Article

MEDICAL LAW INTERNATIONAL

Medical Law International
12(2) 92–120

DOI: 10.1177/0968533212458431
mli.sagepub.com
SAGE

Open science and community norms: Data retention and publication moratoria policies in genomics projects

Yann Joly and Edward S. Dove
Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine, McGill University, Canada

Karen L. Kennedy
Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Martin Bobrow
Department of Medical Genetics, University of Cambridge, Cambridge, UK

B.F. Francis Ouellette
Ontario Institute for Cancer Research, Toronto, Canada

Stephanie O.M. Dyke
Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK

Kazuto Kato
Department of Biomedical Ethics and Public Policy, Graduate School of Medicine, Osaka University, Osaka, Japan

Bartha M. Knoppers
Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine, McGill University, Canada

Corresponding author:
Yann Joly, Centre of Genomics and Policy, Department of Human Genetics, Faculty of Medicine, McGill University, 740 Dr Penfield Avenue, Suite 5101, Montreal, QC, Canada H3A 0G1
Email: yann.joly@mail.mcgill.ca

Abstract
While modern genomics research often adheres to community norms emphasizing open data sharing, many genomics institutes and projects have recently nuanced such norms with a corpus of data release policies. In particular, publication moratoria and data retention policies have been enacted to 'reward' data producers and ensure data quality control. Given the novelty of these policies, this article seeks to identify and analyse the main features of data retention and publication moratoria policies of major genomics institutes and projects around the world. We find that as more collaborative genomics projects are created, and further genomic research discoveries are announced, the need for more sophisticated yet practical and effective policies will increase. Reward systems should be implemented that recognize contributions from data producers and acknowledge the need to remain dedicated to the goals of open data sharing. To this end, in addition to the current choices of employing data retention or publication moratoria policies, alternative models that would be easier to implement or less demanding on open science should also be considered.

Keywords
Biobanks, data sharing, genomics, open science, publication moratorium

Introduction

International law and guidelines play a central role in encouraging individuals, particularly researchers, to share knowledge's progenitor – data – and in ensuring the continued existence of an 'open science' domain. A number of international instruments stress that individuals and states should work to encourage the free circulation of data and share in scientific advancement (e.g. the OECD *Declaration on Access to Research Data From Public Funding* (2004), Article 27 of the *Universal Declaration of Human Rights* (1949), Article 15 of the *International Covenant on Economic, Social and Cultural Rights* (1966), Articles 13 and 19 of the *Universal Declaration on the Human Genome and Human Rights* (1997), Article 18 of the *International Declaration on Human Genetic Data* (2003) and Article 15 of the *Universal Declaration on Bioethics and Human Rights* (2005)). In the United States, copyright law protection is waived on data directly produced by federal government agencies, thereby allowing its deposition into the public domain.[1] Recently, the *Federal Research Public Access Act of 2012* was presented as a bill in the US Congress. The Bill would require federal research-granting agencies to make resulting publications freely available to the public within 6 months of publication in a peer-reviewed journal.[2] The policy of open access has also been echoed in Canada, where the federal government released a consultation paper in 2010 stating that

1. J.H. Reichman and P.F. Uhlir, 'A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment', *Law and Contemporary Problems* 66 (2003), pp. 315–462.
2. U.S., Bill H.R. 4004, *Federal Research Public Access Act of 2012*, 112th Congress, 2012.

'[g]overnments can help by making publicly-funded research data more readily available to Canadian researchers and businesses. Open access is consistent with many national strategies and holds great economic potential for Canadians. . . '[3]

At the same time, laws and government policies can restrict data sharing, be it to protect individual privacy, national security or financial benefits arising from innovation.[4] Indeed, law often works to achieve principled approaches that strike an appropriate balance between openness and other important social values such as privacy, security and the encouragement of entrepreneurial creativity and high-risk research. 'Openness', Professor Sheila Jasanoff writes, 'is a treasured attribute of science, but like most good things, even scientific openness has to be purposefully cultivated and judiciously deployed in order to serve its intended functions well.'[5]

Modern genomics research is a prime example of a field that strives to adhere to the principle of rapid open data sharing.[6] Many believe that data sharing facilitates the achievement of scientific community goals, including replicating results, promoting new research, improving methods of data collection and measurement, educating new researchers and allowing more effective use of researchers' and funding agencies' limited resources.[7] While genomics data sharing is relatively new, open data sharing as a practice has existed in various iterations and modalities in other scientific disciplines for

3. Government of Canada, *Consultation Paper on a Digital Economy Strategy for Canada* (Ottawa, Canada: Government of Canada, 2010). Available at: http://www.digitaleconomy.gc.ca/eic/site/028.nsf/eng/h_00025.html (accessed 6 July 2012).
4. S. Jasanoff, 'Transparency in Public Science: Purposes, Reasons, Limits', *Law and Contemporary Problems* 69 (2006), pp. 21–45.
5. Jasanoff, 'Transparency', p. 42.
6. Human Genome News, January 1993, 'DOE guidelines encouraging sharing of data, resources' Available at: http://www.genome.gov/EdKit/pdfs/1992b.pdf (accessed 9 July 2012); International Strategy Meeting on Human Genome Sequencing, 'Summary of Principles Agreed at the International Strategy Meeting on Human Genome Sequencing, Bermuda, 25–28 February 1996' Available at: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1 (accessed 9 July 2012).
7. B.M. Knoppers et al., 'Towards a Data Sharing Code of Conduct for International Genomic Research', *Genome Medicine* 3 (2011), p. 46; P. Boddington, 'Data Sharing in Genomics', in P. Boddington, *Ethical Challenges in Genomics Research* (New York: Springer, 2012), p. 195; B.A. Fischer and M.J. Zigmond, 'The Essential Nature of Sharing in Science' *Science and Engineering Ethics* 16 (2010), pp. 783–799; T. Silverstein et al., 'The Commercialization of Genomic Academic Research: Conflicting Interests?', in E.R. Gold and B.M. Knoppers, eds. *Biotechnology IP and Ethics* (Markham, Ontario: LexisNexis Canada, 2009), pp. 131–163; Y. Joly, 'Open Sources Approaches in Biotechnology: Utopia Revisited', *Maine Law Review* 59 (2007), pp. 386–405; H.A. Piwowar et al., 'Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers', *PLoS Medicine* 5 (2008), p. e183; F.S. Collins et al., 'A Vision for the Future of Genomics Research', *Nature* 422 (2003), pp. 835–847; R.C. Rockwell and R.P. Abeles, 'Sharing and Archiving Data is Fundamental to Scientific Progress', *The Journals of Gerontology: Series B, Psychological Sciences and Social Sciences* 53 (1998), pp. S5–S8.

decades.[8] Indeed, one may view open data sharing as a continuation of 'scientific progress', a concept of 16th-century origin associated with the ideal of free and open dissemination of knowledge.[9]

Recognizing this historical backdrop, this article situates data sharing in its modern social, legal and scientific context by (1) analysing the key features of data retention policies (strategies that control the rate and amount of data sharing) and publication moratoria policies (strategies that control the rate at which knowledge in the public domain is freely utilizable) of international genomics projects; (2) assessing the strengths and weaknesses of these various policies and (3) exploring the potential for alternative models that may better reflect current and future trends.

Context

Today, open science includes submission of data into online databases that are subject to specific requirements, rules and procedures.[10] With respect to publicly-funded genomics projects, data sharing may be an integral part of fulfilling obligations to both funding agencies and project participants.[11] There are myriad illustrations of successful data sharing in the genomics field, including three established and jointly collaborative global databases: GenBank at the National Center for Biotechnology Information (NCBI) at the US National Institutes of Health (NIH), EMBL-Bank at the European Bioinformatics Institute (EBI), and the Center for Information Biology and DNA Data Bank of Japan (DDBJ).

