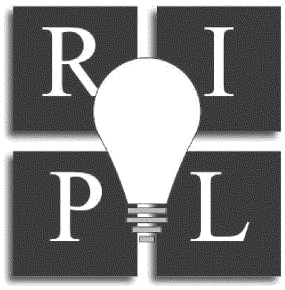
THE JOHN MARSHALL
REVIEW OF INTELLECTUAL PROPERTY LAW




# THE APPLICATION OF DATA ACCESS POLICIES DESIGNED FOR GENOME-WIDE ASSOCIATION STUDIES TO SMALLER SCALE DATABASES

DONNA M. GITTER

ABSTRACT

Scientific progress thrives with open discussion of new ideas and supporting data. To this end, researchers traditionally publish their results in scientific papers—papers that contain the new ideas and the underlying data supporting those ideas. With the advent of large-scale and high-throughput data analysis, however, the creation of scientific databases have replaced the traditional model. For such publically-funded, data-intensive projects, funding agencies typically require that all relevant data be made available on a publicly accessible website at the time of the paper's publication. Against the backdrop of the public accessibility model used in the 1000 Genomes Project, the author recommends that a modified framework be applied to smaller scale data collection projects. Such a framework could overcome the data producers' concern for protecting the data they have created, thereby encouraging researchers to share data from smaller scale studies.



*Cite as* Donna M. Gitter, *The Application of Data Access Policies Designed for Genome-Wide Association Studies to Smaller Scale Databases*, 10 J. MARSHALL REV. INTELL. PROP. L. 476 (2011).

# THE APPLICATION OF DATA ACCESS POLICIES DESIGNED FOR GENOME-WIDE ASSOCIATION STUDIES TO SMALLER SCALE DATABASES

DONNA M. GITTER

# THE APPLICATION OF DATA ACCESS POLICIES DESIGNED FOR GENOME-WIDE ASSOCIATION STUDIES TO SMALLER SCALE DATABASES

DONNA M. GITTER*

## I. INTRODUCTION

Scientific discourse and progress depend upon the open discussion of ideas and full disclosure of supporting facts.[1] This discussion has traditionally occurred through the process of publication, which is the primary means by which researchers achieve recognition for their work.[2] Traditionally, researchers published papers that combined in one entity both their ideas and the underlying data.[3] With the advent of large-scale and high-throughput data analyses, however, the creation of scientific databases replaced the traditional model.[4] Typically, for such data-intensive projects, funding agencies require that all relevant data must be made available on a publicly accessible website at the time of the paper's publication.[5]

The Human Genome Project ("HGP"), completed in 2003, demonstrated to the scientific community that making data broadly available before publication results in valuable benefits to the public.[6] This is particularly true where there is a community of scientists who can use the data more quickly than the data producers themselves, and in ways not originally anticipated at the outset of the project.[7] One successor to the HGP is the 1000 Genomes Project, which provides that project data will be released quickly, prior to publication, into the public domain.[8]

While the open access approach of the 1000 Genomes Project is the norm for large-scale, publicly-funded genomic databases, many smaller projects that likewise produce vast amounts of scientific data nonetheless do not embrace data-sharing.[9]

---

* © Donna M. Gitter 2011. Associate Professor of Law, Baruch College, New York, New York. J.D., University of Pennsylvania Law School; B.A., Cornell University College of Arts and Sciences. E-mail: Donna.Gitter@Baruch.cuny.edu. I would like to thank two colleagues at The Australian National University College of Law for encouraging me to develop this work: Dr. Matthew Rimmer, Associate Professor, Associate Director of Research, and Associate Director, The Australian Centre for Intellectual Property in Agriculture; and Alison McLennan, Vice Chancellor's Scholar. A version of this work will appear in a book they have edited, entitled *Intellectual Property and Emerging Biotechnologies,* which will be published in 2011. I would also like to thank the editors of the *Review of Intellectual Property Law* of The John Marshall Law School for inviting me to contribute this work to their symposium "Biotechnology and Health-Related Issues in IP Law."

[1] Ewan Birney et al., *Prepublication Data Sharing*, 461 NATURE 168, 168 (2009).

[2] Paul N. Schofield et al., *Post-publication Sharing of Data and Tools*, 461 NATURE 171, 171 (2009).

[3] Birney et al., *supra* note 1, at 168.

[4] *Id.*

[5] *Id.*

[6] *Id.*

[7] *Id.*

[8] *See About the 1000 Genomes Project*, 1000 GENOMES, http://www.1000genomes.org/about#ProjectOverview (last visited Mar. 25, 2011).

[9] J.H. Reichman & Paul F. Uhlir, *The Public Domain: A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, 66 LAW & CONTEMP. PROBS. 315, 344 (2003).

