CONSTRUCTING THE GENOME COMMONS

[forthcoming in CONVENING CULTURAL COMMONS (Oxford, 2013)]

Working Draft: Aug. 15, 2012

Jorge L. Contreras[*]

ABSTRACT

Basic scientific research is often viewed as a public good: a non-depletable, non-rival resource that exists in the public domain. The vast collection of genomic data generated since the Human Genome Project, however, belies this description. This valuable resource, though generally accessible to researchers worldwide, is governed by a complex set of rules that have evolved over the past two decades. As such, the "genome commons" more closely resembles a common pool resource described by Elinor Ostrom than a public good. In this chapter, I apply Ostrom's Institutional Analysis and Development (IAD) framework, as modified by Madison, Strandburg and Frischmann, to the genome commons and elucidate the stakeholder interests and negotiations that led to the rules-in-use, both formal and norms-based, that govern this global scientific resource. I conclude that a public goods approach to the genome commons, as has been suggested in some contexts, is overly simplistic and, if pursued, could lead to lessening participation in the creation of this valuable public resource.

I. INTRODUCTION

In his contribution to this volume, Yochai Benckler explores the tension between two competing conceptions of the "commons": that which is generally espoused by the academic legal literature and focuses on regimes of open access and limited propertization, and that which is addressed in the literature of common pool resources pioneered by Elinor Ostrom, which focuses on social structures for the management of shared resources.[1] In economic terms, basic scientific research is typically classified as a "public good", a resource provisioned by the state that is susceptible to neither exclusion nor depletion by use (i.e., the economic characteristics of non-excludability and non-rivalry). Common examples of economic public goods include lighthouses and public highways. Though the analogy is imperfect, in legal terms public goods are often associated with the public domain, a category that implies free access and a lack of ownership by any particular entity. The public domain is often evoked in the context of intangible ideas and scientific discoveries that are free from intellectual property encumbrances, either because the relevant rights have expired or were never procured in the first place. Examples include the works of Shakespeare (in which copyright has expired) and discoveries that have been published but as to which patent applications have not been filed. As such, scientific research (to the extent not protected by intellectual property) generally falls into Benckler's first broad classification of the commons: that which is generally associated with the public domain.

[*] American University – Washington College of Law.

[1] Benckler 2013, p.x.

In at least one significant case, however, basic scientific research has shown itself to be more amenable to analysis as a common pool resource along the lines developed by Ostrom and her colleagues. As the title of this chapter suggests, the case to which I am referring is genomic research: the large-scale study of the genetic make-up of humans and other organisms that began with the Human Genome Project (HGP) in the late 1980s and continues today in numerous government and privately-funded projects. These projects have made a vast quantity of genetic information available in public databases across the globe. This massive accumulation of data is what I refer to collectively as the "genome commons."[2]

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 Genomics, which competed with the public HGP to complete the human genomic sequence from 1998 to 2001.

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.[3] These groundbreaking "Bermuda Principles" required that all DNA sequence data generated by the HGP be released to the public just twenty-four hours after generation, a stark contrast to the months or years that usually preceded the release of scientific data. The Bermuda Principles arose from an early recognition by scientists and policy makers that rapid and efficient sharing of data was necessary to coordinate activity among the geographically dispersed laboratories. But project coordination was not the only factor motivating the unorthodox rapid-release requirement of the Bermuda Principles. More importantly, this approach arose from the conviction among project leaders that rapid release of genomic data was necessary for the advancement of scientific research.[4]

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.[5] Advances in science and technology, however, together with increasingly challenging 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 large data sets to analyze and publish their research before others. The emergence and recognition of these considerations has led to an evolution of genomics data release policies and norms that are more restrictive and complex than those of the HGP, but which nevertheless preserve the fundamental shared nature of the genome commons. In this respect, the genome commons more resembles the

[2] Contreras, 2010a, 2010b, 2011
[3] *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].
[4] *See, e.g.*, HGP Initial Paper, *supra* note x, at 864
[5] Kaye, 2009

common pool resources studied by Ostrom and her colleagues than the simpler models of the public domain/public goods that are typically associated with basic scientific research.

