CONSTRUCTING THE GENOME COMMONS

Working Draft: 18 September 2011

Jorge L. Contreras[*]

I. INTRODUCTION

The human genome project (HGP), which spanned fifteen years and involved over a thousand scientists worldwide, has dramatically changed biomedical science and technology.[1] The genetic basis for most common hereditary diseases is now known, and genetic tests are widely-available for these diseases and other physical traits.[2] Among the HGPs greatest contributions to society is the vast quantity of genetic data that it produced and made freely available in public databases. This global data resource, which has continued to expand at a breathtaking pace since the HGP concluded more than a decade ago, is what I refer to as the "genome commons."[3]

Today the free availability of genomic data is a fundamental feature of the scientific research landscape. But the existence of this invaluable public resource was by no means assured when the HGP was initiated in the early 1990s. In fact, it was widely believed (and feared) that the majority of genomic data would be held in proprietary databases, protected by patents or confidentiality restrictions, and made available to researchers only under costly subscription agreements. This alternative model, in fact, was the one initially proposed by Celera Corporation, which competed with the public HGP to complete the initial human genomic sequence from 1998 to 2001.[4]

The fact that the genome commons is today a global, public resource owes much to a 1996 accord reached in Bermuda by scientific leaders and policymakers. These groundbreaking "Bermuda Principles" required that all DNA sequence data generated by the HGP be released to the public a mere twenty-four hours after generation,[5] a stark contrast to the months or years that

---

[*] American University – Washington College of Law.

[1] *See, e.g.*, FRANCIS S. COLLINS, THE LANGUAGE OF LIFE 3 (2010) ("[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"); NAT'L RESEARCH COUNCIL, REAPING THE BENEFITS OF GENOMIC AND PROTEOMIC RESEARCH 38–40 (2006) [hereinafter NRC, GENOMIC AND PROTEOMIC RESEARCH]; ARTHUR M. LESK, INTRODUCTION TO GENOMICS 305-07 (2007)

[2] 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).

[3] Jorge L. Contreras, *Prepublication Data Release, Latency, and Genome Commons*, 329 SCIENCE 393, 393 (2010) [hereinafter Contreras, *Prepublication Data Release*].

[4] *See* Section x, *infra*.

[5] *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].

usually preceded the release of scientific data.[6] The Bermuda Principles arose from an early recognition by scientists and policy makers that rapid and efficient sharing of data would be necessary to coordinate activity among the geographically distant laboratories working on the massive HGP. But project coordination was not the only factor motivating the unorthodox rapid-release requirement of the Bermuda Principles.[7] More importantly, this approach arose from the conviction among several project leaders that rapid release of genomic data was necessary for the advancement of scientific research, medical discovery and the improvement of human health.[8] Related to this sentiment was a third policy rationale for rapid data release: preventing the encumbrance of DNA sequence data by intellectual property rights, particularly patents.[9] While this policy objective was seldom discussed openly, it reflects a current that runs through many of the early (and recent) debates regarding the release and sharing of technical information.

The Bermuda Principles continue to shape data release practices of the genomics research community today and have established "rapid pre-publication data release" as the norm in this and other fields.[10] Advances in science and technology, however, together with increasingly difficult ethical and legal issues, have complicated the data release landscape and given rise to policy considerations not foreseen in Bermuda. Among these are need to protect human subject data, even at the genomic level, and the desire of scientists who generate data sets to publish their findings before data is accessed and used by others. The emergence and recognition of these considerations has led to an evolution of genomics data release policies that are more restrictive, complex and sophisticated than those of the HGP, but which still preserve the fundamental public nature of the genome commons.

In their recent article, *Constructing Commons in the Cultural Environment*,[11] Michael Madison, Brett Frischmann, and Katherine Strandburg question the prevailing functionalist view of information production, which maintains that non-rivalrous public goods (such as scientific research) are likely to be under-produced absent the private incentives afforded either by intellectual property protection (contributing to what they term the "domain of exclusion") or governmental subsidies (contributing to the "public domain").[12] Madison, Frischmann and

[6] 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.

[7] 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.

[8] *See, e.g.*, HGP Initial Paper, *supra* note x, 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").

[9] 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 x, at 117. The increasing trend toward patenting of genetic material alarmed many of the leaders of the HGP. *See infra* note x and accompanying text.

[10] For a detailed history of the Bermuda Principles and their lasting effect on genomic research data release policies, *see* Jorge L. Contreras, *Bermuda's Legacy: Policy, Patents and the Design of the Genome Commons*, 12 MINN. J. L., SCI. & TECH 61 (2011).

[11] Michael J. Madison, Brett M. Frischmann & Katherine J. Strandburg, *Constructing Commons in the Cultural Environment*, 95 CORNELL L. REV. 657, 659 (2010) [hereinafter Cultural Commons].

[12] *Id*. at 666-67.

Strandburg call for the systematic analysis of "cultural" commons institutions[13] to develop an alternative to this functionalist theory of production, which they criticize as an oversimplification that ignores the many cultural practices and institutions that are not motivated primarily by property-based incentives.[14] In support of this analysis they offer a modified version of the Institutional Analysis and Development (IAD) framework pioneered by Elinor Ostrom in the 1980s and early 1990s in connection with the study of physical resource commons[15] and adapted by her and Charlotte Hess to the analysis of information commons.[16]

In this article, I engage the modified IAD framework proposed by Madison, Frischmann and Strandburg to elucidate the structural and narrative elements of the genome commons as it has evolved over the years. In particular, I assess the manner in which the backers of this public resource vigorously opposed the proprietary models advanced by Celera Genomics, and how, after Celera and most other commercial entities exited the scene, the genome commons has continued to remain relatively free of encumbrances imposed by intellectual property, but has become increasingly burdened by administrative procedures and restrictions on usage imposed to satisfy different constituencies. As such, the genome commons exhibits characteristics of the cultural commons identified by Madison, Frischmann and Strandburg, which govern the production of intellectual assets by means other than traditional intellectual property exclusion and incentives.

[*insert article roadmap*]

II. COMMONS THEORY BACKGROUND

Since the Middle Ages, the term "commons" has denoted shared physical spaces such as fields, pastures and forests that were open and free for exploitation by farmers, herdsmen and other local peoples.[17] Elinor Ostrom and her collaborators conducted the seminal analysis of social and organizational structures governing physical commons in the 1980s.[18] Among Ostrom's many insights was the applicability of the well-known Institutional Analysis and Development (IAD) framework, employed since the 1970s to evaluate organizational characteristics and institutional decision-making, to common-pool resources.[19] Under the IAD framework, commons structures may be examined with respect to three broad sets of characteristics: those of the common resource itself, the "action arena" in which the common

[13] 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.
[14] Cultural Commons, supra note x, at 665, 668.
[15] *See* ELINOR OSTROM, GOVERNING THE COMMONS — THE EVOLUTION OF INSTITUTIONS FOR COLLECTIVE ACTION (1990).
[16] *See* Ostrom & Hess, *supra* note x, at 42-43.
[17] *See* Hess & Ostrom, *Introduction* to KNOWLEDGE AS A COMMONS, *supra* note x, at 12; NANCY KRANICH, THE INFORMATION COMMONS — A PUBLIC POLICY REPORT 10 (2004), *available at* http://www.fepproject.org/policyreports/InformationCommons.pdf. In the U.S., "commons" have also been associated historically with New England's open town squares that served as popular venues for speechifying and pamphleteering. Hess & Ostrom, *supra* at 13. In both cases, "commons" terminology has a strong traditional association with freedom and openness.
[18] OSTROM, *supra* note x.
[19] *See* Charlotte Hess & Elinor Ostrom, *A Framework for Analysing the Microbiological Commons*, 58 INT'L SOC. SCI. J. 335, 339 (2006); Ostrom & Hess, *supra* note x, at 42–43.

resource is utilized, and the desired or actual outcomes of the commons structure.[20] Each of these broad areas may be 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 the controlling rules set, whether legal or norms-based.[21] The application of the IAD framework analysis results in a deeper understanding of the factors that should be considered when structuring or evaluating a commons structure and allows comparison of the attributes of otherwise unrelated common resources. Ostrom and others have persuasively applied the IAD framework to common resource arrangements ranging from fisheries to irrigation systems to environmental governance.[22]

In the mid-1990s scholars began to apply commons theory to intangible shared resources and information.[23] Since then, much has been written about so-called "information commons" in areas including computer software, network capacity, artistic content, scholarly learning and scientific data.[24] Information commons are, of course, different than aggregations of finite physical resources inasmuch as information is generally viewed as "non-rivalrous," meaning that any number of individuals may enjoy its benefits without depleting it. Building upon their earlier work on physical commons, Ostrom and Hess have applied the IAD framework to the analysis of knowledge-based commons structures, reasoning that both physical resource commons and information commons share numerous attributes.[25]

Last year, Michael Madison, Brett Frischmann and Katherine Strandburg critically re-examined the IAD framework in relation to commons in the "cultural environment." In doing so, they recognized that, unlike the farmers and fishermen who exploit physical commons of natural resources, users of information commons not only *use* the common resource, but *produce* it as well.[26] This insight led them to propose a modified framework that more closely links the features of the common resource to its users/producers, as mediated through constructed "rules in use". *Figure 1* illustrates the modified IAD framework that Madison, Frischmann and Strandburg have proposed in the context of cultural commons.

