Contreras JL. Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons. Minnesota Journal of Law, Science & Technology. 2011;12(1):61-125.

Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons

Jorge L. Contreras*

* Acting Director, Intellectual Property Program, and Senior Lecturer in Law, Washington University in St. Louis, School of Law. The author wishes to thank the following for their thoughtful comments on earlier drafts of this paper: Sara Bronin, Rebecca Eisenberg, Terry Fisher, Charlotte Hess, Kimberly Kaphingst, Michael Madison, Charles McManis, Ruth Okediji, the participants in the 2009 SALT Junior Faculty Development Workshop, the 2010 Junior Scholars in Intellectual Property (JSIP) workshop and the 2010 Intellectual Property Scholars Conference (IPSC), and research assistant Jeff Johnson.

I. INTRODUCTION

The multinational effort to sequence the human genome was one of the most ambitious scientific undertakings in history and has been compared to the Apollo manned space program, the Lewis and Clark expedition, and the Manhattan Project.[1] When completed, the Human Genome Project (HGP), which spanned fifteen years[2] and involved over a thousand scientists worldwide, was heralded by President Bill Clinton as "an epoch-making triumph of science and reason."[3] It led to the publication of the first complete human DNA sequence and has resulted in major advances in biochemistry, bioinformatics, and genetics.[4] The project also generated vast quantities of data about the genetic make-up of humans and other organisms, which reside in public databases that are available to any researcher in the world, creating what I refer to as the "genome commons."[5] But, in some respects, even more remarkable than

1. *E.g.*, ARTHUR M. LESK, INTRODUCTION TO GENOMICS 22 (2007); FRANCIS S. COLLINS, THE LANGUAGE OF LIFE 2 (2010); VICTOR K. MCELHENY, DRAWING THE MAP OF LIFE – INSIDE THE HUMAN GENOME PROJECT at ix (2010); James D. Watson, *The Human Genome Project: Past, Present and Future*, 248 SCIENCE 44, 44 (1990).

2. Planning for the HGP began in 1988 and is generally agreed to have concluded in 2002, though work continues to refine the human genomic map. *See, e.g.*, Jeffrey M. Kidd et al., *Mapping and Sequencing of Structural Variation from Eight Human Genomes*, 453 NATURE 56 (2008); Watson, *supra* note 1, at 46; *Major Events in the U.S. Human Genome Project and Related Projects*, U.S. DEPARTMENT OF ENERGY GENOME PROGRAM, http://www.ornl.gov/sci/techresources/Human_Genome/project/timeline.shtml (last visited Oct. 28, 2010).

3. *Reading the Book of Life: White House Remarks on Decoding of Genome*, N.Y. TIMES, June 27, 2000, at F8.

4. *See generally* Francis Collins, *Opinion: Has the Revolution Arrived?*, 464 NATURE 674 (2010) (describing the "profound impact on scientific progress" achieved by the HGP); International Human Genome Sequencing Consortium, *Initial Sequencing and Analysis of the Human Genome,* 409 NATURE 860, 911–13 (2001) [hereinafter *HGP Initial Paper*] (discussing the impact the HGP has had on scientists' ability to find disease genes and drug targets).

5. Jorge L. Contreras, *Prepublication Data Release, Latency, and Genome Commons*, 329 SCIENCE 393, 393 (2010) [hereinafter Contreras, *Prepublication Data Release*]. The term "commons" derives from the traditional designation of shared physical resources such as fields, pastures and forests, but has more recently been applied to intangibles and information, including aggregations of scientific data. *See, e.g.*, Charlotte Hess & Elinor Ostrom, *Introduction: An Overview of the Knowledge Commons*, *in* UNDERSTANDING KNOWLEDGE AS A COMMONS: FROM THEORY TO PRACTICE 4, 12 (Charlotte Hess & Elinor Ostrom eds., 2006) (discussing the evolution of the term "commons"); Michael J. Madison, Brett M. Frischmann & Katherine J. Strandburg, *Constructing*

the impressive *quantity* of data generated by the HGP is the *speed* at which that data has been released to the public.

At a 1996 summit in Bermuda, still early in the HGP, leaders of the scientific community agreed on a groundbreaking set of principles requiring that all DNA sequence data be released in publicly-accessible databases within *twenty-four hours* after generation.[6] These "Bermuda Principles"[7]

Commons in the Cultural Environment, 95 CORNELL L. REV. 657, 659 (2010) [hereinafter *Cultural Commons*] (discussing the governance of cultural and scientific knowledge "commons"); Jorge L. Contreras, *Data Sharing, Latency Variables and Science Commons*, 25 BERKELEY TECH. L.J. (forthcoming 2010) [hereinafter Contreras, *Data Sharing*] (discussing the challenges of creating scientific knowledge commons).

6. *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) [hereinafter *Bermuda Principles*] (reproducing the original report by the Human Genome Organisation (HUGO)). The text of the Bermuda Principles, as reported by the Human Genome Organisation (HUGO), reads, in pertinent part, as follows:

> Primary Genomic Sequence Should be in the Public Domain
>
> It was agreed that all human genomic sequence information, generated by centres funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximise its benefit to society.
>
> Primary Genomic Sequence Should be Rapidly Released
>
> - Sequence assemblies should be released as soon as possible; in some centres, assemblies of greater than 1 Kb would be released automatically on a daily basis.
> - Finished annotated sequence should be submitted immediately to the public databases.
>
> It was agreed that these principles should apply for all human genomic sequence generated by large-scale sequencing centres, funded for the public good, in order to prevent such centres establishing a privileged position in the exploitation and control of human sequence information.

The text of the Bermuda Principles contained in a recent National Research Council report appears to reproduce an earlier, unapproved draft of the Bermuda Principles that contains the apocryphal sentence "[i]t was also agreed that patents should not be sought." NATIONAL RESEARCH COUNCIL, REAPING THE BENEFITS OF GENOMIC AND PROTEOMIC RESEARCH 57, box C (2006) [hereinafter NRC – GENOMIC AND PROTEOMIC RESEARCH]. It is the author's understanding, based on conversations with attendees at the original Bermuda meeting, that this sentence was deleted prior to final approval and does not form part of the generally-accepted text of the Bermuda Principles. Its significance, however, is discussed *infra* Section III.B.

7. These principles are referred to variously in the literature as the Bermuda Principles, the Bermuda Agreement, the Bermuda Resolution, the

contravened the typical practice in the sciences of making experimental data available only *after* publication[8] and were praised by many including President Clinton, who urged "all nations, scientists and corporations to adopt this policy and honor its spirit."[9] The Bermuda Principles represent a significant achievement of private ordering in shaping the practices of an entire industry and establishing a global knowledge resource for the advancement of science. They continue to shape the data release practices of the genomics research community and have established rapid pre-publication data release as the norm in this and other fields.[10] In this

Bermuda Rules, the Bermuda Protocol and the Bermuda Accord. For the sake of consistency, I will use the term "Bermuda Principles" throughout this paper.

8. Prior to the adoption of the Bermuda Principles (and to this day in fields outside of genomics), the data release policies of most government-funded projects allowed researchers to retain their data privately until publication of results or for some specified "exclusivity period", usually in the neighborhood of one year. *See, e.g.*, NATIONAL ACADEMY OF SCIENCES, ENSURING THE INTEGRITY, ACCESSIBILITY, AND STEWARDSHIP OF RESEARCH DATA IN THE DIGITAL AGE 64 (2009) [hereinafter NAS – RESEARCH DATA] (noting that NASA and the European Southern Observatory Administration impose a 12-month proprietary periods and the U.S. National Optical Astronomy Observatory imposes an 18-month proprietary period on the release of data); NATIONAL RESEARCH COUNCIL, SHARING PUBLICATION-RELATED DATA AND MATERIALS: RESPONSIBILITIES OF AUTHORSHIP IN THE LIFE SCIENCES 75 (2003) [hereinafter NRC – SHARING PUBLICATION-RELATED DATA] (describing the one-year "hold allowance" on the deposition of crystallography data into the Protein Data Bank); NATIONAL RESEARCH COUNCIL, BITS OF POWER – ISSUES IN GLOBAL ACCESS TO SCIENTIFIC DATA 80–82 (1997) (describing data release policies of NASA and Global Change Research Program); 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) ("[i]n most cases, publication of research results marks the point at which data produced by government-funded investigators should become generally available").

9. JAMES SHREEVE, THE GENOME WAR 322 (2004) (quoting President Clinton).

10. *See, e.g.*, Collins, *supra* note 4, at 675 (referring to the "radical ethic of immediate data deposit" adopted by the HGP as the current "norm for other community research projects"); Jane Kaye et al., *Data Sharing in Genomics – Re-shaping Scientific Practice*, 10 NATURE REV. GENETICS 331, 332 box 1 (2009) ("[t]hese policies have created a climate in which data sharing has become the default, and [grant] applicants must demonstrate why their data should be exempt from the requirement that it should be deposited for use by other scientists"); Nikos Kyrpides, *Fifteen Years of Microbial Genomics: Meeting the Challenges and Fulfilling the Dream*, 27 NATURE BIOTECHNOLOGY 627, 627 (2009) ("[o]ver time, as the substantial benefits of prepublication release of genome data have been recognized, many funding agencies and most

paper, I offer the first systematic analysis of the social, legal, and political factors that led to the adoption of the Bermuda Principles and the evolution of genomic data release policies over the past two decades.

At the outset of the HGP, policy makers realized that it was necessary to develop efficient systems for coordinating activity among the geographically dispersed laboratories working on the massive project. But project coordination was not the only factor justifying the unorthodox rapid-release requirement of the Bermuda Principles.[11] Rather, this revolutionary approach arose from the belief of several project leaders, both scientists and policy makers, that rapid release of the project's genomic data was desirable for the advancement of scientific discovery and the consequent improvement of human health.[12] Two distinct policy rationales thus emerged to support the rapid data release principles of Bermuda: (1) *project coordination* and (2) *scientific advancement*. Coupled with these, however, was a third distinct policy rationale for rapid data release: (3) *minimizing encumbrances* of DNA sequence data by patents.[13] While this policy objective was seldom stated explicitly, it reflects a current that runs through many of the early (and recent) debates regarding data release.

After the HGP completed its work, the rapid data release principles adopted in Bermuda were exported to other projects involving genomic and related technologies.[14] Advances in science and technology, however, together with increasingly

---

of the large sequencing centers now adhere to the rapid data release policy set forth as the Bermuda Principles in 1996 and renewed in 2003").

11. Though systems for sharing data among participating researchers were used in large-scale scientific projects such as the Manhattan Project and the NASA space launches, the release of data to the *public* was not a priority in these projects.

12. *See, e.g.*, *HGP Initial Paper*, *supra* note 4, at 864 ("[w]e believed that scientific progress would be most rapidly advanced by immediate and free availability of the human genome sequence. The explosion of scientific work based on the publicly available sequence data in both academia and industry has confirmed this judgment").

13. By the late 1980s and the beginning of the HGP, there was already heated debate in the United States regarding the patentability of genetic material. *See* ROBERT COOK-DEEGAN, THE GENE WARS – SCIENCE, POLITICS, AND THE HUMAN GENOME 308–11 (1994); MCELHENY, *supra* note 1, at 117. The increasing trend toward patenting of genetic material alarmed many of the leaders of the HGP. *See infra* note 92 and accompanying text.

14. *See* Kyrpides, *supra* note 10, at 627–28.

complex ethical, legal, and technical issues, have complicated the data release landscape and given rise to additional policy considerations. Among these have been (4) the protection of human subject data that resides in public databases (*data protection*), and (5) the need for scientists generating large data sets to publish their data before it is accessed and used by others in order to facilitate their own career advancement and grant funding (*publication priority*).[15] The emergence and recognition of these considerations has led to an evolution of genomics data release policies. The bold pronouncements made in Bermuda have given way to more nuanced approaches that address differences in types of data and the goals of the projects themselves, as well as the differing and sometimes divergent requirements of data generators and data users.

Elinor Ostrom and her colleagues pioneered the analysis of common resource structures, whether physical or informational, using an organizational theory tool known as the Institutional Analysis and Development (IAD) framework, work that earned her the 2009 Nobel Prize in Economics.[16] More recently, Michael Madison, Brett Frischmann, and Katherine Strandburg have undertaken a thorough re-examination of the IAD framework in relation to commons in the "cultural environment,"[17] seeking to combine the functionalist IAD approach with metaphorical and narrative accounts of commons formation.[18] In this paper, I engage the theoretical framework of Ostrom and Madison, Frischmann, and Strandburg and elucidate both the structural and

15. Among the factors weighing most heavily against rapid data release is the loss by data generators of any "head start" that they might otherwise have had in preparing papers analyzing the released data. That is, under a rapid data release structure, data generators must release their data very shortly after it has been produced, giving competing researchers access to the data at the same time as the scientists who generated it. *See, e.g.*, Contreras, *Data Sharing*, *supra* note 5 (observing that data retention strategies give the data generator a head start with respect to analyzing the data). *See generally infra* Section III.B.2 (discussing the process of publishing of results and the requirements for making the underlying raw data publicly available).

16. Press Release, The Royal Swedish Academy of Sciences (Oct. 12, 2009), *available at* http://nobelprize.org/nobel_prizes/economics/laureates/2009/press.html.

17. *Cultural Commons*, *supra* note 5, at 659. Madison, Frischmann and Strandburg refer to aggregations of shared information as "cultural commons" and include within their far-ranging analysis shared resource structures as varied as patent pools, open source software, Wikipedia, the Associated Press, and jamband fan communities. *Id*. at 660–63.

18. *Id*. at 671–74, 681–83.

narrative elements of the unique developmental history of the genome commons. The IAD methodology offers a systematic means for examining the characteristics of a commons structure: those of the common resource, the "action arena" in which stakeholders interact with the commons and the resulting patterns of interaction.[19] Each of these broad areas is subdivided into further analytical components so that the common resource, for example, is assessed with respect to its bio-physical characteristics, the attributes of the relevant community, and its applicable "rules in use."[20] The application of the IAD framework analysis results in a deeper understanding of the factors that should be considered when structuring or evaluating an information commons.

Consistent with the IAD methodology, I describe in Part II.A the characteristics of the genome commons, including both genomic data and the databases in which it is housed. In Part II.B, I identify and discuss the various stakeholder communities involved in the development and use of the genome commons and their predisposition toward the five principal policy considerations noted above. In Part III, I trace the development of genomic data release policies in the United States, beginning with the HGP and the Bermuda Principles and concluding with current and planned policies both in government-funded and private projects. In Part IV, I analyze the impact of the five policy considerations identified above on the evolving genome commons landscape, particularly in view of the requirements and objectives of the relevant stakeholder communities. I conclude with a number of observations regarding the applicability of these findings to the design of commons in the sciences, generally, and to the future direction of the genome commons.

