 225.

[96] See J Dunn et al, "Complete Nucleotide Sequence of Bacteriophage T7 DNA and the Locations of T7 Genetic Elements" (1983) 166(4) *Journal of Molecular Biology* 477.

[97] See F Sanger et al, "Nucleotide Sequence of Bacteriophage λ DNA" (1982) 162(4) *Journal of Molecular Biology* 729.

[98] See R Roberts, "The Early Days of Bioinformatics Publishing" (2000) 16(1) *Bioinformatics* 2, 2 and the references therein; Smith, n 43, 702; Lenoir, n 37, 27-45.

[99] See, eg M Gužvi, "The History of DNA Sequencing" (2013) 32(4) *Journal of Medical Biochemistry* 301. See also M Morey et al, "A Glimpse into Past, Present, and Future DNA Sequencing" (2013) 110(1) *Molecular Genetics and Metabolism* 3.

[100] See D Dickson, "Consortium Plans 'Public' Map of Genome" (1994) 371(6498) *Nature* 551.

> and to avoid unnecessary duplication. It is also desirable to make the fruits of genome research available to the scientific community as a whole as soon as possible to expedite research in other areas.[101]

The only restriction envisioned at the time was that genomic data must be released within six months of being generated, with the delay allowing the sequence producers adequate time to publish about their results.[102]

The public and private ownership of sequences in the early stages of the Human Genome Project was, however, controversial because of patent claims to short sequences,[103] and the announcement by the private sector company, Human Genome Sciences Inc, of terms for accessing its proprietary sequence database from its independent sequencing project.[104] The terms of access required providing Human Genome Sciences Inc with "the opportunity to retain control over the commercial application of knowledge derived from the sequences, and in particular the discovery of any genes arising directly from this knowledge".[105] This included options on any patents arising, pre-publication review to determine intellectual property issues, an opportunity to make patent applications, and a first right to negotiate over commercialisation.[106] Other competing public and private entities at the time objected to these stringent conditions and the potential privatisation of sequences.[107]

A more pressing problem, at the time, was recognition that the accepted delay of six months before publicly disclosing sequences was too long.[108] The Human Genome Project resolved these concerns by expressly requiring that sequences be disclosed within 24 hours of being generated without restrictions on publications.[109] The available data was publicly released through GenBank (and shared with EMBL-EBI and DDBJ).[110] This implemented the so-called Bermuda Principles in 1996:[111]

> **Primary Genomic Sequence Should be in the Public Domain**
>
> It was agreed that all human genomic sequence information, generated by centres funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society.

[101] Department of Energy, "NIH, DOE Encourage Sharing of DNA, Resources" (1993) 4(5) *Human Genome News* 4.

[102] Department of Energy, n 101. This had been one of the original concerns that had delayed the setting up of GenBank: see Lewin, n 36, 817.

[103] See L Roberts, "Genome Patent Fight Erupts" (1991) 254(5029) *Science* 184; B Healy, "Special Report on Gene Patenting" (1992) 327 *New England Journal of Medicine* 664. See also T Relchhardt, "Patent on Gene Fragment Sends Researchers a Mixed Message" (1998) 396(6711) *Nature* 499; C Anderson, "NIH Drops Bid for Gene Patents" (1994) 263(5149) *Science* 909; L Roberts, "NIH Gene Patents, Round Two" (1994) 255(5047) *Science* 255.

[104] See D Dickson, "HGS Seeks Exclusive Option on All Patents Using its cDNA Sequences" (1994) 371(6497) *Nature* 463. See also E Dorey et al, "TIGR Releases EST Publicly" (1997) 15(5) *Nature Biotechnology* 397; E Marshall, "A Showdown Over Gene Fragments" (1994) 266 (5183) *Science* 208.

[105] Dickson, n 104.

[106] Dickson, n 104.

[107] See D Dickson, "'Gene Map' Plan Highlights Dispute over Public vs Private Interests" (1994) 371(6496) *Nature* 366; D Dickson, "Merck to Back 'Public' Sequencing" (1994) 371(6496) *Nature* 366. See also E Marshall, "Ethics in Science: Is Data-Hoarding Slowing the Assault on Pathogens?" (1997) 275(5301) *Science* 777. See also D Bentley, "Genomic Sequence Information Should Be Released Immediately and Freely in the Public Domain" (1996) 274(5287) *Science* 533; M Adams and C Venter, "Should Non-Peer-Reviewed Raw DNA Sequence Data Release Be Forced on the Scientific Community?" (1996) 274(5287) *Science* 534.

[108] See J Contreras, "Prepublication Data Release, Latency, and Genome Commons" (2010) 329(5990) *Science* 393, 393. See also Marshall, n 107.

[109] See E Marshall, "The Human Gene Hunt Scales Up" (1996) 274(5292) *Science* 1456; E Marshall, "Genome Researchers Take the Pledge" (1996) 272(5261) *Science* 477, 477.

[110] Bentley, n 107, 534.

[111] See E Marshall, "Bermuda Rules: Community Spirit, with Teeth" (2001) 291(5507) *Science* 1192.

Primary Genomic Sequence Should be Rapidly Released

- Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 Kb would be released automatically on a daily basis.
- Finished annotated sequence should be submitted immediately to the public databases.

It was agreed that these principles should apply for all human genomic sequence generated by large-scale sequencing centres, funded for the public good, in order to prevent such centres establishing a privileged position in the exploitation and control of human sequence information.[112]

These principles were re-endorsed a short time later in 1997,[113] and then updated in 2000 to deal with the concern that those producing the sequence should have an opportunity to publish about their sequence before others.[114] Then, in 2003, the Fort Lauderdale summit announced an accord that agreed to no restrictions on using data, with the request that users "act responsibly to promote the highest standards of respect for the scientific contribution of others".[115] This summit also recommended that these pre-publication release principles apply to other large-scale projects, presumably covering the range of large-scale sequencing projects addressing whole genomes, genome-wide associations studies and so on, and that funding agencies provide ongoing funding for database projects.[116] Similar principles were extended to proteomic data in 2008[117] and other biological data sets in 2009.[118] The issue of how long data releases should be postponed to allow publication remains contested, with a range of databases trying various latency or embargo periods from as soon as possible to 12 months.[119]

An early challenge to the Bermuda Principles and open access through the Human Genome Project was the decision by Celera Genomics Corporation, and its chief Craig Venter, to independently sequence the human genome.[120] A key issue was whether the sequence data generated would be publicly available:

> Venter acknowledges that he hears this question a lot. Business people ask where Celera's profits will come from, while dubious academics ask whether the business agenda is compatible with collegial sharing of data. Venter – never one to mince words – responds that the questioners just "don't get it". Celera must succeed in two worlds, he said in a recent interview: "The scientific community thinks this is just a business project, and the business community thinks it's just a science project. The reality is, it's both". The "business model only works if [we do] absolutely world-class science", Venter explains, "and the science model only works if it's world-class business". In his view, he is implementing a "radical change" in biology, an approach that enjoys "the best of both worlds" – private funding and

[112] Human Genome Organisation, *Summary of Principles Agreed Upon at the First International Strategy Meeting on Human Genome Sequencing* (Bermuda, 25-28 February 1996) <http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml>.