Ostrom pioneered the analysis of common resource structures, whether physical or informational, using a conceptual tool known as the Institutional Analysis and Development (IAD) framework. More recently, Michael Madison, Brett Frischmann and Katherine Strandburg have undertaken a thorough re-examination of the IAD framework in relation to commons in the "cultural environment,"[6] seeking to combine the functionalist IAD approach with metaphorical and narrative accounts of commons formation.[7] In this chapter, I engage the theoretical framework of Ostrom and Madison, Frischmann and Strandburg to elucidate both the structural and narrative elements of the genome commons. The IAD methodology offers a systematic means for examining the characteristics of a commons structure: those of the common resource, the "action arena" in which stakeholders interact with the commons, and the resulting patterns of interaction.[8] Each of these broad areas is subdivided into further analytical components, so that the common resource, for example, is assessed with respect to its basic characteristics, the attributes of the relevant community, and its applicable "rules in use." In particular, I chart the evolution of the genome commons from what was initially a public domain vehicle established to deter the proprietary designs of emerging biotechnology companies, into a unique polycentric governance institution for the growth, management and stewardship of a massively-shared public resource.

II. ATTRIBUTES OF THE GENOME COMMONS

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 resource it is useful to consider both the data contained within it and the databases that house it, as well as the underlying legal environment that applies to such aggregations of data.

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

[6] *Cultural Commons*, *supra* note **Erreur ! Signet non défini.**, at 659.
[7] *Id.* at 671–74, 681–83.
[8] *See* Elinor Ostrom & Charlotte Hess, *A Framework for Analyzing the Knowledge Commons*, *in* KNOWLEDGE AS A COMMONS, *supra* note **Erreur ! Signet non défini.**, at 44–45.

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

Some segments of DNA within an organism's cells form functional units called "genes", ranging in size from as few as a hundred to more than two million base pairs. It is currently estimated that each human possesses between 20,000 and 25,000 genes. Genes 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 genomic science has been to associate particular genes or genetic variations (mutations) with phenotypic traits.

Early in the twentieth century, hereditary diseases began to be associated with genes passed from parents to their offspring. But while simple inheritance explained numerous conditions, from benign traits such as hair coloration to debilitating ailments such as cystic fibrosis, Down syndrome and Huntington's disease, it was not until the 1970s that scientists could identify the individual genes responsible for these conditions. Even then, each of these discoveries took years of painstaking research and a healthy dosage of good luck. But in 1986 a revolutionary new process for copying DNA fragments, the polymerase chain reaction (PCR), was developed and enabled the large-scale, rapid sequencing of DNA. PCR technology soon gave rise to ambitious plans to sequence not only genes identified with specific diseases, but the entire human genome.

The HGP, which is discussed in detail below, took a decade to complete, and resulted in the first detailed map of the 3.2 billion base pairs that constitute the human genome. Since the completion of the initial draft of the human genome 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[9] and to sequence the genomes of the multitude of microorganisms residing within the human body.[10]

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 conditions, recent "genome-wide association studies" (GWAS) have been credited with identifying variants in multiple genes that increase susceptibility for complex conditions such as diabetes, cancer, hypertension and numerous other diseases. Such studies, which involve scanning the entire human genome for variants that are common among affected individuals, have been made possible by dramatic advances in sequencing and data analysis technology.

2. *Data and Databases*. For hundreds of years, the traditional means of disseminating scientific information has been the peer-reviewed scholarly journal. Scientists are

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

judged, both for purposes of career advancement and the awarding of government grants, on the quantity of their publications, giving scientists a significant personal incentive to publish and share their data with others.[11] Yet, despite the prevalence of scientific publications, there are two principal reasons that journal publication has proven to be 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. Despite its dependence on biological systems, Hess and Ostrom accurately observe that modern biology has been transformed into an "*information* science."[12] 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 printed pages, roughly equivalent to 200 New York yellow pages directories.[13] Accordingly, a journal article typically includes only a brief presentation of significant experimental findings, often made in summary or tabular fashion, together with the researcher's analysis and conclusions. While the published data are often essential to support the analysis, it typically represents 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.