19. *See* Elinor Ostrom & Charlotte Hess, *A Framework for Analyzing the Knowledge Commons, in* UNDERSTANDING KNOWLEDGE AS A COMMONS: FROM THEORY TO PRACTICE 41, 44–45 (Charlotte Hess & Elinor Ostrom eds., 2006).

20. *Id.* at 45–53.

II. ATTRIBUTES OF THE GENOME COMMONS

A. GENES AND GENOMES

1. Building Blocks[21]

Deoxyribonucleic acid (DNA) is a chemical substance that exists in almost every living organism. Each DNA molecule is composed of four basic building blocks or nucleotides: adenine (A), thymine (T), guanine (G) and cytosine (C). These nucleotides form long strings of linked pairs (A-T and G-C) that are twisted in a ladder-like chain: the famous "double-helix" first described by James Watson and Francis Crick in 1953. Each rung of this ladder is referred to as a "base pair", and the full complement of DNA found within an organism is its "genome". The genome of simple organisms such as the *E. coli* bacterium contains approximately five million base pairs, that of the fruit fly *Drosophila melanogaster* contains approximately 160 million base pairs, and that of *Homo sapiens* contains approximately 3.2 billion base pairs.

The double-helical strands of DNA that exist within an organism's cells are typically bound into discrete units called "chromosomes" (each human carries twenty-three pairs of chromosomes). The DNA on each chromosome is divided into smaller "genes," ranging in size from as few as one hundred to more than two million base pairs. It is currently estimated that humans each possess approximately 25,000 genes, which are generally regarded as the basic functional units of DNA. An organism's genes serve many functions. They are responsible for the inheritance of traits from one generation to the next, and they encode the many proteins responsible for the biochemical functions within the cell. Each human genome is approximately 99.5 percent identical, but very small differences are responsible for the great variability in human physical and physiological traits. The observable characteristics of an individual, including physical, physiological, behavioral, and demographic characteristics, are referred to as that individual's "phenotype." One of the principal goals of genetic science has

21. This Section contains a basic explanation of the scientific terminology and concepts used throughout this paper. Most of this information can be found in any modern biology textbook. In some cases, I have simplified the discussion of complex scientific concepts for the general reader. *See generally* LESK, *supra* note 1; MATTHEW RIDLEY, GENOME 6–10 (1999); WILLIAM S. KLUG & MICHAEL R. CUMMINGS, ESSENTIALS OF GENETICS (3d ed. 1999).

been to associate particular genes, genetic variations, or "mutations" with phenotypic traits.

2. Medical Genetics

As early as 1902, scientists began to associate hereditary diseases with genes passed down from parents to their offspring. But while numerous conditions were associated with patterns of inheritance, from relatively benign traits such as albinism and hair color to debilitating ailments such as cystic fibrosis, Down syndrome, and Huntington's disease, it was not until the 1970s that technology had advanced to a state sufficient to enable scientists to identify the individual genes responsible for these conditions. Even then, each of these discoveries took years of painstaking work and a measure of good luck to achieve. It was not until 1986 that a revolutionary new process for copying DNA fragments called polymerase chain reaction (PCR) enabled the large-scale, rapid sequencing of DNA. The advent of PCR technology soon gave rise to ambitious plans to sequence not only genes identified with specific diseases, but the entire human genome.

3. The Human Genome Project[22]

The Human Genome Project was formally launched in 1990 as a joint project of the National Institutes of Health (NIH)[23] and the U.S. Department of Energy (DOE),[24] with

22. The Human Genome Project, and particularly the race between the publicly-funded HGP and Celera Genomics, has been the subject of numerous popular and scholarly accounts. *See generally* SHREEVE, *supra* note 9; J. CRAIG VENTER, A LIFE DECODED: MY GENOME, MY LIFE (2007); NRC – GENOMIC AND PROTEOMIC RESEARCH, *supra* note 6, at 34–36; INSTITUTE OF MEDICINE & NATIONAL RESEARCH COUNCIL, LARGE-SCALE BIOMEDICAL SCIENCE 31–40 (2003) [hereinafter LARGE-SCALE SCIENCE]; *HGP Initial Paper*, *supra* note 4, at 862–63. The early days of the HGP are extensively chronicled by Robert Cook-Deegan in COOK-DEEGAN, *supra* note 13.

23. The National Institutes of Health (NIH) formed the National Center for Human Genome Research (NCHGR) in 1989, under the direction of James Watson, to carry out its component of the HGP. In 1997, the Department of Health and Human Services elevated NCHGR to the status of a full "institute" within the NIH system, forming the National Human Genome Research Institute (NHGRI). *About NHGRI: A Brief History and Timeline*, NATIONAL HUMAN GENOME RESEARCH INSTITUTE, http://www.genome.gov/10001763 (last visited Oct. 28, 2010).

24. DOE's interest in a genome sequencing project arose from its work on genetic mutations among atomic bomb survivors. *See* COOK-DEEGAN, *supra* note 13, at 93–95. DOE was also the overseer of the GenBank DNA sequence

support from the Wellcome Trust in the United Kingdom and the involvement of groups in the United Kingdom, France, Germany, and Japan.[25] In its initial stages, the HGP sought to build infrastructure, improve sequencing technologies, and sequence the genomes of smaller model organisms. Building on success with these early efforts, the international initiative to sequence the human genome commenced in 1996 with plans to complete the full sequence by 2005.[26]

By 1998, the HGP had spent nearly two billion dollars with relatively little progress other than the sequences for the model organisms.[27] Then, in May, J. Craig Venter, a former NIH scientist, famously proclaimed that he, funded by substantial commercial backers, would utilize a battalion of three-hundred state-of-the-art sequencing machines to sequence the entire human genome in only three years, a full four years before the publicly-funded HGP.[28] Venter's announcement led to a technological "arms race" between his new company, Celera Genomics, and the HGP, a race in which competing claims and accusations became regular features in the scientific literature and the popular press.[29] Ultimately, a truce was declared, and, in June 2000, the leaders of the competing groups made a joint White House announcement that a "first draft" of the human genome sequence had been completed.[30] The draft sequence was published in the public GenBank database in 2001.[31]

---

database at Los Alamos National Laboratory, which it established in 1983. LARGE-SCALE SCIENCE, *supra* note 22, at 31. *See generally* Stephen Hilgartner, *Potential Effects of a Diminishing Public Domain in Biomedical Research Data*, *in* NATIONAL RESEARCH COUNCIL, THE ROLE OF SCIENTIFIC AND TECHNICAL DATA AND INFORMATION IN THE PUBLIC DOMAIN: PROCEEDINGS OF A SYMPOSIUM 137 (2003) (describing the history of GenBank and its predecessor, the Los Alamos Sequence Library).

25. SHREEVE, *supra* note 9, at 45–47.

26. *See* NRC – GENOMIC AND PROTEOMIC RESEARCH, *supra* note 6, at 35.

27. Nicholas Wade, *Scientist's Plan: Map All DNA Within 3 Years*, N.Y. TIMES, May 10, 1998, at 20.

28. SHREEVE, *supra* note 9, at 22–23; Leslie Roberts, *Controversial from the Start*, 291 SCIENCE 1182, 1187; Wade, *supra* note 27, at 1.

29. *See* Roberts, *supra* note 28, at 1188.

30. Roberts, *supra* note 28, at 1188; Nicholas Wade, *Genetic Code of Human Life is Cracked by Scientists: A Shared Success*, N.Y. TIMES, June 27, 2000, at A1.

31. *See HGP Initial Paper*, *supra* note 4.

4. The Post-Genome World

The completion of the human genome sequence has had a significant impact on biomedical science.[32] The genetic basis for thousands of common hereditary diseases is now known, and widely-available genetic tests exist for many common diseases and other physical traits.[33] Related fields such as proteomics (the study of protein expression throughout an organism) have also benefitted from the technological and scientific advances made possible by the HGP.[34] Today, additional international efforts are under way to sequence the genomes of one thousand individual humans to create the most complete and detailed reference map of the human genome to-date (the "1000 Genomes Project")[35] and to sequence the genomes of some of the multitude of microorganisms residing within the human body (the "Human Microbiome Project").[36]

The public human genome map has also enabled researchers to conduct studies to determine complex combinations of genetic factors contributing to disease. Whereas earlier studies took years to identify single genes responsible for specific inherited diseases, recent "genome-wide association studies" (GWAS or GWA studies) have been credited with identifying variants in multiple genes that increase susceptibility for complex conditions such as Type 2 diabetes,[37] breast cancer,[38] prostate cancer,[39] hypertension,[40]

32. *See, e.g.*, COLLINS, *supra* note 1, at 3 ("[v]irtually all biomedical researchers would agree that their approach to understanding how life works has been profoundly and irreversibly affected by access to the complete DNA sequence of the human genome, and that of many other organisms").

33. As of December 6, 2009, NCBI's GeneTests web site identified 1830 different diseases for which genetic tests are available. GENE TESTS, http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests (last visited Dec. 6, 2009).

34. *See* NRC – GENOMIC AND PROTEOMIC RESEARCH, *supra* note 6, at 38–40; LESK, *supra* note 1, at 305–07.

35. Erika Check Hayden, *International Genome Project Launched*, 451 NATURE 378, 378 (2008).

36. Peter J. Turnbaugh et al., *The Human Microbiome Project*, 449 NATURE 804, 804 (2007).

37. Laura J. Scott et al., *A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants*, 316 SCIENCE 1341 (2007); Robert Sladek, *A Genome-Wide Association Study Identifies Novel Risk Loci for Type 2 Diabetes*, 445 NATURE 881(2007).

38. D.F. Easton et al., *Genome-Wide Association Study Identifies Novel Breast Cancer Susceptibility Loci*, 447 NATURE 1087 (2007); D.J. Hunter et al.,

and numerous other diseases.[41] Such studies, which involve scanning the entire human genome for variants that are common among persons with similar diseases or other observable traits, have been made possible by dramatic advances in the technology used to sequence and analyze the vast quantities of data embedded within human DNA and similarly dramatic reductions in the cost of sequencing technology.[42]

B. DATA AND DATABASES

1. Publication of Results

The peer-reviewed journal article is the traditional means of disseminating scientific information.[43] Scientists are judged, both for purposes of career advancement and the awarding of government grants, on the quantity of their publications,

---

A Genome-Wide Association Study Identifies Alleles in FGFR2 Associated with Risk of Sporadic Postmenopausal Breast Cancer, 39 NATURE GENETICS 870 (2007).

39. Meredith Yeager et al., *Genome-Wide Association Study of Prostate Cancer Identifies a Second Risk Locus at 8q24*, 39 NATURE GENETICS 645 (2007).

40. Adebowale Adeyemo et al., *A Genome-Wide Association Study of Hypertension and Blood Pressure in African Americans*, PLOS GENETICS (July 2009), http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000564.

41. *See, e.g.*, The Wellcome Trust Case Control Consortium, *Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls*, 447 NATURE 661 (2007); Monya Baker, *Genetics by Numbers*, 451 NATURE 516 (2008) (discussing GWA study of several common diseases); Lucia A. Hindorff et al., *Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits*, 106 PROCEEDINGS OF THE NAT. ACAD. SCI. 9362 (2009) (discussing an online catalog of GWAS association data that references hundreds of publications identifying more than 100 diseases and traits).

42. In 1985, the cost of sequencing a single human DNA base pair was approximately $10.00. That cost decreased to $1.00 by 1991, $0.10 by 1993, and approximately $0.001 by 2006. LESK, *supra* note 1, at 23. Between 1999 and 2009, the cost of gene sequencing technology dropped by an astonishing factor of 14,000. Collins, *supra* note 4, at 674. The NHGRI is currently funding the development of technology capable of sequencing an entire human genome (approximately 3.2 billion base pairs) for a cost of $1,000. *See* Collins, *supra* note 4, at 675.

43. *See* ROBERT K. MERTON & HARRIET ZUCKERMAN, INSTITUTIONALIZED PATTERNS OF EVALUATION IN SCIENCE (1971), *reprinted in* THE SOCIOLOGY OF SCIENCE 460, 463–65 (Norman W. Storer ed., 1973) (tracing the origin of scientific publication to the advent of printing and the establishment of the first scientific journals in 1665).

making the publication of articles of paramount importance to many scientists and giving scientists a significant personal incentive to publish and, thus, share their data with others.[44] A significant period of time, however, typically elapses between the point at which experimental data are generated and the time that they are published. This delay reflects the time required for the investigators to analyze their results, gather additional data, refine their analysis, prepare a paper based on their findings, and submit the paper to journals; for the journals to conduct their peer review and editorial process; for the investigators to make any revisions required by the journals (including, at times, to conduct additional experiments) or, if the paper is rejected by the journal, to revise and submit it to different journals; and, finally, for the journal to edit, format, and prepare the accepted paper for publication. One recent study reports that the period from completion of scientific work until publication is typically between twelve and eighteen months.[45] Older studies have found comparable or longer delays in other fields of research.[46]

44. ROBERT K. MERTON, PRIORITIES IN SCIENTIFIC DISCOVERY (1957), *reprinted in* THE SOCIOLOGY OF SCIENCE 286, 316 (Norman W. Store red., 1979) (noting the "tendency, in many academic institutions, to transform the sheer number of publications into a ritualized measure of scientific or scholarly accomplishment"); RESEARCH INFO. NETWORK, TO SHARE OR NOT TO SHARE: PUBLICATION AND QUALITY ASSURANCE OF RESEARCH DATA OUTPUTS 25 (2008), *available at* www.rin.ac.uk/data-publication (the assessment of researchers is "perceived to value above all else the publication of papers in high-impact journals").

45. Carlos B. Amat, *Editorial and Publication Delay of Papers Submitted to 14 Selected Food Research Journals. Influence of Online Posting*, 74 SCIENTOMETRICS 379 (2008).

46. *See* William D. Garvey & Belver C. Griffith, *Scientific Information Exchange in Psychology*, 146 SCIENCE 1655, 1656 (1964) (reporting that in the psychology field, their study indicated that the time between hypothesis and publication is between 30 and 36 months, and the time between reportable results and publication is between 18 and 21 months); Charles G. Roland & Richard A. Kirkpatrick, *Time Lapse Between Hypothesis and Publication in the Medical Sciences*, 292 NEW ENG. J. MED. 1273, 1274 (1975) (finding delays of 20 and 24 months between the completion of research and publication, respectively, for medical laboratory research and clinical research studies). Anecdotally, the author has been informed that publication delays are typically even longer in the social sciences.

2. Raw Data

Despite the abundant incentives for scientists to share data via publication, the data set published in most journal articles represents only a small portion of the "raw" data collected in a given research project.[47] This data set is typically presented in a summary fashion and is intended primarily to support the scientist's analysis and conclusions.[48] Yet in order to enable the verification and reproduction of an experiment by other scientists, the full data set is often required. Thus, a growing number of scientific journals now require that authors make the data underlying their published results available to readers as well.[49] In the case of genomic sequence data, journals often require a deposit of the data at the time of publication into a public database,[50] such as NIH's Genbank[51] or dbGaP[52]. These requirements, coupled with the funding agency's data release requirements described below, have enabled the efficient, rapid, and cost-effective sharing of new knowledge and the pursuit of studies and analyses that

47. *See generally* Rebecca S. Eisenberg, *Patents and Data-Sharing in Public Science*, 15 INDUS. & CORP. CHANGE 1013, 1024 (2006).