In recent years, some large-scale publicly-funded genomics projects have crafted a corpus of innovative data release policies that nuance the open data sharing principle.[12] Commentators have expressed concerns about unqualified data sharing,[13] such as the threat of infringement of proprietary rights, feared violations of confidentiality and

8. S. Hilgartner, 'Biomolecular Databases: New Communication Regimes for Biology?', *Science Communication* 17 (1995), pp. 240–263; National Research Council, Committee on National Statistics, *Sharing Research Data* (Washington, DC: National Academy Press, 1985); V.A. de Wolf et al., 'Part I: What is the Requirement for Data Sharing?', *IRB: Ethics and Human Research* 27 (2005), pp. 12–16.
9. Fischer et al., 'Essential Nature' p. 783; Joly, 'Utopia Revisited' p. 386; P.A. David, 'The Historical Origins of 'Open Science': An Essay on Patronage, Reputation and Common Agency Contracting in the Scientific Revolution', *Capitalism and Society* 3 (2008), pp. 1–106.
10. de Wolf et al., 'Requirement', p. 12; S.B. Haga and J. O'Daniel, 'Public Perspectives Regarding Data-Sharing Practices in Genomics Research', *Public Health Genomics* 14 (2011), pp. 319–324.
11. S. Fortin et al., 'Access Arrangements' for Biobanks: A Fine Line between Facilitating and Hindering Collaboration', *Public Health Genomics* 14 (2011), pp. 104–114; J. Kaye et al., 'Data Sharing in Genomics: Re-shaping Scientific Practice', *Nature Reviews Genetics* 10 (2009), pp. 331–335.
12. S.O. Dyke and T.J. Hubbard, 'Developing and Implementing an Institute-Wide Data Sharing Policy', *Genome Medicine* 3 (2011), p. 60.
13. Rockwell and Abeles, 'Sharing and Archiving', p. S5; G. Boulton et al., 'Science as a Public Enterprise: the Case for Open Data', *Lancet* 377 (2011), pp. 1633–1635; C. Tenopir et al., 'Data Sharing by Scientists: Practices and Perceptions', *PLoS One* 6 (2011), p. e21101.

privacy,[14] data misuse/dual use and misinterpretation,[15] predicted harm to the data producer's reputation because of misinterpreted data, and scooping by 'data users' (i.e. those who wish to access the genomic data).[16] In the genomics arena, there has been speculation that reluctance by some researchers to embrace data sharing may be due to increased scientific competitiveness in the field as well as opportunities for commercial application.[17] What is clear is that as the quantity, speed and mode of data release advances and increases in sophistication, along with the creation of more large-scale, internationally collaborative efforts, the role of 'data producers' (i.e. those who produce raw genomic data) in genomics projects has metamorphosed.[18]

Data producers may need to comply with institutional or professional norms, university licensing requirements, guidelines or pressures to retain dominance in their field and protect their contribution.[19] There is also a need to protect the privacy of research contributors' data in the face of large-scale, in silica, international studies and sophisticated technologies that can possibly reidentify individuals.[20] Privacy can be protected by data

14. C.J. Savage and A.J. Vickers, 'Empirical Study of Data Sharing by Authors Publishing in PLoS Journals', *PLoS One* 4 (2009), p. e7078; G. Laurie et al., 'Managing Access to Biobanks: How Can We Reconcile Individual Privacy and Public Interests in Genetic Research?', *Medical Law International* 10 (2010), pp. 315–337.
15. L. Bezuidenhout, 'Data Sharing and Dual-Use Issues', forthcoming in *Science and Engineering Ethics*; E.G. Campbell et al., 'Data Withholding in Academic Genetics: Evidence from a National Survey', *Journal of the American Medical Association* 287 (2002), pp. 473–480.
16. D.M. Gitter, 'The Challenges of Achieving Open-Source Sharing of Biobank Data', *Biotechnology Law Report* 29 (2010), pp. 623–635; D.M. Gitter, 'The Application of Data Access Policies Designed for Genome-Wide Association Studies to Smaller Scale Databases', *The John Marshall Review of Intellectual Property Law* 10 (2011), pp. 476–490; W.O. Hagstrom, *The Scientific Community* (New York, NY: Basic Books, 1965).
17. Tenopir et al., 'Data Sharing', p. e21101.
18. P. Boddington, 'Data Sharing in Genomics', p. 195; Kaye et al., 'Re-shaping' p. 331; C.Heeney and A. Smart, 'Enacting Governance: The Case of Access', in J. Kaye et al., eds. *Governing Biobanks: Understanding the Interplay Between Law and Practice* (Oxford and Portland, OR,: Hart Publishing, 2012), p. 232.
19. de Wolf et al., 'Requirement', p. 12.
20. Z. Lin, A.B. Owen and R.B. Altman, 'Genomic Research and Human Subject Privacy', *Science* 305 (2004), p. 183; N. Homer et al., 'Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays', *PLoS Genetics* 4 (2008), p. e1000167; E.A. Zerhouni and E.G. Nabel, 'Protecting Aggregate Genomic Data', *Science* 322 (2008), p. 44; T. Caulfield et al., 'Research Ethics Recommendations for Whole-Genome Research: Consensus Statement', *PLoS Biology* 6 (2008), p. e73; M. Parker et al., 'Ethical Data Release in Genome-Wide Association Studies in Developing Countries', *PLoS Medicine* 6 (2009), p. e1000143; A.L. McGuire et al., 'To Share or Not to Share: A Randomized Trial of Consent for Data Sharing in Genome Research', *Genetics in Medicine* 13 (2011), pp. 948–955; J.M. Oliver et al., 'Balancing the Risks and Benefits of Genomic Data Sharing: Genome Research Participants' Perspectives', *Public Health Genomics* 15 (2012), pp. 106–114; H.K. Im et al., 'On Sharing Quantitative Trait GWAS Results in an Era of Multiple-omics Data and the Limits of Genomic Privacy', *American Journal of Human Genetics* 90 (2012), pp. 591–598; E.E. Schadt et al., 'Bayesian

users adhering to certain safeguards, but doing so may entail delays in data release due to additional processing steps, such as checking the credentials of data user applicants.[21] While there is a recognized need by data users for oversight mechanisms and associated delays in data access for certain data sets to protect the privacy of research contributors' or patients' data,[22] privacy and issues are beyond the scope of this article, which focuses instead on evidence of increasingly complex restrictions on data release policies. Moreover, unlike data retention policies, which may be enacted for various reasons, organizations implement publication moratoria policies to reward the creativity and labour of data producers, rather than to protect the privacy of research subjects or patients.

Most critically for the purpose of this article, many researchers experience pressure to disseminate their data before it is published. This has led to discussions within the genomics community about striking an appropriate balance between prepublication data dissemination and the desire to secure direct rewards and acknowledgement for scientific output, particularly through the medium of publications.[23] Thus, in the postgenomic era, we are witnessing a change in genomics projects: in the face of an entrenched normative guideline of disseminating data to data users,[24] data dissemination is increasingly subjected to endogenous and exogenous complexities and qualifications.

There is a deep-rooted desire by data producers to be recognized for their contributions and invested effort.[25] In the 1940s, the sociologist Robert K. Merton documented how researchers share data, in part, to inform the community of their discoveries and receive the recognition and accolades of their colleagues.[26] More recent scholarship highlights the competitive elements of the scientific enterprise, the role and motivations

---

Method to Predict Individual SNP Genotypes from Gene Expression Data', *Nature Genetics* 44 (2012), pp. 603–608.

21. Y. Joly et al., 'Data Sharing in the Post-Genomic World: The Experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO)', *PLoS Computational Biology* 8 (2012), p. e1002549.
22. P.N. Ossorio, 'Bodies of Data: Genomic Data and Bioscience Data Sharing', *Social Research: An International Quarterly* 78 (2011), pp. 907–932; Heeney and Smart, 'Enacting Governance', p. 249.
23. Piwowar et al., 'Data Sharing Culture', p. e183; E.G. Campbell and E. Bendavid, 'Data-Sharing and Data-Withholding in Genetics and the Life Sciences: Results of a National Survey of Technology Transfer Officers', *Journal of Health Care Law and Policy* 6 (2003), pp. 241–255.
24. Gitter, 'Challenges', p. 623; Gitter, 'Application', p. 476; J.L. Contreras, 'Data Sharing, Latency Variables and the Science Commons', *Berkeley Technology Law Journal* 25 (2010), pp. 1602–1672; J.L. Contreras, 'Bermuda's Legacy: Patents, Policy and the Design of the Genome Commons', *Minnesota Journal of Law, Science and Technology* 12 (2011), pp. 61–125.
25. B.J. Strasser, 'The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine', *Isis*, 102 (2011), pp. 60–96.
26. R.K. Merton, 'The Normative Structure of Science' (1942), in N.W. Storer, ed., *The Sociology of Science: Theoretical and Empirical Investigations* (Chicago, IL: University of Chicago Press, 1979), pp. 267–278.

of data sharing in science[27] and the importance of intellectual property rights for promoting central norms of the scientific community.[28] These rewards may come in the form of intangible benefits like greater self-esteem[29] and also as tangible benefits such as patents, tenure advancement, endowed chairs and grant funding.[30] It may also be the case that the broader community interest is served by giving researchers adequate rewards, since if this does not happen, the best minds might not be willing to go into the field or share.

