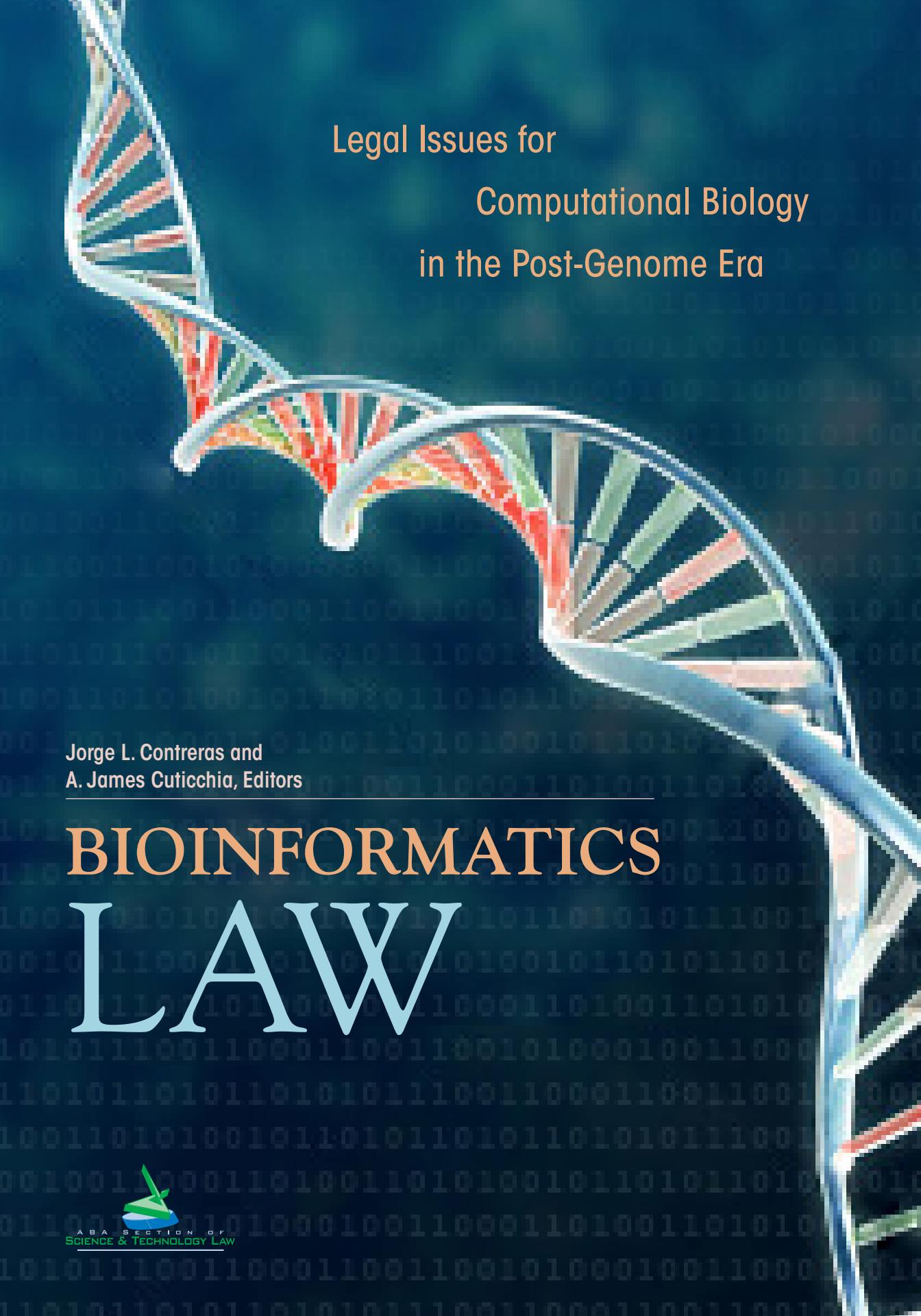

Legal Issues for
Computational Biology
in the Post-Genome Era

Jorge L. Contreras and
A. James Cuticchia, Editors

BIOINFORMATICS LAW

ABA Section of
Science & Technology Law

CHAPTER 12

Data Release and Access

Jorge L. Contreras*

INTRODUCTION

Bioinformatics as a discipline is heavily dependent on the availability of vast public collections of data that can be downloaded and analyzed by researchers worldwide. In general, this data has been assembled by publicly and privately funded research projects, beginning with the Human Genome Project (HGP). The availability of this tremendous public data resource is due, in large part, to the data-release policies developed toward the beginning of the HGP, which have been carried forward, in modified form, to the present. These policies impose requirements on both the generators of data (typically the sequencing centers and other laboratories conducting genetic experiments) and the users of that data (i.e., anyone who downloads and/or uses it). This chapter briefly outlines the history of such data-release policies, including those developed by the U.S. National Institutes of Health (NIH), the private sector, and non-U.S. bodies, and provides an overview of the obligations imposed on both data generators and data users.

THE EXPANDING DATA LANDSCAPE

The principal databases for the deposit of genomic sequence data are GenBank, which is administered by the National Center for Biotechnology Information (NCBI)—a division of the NIH's National Library of Medicine; the European Molecular Biology Library (EMBL) in Hinxton, England; and the DNA Data Bank of Japan (DDBJ). NCBI also maintains the RefSeq database, which consolidates and annotates much of the sequence data found in GenBank. In addition to DNA sequence data, genomic studies generate data relating to the association between particular genetic markers and disease risk and other physiological traits. This type of data, which is more complex to record, search, and correlate than the raw sequence data deposited in GenBank, is housed in databases such as the Database of Genotypes and Phenotypes (dbGaP), operated by NIH's National Library of Medicine. dbGaP can also accommodate phenotypic data, which

*This chapter is adapted from Jorge L. Contreras, *Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons*, 12 Minn. J.L. Sci. & Tech. 61 (2011).

includes elements such as de-identified subject age, ethnicity, weight, demographics, drug exposure, disease state, and behavioral factors, as well as study documentation and statistical results, including linkage and association analyses. Given the potential sensitivity of phenotypic data, dbGaP allows access to data on two levels: open and controlled. Open data access is available to the general public via the Internet and includes nonsensitive summary data, generally in aggregated form. Data from the controlled portion of the database may be accessed only under conditions specified by the data supplier, often requiring certification of the user's identity and research purpose.