Second, there is usually a lengthy delay between the collection of data 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 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. Studies report that the average delay between the completion of scientific work and publication can range from twelve to eighteen months and longer, depending on the field.[14] Clearly, in a field in which rapid access to experimental data is required 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. Many science funding agencies now require that genomic data be released into public databases shortly after it is generated. A growing number of scientific journals also require that authors make the data underlying their published results available to readers on a web site accessible through the journal, through their own institutions or in a government-maintained database. These databases have enabled the efficient, rapid and cost-

[11] Robert K. Merton, *Priorities in Scientific Discovery* (1957), *reprinted in* THE SOCIOLOGY OF SCIENCE 286, 316.
[12] Charlotte Hess & Elinor Ostrom, *A Framework for Analysing the Microbiological Commons*, 58 INTL. SOC. SCI. J. 335, 335 (2006) [hereinafter Hess & Ostrom, *Framework*].
[13] 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).
[14] Carlos B. Amat, *Editorial and Publication Delay of Papers Submitted to 14 Selected Food Research Journals. Influence of Online Posting*, 74 SCIENTOMETRICS 379 (2008). William D. Garvey & Belver C. Griffith, *Scientific Information Exchange in Psychology*, 146 SCIENCE 1655, 1656 (1964); Charles G. Roland & Richard A. Kirkpatrick, *Time Lapse Between Hypothesis and Publication in the Medical Sciences*, 292 NEW ENG. J. MED. 1273, 1274 (1975).

effective sharing of new knowledge and the pursuit of studies and analyses that otherwise might have been impossible.

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

A final important observation regarding the nature of the genome commons is its sheer size and the breathtaking rate at which it is expanding. Over its decade-long existence, the HGP mapped the 3.2 billion base pairs comprising the human genome. To do so, it sequenced tens of billions of DNA bases (gigabases), creating what was then an unprecedented accumulation of genomic data. By way of comparison, the current 1000 Genomes Project is projected to generate 200,000 gigabases of data – approximately 20,000 times the quantity generated by the HGP.[15] In 2010, the Cancer Genome Atlas Project (TCGA) generated data at a rate of 7,300 gigabases per month.[16] Together with the Human Microbiome Project and the 1000 Genomes Project, the total data generation rate from major NIH-funded genome projects in 2010 was nearly 10,000 gigabases per month.[17] According to one 2011 report, "a single DNA sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project".[18] Statistics like these have led to talk of a "data tsunami" in genomic science, in which the capacity to manage and analyze these vast quantities of data will severely lag the rate at which such data is being produced.[19] Given these challenges, researchers have been experimenting with new approaches to the storage and handling of these large data sets. Data from the 1000 Genomes Project, for example, is currently being made available through the "cloud" via Amazon Web Services.[20]

The organizational implications of the rapid growth of the genome commons are significant. The institutional rules and structures surrounding genomic data were created when

[15] http://www.nih.gov/news/health/mar2012/nhgri-29.htm
[16] http://cancergenome.nih.gov/researchhighlights/leadershipupdate/ozenberger
[17] http://cancergenome.nih.gov/researchhighlights/leadershipupdate/ozenberger
[18] http://www.sciencemag.org/content/331/6018/666.full
[19] Green, NATURE 2011 (Feb 10, 2011) at 207-08, and Royal Society, Science as an Open Enterprise (2012) at 88
[20] Emily Waltz, 1000 Genomes on Amazon's Cloud, 30 Nature Biotech 376 (May 2012).

the size of this common resource was several orders of magnitude smaller than it is today. It is not surprising that rules established in the days of the HGP did not contemplate many of the complexities associated with today's genome commons. It is as though rules established among the early American colonists, numbering in the mere tens of thousands, would be expected to govern today's American nation of more than 300 million (particularly if this spectacular growth had occurred over the course of just two decades). In the literature of common pool resources, it is doubtful that any shared resource has grown at a rate anywhere close to this. In this light, the rules and norms established by the HGP, which continue to shape policy today, have fared remarkably well.

3. *Legal Background Environment*. Madison, Frischmann and Strandburg emphasize that an understanding of the "natural" environment in which a cultural commons exists is critical to understanding the attributes and operation of that commons.[21] 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 presents a complex picture, as it embodies both biomedical discoveries, which are typically addressed via the patent system, as well as large aggregations of data, which are typically addressed via access restrictions, contractual obligations 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. Excluded from patentable subject matter, however, are laws of nature and natural phenomena.[22] 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 1980s, shortly after large-scale DNA sequencing became feasible. The U.S. National Institutes of Health (NIH) was among the first to seek patent protection for DNA sequences. In 1991, a group led by NIH researcher J. Craig Venter filed patent applications claiming 337 short genetic sequences known as expressed sequence tags (ESTs), accompanied by an announcement that NIH would seek to patent thousands more ESTs in the coming months.[23] There was a public outcry in response to this announcement, triggering what Robert Cook-Deegan has called "an international firestorm."[24] The debate over gene patenting within NIH was equally vehement and marked a turning point in NIH's attitude toward patents on genetic material. By 1994, the agency elected not to appeal the Patent and Trademark Office's rejection of its initial EST patent applications,[25] and has since adopted a consistently lukewarm, if not outright averse, attitude toward the patenting of genetic sequences. This attitude is clearly reflected in the agency's support for the patent-deterring Bermuda Principles and subsequent policies.