Quite often, researchers involved in "small science," meaning research "performed by individual investigators or small and autonomous research groups operating outside large, organized research programs, often with non-federal sources of funding," are not as influenced by the norms of open access and sharing and instead rely on informal exchanges of data and samples.[10] Scientific researchers, ethicists, lawyers, representatives of funding agencies, and journal editors are all actively engaged in discussions aimed at establishing new scientific norms in order to overcome reluctance to data-sharing in the research community.[11] Close investigation of the 1000 Genomes Project provides guidance in terms of applying the data-sharing lessons learned from large-scale, publicly-funded databases to smaller projects.[12]

## II. THE 1000 GENOMES PROJECT: GOALS AND CHALLENGES

Launched in January 2008, the 1000 Genomes Project aims to develop a new map of the human genome that will provide an extraordinarily detailed view of biomedically relevant DNA variations among individuals.[13] A clearer understanding of these small genetic variations will assist in explaining individual differences in susceptibility to disease, responses to drugs, and reactions to environmental factors.[14] The 1000 Genomes Project builds upon the HapMap Project, a large-scale, publicly-funded genomic project that established in 2005 a public database of common genetic variants in human beings.[15] Using HapMap data and related resources, researchers have succeeded in identifying more than one hundred regions of the genome containing genetic variants that are associated with risks of common human diseases such as diabetes, coronary artery disease, and prostate and breast cancer.[16] The goal of the 1000 Genomes Project is to study these genetic differences in greater depth, thereby increasing the ability of researchers to discern the causal relationships between genetic variations and human disease.[17] Ultimately, the 1000 Genomes Project will expedite efforts to diagnose, treat, and prevent human diseases.[18]

Like the HapMap Project, the 1000 Genomes Project relies upon a combination of public and private support, along with the expertise of an international consortium of research teams.[19] The research teams for the 1000 Genomes Project will sequence the genomes of at least one thousand people worldwide in order to create a highly detailed database cataloguing human genetic variation.[20] As with other large-scale, publicly funded genomic projects, data from the 1000 Genomes Project will be made

---

[10] *See id.* at 322–23.

[11] *See* Birney et al., *supra* note 1, at 168–69.

[12] *See id.* at 169.

[13] *International Consortium Announces the 1000 Genomes Project*, NAT'L INSTS. HEALTH (Jan. 22, 2008), http://www.nih.gov/news/health/jan2008/nhgri-22.htm.

[14] *Id.*

[15] *See generally* Int'l HapMap Consortium, *A Haplotype Map of the Human Genome*, 437 NATURE 1299 (2005) (analyzing data collected from phase one of the HapMap Project).

[16] *International Consortium Announces*, *supra* note 13.

[17] *See About the 1000 Genomes Project*, *supra* note 8.

[18] *International Consortium Announces*, *supra* note 13.

[19] *Id.*; *see Three Sequencing Companies Join 1000 Genomes Project*, NAT'L INSTS. HEALTH (June 11, 2008), http://www.nih.gov/news/health/jun2008/nhgri-11.htm.

[20] *Three Sequencing Companies Join*, *supra* note 19.

available as rapidly as possible to the international scientific community via release into freely accessible public databases.[21] The quantity of data will be immense, posing great challenges for the experts in the fields of bioinformatics and statistical genetics.[22] According to Dr. Gil McVean of the University of Oxford in England, one of the co-chairs of the consortium's analysis group: "At six trillion bases, the 1000 Genomes Project will generate sixty-fold more sequence data over its three-year course than have been deposited into public databases over the past twenty-five years."[23] Certainly, the 1000 Genomes Project Consortium faces a significant challenge in devising the most effective way to analyze this overwhelming quantity of data.[24] What is more, the consortium must ensure that its data access policy will make the data available to the maximum number of users, while simultaneously permitting researchers to seek intellectual property protection that will foster innovation.[25]

## III. THE SCIENCE OF THE 1000 GENOMES PROJECT

A gene is the basic physical and functional unit of heredity for humans and almost all other organisms. Genes are made up of DNA, which is composed in part of chemical compounds called bases.[26] Genes cause the body to make molecules called proteins that guide bodily functioning.[27] In humans, genes vary in size from a few hundred DNA bases to more than two million bases.[28] The Human Genome Project has estimated that humans have between 30,000 and 40,000 genes.[29]

The DNA sequence of any two people is 99.9 percent identical, with very small genetic variations proving quite significant in terms of an individual's susceptibility to disease or response to pharmaceuticals.[30] Sites in the DNA sequence where individuals differ at a single DNA base are called single nucleotide polymorphisms ("SNPs").[31] Sets of nearby SNPs on the same chromosome are inherited in blocks called haplotypes.[32] While a haplotype block may contain a large number of SNPs, researchers are able to uniquely identify a haplotype by using a few SNPs, called tag SNPs.[33]

---

[21] *International Consortium Announces*, *supra* note 13.

[22] *Id.*

[23] *Id.*

[24] *Id.*; *see also Three Sequencing Companies Join*, *supra* note 19; John M. Conley et al., *Enabling Responsible Public Genomics*, 20 HEALTH MATRIX 325, 329–34 (2010).