The sheer size of the public aggregation of genomic data (which I have termed the "genome commons")[1] is matched only by the breathtaking rate at which it is expanding. Over its decade-long existence, the HGP mapped the 3.2 billion base pairs comprising the human genome. To do so, it sequenced tens of billions of DNA bases (gigabases), creating what was then an unprecedented accumulation of genomic data. By way of comparison, the current 1000 Genomes Project is projected to generate 200,000 gigabases of data—approximately 20,000 times the quantity generated by the HGP.[2] In 2010, the Cancer Genome Atlas project (TCGA) generated data at a rate of 7,300 gigabases per month. Together with the Human Microbiome Project and the 1000 Genomes Project, the total data generation rate from major NIH-funded genome projects in 2010 was nearly 10,000 gigabases per month.[3] And the rate at which DNA sequencing can be accomplished is rapidly increasing. According to one 2011 report, "a single DNA sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project."[4] Statistics like these have led to talk of a "data tsunami" in genomic science, in which the capacity to analyze the vast quantities of data being generated will severely lag behind the rate at which such data is produced.

BERMUDA AND THE ORIGINS OF RAPID GENOMIC DATA RELEASE

In 1992, shortly after the HGP was launched, NIH and the Department of Energy (DOE) developed formal guidelines for the sharing of HGP data.[5] The guidelines required that data generated by the HGP be deposited in GenBank, making it available to all scientists worldwide. Recognizing that the sequencing centers working on the HGP would require time to analyze and prepare publications relating to this data, the policy gave researchers a six-month period before the public release of data was required. The year 1996 marked a turning point for the HGP. Not only was it the year in which sequencing of the human genome was scheduled to begin, it also signaled a sea change in the data-release landscape. That February, approximately 50 project leaders met in Hamilton, Bermuda, to deliberate over the speed with which HGP data should be released to the public, and whether the six-month "holding period" approved in 1992 should continue.[6] The resulting "Bermuda Principles" established that all DNA sequence information from large-scale human genomic sequencing projects should be "freely available and in the public domain in order to encourage research and development and to maximize its benefit to society."[7] They went on to define the method by which such data should be shared, requiring that sequence assemblies greater than one kilobase (kb) in length be released *within 24 hours* after assembly, and that finished annotated sequences should be submitted *immediately* to a public database.