While this desire for reward endures, the exponential growth in science and technology outputs, the commercialization of university-based biomedical research,[31] the highly contingent nature of genomic data[32] and the deluge in data volume[33] engender new questions about policies centred on uninhibited open data sharing.[34] For instance, as the scientific community adopts open data sharing principles as a norm, data producers may desire an exclusive period of time to refine and analyse experimental data for publication before allowing data users to use the same data for their own publications.[35] Another

27. Hilgartner, 'Biomolecular Databases', p. 240; S.J. Ceci, 'Scientists' Attitudes toward Data Sharing', *Science, Technology, and Human Values* 13 (1988), pp. 45–52; V. Weil and R. Hollander, 'Normative Issues in Data-Sharing', in J. Sieber, ed., *Sharing Social Science Data: Advantages and Challenges* (London, Sage, 1991), pp. 151–157; S. Hilgartner and S. I. Brandt-Rauf, 'Data Access, Ownership and Control: Toward Empirical Studies of Access Practices', *Science Communication*, 15 (1994), pp. 355–372; National Research Council, Committee on Responsibilities of Authorship in the Biological Sciences *Sharing Publication-related Data and Materials: Responsibilities of Authorship in the Life Sciences* (Washington, DC: National Research Council, 2003); M.M. Wasko and S. Faraj, 'Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice', *MIS Quarterly* 29 (2005), pp. 35–57; D. Blumenthal et al., 'Data Withholding in Genetics and the Other Life Sciences: Prevalences and Predictors', *Academic Medicine* 81 (2006), pp. 137–145; H.A. Piwowar et al., 'Sharing Detailed Research Data is Associated with Increased Citation Rate', *PLoS One* 2 (2007), p. e308; J. Tucker, '*Motivating Subjects: Data Sharing in Cancer Research*' PhD Thesis, Virginia Polytechnic Institute and State University, 2009.
28. S.F. Kieff, 'Facilitating Scientific Research: Intellectual Property Rights and the Norms of Science: A Response to Rai and Eisenberg', *Northwestern University Law Review* 95 (2001), pp. 691–705.
29. Hagstrom, Scientific Community.
30. Contreras, 'Bermuda's Legacy', p. 61.
31. S.H. Harmon, T. Caulfield and Y. Joly, 'Commercialization versus Open Science: Making Sense of the Message(s) in the Bottle', *Medical Law International* 12 (2012), pp. 3–10; A. K. Rai and R.S. Eisenberg, 'Bayh-Dole Reform and the Progress of Biomedicine', *Law and Contemporary Problems* 66 (2003), pp. 289–314.
32. F. Milanovic, D. Pontille and A. Cambon-Thomsen, 'Biobanking and Data Sharing: A Plurality of Exchange Regimes', *Genomics, Society and Policy* 23 (2007), pp. 17–30.
33. C.L. Borgman, 'The Conundrum of Sharing Research Data', *Journal of American Society for Information Science* 63 (2012), pp. 1059–1078.
34. Gitter, 'Application', p. 476; Contreras, 'Data Sharing', p. 1602; Strasser, 'Experimenter's Museum', p. 60.
35. P. Boddington, 'Data Sharing in Genomics', p. 201; L. Rowen et al., 'Publication Rights in the Era of Open Data Release Policies', *Science* 289 (2000), p. 1881; P. Arzberger et al., 'An International Framework to Promote Access to Data', *Science* 303 (2004), pp. 1777–1778.

example is the omnipresent requirement that data users acknowledge the source of their data,[36] a stipulation that 92% of recently surveyed researchers consider important when sharing their data.[37] Nonetheless, many in the science and data sharing community express a need to preserve open data sharing principles as much as possible. As open science advocate John Wilbanks contends, 'nothing other than the public domain really works from the perspective of data integration. And data integration is coming at us at exponential speed'.[38]

Recent internationally adopted principles (such as the Fort Lauderdale Agreement in 2003 and the Toronto Statement in 2009) and large-scale genomics projects attest that the genomics community wants to share or give access to data while preserving credit for data producers. Simultaneously, a wide variety of researchers increasingly desire access to genomic data.[39] Tenopir and colleagues found that 85% of surveyed researchers want to use other researchers' data sets, if those data sets are easily accessible.[40] However, these researchers may not be cognizant of the various norms of rapid prepublication data release endorsed by the genomics community.[41] Consequently, since the mid-2000s, many large-scale genomics projects have created or modified data release policies that are more sophisticated in their approach to addressing the needs of various genomics stakeholders, including the protection of the contributions of data producers.

In particular, two main policy approaches have been adopted following the Fort Lauderdale Agreement that pertain to the 'knowledge commons',[42] which we define as an ecosystem of useful information shared among communities that may be public or quasi-public (such as within an international research consortium) and subject to social dilemmas. First, genomics projects have adopted a 'knowledge latency-based' data retention strategy that controls the rate of data entering the commons. Second, genomics projects have adopted a 'rights latency-based' publication moratorium strategy that controls the rate at which knowledge in the commons is freely utilizable.[43] While, as noted earlier, various modalities of data retention and release in science disciplines are long-standing, owing to its modern nature, genomics data sharing policies have recently changed. Yet, even within this modern domain, mature ideas exist. Data retention policies in large genomic data sets date to the early 1990s, when, for example, 1992 NIH/

36. Fortin et al., 'Access Arrangements', p. 104.
37. Tenopir et al., 'Data Sharing', p. e21101.
38. J. Wilbanks, 'Public Domain, Copyright Licenses and the Freedom to Integrate Science', *Journal of Science Communication* 7 (2008), pp. 1–10.
39. Contreras, 'Bermuda's Legacy', p. 61.
40. Tenopir et al., 'Data Sharing', p. e21101.
41. Contreras, 'Data Sharing', p. 1602; R.M. Cook-Deegan, 'The Urge to Commercialize: Interactions between Public and Private Research and Development', in National Research Council, *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium* (Washington, DC: National Academy Press, 2003), pp. 87–94.
42. C. Hess and E. Ostrom, eds., *Understanding Knowledge as a Commons: From Theory to Practice* (Cambridge, MA: MIT Press, 2006); C. Hess, 'The Unfolding of the Knowledge Commons', *St Antony's International Review* 8 (2012), pp. 13–24.
43. Contreras, 'Bermuda's Legacy', p. 61.

Department of Energy (DOE) guidelines established a 6-month period of time from when Human Genome Project data were generated until they had to be made publicly available.[44] Data retention policies date even earlier in other biomedical domains that contain publicly available data sets. In the early 1980s, the Protein Data Bank at Brookhaven National Laboratory allowed deposition of data without access to external users for 1–4 years after the publication in a journal of the general conclusions derived from the data, in order to protect the authors' ability to exploit it further.[45]

Since the 1996 Bermuda Principles and the growth of genome-wide association studies (GWAS), the data retention strategy has diminished from 6 months to a 'rapid' data release period.[46] More recently, genomics institutes and projects have adopted publication moratorium policies, largely for two reasons. First, data producers expressed concerns regarding a potential loss of publication privileges and general scientific advantage; second, genomics projects were releasing data increasingly in a continuous manner, rather than by a bulk release of entire data sets for analyses. Starting with the Wellcome Trust Case Control Consortium (WTCCC) in 2005 and the US federally funded Genetic Association Information Network (GAIN) in 2006, policies have been implemented that prohibit data users from submitting publications based on accessed data for a specified 'embargo' period. Since WTCCC and GAIN, other genomics projects have implemented data release policies that employ a data retention policy or a publication moratorium policy or both.

Recently, certain large funding agencies and government research institutions have reexamined the issue of publication moratoria for large genomics projects vis-à-vis the sharing of primary data sets. For example, NIH was preparing to release a new data sharing policy for sequencing data sets that would have required the rapid, prepublication, submission of primary data files. In particular, it would have removed the publication moratorium. This was a result of perceived increasing concern about the complexity for funders, data producers and users of tracking data submissions and moratorium periods. The revised policy would have instituted a 'privacy period' for data producers in certain genomic data sets. Rapid data submission would remain the norm and submissions would be deposited in a central database where researchers could have the knowledge of the contents, but the data would not be released for 6 months. At release, the data would be available for use without any restrictions or publication moratorium. However, this 'privacy' or 'blackout period' would not apply to large 'community resource projects', such as The Cancer Genome Atlas (TCGA). 'Community resource projects' are research projects specifically devised and implemented to create a set of data whose primary utility will be as a resource for the broad scientific community. For these projects, data would continue to be made available immediately, in line with the principles of the Fort Lauderdale Agreement.

44. Human Genome News, 'DOE Guidelines' Available at: http://www.genome.gov/EdKit/pdfs/1992b.pdf (accessed 8 July 2012).
45. Strasser, 'Experimenter's Museum', p. 60.
46. GAIN Collaborative Research Group, 'New Models of Collaboration in Genome-Wide Association Studies: The Genetic Association Information Network', *Nature Genetics* 39 (2007), pp. 1045–1051.

Despite reexamination, NIH never adopted the new data sharing policy. NIH postponed its release following the announcement of the reduction in archive services at NCBI's Sequence Read Archive in February 2011 due to budget constraints. In June 2011, NCBI reversed its decision to close the archive, stating that with a commitment of interim funding and a plan for future support, it would continue to accept submissions and maintain the Sequence Read Archive and Trace Archive repositories for high-throughput sequence data.[47] Nonetheless, the new policy reflects a nascent debate in the genomics community about whether to impose 'community resource' standards on all data producers or whether to permit some delay in data release, during which time only data producers have access to their data. Indeed, some may view as insurmountable the practical challenges to implementing and enforcing publication moratoria on a 'rolling' basis (i.e. releasing data continuously over long periods of time) and in complex projects such as those in international consortia.