[113] See Human Genome Organisation, *Summary of the Report of the Second International Strategy Meeting on Human Genome Sequencing* (Bermuda, 27 February-2 March 1997) <http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml>. Notably scientists attending a 1998 meeting extended their individual commitment when conducting other large-scale sequencing projects, although this commitment did not extend to funding bodies: see M Guyer, "Statement on the Rapid Release of Genomic DNA Sequence" (1998) 8(5) *Genome Research* 413.

[114] *NHGRI Policy for Release and Database Deposition of Sequence Data* <http://www.genome.gov/page.cfm?pageID=10000910>. See, eg R Hyman, "Sequence Data: Posted vs Published" (2001) 291(5505) *Science* 827; E Bell, "Publication Rights for Sequence Data Producers" (2000) 290(5497) *Science* 1696; L Rowen et al, "Publication Rights in the Era of Open Data Release Policies" (2000) 289(5486) *Science* 1881.

[115] Wellcome Trust, *Report of Meeting organized by the Wellcome Trust, Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* (14-15 January 2003) 4 <http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf>. See also C Dennis, "Draft Guidelines Ease Restrictions on Use of Genome Sequence Data" (2003) 421(6926) *Nature* 877; L Roberts, "New Policy Reaffirms Pledge to Share Genome Data" (2003) 299(5611) *Science* 1293.

[116] Wellcome Trust, n 115, 3-4.

[117] See H Rodriguez et al, "Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: The Amsterdam Principles" (2009) 8(7) *Journal of Proteome Research* 3689.

[118] See Toronto International Data Release Workshop Authors, "Prepublication Data Sharing" (2009) 461(7261) *Nature* 168.

[119] See Contreras, n 108, 393-394 and the references therein.

[120] See E Marshall and E Pennisi, "Hubris and the Human Genome" (1998) 280(5366) *Science* 994.

> academic freedom. As a result, he says, he will be working more openly than most companies or academic labs, for both the science and the finances will be open to scrutiny: "Everything will be out in the open". This, Venter insists, "is the opposite of secret".[121]

The Celera Genomics Corporation business model was to charge a fee for early access to its sequence data and then "to patent several hundred human genes and a large set of human single nucleotide polymorphisms for use in individually tailored medicine".[122] The remaining issue was how the sequence data was to be made publicly available, and whether this included depositing the data in the publicly available databases of GenBank, EMBL-EBI and the DDBJ.[123] Initially Celera Genomics Corporation proposed releasing the data on its own website.[124] To get its research published, however, it entered into negotiations with the journal *Science* and eventually agreed that its data would be available from its own website for free only if it was agreed not to redistribute the data or commercialise the data, and any commercialisation required another negotiated agreement.[125] Years later when Celera Genomics Corporation's business, under the new owner Applera Corporation, moved into drug discovery, the sequence information was of less value to the business and all the data was provided to GenBank.[126] The significance of this series of events involving Celera Genomics Corporation was to confirm that the Bermuda Principles and open access are not absolute, and that journals against their stated policies can and do make concessions to non-open access sequence databases to maintain confidentiality. Notably this was repeated with the rice genome sequence being published separately on the Syngenta Corporation website.[127] These limitations, however, appear as exceptions to a generality of favouring open access.

The effect of the Human Genome Project was the modified Bermuda Principles that essentially required pre-publication as soon as possible and "respect" for the sequence producers in using the database materials. Despite this broad agreement, at the time there remained concerns that sequence producers did not have sufficient time to publish about their sequences.[128] As set out above, this has been addressed through allowing some delay to enable publication. Further, there has been recognition of the need for privacy for some human sequences.

OTHER RECENT DEVELOPMENTS

In addition to the GenBank, EMBL-EBI and DDBJ databases, there has been a proliferation of alternative and independent specialised sequence databases, now numbering in the hundreds,[129] directed to particular model organisms, gene groupings, diseases and so on.[130] With the GenBank, EMBL-EBI and DDBJ databases absolving themselves of responsibility for assessing the ownership, conditions of use and privacy,[131] these other databases and funding and research organisations have started to develop specific policies and regulations.[132] The most significant recent development has been the need to respect personal privacy and confidentiality. The concern is that sequence data in

[121] E Marshall, "A High-Stakes Gamble on Genome Sequencing" (1999) 284(5422) *Science* 1906, 1906.

[122] Marshall, n 121, 1909.

[123] For a similar position asserted by Syngenta Corporation in dealing with the rice genome, see P Moore, "Publication with a Pinch of Privatization" (2002) 3(4) *Genome Biology*, DOI: 10.1186/gb-spotlight-20020404-02.

[124] See Marshall, n 121, 1909.

[125] See E Marshall, "Celera and Science Spell Out Data Access Provisions" (2001) 291(5507) *Science* 1191. See also E Marshall, "Sharing the Glory, Not the Credit" (2001) 291(5507) *Science* 1189.

[126] J Kaiser, "Celera to End Subscriptions and Give Data to Public GenBank" (2005) 308(5723) *Science* 775.

[127] See Moore, n 123.

[128] See E Pennisi, "Group Calls for Rapid Release of More Genomic Data" (2009) 324(5930) *Science* 1000, 1001.

[129] See <http://www.oxfordjournals.org/our_journals/nar/database/a>.

[130] The journal *Nucleic Acids Research* publishes an annual review of new and updated databases: see Galperin et al, n 86.

[131] See NCBI, n 86.