The Bermuda Principles were revolutionary in that they established, for the first time, that data from public genomic projects should be released to the public almost immediately after their generation. Prior to 1996, the position of the National Human Genome Research Institute (NHGRI) with respect to data release and intellectual property was not

very different than that of other federal agencies. By and large, such projects required data release after the publication of study results, which often occurred one to two years following the completion of research.[8] But in the negotiations at and leading up to the Bermuda meeting, the scientific community's acknowledgment of the collective norms of data sharing and the public domain, bolstered by the gravitas of several Nobel laureates and other leading figures, seems to have captured the agency's imagination. These norms have since become ingrained as part of NHGRI's basic position treating genomic data as a public good that should be widely available and unencumbered.

DATA RELEASE POST-HGP

The initial draft of the human genome sequence was published by the HGP and a competing private effort in 2001. In 2003, the Wellcome Trust convened a meeting in Fort Lauderdale, Florida, to revisit rapid data-release issues in the "post-genome" world.[9] While the Fort Lauderdale participants "enthusiastically reaffirmed" the 1996 Bermuda Principles, they also expressed concern over the inability of data-generating scientists to study their results and publish analyses prior to the public release of data. The most significant outcome of the Fort Lauderdale meeting was a consensus that the Bermuda Principles should apply to each "community resource project" (CRP), meaning "a research project specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community." Under this definition, the 24-hour rapid release rules of Bermuda would be applicable to large-scale projects generating nonhuman sequence data, other basic genomic data maps, and other collections of complex biological data such as protein structures and gene expression information. In order to effectuate this data-release requirement, funding agencies were urged to designate appropriate efforts as CRPs and to require, as a condition of funding, that rapid, prepublication data release be required in such projects.

Notwithstanding this show of support, the Fort Lauderdale participants acknowledged that rapid, prepublication data release might not be feasible or desirable in all situations, particularly for projects other than CRPs. In particular, the notion of a CRP, the primary goal of which is to generate a particular data set for general scientific use, is often distinguished from "hypothesis-driven" research, in which the investigators' primary goal is to solve a particular scientific question, such as the function of a specific gene or the cause of a specific disease or condition.[10] In hypothesis-driven research, success is often measured by the degree to which a scientific question is answered rather than the completion of a quantifiable data set. Thus, the early release of data generated by such projects would generally be resisted by the data-generating scientists who carefully selected their experiments to test as-yet-unpublished theories. Giving such data away before their theories are published could potentially enable a competing group to "scoop" the originating group, a persistent fear among highly competitive scientists.

In the years following the Fort Lauderdale summit, numerous large-scale genomic research projects were launched with increasingly sophisticated requirements regarding data release. These policies implement their requirements through contractual mechanisms that are more tailored and comprehensive than the broad policy statements of the HGP era. Moreover, increasingly sophisticated database technologies have enabled the provision of differentiated levels of data access, the screening of user applications for data access, and improved tracking of data access and users.

DATA-RELEASE REQUIREMENTS TODAY

Today, all major publicly funded genomic research projects contain detailed data-reporting and data-release requirements. Below is a summary of just a few current programs and their associated requirements.

The Cancer Genome Atlas

In 2006, the National Cancer Institute (NCI) and NHGRI launched a project to catalog genomic changes relating to cancer.[11] The TCGA project, which continues today, generates genomic sequence and related data, but also keeps track of a large amount of clinical data, including patient diagnosis, treatment history, and ongoing status.[12] Due to the specialized nature of the project data, deposits are made both in dbGaP as well as a TCGA-specific database administered by NCI.[13] Given the potential for identifying individual patients from their genomic and phenotypic data, great attention was paid to controlling access to TCGA data. TCGA data is available in an open-access tier and a controlled-access tier.[14] Open-access is provided for data that cannot be aggregated to generate an individually identifiable data set, whereas controlled-access enables researchers to access clinical and individually unique data. Access to the controlled-access data tier requires the user's acknowledgement of a Data Access Certification containing restrictions on research use, security, transferability, and other matters.