---

[20] *See* Ostrom & Hess, *supra* note x, at 44–45.
[21] *Id.* at 45–53.
[22] *See* Hess & Ostrom, *supra* note 19, at 339.
[23] Hess & Ostrom, *supra* note 17, at 4 (noting the "explosion" of information commons scholarship beginning around 1995).
[24] *See, e.g.*, HAL ABELSON, KEN LEDEEN & HARRY LEWIS, BLOWN TO BITS — YOUR LIFE, LIBERTY, AND HAPPINESS AFTER THE DIGITAL EXPLOSION 277 (2008) (discussing the application of commons theory to broadcast spectrum); LAWRENCE LESSIG, THE FUTURE OF IDEAS 85-86 (2001) (arguing that commons systems have encouraged innovation, specifically with respect to software, telecommunications and the Internet); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET AND HOW TO STOP IT 78–79 (2008) (discussing commons approaches both to Internet content and hardware); Yochai Benkler, *Coase's Penguin, or Linux and the Nature of the Firm*, 112 YALE L.J. 369 (2002) (arguing that "commons-based peer production" of software has proven to be both viable and efficient, as demonstrated by the model of the Linux operating system); James Boyle, *The Second Enclosure Movement and the Construction of the Public Domain*, 66 LAW & CONTEMP. PROBS. 33, 44–49 (2003) (discussing open source software).
[25] Ostrom & Hess, *supra* note x, at 43.
[26] *Id.* at 681. In this respect, they echo the well-known principle that users of intellectual property are also its producers. *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 13–14 (2003).

Figure 1[27]

[INSERT REVISED IAD PICTURE]

The modified IAD framework proposed by Madison, Frischmann and Strandburg

Madison, Frischmann and Strandburg offer this modified IAD framework to provide a basis for the systematic theoretical and empirical study of cultural commons systems.[28] They hope that the collection of information using this common framework will demonstrate the existence of "a wide variety of formal and informal institutional arrangements" that operate according to principles beyond the functionalist intellectual property-based account that is often used to explain and justify the production of cultural assets.[29]

III. ATTRIBUTES OF THE GENOME COMMONS

In this Section, I describe the principal definitional attributes of the genome commons as conceptualized under the modified IAD framework: the characteristics of the common resource and the attributes of the community engaged in creating and using it. The third descriptive element of the IAD framework, the "rules-in-use" of the commons, are discussed in Section IV below.

A. RESOURCE CHARACTERISTICS

The genome commons is, at its most basic level, a massive collection of data stored in publicly-managed electronic databases across the world. In order to understand the unique nature of this data resource it is useful to describe both the data contained within it (i.e., human DNA information) and the databases that house it, as well as the underlying legal environment that applies to such aggregations of data.

1. Genomic Data.[30] 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

[27] *Cultural Commons*, supra note x, at x.
[28] *Id.* at 678 ("[a] research framework such as ours aims to systematize the investigation, facilitate a more rigorous evaluation by matching and testing of theories and models with observed phenomena, and, most generally, enable learning over time").
[29] *Id.* at 665.
[30] 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 x; MATTHEW RIDLEY, GENOME, 6–10 (1999); WILLIAM S. KLUG & MICHAEL R. CUMMINGS, ESSENTIALS OF GENETICS (3rd ed. 1999).

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. Each human genome is approximately 99.5% identical, but very small differences are responsible for the great variability in human physical and physiological traits.

Some of the segments of DNA strands within an organism's cells form functional units called "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 between 20,000 and 25,000 genes. 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. 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 been to associate particular genes or genetic variations or "mutations" with phenotypic traits.

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.

The race to sequence the human genome is described in Section x, *infra*. For purposes of this discussion, suffice it to say that since the completion of the initial draft of the human genome sequence in 2001, the HGP and follow-on projects have generated vast amounts of genomic data, including the full genomic sequences of hundreds of individual humans and thousands of other organisms. Today, additional international efforts are under way to sequence the genomes of thousands of additional individuals to create still more complete and detailed reference maps of the human genome[31] and to sequence the genomes of the multitude of microorganisms residing within the human body.[32]

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) have been credited with identifying variants in

[31] Erika Check Hayden, *International Genome Project Launched*, 451 NATURE 378, 378 (2008); cite UK 10,000 genome project.
[32] Peter J. Turnbaugh, et al., *The Human Microbiome Project*, 449 NATURE 804, 804 (2007).

multiple genes that increase susceptibility for complex conditions such as Type 2 diabetes,[33] breast cancer,[34] prostate cancer,[35] hypertension[36] and numerous other diseases.[37] 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.[38] According to most predictions, the data comprising the genome commons is expected to continue to expand at a rapidly-increasing rate for the foreseeable future.[39]

2. *Data and Databases*. For hundreds of years, the traditional means of disseminating scientific information has been the peer-reviewed journal article. Scientists are judged, both for purposes of career advancement and the awarding of government grants, on the quantity of their publications, making the publication of scholarly articles of paramount importance to many scientists and giving scientists a significant personal incentive to publish and thus share their data with others.[40] Yet, despite the prevalence of scientific publications, there are two principal reasons that journal publication has proven to be wholly inadequate for the dissemination of genomic data.

First, the sheer quantity of genomic data is far too large to be published in any reasonable format, and is only useful if available for electronic manipulation and analysis. One source estimates that if the entire human genome of 3.2 billion base pairs were printed in paper format, it would occupy 200,000 pages, roughly equivalent to 200 New York yellow pages directories.[41]

---

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

[34] D.F. Easton, et al., *Genome-Wide Association Study Identifies Novel Breast Cancer Susceptibility Loci*, 447 NATURE 1087 (2007); D.J. Hunter, et al., *A Genome-Wide Association Study Identifies Alleles in FGFR2 Associated with Risk of Sporadic Postmenopausal Breast Cancer*, 39 NATURE GENETICS 870.

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

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

[37] *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 approximately hundreds of publications identifying more than 100 diseases and traits).

[38] 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 x, at 23. Between 1999 and 2009, the cost of gene sequencing technology dropped by an astonishing factor of 14,000. Collins, *supra* note x, 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 x at 675.

[39] Cite.

[40] Robert K. Merton, *Priorities in Scientific Discovery* (1957), *reprinted in* THE SOCIOLOGY OF SCIENCE 286, 316 (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").

[41] U.S. Dept. of Energy – Office of Science – Office of Biological & Environmental Research, Human Genome Project Information (available at http://www.ornl.gov/sci/techresources/Human_Genome/faq/faqs1.shtml) (visited Sept. 10, 2011).

Accordingly, a journal article typically includes only a brief presentation of significant experimental findings, often made in summary or tabular fashion, together with the scientist's analysis and conclusions based upon those findings.[42] While the published data are usually essential to support the scientist's analysis, the data reported in a journal article typically represent only a small fraction of the "raw" data set. Yet in order to enable the verification and reproduction of an experiment by other scientists, the full data set is often required in a usable, electronic format.[43]

Second, there is usually a lengthy delay between the completion of data collection and publication in a journal. 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.[44] Older studies have found comparable or longer delays in other fields of research.[45] Clearly, in a field in which access to experimental data is required quickly so as to enable additional studies and analysis, these lengthy delays are highly undesirable.