[25] *Three Sequencing Companies Join*, *supra* note 19; *see* The 1000 Genomes Project Consortium, *A Map of Human Genome Variation from Population-Scale Sequencing*, 467 NATURE 1061, 1062 (2010).

[26] LEROY WALTERS & JULIE GAGE PALMER, THE ETHICS OF HUMAN GENE THERAPY 5 (1997).

[27] *Id.* at 4.

[28] *Id.* at 5–7.

[29] Donna M. Gitter, *International Conflicts Over Patenting Human DNA Sequences in the United States and the European Union: An Argument for Compulsory Licensing and a Fair-Use Exemption*, 76 N.Y.U. L. REV. 1623, 1633 (2001).

[30] *International Consortium Announces*, *supra* note 13.

[31] *International HapMap Project*, NAT'L HUM. GENOME RES. INST., http://www.genome.gov/10001688 (last visited Mar. 25, 2011).

[32] *Id.*

[33] *Id.*

The HapMap is a map of the haplotype blocks.[34] Its value lies in the fact that the HapMap reduced the number of SNPs required to examine the entire genome for association with a phenotype from the ten million SNPs that exist to roughly 500,000 tag SNPs.[35] Thus, researchers can now locate much more efficiently and comprehensively the regions of the genome with genes that affect disease, because it is not necessary to study more SNPs than necessary.[36]

The HapMap catalogue, however, identifies only those genetic variations that occur in at least five percent of the population.[37] The 1000 Genomes Project will use new sequencing technologies to "produce a catalogue of variants present at 1 percent or greater frequency in the human population across most of the genome, and down to 0.5 percent or lower within genes."[38] This will permit researchers to gain a better understanding of some diseases that arise less often.[39]

The 1000 Genomes Project will also surpass the HapMap Project in that it will map not only SNPs, the single-letter differences in people's DNA, but also map larger differences in genome structure called structural variants.[40] Structural variants are rearrangements, deletions, or duplications of segments of the human genome which are believed to be associated with predisposition to conditions such as mental retardation and autism.[41]

In order to achieve its scientific goals, the 1000 Genomes Project will sequence the genomes of at least one thousand individuals who gave informed consent for their DNA to be analyzed and placed in public databases.[42] In fact, the first thousand samples will come from those used for the HapMap and from additional samples in the extended HapMap set.[43] These participants are drawn from a geographically wide-ranging group of people of African, East Asian, and European descent.[44] In addition, an effort began in 2010 to expand data collection to include individuals of Hispanic and African-American descent.[45] All of the research participants will remain anonymous and will not have any individual medical information collected from them, because the project is intended to catalogue genetic variation rather than to study the disease profile of the individual research participants.[46] In the future, researchers will be able to use the resource that is developed from the 1000 Genomes Project in order to conduct studies of people affected by various diseases.[47]

The institutions supporting and conducting the genetic sequencing work of the 1000 Genomes Project are the Wellcome Trust Sanger Institute in England, the

---

[34] *Id.*
[35] *Id.*
[36] *Id.*
[37] *Three Sequencing Companies Join*, *supra* note 19.
[38] *International Consortium Announces*, *supra* note 13.
[39] *Three Sequencing Companies Join*, *supra* note 19.
[40] *International Consortium Announces*, *supra* note 13.
[41] *Id.*
[42] *Id.*
[43] *Id.*
[44] *Id.*
[45] Julia Karow, *1000 Genomes Project to Sequence Nearly 1,000 More Samples by Early 2010; New Samples Collected*, GENOMEWEB (Sept. 8, 2009), http://www.genomeweb.com/sequencing/1000-genomes-project-sequence-nearly-1000-more-samples-early-2010-new-samples-co.
[46] *International Consortium Announces*, *supra* note 13.
[47] *Id.*

Beijing Genomics Institute in China, and the National Human Genome Research Institute, part of the National Institutes of Health ("NIH") in the United States.[48] Most of the funding for the 1000 Genomes Project, which will cost between U.S. $30 to $50 million, will come from the existing sequencing budgets of the participating institutes.[49] In addition, three private U.S. firms that have pioneered development of new sequencing technologies are contributing a significant portion of the sequence data, indicating their belief in the value of the large datasets to be produced and made available through publicly accessible databases.[50] Indeed, these firms, 454 Life Sciences, a Roche Company; Applied Biosystems, an Applera Corp. business; and Illumina Inc., like all other participants in the 1000 Genomes Project, have consented to the open access policies established by the 1000 Genomes Project Steering Committee.[51] These policies include "rapid public release of the data; a prohibition on project participants gaining early access to the data; an intellectual property policy that precludes any participants from controlling the information provided by the project or filing for intellectual property rights in the sequences they produce; regular progress reporting; and coordination of scientific publications with the rest of the consortium."[52]