[132] See D Resnik, "Genomic Research Data: Open vs Restricted Access" (2010) 32(1) *IRB: Ethics & Human Research* 1.

some forms (and particularly whole genome sequences) can be used to identify individuals.[133] GenBank addresses this concern superficially: "If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source."[134] Some of these other databases have provided a more sophisticated approach with restrictions on using the databases.[135] For example, the database of Genotypes and Phenotypes (dbGaP)[136] is subject to the 2007 NIH *Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies* (GWAS Policy).[137] dbGaP was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype, including sequence data.[138] The NIH GWAS Policy established a framework for ensuring data submitted to dbGaP was de-identified and that sharing genome-wide association studies data from dbGaP relied on a two-tiered system of unrestricted (summary-level information and aggregate genotype data) or controlled access (individual-level genotypes and phenotypes); data sets are only released according to the level of consent that has been provided and there is ongoing oversight of the uses of the accessed data.[139] More recently in 2014 the *National Institutes of Health Genomic Data Sharing Policy* distinguished between human and non-human data.[140] As a generalisation, the expectation is that non-human sequence data will be made "publicly available" no later than the date of initial publication through "any widely used data repository",[141] and de-identified human sequence data must be deposited with "an NIH-designated data repository" and will be "released" on the first of either six months after submission or publication.[142] A two-tiered system of release, however, is applied to human sequence data distinguishing between unrestricted and controlled data access mechanisms:

> Respect for, and protection of the interests of, research participants are fundamental to NIH's stewardship of human genomic data. The informed consent under which the data or samples were collected is the basis for the submitting institution to determine the appropriateness of data submission to NIH-designated data repositories, and whether the data should be available through unrestricted or controlled access. Controlled-access data in NIH-designated data repositories are made available for secondary research only after investigators have obtained approval from NIH to use the requested data for a particular project. Data in unrestricted-access repositories are publicly available to anyone.[143]

The controlled data are submitted according to the terms of an "Institutional Certification" that addresses whether the data collection complies with laws and regulation (including cultural norms and institutional policies), informed consent, and that the data was appropriately collected.[144] Where access is sought to controlled data the request is considered by an NIH Data Access Committee to confirm compliance with the terms of the "Institutional Certification" (particularly the terms of

[133] See N Homer et al, "Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-density SNP Genotyping Microarrays" (2008) 4(8) *PLoS Genetics* e1000167. See also J Couzin, "Whole-Genome Data Not Anonymous, Challenging Assumptions" (2008) 321(5894) *Science* 1278; E Zerhouni and E Nabel, "Protecting Aggregate Genomic Data" (2008) 322(5898) *Science* 44; Z Lin et al, "Genomic Research and Human Subject Privacy" (2004) 305(5681) *Science* 183.

[134] NCBI, *GenBank Overview: Privacy* <http://www.ncbi.nlm.nih.gov/genbank>.

[135] See Resnik, n 132.

[136] See NCBI, *dbGaP* <http://www.ncbi.nlm.nih.gov/gap>.

[137] See D Paltoo et al, "Data Use under the NIH GWAS Data Sharing Policy and Future Directions" (2014) 46(9) *Nature Genetics* 934.

[138] See NCBI, n 136. See also K Tryka et al, "NCBI's Database of Genotypes and Phenotypes: dbGaP" (2014) 42(Database issue) *Nucleic Acids Research* D1, D975-D979.

[139] For an overview see Paltoo et al, n 137, 934-936; McEwen et al, n 83, 377.

[140] National Institute of Health, *National Institutes of Health Genomic Data Sharing Policy* (2014) [IV(B), (C)] <http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf>.

[141] National Institute of Health, n 140, [IV(B)(1), (2)].

[142] National Institute of Health, n 140, [IV(C)(1)].

[143] National Institute of Health, n 140, [IV(C)(3)].

[144] National Institute of Health, n 140, [IV(C)(5)].

consent) and then used according to a "Data Use Certification".[145] The "Data Use Certification" imposes obligations on the users to: protect data confidentiality; comply with laws and regulation (including cultural norms and institutional policies); not identify individual participants from their data; not sell the data; not share the data; and provide reports about data use.[146] And while "the NIH discourages the use of patents to prevent the use of or to block access to genomic or genotype-phenotype data developed with NIH support", it does "encourage[] patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs without impeding research".[147]

The other major development has been the Organisation for Economic Co-operation and Development (OECD) Committee on Scientific and Technology Policy meeting at ministerial level adopting the *Declaration on Access to Research Data from Public Funding* in 2004.[148] As required by the Declaration, the OECD administration developed and released in 2007 the *OECD Principles and Guidelines for Access to Research Data from Public Funding* with the objective, in part, of "[p]romot[ing] a culture of openness and sharing of research data among the public research communities within member countries and beyond" and "provid[ing] a commonly agreed upon framework of operational principles for the establishment of research data access arrangements in member countries".[149] The various OECD members have implemented these principles and guidelines to varying degrees and generally require research data and publications to be made publicly available.

There has also been a massive expansion in the numbers of databases for specific collections of DNA, RNA and amino acid sequences with some specific rules developing for some disciplines. For example, the *Principles for Proteomic Data Release and Sharing* (also known colloquially as the "Amsterdam Principles 2008") were developed among members of the international proteomics community to address the lack of policies dealing with the rapid release of large-scale proteomic data into the public domain.[150] The outcome was a number of principles:

> 1. Timing. The timing with which proteomic data is released into the public domain should depend on the nature of the effort generating the data and should take into account the legitimate concerns of data producers … that data generated by individual investigators should be released into the public domain at the latest upon publication while data generated by community resource projects should be released upon generation following appropriate [quality assurance and control] procedures.
> 2. Comprehensiveness. For data to be valuable to the proteomics community and other interested scientists, they must be released in a format that, as comprehensively as possible, captures the results of an experiment and the conditions under which the experiment was run …
> 3. Format. Open access to proteomic data requires community-supported standardized formats, controlled vocabularies, reasonable reporting requirements, and publicly available central repositories.
> 4. Deposition to repositories. Central repositories should make it attractive for depositors to use them …
> 5. Quality metrics. Central repositories should develop threshold metrics for assessing data quality …
> 6. Responsibility. Scientists, funding agencies, and journals share a joint responsibility for ensuring that all parties adhere to community standards for data release.[151]

[145] National Institute of Health, n 140, [V(A), (B)].

[146] National Institute of Health, n 140, [V(B)].

[147] National Institute of Health, n 140, [VI].