These two considerations have led to the practice of making large scientific data sets available independently of journal articles. As discussed in greater detail in Section x below, many science funding agencies now require that genomic data be released into public databases shortly after it is generated.[46] A growing number of scientific journals now also require that authors make the data underlying their published results available to readers on a web site

---

[42] *See generally* Rebecca S. Eisenberg, *Patents and Data-Sharing in Public Science*, 15 INDUS. & CORP. CHANGE 1013, 1024 (2006). By way of example, one recently-published study identifies the *fgf4* gene as a factor leading to short-leggedness in dogs such as the Welsh corgi and the dachshund. Heidi G. Parker, *An Expressed Fgf4 Retrogene Is Associated with Breed-Defining Chondrodysplasia in Domestic Dogs*, 325 SCI. 995 (2009). The association of fgf4 with the physical or "phenotypic" trait of short-leggedness is an experimental *result*. A vast quantity of *data* had to be collected and generated in order to arrive at this result, including raw genetic sequence reads for numerous dogs across different breeds, associated phenotypic data for each of the subjects, and a complex of statistical analyses, associations and computations.

[43] The actual biological samples from which DNA is extracted, and the extracted DNA itself, play a relatively minor role in most genomic research and analysis. The focus of most genomic research today is on computational, computer-based statistical analysis, rather than chemical or biochemical analysis of DNA samples. As Ostrom and Hess have observed, modern biology is an "*information* science". Charlotte Hess & Elinor Ostrom, *A Framework for Analysing the Microbiological Commons*, 58 INTL. SOC. SCI. J. 335, 335 (2006) [hereinafter Hess & Ostrom, *Framework*].

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

[45] *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.

[46] See x

accessible through the journal, through their own institutions or in a government-maintained database.[47] These databases have enabled the efficient, rapid and cost-effective sharing of new knowledge and the pursuit of studies and analyses that otherwise might have been impossible.[48]

The principal databases for the deposit of genomic sequence data are GenBank, which is administered by the National Center for Biotechnology Information (NCBI) a division of the NIH's National Library of Medicine, the European Molecular Biology Library (EMBL) in Hinxton, England, and the DNA Data Bank of Japan (DDBJ).[49] NCBI also maintains the RefSeq database, which consolidates and annotates much of the sequence data found in GenBank.[50] In addition to sequence data, genomic studies generate data relating to the association between particular genetic markers and disease risk and other physiological traits.[51] This type of data, which is more complex to record, search and correlate than the raw sequence data deposited in GenBank, is housed in databases such as the Database of Genotypes and Phenotypes (dbGaP), operated by NIH's National Library of Medicine. dbGaP can also accommodate phenotypic data, which includes elements such as de-identified subject age, ethnicity, weight, demographics, exposure, disease state, and behavioral factors, as well as study documentation and statistical results, including linkage and association analyses.[52] Given the potential sensitivity of phenotypic data, dbGaP allows access to data on two levels: open and controlled. Open access data is available to the general public via the Internet and includes non-sensitive summary data, generally in aggregated form. Data from the controlled portion of the database may be accessed only under conditions specified by the data supplier, often requiring certification of the user's identity and research purpose.

3. *Legal Background Environment*. Madison, Frischmann and Strandburg suggest that an understanding of the "natural" environment in which a cultural commons exists is critical to understanding the attributes and operation of that commons.[53] In the case of collections of intangibles, this natural environment necessarily includes the intellectual property rules that govern rights and permissions with respect to the elements of the common resource. The genome commons thus presents a complex picture, as it embodies both biomedical discoveries, which are

[47] *See, e.g.*, *Guide to Publication Policies of the Nature Journals*, NATURE (last updated Apr. 30, 2009), http://www.nature.com/authors/gta.pdf; *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).

[48] *See* Eisenberg, *supra* note x, at 1020.

[49] *See* LESK, *supra* note x, 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).

[50] See http://www.ncbi.nlm.nih.gov/RefSeq/. *See also* Heidi Williams, *Intellectual Property Rights and Innovation: Evidence from the Human Genome* 40 (Aug. 20, 2010) (cite).

[51] 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 accessed 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.*

[52] *See generally*, dbGaP, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gap (last visited Oct. 27, 2010).

[53] *Cultural Commons*, *supra* note 3, at 684-88.

typically addressed via the patent system, as well as large aggregations of data, which are typically addressed via access restrictions and copyright rules.

a. *Patents and DNA.* Patents may be obtained in most countries to protect novel and inventive articles of manufacture, compositions of matter and processes.[54] Excluded from patentable subject matter, however, are laws of nature and natural phenomena.[55] The fundamental question, thus, is whether DNA sequence information and medical conclusions drawn from DNA information are more akin to "inventions" that are protectable by patents, or "products of nature" that are not.

The debate regarding the patentability of DNA sequence information began in earnest in the early 1990s, shortly after DNA sequencing became practical at large scales. NIH was among the first to seek patent protection for DNA sequences. In 1991, a group led by J. Craig Venter, then a scientist at the National Institute of Neurological Disorders and Stroke (NINDS), filed patent applications claiming 337 short genetic sequences known as expressed sequence tags (ESTs). NIH announced this filing as well as its intention to continue to file EST patent applications on a monthly basis.[56] The public response to this announcement was vociferous and triggered what Robert Cook-Deegan describes as "an international firestorm."[57] The debate within NIH was equally vehement and ultimately led to James Watson's resignation in protest from the agency that oversaw the HGP.[58] 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 rejection of its initial EST patent applications,[59] and since then it has adopted a consistently lukewarm, if not outright averse, attitude toward the patenting of genetic sequences.[60] This attitude is reflected in the agency's support for the Bermuda Principles and the data release and patent policies adopted by the agency in the years thereafter.[61]

[54] *See* 35 U.S.C. §101 (2006). This requirement has been broadly interpreted by the United States Supreme Court to include "anything under the sun that is made by man." *Diamond v. Chakrabarty*, 447 U.S. 303, 309 (1980).
[55] *Diamond v. Diehr*, 450 U.S. 175, 185 (1981) (recognizing exclusions to patentability for "laws of nature, natural phenomena, and abstract ideas.")
[56] *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).
[57] *See* COOK-DEEGAN, *supra* note **9**, 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).
[58] Watson decried NIH's EST patenting program as "sheer lunacy." SHREEVE, *supra* note x, 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 **9**, at 317.
[59] *See* LARGE-SCALE SCIENCE, *supra* note x, 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).
[60] 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 x, at 53. For an overview of legal objections to the practice of patenting ESTs, see *id*. at 52, and Barry, *supra* note x, at 18–21.
[61] See discussion at Section x, infra.

Nevertheless, the patenting of genetic information by academic research institutions and private enterprises has continued, leading to one oft-quoted estimate that, by 2005, a full 20% of human genes were covered by U.S. patents.[62] [*Add short discussion of current status – Myriad and Prometheus*] [63]

2. *Data and Databases*. Under U.S. law it has long been held that "facts" such as scientific data are not subject to copyright protection,[64] and databases that merely contain compilations of factual information similarly lack any cognizable legal protection.[65] Nevertheless, access to data that is contained in electronic databases can be controlled by the database operator using technical means such as password-restricted access. And while the data itself may not be subject to legal protection, circumventing such technical protection measures can be prosecuted under a number of legal theories.[66] Thus, scientific information that might otherwise be in the public domain can become encumbered when compiled in proprietary databases.[67] This approach was the one initially adopted by Celera Genomics when it announced its intention to sequence the human genome in competition with the publicly-funded HGP,[68] and the looming threat of propertization of the genome in this manner has fueled continuing public support for GenBank, dgGaP and other publicly-accessible repositories for genomic data.