[21] *Cultural Commons* at 684-88.
[22] *Diamond v. Diehr*, 450 U.S. 175, 185 (1981)
[23] Christopher Anderson, *US Patent Application Stirs Up Gene Hunters*, 353 NATURE 485, 485 (1991) and Leslie Roberts, *Genome Patent Fight Erupts*, 254 SCIENCE 184, 184 (1991).
[24] *See* COOK-DEEGAN, at 330-31.
[25] *See* LARGE-SCALE SCIENCE, *supra* note x, at 36-37.

Nevertheless, the patenting of genetic information by academic research institutions and private enterprises is the subject of continuing controversy. According to many sources, the number of such patents has continued to rise.[26] In *Prometheus Laboratories, Inc. v. Mayo Collaborative Services*, the U.S. Supreme Court recently cast doubt on patents that claimed so-called "laws of nature",[27] and in *Ass'n for Molecular Pathology v. US Patent and Trademark Office*, the Court of Appeals for the Federal Circuit [*describe new decision when available*].[28] Elsewhere, I have analyzed the specific patent deterrent effects of the genomic data release policies adopted by NIH during and after the HGP.[29] While a full discussion of this topic is beyond the scope of this chapter, suffice it to say that policies requiring rapid release of genomic data, together with the evolving understanding of the "utility requirement" under U.S. patent law, are likely to have had a substantial dampening effect on the issuance of patents covering "raw" DNA sequence information, though not on patents claiming specific functions of DNA sequences or the use of particular DNA sequences in diagnostic tests or as the basis for treatment decisions. It is these later categories that have, in recent years, constituted the bulk of so-called "gene patents" and which have come under increasing fire in the courts. Thus, while the raw sequence data contained in public repositories such as Genbank is not itself generally subject to patent protection (save for a dwindling number of early "composition of matter" patents, such as the ones litigated in the *Myriad* case), the practical uses of such sequence data might be constrained by patents covering specific diagnostic or therapeutic uses of that data. As noted above, the debate regarding the proper legal scope of such restrictions continues.

2. *Protection of Data and Databases*. Under U.S. law it has long been held that "facts" such as scientific data are not subject to copyright protection, and databases that merely contain compilations of factual information similarly lack formal legal protection.[30] Nevertheless, access to data that is contained in electronic databases can be controlled by the database operator via technical means such as password-restricted access as well as limitations built into contractual access agreements. Thus, while data itself may not be subject to legal protection, circumventing such technical protection or contractual measures can be prosecuted under a number of legal theories. In this way, scientific information that might otherwise be in the public domain can become encumbered when compiled in proprietary databases.[31] Such restrictions were initially adopted by Celera Genomics when it announced its intention to sequence the human genome in competition with the publicly-funded HGP and offer the resulting data to commercial users pursuant to license agreements. The threat of propertization of the genome in this manner has fueled continuing public support for GenBank, dbGaP and other publicly-accessible repositories for genomic data.

B. ACTORS AND STAKEHOLDERS

Much early work regarding common resource governance was devoted to understanding the attributes of the communities that shared the commons, whether herdsmen grazing cattle on a common pasture or fishermen trolling ocean stocks. This analysis is equally important in the

[26] Kyle Jensen & Fiona Murray, *Intellectual Property Landscape of the Human Genome*, 310 Science 239 (2005). Isabelle Huys, *et al.*, *Legal Uncertainty in the Area of Genetic Diagnostic Testing*, 27 NATURE BIOTECH. 903 (2010)
[27] S.Ct. cite
[28] Fed. Cir. cite.
[29] Contreras – Two Narratives.
[30] Feist v Rural Tel.
[31] See Reichman & Uhlir.

context of the genome 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, while also seeking to represent the interests of the general public. Thus, the role and influence of other stakeholder groups is not to be underestimated: the development of data release policies in the genome sciences has been a polycentric 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 approximately $2.7 billion (in 1991 dollars),[32] has been called "the largest and most visible large-scale science project in biology to date."[33] As such, NIH and the U.S. 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 significant influence over the project's technical and policy direction. Many of the scientists involved in the early planning and execution stages of the project were globally prominent and included numerous Nobel Prize winners. 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.[34]