The NIH Genome-Wide Association Studies Policy

In response to the growing number of genome-wide association (GWA) studies being conducted and the large amount of genomic data generated by such studies, in August 2007 the NIH released a new policy regarding the generation, protection, and sharing of data generated by all federally funded GWA studies.[15] The NIH Genome-Wide Association Studies (GWAS) policy requires that researchers submit descriptive information about each GWA study for inclusion in the "open access" portion of dbGaP. Researchers are also "strongly encouraged" to submit study results, including phenotypic, exposure, and genotypic data, for inclusion in the "controlled access" portion of the database "as soon as quality control procedures have been completed."[16]

Among the principal concerns raised regarding GWA study data were those surrounding the public release of phenotypic or clinical information that could eventually be traced back to individual subjects.[17] To address this concern, the NIH GWAS policy requires that GWAS data be de-identified in accordance with HIPAA guidelines.[18] Moreover, the data in the controlled-access portion of the database may be released only after approval of the proposed research use by a Data Access Committee,[19] and then only under a signed Data Use Certification. The Data Use Certification requires researchers and their institutions to agree, among other things, to use data only for the approved research purpose, protect data confidentiality, implement appropriate data security measures, not attempt to identify individual data subjects, not sell any data, not share data with third parties, and to report violations to the committee. Finally, the NIH sets forth its position that a request under the federal Freedom of Information Act (FOIA)[20] for the release of individually identifiable GWAS information would constitute an "invasion of personal privacy" under FOIA, and will be denied by NIH.[21]

The NIH GWAS policy was amended in August 2008 following the publication of a scientific paper demonstrating that inferences regarding individual identity could be drawn by analyzing allele frequency data in aggregated genomic data sets and other statistical

techniques.[22] Due to concerns relating to potential identification of GWAS subjects, NIH withdrew certain GWAS-generated single nucleotide polymorphism (SNP) data from the publicly accessible portions of dbGaP and certain NCI databases and placed them in the controlled-access portions of these databases.[23]

The NIH GWAS policy addresses the publication priority concerns of data generators by stating an expectation that users of GWAS data 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. NIH also expresses a "hope" and expectation that "genotype-phenotype associations identified through NIH-supported and NIH-maintained GWAS datasets and their obvious implications will remain available to all investigators, unencumbered by intellectual property claims."[24] It goes on to explain that "[t]he filing of patent applications and/or the enforcement of resultant patents in a manner that might restrict use of NIH-supported genotype-phenotype data could diminish the potential public benefit they could provide."[25] However, in an effort to show some support for patent seekers, the GWAS policy also "encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs."[26]

ENCODE and modENCODE

In 2007 NIH launched the ENCODE and modENCODE projects to identify functional genomic elements in humans and two simpler model organisms.[27] The ENCODE data-release policy[28] designates the project as a "community resource project" but also recommends a nine-month embargo period during which users of released data are asked not to publish or present results based on that data. The ENCODE Policy distinguishes between published and unpublished data, as well as verified and unverified data, and offers several examples of the data-use implications for different types of studies. The length and complexity of the policy evidences the agency's and the participants' desire for clear guidelines and the avoidance of misunderstandings regarding the release of data, as the diversity of participants, organisms, and data types has expanded dramatically beyond those originally considered by the framers of the Bermuda Principles. This year, ENCODE scientists released a massive series of findings in 30 simultaneously published scientific papers.[29]

The Human Microbiome Project

The Human Microbiome Project (HMP) is a large-scale community resource project initiated in 2008 that is designed to identify and sequence the genomes of a handful of the host of microorganisms inhabiting the human body.[30] While much HMP data is subject to rapid Bermuda-like disclosure requirements, investigators are permitted to withhold certain other data from the public for a period of several months.[31] This hold-back period is intended to permit HMP researchers to analyze and prepare publications on their data before it is released to competing researchers. The reasons that researchers, who are driven by intense competitive pressure to publish and claim credit for discoveries, have pushed for such hold-back periods are clear. However, it also appears, at least in the case of HMP, that NIH has not vigorously advanced the patent deterrent arguments that previously motivated policy decisions during the HGP and its immediate aftermath. Whether the experience of the HMP indicates a new direction for NIH, or simply a minor deviation from its overall policy mission, is not clear.