B. ACTORS AND STAKEHOLDERS

Much 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. This analysis is equally valuable in the context of the information commons. While genomic data release policies are typically drafted and adopted by funding agencies, NIH in particular has given substantial deference to the views and opinions of the scientific community when developing policy, while also seeking to represent the interests of the general public.[69] Thus, the role and influence of other stakeholder

---

[62] Kyle Jensen & Fiona Murray, *Intellectual Property Landscape of the Human Genome*, 310 SCIENCE 239, 239 (2005). Since it was published, various attempts have been made to refute the claims made in this paper. *See, e.g.,* [add counter-cites]

[63] [Note: update citations] *See, e.g.*, *Ass'n for Molecular Pathology v. US Patent and Trademark Office*, No. 09 Civ. 4515, 2010 WL 1233416 (S.D.N.Y. Mar. 29, 2010) (ruling that patent claims which claimed isolated human DNA sequences (genes) or methods of comparison of these genes to those of a human sample were invalid under 35 U.S.C. §101 as they were products of nature, and thus natural phenomena); *Prometheus Laboratories, Inc. v. Mayo Collaborative Services*, No. 09-490, 2010 WL 2571881 (U.S. June 29, 2010) (claims at issue involve simple methods of medical diagnosis and the case was remanded to the appellate court for reconsideration of whether these simple method claims meet the subject matter requirements of 35 U.S.C. §101 in light of the Supreme Court's decision in *Bilski v. Kappos*). Bilski v. Kappos, 130 S. Ct. 3218 (U.S. June 28, 2010) (affirming the Federal Circuit's finding that a simplistic business method claim of hedging risk in commodities was not patentable subject matter).

[64] INS v. AP. Note

[65] Feist (holding that a "white pages" directory of names and telephone numbers lacked sufficient creative content to merit copyright protection).

[66] E.g., the anti-circumvention provisions of the Digital Millennium Copyright Act, and the Computer Fraud and Abuse Act.

[67] See Reichman & Uhlir, supra note x, for a discussion of this phenomenon.

[68] See discussion at notes x, infra, and accompanying discussion.

[69] *See, e.g.*, Collins et al., *supra* note x, at 846 (2003) (describing community involvement in setting goals for national genomics research program). NIH has most recently requested public comment and feedback on data release policies in October 2009. Press Release, Nat'l Insts. of Health, Notice on Development of Data Sharing Policy for Sequence and Related Genomic Data, NOT-HG-10-006 (Oct. 19, 2009), *available at* http://grants.nih.gov/grants/guide/notice-files/NOT-HG-10-006.html.

groups is not to be underestimated: the development of data release policies in the genome sciences has been a process of negotiation and compromise. 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."[70] As such, the U.S. National Institutes of Health (NIH) and Department of Energy (DOE), which funded the bulk of the massive project, together with their counterparts at the Wellcome Trust in the U.K., exerted a significant degree of influence over both its technical and policy dimensions.[71] 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 Center for Human Genome Research in 1988. Other scientists involved in the early planning and execution stages of the project were also globally prominent and included numerous Nobel Prize winners.[72] 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 nature of several HGP policies.[73]

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 genetically-linked diseases. The HGP, in contrast, treated the mapping of the human genome as a campaign of large-scale data production.[74] 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).[75] The intensity of this work, the amount of capital equipment required to undertake it, and the degree of specialization demanded by the emerging 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

[70] INSTITUTE OF MEDICINE & NATIONAL RESEARCH COUNCIL, LARGE-SCALE BIOMEDICAL SCIENCE 29 (2003) [hereinafter LARGE-SCALE SCIENCE].

[71] 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 x, at 39.

[72] In addition to Watson (Chemistry, 1962), the HGP leadership group included Nobelists 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).

[73] 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; *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).

[74] *Id.* at 1182.

[75] *See* LARGE-SCALE SCIENCE, *supra* note x, at 39.

the biosciences continues to increase.[76] Like other scientists, data-generating scientists share two principal concerns: (a) obtaining funding for their work and (b) 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.[77]

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. The data generated by the HGP and subsequent projects have eliminated the need for researchers to conduct much of this groundwork. Unlike the close-knit community of data generators at large-scale sequencing centers, there is no coherent community of data users. These comprise scientists across the world whose research may benefit from the use of genomic data.

4. Data Intermediaries. Individual scientists and laboratories that generate data are seldom the ones that make this 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)[78]). 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. Recently, the critical role of scientific journals in the creation and sustainability of the genome commons has been recognized, particularly with respect to the need to offer meaningful and career-enhancing publication opportunities to data generating scientists.[79]

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.[80] Similar characteristics applied to other early genomic projects such as the HapMap Project.[81] 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

[76] The implications of participating in large-scale data generating work on the careers of junior scientists has been the subject of much discussion. *See* LARGE-SCALE SCIENCE, *supra* note x, 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].
[77] *See* Contreras, *Prepublication Data Release*, *supra* note 3, at 393; Contreras, *Data Sharing*, *supra* note 3, at 38.
[78] *See* Contreras, *Data Sharing*, *supra* note 3, at 38.
[79] *See* Ft. Lauderdale Principles, supra note x, at 4; Toronto Authors, *supra* note x, at 170.
[80] *The Human Genome Project Completion: Frequently Asked Questions*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/11006943 (last visited Oct. 28, 2010).
[81] *See* Eisenberg, *supra* note x, at 1026.

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.[82] 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.[83] 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.[84] 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.[85]

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. Beyond general interest, however, are several 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 intense interest in the direction and results of research.[86] 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 both the motivation and the financial wherewithal to lobby for changes in research policy.[87] Finally, members of the general public beyond patient advocacy groups have begun to take an interest in, and to express concern regarding, genomic research and the data sharing practices of genomics researchers.[88]

---

[82] *See* Toronto Report, *supra* note 76, at 170.

[83] 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).

[84] 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.

[85] *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.

[86] *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).

[87] *See* Lee, *supra* note 73, 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).

[88] See S.B. Haga & J.O'Daniel, *Public Perspectives Regarding Data-Sharing Practices in Genomics Research*, Public Health Genomics, Mar. 24, 2011 (noting public concern regarding the potential loss of privacy resulting from data sharing practices of genomics researchers).

IV. INITIAL RULES-IN-USE OF THE GENOME COMMONS

Under the IAD framework, the "rules-in-use" or governance structure of a commons system constitute its third primary attribute. When considering physical resource commons, the common resource, whether a forest, a pasture or a body of water, typically exists prior to the imposition of rules regarding its use. Rules-in-use, in this case, typically allocate access and usage rights with respect to this pre-existing commons and, while such rules necessarily affect the sustainability of the common resource and the rate at which it is depleted and replenished, they do not create or define it. As observed by Madison, Frischmann and Strandburg, however, the rules governing constructed cultural commons dictate the commons' very nature, from the size and nature of the common resource, to the speed at which data is deposited in it, to when and how it can be accessed and used. In the case of the genome commons, formal rules-in-use were established at the outset of the HGP and have steadily evolved since then.

For purposes of analyzing the genome commons, I consider the data release policies of the HGP (1992-2001) to constitute the initial conditions imposed on the commons. In Section V, I discuss the evolution of these policies after the completion of the HGP, illustrating the "feedback" loop between the commons action arena and rules-in-use.

A. EARLY YEARS OF THE HGP

Several factors contributed to the call, from the initiation of the HGP, 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 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.[89] Moreover, and perhaps more importantly, there was a sense among the leadership of the project that genomic data, in the words of Ari Patrinos, the DOE's Associate Director for Biological and Environmental Research, that "the genome belongs to everybody."[90] 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."[91]

---

[89] *See* HGP Initial Paper, *supra* note x, at 864; MCELHENY, *supra* note , at xi ("Openness was at the core of the [bacteriophage] ethos, and it soon propagated to the genetic research systems of the future."); NRC – PUBLIC DOMAIN, *supra* note x, 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 x, at 54–56.

[90] 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 x, at 48.

[91] NATIONAL 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).