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.[35] 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).[36] 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 data rather than the development and testing of hypotheses. This distinction persists today as the number of data-generating projects in the biosciences continues to increase. 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 production projects continues, data generating scientists face challenges when it comes to publishing their work in traditional scientific journals, as the creation of large data sets has not traditionally been viewed as meriting recognition in the most prestigious journals.[37]

Another trend having a significant impact on the generation of genomic data is the continuing decline in the cost of DNA sequencing equipment. During the HGP and through the early 2000s, genomic sequencing work was typically carried out at large-scale, specialized

[32] http://www.genome.gov/11006943
[33] INSTITUTE OF MEDICINE & NATIONAL RESEARCH COUNCIL, LARGE-SCALE BIOMEDICAL SCIENCE 29 (2003) [hereinafter LARGE-SCALE SCIENCE].
[34] 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.
[35] *Leslie Roberts, Controversial from the Start, 291* Science *1182 (year).*
[36] *See* LARGE-SCALE SCIENCE, *supra* note x, at 39.
[37] Toronto Int'l Data Release Workshop Authors, *Pre-Publication Data Sharing*, 461 Nature 168, 169–70 (2009)

research centers. But according to recent estimates, since 2004 the cost of DNA sequencing has dropped by 50% every five months.[38] This precipitous price decline has enabled even the smallest labs to afford sophisticated gene sequencing equipment, and has shifted much sequencing work from large specialized centers to disaggregated labs across the world. As the community of data generating researchers expands, the willingness of this new corps of data generators to abide by the policies and norms forged by old-guard sequencing centers may be tested.

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 original 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 in nearly every biological discipline whose research may benefit from the use of genomic data. The emergence and growth of this large constituency of data users, and its divergence from the more tightly-knit community of data generators, has had a significant impact on policies for the release and use of genomic data.

4. Data Intermediaries. Individual researchers and laboratories that generate genomic data are seldom the ones that make it available to others. 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)). Not surprisingly, the interests of commercial and non-commercial data intermediaries differ in several regards, most notably in the area of pricing access to information. Nevertheless, these stakeholders also share a number of common motivations, including the desire to disseminate information in ways that are effective, secure and accurate and the need to maintain some level of financial sustainability. 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.[39]

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.[40] In later projects, however, and particularly with the commencement of large-scale GWA studies, concerns with the potential identification of human subjects has grown.[41] This is because a GWA study seeks to *associate* genotypic information (e.g., genetic markers) with disease risk,

[38] http://www.sciencemag.org/content/331/6018/666.full, Royal Society at 88
[39] *See* Ft. Lauderdale Principles, supra note x, at 4; Toronto Authors, *supra* note x, at 170.
[40] *The Human Genome Project Completion: Frequently Asked Questions*, NAT'L HUMAN GENOME RESEARCH INST., http://www.genome.gov/11006943.
[41] *See* Toronto Report, *supra* note x, at 170.

information regarding donor demographics, disease state and treatment are necessary to interpret 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. 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 statistically inferred from a group of otherwise anonymous samples.[42] Such findings increasingly 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.[43]

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. Second, members of the public who are themselves affected, directly or indirectly, by genetic disorders may 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.[44] Finally, even 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.[45] Thus, the public is an important stakeholder in both the creation and the use of the genome commons.

III. RULES-IN-USE OF THE HGP

Under the IAD framework, the "rules-in-use" or governance structure of a commons system constitutes 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 a 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.

---

[42] 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.

[43] *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. Ossorio, 2011. Also cite Presidential Commission for the Study of Bioethical Issues, Aug 2012 report.

[44] *See* Lee, *supra* note 34, at 986–90 and Sharon F. Terry, et al., *Advocacy Groups as Research Organizations: The PXE International Example*, 8 NATURE REVIEWS GENETICS 157, 157–162 (2007).

[45] See S.B. Haga & J.O'Daniel, *Public Perspectives Regarding Data-Sharing Practices in Genomics Research*, Public Health Genomics, Mar. 24, 2011.

Ostrom defines the "rules in use" of a commons as both its formal *de jure* rules coupled with the informal (but often forceful) norms that govern its members' behavior. As explained by Peter Boetke, "it is the 'rules in use' (the lived practice of everyday life) that matter for social cooperation, not so much the 'rules in form' (on the books)".[46] In the case of the genome commons, formal rules-in-use were established at the outset of the HGP, but were strongly influenced by the norms of the scientific community and continue to play an important role.