PRIVATE SECTOR INITIATIVES

In addition to the public-sector projects described above, a number of private-sector research projects have resulted in the creation of large aggregations of genomic data. During the HGP era, privately funded projects such as the Merck Gene Index[32] and the SNP Consortium[33] attracted attention and kudos for releasing significant amounts of data into the public domain. Today, successor projects such as the International Serious Adverse Events (SAE) Consortium (SAEC) continue to fund academic research in targeted areas. SAEC's particular focus is the discovery of genetic markers potentially associated with severe adverse drug reactions.[34] The SAEC works with academic collaborators to collect DNA samples and associated phenotypic data, and then funds GWAS, targeted sequencing, and statistical analyses to identify potential markers and associations of interest. SAEC studies have investigated potential DNA markers relating to drug-induced liver injury, serious skin rash, excessive weight gain, osteonecrosis of the jaw, and other conditions, and all data is released publicly on the SAEC website. The SAEC seeks to minimize patent encumbrances on genetic markers and associations that it identifies via a "protective" patent strategy in which patent applications claiming various DNA markers have been filed to act as prior art against third-party patent filings. Like the public sector policies discussed above, the SAEC imposes various security, research purpose, and nonpatenting restrictions on data that is publicly released.[35] It also secures for data-generating scientists a period of exclusivity (up to 12 months), during which they have sole access to the data.[36] During this time they have the ability to analyze data and prepare papers for publication without the threat of being preempted by competing groups.

Another recent development is the Harvard-led Personal Genome Project (PGP).[37] The PGP, launched in 2008 to significant press coverage, solicits volunteers to submit tissue samples and accompanying phenotypic data. Researchers are then authorized to analyze the submitted samples and publish any resulting genomic information on the PGP website. All such data is released without restriction under the "CC0" Creative Commons copyright waiver.[38] The PGP approach differs markedly from that of the government and privately funded projects described above, in that it dispenses entirely with any attempt to restrict the use of its genomic data. PGP requires its contributors to waive all privacy-related rights when contributing their tissue samples to the project, and gives no preference to use of the data by researchers of any kind.

POLICIES OUTSIDE THE UNITED STATES

Although the HGP and subsequent genome sequencing projects relied on international cooperation and collaboration, the data-release policies adopted by groups outside the United States have differed in material ways from corresponding policies adopted by NIH and NHGRI. In particular, non-U.S. funding agencies have generally exhibited less concern with patenting issues, and have remained more flexible with respect to the time frames both for release of data by data generators and embargo periods on publication for data users. A few examples of recent non-U.S. data-release policies are described next.

Genome Canada

Genome Canada, a participant in the HGP, adopted its first formal data-release policy in 2005.[39] While acknowledging the Fort Lauderdale principles, the Canadian policy does not adopt the 24-hour release requirement set forth in the earlier Bermuda Principles. With respect to data generators, Genome Canada "expects data to be released and shared no later

than the original publication date" of the researchers' results, provided that all data must be released "without restriction" by the end of a project. As for patents, Genome Canada "recognizes the need to protect patentable and other proprietary data," and thus requires that data generators' obligation to release data occurs upon the earlier of publication or the filing of a patent application.[40]

Wellcome Trust Case Control Consortium (WTCCC)

The Wellcome Trust is the largest charity in the United Kingdom and the second-largest biomedical research funding charity in the world. Since the beginning of the HGP, the Wellcome Trust has supported genomics initiatives, both through direct funding and through its Sanger Institute in Cambridge, England, a leading sequencing center. In 2006, the Trust funded a large-scale GWA study of seven complex human diseases that was conducted by more than 50 research groups from institutions across the United Kingdom. The study generated a large quantity of data, including aggregated and individual-level genotypic and phenotypic information. Most of this data was made available to the public in accordance with the Fort Lauderdale Principles, and the project designated itself as a CRP.[41] In order to ensure appropriate use of released data, the consortium requires all prospective data users to apply to the consortium's Data Access Committee and sign a written Data Access Agreement.[42] Access to data is granted only to qualified investigators for "appropriate use," as determined by the committee.[43] The data access agreement requires security, acknowledgment, transfer, and use restrictions comparable to those found in the Genetic Association Information Network (GAIN)[44] and other recent policies.[45] It also includes some restrictions that are specific to the study samples, such as a prohibition on any use of data from the 1958 British Birth Cohort for commercial purposes. The agreement does not, however, contain any specific embargo on publication or any restriction on patenting activity.