48. *Id.*

49. *See, e.g.*, *Guide to Publication Policies of the Nature Journals*, NATURE, http://www.nature.com/authors/gta.pdf (last updated Apr. 30, 2009); *General Information for Authors*, AM. ASS'N FOR THE ADVANCEMENT OF SCI., http://www.sciencemag.org/about/authors/prep/gen_info.dtl (last visited Oct. 27, 2010); *Information for Authors*, PROCEEDINGS OF THE NAT'L ACAD. OF SCI., http://www.pnas.org/site/misc/iforc.shtml#viii (last visited Oct. 27, 2010).

50. *See* Hilgartner, *supra* note 24, at 137.

51. Today, GenBank is administered by the National Center for Biotechnology Information (NCBI) and forms one of three international nucleotide libraries that work in close partnership, and the others are the European Molecular Biology Library (EMBL) in Hinxton, England, and the DNA Data Bank of Japan (DDBJ). *See* LESK, *supra* note 1, at 251. The quantity of data in GenBank increased from about 2 billion base pairs in 1999 to 86 billion in 2008. Mike May, *Sharing the Wealth of Data*, SCI. AM. WORLDVIEW 88, 89 (2009).

52. The Database of Genotypes and Phenotypes (dbGaP) operated by the NIH's National Library of Medicine can accommodate phenotypic data, which includes elements such as de-identified subject age, ethnicity, weight, demographics, exposure, disease state, and behavioral factors, which is far more complex to record, search and correlate than the raw sequence data deposited in GenBank, and in addition to genotypic and phenotypic data, dbGaP can accommodate study documentation and statistical results, including linkage and association analyses. *See generally* DBGAP, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gap (last visited Oct. 27, 2010).

otherwise might have been impossible.[53] As Hess and Ostrom observe, modern biology has been transformed into an "*information* science."[54]

C. ACTORS AND STAKEHOLDERS

Much of the early work regarding common resource structures was devoted to understanding the attributes of the different communities that shared the commons, whether herdsmen grazing cattle on a common pasture or fishermen trolling ocean stocks.[55] This analysis is equally valuable in the context of the information commons, and in designing "rules in use," policy makers must consider the interests of the different communities that both use and develop the common resource, including which interests may be overlapping, divergent, and, sometimes, contradictory.[56] The principal stakeholder communities relevant to the genome commons, both initially and as it has evolved over time, include the following:

1. Funders

The HGP, which cost over $2 billion to complete, has been called "the largest and most visible large-scale science project in biology to date."[57] As such, the U.S. governmental agencies that funded the bulk of the massive project, together with their counterparts at the Wellcome Trust in the United Kingdom, exerted a significant degree of control over both its technical and policy dimensions.[58] Consistent with the perceived importance of the project, NIH appointed "James Watson, Nobel laureate and co-discoverer of the [double-]helical structure of DNA," to oversee the newly-formed National

53. *See* Eisenberg, *supra* note 47, at 1020.

54. Charlotte Hess & Elinor Ostrom, *A Framework for Analysing the Microbiological Commons*, 58 INTL. SOC. SCI. J. 335, 335 (2006).

55. *See, e.g.*, *Cultural Commons*, *supra* note 5.

56. Both Ostrom and Hess and Madison, Frischmann and Strandburg emphasize the importance of identifying the various constituencies connected with a cultural commons. Ostrom & Hess, *supra* note 19, at 48–50; *Cultural Commons*, *supra* note 5, at 690. *See also* Contreras, *Data Sharing*, *supra* note 5, at 40–41.

57. LARGE-SCALE SCIENCE, *supra* note 22, at 29.

58. The Wellcome Trust in the U.K., at that time the world's largest private medical charity, also contributed substantial funding and support to the project, primarily to the work conducted at the Sanger Centre in Cambridge, England. LARGE-SCALE SCIENCE, *supra* note 22, at 39.

Center for Human Genome Research in 1988.[59] Other scientists involved in the early planning and execution stages of the project were also globally prominent and included a significant number of Nobel Prize winners.[60] This leadership by preeminent and respected scientists was critical to the HGP and gave the group's decisions a *gravitas* that they otherwise might have lacked. It also engendered among the project's leadership a sense of public stewardship that contributed to the public-spirited nature of several HGP policies.[61]

2. Data Generators

Prior to the HGP, genetic research was conducted in hundreds of academic laboratories across the world and funded primarily by small grants directed toward the investigation of specific hypotheses.[62] The HGP, by contrast, treated the mapping of the human genome as a campaign of large-scale data production.[63] The NIH funded three major genome centers (Baylor College of Medicine, Washington University, and the Whitehead Institute) that worked closely with the DOE's Joint Genome Institute and the Sanger Centre in Cambridge, England (funded by the Wellcome Trust).[64] These five centers produced the majority of the data that resulted from the

59. *Id.* at 35. When Watson resigned in 1992 following a dispute over the NIH's attempts to patent small DNA fragments known as expressed sequence tags (ESTs), Francis Collins, another high-profile scientist, was appointed to replace him. *Id.* at 36–37.

60. In addition to Watson (Chemistry, 1962), the HGP leadership group included Fred Sanger (Chemistry, 1958 and 1980), Hamilton Smith (Medicine, 1978) and Walter Gilbert (Chemistry, 1980). Other scientists involved in the HGP won the Nobel Prize after the commencement of the project (e.g., John Sulston (Medicine, 2002)). *See generally* Robert Mullan Cook-Deegan, *Origins of the Human Genome Project*, 5 RISK 97 (1994).

61. For instance, in 1988, James Watson allocated 3% of the HGP budget to investigate the ethical and social implications of sequencing the human genome, creating the Ethical, Legal and Social Implications (ELSI) group within the HGP, and the budget for ELSI was later raised to 5% of the HGP budget, indicating the importance HGP leadership placed on the social impact of the HGP. *See* James D. Watson, *Genes and Politics*, 75 J. MOLECULAR MED. 624, 633-34 (1997); Eric T. Juengst, *Self-Critical Federal Science? The Ethics Experiment Within the U.S. Human Genome Project*, 13 SOC. PHIL. & POL'Y 63, 63 (1996); *see also* Peter Lee, *Toward a Distributive Commons in Patent Law*, 2009 WIS. L. REV. 917, 950-67 (2009) (analyzing the distributive justice interests of public institutions which fund scientific research).

62. *See* Roberts, *supra* note 28, at 1185.

63. *Id.* at 1182.

64. *See* LARGE-SCALE SCIENCE, *supra* note 22, at 39.

HGP.[65] The intensity of this work, the amount of capital equipment required to undertake it, and the degree of specialization required by the new science of genomics led to the creation of a new breed of scientist: one whose principal research aim was the generation of large data sets rather than the development and testing of hypotheses. This distinction persists today as the number of data-generating projects in the biosciences continues to increase.[66] The factors motivating these data-generating scientists are twofold: (1) obtaining continued grant funding for their work and (2) advancing their careers through publication and peer recognition. But while governmental funding of new data generation projects continues, data generating scientists face challenges when it comes to publishing their work in traditional scientific journals.[67]

3. Data Users

Prior to the completion of the HGP, researchers studying a particular genetic disease devoted substantial time and effort to isolating and sequencing the relevant gene—work that would often take years of painstaking trial-and-error experimentation.[68] The data generated by the HGP and its follow-up projects have eliminated the need for researchers to conduct much of this groundwork.[69] Unlike the close-knit community of data generators at large-scale sequencing centers, there is no coherent community of data users. These users comprise all scientists across the world whose research may benefit from the use of genomic data.

65. *Id.*

66. The implications of participating in large-scale data generating work on the careers of junior scientists have been the subject of much discussion. *See* LARGE-SCALE SCIENCE, *supra* note 22, at 26–27; Kaye et al., *supra* note 10, at 332–33; Toronto Int'l Data Release Workshop Authors, *Pre-Publication Data Sharing*, 461 NATURE 168, 169–70 (2009) [hereinafter *Toronto Report*].

67. *See* Contreras, *Prepublication Data Release*, *supra* note 5, at 393; Contreras, *Data Sharing, supra* note 5, at 38.

68. *See* SHREEVE, *supra* note 9, at 40.

69. *Id.*

4. Data Intermediaries

Individual scientists and laboratories that generate data are seldom the ones that make such data available to others, except in limited one-on-one interactions with colleagues. In most cases, scientists rely on data intermediaries, whether scientific journals that publish their analyses and results or centralized database managers that host large quantities of raw data. Data intermediaries may operate either as commercial entities (as in the case of commercial publishers and paid database services) or non-profit/governmental entities (such as the GenBank and dbGaP databases and "open access" journals such as those published by the Public Library of Science (PLoS)[70]). Not surprisingly, the interests of commercial and non-commercial data intermediaries differ in several regards, most notably in the area of pricing for access to information. Nevertheless, these stakeholders also share a number of common traits, including the desire to disseminate information in ways that are effective, secure, and accurate and the need to maintain some level of financial stability.[71]

5. Data Subjects

Human genomic information, by definition, is derived from human subjects. Because the goal of the HGP was to generate a baseline map of the human genome without regard to the particular physiological and pathological traits associated with genetic variation among individuals, the genomic sequence data generated by the HGP was anonymous and retained no association with the individual subjects whose DNA was sequenced.[72] Similar characteristics applied to other early

70. *See* Contreras, *Data Sharing*, *supra* note 5, at 38.

71. Subscription costs of scientific journals, particularly those of commercial publishers, have risen sharply in recent years, and the scientific publishing industry often complains of thinning margins and rising expenses, issues not unique to the commercial sector while some publicly-funded databases also suffer from funding shortfalls and are in danger of discontinuation. *See, e.g.*, Joan B. Schlimgen et al., *Update on Inflation of Journal Prices: Brandon/Hill List Journals and the Scientific, Technical, and Medical Publishing Market*, 92 J. MED. LIBR. ASS'N 307 (2004) (analyzing the rising costs of scientific journals and the pressures causing that rise); Editorial, *Access Denied?*, 462 NATURE 252 (2009) (describing the threatened demise of the NSF-funded TAIR arabidopsis plant genome database).

72. *The Human Genome Project Completion: Frequently Asked Questions*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/11006943 (last visited Oct. 28, 2010).

genomic projects such as the HapMap Project.[73] These data were intended to elucidate non-individualized information applicable to the human genome in general. In later projects, however, and particularly with the commencement of large-scale GWA studies, concerns with the potential identification of human subjects grew because the genotypic data generated by a GWA study is not meaningful without the associated phenotypic data.[74] That is, because a GWA study often seeks to *associate* genotypic information (e.g., genetic markers) with particular disease states, information regarding donor demographics, disease state, and treatment are necessary to interpret the genotypic findings. The prospect of releasing clinical and phenotypic data to the public sparked substantial concern and has led to the recognition of human data subjects as important stakeholders in the genomic data equation.[75] Public concern has only been heightened by the publication in 2008 of a paper suggesting that the presence of an identifiable individual's DNA can be inferred from a group of samples using statistical techniques.[76] Such findings suggest that the interests of data subjects may require substantial attention as genomic science advances and have led to numerous proposals for heightened protection of individual identity in publicly-released genomic data.[77]

73. *See* Eisenberg, *supra* note 47, at 1026.

74. *See Toronto Report*, *supra* note 66, at 170.

75. For a general discussion of the protection of human subjects data in genomic studies, a topic that is beyond the scope of this paper, but which has been extensively addressed in the literature. *See, e.g.*, LORI B. ANDREWS, MAXWELL J. MEHLMAN & MARK A. ROTHSTEIN, GENETICS: ETHICS, LAW AND POLICY 592–630 (1st ed. 2002); Domenic A. Crolla, *Reflections on the Legal, Social, and Ethical Implications of Pharmacogenomic Research*, 46 JURIMETRICS 239, 241–47 (2006); John A. Robertson, *Privacy Issues in Second Stage Genomics*, 40 JURIMETRICS 59 (1999).

76. 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.

77. *See, e.g.*, P3G Consortium et al., *Public Access to Genome-Wide Data: Five Views on Balancing Research with Privacy and Protection*, PLOS GENETICS (Oct. 2009), http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000665.

6. The Public

The general public cannot be ignored as a key stakeholder with respect to genomic research. The HGP generated significant public interest and was regularly covered by the popular news media.[78] Beyond general interest, however, are two significant aspects of public engagement with genomics. First, government-sponsored research is largely taxpayer-funded, meaning that public taxpayers and their representatives in Congress have a legitimate and significant interest in the direction and results of research.[79] Second, members of the public who are themselves affected, directly or indirectly, by genetic disorders or diseases often form patient advocacy and disease interest groups. These groups frequently possess a high degree of familiarity with the relevant scientific literature and have both the motivation and the financial means to lobby for changes in research policy.[80]

III. THE EVOLUTION OF RAPID, PRE-PUBLICATION DATA RELEASE IN THE GENOME SCIENCES

A. EARLY YEARS OF THE HGP

Since the initiation of the HGP, several factors contributed to the call to release the data generated by the project to the public. First, the early work of the HGP involved sequencing the genomes of simple model organisms including the roundworm (*C. elegans*) and mouse (*Mus musculus*). The

78. *See, e.g.*, Nicholas Wade, *Genome's Riddle: Few Genes, Much Complexity*, N.Y. TIMES, Feb. 13, 2001, at F1, *available at* http://www.nytimes.com/2001/02/13/health/13HUMA.html; Justin Gillis, *Scientists Speed Up Timetable for Mapping Human Genes*, WASH. POST, Sept. 15, 1998, at A2; Leon Jaroff et al., *Science: The Gene Hunt*, TIME, Mar. 20, 1989, at 62, *available at* http://www.time.com/time/magazine/article/0,9171,957263,00.html.

79. *See, e.g.*, Jonathan Karl et al., *Stimulus Slammed: Republican Senators Release Report Alleging Waste*, ABC NEWS, August 3, 2010, *available at* http://abcnews.go.com/GMA/stimulus-slammed-republican-senators-release-report-alleging-waste/story?id=11309090 (detailing public and Congressional criticism of research on topics such as cocaine use in monkeys, collection of exotic ants and the use of yoga among cancer survivors).