In 1992, shortly after the project was launched, NIH and DOE developed formal guidelines for the sharing of HGP data.[92] These guidelines were viewed as essential to achieve the program's goals, avoid unnecessary duplication of effort and expedite research in other areas.[93] 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.[94] But the need for project coordination did not require immediate *public* release of the 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.[95] This "advantage" manifested itself in a 6-month maximum period from the time that HGP data are generated until the time that they must be made publicly available. During this 6-month period, HGP researchers could analyze their data and prepare publications, and only after the end of the 6-month period were they required to release the data to the public.[96]

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.[97] This pro-patent attitude arose contemporaneously with NIH's nearly disastrous attempt to seek patents on ESTs, and had waned significantly by the mid-1990s.[98]

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 from the U.S., Europe and Japan met in Bermuda[99] to deliberate over the speed with which HGP data should be released to the public, and whether the 6-month "holding period" approved in 1992 should continue.[100] The resulting Bermuda Principles established that all DNA sequence information from large-scale human genomic sequencing projects should be "freely available and in the public domain in order to encourage research and development and to maximize its benefit to society."[101] They went on to

---

[92] *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].
[93] *Id.*
[94] *Id.*
[95] *Id.*
[96] *Id.*
[97] *Id.*("[I]ntellectual property protection may be needed for some of the data and materials.").
[98] See notes x, supra, and accompanying text.
[99] 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.
[100] *See* Marshall, *supra* note 90, 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.
[101] Bermuda Principles, *supra* note x.

define the method by which such data should be shared, requiring that sequence assemblies greater than one kilobase (Kb) in length[102] should be released automatically *within twenty-four hours*, and that finished annotated sequences should be submitted *immediately* to a public database.[103] 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 6-month data holding period established in 1992 was supported by both the NIH and DOE and had significant international ramifications.[104]

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.[105] 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 project leaders, 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).[106]

2. *Rapid Data Release and Patents*. 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 and reversing the pro-patent position espoused in the 1992 guidelines.[107] In particular, the Bermuda Principles ensured that HGP data would be made

[102] *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. SUMMARY OF THE REPORT OF THE SECOND INTERNATIONAL STRATEGY MEETING ON HUMAN GENOME SEQUENCING, BERMUDA, 27TH FEBRUARY – 2ND MARCH 1997 [hereinafter Bermuda 1997 Report], *available at* http://www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml.

[103] Bermuda Principles, *supra* note x.

[104] 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).

[105] 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).

[106] *See* Bentley, *supra* note 105, 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 100 ("[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.").

[107] Bentley, *supra* note 105, at 533-34; *see also* Marshall, *supra* note 90; JAMES D. WATSON et al., RECOMBINANT DNA 295 (3rd ed. 2005); Rebecca S. Eisenberg, *Genomics in the Public Domain: Strategy and Policy*, 1 NATURE REVIEWS – GENETICS 70, 72 (2000).

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.[108] This result, 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.[109]

In response to this criticism, NHGRI's 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."[110] The consequences if such patenting activity is discovered are left unstated, but the clear implication is that the agency may view future grant applications by "violators" unfavorably.[111]

The significance of NHGRI's implementation of the Bermuda Principles[112] 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.[113] 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

---

[108] 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., 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 , at 1025–26 (discussing the creation of "patent-defeating" prior art through the HGP's data release rules).

[109] 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."). Commentators have argued that NIH's adoption of the Bermuda rapid data release requirements 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 x, at 46 ("Strictly speaking, the policy directly contradicted the Bayh-Dole Act.").

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

[111] For a general critique of the NIH's "hortatory" approach to this issue, see Rai & Eisenberg, *supra* note 109, at 293-94, 306.

[112] *See* NHGRI 1996 Policy, *supra* note 110; 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.

[113] *See* discussion of NASA and other federal policies *supra* note x.

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.

C. PUBLIC VERSUS PRIVATE: THE RACE WITH CELERA

By 1998, the HGP had begun the monumental task of sequencing the human genome at research centers in the U.S., Europe and Japan[114] and which had, to that point, already cost nearly $2 billion.[115] Then, in May of that year, J. Craig Venter, a former NIH scientist,[116] famously proclaimed that he, funded by substantial commercial backers, would utilize a novel technological approach called "whole-genome shotgun" sequencing and a battalion of 300 state-of-the-art machines to complete the sequence of the entire human genome by 2001, a full four years before the publicly-funded HGP was scheduled to complete its work.[117] Venter's announcement, which shocked the scientific establishment, quickly 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.[118] Ultimately, a truce was brokered by the preeminent scientific *Science*, which agreed to publish the genomic sequence generated by Celera, while *Nature* would publish the sequence assembled by the public HGP.[119] In June 2000, Francis Collins, Director of the HGP, and Craig Venter joined President Bill Clinton at the White House to announce that a "first draft" of the human genome sequence had been completed.[120] President Clinton heralded the accomplishment as "an epoch-making triumph of science and reason," and both sides declared a major scientific victory.[121]

Despite the eventual détente between Celera and the HGP, the two sequencing efforts approached the release of their genomic data very differently. Unlike the public HGP, Celera initially deposited its data on its own commercial web site, rather than the public GenBank database. [122] The company allowed scientists from non-profit and academic institutions to access the data without charge but required that scientists who wished to use the data for commercial purposes enter into a license agreement.[123] This approach outraged much of the scientific community and led to a highly-publicized debate regarding public access to human sequence data. Prominent in this debate were contentions regarding the need to release data broadly and publicly in order to promote scientific advancement and medical breakthroughs, sentiments that Celera found hard to contest. Ultimately, in the settlement brokered by the journal *Science*,

---

[114] See note x, supra [describing seq centers].
[115] Nicholas Wade, *Scientist's Plan: Map All DNA Within 3 Years*, N.Y. TIMES, May 10, 1998, at 20.
[116] Craig Venter left the National Institute of Neurological Disorders and Stroke (NINDS) at NIH in 1992 to found The Institute for Genomic Research (TIGR). In 1998 he left TIGR to found Celera Genomics. *See, generally,* LARGE-SCALE SCIENCE, *supra* note x, at 38, and SHREEVE, *supra* note x, at 86, 117-18.
[117] SHREEVE, *supra* note x, at 22–23; Leslie Roberts, *Controversial from the Start*, 291 SCIENCE 1182, 1187; Wade, *supra* note 27, at 1.
[118] *See* Roberts, *supra* note 117, at 1188. Add cites to Shreve, Venter, Ridley.
[119] cite
[120] Roberts, *supra* note 117, at 1188; Nicholas Wade, *Genetic Code of Human Life is Cracked by Scientists: A Shared Success*, N.Y. TIMES, June 27, 2000, at A1.
[121] *Reading the Book of Life: White House Remarks on Decoding of Genome*, N.Y. TIMES, June 27, 2000, at F8.
[122] Despite Celera's intention to earn subscription fees from its genomic sequence data, Celera did not actively pursue patent protection for the data it generated. Rather, Celera protected its commercial position through a combination of contractual restrictions on users and limited access via its controlled web site.
[123] Eliot Marshall, *Storm Erupts over Terms for Publishing Celera's Sequence*, 290 SCIENCE 2042, 2042 (2000).

Celera agreed to make its data broadly available on its own corporate web site under a somewhat less restrictive licensing agreement.[124] The HGP draft sequence was published in GenBank in 2001,[125] and by 2003 most of the genes contained in Celera's database had also been resequenced and released publicly by the HGP. 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.[126]

V. THE ACTION ARENA: EVOLUTION OF PUBLIC DATA RELEASE RULES

Under the IAD framework, the "action arena" constitutes the set of scenarios in which the participants interact with respect to the common resource.[127] "Patterns of interaction" emerge from these exchanges, resulting in outcomes that in turn affect the characteristics of the community, the common resource and its rules-in-use. Madison, Frischmann and Strandburg equate these outcomes and patterns of interaction in the context of cultural commons, arguing that "[h]ow people interact with rules, resources, and each other .. is itself an outcome that is inextricably linked with the form and content of the knowledge or informational output of the commons."[128]

In the case of the genome commons, interactions occur at both scientific and policy levels. The vast majority of day-to-day scientific interactions – involving the generation and analysis of scientific data, the securing of funding for research projects, and the publication of results – occur relatively independently of the policy-level debates described above. Yet policy decisions fundamentally affect the manner in which the scientific enterprise is conducted. Data must be released to public databases on a frequent basis, these databases are consulted regularly both to supplement and validate collected data, and the preparation and submission of publications is constrained by the rules of the commons. During the conduct of this day-to-day scientific work, scientists and researchers accumulate experiences and preferences regarding the rules under which they must operate. They form opinions and draw conclusions regarding the difficulty of regularly depositing data into public databases, the ease with which this data may be used, the usefulness of public data, and the rate at which competing groups seem to be utilizing "their" data to compete with them. These opinions and conclusions manifest themselves in the next set of policy discussions regarding the next project to be proposed. Thus, as anticipated by Ostrom and Madison, Frischmann and Strandburg, a feedback loop develops, in which policy-level decisions affect interactions within the Action Arena and cause participants to seek policy-level changes in subsequent iterations of policy-making.