A. EARLY YEARS OF THE HGP

Several factors contributed to the impetus, 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 small scientific communities that worked on these organisms traditionally abided by strong "open science" norms and were accustomed to sharing data freely, laying a strong precedent for the HGP.[47] Moreover, and perhaps more importantly, there was a sense among the leadership of the project that genomic data possessed a special and unique character. In the words of Ari Patrinos, the DOE's Associate Director for Biological and Environmental Research, the genome "belongs to everybody."[48] 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."[49]

In 1992, shortly after the project was launched, NIH and DOE developed formal guidelines for the sharing of HGP data.[50] These guidelines were viewed as essential to achieve the program's goals, avoid unnecessary duplication of effort and expedite research in other areas. 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. 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 hold-back 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.

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. This pro-patent attitude arose contemporaneously with NIH's unsuccessful attempt to seek patents on ESTs, and had waned significantly by the mid-1990s.

B. THE BERMUDA PRINCIPLES

[46] Boettke, Living Economics, p.x

[47] *See* HGP Initial Paper, *supra* note x, at 864; NRC – PUBLIC DOMAIN, *supra* note x, at 89; NRC - GENOMIC AND PROTEOMIC RESEARCH, *supra* note x, at 54–56.

[48] Eliot Marshall, *Bermuda Rules: Community Spirit, With Teeth*, 291 SCIENCE 1192 (2001).

[49] NATIONAL RESEARCH COUNCIL, MAPPING AND SEQUENCING THE HUMAN GENOME 8 (1988) [hereinafter NRC – HUMAN GENOME].

[50] *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].

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 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.[51] 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."[52] Most importantly, the Bermuda Principles required that this data should be released in public databases within a mere *twenty-four hours*.

The Bermuda Principles achieved several of the most important policy objectives held by the HGP funders. First, they greatly enhanced project coordination by enabling HGP sequencing centers to avoid duplication of effort and optimize their respective tasks.[53] 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., making sequence data as broadly available as possible as quickly as possible to accelerate discoveries for the benefit of society).[54]

2. *Rapid Data Release and Patents*. In addition to the effects described above, rapid data release under the Bermuda Principles was also believed to limit the ability of researchers to obtain patent protection on data generated by the HGP. In particular, the Bermuda Principles ensured that HGP data would be made publicly-available before data generators could file patent applications covering "inventions" arising from that data in most countries, and in a manner that ensured its availability as prior art against third party patent filings at the earliest possible date.[55] This result, though lauded 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 allows the patenting of federally-funded inventions for the benefit of the U.S. economy.[56] In response to this criticism, NIH's 1996 policy adopting the Bermuda Principles pays lip service to the Bayh-Dole Act, acknowledging that recipients of NIH funding have the right to seek patents on inventions that "reveal convincing evidence for utility." But in the next breath the agency warns that it "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."[57] The consequences of violating this prescription are left unstated. NIH's approach is

[51] *See* Marshall, *supra* note 48, 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.
[52] Bermuda Principles, *supra* note x.
[53] 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).
[54] *See* Bentley, *supra* note 53, at 533; Cook-Deegan & McCormack, *supra* note 51.
[55] Contreras, 2 Narratives
[56] Bayh-Dole Act of 1980, 35 U.S.C. §§ 200-12 (2006).
[57] 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.

thsu one of norms-setting rather than the imposition of legally-enforceable penalties, a tactic that will be seen repeated throughout the evolution of the genome commons.

The significance of NIH's implementation of the Bermuda Principles[58] cannot be overstated. Prior to 1996, NIH's position with respect to data release and intellectual property was not very different than that of other federal agencies. But in the negotiations at and leading up to the Bermuda meeting, the scientific community's acknowledgement of the collective norms of data sharing seems to have captured the agency's imagination. These norms have since become ingrained as part of NIH's basic position that genomic data 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. Then, in May of that year, J. Craig Venter, a former NIH scientist, 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 a full four years before the publicly-funded HGP.[59] 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.[60] Ultimately, a truce was brokered by the preeminent scientific journal *Science*, which agreed to publish the genomic sequence generated by Celera, while its rival *Nature* would publish the sequence assembled by the public HGP.[61] In June 2000, Francis Collins, Director of the HGP, and Venter joined President Bill Clinton at the White House to announce that a "first draft" of the human genome sequence had been completed and both sides declared a major scientific victory. [62] [63]

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 commercial web site, rather than in GenBank. 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.[64] 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.