80. *See* Lee, *supra* note 61, at 986–90 (addressing the interests and policy concerns of disease advocacy groups); and *see e.g.*, Sharon F. Terry et al., *Advocacy Groups as Research Organizations: The PXE International Example*, 8 NATURE REVIEWS GENETICS 157, 157–162 (2007) (describing the experience of an advocacy organization for the disease pseudoxanthoma elasticum and the methods the group used to advance a scientific agenda).

groups that worked on these organisms abided by strong "open science" norms and were accustomed to sharing their data freely with one another, laying a strong precedent for the HGP.[81] Moreover, and perhaps more importantly, there was a sense among the leadership of the project, in the words of Ari Patrinos, the DOE's Associate Director for Biological and Environmental Research, that "the genome belongs to everybody."[82] Accordingly, in 1988 the National Research Council recommended that all data generated by the HGP "be provided in an accessible form to the general research community worldwide."[83]

In 1992, shortly after the project was launched, NIH and DOE developed formal guidelines for the sharing of HGP data.[84] These guidelines were viewed as essential to achieve the program's goals, avoid unnecessary duplication of effort, and expedite research in other areas.[85] In other words, the putative purpose of these guidelines was to facilitate the straightforward policy goal of *project coordination.* The guidelines required that data generated by the HGP be deposited in public databases (e.g., GenBank), making it available to all scientists worldwide.[86] But the need for project coordination did not require immediate *public* release of the

81. *See HGP Initial Paper*, *supra* note 4, at 864; MCELHENY, *supra* note 1, at xi ("Openness was at the core of the [bacteriophage] ethos, and it soon propagated to the genetic research systems of the future."); Hilgartner, *supra* note 24, at 89 ("There were . . . communities doing molecular biology . . . on yeast and Drosophila that had "open science" norms. Those norms were the ones adopted as the models for the Human Genome Project."). The evolution of the open science culture among *C. elegans* researchers is described in some detail in NRC - GENOMIC AND PROTEOMIC RESEARCH, *supra* note 6, at 54–56.

82. Eliot Marshall, *Bermuda Rules: Community Spirit, With Teeth*, 291 SCIENCE 1192 (2001). James Watson, then-director of the National Center for Human Genome Research, wrote in 1990 that "making the sequences widely available as rapidly as practical is the only way to ensure that their full value will be realized and is the only acceptable way to handle information produced at public expense." Watson, *supra* note 1, at 48.

83. NAT'L RESEARCH COUNCIL, MAPPING AND SEQUENCING THE HUMAN GENOME 8 (1988) [hereinafter NRC – HUMAN GENOME] (arguing that the project's mapping and sequencing data will be "of little value" if not made accessible to the general research community).

84. *NIH, DOE Guidelines Encourage Sharing of Data, Resources*, HUMAN GENOME NEWS (Oak Ridge Nat'l Laboratory, Oak Ridge, Ten.), Jan. 1993, at 4 [hereinafter *NIH/DOE Guidelines*].

85. *Id.*

86. *Id.*

HGP data. The HGP policy makers in 1992 recognized the need to provide data generators with "some scientific advantage from the effort they have invested" in generating the data.[87] This "advantage" manifested itself in a six-month maximum period from the time that HGP data are generated until the time that they must be made publicly available. During this six-month period, HGP researchers could analyze their data and prepare publications. Only after the end of the six-month period were they required to release the data to the public.[88]

The 1992 guidelines, in sharp contrast with later policies, also indicate that the agencies would not disfavor investigators that wished to secure patent rights in HGP-funded discoveries.[89] This patent-friendly attitude manifested itself in NIH's nearly disastrous attempt to seek patents on short genetic sequences known as expressed sequence tags (ESTs). This effort began in 1991, when NIH filed patent applications claiming 337 ESTs identified, ironically, by Craig Venter's research group. NIH announced this filing as well as its intention to continue to file EST patent applications on a monthly basis.[90] The public response to this announcement was vociferous and triggered what Robert Cook-Deegan describes as "an international firestorm."[91] The debate within NIH was equally vehement and ultimately led to James Watson's resignation from the agency that oversaw the HGP.[92] The EST debacle marked a turning point in NIH's attitude toward patents on genetic material. By 1994, a significantly cowed NIH elected not to appeal the Patent and Trademark Office's

87. *Id.*

88. *Id.*

89. *Id.* ("[I]ntellectual property protection may be needed for some of the data and materials.").

90. *See* Thomas Barry, *Revisiting Brenner: A Proposed Resolution to the Debate Over the Patentability of Expressed Sequence Tags Using the Concept of Utility Control*, 35 AIPLA Q.J. 1, 11 (2007).

91. *See* COOK-DEEGAN, *supra* note 13, at 330–31 (detailing international responses to NIH's EST patent applications including UK threats to file countervailing patent applications, UK and French efforts to forge an international anti-patenting agreement, public commitments by Japanese investigators not to pursue patents and pronouncements from various international scientific conferences).

92. Watson decried the EST patenting plan as "sheer lunacy." SHREEVE, *supra* note 9, at 84–85. The NIH's and DOE's own advisory committees were "unanimous in deploring the decision to seek such patents." COOK-DEEGAN, *supra* note 13, at 317.

rejection of its initial EST patent applications,[93] and, since then, it has adopted a consistently lukewarm, if not outright averse, attitude toward the patenting of genetic sequences.[94] This attitude is reflected in NIH's support for the Bermuda Principles and in the data release and patent policies adopted by NIH in the years thereafter.

B. THE BERMUDA PRINCIPLES

1. The Birth of Rapid Pre-Publication Data Release

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 fifty scientists and policy-makers met in Hamilton, Bermuda[95] 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.[96] The resulting Bermuda Principles established that all DNA sequence information from large-scale human genomic sequencing

93. *See* LARGE-SCALE SCIENCE, *supra* note 22, at 36–37. The patentability of ESTs has subsequently been addressed by the U.S. Court of Appeals for the Federal Circuit in *In re Fisher*, 421 F.3d 1365, 1374 (Fed. Cir. 2005) (holding that the claimed ESTs do not meet the utility requirement of 35 U.S.C. § 101 because they do not identify the function for the underlying protein-encoding genes).

94. In 1999, based partially on its experience with the EST patent applications, NIH formally urged the PTO to impose stricter utility standards when considering DNA-based patents. *See* NRC - GENOMIC AND PROTEOMIC RESEARCH, *supra* note 6, at 53. For an overview of legal objections to the practice of patenting ESTs, see *id*. at 52, and Barry, *supra* note 90, at 18–21.

95. The International Strategy Meeting on Human Genome Sequencing meeting was sponsored by the Wellcome Trust and included representatives of NIH and DOE, the Wellcome Trust, UK Medical Research Council, the German Human Genome Programme, the European Commission, the Human Genome Organisation (HUGO) and the Human Genome Projects of France and Japan. In addition to the data release issues addressed in this paper, and for which the meeting is best known, attendees also discussed and debated issues relating to sequencing strategies, software tools and informatics methodologies. *See International Large-Scale Sequencing Meeting*, HUMAN GENOME NEWS (Oak Ridge Nat'l Laboratory, Oak Ridge, Ten.), Apr.–June 1996, at 19.

96. *See* Marshall, *supra* note 82, at 1192; Robert Cook-Deegan & Stephen J. McCormack, *A Brief Summary of Some Policies to Encourage Open Access to DNA Sequence Data*, 293 SCIENCE 217 supp. (2001), *available at* http://www.sciencemag.org/cgi/content/full/293/5528/217/DC1.

projects should be "freely available and in the public domain in order to encourage research and development and to maximize its benefit to society."[97] 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[98] should be released automatically *within twenty-four hours* and that finished annotated sequences should be submitted *immediately* to a public database.[99]

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. Elimination of the six-month data holding period established in 1992 was supported by both the NIH and DOE and had significant international ramifications.[100] Even Craig Venter and Celera Genomics eventually agreed to make the data from their competing effort to sequence the human genome available to the public.[101]

97. *Bermuda Principles*, *supra* note 6.

98. *Id.* One kilobase (Kb) represents 1,000 base pairs. The human genome consists of approximately 3.2 billion base pairs. One Kb is thus a very small increment of the genetic code that corresponds to an initial "read" by gene sequencing technology of the 1990s. At a follow-up meeting held in Bermuda in 1997, this requirement was changed to apply to sequence assemblies of 2 Kb or more in size to ensure that the released sequences include at least two sequence reads for greater reliability.

99. *Id.*

100. Among other things, the Bermuda Principles contributed to the German government's 1997 decision to revoke its rule granting German companies three months privileged access to human genome sequence data generated with German government funding. Allison Abbott, *Germany Rejects Genome Data 'Isolation'*, 387 NATURE 536, 536 (1997).

101. Though Celera ultimately made its sequence data publicly-available, the path that led to this result was bumpy and circuitous. Unlike the public HGP, Celera offered its data on a commercial web site, rather than the public GenBank database. Celera allowed scientists from non-profit and academic institutions to access it without charge but required that scientists who wished to use the data for commercial purposes enter into a license agreement. Eliot Marshall, *Storm Erupts over Terms for Publishing Celera's Sequence*, 290 SCIENCE 2042, 2042 (2000). This approach outraged much of the scientific community and led to a highly-publicized debate. Ultimately a settlement was brokered by the journal *Science*, which published Celera's article announcing its draft of the human genome sequence, provided that the company make its data broadly available (the competing HGP article was published by *Nature* on the same day). Eliot Marshall, *Sharing the Glory, Not the Credit*, 291 SCIENCE 1189–93 (2001). Celera's subscription-based data business was ultimately unsuccessful and, in 2005, the company finally released its human, rat and mouse genomic data to GenBank. Jocelyn Kaiser, *Celera to End Subscriptions and Give Data to Public GenBank*, 308 SCIENCE

The Bermuda Principles achieved several of the most important policy objectives held by the HGP funders. First, they critically enhanced *project coordination* by enabling the HGP sequencing centers to obtain regularly-updated data sets from one another to avoid duplication of effort and to optimize their respective tasks.[102] Waiting six months to obtain data under the 1992 policy was simply not practical if the project were to function effectively. Second, the funders, particularly the prominent leaders chosen to lead the HGP, argued that rapid data release was the best way to maximize *scientific advancement* (i.e., putting sequence data into the hands of as many laboratories as possible as quickly as possible to accelerate the solution of problems for the benefit of society).[103]

Finally, rapid data release under the Bermuda Principles severely limited the ability of private parties to obtain patent protection on data generated by the HGP, thus satisfying the policy goal of *minimizing encumbrances* that was deeply held by several HGP leaders.[104] In particular, the Bermuda Principles ensured that HGP data would be made publicly-available before data generators could file patent applications covering "inventions" arising from that data and in a manner that ensured its availability as prior art against third-party patent filings at the earliest possible date.[105] This result,

775, 775 (2005).

102. David R. Bentley, *Genomic Sequence Information Should be Released Immediately and Freely in the Public Domain*, 274 SCIENCE 533, 533 (1996); *see also* Adam Bostanci, *Sequencing Human Genomes*, *in* FROM MOLECULAR GENETICS TO GENOMICS 174 (Jean-Paul Gaudilliére & Hans-Jörg Rheinberger eds., 2004) (arguing that the immediate publication requirement was successful in reducing the risk of duplication posed by researchers' tendency to focus on lucrative genes).

103. *See* Bentley, *supra* note 102, at 533 (insisting that, because sequences derive their value from effective interpretation and use, the public good requires that raw sequences be made available to the greatest number of scientists as quickly as possible); Cook-Deegan & McCormack, *supra* note 96 ("[W]ithout [the Bermuda Principles], the wait for information sufficient to meet patent criteria from high throughput sequencing programs would lead to long delays, and thus be a serious drag on science, undermining the publicly funded sequencing programs' very purpose.").

104. Bentley, *supra* note 102, at 533-34; *see also* Marshall, *supra* note 82; JAMES D. WATSON ET AL., RECOMBINANT DNA 295 (3d ed. 2005).

105. In jurisdictions such as the European Union and Japan that have so-called "absolute novelty" requirements, an invention may not be patented if it has been publicly disclosed prior to the filing of a patent application. *See* JOHN GLADSTONE MILLS III ET AL., PATENT LAW FUNDAMENTALS §2:30 (perm. ed.,

though praised by many, was also criticized by those who believed that the NIH's adoption of this anti-patenting approach contravened the requirements of the Bayh-Dole Act of 1980, which expressly favors the patenting of federally-funded inventions for the benefit of the U.S. economy.[106]

In response to this criticism, the National Human Genome Research Institute's (NHGRI) 1996 policy adopting the Bermuda Principles explicitly acknowledges the Bayh-Dole Act, noting that recipients of NIH funding have the right to choose to apply for patents on inventions that "reveal convincing evidence for utility," but it goes on to warn that "NHGRI will monitor grantee activity in this area to learn whether or not attempts are being made to patent large blocks of primary human genomic DNA sequence."[107] The consequences if such

---

rev. vol. May 2009). In such countries, a description of the invention in a scientific journal could preclude the inventor from obtaining patent protection for his or her invention. In the United States, a patent application may be filed with respect to an invention that has been disclosed in a printed publication, but only if the publication occurred less than one year before the filing of the patent application. 35 U.S.C. § 102(b) (2006). Thus, if an inventor wishes to seek patent protection for his or her invention, he or she must file a patent application prior to the disclosure of the invention in a publication (or, in the United States, no more than one year following publication). *See* Eisenberg, *supra* note 47, at 1025–26 (discussing the creation of "patent-defeating" prior art through the HGP's data release rules).

106. Bayh-Dole Act of 1980, 35 U.S.C. §§ 200-12 (2006) ("It is the policy and objective of the Congress to use the patent system to promote the utilization of inventions arising from federally supported research or development."). The Act rationalized the previously chaotic rules governing federally-sponsored inventions and allowed researchers to obtain patents on inventions arising from government-funded research. Penalties, including forfeiture of rights, could result from an institution's failure to pursue patent protection for a federally-funded invention. 35 U.S.C. § 202(c)(3). Commentators have argued that NIH's adoption of the rapid data release requirements of Bayh-Dole deliberately thwart patent protection on genomic inventions. *See* Arti K. Rai & Rebecca S. Eisenberg, *Bayh-Dole Reform and the Progress of Biomedicine*, 66 LAW & CONTEMP. PROBS. 289, 308 (2003) ("Arguably, NIH has acted outside the scope of its statutory authority . . . at least with respect to patentable inventions."); SHREEVE, *supra* note 9, at 46 ("Strictly speaking, the policy directly contradicted the Bayh-Dole Act.").

107. NATIONAL HUMAN GENOME RESEARCH INSTITUTE, NHGRI POLICY REGARDING INTELLECTUAL PROPERTY OF HUMAN GENOMIC SEQUENCE (April 9, 1996) [hereinafter NHGRI 1996 POLICY], *available at* http://www.genome.gov/10000926. In a 1999 NIH-wide policy applicable to all biomedical research tools, the agency expressly stated that the goals of the Bayh-Dole Act can be met through publication or databank deposit of generally-applicable research tools, and that restrictive licensing of such inventions would be "antithetical" to the goals of the Act. Principles and Guidelines for Recipients of NIH Research Grants and Contracts on Obtaining

patenting activity is discovered are left unstated, but the clear implication is that the agency may view future grant applications by "violators" unfavorably.[108]

The significance of NHGRI's implementation of the Bermuda Principles[109] cannot be overstated. Prior to 1996, NHGRI's position with respect to data release and intellectual property was not very different than that of other federal agencies.[110] But in the negotiations at and leading up to the Bermuda meeting, the scientific community's acknowledgement 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.