Given that genomic data sets are becoming more complex amidst the growth of internationally collaborative projects,[48] it is not surprising that the policies that govern their release are becoming more complex as well. This article continues – and broadens – the preliminary publication moratorium and data retention discussion and analysis undertaken in previous scholarly research.[49]

Methods

To identify a broad spectrum of data release policies concerning prepublication knowledge retention and publication moratoria, we used three complementary search strategies: (1) identify data retention and publication policies mentioned in access policies of large-scale genomics institutes and projects (this includes genetic databases, data sets, archives, repositories and biobanks); (2) identify data retention and publication policies generated by international consensus statements or mentioned in the guidelines of significant government research institutions or funding agencies; and (3) critically review the literature on data retention and publication policies produced by the legal, ethics and scientific community so as to inform our analysis and discussion.

Policies of genomics institutes and projects

We prepared a list of genomics projects by identifying the Charter members of the Public Population Project in Genomics (P^3G) and Society Consortium[50] and partners of the ENGAGE Consortium.[51] Both of these international consortia foster collaboration and

47. P. Cooper and R. Morris, eds. 'NCBI News' (June 2011) Available at: http://www.ncbi.nlm.nih.gov/books/NBK56224/pdf/June11.pdf (accessed 8 July 2012).
48. J.A. Singh and A.S. Daar, 'Intra-Consortium Data Sharing in Multi-National, Multi-Institutional Genomic Studies: Gaps and Guidance', *The HUGO Journal* 3 (2009), pp. 11–14.
49. Parker et al., 'Ethical Data Release', p. e1000143; Gitter, 'Challenges', p. 623; Gitter, 'Application', p. 476; Contreras, 'Data Sharing', p. 1602; Contreras, 'Bermuda's Legacy', p. 61.
50. http://www.p3g.org (accessed 8 July 2012).
51. http://www.euengage.org (accessed 8 July 2012).

harmonization of genomics projects and data on a global scale. We identified additional current, large-scale retrospective and prospective genomics projects by consulting the multidisciplinary academic literature. In concordance with systematic, web-based review methodology, we reviewed the PubMed,[52] SciVerse Scopus,[53] Social Science Research Network,[54] JSTOR,[55] LexisNexis Academic[56] and HeinOnline[57] databases, as well as Google Scholar, with an open date range. The inclusion criteria were (1) the presence of an official Web site for the project, (2) the availability of English language policies on the Web site or through personal correspondence, and (3) publication policies with application and reference to external data users. ('Publication policies' included stand-alone policies that addressed publications or data access and policies that were embedded in the clauses of data or material transfer agreements.) Using this methodology, we identified 77 genomics projects and retained 53 projects among them. Appendix 1 lists all genomics projects surveyed for data release policies (Full supplemental information, including extracts from genomics project or consortia Web sites that address publication moratorium, data retention and/or data access policies, is available on file with YJ and ESD).

Policy, consensus statements and literature research

Using the HumGen database[58] and using the keywords 'biobank', 'database', 'data set' and 'policy', we prepared a list of data retention and publication policy guidelines of government research institutions or funding agencies (both government and nongovernment funded) that contained recommendations on the governance of genomics projects. Coupled with the HumGen international search, we developed a purposive sampling frame that included government research institutions or funding agencies that have made significant contributions in the genomics policy-making field: National Institutes of Health (United States), Canadian Institutes of Health Research (Canada), Genome Canada (Canada), INMEGEN (Mexico), Medical Research Council (United Kingdom), Biotechnology and Biomedical Sciences Research Council (United Kingdom), Wellcome Trust (United Kingdom), INSERM (France) and the National Health and Medical Research Council (Australia). In order to identify relevant policies, we used the keywords 'biobank', 'data set', 'database', 'data' and 'publication' in the search engines of each Web site. Using the same inclusionary criteria as in our genomics institutes and projects research above, we identified 18 data retention and publication policies and retained 7 policies among them. Appendix 2 lists the 7 government research institutions or funding agencies.

Using the same databases and methodology described above, we also critically reviewed international consensus statements and the literature from the legal, ethics, social

52. http://www.ncbi.nlm.nih.gov/pubmed/ (accessed 8 July 2012).
53. http://www.info.sciverse.com/scopus (accessed 8 July 2012).
54. http://www.ssrn.com/ (accessed 8 July 2012).
55. http://www.jstor.org/ (accessed 8 July 2012).
56. http://www.lexisnexis.com (accessed 8 July 2012).
57. http://home.heinonline.org/ (accessed 8 July 2012).
58. http://www.humgen.org (accessed 8 July 2012).

sciences and scientific community so as to inform our discussion and analysis. Appendix 3 lists some of the major international consensus statements that address life sciences governance as it pertains to publications and data retention by data producers or users.

Results

Data release policies and guidelines

A large majority (75%) of projects had some version of data release policy or guideline in place that may have also contained a publication moratorium or data retention period. This likely reflects an adherence to international norms like the Bermuda Principles and Fort Lauderdale Agreement, which stress open access to data, as well as the fact that many genomics projects are publicly funded and therefore data sharing obligations are imposed on them by funding agencies. This may be indicated in various iterations such as a 'data management and data sharing plan' (Wellcome Trust), a 'data sharing policy' (Biotechnology and Biomedical Sciences Research Council), a 'data sharing and data preservation plan' (Medical Research Council) or a 'data release plan' (Genome Canada), all of which must be prepared by a project before funds are released.[59]

The vast majority of genomics projects (87%) had publications and acknowledgement policies in place, which may reflect a desire to increase international exposure within the scientific community. On the other hand, policies imposing prepublication review of a user's draft publication stemming from a project's data set were imposed by only a minority (28%) of projects. This may reflect a consensus that issues viewed by genomics projects as potentially germane to prepublication review, such as research contributor confidentiality, data accuracy, respect of contributors' choices expressed in the informed consent process and overlap in scientific investigations, are sufficiently addressed in alternative governance, security and quality control methods.

Most importantly, however, we found that funding agencies also acknowledged that restrictions on data sharing, particularly in the form of publication moratorium and data retention policies, may be appropriate mechanisms to restrict uninhibited data sharing. We explore these mechanisms below.

Publication moratorium synthesis

Of the genomics projects, 25% contained a publication moratorium policy. Table 1 illustrates the various publication moratorium periods imposed on data users.

Prepublication data retention period synthesis

Tables 2 and 3 illustrate the various prepublication data retention policies. Nearly all genomics projects had a policy regarding data retention, whether it was a 'short'-period policy or guideline stating that data producers must release their data very shortly after it has been produced (Table 2) or a defined period of time. Additionally, data retention periods may not be expressed as a numerical value. Very often, policies or

59. Fortin et al., 'Access Arrangements', p. 104.

Table 1. Publication moratoria policies of genomics institutes or projects

Genomics project (year policy created)	Information regarding publication moratorium
Wellcome Trust Case Control Consortium (WTCCC) (2005)	In accordance with the Fort Lauderdale Agreement, WTCCC asked data users to respect data producers' opportunity to publish initial analyses using the data and to refrain from publishing analyses which used the WTCCC data prior to the initial WTCCC publication or publications. WTCCC did not specify a publication moratorium time period.
Genetic Association Information Network (GAIN) (2006)	GAIN imposes an embargo on publication and presentation of data for 9 months. The GAIN policy is the first genomic data release policy to introduce a time limit on the publication moratorium (as opposed to the temporal release requirements imposed on data producers since the Bermuda Principles).
NIH Genome Wide Association Studies (GWAS) and dbGaP (2007)	The GWAS Policy addresses the publication priority concerns of data producers by stating an expectation that GWAS data users refrain from submitting their analyses and conclusions for publication, or otherwise presenting them publicly, during an 'exclusivity' period of up to 12 months from the date that the data set is made available.
International SAE Consortium (iSAEC) (2007)	The iSAEC imposes a publication moratorium of up to 9 months.
NIH Human Microbiome Project (HMP) (2007)	The NIH Human Microbiome Project Consortium publication policy states that an HMP publication moratorium period for HMP Demonstration Projects is up to 12 months from the date of data submission or upon publication by the data producer, whichever is earlier.
International Human Microbiome Consortium (IHMC) (2007)	The IHMC imposes a publication moratorium of up to 12 months on all data.
Kaiser Permanente Research Program on Genes, Environment and Health (2007)	The RPGEH Access and Collaborations Policy states that investigators will have publication rights for 12 months after the release of data and/or biospecimens. During this time, other investigators may be approved to use data and/or biospecimens for the same hypothesis, but will not be permitted to publish. Investigators will be permitted to apply for extensions of exclusivity for good cause.
The Full ENCODE Project and modENCODE (2008)	The ENCODE Publication Policy uses the Fort Lauderdale Principles terminology in designating itself a 'community resource project', and recommends a 9-month embargo period during which users of released data are requested not to publish or present results based on that data.
MalariaGEN (2008)	MalariaGEN's publication policy imposes a publication moratorium on data users. Their policy states that investigators retain the exclusive right to publish planned analyses of the released data sets for 'a defined period of time, as advised on the MalariaGEN Web site.' The MalariaGEN Web site references the Fort Lauderdale Agreement and states that it adheres to its principles.