[148] Organisation for Economic Co-operation and Development, *Science, Technology and Innovation for the 21st Century: Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004, Final Communiqué* (2004) [17] <http://www.oecd.org/science/sci-tech/sciencetechnologyandinnovationforthe21stcenturymeetingoftheoecdcommitteeforscientificandtechnologicalpolicyatministeriallevel29-30january2004-finalcommunique.htm>.

[149] Organisation for Economic Co-operation and Development, *OECD Principles and Guidelines for Access to Research Data from Public Funding* (2007) 11. See also D Pilat and Y Fukasaku, "OECD Principles and Guidelines for Access to Research Data from Public Funding" (2007) 6 *Data Science Journal* OD4.

[150] Rodriguez et al, n 117, 3689-3690.

[151] Rodriguez et al, n 117, 3691-3692.

More recently, the Toronto 2009 Data Release Workshop developed some "best practices", including a statement, providing in part:

1. Rapid pre-publication data release should be encouraged for projects with the following attributes: large-scale …; broad utility; creating reference datasets; associated with community buy-in …
2. Funding agencies should facilitate the specification of data release policies for relevant projects …
3. Data producers should state their intentions and enable analyses of their data …
4. Data analysts/users should freely analyze released pre-publication data and act responsibly in publishing analyses of those data.
5. Scientific journal editors should engage the research community about issues related to pre-publication data release and provide guidance to authors and reviewers on the third-party use of pre-publication data in manuscripts.[152]

More concrete advice was provided in the form of a table (see Table 1). These developments illustrate that there is an imperative for open access, while recognising that limitations may be justified in some circumstances, such as where credit and attribution are sought or privacy protections are necessary.

TABLE 1 Examples of Pre-publication Data Release Guidelines for Different Project Types

Project Type	Pre-publication Data Release Recommended	Pre-publication Data Release Optional
Genome sequencing	Whole-genome or mRNA sequence(s) of a reference organism or tissue	Sequences from a few loci for cross-species comparisons in a limited number of samples
Polymorphism discovery	Catalogue of variants from genomic and/or transcriptomic samples in one or more populations	Variants in a gene, a gene family, or a genomic region in selected pedigrees or populations
Genetic association studies	Genome-wide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or whole-genomes of tumor and non-tumor samples	Somatic mutations of a specific locus or limited set of genomic regions
Microbiome studies	Whole-genome sequence of microbial communities in different environments	Sequencing of target locus in a limited number of microbiome samples
RNA profiling	Whole-genome expression profiles from a large panel of reference samples	Whole-genome expression profiles of a perturbed biological system(s)
Proteomic studies	Mass spectrometry datasets from large panels of normal and disease tissues	Mass spectrometry datasets from a well-defined and limited set of tissues
Metabolomic studies	Catalogue of metabolites in one or more tissues of an organism	Analyses of metabolites induced of a perturbed biological system(s)
RNAi or chemical library screen	Genome-wide screen of a cell line or organism analyzed for standard phenotypes	Focused screens used to validate a hypothetical gene network
3D structure elucidation	Large-scale cataloguing of 3D structures of proteins or compounds	3D structure of a synthetic protein or compound elucidated in the context of a focused project

Source: Toronto International Data Release Workshop Authors, "Prepublication Data Sharing" (2009) 461(7261) *Nature* 168, 168.

[152] Toronto International Data Release Workshop Authors, n 118, 169.

DATA AND INFORMATION OBLIGATIONS UNDER THE CBD AND NAGOYA PROTOCOL, PLANT TREATY AND PIP FRAMEWORK

The CBD provides, independent of ABS contracting, a general obligation for the exchange of information about the "results of technical, scientific and socio-economic research", "training and surveying programmes", "specialized knowledge", "indigenous and traditional knowledge as such and in combination with the technologies ['that are relevant to the conservation and sustainable use of biological diversity or make use of genetic resources']",[153] and "where feasible, include repatriation of information".[154] This is assumed to include data and information in the form of DNA, RNA and amino acid sequences.[155] There is also a Clearing House Mechanism "to promote and facilitate technical and scientific cooperation".[156] The Clearing House Mechanism is considered to be essential to implementing the CBD[157] and is currently being realised through a decentralised collection of information hubs (databases and websites) and national government websites with very little formal regulation.[158] The separate and centrally located Nagoya Protocol Access and Benefit-sharing Clearing-House is also being developed as a part of the CBD's Clearing House Mechanism.[159] In co-operation with the CBD's Clearing House Mechanism,[160] the Plant Treaty and its SMTA impose various information obligations (including potentially information about traditional knowledge)[161] and provide for a "Global Information System" (GLIS) to allow the data and information about plant materials to be collected, made available and shared.[162] Like the CBD's Clearing House Mechanism, the GLIS is conceived as a decentralised network of databases and websites with some non-binding standards.[163] These forms of data and information are considered to be an essential part of benefit sharing under the Plant Treaty.[164]

[153] CBD, n 7, Art 16(1).

[154] CBD, n 7, Art 17(2).

[155] See Ad Hoc Open-Ended Working Group on Access and Benefit-Sharing, *The Concept of "Genetic Resources" in the Convention on Biological Diversity and How it Relates to a Functional International Regime on Access and Benefit-Sharing*, UNEP/CBD/WG-ABS/9/INF/1 (2010) 36.

[156] CBD, n 7, Art 18(3). See <http://www.chm-cbd.net>.

[157] See UNEP/CBD/COP/10/27, n 11, [217] and Annex (Decision X/15, 163-165).

[158] See Ad Hoc Open-Ended Working Group on Review of Implementation of the Convention, *Progress Report on the Clearing House Mechanism*, UNEP/CBD/WGRI/5/3/Add 2 (2014); Conference of the Parties to the Convention on Biological Diversity, *Progress Report on Technical and Scientific Cooperation and the Clearing House Mechanism*, UNEP/CBD/COP/12/11 (2014). The main direction for the Clearing House Mechanism is in the implementation of the *Strategic Plan for Biodiversity 2011-2020* and the achievement of the *Aichi Biodiversity Targets*: see Conference of the Parties to the Convention on Biological Diversity, *Report of the Twelfth Meeting of the Conference of the Parties to the Convention on Biological Diversity*, UNEP/CBD/COP/12/29 (2014) [149] and Decision XII/2 (12-18).