2. Data Generators versus Data Users

In their effort to promote the policy goals of *project coordination*, *scientific advancement*, and *minimizing encumbrances*, the HGP organizers sacrificed the interests of data generators. That is, the rapid data release requirements of the Bermuda Principles effectively eliminated the ability of

and Disseminating Biomedical Research Resources: Final Notice, 64 Fed. Reg. 72,090, 72,093 (Dec. 23, 1999).

108. For a general critique of the NIH's "hortatory" approach to this issue, see Rai & Eisenberg, *supra* note 106, at 293-94, 306. Interestingly, the National Cancer Institute (NCI) followed a different approach when addressing concerns over the release of genomic data and the requirements of the Bayh-Dole Act with its Cancer Genome Anatomy Program (CGAP) in 2000. Rather than issuing policy statements along the lines of NHGRI, NCI invoked a seldom-used provision of the Bayh-Dole Act seeking a declaration of "exceptional circumstances" to retain the intellectual property rights in cDNA sequences generated by CGAP's contractors. *See* NCI FREDERICK CANCER RESEARCH AND DEVELOPMENT CENTER, MOLECULAR TARGET LIBRARIES (MTLS), *reprinted in* COM. BUD. DAILY, Feb. 24, 2000. Because NCI did not pursue patent protection for the sequences, they were effectively contributed to the public domain. The procedures relating to such declarations of "exceptional circumstances" are involved and time-consuming and, because they have not been widely utilized, unpredictable. *See* Rai & Eisenberg, *supra* note 106.

109. *See* NHGRI 1996 POLICY, *supra* note 107; NATIONAL HUMAN GENOME RESEARCH INSTITUTE, CURRENT NHGRI POLICY FOR RELEASE AND DATABASE DEPOSITION OF SEQUENCE DATA (Mar. 7, 1997) [hereinafter NHGRI 1997 POLICY], *available at* http://www.genome.gov/page.cfm?pageID=10000910.

110. *See* discussion of NASA and other federal policies *supra* note 8.

data generators to publish analyses and conclusions based on their data before others could access it via public means.[111] The implications of this effect were not realized immediately, but, in the years immediately following the completion of the HGP, a number of large-scale, publicly-funded genomics projects adopted data release policies that reflect an increasing recognition of the inherent tension between data generators and data users. This distinction was first codified in a new NHGRI data release policy adopted shortly after the Third International Strategy Meeting on Human Genome Sequencing held at Cold Spring Harbor in May 2000.[112] The NHGRI 2000 policy reaffirmed the Institute's 1997 Bermuda-based requirement that initial genomic sequence assemblies be deposited into GenBank within twenty-four hours of assembly and extended the earlier policy to later-stage data. For the first time, however, it also imposed formal requirements on *users* who accessed and downloaded the released data. The policy acknowledges "the widely accepted ethic in the scientific community that those who generate the primary data freely should have both the right and responsibility to publish the work in a peer-reviewed journal."[113] Thus, the policy expressly prohibits users from employing the public data "for the initial publication of the complete genome sequence assembly or other large-scale analyses,"[114] thereby reserving this right to the data generators. Moreover, when data users do utilize the publicly-available sequence data, they are required to acknowledge its source.

111. Deanna M. Church & LeDeana W. Hillier, *Back to Bermuda: How is Science Best Served?* 10 GENOME BIOLOGY 105, 105.1 (Apr. 24, 2009) ("[T]here was some concern that [the policy] would jeopardize the genome center's ability to analyze and publish the data they had produced.").

112. *See* NATIONAL HUMAN GENOME RESEARCH INSTITUTE, NHGRI POLICY FOR RELEASE AND DATABASE DEPOSITION OF SEQUENCE DATA (Dec. 21, 2000) [hereinafter NHGRI 2000 POLICY], *available at* www.genome.gov/page.cfm?pageID=10000910.

113. *Id.*

114. *Id.* While this prohibition represents an important gain for data generators, it does not address their more fundamental concern with the publication of *analyses* based on the data they have generated, as opposed to the raw data itself.

C. FT. LAUDERDALE AND COMMUNITY RESOURCE PROJECTS (CRPS)

1. Reaffirmation of Bermuda

Questions regarding the ongoing validity of the Bermuda Principles began to emerge following the completion of the HGP. In order to address these concerns, the Wellcome Trust sponsored a 2003 meeting in Ft. Lauderdale, Florida to revisit rapid data release issues in the "post-genome" world. The meeting was attended by representatives of funding agencies, sequencing centers, database managers, biological laboratories, and scientific journals, many of whom were involved in the original HGP.[115] The Ft. Lauderdale participants "enthusiastically reaffirmed" the 1996 Bermuda Principles.[116] The most significant outcome of the Ft. 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."[117]

Under this definition, the twenty-four hour rapid release rules of Bermuda would be applicable to large-scale projects generating non-human sequence data (e.g., the Mouse Genome Consortium), other basic genomic data maps (e.g., the SNP Consortium and International HapMap Consortium), and other collections of complex biological data, such as protein structures and gene expression information.[118] 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 pre-publication data release be required in such projects.[119]

115. 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), [hereinafter *Ft. Lauderdale Principles*], *available at* http://www.genome.gov/Pages/Research/WellcomeReport0303.pdf.

116. *Id.* at 2 (recognizing, in addition, that considerations might apply to data other than large-scale genomic sequences).

117. *Id.*

118. *Id.* at 2, 5.

119. *Id.* at 3.

2. Different Data Types and Release Considerations

Notwithstanding this show of support, the Ft. Lauderdale participants acknowledged that rapid pre-publication 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.[120] 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 or other product. 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 finalized or published could potentially enable a competing group to "scoop" the originating group, a persistent fear among highly competitive scientists.

3. Adoption by NHGRI

As with the Bermuda Principles, the refinements agreed to in Ft. Lauderdale were widely adopted, both by NHGRI and the major sequencing laboratories. The NHGRI 2003 policy, issued just a few weeks after the Ft. Lauderdale meeting, reiterates the agency's commitment to the Bermuda Principles "for all types of large-scale DNA sequence data sets."[121] In the policy, NHGRI recognizes the need for data generators to achieve publications from the data they have released.[122] Despite this acknowledgement, the agency declines to impose any time

120. *See, e.g.*, Kaye et al., *supra* note 10. An analogy to the distinction between CRP and hypothesis-driven projects in biomedical science may be drawn from geology. In geology, a CRP might be the U.S. Geological Survey's creation of a geophysical map of a region for the use of all interested geologists, while a hypothesis-driven project might seek to determine whether shale oil can be extracted from a particular valley in that region.

121. *Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects*, NAT'L HUMAN GENOME RESEARCH INST. (Feb. 2003), http://www.genome.gov/10506537 [hereinafter NHGRI 2003 POLICY].

122. *Id.* ("[T]he sequence producers have a legitimate interest in publishing peer-reviewed reports describing and analyzing the sequence they have produced.").

limitation or other restriction on users of the released sequence data. Instead, the policy strongly reaffirms NHGRI's position that DNA sequence data "should be available for all to use without restriction" and urges data users to act in accordance with "standard scientific norms" and to acknowledge data generators in published analyses based on their data.[123] These recommendations, though indicative of NHGRI's desired policy, lack binding effect, which NHGRI acknowledges but fails to remedy, stating that "even if the sequence data are occasionally used in ways that violate normal standards of scientific etiquette, unconditional release of sequence data from large-scale sequence production centers is a necessary risk set against the considerable benefits of immediate data release."[124] This statement provides little comfort to data generators who are given no effective recourse if their data are used in a manner that violates these standards.

4. The International HapMap Project, a New CRP

Beginning in 2002, a group of scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States began a project to develop a haplotype map of the human genome.[125] The data release policy of the HapMap Project is based on the Ft. Lauderdale principles, and the project self-designates itself as a CRP.[126] Data generated by the project "[were] released rapidly into" publicly accessible databases,[127] but access was subject to the user's consent to the terms of a standardized, online click-wrap agreement.[128]

123. *Id.*

124. *Id.* (emphasis added).

125. *See generally* The Int'l HapMap Consortium, *The International HapMap Project*, 426 NATURE 789, 790 (2003) (noting that a haplotype map shows genomic "markers" that tend to recur in groups).

126. *Id.* at 793.

127. *Id.* SNP data were deposited in the NIH's dbSNP database (a public database), while genotype and haplotype data were made available through the project's data coordination center.

128. *Id.* A click-wrap agreement (alternatively referred to as a "click-through" or "click-to-accept" agreement or license) is "an electronic form agreement to which [a] party may assent by clicking an icon or a button or by typing in a set of specified words." Christina L. Kunz et al., *Click-Through Agreements: Strategies for Avoiding Disputes on Validity of Assent*, 57 BUS. LAW. 401 (2001–2002). Rebecca Eisenberg, who analogizes the HapMap Agreement to the open source software General Public License (GPL) raises

The HapMap Project took several affirmative steps to ensure that patents would not be filed by data generators, data users claiming haplotypes, or other data generated by the project.[129] Most importantly, each user of HapMap data (including data generators) was expressly prohibited from restricting access to the HapMap database and, in particular, from filing patent applications on the haplotypes or other scientific data generated by the project.[130] The HapMap Consortium's non-patenting requirement was viewed with admiration by many, including policy makers at NHGRI.[131]

As a corollary to the provisions of its click-wrap agreement, the HapMap Project adopted a "Data Release Policy," setting forth the participants' somewhat conclusory position that raw SNP and haplotype data lack "specific utility" necessary for patent protection.[132] The Policy also stated that, because the Project will not relate genetic variants to medically relevant conditions, "results that might be patentable can be obtained only through additional studies not connected with the HapMap Project."[133]

5. The ENCODE Pilot Project

The Encyclopedia of DNA Elements (ENCODE) pilot project was launched by NHGRI in 2003 as an effort to elucidate the biological functions of various genetic elements.[134] NHGRI issued a data release policy for the ENCODE pilot

questions about the enforceability of such agreements. Eisenberg, *supra* note 47, at 1028. For a general discussion of the enforceability of click-wrap agreements, *see generally* GEORGE G. DELTA & JEFFREY H. MATSUURA, LAW OF THE INTERNET § 10.05 (2d ed. 2008).

129. *See* The International HapMap Consortium, *supra* note 125, at 793.

130. *Registration for Access to the HapMap Project Genotype Database*, INT'L HAPMAP PROJECT, http://hapmap.ncbi.nlm.nih.gov/cgi-perl/registration (last visited Jan. 18, 2011) [hereinafter *HapMap Agreement*].

131. *See ENCODE Project Data Release Policy (2003-2007)*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/12513440 (last visited Oct. 18, 2010) [hereinafter *ENCODE 2003 Pilot Policy*]. (referring to the HapMap Project's successful policy of discouraging "parasitic patents").

132. *Data Release Policy*, INT'L HAPMAP PROJECT, http://www.hapmap.org/datareleasepolicy.html (last visited Jan. 18, 2011).

133. *Id.* Though unclear from the HapMap project web site, Rebecca Eisenberg reports that the Data Release Policy was adopted as late as 2004 and was intended to supersede the click-wrap structure. Eisenberg, *supra* note 47, at 1026.

134. The ENCODE Project Consortium, *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*, 447 NATURE 799 (2007).

project closely following the Ft. Lauderdale principles.[135] The NHGRI designated the project as a CRP.[136] As recommended in Ft. Lauderdale, users of the data were urged to cite the data generators in their analyses and to consider research collaborations with them.[137]

With respect to intellectual property issues, the agency first acknowledges the requirements of the Bayh-Dole Act by, on one hand, stating that it has complied with those requirements and, on the other, expressing its view that patent protection for genomic sequence data is inappropriate.[138] With this preface, NHGRI acknowledges that the data created by the ENCODE project will differ in character from the raw sequence data generated by the HGP and HapMap project. That is, the DNA sequence elements identified by ENCODE will, by definition, "have biological function, and therefore might be considered to have utility and be able to be patented."[139] Constrained by Bayh-Dole from expressly requiring researchers to forego the opportunity to patent their federally-funded inventions, NHGRI strongly "encourages all ENCODE data producers to consider placing all information generated from their project-related efforts in the public domain . . . ."[140] In addition, if grantees elect *not* to place their results in the public domain, the agency encourages them to consider "maximal use of non-exclusive licensing of patents to allow for broad access and stimulate the development of multiple products."[141] This language seems to represent NHGRI's perception of the greatest extent of its ability to promote the public domain over patenting while remaining compliant with the letter of the Bayh-Dole Act.

135. *See ENCODE 2003 Pilot Policy*, *supra* note 131.
136. *Id.*
137. *Id.*
138. *Id.*
139. *Id.*
140. *Id.*
141. *Id.*

D. EARLY PRIVATE SECTOR INITIATIVES

In addition to the HGP and other public sector sequencing efforts described above, a number of private sector projects made substantial contributions to the genome commons, many with data release policies informed by the principles established in Bermuda and Ft. Lauderdale.

1. The Merck Gene Index

As early as 1994, pharmaceutical manufacturer Merck, collaborating with Lawrence Livermore National Laboratory and Washington University, compiled the so-called "Merck Gene Index," a publicly accessible database of expressed sequence tags (ESTs).[142] By 1998, the Merck Gene Index had released over 800,000 ESTs through GenBank.[143] Merck's stated reason for contributing this potentially valuable data to the public was the expansion of basic knowledge in the interest of combating disease.[144] While this goal is laudable, it was generally acknowledged that another motivation for placing these ESTs into the public was the pre-emption of patent filings by biotech companies, several of which had already announced business plans that involved the patenting and licensing of ESTs and other genetic information.[145]

2. The SNP Consortium

An interesting and oft-cited parallel to the post-HGP government-funded projects discussed above is that of the SNP Consortium. This non-profit entity was formed in 1999 by a

142. *See* 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), [hereinafter Merck Gene Index Press Release], *available at* http://www.bio.net/bionet/mm/bionews/1995-February/001794.html; *see also supra* notes 92–93and accompanying text (discussing ESTs and the patenting debate surrounding them).