(continued)

Table 1. (continued)

Genomics project (year policy created)	Information regarding publication moratorium
1000 Genomes Project (2010)	The 1000 Genomes Project imposes a publication moratorium on data users that does not specify a time limit. Their policy states that data users may use the data for many studies, but are expected to allow the data producers to make the first presentations and to publish the first paper with global analyses of the data.
International Cancer Genome Consortium (ICGC) (2010)	ICGC imposes a publication moratorium on global analyses that lasts until data is published by an ICGC member project or 12 months after a specified quantity of data has been released via the ICGC database or other public databases. In all cases, data will be free of a publication moratorium 2 years after its initial release.
UK10K (inaugurated in 2010)	Expressing its commitment to the principles of rapid data release to the scientific community, the UK10K's Data Sharing Policy imposes a publication moratorium on certain data analyses (here, it covers the description of genetic variants and their use in association tests for the named phenotypes for which the samples were selected into the project). The UK10K publication embargo follows ICGC policy's temporal restriction, whereby all data will no longer be subject to the publication moratorium if the data has been published, or if 12 months passed since the full data set required for analysis was released, and in all cases data will no longer be subject to a publication moratorium 2 years after its initial release.

guidelines spoke to a norm of data deposition in terms similar to 'as rapidly as possible'. Table 3 illustrates genomics projects without data retention periods, policies with defined periods of time and policies expressing non-numerical norms of data deposition.

There does not appear to be any particular association between the modalities of these policies and either the geography or form of funding of these projects. However, several recently constituted projects, such as the International SAE Consortium (iSAEC), inaugurated in 2007, and MalariaGEN, inaugurated in 2008, have enacted detailed data retention policies that specify periods of time (up to 12 and 9 months, respectively) in which data will be withheld from public release. This suggests that, as with publication moratoria policies, genomics projects are imposing specific, defined temporal restrictions on the open access rights of data users with more frequency, although UK10K has not imposed a specific temporal period, instead opting for a general principle of rapid release.

Discussion

In few years since WTCCC became the first genomics project to institute an embargo on publication and presentation of data (in 2005), genomics projects have already

Table 2. Genomics institutes or projects with 'short' prepublication data retention period policies

International HapMap Project (International)
International Cancer Genome Consortium (International)
International Human Microbiome Consortium (International)
1000 Genomes Project (International)
GenomEUtwin (Europe)
UK10K (UK)
Wellcome Trust Sanger Institute (UK)
The Cancer Genome Atlas (USA)
ENCODE & modENCODE (USA)
Genetic Association Information Network (GAIN) (USA)
Human Microbiome Project (USA)
NIH Genome-wide association studies (GWAS) in dbGaP database (USA)
Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) (USA)
Atherosclerosis Risk In Communities (ARIC) Study (USA)
Framingham Heart Study (USA)

introduced a number of modifications and policy gradations. In 2006 and 2007, the six projects identified as having a publication moratorium policy imposed a general moratorium of up to 12 months. After 2008, however, projects introduced variations that sought to nuance this general policy.

For example, the ENCODE Consortia Data Release, Data Use and Publication Policies are more sophisticated than previous data release policies, as they distinguish between published and unpublished data, verified and unverified data and offer several examples of data use implications for different types of studies conducted with ENCODE data. The ENCODE policies suggest that both NIH and the consortium participants desire clear guidelines that avoid misunderstandings regarding the release of data, and also reflect the growing diversity of stakeholders and data beyond those originally considered by the Bermuda Principles.[60]

Similarly, the publication policies of the International Cancer Genome Consortium (ICGC) and UK10K impose a more complex publication moratorium because of their 'rolling' data releases. ICGC prohibits data users from publishing their research based on global analyses of ICGC data until such data is published by an ICGC member project, or 12 months after a specified quantity of data has been released via the ICGC database or other public databases. In all cases, ICGC data will be free of a publication moratorium 2 years after its initial data release. UK10K imposes a similar policy. Both consortia's policies evidence the publication moratorium principle espoused in the 2009 Toronto Statement, which seeks to clarify the Fort Lauderdale Agreement around the 'rights' of first publication for data producers. In doing so, it states that the data users may freely analyse released prepublication data, but they must act responsibly in publishing analyses of those data by, inter alia, respecting the scientific etiquette that allows data producers to publish the first global analyses of their data set/data sets. The consortia's policies also display the

60. Contreras, 'Data Sharing', p. 1602.

Table 3. Genomics institutes or projects that lack or have defined period or non-numerical normative prepublication data retention period policies

Genomics institute or project (Jurisdiction)	Prepublication data retention period
International Serious Adverse Event Consortium (Intl.)	All research results are made available publicly within 12 months of the completion of the study group's genotyping and data quality control efforts.
Type 1 Diabetes Genetics Consortium (Intl.)	Reporting to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repository must occur within 12 months after receipt of samples and/or data.
MalariaGEN (Intl.)	Each data set will be made available 9 months after MalariaGEN investigators at the local study site first have access to that data set. Where principal investigators from the study site agree, data may nevertheless be released immediately along with notification of areas of research the MalariaGEN Network and individual principal investigators are undertaking with the data set.
European Genome-phenome Archive (EGA) (Intl.)	No specific policy, but generally supports data access that is restricted to named consortium members prior to publication; typically 6–12 months pre-publication.
UK Data Banking Network	No specific policy – only confidential access by UDBN to data on all tests performed on any sample requested from the UDBN archive within 10 days of acceptance for publication reporting on any of the tests on any of the samples.
ALSPAC (UK)	No specific policy. Users of ALSPAC data must provide ALSPAC with a fully documented electronic copy of the full results before publication in any form, or within 6 months of the completion of the research, whichever is sooner.
Generation Scotland (UK)	No specific policy. Users of Generation Scotland data must, where requested to do so, provide Generation Scotland with a fully documented electronic copy of any data prior to publication in any form or within 6 months of the completion of the research, whichever is sooner.
UK Biobank	No specific policy. Users of UK Biobank data must endeavour to publish the findings within 6 months of representing to UK Biobank that the research project will be completed in either an academic journal or on an open source publication site. Within 6 months of publication, or 12 months of when the research project was to be completed, users are required to provide the results of the research and the raw data behind them, for inclusion in UK Biobank (subject to requests for reasonable extension).
Wellcome Trust Case-Control Consortium (WTCCC) (UK)	Data generated from WTCCC will be made available 6 months after data generation and quality control is complete, or upon publication, whichever is sooner.
Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH) (USA)	No specific policy – users of RPGEH data are obligated to furnish a full written report containing all research results promptly upon study completion.

(continued)

Table 3. (continued)

Genomics institute or project (Jurisdiction)	Prepublication data retention period
NIH/NCBI GenBank (USA)	No specific policy – there is an opportunity to keep data confidential for a specific period of time, until publication.
NIH/NCBI dbSTS (USA)	No specific policy – data can be withheld from public view until the accession number is published.
NIH/NCBI dbMHC (USA)	No specific policy.
NIH/NCBI Gene Expression Omnibus (GEO) (USA)	Submissions may be held private for a maximum allowable limit of 3 years, until publication; this date may be brought forward or pushed back at any time.
NIH/NCBI SKY/M-FISH and CGH Database (USA)	Maximum of 2 years.
Marshfield Clinic Personalized Medicine Research Project (USA)	All investigators are expected to return data and analyses to the PMRP database for other investigators to use within 6 months after final data analysis.
CARTaGENE (Canada)	No specific policy – the time period will be determined by the Samples and Data Access Committee and the Executive Committee on a project-by-project basis.
Helmholtz Zentrum München (German Research Center for Environmental Health)	No specific policy.
LifeGene (Sweden)	Non-numerical normative statement ('within a reasonable time limit').
EpiHealth (Sweden)	Non-numerical normative statement ('for a limited and reasonable period').
DNA Databank of Japan	No specific policy – at discretion of data producer but released when (1) data producer requests to release the data; (2) data producer has published own accession number/numbers and it has been confirmed; or (3) specified hold-date has come.
Singapore Bio-Bank	No specific policy.

increasing level of sophistication governing publication moratorium policies, as certain types of data may be subjected to a publication moratorium, whereas others are subjected to no moratorium.