UK Medical Research Council

In 2008, the UK's Medical Research Council (MRC) released a comprehensive set of guidelines surrounding release of data from MRC-funded research.[46] In a set of broad "data access principles," the MRC announced that data generated by publicly funded research would be a public good and, as such, "must be made available for new research purposes in a timely, responsible manner." Following the reasoning behind the Fort Lauderdale principles, the MRC states that access to data "must balance the interests of data creators, custodians, users and data subjects," and acknowledges that "limited, defined" periods of exclusive use "will often be justifiable." Beyond these broad pronouncements, however, the MRC gives little specific guidance with respect to the timing or manner of data release. Like the WTCCC guidelines, the MRC guidelines place a high value on formal, written agreements to govern the relationships between data generators and data users. Such agreements "must" be used if restrictions on the use of data are to be imposed, and are "particularly important" when publication rights and intellectual property are implicated.[47] By and large, however, the MRC allows individual parties to define the specific requirements of their data-sharing agreements and does not attempt to impose overarching rules regarding the timing of data release.

BEYOND GENOMICS

The success and broad adoption of genomics data-release policies incorporating the Bermuda and Fort Lauderdale principles have led scientists in related fields to consider the

adoption of analogous principles in their own research. One prominent example occurred in 2008, when the NCI convened a meeting of proteomics researchers in Amsterdam to "identify and address potential roadblocks to rapid and open access to [proteomics] data."[48] Participants identified technical, infrastructure, and policy challenges to the rapid release of proteomic data. Technical challenges included the wide variety of disparate platforms and techniques used to generate proteomic data, making "raw" data from experimental instruments difficult to interpret by scientists unfamiliar with, or lacking access to, the instruments used to generate the data. Proteomics also lacks the established public database infrastructure of genomics. Whereas DNA sequence data can be deposited readily in GenBank, the EMBL, or DDBJ, and is often deposited in all three, there is no common public data repository for proteomic data, and existing proteomic databases suffer from inconsistent and sometimes incompatible data formats. Finally, unlike genomics, in which the entire field focused for several years on the single HGP project, proteomics research lacks a unifying policy core and proteomics-focused journals have each developed their own, sometimes inconsistent, guidelines for data submission.

Notwithstanding these difficulties, the Amsterdam participants articulated six data-release and data-sharing principles that reflect the spirit of the Bermuda and Fort Lauderdale principles: (1) timing (should depend on the nature of the effort generating the data, but should in no event be later than publication or, for community resource projects, following appropriate quality-assurance procedures), (2) comprehensiveness (full raw data sets should be released together with associated metadata and quality data), (3) format (standardized formats are encouraged), (4) deposition to repositories (central repositories for proteomic data should be established), (5) quality metrics (central repositories should develop metrics for assessing data quality), and (6) responsibility (scientists, funding agencies, and journals share responsibility for ensuring adherence to community data-release standards).

In 2009, more than one hundred scientists, journal editors, legal scholars, and representatives of governmental and private funding agencies met in Toronto to assess the current state of rapid prepublication data release and the applicability of the Bermuda Principles in projects well beyond the generation of genomic sequence data. The participants reaffirmed a general community commitment to rapid prepublication data release, expanding the scope of projects as to which these principles should apply to all biomedical datasets that "[have] broad utility, are large in scale . . . and are 'reference' in character." Specifically, they cited, in addition to genomic and proteomic studies, structural chemistry, metabolomics, and RNAi datasets, as well as annotated clinical resources such as cohorts, tissue banks, and case-control studies.