143. DON TAPSCOTT & ANTHONY D. WILLIAMS, WIKINOMICS: HOW MASS COLLABORATION CHANGES EVERYTHING 166 (2006).

144. Merck Gene Index Press Release, *supra* note 142.

145. Marshall, *supra* note 82. Companies such as Incyte Pharmaceuticals in Palo Alto, California, and Human Genome Sciences in Rockville, Maryland, were then actively pursuing a business strategy of patenting, and licensing, ESTs and other genetic data. *Id.*; *See* TAPSCOTT & WILLIAMS, *supra* note 143; Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 134 (1999–2000).

group of ten pharmaceutical companies[146] and the Wellcome Trust to identify and map genetic markers referred to as "single nucleotide polymorphisms" (SNPs) and to release the resulting data to the public domain.[147] SNP data were publicly released on the Consortium's web site on a quarterly, and later monthly, basis during the two-year research program, and also deposited in GenBank.[148] The Consortium ultimately mapped 1.4 million SNPs and created a genome-wide SNP-based human linkage map, all of which were made publicly available along with a number of query and search tools.[149] Like the Merck Gene Index, the SNP Consortium aimed to generate data for the use of all researchers, unencumbered by patents.[150] It accomplished this goal by filing U.S. patent applications covering SNPs that it discovered and then contributing these

146. The SNP Consortium Ltd. was incorporated in March 1999 with the following sponsoring (i.e., dues-paying) members: The Wellcome Trust Limited, Pfizer Inc, Glaxo Wellcome Inc., Hoechst Marion Roussel, Zeneca Inc., Hoffman-La Roche Inc., Novartis Pharmaceuticals Corporation, Bristol-Myers Squibb Company, SmithKline Beecham Corporation, Bayer Corporation and Monsanto Corporation. Technology giants Motorola, Inc. and International Business Machines Corporation joined as sponsoring members in November 1999 and Amersham Pharmacia Biotech Inc. became a sponsoring member in 2001. Jorge Contreras, Personal Files (on file with author).

147. SNPs are instances in which single base pairs in the genome differ among individuals and occur roughly once per thousand base pairs. Though the presence of certain SNPs has been associated with diseases, the purpose of generating so-called SNP maps is to establish a uniform set of "mile markers" along the vast genome. *See* Arthur Holden, *The SNP Consortium: Summary of a Private Consortium Effort to Develop an Applied Map of the Human Genome*, 32 BIOTECHNIQUES 22 (2002).

148. *See* Holden, *supra* note 147, at 25–26 and *SNP Fact Sheet*, U.S. DEPARTMENT OF ENERGY GENOME PROGRAM, http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml#when (last visited Jan. 18, 2011). The SNP Consortium's data is currently hosted on the International HapMap Project's web site.

149. Holden, *supra* note 147, at 25–26. *See also* Gudmundur A. Thorisson & Lincoln D. Stein, *The SNP Consortium website: past, present and future*, 31 NUCLEIC ACIDS RES. 124, 124–27 (2003) (providing a detailed description of how the public can utilize the consortium's website).

150. *See, e.g.*, Holden, *supra* note 147, at 26 ("[t]he overall IP objective is to maximize the number of SNPs that (*i*) enter the public domain at the earliest possible date, and, (*ii*) are free of third-party encumbrances such that the map can be used by all without financial or other IP obligations."); TAPSCOTT & WILLIAMS, *supra* note 143, at 168 (noting consortium members' concerns about biotech companies' plans to patent SNPs and "sell them to the highest bidder.").

applications to the public domain prior to issuance.[151] This approach ensured that the Consortium's discoveries would act as prior art defeating subsequent third-party patent applications, with a priority date extending back to the initial filings. The SNP Consortium's innovative "protective" patenting strategy has been cited as a model of the private industry's potential to contribute to the public genome commons.[152]

E. SECOND GENERATION GENOMIC DATA RELEASE POLICIES

In the years following the Ft. Lauderdale meeting, numerous large-scale genomic research projects have been 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.

1. Genetic Association Information Network (GAIN)

The Genetic Association Information Network (GAIN) was established in 2006 by the Foundation for the National Institutes of Health (FNIH), the NIH, and several corporations.[153] GAIN's purpose was to conduct GWA studies of

151. The SNP Consortium's patenting strategy included the filing of patent applications covering all mapped SNPs and then converting those applications into statutory invention registrations (SIRs) or abandoning the applications after publication. *See* Identification and Mapping of Single Nucleotide Polymorphisms in the Human Genome, U.S. Statutory Invention Registration, No. H2220 (filed Aug. 8, 2001); Identification and Mapping of Single Nucleotide Polymorphisms in the Human Genome, U.S. Statutory Invention Registration, No. H2220 (filed Nov. 21, 2002).

152. *See, e.g.,* Marshall, *supra* note 82, at 1192 (noting the consortium's "defensive move" deriving from the Merck Gene Index's earlier strategy); Cook-Deegan & McCormack, *supra* note 96 (describing the consortium's "unusual and sophisticated approach to keeping data in the public domain."); Allen C. Nunnally, *Intellectual Property Perspectives in Pharmacogenomics*, 46 JURIMETRICS 249, 252–53 (2006) (noting that the consortium members' placement of the raw SNP map into the public domain did not necessarily preclude their, or anybody else's, patenting of subsequent discoveries made using the basic research funded by the consortium).

153. *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

the genetic basis for six common diseases.[154] Data generators in the GAIN program were required to sign an applicant agreement agreeing to various program commitments, including "immediate" release of data generated by the project.[155] Over the course of the three-year project, approximately 18,000 human DNA samples were genotyped.[156] The resulting data was deposited in the Database of Genotypes and Phenotypes (dbGaP) within the National Library of Medicine at NIH.[157]

The dbGaP allows access to data on two levels: open and controlled.[158] Open access is available to the general public via the Internet and includes non-sensitive summary data, generally in aggregated form.[159] Researchers wishing to access data from the controlled portion of the database must register with and be approved by the GAIN Data Access Committee (DAC).[160] They must also agree to keep the data secure, use it only for approved research purposes, refrain from patenting the data or conclusions drawn directly from the data, acknowledge data generators, and refrain from attempting to identify

selection and characteristics of initial GAIN studies, the structure of GAIN, and defining who has access to GAIN data).

154. The diseases studied were Attention Deficit Hyperactivity Disorder (ADHD), diabetic nephropathy in Type 1 diabetes, major depression, psoriasis, schizophrenia and bipolar disorder. *Genetic Association Information Network (GAIN)*, FOUND. FOR THE NAT'L INST. OF HEALTH, http://www.fnih.org/work/past-programs/genetic-association-information-network-gain (last visited Oct. 28, 2010) [hereinafter *FNIH Gain Information Sheet*].

155. The GAIN Collaborative Research Group, *supra* note 153, at 1048 (Box 1).

156. Teri A. Manolio, *Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics*, 10 PHARMACOGENOMICS 235, 236 (2009).

157. The combination of phenotypic data with genomic data is critical to understanding disease and physiological traits having genetic influences. *See generally DbGaP Overview*, DBGAP-GENOTYPES & PHENOTYPES, http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html (last visited Oct. 28, 2010). However, phenotypic data, which includes elements such as de-identified subject age, ethnicity, weight, demographics, exposure, disease state and behavioral factors, are far more complex to record, search and correlate than raw sequence data deposited in GenBank. *Id.* In addition to genotypic and phenotypic data, dbGaP can accommodate study documentation and statistical results, including linkage and association analyses. *Id.*

158. *Id.*

159. *Id.*

160. The Gain Collaborative Research Group, *supra* note 153, at 1049.

individual study participants.[161]

Perhaps most importantly, the GAIN policy is the first genomic data release policy to introduce a temporal restriction on the *users* of the data (as opposed to the temporal release requirements imposed on data *generators* by the Bermuda Principles). That is, in order to secure a period of exclusive use and publication priority for the data generators, data users are prohibited from submitting abstracts and publications and from making presentations based on GAIN data for a specified embargo period.[162] The duration of the embargo period for a given data set is identified in the relevant data repository and may vary by data set, but has generally been set at nine months.[163]

2. The Cancer Genome Atlas (TCGA)

In 2006, the National Cancer Institute (NCI) and NHGRI launched a pilot project to catalog genomic changes relating to cancer.[164] The Cancer Genome Atlas (TCGA) project generates genomic sequence and related data, but also keeps track of large amounts of clinical data, including patient diagnosis, treatment history, and ongoing status.[165] Due to the specialized nature of the project data, deposits are made in both dbGaP and a TCGA-specific database administered by NCI.[166]

Given the potential for identifying individual patients from their genomic and phenotypic data, great attention was paid to controlling access to TCGA data.[167] Like GAIN data, TCGA

161. *Data Use Certification Agreement*, GENETIC ASS'N INFO. NETWORK (GAIN) (Dec. 3, 2008) https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000021.v1.p1 [hereinafter *GAIN Data Use Agreement*].

162. The GAIN Collaborative Research Group, *supra* note 153, at 1049.

163. *Id.*

164. *See generally* Francis S. Collins & Anna D. Barker, *Mapping the Cancer Genome*, SCI. AM., Mar. 2007, at 50. The pilot project is scheduled to conclude in October 2009. *Id.*

165. *Types of Data*, THE CANCER GENOME ATLAS DATA PORTAL, http://cancergenome.nih.gov/dataportal/data/about/types/clinical/ (last visited Oct. 28, 2010).

166. *Data Use Certification Agreement*, 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.

167. A multi-constituency workshop was convened in May 2006 to discuss proposed TCGA data access policies and practices. *See generally Policies and Guidelines*, THE CANCER GENOME ATLAS,

data is available in an open-access tier and a controlled-access tier.[168] Open-access is provided for data that cannot be aggregated to generate an individually-identifiable dataset, whereas controlled-access enables researchers to access clinical and individually-unique data.[169] 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 that are nearly identical to those in the GAIN agreement.[170] One significant difference from the GAIN agreement, however, is the absence in the TCGA certification of a protected period for data generators. Thus, while data users are requested to acknowledge the TCGA in publications based on TCGA data,[171] there is no embargo restriction on the right of data users to submit abstracts or publications derived from TCGA data.

3. The NIH GWAS Policy

In response to the growing number of 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.[172] The NIH GWAS Policy requires that grantees submit descriptive information about each GWA study for inclusion in the "open

http://cancergenome.nih.gov/about/policies/informed_consent.asp (last visited Oct. 28, 2010) (detailing the many considerations taken into account in creating the policies for data access).

168. *Data Access*, THE CANCER GENOME ATLAS DATA PORTAL, http://cancergenome.nih.gov/dataportal/data/access/ (last visited Oct. 28, 2010).

169. *Id.*

170. *Compare Data Use Certification Agreement, supra* note 166, *with GAIN Data Use Agreement*, *supra* note 153.

171. *TCGA Data Use Certification*, *supra* note 166, at 7.

172. Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS), 72 Fed. Reg. 49290, 49294–97 (Aug. 28, 2007) [hereinafter NIH GWAS Policy]. Though the HGP and other early genomic studies were conducted under the auspices of NHGRI, by 2006 most of the NIH Institutes were funding genomic research and GWA studies of their own in support of their individual research missions. *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*].

access" portion of dbGaP.[173] Grantees 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."[174]

Among the principal concerns raised concerning GWAS data were those surrounding the public release of phenotypic or clinical information that could eventually be traced back to individual subjects.[175] To address this concern, the NIH GWAS Policy requires that GWAS data be de-identified in accordance with HIPAA guidelines.[176] 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[177] and then only under a signed Data Use Certification that contains stringent protective clauses.[178] Finally, the NIH sets forth its position that a request under the

173. Descriptive information includes the study protocol, questionnaires, manuals, variables measured and other supporting documentation. NIH GWAS Policy, *supra* note 172, at 49, 295. 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. *Modifications to GWAS Data Access*, *supra* note 172. Due to concerns relating to potential identification of GWAS subjects, NIH withdrew certain GWAS-generated SNP data from the publicly-accessible portions of dbGaP and certain NCI databases and placed them in the controlled-access portions of these databases. *Id.*

174. NIH GWAS Policy, *supra* note 172, at 49295. As in the GAIN Policy, access to the controlled-access portion of the database is regulated by a Data Access Committee and carries stringent protective measures on the use of data. *Id.* at 49296.

175. *Id.* at 49292 (summarizing public concerns over the availability of personally-identifiable data). The NIH acknowledges that technologies either in existence or likely to be available soon would make the identification of individuals from raw genotypic and phenotypic data "feasible and increasingly straightforward." *Id.*

176. *Id.* at 49295 (citing the HIPAA Privacy Rule, 45 CFR 164.514(b)(2)).

177. The DAC is comprised primarily of NIH staff with expertise in the relevant scientific disciplines, data privacy and data subject protection. *Id.* at 49296.

178. Like the certification required under the GAIN program, *see supra* Section III.E.1, the GWAS 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. NIH GWAS Policy, *supra* note 172, at 49296.

Federal Freedom of Information Act (FOIA)[179] for the release of individually-identifiable GWAS information would constitute an "invasion of personal privacy" under FOIA and will be denied by NIH.[180]

The 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 twelve months from the date that the data set is made available.[181] The agency 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."[182] 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."[183] 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."[184]

4. International SAE Consortium

Since the successful completion of the SNP Consortium project, several other privately-funded research collaborations have adopted data release models that are similarly intended to place large quantities of genomic data into the public domain. One of these is the International Serious Adverse

179. *See generally* Federal Freedom of Information Act, 5 U.S.C. § 552 (2006).

180. NIH GWAS Policy, *supra* note 172, at 49292 (citing FOIA Exemption 6, 5 U.S.C. §552(b)(6)).

181. This exclusivity period was originally nine months when the GWAS Policy was released for public comment, but was subsequently lengthened to twelve months. *Request for Information (RFI): Proposed Policy for Sharing of Data obtained in NIH supported or conducted Genome-Wide Association Studies (GWAS)*, NAT'L INST. OF HEALTH (Aug. 30, 2006) http://grants.nih.gov/grants/guide/notice-files/NOT-OD-06-094.html.

182. NIH GWAS Policy, *supra* note 172, at 49296.

183. *Id*. at 49297

184. *Id*. at 49296.

Event Consortium (iSAEC), a group of pharmaceutical companies formed in 2007 to fund research toward the identification of DNA markers for drug-induced serious adverse events.[185] The Consortium works with academic collaborators to collect DNA samples and associated phenotypic data and to then conduct GWA studies, targeted sequencing, and statistical analyses to identify potential markers and associations of interest.[186] Since its formation, iSAEC studies have identified DNA markers relating to drug-induced liver injury (DILI)[187] and serious skin rash (SSR). The iSAEC seeks to minimize patent encumbrances on genetic markers and associations that it identifies via a "protective" patent strategy modeled on that of the SNP Consortium. To date, patent applications claiming various DNA markers relevant to DILI and SSR have been filed with the intention that they will be abandoned following publication.[188] Like the GAIN and other policies discussed in this section, the iSAEC imposes various security, research purpose, and non-patenting restrictions on data that is publicly released. It also secures for data-generating scientists a period of exclusivity (up to twelve months) during which they have sole access to the data.[189] During this time, they have the ability to analyze data and prepare papers for publication without the threat of being scooped by competing groups. While the research funded by iSAEC would not typically be considered a "community resource project" as defined in Ft. Lauderdale (as its goal is not the creation of a large, generally-applicable data set),[190] the Consortium has still committed to release its data to the public, albeit on a delayed basis. This approach illustrates an effective compromise among the interests of data generators in a hypothesis-driven research

185. *iSAEC's Background and Organizational Structure*, INT'L SAE CONSORTIUM, http://www.saeconsortium.org/ (last accessed Oct. 28, 2010).