[159] UNEP/CBD/COP/10/27, n 11, [103] and Annex (Art 14(1); Decision X/1, 85-109). See also Open-Ended Ad Hoc Intergovernmental Committee for the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization, *Report of the Third Meeting of the Open-Ended Ad Hoc Intergovernmental Committee for the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization*, UNEP/CBD/COP/12/6 (2014) [51]-[58]; Open-Ended Ad Hoc Intergovernmental Committee for the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization, *Report on Progress in the Implementation of the Pilot Phase of the Access and Benefit-Sharing Clearing-House*, UNEP/CBD/ICNP/3/6 (2014).

[160] See Plant Treaty, n 12, Art 17(1).

[161] Plant Treaty, n 12, Arts 12(4), 17(1); IT/GB-1/06/Report, n 25, [12] (Resolution 2/2006) and Appendix H.

[162] See Plant Treaty, n 12, Art 17(1). See also C Lawson, "Information Intellectual Property and the Global Information System for Plant Genetic Resources for Food and Agriculture" (2015) 26 *Australian Intellectual Property Journal* 27.

[163] See Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture, *Report of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture*, IT/GB-6/15/7 (2015) and the references therein. See also Expert Consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture, *First Meeting of the Expert Consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture*, IT/COGIS-1/15/Report (2015).

[164] Plant Treaty, n 12, Art 17(1).

Under the CBD's Clearing House Mechanism (including the Nagoya Protocol's Benefit-sharing Clearing-House) and the Plant Treaty's GLIS, there are, as yet, no specific arrangements for dealing with DNA, RNA and amino acid sequences. It is clear, however, that realising the CBD and Plant Treaty obligations is through a network of existing resources.[165] These existing resources, and a number of evolving frameworks dealing with information about genetic resources including sequence data and information, include the Global Biodiversity Informatics Outlook,[166] Digital Seed Bank,[167] European Network of Gene Banks (Eurisco),[168] Gateway to Genetic Resources (GenSys),[169] World Information Sharing Mechanism for the implementation of the Global Plan of Action (WISM-GPA),[170] Germplasm Resource Information Network (GRIN-Global)[171] and Diversity Seek (DivSeek).[172] The challenge for both the CBD's Clearing House Mechanism (including the Nagoya Protocol's Benefit-sharing Clearing-House) and the Plant Treaty's GLIS will be developing coherent listings of what is maintained in the databases and websites, and permanent unique identifiers to be able to distinguish what has been stored.[173] Under both the CBD's Clearing House Mechanism (and the Nagoya Protocol's Benefit-sharing Clearing-House) and Plant Treaty's GLIS, the problem of open access to DNA, RNA and amino acid sequences and ABS has not been addressed,[174] although this may be expected soon.

Meanwhile the PIP Framework provides for sharing samples of "influenza viruses with human pandemic potential",[175] including human clinical specimens, influenza virus isolates, extracted RNA, cDNA, and influenza candidate vaccine viruses.[176] These materials are shared with a WHO co-ordinated network of laboratories,[177] with an obligation to share "[g]enetic sequence data and analyses arising from that data",[178] and recognising "that greater transparency and access concerning influenza virus genetic sequence data is important to public health" and that "there is a movement towards the use of public domain or public access databases such as GenBank and [the Global Initiative on Sharing Avian Influenza Data (GISAID) databases]".[179] The WHO co-ordinated network of laboratories involved in sharing the viruses are engaged with terms of reference[180] that require that each laboratory "submit genetic sequence data to GISAID and GenBank or similar database in a timely manner consistent with the [SMTA]".[181] GISAID is a consortium of existing networks within

[165] See CBD: UNEP/CBD/COP/12/11, n 158, [7]; IT/GB-6/15/7, n 163, Annex 1.

[166] See Global Biodiversity Informatics Outlook <http://www.biodiversityinformatics.org>.

[167] See Digital Seed Bank <http://www.globalplantcouncil.org>.

[168] See European Network of Gene Banks (Eurisco) <http://www.ecpgr.cgiar.org>.

[169] See Gateway to Genetic Resources (GenSys) <http://www.genesys-pgr.org>.

[170] See World Information Sharing Mechanism for the implementation of the Global Plan of Action (WISM-GPA) <http://www.fao.org/pgrfa-gpa-archive/selectcountry.jspx>.

[171] See Germplasm Resource Information Network (GRIN-Global) <http://www.ars-grin.gov>.

[172] See Diversity Seek (DivSeek) <http://www.divseek.org>.

[173] See, eg IT/GB-6/15/7, n 163, [18].

[174] See Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture, *Report of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture*, IT/GB-6/15/Report (2015) [31] and Appendix A.3; Meeting of the Informal Advisory Committee to the Clearing-House Mechanism of the Convention on Biological Diversity, *Information Services of the Central Clearing-House Mechanism*, UNEP/CBD/CHM/IAC/2010/1/3 (2010).

[175] PIP Framework, n 13, Art 4.2.

[176] PIP Framework, n 13, Arts 2.1(i), 4.1.

[177] PIP Framework, n 13, Art 5.1.1.

[178] PIP Framework, n 13, Art 5.2.1. See also World Health Assembly, n 20, Annex ([4]-[5]).

[179] PIP Framework, n 13, Art 5.2.2.

[180] See PIP Framework, n 13, Annex 4.

[181] PIP Framework, n 13, Annex 4, [9].

the international scientific community that has agreed to share their sequence data deposited in GenBank, EMBL and the DDBJ as soon as possible after analysis and validation, and with a maximum delay of six months.[182]

Like the CBD and Plant Treaty, sharing data and information under the PIP Framework is considered to be a part of the PIP Benefit Sharing System through making sequences available as part of pandemic surveillance and risk assessment.[183] At the time the PIP Framework was negotiated, it was appreciated that sequence data might be used independently of the physical sample to synthesise candidate vaccine viruses, virus proteins and antibodies.[184] While under the PIP Framework there already exists a means for tracking physical samples through the Influenza Virus Traceability Mechanism[185] – the challenge has been to find processes that track uses of the sequences where physical samples have not been provided. This has started with the PIP Framework expressly providing for the Director-General to consult the Pandemic Influenza Preparedness (PIP) Framework Advisory Group,[186] and they in turn have convened the Technical Expert Working Group (TEWG) on Genetic Sequence Data.[187] The outcome of this technical group so far has been to identify some key elements in any process:

- Industry stated its concern that placing restrictions on the access/use of genetic sequence data would delay development of pandemic products.
- Industry indicated that assuring the sharing of benefits associated with the use of genetic sequence data might be better approached through the monitoring of products derived from the use of genetic sequence data.
- Industry and other stakeholders raised the issue of potential biosecurity/biosafety risks related to use of genetic sequence data.[188]

Most recently the Advisory Group has started to consider the optimal characteristics of a sequence-sharing system under the PIP Framework[189] and recommended that future work specifically address how sequence data is used under the PIP Framework consistent with its identified key elements.[190] At this stage, there has been some consideration of how sequence data might be addressed, including:

- The objective of benefit-sharing may be met by monitoring use of [genetic sequence data] and/or tracing [genetic sequence data] or by other mechanisms related to influenza-related products.
- While monitoring and tracing the use of [genetic sequence data] is limited by the medium used to share it, technical mechanisms to trace or monitor downloading of [genetic sequence data] from databases may be implemented.
- [genetic sequence data] of PIP biological material can also be generated by non-[WHO] laboratories. In that case, WHO will likely not know of this, and the sharing of such will be more difficult to monitor.