The expansion of rapid prepublication data-release principles beyond genomics and proteomics projects, which often have as their ultimate goal the generation of a large data set, to these other areas necessarily raises issues concerning the appropriateness of rapid data release in hypothesis-driven research. Accordingly, the Toronto participants concurred that, while funding agencies should *require* rapid prepublication data release for "broad utility" projects, rapid data release "should not be mandated" for projects that are generally hypothesis-driven. The Toronto participants also addressed the priority concerns of data generators versus data users, observing anecdotally that in many cases data users have, in fact, published papers based on publicly released data sets *before* the publication of the data generators' papers analyzing the data sets themselves, and that this situation caused no "serious damage" to the data generators' subsequent publications. Nevertheless, the participants acknowledged the acceptability of a "protected period" during which data

users could be restricted from publishing on released data sets, cautioning, however, that this period should never exceed one year. The Toronto participants produced a set of "best practices" embodying these principles and applying them to the three constituencies originally identified in Fort Lauderdale—funding agencies, data generators, and data users—as well as to the scientific journals, which were urged to monitor and provide guidance relating to data-release issues.

NOTES

1. Jorge L. Contreras, *Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons*, 12 Minn. J.L. Sci. & Tech. 63 (2011).

2. National Human Genome Research Institute, *1000 Genomes Project data available on Amazon Cloud*, NIH News (Mar. 29, 2012), http://www.nih.gov/news/health/mar2012/nhgri-29.htm.

3. Brad Ozenberger, *TCGA: A Future Arrived*, The Cancer Genome Atlas, http://cancergenome.nih.gov/researchhighlights/leadershipupdate/ozenberger (last visited Sept. 10, 2012).

4. Elizabeth Pennisi, *Will Computers Crash Genomics?*, 331 Sci. 666, 666 (2011).

5. NIH, *DOE Guidelines Encourage Sharing of Data, Resources*, Human Genome News (Oak Ridge Nat'l Laboratory, Oak Ridge, Tenn.), Jan. 1993, at 4.

6. *International Large-Scale Sequencing Meeting*, Human Genome News (Oak Ridge Nat'l Laboratory, Oak Ridge, Tenn.), Apr.–June 1996, at 19.

7. *Summary of Principles Agreed at the First International Strategy Meeting on Human Genome Sequencing,* U.S. Department of Energy Genome Program, http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml (last visited Oct. 28, 2010).

8. National Academy of Sciences, Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age 64 (2009); National Research Council, Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences 75 (2003); National Research Council, Bits of Power—Issues in Global Access to Scientific Data 80–82 (1997); J.H. Reichman & Paul F. Uhlir, *A Contractually Reconstructed Research Commons for Scientific Data in a Highly Protectionist Intellectual Property Environment*, 66 Law & Contemp. Probs. 315, 335 (2003).

9. Report of Meeting organized by the Wellcome Trust, *Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility* (Jan. 14–15, 2003), *available at* http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.

10. Jane Kaye et al., *Data Sharing in Genomics—Re-shaping Scientific Practice*, 10 Nature Rev. Genetics 331, 332 box 1 (2009).

11. Francis S. Collins & Anna D. Barker, *Mapping the Cancer Genome*, Sci. Am., Mar. 2007, at 50.

12. *Types of Data,* The Cancer Genome Atlas Data Portal, http://cancergenome.nih.gov/dataportal/data/about/types/clinical/ (last visited Oct. 28, 2010).

13. *Data Use Certification* 1, The Cancer Genome Atlas Pilot Project (Feb. 22, 2010), http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=DUC&view_pdf&stacc=phs000178.v1.p1.

14. *Data Access*, The Cancer Genome Atlas Data Portal, http://cancergenome.nih.gov/dataportal/data/access/ (last visited Oct. 28, 2010).

15. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS), 72 Fed. Reg. 49,290, 49,294–97 (Aug. 28, 2007) [hereinafter NIH GWAS Policy]; *Modifications to Genome-Wide Association Studies (GWAS) Data Access*, Nat'l Inst. of Health (Aug. 28, 2008), http://grants.nih.gov/grants/gwas/data_sharing_policy_modifications_20080828.pdf [hereinafter *Modifications to GWAS Data Access*].

16. NIH GWAS Policy, *supra* note 15, at 49,295.

17. *Id.* at 49,292.

18. *Id.* at 49,295.

19. *Id.* at 49,296.

20. 5 U.S.C. § 552 (2006).