These patterns emerge in the successive genomics projects that followed the HGP.

A. DATA GENERATORS VERSUS DATA USERS

In their effort to promote the policy goals of project coordination, scientific advancement and minimizing encumbrances, the HGP organizers knowingly subrogated the interests of data generators to those of the public. That is, the rapid data release requirements of the Bermuda

[124] Eliot Marshall, *Sharing the Glory, Not the Credit*, 291 SCIENCE 1189-93 (2001).
[125] *See* HGP Initial Paper, *supra* note **Error! Bookmark not defined.**.
[126] Jocelyn Kaiser, *Celera to End Subscriptions and Give Data to Public GenBank*, 308 SCIENCE 775, 775 (2005).
[127] Ostrom & Hess, *supra* note x, at 53-59.
[128] Cultural Commons, *supra* note x, at 682.

Principles effectively eliminated the ability of data generators to publish analyses and conclusions based on their data before others could access it via public means.[129] The implications of this effect were not realized immediately, but in the years 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.[130] 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."[131] 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,"[132] 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.

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

1. Reaffirmation of Bermuda. In 2003, the Wellcome Trust sponsored a 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.[133] While the Ft. Lauderdale participants "enthusiastically reaffirmed" the 1996 Bermuda Principles, they also expressed reservations about extending these broad principles to every aspect of scientific research and discovery. Thus, they drew a distinction between ordinary "hypothesis-driven" scientific research, in which the investigators' primary goal is to solve a particular scientific question, and "community resource projects" (CRPs) that were "specifically devised and implemented to create a set of data, reagents or other material whose primary utility

[129] 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.").

[130] *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.

[131] *Id.*

[132] *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.

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

will be as a resource for the broad scientific community."[134] In hypothesis-driven research, success is typically 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. This risk, and the "legitimate interest" of data generating scientists to publish the results of their work in peer-reviewed journals, was explicitly recognized by NHGRI.[135] Accordingly, the Ft. Lauderdale participants concurred that while the twenty-four hour rapid-release rules of Bermuda would continue to apply to CRPs, there would be no requirement that the Bermuda Principles apply to scientific research other than CRPs.

2. The International HapMap Project, a New CRP. Beginning in 2002, an international group of scientists and funding agencies began a project to create a haplotype map of the human genome.[136] The data release policy of the HapMap Project is based on the Ft. Lauderdale principles and the project self-designates itself as a CRP.[137] Data generated by the project "[were] released rapidly into" publicly accessible databases,[138] but access was subject to the user's consent to the terms of a standardized, online click-wrap agreement.[139]

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.[140] 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

---

[134] *Id.* 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. *See, e.g.*, Kaye et al., *supra* note x. As envisioned by the Ft. Lauderdale participants, CRPs would include large-scale projects generating human and non-human sequence data, other basic genomic data maps, and other collections of complex biological data such as protein structures and gene expression information. Ft. Lauderdale Principles, supra note x, at 2, 5.

[135] *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].

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

[137] *Id.* at 793.

[138] *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.

[139] *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). A copy of the HapMap Project's click-wrap agreement is available at http://www.hapmap.org/cgi-perl/registration. REGISTRATION FOR ACCESS TO THE HAPMAP PROJECT GENOTYPE DATABASE, http://hapmap.ncbi.nlm.nih.gov/cgi-perl/registration [hereinafter HapMap Agreement]. Rebecca Eisenberg, who analogizes the HapMap Agreement to the open source software General Public License (GPL) raises questions about the enforceability of such agreements. Eisenberg, *supra* note **Error! Bookmark not defined.**, 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).

[140] *See* The International HapMap Consortium, *supra* note 136 at 793.

applications on the haplotypes or other scientific data generated by the project.[141] The HapMap Consortium's non-patenting requirement was viewed with admiration by many, including policy makers at NHGRI.[142]

As a corollary to the provisions of its click-wrap agreement, the HapMap Project adopted a "Data Release Policy," setting forth the drafters' somewhat conclusory position that raw SNP and haplotype data lack "specific utility" necessary for patent protection.[143] 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."[144]

3. 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.[145] NHGRI issued a data release policy for the ENCODE pilot project closely following the Ft. Lauderdale principles.[146] The NHGRI designated the project as a CRP.[147] As recommended in Ft. Lauderdale, users of the data were urged to cite the data generators in their analyses and were encouraged to consider research collaborations with them.[148]

With respect to intellectual property issues, the agency first acknowledges the requirements of the Bayh-Dole Act; 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.[149] 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".[150] 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 . . . ."[151] 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."[152] This

[141] HapMap Agreement, *supra* note 139.
[142] *See ENCODE Project Data Release Policy (2003-2007)*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/12513440 (last Reviewed Oct. 18, 2010) [hereinafter ENCODE 2003 Pilot Policy]. (referring to the HapMap Project's successful policy of discouraging "parasitic patents").
[143] Int'l HapMap Project, *Data Release Policy*, http://www.hapmap.org/datareleasepolicy.html).
[144] *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 x, at 1026.
[145] 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).
[146] *See* ENCODE 2003 Pilot Policy, *supra* note 142.
[147] *Id.*
[148] *Id.*
[149] *Id.*
[150] *Id.*
[151] *Id.*
[152] *Id.*

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.

C. 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. The effect of these private sector initiatives is important, as they both reacted to, and were closely observed by, the publicly-funded projects that continued to operate alongside them.

1. The Merck Gene Index. Beginning in 1994, pharmaceutical giant Merck, collaborating with Lawrence Livermore National Laboratory and Washington University in St. Louis, began to assemble a database of expressed sequence tags (ESTs) known as the "Merck Gene Index," which it intended to release to the public.[153] By 1998, the Merck Gene Index had released over 800,000 ESTs through GenBank.[154] Merck's stated rationale for publicly releasing this potentially valuable data was the expansion of basic knowledge in the interest of combating disease.[155] While this goal is laudable, it is believed 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.[156]

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 group of ten pharmaceutical companies[157] and the Wellcome Trust to identify and map genetic markers referred to as "single nucleotide polymorphisms"

[153] *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), *available at* http://www.bio.net/bionet/mm/bionews/1995-February/001794.html [hereinafter Merck Gene Index Press Release]; *See also supra* notes x and accompanying text (discussing ESTs and the patenting debate surrounding them).
[154] DON TAPSCOTT & ANTHONY D. WILLIAMS, WIKINOMICS: HOW MASS COLLABORATION CHANGES EVERYTHING 166 (2006).
[155] Merck Gene Index Press Release, *supra* note 153.
[156] Marshall, *supra* note 90. 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 154; Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 134 (1999–2000).
[157] 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. (Author's personal files). For an informative perspective on the interest that information technology providers such as IBM took in the emerging field of genomics, *see* cite IBM article.

(SNPs) and to release the resulting data to the public domain.[158] SNP data were publicly released on the consortium's web site on a quarterly, then on a monthly, basis during the two-year research program, and also deposited in GenBank.[159] 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.[160] The SNP Consortium wished to generate data for the use of all researchers, unencumbered by patents.[161] It accomplished this goal by filing U.S. patent applications covering SNPs that it discovered, and then contributing these applications to the public domain prior to issuance.[162] 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.[163]

D. 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.

---

[158] 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).

[159] *See* Holden, *supra* note 158, at 25–26 and U.S. Dept. of Energy – Office of Science, Human Genome Project Information – SNP Fact Sheet, *available at* http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml#when) (last visited July 22, 2009). The SNP Consortium's data is currently hosted on the International HapMap Project's web site.

[160] Holden, *supra* note 158, 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).

[161] *See, e.g.*, Holden, *supra* note 158, 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 **154**, at 168 (noting consortium members' concerns about biotech companies' plans to patent SNPs and "sell them to the highest bidder").

[162] 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).

[163] *See, e.g.*, Marshall, *supra* note 90, at 1192 (noting the consortium's "defensive move" deriving from the Merck Gene Index's earlier strategy); Cook-Deegan & McCormack, *supra* note 100 (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).