[182] P Bogner et al, "A Global Initiative on Sharing Avian Flu Data" (2006) 442(7106) *Nature* 981.

[183] PIP Framework, n 13, Art 6.1.2(i). See also World Health Assembly, n 20, Annex ([13(b)]); Pandemic Influenza Preparedness (PIP) Framework Advisory Group Technical Expert Working Group (TEWG) on Genetic Sequence Data, *Final Report to the PIP Advisory Group (Revised)* (2014) 2-3 <http://www.who.int/influenza/pip/advisory_group/PIP_AG_Rev_Final_TEWG_Report_10_Oct_2014.pdf>.

[184] PIP Framework, n 13, Arts 5.2.1-5.2.4.

[185] See World Health Assembly, n 20, Annex 2 ([2.2]).

[186] PIP Framework, n 13, Art 5.2.4.

[187] See World Health Assembly, n 20, Annex 1 ([6]) and Annex ([2]-[3]).

[188] World Health Assembly, n 20, Annex ([13(b)]).

[189] See PIP Framework Advisory Group, *Briefing Note for the PIP Advisory Group Special Session* (2015) [10]-[11] <http://www.who.int/influenza/pip/advisory_group/Briefing_NoteAGSS.pdf?ua=1>.

[190] See PIP Framework Advisory Group, *Report to the Director-General* (2015) [8(a)] and [18(a)] <http://www.who.int/influenza/pip/advisory_group/ag_spec_session_report.pdf?ua=1>.

- Notwithstanding, there are other potential mechanisms that could be developed to monitor the use of [genetic sequence data], such as processes related to influenza-related products (eg regulatory approval files and patent applications).[191]

In advancing these considerations, the Advisory Group considered that "[t]he objective of benefit-sharing may be met by mechanisms related to monitoring products generated using influenza [genetic sequence data], rather than by monitoring use of [genetic sequence data] and/or tracing [genetic sequence data], noting that source identification is critical".[192] The work continues with the development of a search engine to monitor the use of genetic sequence data in end products, a review of existing data-sharing systems and the development of an options paper on monitoring the use of sequence data in end products.[193] As this analysis shows, the favoured method appears, at this stage at least, to be monitoring and tracing the use of sequence data in end products.[194]

DISCUSSION

This article has traced the evolution of open access to DNA, RNA and amino acid sequences and showed that the open access project has been incredibly successful in enabling broad access to valuable DNA, RNA and amino acid sequence data and information.[195] So far the key concerns have been the professional dilemmas about sharing data[196] and respecting the privacy of human subjects contributing the data.[197] Despite the apparent success of open access, however, the article has also shown that the ideals of open access have been negotiable with Celera Genomics Corporation restricting their Human Genome Project sequence data for commercial reasons and still gaining all the benefits of publishing their research in reputable international journals.[198] Most significantly, however, this article has shown that the use of GenBank, EMBL-EBI and DDBJ is not free of charge and without restrictions (gratis),[199] but rather is subject to some limits on how the accessed data and information might be used (libre). So, for example, NCBI (the host of GenBank) states that it "cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the [data

[191] PIP Framework Advisory Group TEWG, n 183, 12.

[192] PIP Framework Advisory Group, *Report to the Director General (Corrected)* (2014) [31(c)] <http://www.who.int/influenza/pip/combined_pipagmroct2014corr.pdf?ua=1>.

[193] See PIP Framework Advisory Group, n 189, [10]-[11].

[194] See PIP Framework Advisory Group Technical Expert Working Group (TEWG) on Genetic Sequence Data, *Draft Optimal Characteristics of an Influenza Genetic Sequence Data Sharing System under the PIP Framework* (2015) <http://www.who.int/influenza/pip/advisory_group/draft_twg_doc.pdf?ua=1>; PIP Framework Advisory Group TEWG on Genetic Sequence Data, *Best Process to Handle Genetic Sequence Data from Influenza Viruses with Human Pandemic Potential (IVPP GSD) Under the PIP Framework* (2015) <http://www.who.int/influenza/pip/advisory_group/gsdoptionspaper.pdf?ua=1>.

[195] While the outcome has been clearly beneficial, there have been challenges and disagreements along the way: see, eg S Nanda and M Kowalczuk, "Unpublished Genomic Data – How to Share?" (2014) 15(5) *BMC Genomics*, DOI: 10.1186/1471-2164-15-5; B Jasny, "Realities of Data Sharing Using the Genome Wars as Case Study: An Historical Perspective and Commentary" (2013) 2(1) *EPJ Data Science* 1; A McGuire et al, "Ethical and Practical Challenges of Sharing Data from Genome-wide Association Studies: The eMERGE Consortium Experience" (2011) 21 *Genome Research* 1001; D Blumenthal et al, "Data Withholding in Genetics and the Other Life Sciences: Prevalences and Predictors" (2006) 81(2) *Academic Medicine* 137.

[196] See, eg Blumenthal et al, n 195; E Campbell et al, "Data Withholding in Academic Genetics: Evidence from a National Survey" (2002) 287(4) *Journal of the American Medical Association* 473. See also J Kaye et al, "Data Sharing in Genomics – Re-shaping Scientific Practice" (2009) 10(5) *Nature Reviews Genetics* 331, 332-333.