## IV. THE DATA ACCESS POLICY OF THE 1000 GENOMES PROJECT

In order to advance research and encourage international scientific collaboration, participants in large-scale, publicly-funded sequencing projects have adopted broad data-sharing policies.[53] In February 1996, at the First International Strategy Meeting on Human Genome Sequencing in Bermuda, representatives of laboratories involved in human genome sequencing and of funding agencies unanimously agreed to implement the Bermuda Statement, an international agreement favoring release into the public domain of genetic databases achieved through public funding.[54] The Bermuda Statement provides 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 maximize its benefit to society."[55] One of the fundamental requirements of the Bermuda Statement is the stipulation that sequences longer than 1,000 base pairs must be made freely available to the public, preferably within twenty-four hours of generation.[56] For the HGP, the highest profile data set rapidly released before publication, the international sequencing centers involved in the

---

[48] *Id.*

[49] Jocelyn Kaiser, *A Plan to Capture Human Diversity in 1000 Genomes*, 319 SCI. 395, 395 (2008).

[50] *Three Sequencing Companies Join*, *supra* note 19.

[51] *Id.*

[52] *Id.*

[53] Schofield et al., *supra* note 2, at 171.

[54] *See* Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing (held in Bermuda on Feb. 25–28, 1996), http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml#1 [hereinafter Bermuda Statement].

[55] *Id.*

[56] Bryn Nelson, *Empty Archives*, 461 NATURE 160, 162 (2009).

HGP retained the right to be the first to describe and analyze their complete data sets in peer-reviewed publications, in return for the early release of their data.[57]

According to its proponents, the Bermuda Statement data release policy fosters independent checking of the sequence data by other researchers,[58] leading to huge gains in life sciences research.[59] With respect to the HGP, this policy resulted in the publication of new information about thirty genes associated with disease even prior to the publication of the draft sequences.[60] In addition, a broad data access policy prevents publicly funded large-scale sequencing centers from "establishing a privileged position in the exploitation and control of human sequence information."[61] What is more, the policy also delivers a symbolic message that "the genome belongs to everybody."[62] When sequencers met again in 2003 in Florida, they reaffirmed their commitment to the Bermuda Statement in a statement referred to as the Fort Lauderdale Principles.[63] Consequently, organizations such as the National Institutes of Health's National Human Genome Research Institute, the Department of Energy, and the Wellcome Trust require compliance with this policy as a condition of receiving funding.[64]

The 1000 Genomes Project has announced its adherence to the Fort Lauderdale principles on its website.[65] The site emphasizes that data producers will "release the Project data quickly, prior to publication, in the expectation that they will be valuable for many researchers" and 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."[66]

In October 2010, the 1000 Genomes Project published in the journal *Nature* an analysis of its completed pilot phase.[67] After the completion of the pilot phase, the 1000 Genomes Project began full-scale studies, which will require an additional two years.[68] Data from the pilot studies and the full-scale project are freely available on the project's website, www.1000genomes.org.[69]

The data release policy of the 1000 Genomes Project aims to make the genotype data freely available, while reserving to data producers the right to make the first

---

[57] Birney et al., *supra* note 1, at 168.

[58] David R. Bentley, *Genomic Sequence Information Should Be Released Immediately and Freely in the Public Domain*, 274 SCI. 533, 533 (1996).

[59] *Id*. at 534.

[60] Birney et al., *supra* note 1, at 168.

[61] Bermuda Statement, *supra* note 54.

[62] Eliot Marshall, *Bermuda Rules: Community Spirit, with Teeth*, 291 SCI. 1192, 1192 (2001).

[63] *See* Wellcome Trust, *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* 2 (conference of Jan. 2003), http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.

[64] Lee Rowen et al., *Publication Rights in the Era of Open Data Release Policies*, 289 SCI. 1881, 1881 (2000).

[65] *Use of the Project Data, Presentations and Publications, and Authorship*, 1000 GENOMES, http://www.1000genomes.org/data#DataUse (last visited Mar. 25, 2011).

[66] *Id*.

[67] *1000 Genomes Project Publishes Analysis of Completed Pilot Phase*, NAT'L HUM. GENOME RES. INST. (Oct. 27, 2010), http://www.genome.gov/27541917; *see Map of Human Genome Variation*, *supra* note 25.

[68] *Analysis of Completed Pilot Phase*, *supra* note 67.