Given that only two large funding agencies were identified that have a general, institution-wide policy regarding publication moratoria (the United Kingdom's MRC and the Wellcome Trust), and both provide broad support for a reasonable, but not unlimited, period of exclusive use for the research data that data producers generate, it is unsurprising that the managers of large projects have committed themselves to a self-governing regime[61] centred around international consensus statements and similarly drafted publication policies. Indeed, another identified point of convergence is

61. M.T. Mayrhofer and B. Prainsack, 'Being a Member of the Club: The Transnational (Self-) Governance of Networks of Biobanks', *International Journal of Risk Assessment and Management* 12 (2009), pp. 64–81.

that it is the large, often internationally focused genomics projects that have instituted these publication moratorium periods. This may well reflect the fact that these budding, collaborative entities contain large volumes of data that a wide variety of international researchers wish to access. Erecting robust policies to protect data producers who contribute to a large genomics project's success may be viewed as a necessity for the long-term health and success of the project itself.[62]

Conversely, few of the smaller genomics projects employ publication moratoria policies because they are usually not as internationally collaborative.[63] Furthermore, these projects may not have the resources to develop and implement data sharing plans.[64] One can speculate that these smaller projects may keep their data confidential, subject to informal exchanges, rather than employing an open data sharing and publication moratorium policy. However, the environment is more nuanced than a simple 'large-scale' versus 'small-scale' genomics project dichotomy indicates. Many funding agencies now require data sharing plans for investigator-initiated grant applications, and many peer-reviewed journals require evidence of sequence submission to data sets, such as an accession number, prior to publication.

The youth of publication moratorium policies should be contrasted with the mature data retention policies regarding release of data into a greater knowledge commons, be it a publicly accessible data set or a controlled access data set overseen by a data access committee. These data retention policies may need to be revised since the literature and international consensus statements clearly espouse a belief in the rapid dissemination of prepublication data. Furthermore, this principle has broadened from genomic data[65] to 'community resource projects'[66] to proteomics data[67] to all biomedical data sets,[68] and, most recently, to stem cell science[69] and public health research.[70]

62. Kaye et al., 'Re-shaping', p. 331.
63. Gitter, 'Challenges', p. 623; Gitter, 'Application', p. 476.
64. Dyke and Hubbard, 'Developing', p. 60; Heeney and Smart, 'Enacting Governance', p. 245.
65. International Strategy Meeting, 'Bermuda Principles' Available at: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1 (accessed 8 July 2012).
66. The Wellcome Trust, 'Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility. Report of a Meeting Organized by the Wellcome Trust, 14–15 January 2003, Fort Lauderdale, USA' Available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf (accessed 8 July 2012).
67. H. Rodriguez et al., 'Recommendations from the 2008 International Summit on Proteomics Data Release and Sharing Policy: the Amsterdam principles', *Journal of Proteome Research* 8 (2009), pp. 3689–3692.
68. Toronto International Data Release Workshop Authors et al., 'Prepublication Data Sharing', *Nature* 461 (2009), pp. 168–170.
69. The Hinxton Group, 'Statement on Policies and Practices Governing Data and Materials Sharing and Intellectual Property in Stem Cell Science' Available at: http://www.hinxtongroup.org/Consensus_HG10_FINAL.pdf (accessed 10 July 2012).
70. M. Walport and P. Brest, 'Sharing Research Data to Improve Public Health', *Lancet* 377 (2011), pp. 537–539.

Yet, the growing importance of open sharing has somewhat paradoxically led to longer and more sophisticated data retention policies that reflect context-dependent facts. For example, it is common for privately funded projects or consortia to employ a 'club good'-type variant of a data retention policy by making their data available only to collaborators, who may need to sign a data transfer agreement in advance. We find, at the same time, a large number of projects identified in our research that ascribe to a 'short' data retention period. That is, data producers are obligated to rapidly release pre-publication data into the knowledge commons, in accordance with the international consensus statements identified in Appendix 3. Generally, the reason for a short data retention period is to give sufficient time to genomics projects and producers to ensure quality control of the data. This policy has held sway since the first large-scale genomic studies were created and it continues to apply to ongoing genomic research projects. On the other hand, we acknowledge that not all genomics projects ascribe to the 'short' data retention period, particularly if they face less funding agency requirements or are tied to developing regions. Where a project's policy-making body decides to move beyond a short retention period, it is usually to protect the data producers 'rights'. Indeed, MalariaGEN and iSAEC impose quantifiable data retention periods, which may be viewed as a means to provide participating researchers with the requisite time to analyse the data and generate publications.[71]

One possible explanation for this may be that the organizational structure of the project dictates the data retention policy implemented. Commentators have noted that funding agency policies can be a critical factor in how genomics projects enact strategies to restrict or promote access to their data.[72] The iSAEC is a nonprofit organization largely financially and scientifically supported by pharmaceutical companies that identify and validate DNA-variants useful in predicting the risk of rare, drug-induced serious adverse events. The research funded by iSAEC is probably not a 'community resource project' as defined in the Fort Lauderdale Agreement since the Consortium's goal is not the creation of a large, generally applicable data set, but it has still committed to release its data to the public, albeit on a delayed basis. Unlike publicly funded genomics projects that may be subject to funding policies or guidelines that mandate rapid data release, iSAEC is able to enact a robust policy that affords data producers maximum protection (including for potential intellectual property rights, which are critical for pharmaceutical companies), although it might impede open data integration. Professor Jorge Contreras notes that data retention gives a data producer a veritable 'head start' with respect to the data, during which time no other party may access or analyse the data, whereas a publication moratorium gives data users the ability to analyse and build upon the data during the moratorium period.[73] By imposing a 12-month data retention period, iSAEC is able to protect its data producers from any potential breaches of data usage. After all, if data users cannot access the data, there is no possibility of breaching a publication moratorium policy.

71. A.L. Holden, 'The Innovative Use of a Large-Scale Industry Biomedical Consortium to Research the Genetic Basis of Drug Induced Serious Adverse Events', *Drug Discovery Today: Technologies* 4 (2007), pp. 75–87.
72. Fortin et al., 'Access Arrangements', p. 104; Tenopir et al., 'Data Sharing', p. e21101.
73. Contreras, 'Data Sharing', p. 1602.

Another possible explanation may be that the executive, ethics or data access committee members of a given project may be more inclined to address the concern of 'scooping' by data users, particularly when the data or the project may be seen as sensitive or tied to socioeconomic and scientific development. MalariaGEN's policy states that each data set will be made available 9 months after MalariaGEN investigators at the local study site first have access to that data set. MalariaGEN, which from the outset of its existence established an in-house ethics team, adopted this data retention policy out of concern that genomic data produced by researchers in developing countries could be 'scooped' by those from richer countries.[74] Unlike iSAEC, MalariaGEN is an entirely publicly funded consortium and thus must follow the data sharing policies and/or guidelines set by its numerous funders. This may explain why its data retention period is shorter than iSAEC's. Yet, because the project is largely based in developing areas of the world, especially sub-Saharan Africa, there is a strong desire to impose some period greater than an undefined 'short' length of time so as to promote the development of scientific infrastructure and capacities. Indeed, a MalariaGEN press release confirmed that the 9-month window is 'intended to build scientific capacity in developing countries without compromising the team's commitment to early, open data access'.[75]

These examples demonstrate the difficulty in attempting to distil broad generalizations about publication moratoria or data retention policies for genomics projects. However, the results of our analysis lend credence to the view that as more collaborative genomics projects are created and further genomic research discoveries are announced, the need for more nuanced policies that consider the views of all stakeholders will increase.[76] To aid in the development of these nuanced policies, it is necessary to conduct more extensive surveys on data sharing policies in other parts of the world, which were not covered in this analysis.

In particular, analysing data retention policies or publication moratoria policies in Asian countries, where genomics research has been rapidly expanding, is important. There has been some recent discussion about data sharing policies in the Japanese genomics research community.[77] A 2007 survey of 1200 researchers in Japan found that a quarter had delayed publication of research results, and 19% reported difficulties in gaining access at least once in the past 5 years, although almost three-quarters reported that access was becoming easier.[78] This emerging evidence aside, there remains a dearth of

74. Parker et al., 'Ethical Data Release', p. e1000143.
75. MalariaGEN, 'MalariaGEN develops data policies for developing country GWAS' Available at: http://www.malariagen.net/node/223 (accessed 9 July 2012).
76. Milanovic et al., 'Biobanking and Data Sharing', p. 17.
77. K. Takahashi and K. Kato, 'Importance of GWAS Data Sharing and Public Dialogue in Human Genome Research', *Igaku no ayumi* or *Journal of Clinical and Experimental Medicine* 225 (2008), pp. 891–894 (in Japanese).
78. J.P. Walsh and H.I. Huang, *Research Tool Access in the Age of the IP Society: Results from a Survey of Japanese Scientists. A Report Prepared for the Project on Science and Intellectual Property in the Public Interest* (Washington, DC: American Association for the Advancement of Science, 2007).

literature substantively discussing these topics. In order to maximize the global benefits of genomics, the genomics research community must work together to facilitate data sharing in all parts of the world.

Limitations

Our policy review is subject to some limitations that may confine the generalization of the findings. First, although we used a structured approach to review policies and literature, as with all studies, it is subject to error and bias, including language bias. Indeed, a second limitation is that genomics projects that did not publish data release policies in English are not represented. Third, our web-based method of analysis may give the false impression that genomics projects not included in the review lack a data release policy. However, we note that internet access is likely the mode of access data users would use if, for example, a data user wished to understand how to access the projects' data. Fourth, we purposely employed a broad definition of 'genomics projects' so as to obtain a large sample that would enable a robust analysis, although some may view the definition as overly inclusive.