[197] See, eg McEwen et al, n 83, 378-380; S Haga and J O'Daniel, "Public Perspectives Regarding Data-sharing Practices in Genomics Research" (2011) 14 *Public Health Genomics* 319; M Gymrek et al, "Identifying Personal Genomes by Surname Inference" (2013) 339(6117) *Science* 321; D Craig et al, "Assessing and Managing Risk when Sharing Aggregate Genetic Variant Data" (2011) 12(10) *Nature Reviews Genetics* 730; J Robinson et al, "Participants' Recall and Understanding of Genomic Research and Large-scale Data Sharing" (2013) 8(4) *Journal of Empirical Research on Human Research Ethics* 42. See also Kaye et al, n 196, 333-334.

[198] See Marshall, n 125. There was a similar outcome for the Syngenta Corporation rice genome sequence data: Moore, n 123.

[199] Noting many specialised databases have started to charge for access: P Schofield et al, "Sustaining the Data and Bioresource Commons" (2010) 330(6004) *Science* 592. Perhaps unsurprisingly, there are no long-term successful models of self-funding databases: C Chandras et al, "Models for Financial Sustainability of Biological Databases and Resources" [2009] *Database* bap017. Although the ideal remains for a culture of sharing: P Schofield et al, "Post-publication Sharing of Data and Tools" (2009) 461(7261) *Nature* 171.

and] information contained in GenBank".[200] NCBI also states that "[a]ll persons reproducing, redistributing, or making commercial use of this information are expected to adhere to the terms and conditions asserted by the copyright holder".[201] The effect of this policy position is that those using GenBank (and EMBL-EBI and DDBJ) data and information are required to determine the pedigree and freedom to operate for any uses they might make of these databases, and the data and information derived from those databases. This distinction between gratis and libre open access is important and provides a possible avenue to the apparent conflict between open access to DNA, RNA and amino acid sequence data and ABS regulatory schemes.

The article has asserted that benefit sharing under the CBD (and its Nagoya Protocol), the Plant Treaty and the PIP Framework faces a problem because data and information can be used without necessarily controlling access to the physical sample. So far this has only been a problem for ABS under the PIP Framework where vaccine production using only sequence data without access to a physical sample has been demonstrated.[202] For example, an avian influenza A(H7N9) virus vaccine has been produced using virus-like particles derived from biochemically synthesised viral DNA sequences.[203] This example demonstrates that by using only the DNA, RNA and amino acid sequence data accessed from databases such as GenBank and GISAID, the PIP Framework ABS arrangements involving a SMTA and benefit sharing can be avoided. This is, as the PIP Framework Advisory Group accepts,[204] a problem for ABS under the PIP Framework,[205] and their proposed solution is to monitor and trace the exploitation of sequence data in end products.[206] The details about how this might be achieved continue to be addressed under the PIP Framework.[207] For this article's purposes, however, this demonstrates that DNA, RNA and amino acid sequence data accessed from databases does present a problem for ABS schemes under the PIP Framework, and that the same or a similar problem is likely under the CBD (and its Nagoya Protocol) and the Plant Treaty.[208]

It must also be assessed as to whether DNA, RNA and amino acid sequence data can be exploited and information monitored and traced in end products. The automated detection of misappropriated DNA, RNA and amino acid sequence data is not quite as simple as it might first appear. A study interrogating patent applications using informatics techniques for evidence of the use of plant genetic resources derived from plants covered under the Plant Treaty demonstrated that monitoring and tracing is possible.[209] The study used text mining to interrogate patent databases for key data, such as varieties, accession codes and for UPOV (International Union for the Protection of New Varieties of

[200] NCBI, n 86.

[201] NCBI, n 88.

[202] See PIP Framework Advisory Group TEWG, n 183, 1-2.

[203] T Hahn et al, "Rapid Manufacture and Release of a GMP Batch of Avian Influenza A(H7N9) Virus-Like Particle Vaccine Made Using Recombinant Baculovirus-Sf9 Insect Cell Culture Technology" (2013) 12(2) *BioProcessing* 1538. See also P Dormitzer et al, "Synthetic Generation of Influenza Vaccine Viruses for Rapid Response to Pandemics" (2013) 5(185) *Science Translational Medicine* 185ra68; PIP Framework Advisory Group TEWG, n 183, 1-2.

[204] See PIP Framework Advisory Group TEWG, n 194.

[205] See PIP Framework, n 13, Art 5.2.4. See also World Health Assembly, n 20, Annex 1 ([6]) and Annex ([2]-[3]).

[206] PIP Framework Advisory Group TEWG, n 194, 2.

[207] See PIP Framework Advisory Group, n 190.

[208] See Subsidiary Body on Scientific, Technical and Technological Advice, *New and Emerging Issues Relating to the Conservation and Sustainable Use of Biodiversity – Possible Gaps and Overlaps with the Applicable Provisions of the Convention, its Protocols and other Relevant Agreements Related to Components, Organisms and Products Resulting from Synthetic Biology Techniques*, UNEP/CBD/SBSTTA/18/INF/4 (2014) [177]-[180] and the references therein; *First Meeting of the Ad Hoc Open-Ended Working Group to Enhance the Functioning of the Multilateral System*, Report, IT/OWG-EFMLS-1/14/Report (2014) [8].

[209] P Oldham and S Hall, "Intellectual Property, Informatics and Plant Genetic Resources" in N Moeller and C Stannard (eds), *Identifying Benefit Flows: Studies on the Potential Monetary and Nonmonetary Benefits Arising from the International Treaty on Plant Genetic Resources for Food and Agriculture* (Food and Agriculture Organization of the United Nations, 2013) 162-223. See also P Oldham et al, "Biological Diversity in the Patent System" (2013) 8(11) *PLoS One* e78737.

Plants) variety denomination names.[210] Despite findings mired by "large-scale problems with noise" and not being able to "establish a clear linkage between plant germplasm in public collections and patent data", the study concluded that the informatics techniques were feasible when refined and could identify plant genetic resources in commercial research and development.[211] The study is a salient proof-of-concept, although not technologically simple, requiring sophisticated algorithms, computing power and large-scale manual data cleaning to produce coherent results.[212] At best the study probably suggests that informatics techniques can narrow the field to suspicious patent applications, from which point further detailed scrutiny would be required to determine the particular contributions. And while there is an ongoing debate about including disclosure requirements in patent applications that could assist this monitoring and tracing DNA, RNA and amino acid sequence data,[213] patent databases are only a small portion of potential uses of sequence data. The other places and forms using sequence data make monitoring and tracing very complicated.[214] For example, the use of genetic derivatives, such as sequences altered for codon preferences in the preparation of an avian influenza A(H7N9) virus vaccine using virus-like particles derived from biochemically synthesised viral DNA sequences in an insect cell culture,[215] may be especially difficult to detect. Despite the potential for monitoring and tracing exploited sequence data in end products this does not seem to be a simple or complete solution.