[69] *How to Access 1000 Genomes Data*, 1000 GENOMES, http://www.1000genomes.org/data/#DataAccess (last visited Mar. 25, 2011).

presentations and to publish the first paper with global analyses of the data.[70] Examination of the 1000 Genomes Project data release policy is particularly important in light of the recommendation arising from the Fort Lauderdale meeting that rapid prepublication data release be applied to other data sets created primarily as a resource for the scientific community.[71] Experts have observed that data producers are no longer only large genomics centers, and that, increasingly, smaller groups are able to produce a large amount of data.[72] Although funding agencies anticipated that the policies from the HGP and 1000 Genomes Project would be automatically adopted by these smaller projects, this did not occur.[73] For this reason, between eighty to one hundred scientists, ethicists, lawyers, representatives of funding agencies, and journal editors convened at a data release workshop in Toronto in May 2009, hosted by Genome Canada, in order to discuss the extension of such data release policies to fields such as proteomics and biobanking.[74] The product of that meeting was a set of proposals published in *Nature* in 2009, referred to as "the Toronto Statement," which aims to establish a set of "best practices" for scientists, both data producers and data users, as well as funding agencies and journal editors.[75] The Toronto Statement is meant to combat the many factors that inhibit researchers from sharing their data.[76] Particularly in the realm of small science, these include data producers' rational desire to be recognized for producing and analyzing data, thereby preserving their ability to win future research grants; the challenge of crafting appropriate publication and intellectual property policies; and the complexity of establishing a suitable infrastructure to house the data.[77]

## V. OBSTACLES TO DATA-SHARING AND POTENTIAL SOLUTIONS OFFERED BY THE TORONTO STATEMENT

### *A. Lack of Recognition of Data Producers*

In pursuing career advancement, data producers are understandably reluctant to grant data users access to research results achieved after devoting years of hard work to "planning a project, securing funding and ethics approval, recruiting the participants, collecting the data and materials, performing analyses, managing the collection and infrastructure, controlling technical quality, and generally nourishing the project through waves of funding and maturation."[78] In addition, data producers

---

[70] *Use of the Project Data*, *supra* note 65.

[71] Wellcome Trust, *supra* note 63, at 2.

[72] Ciara Curtin, *When to Share*, GENOMEWEB (May 2010), http://www.genomeweb.com/when-share.

[73] *Id.*

[74] *Id.*

[75] Birney et al., *supra* note 1, at 168–70.

[76] *See* Donna M. Gitter, *The Challenges of Achieving Open Source Sharing of Biobank Data*, 29 BIOTECH. L. REP. 623, 625 (2010).

[77] *Id.*

[78] William W. Lowrance, *Access to Collections of Data and Materials for Health Research* 24 (Mar. 2006), http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtx030842.pdf.

fear being "scooped" by data users who may mine data and discover relationships in it that the producer did not discern.[79] As noted in the Toronto Statement, an emphasis on early data release gives rise to tension between data producers and users.[80] The former typically wish to publish a first description of a data set, while the latter desire to publish their own analyses of the data.[81] One way to reduce this tension is to ensure proper attribution and recognition for data producers.[82]

One means of acknowledging the contributions of data producers is to track the usage and citation of data sets through the use of electronic systems similar to those used for traditional publications.[83] In order to quantify the usefulness of particular data sets, Professor Boyle has offered the example of a music site associated with Creative Commons,[84] known as ccMixter.[85] This site allows users to download music from the site and remix the samples into new tracks, while simultaneously maintaining a record of the credits due to each musician.[86] A comparable system would permit attribution to the appropriate data producer and also allow universities and funding agencies to track the number of uses, and therefore the value, of a researcher's data.[87] This satisfies researchers' desire for attribution, and assists them in demonstrating their productivity to funding agencies.[88]

Another expert, Myles Axton of *Nature*, has proposed recognizing contributors via "microattribution," meaning the tracking of researchers' contributions "down to the very smallest meaningful unit (database record, gene), as opposed to whole papers as is tradition."[89] Microattribution could minimize the shortcomings of the Fort Lauderdale Agreement, which Axton has critiqued for failing to reward data producers with credit for uses of their data.[90] Nor does the Agreement make any provision for crediting data generators, such as genomics "factory" laboratories.[91] In addition, notes Axton, the Fort Lauderdale Agreement is not truly enforceable.[92]

The Toronto Statement also suggests that funding agencies have a role in fostering data-sharing.[93] First, the Statement advises that funding agencies should mandate rapid prepublication data release for projects that "have broad utility, are large in scale, are 'reference' in character and typically have community 'buy-in.'"[94]

---

[79] Nelson, *supra* note 56, at 163.

[80] Birney et al., *supra* note 1, at 169.

[81] *Id*.

[82] *Id*. at 169–70.

[83] Nelson, *supra* note 56, at 163.

[84] Creative Commons is a nonprofit corporation that aims to facilitate content-sharing in accordance with the law of copyright. *See About Creative Commons*, CREATIVE COMMONS, http://creativecommons.org/about/ (last visited Mar. 25, 2011).

[85] Nelson, *supra* note 56, at 163 (citing Professor James Boyle of Duke Law School).

[86] *See About*, ccMixter, http://ccmixter.org/about (last visited Mar. 25, 2011).

[87] Nelson, *supra* note 56, at 163.

[88] *Id*.

[89] Gudmundur A. Thorisson, & Anthony J. Brookes, *Meeting Minutes, IRBW2009 Workshop of May 13–14, 2009*, at 5 (Aug. 27, 2009), http://www.gen2phen.org/system/files/private/IRBW2009%20meeting%20minutes%20v2.pdf.