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.[164] GAIN's purpose was to conduct GWA studies of the genetic basis for six common diseases.[165] 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.[166] Over the course of the three-year project, approximately 18,000 human DNA samples were genotyped.[167] The resulting data was deposited in dbGaP. As described above, dbGaP allows the data producer to segregate the data into open and controlled access portions. Researchers wishing to access GAIN data from the controlled portion of the database must register with, and be approved by, the GAIN Data Access Committee (DAC)[168] and 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 individual study participants.[169]

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 making presentations based on GAIN data for a specified embargo period.[170] 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.[171]

2. The Cancer Genome Atlas (TCGA). In 2006, NCI and NHGRI launched a pilot project to catalog genomic changes relating to cancer.[172] The Cancer Genome Atlas (TCGA) project generates genomic sequence and related data, but also keeps track of a large amount of clinical data, including patient diagnosis, treatment history and ongoing status.[173] Due to the

---

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

[165] 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]*.

[166] The GAIN Collaborative Research Group, *supra* note 164, at 1048 (Box 1).

[167] 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).

[168] Gain Collaborative Research Group, *supra* note 164, at 1049.

[169] *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*].

[170] GAIN Collaborative Research Group, *supra* note 164, at 1049.

[171] *Id.*

[172] *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.*

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

specialized nature of the project data, deposits are made both in dbGaP as well as a TCGA-specific database administered by NCI.[174]

Given the potential for identifying individual patients from their genomic and phenotypic data, great attention was paid to controlling access to TCGA data.[175] Like GAIN data, TCGA data is available in an open-access tier and a controlled-access tier.[176] 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.[177] 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.[178] 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,[179] 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.[180] The NIH GWAS Policy requires that grantees submit descriptive information about each GWA study for inclusion in the "open access" portion of dbGaP.[181] Grantees are also "strongly encouraged" to submit study results, including phenotypic,

---

[174] *Data Use Certification 1* THE CANCER GENOME ATLAS PILOT PROJECT (Feb. 22, 2010) http://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=DUC&view_pdf&stacc=phs000178.v1.p1 [hereinafter *TCGA Data Use Certification*].

[175] 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 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).

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

[177] *Id.*

[178] *Compare TCGA Data Use Certification*, *supra* note 174 *with GAIN Data Use Agreement*, *supra* note 164.

[179] *TCGA Data Use Certification*, *supra* note 174, at 7.

[180] 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*].

[181][181] Descriptive information includes the study protocol, questionnaires, manuals, variables measured and other supporting documentation. NIH GWAS Policy, *supra* note 180, 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 174. 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.*

exposure and genotypic data, for inclusion in the "controlled access" portion of the database "as soon as quality control procedures have been completed."[182]

Among the principal concerns raised concerning GWA study data were those surrounding the public release of phenotypic or clinical information that could eventually be traced back to individual subjects.[183] To address this concern, the NIH GWAS Policy requires that GWAS data be de-identified in accordance with HIPAA guidelines.[184] 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,[185] and then only under a signed Data Use Certification that contains stringent protective clauses.[186]

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.[187] 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."[188] 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."[189] However, in an effort to show some support for patent seekers, the GWAS Policy also "encourages patenting of technology

[182] *NIH GWAS Policy*, *supra* note 180, 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.

[183] *NIH GWAS Policy*, *supra* note 180, 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.*

[184] *NIH GWAS Policy*, *supra* note 180, at 49295 (citing the HIPAA Privacy Rule, 45 CFR 164.514(b)(2)).

[185] The DAC is comprised primarily of NIH staff with expertise in the relevant scientific disciplines, data privacy and data subject protection. *NIH GWAS Policy*, *supra* note 180, at 49296.

[186] 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 180, at 49296.

[187] 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. However, negotiation among competing stakeholder groups eventually led to the imposition of the 12-month exclusivity period, a period has been criticized as potentially being too *short*. Michael Krawczak, et al., *Is the NIH Policy for Sharing GWAS Data Running the Risk of Being Counterproductive?* at 4, INVESTIGATIVE GENETICS, Sept. 1, 2010.

[188] *NIH GWAS Policy*, *supra* note 180, at 49296.

[189] *Id*. at 49297

suitable for subsequent private investment that may lead to the development of products that address public needs."[190]

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 SAE Consortium (SAEC), a group of pharmaceutical companies formed in 2007 to fund research toward the identification of DNA markers for drug-induced serious adverse events.[191] The Consortium works with academic collaborators to collect DNA samples and associated phenotypic data, and then to conduct GWAS, targeted sequencing and statistical analyses to identify potential markers and associations of interest.[192] Since its formation, SAEC studies have identified DNA markers relating to drug-induced liver injury (DILI)[193] and serious skin rash (SSR). The SAEC 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.[194] Like the GAIN and other policies discussed in this section, the SAEC 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.[195] During this time they have the ability to analyze data and prepare papers for publication without the threat being scooped by competing groups. While the research funded by SAEC 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),[196] 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 model and the community of data users and funders.[197]

5. The Full ENCODE Project and modENCODE. In 2007 NHGRI expanded the ENCODE pilot project[198] to cover the entire human genome and launched a corollary project (modENCODE) to identify the functional genomic elements of two common model organisms,

---

[190] *Id.* at 49296.
[191] *iSAEC's Background and Organizational Structure* INT'L SAE CONSORTIUM http://www.saeconsortium.org/ (last accessed Oct. 28, 2010).
[192] *Id.*
[193] *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).
[194] 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).
[195] Int'l SAE Consortium Ltd., DATA RELEASE AND INTELLECTUAL PROPERTY POLICY (last amended Nov. 5, 2009) (on file with author). *[has been provided by author]*
[196] *See supra* Section III.C.1.
[197] The compromises and negotiation strategy inherent in this approach is discussed in greater detail in Contreras, *Data Sharing*, *supra* note 3, at 11.
[198] *See supra* Section III.C.5.

the roundworm (*c. elegans*) and fruit fly (*drosophila melanogaster*).[199] This expansion entailed 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.[200] 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.[201]

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.[202] The length and complexity of the policy evidences the agency's and the participants' desire for clear guidelines and the avoidance of misunderstandings regarding the release of data, as the diversity of participants, organisms and data types has expanded dramatically beyond those originally considered by the framers of the Bermuda Principles.

E. RECENT DEVELOPMENTS.

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 National Cancer Institute convened a meeting of proteomics[203] researchers in Amsterdam to "identify and address potential roadblocks to rapid and open access to [proteomics] data."[204]

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.[205] 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

---

[199] *See* Susan E. Celniker et al., *Unlocking the Secrets of the Genome*, 459 NATURE 927 (2009) (describing the modENCODE project methodology and goals).
[200] ENCODE Consortia, DATA RELEASE, DATA USE, AND PUBLICATION POLICIES (2008), *available at* http://www.genome.gov/Pages/Research/ENCODE/ENCODEDataReleasePolicyFinal2008.pdf [hereinafter "ENCODE 2008 Policy"].
[201] *Id.* at 4.
[202] *Id.* at 5-7.
[203] 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 **Error! Bookmark not defined.**, at 312-22.
[204] 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).
[205] *See Id.* at 3689-90.

proteomic data, and existing proteomic databases suffer from inconsistent and sometimes incompatible data formats.[206] Finally, unlike genomics, in which the entire field focused for several years on the single HGP project, proteomics research lacks a unifying policy core and proteomics-focused journals have each developed their own, sometimes inconsistent, guidelines for data submission.[207]

Notwithstanding these potential 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 should in no event be later than publication or, for community resource projects, following appropriate quality assurance procedures), (2) Comprehensiveness (full raw data sets should be released together with associated metadata and quality data), (3) Format (standardized formats are encouraged), (4) Deposition to repositories (central repositories for proteomic data should be established), (5) Quality metrics (central repositories should develop metrics for assessing data quality), and (6) Responsibility (scientists, funding agencies and journals share responsibility for ensuring adherence to community data release standards).[208]

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.[209] The participants reaffirmed a general community commitment to rapid pre-publication data release, expanding the scope of projects as to which these principles should apply to all biomedical datasets having "broad utility, are large in scale … and are 'reference' in character".[210] Specifically, they cited, in addition to genomic and proteomic studies, structural chemistry, metabolomics and RNAi datasets as well as annotated clinical resources such as cohorts, tissue banks and case-control studies.[211]

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 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 (evoking the CRP designation developed in Ft. Lauderdale), rapid data release "should not be mandated" for projects that are generally hypothesis-driven.[212] The Toronto participants also addressed the priority concerns of data generators versus data users, observing anecdotally that in many cases data users have, in fact, published papers based on publicly-released data sets *before* the publication of the data generators' papers analyzing the data sets themselves, and that this

[206] *Id.* at 3690. Existing proteomic databases include GPMDB, UniProtKB, Peptide Atlas, PRIDE and NCBI's Peptidome. *Id.*
[207] *Id.*
[208] *Id.* at 3690-91.
[209] *See* Toronto Report, *supra* note 76.
[210] *Id.* at 168. To some degree, this characterization is a restatement of the Ft. Lauderdale definition of "community resource projects".
[211] *Id.*
[212] Toronto Report, *supra* note 76, at 169.

situation caused no "serious damage" to the data generators' subsequent publications.[213] 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.[214] 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.[215]

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 date generators' ability to secure patent protection, with concomitant implications for U.S. funding agencies subject to Bayh-Dole requirements.[216] 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 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, [217] the International Cancer Genome Consortium[218] and the Human Microbiome Project.[219] NIH and NHGRI are in the process of considering yet further revisions to their institutional data release policies and collecting feedback from various stakeholder groups.[220] 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.