There are other ways of monitoring and tracing the use of DNA, RNA and amino acid sequence data that impose obligations on those using sequences from publicly accessible databases like GenBank, EMBI-EBI and DDBJ to comply with ABS obligations consistent with the CBD (and its Nagoya Protocol), the Plant Treaty and the PIP Framework:

(a) *The contract model* – Each of the CBD (and its Nagoya Protocol), the Plant Treaty and PIP Framework adopts a contract model for ABS requiring a binding agreement between the provider of the materials and the users.[216] The terms and conditions of these agreements could impose binding obligations requiring users (and subsequent users where those terms and conditions are passed on) to disclose or report on their uses of the supplied materials, including the non-material uses. The terms and conditions would need to build in incentives and penalties to promote benefit sharing and avoid uses of the materials contrary to the ABS obligations. A relatively easy option would be to require those submitting DNA, RNA and amino acid sequences to databases to assert conditions on the submission that the samples are made available subject to restricted permission concerning the use, copying, or distribution of the data and information obtained from the database, and requiring those exploiting the sequence for commercial purposes to seek specific permissions that address ABS.

(b) *The copyright and database right model* – Sequencing DNA, RNA and amino acids requires the documentation of the sequence in a material form and this may mean copyright subsists in the written representation of the sequence under copyright laws[217] and database laws.[218] Whether

[210] Oldham and Hall, n 209, 208.

[211] Oldham and Hall, n 209, 208.

[212] See Oldham and Hall, n 209, 205-208.

[213] See, eg Conference of the Parties to the Convention on Biological Diversity, *Defusing Disclosure in Patent Applications: Strengthening Legal Certainty in the International Regime on Access to Genetic Resources and Benefit-Sharing and Supporting WIPO's Intergovernmental Committee on Intellectual Property and Genetic Resources, Traditional Knowledge and Folklore*, UNEP/CBD/COP/10/INF/44 (2010).

[214] See PIP Framework Advisory Group TEWG, n 183, 11-12.

[215] See Hahn et al, n 203.

[216] See CBD, n 7, Art 15; Nagoya Protocol, n 11, Art 6; Plant Treaty, n 12, Art 12.4; PIP Framework, n 13, Art 5.4.

[217] See W Stemmer, "How to Publish DNA Sequences with Copyright Protection" (2002) 20(3) *Nature Biotechnology* 217; N Derzko, "Protecting Genetic Sequences under the Canadian Copyright Act" (1993) 8(1) *Intellectual Property Journal* 31. Noting that the DNA, RNA and amino acid molecule itself might be copyrightable: see also A Torrance, "DNA Copyright" (2011) 46(1) *Valparaiso University Law Review* 1; C Holman, "Copyright for Engineered DNA: An Idea Whose Time Has Come?" (2011) 13(3) *West Virginia Law Review* 699; S Coke, "Copyright and Gene Technology" (2002) 10 JLM 97; S McBride, "Bioinformatics and Intellectual Property Protection" (2002) 17(4) *Berkeley Technology and Law Journal* 1331;

copyright subsists in scientific discoveries and facts (like sequence data in GenBank, EMBL-EBI and DDBJ) remains uncertain, although the preferable view is probably that these data will not be copyright protected.[219] Where copyright and database rights do not apply to the sequence data and information, this may be addressed by changing the form of expression, such as adopting a music format[220] or including watermarking.[221] Where a copyright or a database right exists, this prevents another copying without the permission of the rights holder subject to some exceptions. The rights holder would then need to make it easy for those seeking permission and set out the clear terms and conditions of permission. Again, a relatively easy option would be to restrict permission concerning the use, copying, or distribution of the data and information obtained from the database, and requiring those exploiting the sequence for commercial purposes to seek specific permissions that address ABS. Any use of the sequence without the permission of the copyright or database right holder will be an infringement and subject to an action for damages, and so on.

There seems little doubt, however, that developments in science and technology are rapidly advancing and that increasingly the DNA, RNA and amino acid sequence data and information accessed from databases can be used without the physical sample.[222] This does mean that open access to DNA, RNA and amino acid sequences can undermine the ABS schemes set out in the CBD and Nagoya Protocol, Plant Treaty and PIP Framework. None of the proposed solutions – monitoring and tracing favoured under the PIP Framework, the contract model, and the copyright and database right model – provides a perfect solution. Each model does, however, suggest that open access to these sequences might be at least partially reconciled with the benefit-sharing obligations under the CBD and Nagoya Protocol, Plant Treaty and PIP Framework. And this really goes to the heart of the broader problem – within the market there are competing claims for open access and contrary claims for maintaining secrecy. The challenge is, and will continue to be, finding a compromise that balances the needs of open access and fair and equitable benefit sharing.

D Burk, "Copyrightability of Recombinant DNA Sequences" (1989) 29(3) *Jurimetrics* 469; D Smith, "Copyright Protection for the Intellectual Property Rights to Recombinant Deoxyribonucleic Acid: A Proposal" (1988) 19 *St Mary's Law Journal* 1083; J Goldstein, "Copyrightability of Genetic Works" (1984) 2 *Bio/Technology* 138; I Kayton, "Copyright in Living Genetically Engineered Works" (1982) 50(2) *George Washington Law Review* 191.

[218] See R Harris and S Rosenfield, "Copyright Protection for Genetic Databases" (2005) 45(2) *Jurimetrics* 225; A Hasan, "Sweating in Europe: The European Database Directive" (2005) 9(3) *Computer Law Review and Technology Journal* 479; E Baba, "From Conflict to Confluence: Protection of Databases Containing Genetic Information" (2003) 30(1) *Syracuse Journal of International Law and Commerce* 121.

[219] See, eg E Howard and G Ramsey, "Bioinformatics Databases: Questions of Copyright" [November 2002] *BioPharm International* 46, 46.

[220] See Stemmer, n 217.

[221] See S-H Lee, "DWT Based Coding DNA Watermarking for DNA Copyright Protection" (2014) 273 *Information Sciences* 263.

[222] There is a huge literature heralding many examples: see eg S Laird and R Wynberg, *Bioscience at a Crossroads: Implementing the Nagoya Protocol on Access and Benefit Sharing in a Time of Scientific, Technological and Industry Change* (Secretariat of the Convention on Biological Diversity, 2012).