[90] *Id*.

[91] *Id*.

[92] *Id*.

[93] Birney et al., *supra* note 1, at 169.

[94] *Id*.

Second, according to the Toronto Statement, funding agencies should make explicit the requirement for prepublication data release and also become actively involved throughout the data release process, as demonstrated by the International HapMap Project and the 1000 Genomes Project, among others.[95] Among the essential roles for funding agencies is participation in the development of effective consent, security, access, and governance mechanisms that protect research participants while encouraging prepublication data release.[96]

Third, for projects generating large data sets, the Toronto Statement advises that funding agencies require that grant applications include data-sharing plans and undergo peer review, and notes that this is not commonly the case at the present time.[97] Experts have recommended, in the context of post-publication data release (and it is equally applicable prepublication) that, where they do not yet exist, clear criteria should be developed for reviewers of grants to help them evaluate applicants' plans for sharing data and material.[98] Among examples of good practice cited in this regard are the NIH,[99] the U.K. Wellcome Trust,[100] and the U.K. Medical Research Council ("MRC"),[101] a government-funded research agency akin to the NIH.[102] Commentators note that, for these agencies, "[d]ata-sharing plans are required in proposals, efforts are made to facilitate sharing, such as putting investigators in touch with repositories and, for some organizations, compliance is an important consideration in funding renewal."[103]

Fourth, in terms of compliance, funding agencies should clearly express the rewards for honoring, as well as the sanctions for ignoring, their policies relating to prepublication data release, and also consistently enforce such policies.[104] One possibility is to condition grant renewals and promotions to some degree on the extent to which the investigator shares data.[105]

### *B. The Challenge of Crafting Appropriate Publication Policies*

In light of researchers' desires to protect the data they have gathered and advance their own careers, efforts to create a culture of data-sharing are inextricably linked to the question of the optimal publication policies to support data-sharing

---

[95] *Id.*

[96] *Id.*

[97] *Id.*

[98] Schofield et al., *supra* note 2, at 172.

[99] *See generally Final NIH Statement on Sharing Research Data*, NAT'L. INSTS. HEALTH (Feb. 26, 2003), http://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html.

[100] *See generally Policy on Data Management and Sharing*, WELLCOME TRUST (Aug. 2010), http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/wtx035043.htm.

[101] *See generally MRC Policy on Data Sharing and Preservation*, MED. RES. COUNCIL, http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Datasharinginitiative/Policy/index.htm (last visited Mar. 25, 2011).

[102] *About Us*, MED. RES. COUNCIL, http://www.mrc.ac.uk/About/Structure/index.htm (last visited Mar. 25, 2011).

[103] Schofield et al., *supra* note 2, at 172.

[104] *Id.*

[105] Birney et al., *supra* note 1, at 169.

efforts.[106] According to the Toronto Statement, although many scholarly works have been published by third parties reporting research findings based upon data sets released before publication, few data producers were affected by these publications that pre-dated their own.[107] Nevertheless, the Toronto Statement proposes that "this ongoing concern is best addressed by fostering a scientific culture that encourages transparent and explicit cooperation among data producers, data analysts, reviewers, and journal editors."[108]

One way of achieving such cooperation is to permit investigators who contribute data to enjoy some period of exclusivity during which they will have the sole right to publish analyses of the data.[109] The Toronto Statement declares that data producers should, as early as possible, and ideally before large-scale data generation begins, produce a citable statement or "marker paper" in which they describe the data set and their intentions in respect of analysis and publication.[110] This statement should include "clear details about the data set to be produced, the associated metadata, the experimental design, pilot data, data standards, security, quality-control procedures, expected timelines, data release mechanisms, and contact details for lead investigators."[111] In the event that data producers request a protected time period to allow them to be the first to publish the data set, data producers should expect this period to be used only for the publication of a global analysis of the data and to expire within one year.[112]

The NIH and the U.K. Medical Research Council have both adopted a period of data exclusivity for data producers.[113] The NIH Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies ("GWAS") declares that investigators who contribute data to a NIH GWAS data repository will retain the exclusive right to publish analyses of the dataset for a maximum of twelve months following its release via the NIH GWAS data repository.[114] During this period of exclusivity, the NIH grants data access through data access committees to other investigators, who may analyze the data, but not submit for publication their analyses or conclusions, until the expiration of the exclusivity period.[115] The NIH also expects all investigators who access GWAS datasets to acknowledge in all publications the data producers who conducted the original study, along with the funding organization that supported the work and the NIH GWAS data repository.[116] Similarly, the U.K. Medical Research Council provides for a period of interim exclusivity for data producers.[117] The MRC data-sharing policy provides that "[a] limited, defined period of exclusive use of data for primary research is reasonable."[118]

---

[106] *Id.* at 169–70.