186. *Id.*

187. *See generally* Ann K. Daly et al., *HLA-B*5701 Genotype is a Major Determinant of Drug-Induced Liver Injury due to Flucloxacillin*, 41 NATURE GENETICS 816 (July 2009) (discussing the genetic basis for susceptibility to drug-induced liver injury from flucloxacillin).

188. Biomarkers for Drug-Induced Liver Injury, U.S. Patent App. 12/505,058 (filed Jul. 17, 2009); Biomarkers for Serious Skin Rash, U.S. Patent App. 61/112,983 (filed Nov.10, 2009); Biomarkers for Serious Skin Rash, U.S. Patent App. 61/168,875 (filed Nov. 10, 2009).

189. Int'l SAE Consortium Ltd., DATA RELEASE AND INTELLECTUAL PROPERTY POLICY (last amended Nov. 5, 2009) (on file with author).

190. *See supra* Section III.C.1.

model and the community of data users and funders.[191]

5. The Full ENCODE Project and modENCODE

In 2007 NHGRI expanded the ENCODE pilot project[192] to cover the entire human genome and launched a corollary project (modENCODE) to identify the functional genomic elements of two common model organisms, the roundworm (*C. elegans*) and fruit fly (*Drosophila melanogaster*).[193] This expansion involved an overhaul of the 2003 ENCODE data release policy and resulted in a new policy in 2008 covering both the expanded ENCODE project and modENCODE.[194] The ENCODE 2008 Policy has much in common with its 2003 predecessor, though it also introduces some of the policy features added by the later GAIN and GWAS policies. Thus, while the ENCODE 2008 Policy continues to use the Ft. Lauderdale terminology in designating itself a "community resource project," it also recommends a nine month embargo period during which users of released data are requested not to publish or present results based on that data.[195]

The ENCODE 2008 Policy is among the most complex data release policies to date, as it distinguishes between published and unpublished data, verified and unverified data, and offers several examples of the data use implications for different types of studies conducted with ENCODE data.[196] The length and complexity of the policy evidences the desire of the agency and the participants for clear guidelines and 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.

191. The compromises and negotiation strategy inherent in this approach is discussed in greater detail in Contreras, *Data Sharing*, *supra* note 5, at 11.

192. *See supra* Section III.C.5.

193. *See* Susan E. Celniker et al., *Unlocking the Secrets of the Genome*, 459 NATURE 927 (2009) (describing the modENCODE project methodology and goals).

194. *ENCODE Consortia Data Release, Data Use, and Publication Policies (2008)*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf [hereinafter *ENCODE 2008 Policy*].

195. *Id.* at 4.

196. *Id.* at 5–7.

F. 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 timeframes for both release of data by data generators and embargo periods on publication for data users.[197] A few examples of recent non-U.S. data release policies are described below.

1. Genome Canada

Genome Canada, a participant in the HGP, adopted its first formal data release policy in 2005.[198] While acknowledging the Ft. 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.[199] For patents, Genome Canada "recognizes the need to protect patentable and other proprietary data" and, thus, requires that the data generators' obligation to release data occur upon the publication or the filing of a patent application, whichever is earlier.[200]

2. 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

197. The reason for this divergence is not clear, though it is possible that the absence of an equivalent to the Bayh-Dole Act in most countries, as well as a patenting landscape that is generally more restrictive outside the U.S., has made patent and data release issues less central to policy discussions outside the U.S.

198. Genome Canada, DATA RELEASE AND RESOURCE SHARING (Sept. 18, 2008), *available at* http://www.genomecanada.ca/medias/PDF/EN/DataReleaseandResourceSharingPolicy.pdf.

199. *Id.*

200. *Id.* (also allowing for extensions of up to 90 days in the event of "extenuating circumstances").

Wellcome Trust has supported genomics initiatives both through direct funding and through its Sanger Institute in Cambridge, England, a leading sequencing center.[201] In 2006, the Trust funded a large-scale GWA study of seven complex human diseases that was conducted by more than fifty research groups from institutions across the United Kingdom.[202] 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 Ft. Lauderdale Principles, and the project self-designated itself as a CRP.[203]

In order to ensure appropriate use of released data, the WTCCC requires all prospective data users to apply to the Consortium's Data Access Committee and sign a written Data Access Agreement.[204] Access to data is granted only to qualified investigators for "appropriate use," as determined by the committee.[205] The data access agreement requires security, acknowledgement, transfer, and use restrictions comparable to those found in the GAIN and other recent policies.[206] 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.[207] The agreement does not, however, contain any specific embargo on publication or any restriction on patenting activity.

---

201. *See generally* WELLCOME TRUST, http://www.wellcome.ac.uk/index.htm (last visited Oct. 26, 2010).

202. The main study covered 2,000 cases and 3,000 shared controls drawn primarily from British subjects for the following seven conditions: bipolar disorder, coronary artery disease, Chron's disease, hypertension, rheumatoid arthritis, type 1 diabetes and type 2 diabetes. The Wellcome Trust Case Control Consortium, *supra* note 41.

203. *Publications Policy,* WELLCOME TRUST CASE CONTROL CONSORTIUM, https://www.wtccc.org.uk/ccc1/publications_policy_ext.shtml (last visited Oct. 26, 2010).

204. *See 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).

205. *WTCCC: Access to Genotype Data*, *supra* note 204, at §4.

206. *Data Access Agreement*, *supra* note 204.

207. *Id.* at §2.

3. UK Medical Research Council

In 2008 the United Kingdom's Medical Research Council (MRC) released a comprehensive set of guidelines surrounding release of data from MRC-funded research.[208] In a set of broad "data access principles," the MRC announces that data generated by publicly-funded research are a public good and, as such, "must be made available for new research purposes in a timely, responsible manner."[209] Following the reasoning behind the Ft. Lauderdale principles, the MRC states that access to data "must balance the interests of data creators, custodians, users and data subjects,"[210] and acknowledges that "limited, defined" periods of exclusive use "will often be justifiable."[211] 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.[212] The MRC, however, by and large allows individual parties to define the specific requirements of their data sharing agreements and does not attempt to impose over-arching rules regarding the timing of data release.

208. *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 (last visited Oct. 26, 2010) (Because the MRC broadly addresses biomedical research across many fields, the MRC guidelines do not focus on genomic data, nor do they expressly reference the Bermuda or Ft. Lauderdale principles, despite MRC's participation in the original Bermuda meeting).

209. *Id.* at 1.

210. *Id.*

211. *Id.* at 4.

212. *Id.* at 5 para. 8.

G. RECENT DEVELOPMENTS IN RAPID PRE-PUBLICATION DATA RELEASE

1. Amsterdam: Proteomics Joins the Fray

The success and broad adoption of genomics data release policies incorporating the Bermuda and Ft. Lauderdale Principles have recently 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[213] researchers in Amsterdam to "identify and address potential roadblocks to rapid and open access to [proteomics] data."[214]

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.[215] 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.[216] 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

213. Proteomics is the study of protein structures. Unlike DNA sequences, which are linear arrangements of the four basic nucleotides, A, C, T and G, proteins consist of intricately-folded, three-dimensional structures formed from twenty different amino acids. Unlike today's relatively straightforward and automated DNA sequencing technologies, the techniques for elucidating protein structures include electrophoresis, various forms of mass spectrometry and an increasing number of other methods. *See generally* LESK, *supra* note 1, at 312–22.

214. 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).

215. *See id.* at 3689–90.

216. *Id.* at 3690. Existing proteomic databases include GPMDB, UniProtKB, Peptide Atlas, PRIDE and NCBI's Peptidome. *Id.*

data submission.[217]

Notwithstanding these difficulties, the Amsterdam participants articulated six data release and sharing principles that reflect the spirit of the Bermuda and Ft. Lauderdale Principles, but which lack the specificity of the genomics policies. The six Amsterdam principles are: (1) Timing (should depend on the nature of the effort generating the data, but in no event should 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).[218]

2. The Toronto Data Release Workshop

In 2009, more than a hundred scientists, journal editors, legal scholars, and representatives of governmental and private funding agencies met in Toronto to assess the current state of rapid pre-publication data release and the applicability of the Bermuda Principles in projects well beyond the generation of genomic sequence data.[219] The participants reaffirmed a general community commitment to rapid pre-publication data release, expanding the scope of projects as to which of these principles should apply to all biomedical datasets having "broad utility, are large in scale . . . and are 'reference' in character."[220] Specifically, in addition to genomic and proteomic studies, they cited structural chemistry, metabolomics, and RNAi datasets, as well as annotated clinical resources such as cohorts, tissue banks, and case-control studies.[221]

The expansion of rapid pre-publication data release principles beyond genomics and proteomics projects, which often have as their ultimate goal the generation of a large data

217. *Id.*

218. *Id.* at 3690–91.

219. *See Toronto Report*, *supra* note 66.

220. *Id.* at 168. To some degree, this characterization is a restatement of the Ft. Lauderdale definition of "community resource projects".

221. *Id.*

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 pre-publication data release for "broad utility" projects, rapid data release "should not be mandated" for projects that are generally hypothesis-driven.[222] The Toronto participants also addressed the priority concerns of data generators versus data users, observing anecdotally that data users have in many cases 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.[223] 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.[224] The Toronto participants produced a set of "best practices" embodying these principles and applying them to the three constituencies originally identified in Ft. 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.[225]

Discussions in Toronto also addressed issues of intellectual property. In particular, it was observed that, as data sets subject to rapid pre-publication release expand beyond genomic and proteomic "basic science" and begin to embody greater functional content and clinical utility, the patentability of this information will be less open to debate, and the early release of such information will have a greater impact on the data generators' ability to secure patent protection with concomitant implications for U.S. funding agencies subject to Bayh-Dole requirements.[226] Given the controversial nature of this subject and the lack of consensus on this issue, the subject of intellectual property was ultimately excluded from the published meeting report. It is inevitable, however, that

222. *Id.* at 169.
223. *Id.* at 169–70.
224. *Id.* at 170.
225. *Id.*
226. Author's personal notes, The Toronto Data Release Workshop (May 13-14, 2009) (on file with author).

intellectual property issues will play an increasingly important role in discussions of rapid pre-publication data release in fields of medical significance.

3. New Policies and Projects

The influence of the Bermuda/Ft. Lauderdale Principles has been lasting and pervasive. The list of new biomedical research projects that are currently developing or have recently adopted data release policies based on these principles or their progeny is too long to list here, but includes projects such as the 1000 Genomes Project,[227] the International Cancer Genome Consortium,[228] and the Human Microbiome Project.[229] NIH and NHGRI are in the process of considering further revisions to their institutional data release policies and collecting feedback from various stakeholder groups.[230] Though the result of this latest round of revisions have not yet been released, it is likely that any new NIH data release policy will continue to refine the rules of rapid pre-publication data release to take into account the policy considerations and objectives described above.

IV. POLICY CONSIDERATIONS IN GENOMIC DATA RELEASE POLICIES

A. ELEMENTS OF POLICY DESIGN

The preceding Section describes the major genomics-related data release milestones and policies from the beginning of the HGP through today, a span of nearly two decades. Table 1 below summarizes the manner in which each of these policies handles issues relating to the speed of data release, restrictions on data use, and intellectual property.

227. *See 1000 Genomes Data and Sample Information*, 1000 GENOMES, http://www.1000genomes.org/page.php?page=data (last visited Oct. 28, 2010).

228. *See* INTERNATIONAL CANCER GENOME CONSORTIUM, GOALS, STRUCTURES, POLICIES & GUIDELINES 15 (2008), *available at* http://www.icgc.org/files/icgc/ICGC_April_29_2008_en.pdf.

229. *See HMP Data Release and Resource Sharing Guidelines for Human Microbiome Project Data Production Grants*, NIH COMMON FUND, http://commonfund.nih.gov/hmp/datareleaseguidelines.asp (last visited Oct. 28, 2010).

230. National Institutes of Health, Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data (Oct. 19, 2009), *available at* http://grants1.nih.gov/grants/guide/notice-files/NOT-HG-10-006.html.

Table 1: Comparative Summary of Genomics Data Release Policies

	Type of Data	Release Speed	User Restrictions	Patent Considerations
NIH-DOE 1992 Guidelines	Materials and data produced by the HGP[231]	Within 6 months after generation[232]	None[233]	IP protection "may be needed" for some data and materials[234]
Bermuda Principles (1996)	Initial genome sequence reads > 1Kb[235]	24 hours after generation[236]	None[237]	Not specified[238]
NHGRI 1996 Policy	Human genomic DNA sequence data produced under the HGP[239]	As rapidly as possible[240]	None[241]	"raw human genomic DNA sequence . . . is an inappropriate material for patent filing"[242]

231. *NIH/DOE Guidelines*, *supra* note 84.
232. *Id.*
233. *Id.*
234. *Id.*
235. *Bermuda Principles*, *supra* note 6. Initial genome sequence reads were increased to 2 Kb under the Bermuda 1997 Report. *Id.*.
236. Contreras, *Prepublication Data Release, supra* note 5, at 393.
237. *Id.*
238. *See id.*
239. NHGRI 1996 POLICY, *supra* note 107.
240. *Id.*
241. *Id.*
242. *Id.*

	Type of Data	**Release Speed**	**User Restrictions**	**Patent Considerations**
NHGRI 1997 Policy	Large-scale genomic DNA sequence data[243]	24 hours after generation[244]	None[245]	Not specified[246]
SNP Consortium (1998)[247]	SNP map and association data[248]	Monthly/quarterly releases to web site[249]	None[250]	protective filing strategy to release data to public domain[251]
NHGRI 2000 Policy	Sequence trace data and ancillary information[252]	Deposited weekly into the NCBI Trace Repository[253]	Users may not use data for the initial publication of the complete genome sequence assembly or other large-scale analyses[254]	Not specified[255]

243. NHGRI 1997 POLICY, *supra* note 107.
244. *Id.*
245. *Id.*
246. *See id.*
247. Holden, *supra* note 147 at 22–26.
248. *Id.*
249. *Id.* at 26.
250. *See id.* at 22–23.
251. *Id.* at 26.
252. NHGRI 2000 Policy, *supra* note 112.
253. *Id.*
254. *Id.*
255. *See id.*

	Type of Data	**Release Speed**	**User Restrictions**	**Patent Considerations**
Ft. Lauderdale Principles (2003)	All data from Community Resource Projects[256]	Reaffirms Bermuda Principles for sequence assemblies > 2kB[257]	Citation of data producers[258]	Not specified[259]
NHGRI 2003 Policy	Large-scale sequence data[260]	24 hours after generation[261]	Citation of data producers[262]	Not specified[263]
Intl. HapMap Project (2003)	SNP and Haplotype data[264]	Rapidly[265]	Citation of data producers[266]	User click-wrap agreement prohibits filing of patent applications on project data or uses thereof, unless unrestricted use is allowed[267]

256. *Ft. Lauderdale Principles*, *supra* note 115.
257. *Id.*
258. *Id.*
259. *See id.*
260. NHGRI 2003 POLICY, *supra* note 121.
261. *Id.*
262. *Id.*
263. *See id.*
264. The Int'l HapMap Consortium, *supra* note 125.
265. *Id.*
266. *Id.*
267. *Id.*; *HapMap Agreement*, *supra* note 130.