21. FOIA Exemption 6, 5 U.S.C. § 552(b)(6).

22. Modifications to GWAS Data Access, *supra* note 15; Nils Homer et al., *Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays*, PLoS Genetics (Aug. 2008), http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000167.

23. *Modification to GWAS Data Access, supra* note 15.

24. NIH GWAS Policy, *supra* note 15, at 49,296.

25. *Id.* at 49,297.

26. *Id*. at 49,296.

27. Susan E. Celniker et al., *Unlocking the Secrets of the Genome*, 459 Nature 927 (2009).

28. ENCODE Consortia, Data Release, Data Use, and Publication Policies (2008), *available at* http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf.

29. B. Maher, *The Human Encyclopaedia*, 489 Nature 46 (2012).

30. Peter J. Turnbaugh et al., *The Human Microbiome Project*, 449 Nature 804 (2007); *Human Microbiome Project Awards Funds for Technology Development, Data Analysis and Ethical Research*, NIH News (Oct. 7, 2008), http://www.genome.gov/27528386.

31. *HMP Data Release and Resource Sharing Guidelines for Human Microbiome Project Data Production Grants*, NIH Common Fund, http://commonfund.nih.gov/hmp/datareleaseguidelines.asp.

32. Press Release, Merck & Co., Inc., First Installment of Merck Gene Index Data Released to Public Databases: Cooperative Effort Promises to Speed Scientific Understanding of the Human Genome (Feb. 10, 1995), *available at* http://www.bio.net/bionet/mm/bionews/1995-February/001794.html.

33. Arthur Holden, *The SNP Consortium: Summary of a Private Consortium Effort to Develop an Applied Map of the Human Genome*, 32 BioTechniques 22 (2002).

34. *iSAEC's Background and Organizational Structure*, Int'l SAE Consortium http://www.saeconsortium.org/ (last visited Oct. 28, 2010).

35. *See* Jorge L. Contreras, Aris Floratos & Arthur L. Holden, *The International Serious Adverse Events Consortium's Data Sharing Model*, 31 Nature Biotech. 17–19 (2013).

36. Int'l SAE Consortium Ltd., Data Release and Intellectual Property Policy (last amended Nov. 5, 2009) (on file with author).

37. Personal Genome Project, www.personalgenomes.org/mission.html (last visited Sept. 10, 2012); *see generally* Misha Angrist, Here Is a Human Being: At the Dawn of Personal Genomics (2010) (describing the author's participation in the Personal Genome Project, together with its history and personalities).

38. Creative Commons, http://creativecommons.org/publicdomain/zero/1.0/legalcode (last visited Sept. 10, 2012).

39. Genome Canada, Data Release and Resource Sharing (Sept. 18, 2008), *available at* http://www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf.

40. *Id*.

41. *Publications Policy,* Wellcome Trust Case Control Consortium, https://www.wtccc.org.uk/ccc1/publications_policy_ext.shtml (last visited Oct. 26, 2010).

42. *WTCCC: Access to Genotype Data*, Wellcome Trust Case Control Consortium, https://www.wtccc.org.uk/docs/CDAC_Guidelines_and_Information_July09.pdf (last visited Oct. 26, 2010); *Data Access Agreement*, The Wellcome Trust Case Control Consortium, https://www.wtccc.org.uk/docs/Data_Access_Agreement_v15.pdf (last visited Oct. 26, 2010).

43. *WTCCC: Access to Genotype Data*, *supra* note 42, § 4.

44. *See generally* The GAIN Collaborative Research Group, *New models of collaboration in genome-wide association studies: The Genetic Association Information Network,* 39 Nature Genetics 1045 (2007) (explaining the selection and characteristics of initial GAIN studies, the structure of GAIN, and defining who has access to GAIN data).

45. *Data Access Agreement, supra* note 42.

46. *Principles for Access to, and Use of, MRC Funded Research Data*, Medical Research Council, http://www.mrc.ac.uk/consumption/groups/public/documents/content/mrc003759.pdf.

47. *Id*. at 5 para. 8.

48. Henry Rodriguez et al., *Recommendations From the 2008 International Summit on Proteomics Data Release and Sharing Policy: A Summit Report*, 8 J. Proteomics Res. 3689 (2009).