Conclusions and outlook

This article surveyed data release policies for genomics projects across the international spectrum, with the goal of identifying points of convergence and divergence. Our analysis of projects and major research institution/funding agency policies, joined with a critical review of the literature, reveals that the rapid evolution of genomics research and biobanking has led to increasingly sophisticated policies that seek to balance the interests of various stakeholders. This is particularly evident in the publication moratorium policies established in large genomics projects. We consider below likely developments regarding three aspects of data sharing policies: data retention, publication moratoria and other potential models.

Data retention

We consider data retention to generally run counter to the open science principle and difficult to justify only on the basis of protecting data producers' priority rights in a publicly funded genomics project.[79] Only in limited circumstances, such as a developing country-focused consortium like MalariaGEN for developing world infrastructure and capacity improvement, could this imposition on open data sharing be justified.

At the same time, we acknowledge the need for a realistic and pragmatic perspective: 'while data sharing may be in the interests of society, in the competitive world of scientific research, data sharing does not just happen'.[80] As Smith and colleagues point out,

79. Dyke and Hubbard, 'Developing', p. 60.
80. Tenopir et al., 'Data Sharing', p. e21101.

and as confirmed by empirical study,[81] researchers recognize the benefits of sharing data, yet not all researchers wish to share their own work; after all, sharing requires much more effort than not sharing.[82] On a financial level, open data sharing costs money in order to turn the data into a suitable form that can be loaded on accessible platforms.[83] On a logistics level, it is not always clear when the clock starts ticking on data release; instantaneous data sharing without repetitive rounds of cleaning, annotation and validation could jeopardize the quality of data in the public domain and do a disservice to data users and the public at large.

Better education on data sharing and analysis could facilitate a more collaborative data sharing culture,[84] but an unadulterated belief in altruistic data sharing – absent any incentive – may be precarious[85] and an untenable bridge in the march towards open science. Strategies other than data retention that create less friction with open science principles, but are still cognizant of the competitive world of scientific research and the logistical and financial aspects of data sharing, should be strongly encouraged.

Publication moratoria

Moratoria policies are powerful self-regulatory instruments that can be applied not only to publications but also to research itself: witness the 60-day moratorium on H5N1 research in January 2012[86] and the 1975 Asilomar Conference on Recombinant DNA.[87] This imparts overtly sociolegal and political dimensions to science and technology issues. With respect to publications, an interesting solution is that of time-limited publication moratoria periods that are limited to specific genomic analyses, although additional evidence is needed to assess its overall effectiveness.[88] Data users themselves appear to support publication moratoria over data retention policies,[89] since they are able to work on data analyses in real-time, as opposed to a data dissemination delay of several months that would retard research and validation by peers.

---

81. International Strategy Meeting, 'Bermuda Principles' Available at: http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1 (accessed 8 July 2012); Fort Lauderdale Agreement, Available at: http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtd003207.pdf (accessed 8 July 2012); Toronto International Data Release Workshop Authors et al., 'Prepublication Data Sharing', p. 168.
82. A. Smith et al., 'Biology and Data-Intensive Scientific Discovery in the Beginning of the 21st Century', *OMICS: A Journal of Integrative Biology* 15 (2011), pp. 209–212.
83. Heeney and Smart, 'Enacting Governance', p. 247.
84. Piwowar et al., 'Data Sharing Culture', p. e183.
85. Bezuidenhout, 'Data Sharing', forthcoming.
86. R.A. Fouchier et al., 'Pause on Avian Flu Transmission Research', *Science* 335 (2012), pp. 400–401.
87. P. Berg et al., 'Summary Statement of the Asilomar Conference on Recombinant DNA Molecules', *Proceedings of the National Academy of Sciences USA* 72 (1975), pp. 1981–1984.
88. Dyke and Hubbard, 'Developing', p. 60.
89. Silverstein et al., 'Commercialization', p. 131.

Genomics projects continue to increase in size and scope across national boundaries, rapidly displacing extant frameworks: 'data production evolves in magnitude and type; by the time a large project publishes its analysis, the state of the art is usually a larger or different type of data set'.[90] Consequently, genomics projects will need to introduce further publication moratorium policy subtlety.

There are also many questions to explore. For example, if a publication moratorium policy is defined by the submission of a certain amount of data, how does one decide upon suitable quantification? Does the embargo period begin anew with each submission? Is the confusion over the timing of an embargo period with ongoing data release simply the cost of encouraging early data release (i.e. not waiting for complete data sets before release)? Whose responsibility is it to tell the data access committee or data coordination centre the current state of the data? Should a committee even be charged with this task, as opposed to the data producers? Will genomics projects have sufficient infrastructure and capacity to maintain publication moratoria policies, given that they are difficult to manage and entail significant costs and other resources? What happens if data users fail to respect the moratorium? More broadly, how will data producers react if a moratorium that commences at the first submission runs out before they have published their analysis? In light of these challenges, further developments are needed to support the publication moratorium system in the long term, including further standardization efforts, and we therefore also suggest envisioning other models to reward data sharing as a way forward.

Other potential models

As data-intensive science has now been christened the 'fourth paradigm of science',[91] open access to a data 'commons' is a prerequisite for discoveries in a 21st century science ecosystem.[92] Self-regulated, community norm-driven reward systems should be implemented that recognize genuine contributions from data producers, data users and other stakeholders such as research contributors and funding agencies, without unduly affecting open science principles.[93] In recognition of the desire for an appropriate balance between risk/reward and 21st century data-intensive science requirements, and given the concerns about data retention periods and the structural challenges to implementing and enforcing publication moratoria on 'rolling' genomics projects of increasing scale and complexity, other complementary methods that reward researchers for sharing data should be explored.

Recognition and reward systems should promote data sharing and meet the expectations of data producers. This should be encouraged, as a recent survey indicates that

90. Nature Genetics Staff, 'Data Producers Deserve Citation Credit', *Nature Genetics* 41 (2009), p. 1045.
91. T. Hey et al., eds., *The Fourth Paradigm: Data-intensive Scientific Discovery* (Redmond, Washington: Microsoft Research, 2009).
92. V. Özdemir et al., 'Policy and Data-Intensive Scientific Discovery in the Beginning of the 21st century', *OMICS: A Journal of Integrative Biology* 15 (2011), pp. 221–225.
93. Özdemir et al., 'Policy', p. 221.

younger researchers (ages 20–39 years) are less likely than older researchers to agree to share their data without restrictions, but are more likely to share if restrictions are in place.[94] Some may view this as a disappointing prospect, but given that a majority of these younger researchers also agree that scientific progress is inhibited by a lack of access to data,[95] it could mean that they simply require more robust motivations and systems to fully embrace data sharing. Just as data producers often appreciate publication moratoria and data retention periods to obtain recognition for their contributions and effort invested, models can be devised that provide more incentive and credit for contributing to genomic data sets. To address but one area of many, we think that data producers should be recognized and rewarded for sharing their data through open access platforms. If reward systems were sufficiently modified, publication moratoria policies may not seem as important to data producers.

Nonetheless, operationalizing the Bermuda Principles, Fort Lauderdale Agreement and Toronto Statement norms into the current fourth paradigm of science may be as much art as science. Any successful, prospectively negotiated, and, ideally, global model[96] that simultaneously promotes equitable access, open data sharing and accommodation for data producers' desires must be the result of at least four key factors: (1) open and transparent dialogue among all stakeholders,[97] (2) the consideration and incorporation of varying perceptions, contexts and practices of researchers worldwide,[98] (3) infrastructural support and investment by funding agencies to facilitate openly accessible data sets, and (4) evidence-based studies on current problems, the efficacy of current 'community resource' standards and the efficiency of proposed revisions.

Several potential alternative models that work to address these factors are emerging. Myles Axton of *Nature* and Giardine and colleagues have proposed a 'microattribution' model where data curators' and producers' contributions are traced and recognized down to the very smallest meaningful unit (e.g. database record, gene), as opposed to the traditional method of whole articles.[99] Mons and colleagues have proposed a 'Concept Web' and 'semantically coded nanopublications' model that could more suitably represent the relationships between research data and efficient exchange of knowledge.[100]

94. Tenopir et al., 'Data Sharing', p. e21101.
95. Tenopir et al., 'Data Sharing', p. e21101.
96. D. Field et al., 'Omics Data Sharing', *Science* 326 (2009), pp. 234–236.
97. Nature Genetics Staff, 'Data Producers', p. 1045; A.L. McGuire et al., 'Ethical and Practical Challenges of Sharing Data from Genome-Wide Association Studies: The eMERGE Consortium Experience', *Genome Research* 21 (2011), pp. 1001–1007; M.W. Foster and R.R. Sharp, 'Share and Share Alike: Deciding How to Distribute the Scientific and Social Benefits of Genomic Data', *Nature Review Genetics* 8 (2007), pp. 633–639.
98. P. Boddington, 'Data Sharing in Genomics', p. 205; Singh and Daar, 'Intra-Consortium', p. 11.
99. Gitter, 'Application', p. 476; Nature Genetics Staff, 'Human Variome Microattribution Reviews', *Nature Genetics* 40 (2008), p. 1; B. Giardine et al., 'Systematic Documentation and Analysis of Human Genetic Variation in Hemoglobinopathies Using the Microattribution Approach', *Nature Genetics* 43 (2011), pp. 295–301.
100. B. Mons et al., 'The Value of Data', *Nature Genetics* 43 (2011), pp. 281–283.