VI. TRENDS AND EVALUATION

A. GENOME COMMONS DESIGN TRENDS AND INFLUENCES

The genome commons, which originated with the HGP and the sweeping Bermuda Principles, has experienced rapid and unanticipated growth over the past decade. But this growth

[213] *Id.* at 169–70.
[214] *Id.* at 170.
[215] *Id.*
[216] Author's personal notes, The Toronto Data Release Workshop (May 13-14, 2009) (on file with author)..
[217] *See 1000 Genomes Data and Sample Information*, 1000 GENOMES, http://www.1000genomes.org/page.php?page=data (last visited Oct. 28, 2010).
[218] *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.
[219] *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).
[220] 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.

has not been without controversy, and each iteration of the policies governing the commons has increased in detail and complexity. The reasons for this increasing complexity 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 sequences) generated by a specific, unique project (the HGP). 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. [221] Similar characteristics applied to data generated by the HapMap Project[222] and the SNP Consortium.[223] These data were intended to elucidate non-individualized information applicable to the human genome generally. Accordingly, in these early projects concerns regarding the identifiability of human subjects from data released to the public, while addressed, were not paramount.

In later projects, and particularly with the commencement of large-scale GWA studies, concerns with the potential identification of human subjects grew.[224] 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.[225]

2. *Scientific Advancement and Publication Priority*. As discussed above, many policy makers believed that 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

[221] *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).
[222] *See What is the HapMap*, INTERNATIONAL HAPMAP PROJECT, http://hapmap.ncbi.nlm.nih.gov/thehapmap.html.en (last visited Oct. 28, 2010).
[223] Holden, *supra* note 158.
[224] *See* Toronto Report, *supra* note 76, at 170.
[225] 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 83, at ch. 13; Crolla, *supra* note 83, at 241–47.

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 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?[226] Even Merton, who championed the norm of scientific communalism, did not specify how *quickly* the sharing of data should occur.

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 immediate access to data and let them perform research, but prohibit them from making related presentations or submitting related papers during the embargo period.[227] The approach taken by private consortia, in contrast, protects data generator priority by allowing data generators to retain data privately for a specified period, but then requires the release of this data to the public without restriction. The patent-related trade-offs between these differing approaches is discussed below.[228]

3. *Patent Encumbrances*. Patent protection is related to, but distinct from, the issue of publication priority. As discussed above, early in the HGP, following the EST patenting debate, NIH representatives adopted a position that patent protection is inappropriate for DNA sequence information. This stance, also held by leaders of the scientific community and international funding agencies, is reflected in the Bermuda Principles. 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 the patenting of results generated by their research.

NHGRI, however, must operate within the constraints of the Bayh-Dole Act. Thus, while NHGRI's various post-Bermuda data release policies all acknowledge the requirements of the Act, they demonstrate a general bias against the placement of patent encumbrances on genomic data.[229] The enforceability, however, of policy provisions that merely "urge" or "encourage" data

[226] *See* Eisenberg, *Patents and Data-Sharing*, *supra* note x, 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 . . . .").
[227] *Supra* Table 1.
[228] For a detailed analysis of the use of time-based "latency" approaches to achieving compromise in the structuring of commons of scientific information, *see* Jorge L. Contreras, *Data Sharing, Latency Variables, and Science Commons*, 25 BERKELEY TECH. L.J. 1601 (2010).
[229] *See ENCODE Pilot Policy*, *supra* note 142; *NHGRI 1996 Policy*, *supra* note 110 and accompanying text; *NIH GWAS Policy*, *supra* note 180.

generators and users not to seek patents on inappropriate subject matter is open to some doubt.[230] 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 to achieve a similar result. The Bermuda Principles, in particular, and their adoption by NHGRI in 1997 and reaffirmation in 2003, 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 "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.[231]

When publication priority issues began to emerge with the movement toward GWAS and other studies involving phenotypic data components, the publication embargo was offered by NIH 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 rapid pre-publication data release.

B. THE GENOME COMMONS AS A CULTURAL COMMONS

Madison, Frischmann and Strandburg offer their modified IAD framework in order to encourage the broad analysis of resource commons in the cultural environment and to counter the prevailing functionalist account of cultural production. In particular, they challenge the notion that the majority of cultural production can be explained in terms of incentive/exclusion-based intellectual property rules or governmental subsidy. To this end, they claim that "[i]nnovation and creativity are matters of governance of a highly social cultural environment."[232]

Scientific research has not typically been viewed as a form of cultural production. In fact, Madison, Frischmann and Strandburg point to scientific research as an area potentially reinforcing the traditional functionalist view of "IP rights and government subsidies".[233] But in this regard, their view of scientific research may be too narrow. The enterprise of science is characterized by a pervasive and complex set of norms that govern both the incentives and behaviors of its participants.[234] My analysis of the genome commons supports this view.

From the early days of the HGP, NIH policy makers and scientific leaders expressed a strong aversion to the encumbrance of genomic information, either through patent protection (as evidenced by the EST patenting debate) or database access restrictions (as evidenced by the HGP's competition with Celera Genomics). While the HGP and subsequent public genomics projects were funded, in large part, by federal grants, private efforts such as the SNP Consortium and the SAE Consortium exhibited similar values. This level of consistency suggests that neither the traditional account of economic property-based incentives or government subsidies fully explains the dynamics observed in the genome commons.

---

[230] *See* Rai & Eisenberg, *supra* note 109, at 309.
[231] 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, *Patents and Data-Sharing*, *supra* note x, at 1026.
[232] Cultural Commons, *supra* note x, at 669.
[233] *Id.* at 665, 666 ("[w]e suspect that over time the constructed cultural commons framework will yield a far larger and richer set of commons cases in the cultural context than one might discover by focusing only on patent law or scientific research of software development.")
[234] *See, e.g.*, Merton, *supra* note x, and Rai, *supra* note x.

In the years following the completion of the HGP, genomic data release policies became more complex and, to a degree, more restrictive. However, these restrictions arose not from efforts to impose traditional intellectual property restrictions on the fruits of genomic research, but from competition among scientific groups to achieve publication priority from their data. This critical aspect of the scientific enterprise, which is abundantly covered in the sociology of science literature,[235] has not generally been given much weight in the economics-oriented discussion of commons formation.[236] In this sense, the genome commons is a cultural commons of the kind sought by Madison, Frischmann and Strandburg, one in which innovation and creativity arise in "a highly social cultural environment",[237] and which can be counted among other examples of non-property-focused institutions for the generation of valuable intellectual assets.

[235] Cite Merton and followers.
[236] *But see* Jonathan M Barnett, *The Illusion of the Commons*, 25 Berkeley Tech. L.J. 1751 (2010) (arguing that social norms play a significant role in commons formation in many areas) and Robert P. Merges, *Property Rights Theory and the Commons: The Case of Scientific Research*, in SCIENTIFIC INNOVATION, PHILOSOPHY, AND PUBLIC POLICY (Ellen Frankel Paul, et al., eds. 1996) (recognizing that social norms affect scientific behavior, even in the presence of strong intellectual property incentives).
[237] Cultural Commons, *supra* note x, at 669.