[107] *Id.*

[108] *Id.* at 169–70.

[109] *Id.* at 170.

[110] *Id.*

[111] *Id.*

[112] *Id.*

[113] *MRC Policy*, *supra* note 101; *see also* Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies, 72 Fed. Reg. 49,290, 49,293 (Aug. 28, 2007).

[114] *See* Policy for Sharing, 72 Fed. Reg. at 49,296.

[115] *Id.*

[116] *Id.*

[117] Lowrance, *supra* note 78, at 25.

[118] *MRC Policy*, *supra* note 101.

The U.K. Biobank follows suit, asking researchers to recruit 500,000 participants and collect their data and samples, in return for which the researchers will gain exclusive access for a period of time or for research into certain specialty areas.[119] In general, the purpose of the period of exclusivity is to provide the data producer sufficient time to produce, organize, document, verify, and analyze the data in preparation for publication in a scientific journal.[120] After publication, the data is made available more broadly, either via a publicly available database or via an application system.[121]

The Toronto Statement emphasizes that data users should honor the "scientific etiquette" that allows data producers to publish the first global analyses of their data set.[122] Participants at the Toronto meeting urged, in cases where the publication plans of data users overlap the data producers' proposed analyses, that the data users approach the producers with the goal of "creating a mutually agreeable publication schedule, such as co-publication or inclusion within a set of companion papers."[123] In addition, data users should acknowledge any data or materials used and the originating sources.[124] Some of the potential methods to achieve this include the addition of metadata tags linking to data and bioresources, or a digital object identifier for resources in public repositories, which would facilitate the searching of the literature for specific bioresources.[125] These approaches also offer appropriate recognition of data producers who comply with data release and deposition policies, making it possible to offer incentives such as funding contingent upon such sharing.[126]

Like data users, journal editors and reviewers also have an important role in the data-sharing process.[127] Participants at the Toronto meeting recommended that journals engage actively in the dialogue about rapid prepublication data release, both in their formal guide to authors and informal instructions to reviewers.[128] Journal editors should work in tandem with reviewers to consider carefully the specific policies regarding citation and use of the data that may be associated with certain large-scale data sets.[129] This will help to raise the quality of analysis and also promote fairness in citation of published studies.[130]

Even a decade before the Bermuda principles, the editors of journals, such as *Nucleic Acids Research,* fostered the early development of GenBank and other genomic repositories by requiring researchers to deposit their data there as a precondition for publishing.[131] Newer journals, such as the open-access Public Library of Science journals, have made publication contingent on making the data "freely available without restriction, provided that appropriate attribution is given

---

[119] Lowrance, *supra* note 78, at 25.
[120] *See* Reichman & Uhlir, *supra* note 6, at 333.
[121] *Id.*
[122] Birney et al., *supra* note 1, at 169.
[123] *Id.* at 170.
[124] *Id.*
[125] *Id.* at 169.
[126] Schofield et al., *supra* note 2, at 171–72.
[127] Birney et al., *supra* note 1, at 170.
[128] *Id.*
[129] *Id.*
[130] *Id.*
[131] Nelson, *supra* note 56, at 162–63.

and that suitable mechanisms exist for sharing the data used in a manuscript."[132] *Nature* journals require authors to "'make materials, data and associated protocols promptly available to readers without preconditions.'"[133]

Despite the increased call in many journals for data-sharing, the ability of journals to compel such behavior is limited.[134] In March 2009, for example, the journal *Epidemiology* declared that "'We invite our authors to share their data and computer code when the burden is minimal,'" according to an editorial in that issue.[135] Miguel Hernán, an epidemiologist at Harvard University and a co-author of that editorial, acknowledged the trend toward data-sharing but contended that forcing a sharing requirement on authors "would be suicidal", especially with unresolved concerns over patient confidentiality, and would likely cause authors to submit their papers elsewhere.[136] Moreover, many journals have no written policy on the availability of either resources or primary data.[137]

### *C. The Complexity of Establishing a Suitable Infrastructure*

The Toronto Statement also recommends that funding agencies offer infrastructural support by funding the creation and long-term maintenance of databases.[138] Among the costs associated with creating a suitable infrastructure to house and maintain the data and any associated physical specimens are: the maintenance of physical premises; the development of appropriate information technology; the preparation of data for storage or archiving, such as anonymising the data and documenting the variables; salaries for administrators, managers, and staff; and the creation and maintenance of an accessible database.[139] These costs are all ongoing ones.[140] Because granting agencies understandably focus primarily on research, they therefore frequently fail to invest in the infrastructural support necessary to support their archiving requirements, in a sort of "tragedy of the commons."[141]

The internationally collaborative nature of biomedical research exacerbates this problem by allowing each funding agency to leave the problem to another agency. Yet overcoming the infrastructure issue is crucial in order to optimize the benefits of technological progress in the life sciences.[142] Taking human genome sequencing as one example, one source estimates that the cost of storing all known DNA sequence information in openly accessible databases costs less than one percent of the sum necessary to generate such sequence data.[143]