	Type of Data	Release Speed	User Restrictions	Patent Considerations
ENCODE Pilot (2003)	Various data types generated by the project[268]	Deposited as soon as data is verified[269]	Citation of data producers [270]	NHGRI encourages all data producers to consider placing information in the public domain[271]
Wellcome Trust Case Control Consortium (WTCCC) (2006)	GWA study data[272]	Not specified [273]	Signed data access agreement Citation of data producers; Security, transfer and use restrictions [274]	None[275]

268. *ENCODE 2003 Pilot Policy*, *supra* note 131.
269. *Id.*
270. *Id.*
271. *Id.*
272. *Publications Policy*, *supra* note 203.
273. *See id.*
274. *WTCCC: Access to Genotype Data*, *supra* note 204.
275. *See id.*; *Publications Policy*, *supra* note 203.

	Type of Data	**Release Speed**	**User Restrictions**	**Patent Considerations**
GAIN (2006)	GWA study data[276]	Immediate[277]	Data use certification[278] Variable-length embargo on publication and presentation of data (generally 9 months),[279] Security, transfer and use restrictions[280]	Data users *agree* not to pursue patents that would block access to data or conclusions drawn directly from data[281]
The Cancer Genome Atlas (TCGA) (2006)	Tumor genomic and clinical data[282]	As rapidly as possible[283]	data use certification[284] Security and transfer[285] restrictions	Urges users to avoid making IP claims[286]

276. The GAIN Collaborative Research Group, *supra* note 153, at 1045.
277. *Id.* at 1048.
278. *Id.* at 1049.
279. *Id.*
280. *Id.* at 1046.
281. *Id.* at 1050.
282. *Types of Data*, *supra* note 165.
283. *Data Use Certification*, *supra* note 166, at 4.
284. *Id.*
285. *Id.* at 2–3.
286. *Id.*

	Type of Data	**Release Speed**	**User Restrictions**	**Patent Considerations**
NIH GWAS Policy (2007)	GWA study data[287]	Strong encouragement to submit data as soon as quality control procedures completed[288]	Signed data use certification[289] 12-month embargo expectation on publication and presentation of data,[290] Citation of data producers;[291] Security, transfer and use restrictions[292]	Patenting of results discouraged[293]

287. 72 Fed. Reg. 49290, 49290 (Aug. 28, 2007).
288. *Id.* at 49293.
289. *Id.* at 49297.
290. *Id.* at 49294.
291. *Id.*
292. 72 Fed. Reg. 49290, 49296 (Aug. 28, 2007).
293. *Id.*

	Type of Data	**Release Speed**	**User Restrictions**	**Patent Considerations**
Intl. SAE Consortium (2007) [294]	Human genotypic and phenotypic data	12 months after data validation	Signed member-ship agreement Up to 9-month embargo on publication and presentation of data Citation of data producers; Security, transfer and use restrictions	Data users agree not to pursue patents that would block access to data or conclusions drawn directly from data
ENCODE + modENCODE (2008)	Various data types generated by the project[295]	Deposited as soon as data is verified[296]	9-month embargo on publication and presentation of data[297] Citation of data producers [298]	NHGRI monitoring of patenting activity and potential consid-eration of click-wrap agreements (e.g., HapMap Project[299])

294. *See supra* Part III.E.4.
295. The ENCODE Project Consortium, *supra* note 134, at 799.
296. *ENCODE 2003 Pilot Policy*, *supra* note 131.
297. *ENCODE 2008 Policy*, *supra* note 194, at 1.
298. *Id.* at 4.
299. *Compare id.* at 6–9, *with HapMap Agreement, supra* note 130.

B. POLICY DESIGN TRENDS

Even a cursory inspection of *Table 1* reveals several points regarding the evolution of genomics data release policies over the past two decades. Perhaps most obviously, these policies have grown more detailed and complex over time. The reasons for this growth are not difficult to guess. The Bermuda Principles introduced a sea change to scientific data release. Despite their groundbreaking significance and lasting influence, the Bermuda Principles were drafted to address one specific type of data (genomic sequence reads) generated by a specific, unique project (the HGP).[300] It soon became clear that, while the spirit and intent of the Bermuda Principles were attractive to many, the extension of these principles to different projects and data types required additional explication and, in some cases, compromise. Below is a summary of the ways in which policy designers addressed the various policy considerations associated with the genome commons over this period.

1. Protection of Human Subject Data

Because the goal of the HGP was to generate a baseline map of the human genome without regard to the particular physiological and pathological traits associated with genetic variation among individuals, the genomic sequence data generated by the HGP was anonymous and retained no association with the individual subjects whose DNA was being sequenced.[301] Similar characteristics applied to data generated by the HapMap Project[302] and the SNP Consortium.[303] These data were intended to elucidate non-individualized information applicable to the human genome generally. Accordingly, concerns regarding the identifiability of human subjects from data released to the public, while addressed, were not

300. *Policies on Release of Human Genomic Sequence Data, Bermuda Quality Sequence,* HUMAN GENOME PROJECT INFORMATION, http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml (last visited Oct. 28, 2010).

301. *About the Human Genome Project*, HUMAN GENOME PROJECT INFORMATION, http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml (last visited Oct. 28, 2010).

302. *See What is the HapMap*, INTERNATIONAL HAPMAP PROJECT, http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en (last visited Oct. 28, 2010).

303. Holden, *supra* note 147.

paramount in these early projects.

In later projects, and particularly with the commencement of large-scale GWA studies, concerns with the potential identification of human subjects grew.[304] The genotypic data generated by a GWA study is not meaningful without the associated phenotypic data. Because a GWA study often seeks to associate genotypic information (e.g., particular markers) with particular disease states, information regarding donor demographics, disease state and treatment are necessary to interpret the genotypic findings. The prospect of releasing clinical and phenotypic data to the public raised concern and led to the imposition of various policy restrictions on data users' ability to disclose and transfer data, as well as the controlled-access mechanisms enabled through repositories such as dbGaP.[305]

2. Scientific Advancement and Publication Priority

As discussed above, the more quickly scientific data is disseminated, the more quickly science will progress. Conversely, when the release of data is delayed due to the length of the publication cycle and patenting concerns, it can be argued that the progress of scientific advancement is retarded, or at least that it may not achieve its greatest potential. If data were not withheld until a researcher's conclusions were published, but released prior to publication, the months-long delays associated with the publishing process could be avoided. Following this line of argument, in an ideal world, maximum scientific efficiency could be achieved by reducing the delay between data generation and data release to zero. That is, the most rapid pace of innovation, discovery of new therapies, development of new technologies, and understanding of natural phenomena could be achieved by releasing scientific data to the community the moment it is generated.

Publication is, however, of crucial importance to scientific careers. Scientists typically spend months validating and analyzing their data, formulating hypotheses, re-running procedures, refining data, and then preparing the manuscript

304. *See Toronto Report*, *supra* note 66, at 170.

305. For a general discussion of the protection of human subjects data in genomic studies, a topic that is beyond the scope of this paper but which has been extensively addressed in the literature, see for example, ANDREWS ET AL., *supra* note 75, at ch. 13; Crolla, *supra* note 75, at 241–47.

of the paper that will present their results to the community. What rational scientist would wish to give this data away before he or she has had a chance to analyze it? Why would he or she enable competitors who have done none of the work to benefit from the data to the same degree as he or she?[306] Even Merton, who championed the norm of scientific communalism, did not specify how *quickly* the sharing of data should occur.[307]

Thus, a clash of cultures has arisen, with the result being a heightened focus on the extent to which users of publicly released data may be restricted in their ability to present or publish results based on that data. The compromise in several recent cases has been time-based. That is, the "embargo" periods in the GAIN Policy, NIH GWAS Policy, and ENCODE 2008 Policy all give users access to data and let them perform research, but prohibit them from making related presentations or submitting related papers during the embargo period.[308] The approach taken by private consortia, in contrast, protects data generator priority by allowing data generators to retain data privately for a specified period. The trade-offs between these differing approaches is discussed below.

3. Patent Encumbrances

Patent protection is related to, but distinct from, the issue of publication priority. As discussed previously, early in the HGP, following the EST patenting debate, NIH representatives adopted a position that patent protection is inappropriate for DNA sequence information.[309] This stance, also held by leaders of the scientific community and international funding agencies, is reflected in the Bermuda Principles.[310] Accordingly, a number of the data release policies developed by private and academic consortia, such as those adopted by the International HapMap Consortium, GAIN, the SNP Consortium, and International SAE Consortium, take explicit steps to prevent

306. *See* Eisenberg, *Patents and Data-Sharing*, *supra* note 47, at 1021 ("Scientists who share their data promptly and freely may find themselves at a competitive disadvantage relative to free riders in the race to make and publish future observations . . . .").

307. *See* Margo A. Bagley, *Academic Discourse and Proprietary Rights: Putting Patents in their Proper Place*, 47 B.C. L. REV. 217, 227 (2006) (quoting Robert K. Merton, *The Normative Structure of Science*, *in* THE SOCIOLOGY OF SCIENCE 274–75 (1973)).

308. *Supra* Table 1.

309. *See supra* notes 93–94 and accompanying text.

310. Contreras, *Prepublication Data Release*, supra note 5, at 393.

the patenting of results generated by their research.[311]

NHGRI, however, must operate within the constraints of the Bayh-Dole Act.[312] Thus, while NHGRI's various post-Bermuda data release policies all acknowledge the requirements of the Bayh-Dole Act, they demonstrate a general bias against the placement of patent encumbrances on genomic data.[313] The enforceability, however, of policy provisions that merely "urge" or "encourage" data generators and users not to seek patents on inappropriate subject matter is open to some doubt.[314]

Lacking a strong policy tool with which to limit expressly the patenting of genomic information, NHGRI policy makers have employed rapid pre-publication data release requirements as a surrogate for achieving the same result. In particular, the Bermuda Principles and their adoption and reaffirmation by NHGRI in 1997 and 2003, respectively, ensured that genomic data from the HGP and other large-scale sequencing projects would be made publicly-available before data generators had an opportunity to file patent applications covering any "inventions" arising from that data and in a manner that ensured its availability as prior art against third party patent filings at the earliest possible date.[315]

When publication priority issues began to emerge with the movement toward GWAS and other studies involving phenotypic data components, the publication embargo was offered as a solution that both protected the publication interests of data generators, but still ensured the early release of data and, consequently, the patent-frustrating effects produced by the rapid pre-publication data release principles espoused by the Bermuda Principles.

311. *Supra* Table 1.

312. NHGRI 1996 POLICY, *supra* note 107.

313. *See ENCODE Pilot Policy*, *supra* note 131; NHGRI 1996 POLICY, *supra* note 107 and accompanying text; NIH GWAS Policy, *supra* note 172.

314. *See* Rai & Eisenberg, *supra* note 106, at 309.

315. Interestingly, Rebecca Eisenberg suggests that, in some cases, the early release of experimental data may actually encourage more patent filings by third parties who are thereby enabled to combine public data with proprietary improvements and patent the combination thereof. *See* Eisenberg, *supra* note 47, at 1026.

V. CONCLUSION

The twenty-year evolution of the genome commons illustrates the ways in which the distinct policy objectives of the relevant stakeholder communities have interacted over time to shape the formal and informal rules that govern the commons. Although governmental agencies played a significant role in the ongoing development of the policies governing the genome commons, other stakeholder groups including data generators, data users, and the public have strongly influenced the direction that these rules have taken. While the groundbreaking Bermuda Principles were straightforward in their implementation and effect, subsequent policies reflect an increased complexity that has arisen from the need to balance the competing and sometimes contradictory interests of these stakeholder groups.[316]

The policy considerations described in this paper are by no means unique to the genome commons. Issues relating to the advancement of science, the appropriate level of patent protection for scientific discoveries, and value-maximizing rewards for researchers are pervasive in many fields of study. Thus, the lessons learned, and the compromises reached, by the designers of the genome commons can inform the discussion and analysis of scientific commons in a variety of fields. Moreover, as the genome commons continues to mature and expand into areas such as proteomics and metabolomics, the policies in existence today will likewise evolve. It is hoped that policy makers considering the design of new commons in these areas will look to the past to understand the complex compromises and rationales behind the policies that flowed from the Bermuda Principles, the legacy of which is likely to remain influential for years to come.

316. For a discussion of the use of timing or "latency" variables as effective means for mediating among these competing stakeholder interests, see generally Contreras, *Data Sharing*, *supra* note 5.

APPENDIX – GLOSSARY OF TERMS

CRP—Community Resource Project (description of a type of research project developed at the Ft. Lauderdale data release meeting)

DAC—Data Access Committee

dbGaP—Database of Genotypes and Phenotypes (database administered by the NIH's National Library of Medicine)

DDBJDNA—Databank of Japan (leading international DNA sequence repository)

DOE—U.S. Department of Energy

EMBL—European Molecular Biology Laboratory (leading European genomics research center and host of a large DNA sequence repository in Hinxton, England)

ENCODE—Encyclopedia of DNA Elements (an NIH-funded research project seeking to identify functional elements of the human genome)

EST—Expressed Sequence Tag (a short fragment of DNA)

GAIN—Genetic Association Information Network (a consortium formed to conduct GWA studies on six common human diseases)

GWAS—Genome-Wide Association Study (a study that seeks genetic markers for specified physiological or pathological traits)

GenBank—A publicly-accessible database of genetic sequences maintained by NCBI

HGP—Human Genome Project (the U.S.-led international project to sequence the human genome)

HIPAA—Health Insurance Portability and Accountability Act (U.S. legislation governing, among other things, privacy of patient healthcare data)

HUGO—Human Genome Organisation (an international, policy-oriented group formed near the beginning of the HGP)

Kb—Kilobase (unit of measurement equal to 1,000 DNA base pairs)

MRC—UK Medical Research Council (principal medical funding agency in Britain)

NAS—National Academy of Sciences (non-governmental agency that advises the U.S. government on scientific matters)

NCBI—National Center for Biotechnology Information (NIH center that operates bioinformatics resources such as GenBank)

NCI—National Cancer Institute (NIH institute dedicated to cancer research, including cancer genomics)

NHGRI—National Human Genome Research Institute (the principal U.S. funding agency for genomic research and one of the NIH institutes)

NIH—National Institutes of Health (the principal U.S. funding agency for biomedical research, comprised of numerous different institutes)

NRC—National Research Council of the U.S. National Academy of Sciences

SNP—Single-Nucleotide Polymorphism (a "marker" in the genetic code)

TCGA—The Cancer Genome Atlas (an NIH-funded pilot project relating to cancer genomics)

WTCCC—Wellcome Trust Case Control Consortium (a large-scale UK-based GWA study)