Although bioresources are not synonymous with data as they are easier to track, one potential model currently being investigated is the BioResource Impact Factor (BRIF), which seeks to provide a global register for databases and allows for an operational approach to promote data sharing. BRIF would allow a unique identifier to cite and acknowledge the use of bioresources in publications and thereby measure their impact; more importantly, it would assist groups in tracking submissions and data release.[101] Another model is the Open Researcher and Contributor ID (ORCID) researcher identifier, which proposes to use data citation and associated metrics for data sets as a means to recognize bioresources, individuals and their roles and consequently promote data sharing.[102] In these models, exploration and discussion of the criteria necessary to determine the kinds of data that would be recorded and counted in citations should be considered. It will also be useful to monitor how these models, which are international in flavour and orientation, will mesh with reward and promotion systems for researchers that are generally determined by local university or research institute policies. Obtaining traction with these models on a larger stage will take time and effort, and may particularly pose hurdles for international consortia that must satisfy the needs of data producers who come from diverse and localized academic backgrounds. Nevertheless, BRIF and ORCID may signal the commencement of nuanced, balanced data sharing tools for 21st century genomic science, although given their embryonic state, their evolution should be monitored closely.

Ultimately, the translational promise of all genomics projects depends on communication of expectations, ongoing dialogue, empirical analyses and analytical innovations. The BRIF and ORCID models discussed above are promising, as they offer a natural incentive to researchers with little impediment to open science principles, although further exploration, development and testing of these models should be encouraged. Once a given model has been appropriately deliberated, adopted and implemented into a policy or guideline, it should, like the data it regulates, be put into the public domain. This would signal, visibly and symbolically, the values of openness and collaboration that are increasingly becoming the bedrock of a 21st century genomic research ecosystem.

Authors' Note

The first two authors (YJ and ESD) conceived of this article and contributed equally to it. All of the authors, who comprise the entire membership of the International Cancer Genome Consortium's Publication Moratorium Subgroup, drafted the article, revised it critically for important intellectual content, and approved the final version submitted for publication.

101. A. Cambon-Thomsen et al., 'The Role of a Bioresource Research Impact Factor as an Incentive to Share Human Bioresources', *Nature Genetics* 43 (2011), pp. 503–504; J. Kaye, 'From Single Biobanks to International Networks: Developing e-Governance', *Human Genetics* 130 (2011), pp. 377–382.
102. Cambon-Thomsen et al., 'Bioresource', p. 503; A.J. Webb et al., 'An Informatics Project and Online 'Knowledge Centre' Supporting Modern Genotype-to-Phenotype Research', *Human Mutation* 32 (2011), pp. 543–550; Nature Genetics Staff, 'Standard Cooperating Procedures', *Nature Genetics* 43 (2011), p. 501.

Funding

All of the authors are grateful for the funding support of the Ontario Institute of Cancer Research (OICR). The funding for this work was also provided by the Public Population Project in Genomics (P^3G) and Society by Genome Canada and Genome Quebec, the Canadian Institutes of Health Research (CIHR), the Ministère du Développement économique, de l'Innovation et de l'Exportation (MDEIE) of Quebec, the Wellcome Trust and the European Network for Genetic and Genomic Epidemiology (ENGAGE) by the European Commission under the 7th Framework Programme-Health Theme.

Acknowledgments

The authors thank Bradley Ozenberger (National Human Genome Research Institute, NIH) for his helpful and insightful comments.

Conflict of interests

The authors declared no conflicts of interest.

Appendix 1

List of genomics institutes or projects surveyed for publication moratorium policies

1000 Genomes Project.
Atherosclerosis Risk In Communities (ARIC) Study.
Australasian Biospecimen Network.
Avon Longitudinal Study of Parents and Children (ALSPAC).
Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC).
Biomarkers Consortium.
Canadian Partnership for Tomorrow Project (CPTP).
CARTaGENE.
Cohort of Norway (CONOR).
DNA Databank of Japan.
ENCODE and modENCODE.
EpiHealth.
Estonian Genome Project.
European Genome-phenome Archive (EGA).
Framingham Heart Study.
Generation Scotland.
Genetic Association Information Network (GAIN).
GenomEUtwin.
Helmholtz Zentrum München (German Research Center for Environmental Health).
Human Microbiome Project.
Integrated BioBank of Luxembourg.
International HapMap Project.
International Cancer Genome Consortium (ICGC).

International Human Microbiome Consortium.
International Serious Adverse Event Consortium (iSAEC).
Kaiser Permanente Research Program on Genes, Environment and Health (RPGEH).
Karolinska Institute Biobank.
King's College London Infectious Diseases BioBank.
KORA-gen.
LifeGene.
LifeLines.
MalariaGEN.
Marshfield Clinic Personalized Medicine Research Project.
NIH Genome-wide association studies (GWAS) in dbGaP database.
NIH/NCBI GenBank.
NIH/NCBI dbSNP.
NIH/NCBI dbSTS.
NIH/NCBI dbMHC.
NIH/NCBI Gene Expression Omnibus (GEO).
NIH/NCBI SKY/M-FISH and CGH Database.
Norwegian Mother and Child Cohort Study (MoBa).
Norwegian Twin Registry.
Ontario Tumour Bank.
Singapore Bio-Bank.
String of Pearls Initiative.
The Cancer Genome Atlas (TCGA).
Type 1 Diabetes Genetics Consortium.
UK10K.
UK Biobank.
UK Data Banking Network.
Wellcome Trust Case-Control Consortium (WTCCC).
Wellcome Trust Sanger Institute.
Western Australian DNA Bank.

Appendix 2

List of government research institutions or funding agencies (both government and nongovernment funded) that contain recommendations on the governance of genomics projects

Biotechnology and Biomedical Sciences Research Council (UK).
Canadian Institutes of Health Research.
Genome Canada.
Medical Research Council (UK).
National Health and Medical Research Council (Australia).
National Institutes of Health (USA).
Wellcome Trust (UK).

Appendix 3.

International consensus statements addressing publication moratoria and/or data retention policies

Article/statement	Year released	Position/positions on data retention	Position/positions on publication moratorium
NIH/DOE Guidelines	1992	A period of 6 months from the time data or materials are generated to the time they are made publicly available is a reasonable maximum, although more rapid sharing is encouraged	None
Bermuda Principles	1996	• Sequence assemblies greater than one kilobase (kb) in length (i.e. 1000 base pairs) should be released automatically within 24 h • Finished annotated sequences should be submitted immediately to a public database	None
Fort Lauderdale Agreement	2003	• 24-h rapid release rules of Bermuda Principles should apply to each 'community resource project' (CRP) • Rapid release may not be appropriate for non-CRPs, such as hypothesis-driven research	None
OECD Principles and Guidelines for Access to Research Data from Public Funding	2007	Research data from public funding should be openly and timely released	• Access arrangements should promote explicit, formal institutional practices regarding the responsibilities of the various activities in data-related activities • Consideration should be given to publication agreements as a means of setting out responsibility for data access and management

(continued)

Appendix 3. (continued)

Article/statement	Year released	Position/positions on data retention	Position/positions on publication moratorium
Amsterdam Principles	2008	• Rapid data release should be extended to proteomics data • Timing should depend on the nature of the effort generating the data and should take into account the legitimate concerns of producers • Data generated by individual investigators should be released into the public domain at the latest upon publication • Data generated by community resource projects should be released following appropriate QA/QC procedures	None
Toronto Statement	2009	• Rapid prepublication data release should include all biomedical data sets having 'broad utility, are large in scale . . . and are "reference" in character.' • Rapid data release 'should not be mandated' for hypothesis-driven projects	'Protected time period' during which data users can be restricted from publishing on released data sets, but period should 'ideally expire' within 1 year
Hinxton Group Statement on Policies and Practices Governing Data and Materials Sharing and Intellectual Property in Stem Cell Science	2010	Funders, research institutions and journals should insist on the deposition of data into databases and 'hubs', with release on (or in a specified time after) publication	None
Wellcome Trust Funders Statement	2011	• Rapid data release should be extended to public health research • Data sharing approaches should balance the needs of data producers, data users, communities, and funders	None

NIH: National Institutes of Health; DOE: Department of Energy; CRP: community resource project; OECD: Organization for Economic Co-operation and Development.