---

[132] *Id.* at 163.
[133] *Id.*
[134] *Id.*
[135] *Id.*
[136] *Id.*
[137] Schofield et al., *supra* note 2, at 171.
[138] Birney et al., *supra* note 1, at 169.
[139] Gitter, *supra* note 76.
[140] *Id.*
[141] Nelson, *supra* note 29, at 162.
[142] *Id.*
[143] EUR. LIFE SCI. INFRASTRUCTURE FOR BIOLOGICAL INFO., ELIXIR: DATA FOR LIFE, http://www.elixir-europe.org/bcms/elixir/Documents/Elixir_brochure.pdf (last visited Mar. 25, 2011).

## VI. THE NEED FOR RESTRICTED ACCESS DATABASES

Notwithstanding the prescriptions in the Toronto Statement for encouraging data-sharing, not all data is suitable for an open access approach.[144] Indeed, samples and data that retain their link with an individual research participant must be protected with additional procedures.[145] Examples of such data include personal information such as individual-level genotypic and phenotypic data; exposure to drugs and environmental factors; and pedigree data, including information about familial relationships, along with analyses of such data.[146] Experts have offered several technical and policy approaches to preserve patient privacy.[147] They recommend, inter alia, the establishment of policies to assess credentials of data users; the execution of clear contracts with data users that define the appropriate use of data; formalization of liability rules for misuse of data; and the use of a technical data management approach that increases the number of research participants whose data will be aggregated where the data is considered more sensitive.[148] Experts note that subjecting data to access controls erects a barrier that limits the number of researchers who will reuse the resources.[149] What is more, researchers may decline to abide by overly complicated data access agreements.[150]

Nevertheless, restricted access databases are also useful in terms of building upon the collaborations already established by scientific researchers, particularly those involved in small science.[151] One such biobank is the U.K. DNA Banking Network ("UDBN"), which is a secondary biobank, meaning that it aggregates and manages tissue samples and associated data gathered by clinicians who gather the samples in the course of studying particular diseases.[152]

In order to establish UDBN, the U.K. Medical Research Council offered grants to centers that housed DNA collections.[153] These awards required the collections to be maintained as "shared national resources," and "made available to collaborators."[154] In addition, awardees were "'required to transfer a portion of each sample'" to the network and "'to add any genotype data they obtain to the common database."[155]

In granting access to its collections, the UDBN explicitly rejected the unrestricted "open access" model, instead developing a "fair access" regime, which

---

[144] Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies, 72 Fed. Reg. 49,290, 49,291 (Aug. 28, 2007).

[145] *Id.* at 49,292.

[146] *Id.* at 49,293.

[147] *See* Bradley Malin et al., *Technical and Policy Approaches to Balancing Patient Privacy and Data Sharing in Clinical and Translational Research*, 58 J. INVESTIGATIVE MED. 11, 15–17 (2010).

[148] *Id.*

[149] *Id.*

[150] Presentation at the 2010 Advances in Genomics Symposium, Paul Flicek, Vertebrate Genomics, Challenges for the Data Management and Analysis of Large-Scale Human Genome Sequencing, http://www.advances-in-genomics.org/presentations/Flicek.pdf (last visited Mar. 25, 2011).

[151] *See generally* Martin Yuille et al., *The UK DNA Banking Network: A "Fair Access" Biobank*, 11 CELL TISSUE BANK 241–51 (2010), *available at* http://www.springerlink.com/content/l1082h68g0645517/fulltext.pdf.

[152] *Id.* at 241.

[153] *Id.* at 242–43.

[154] *Id.* at 243.

[155] *Id.*

derives inspiration from the 2003 United Nations Educational Scientific and Cultural Organization ("UNESCO") International Declaration on Human Genetic Data.[156] The UDBN distributes data via the project website to third party researchers, who apply online for registration.[157] After verifying the researcher's credentials, UDBN grants access to a restricted area of the website.[158] The third party researchers can then communicate online with the data collectors in order to negotiate collaboration.[159] If a collaborative relationship is successfully negotiated, UDBN permits the data collector to grant the third party access to fuller data.[160]

The fair access model builds upon researchers' willingness to share data with collaborators, and concomitantly denies access to non-collaborators.[161] This approach also accepts a collector's right to "exclusive access to his/her collection for the purposes of the investigational goals stated in the initial collection proposal," recognizing that granting a "first mover" advantage is likely to motivate scientific discovery.[162] This model also acknowledges that tensions frequently arise when a potential collaborator requests data access from a data producer.[163] Thus, the restricted access model provides a necessary complement to the open access databank, particularly in the realm of "small science."

---

[156] *Id.*
[157] *Id.*
[158] *Id.* at 246.
[159] *Id.*
[160] *Id.*
[161] *Id.* at 244–46.
[162] *Id.*
[163] *Id.*