---

[58] *See* NHGRI 1996 Policy, *supra* note 57; 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.

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

[60] *See* Roberts, *supra* note 59, at 1188. Add cites to Shreve, Venter, Ridley.

[61] cite

[62] Roberts, *supra* note 59, at 1188; Nicholas Wade, *Genetic Code of Human Life is Cracked by Scientists: A Shared Success*, N.Y. TIMES, June 27, 2000, at A1.

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

[64] Eliot Marshall, *Storm Erupts over Terms for Publishing Celera's Sequence*, 290 SCIENCE 2042, 2042 (2000).

Ultimately, in the settlement brokered by the journal *Science*, Celera agreed to make its data broadly available under a somewhat less restrictive licensing agreement.[65] The HGP draft sequence was published in GenBank in 2001,[66] and by 2003 most of the genes contained in Celera's database had been resequenced and released publicly by the HGP. Celera's subscription-based data business was ultimately unsuccessful and, in 2005, the company released its genomic data to GenBank.[67]

IV. THE ACTION ARENA: EVOLUTION OF RULES AND NORMS

Under the IAD framework, the "action arena" constitutes the set of scenarios in which the participants interact with respect to the common resource.[68] "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."[69]

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, whether publicly and privately funded.

A. DATA GENERATORS VERSUS DATA USERS

In their effort to promote the policy goals of the massive HGP, the projects's organizers knowingly subrogated the interests of data generators to those of the public. That is, the rapid data release requirements of Bermuda effectively eliminated the ability of data generators to

[65] Eliot Marshall, *Sharing the Glory, Not the Credit*, 291 SCIENCE 1189-93 (2001).
[66] *See* HGP Initial Paper.
[67] Jocelyn Kaiser, *Celera to End Subscriptions and Give Data to Public GenBank*, 308 SCIENCE 775, 775 (2005).
[68] Ostrom & Hess, *supra* note x, at 53-59.
[69] Cultural Commons, at 682.

publish their analyses and conclusions before others could access "their" data.[70] 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 recognize the inherent tension between data generators and data users. This distinction was first codified in a new NIH policy adopted in 2000.[71] The policy reaffirmed the Institute's Bermuda-based requirement that DNA sequences be deposited into GenBank within twenty-four hours of assembly. For the first time, however, it also imposed formal requirements on *users* who downloaded this data. The policy references "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." The policy goes on to prohibit users from using publicly-released HGP data "for the *initial* publication of the complete genome sequence assembly or other large-scale analyses," thereby reserving this right to the data generators. This concession to the requirements of data generators may not be large, but it laid the groundwork for many of the post-HGP policy shifts that followed.

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

The HGP largely completed its work in 2001. Two years later, in 2003, the Wellcome Trust convened a summit of funding agencies, sequencing centers, database managers, biological laboratories and scientific journals in Ft. Lauderdale, Florida to revisit rapid data release issues in the "post-genome" world.[72] The Ft. Lauderdale meeting coincides with the spread of genomic research beyond the traditional genomics community (e.g., to researchers in oncology, virology, microbiology). These researchers, who were trained in non-genomics research traditions, did not share the same basic norms of data sharing and openness as the original HGP research groups.[73] Thus, 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. They drew a distinction between HGP-like "community resource projects" (CRPs) that were "specifically devised and implemented to create a set of data, reagents or other material whose primary utility will be as a resource for the broad scientific community" and "hypothesis-driven" research, in which the goal is to answer a particular scientific question through the interrogation of experimental data. In hypothesis-driven research, success is typically measured by the degree to which the scientific question is answered rather than the completion of a quantifiable data set or other product. Scientists engaged primarily in hypothesis-driven research generally resisted the early release of data. Giving data away before theories were finalized or published might enable a competing group to "scoop" the data generator, a persistent fear among highly competitive scientists. This risk, and the "legitimate interest" of data generating scientists to be the first to publish the results of their

[70] Deanna M. Church & LeDeana W. Hillier, *Back to Bermuda: How is Science Best Served?* 10 GENOME BIOLOGY 105, 105.1 (Apr. 24, 2009).

[71] *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.

[72] 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].

[73] Cook-Deegan & x, x

work, were also recognized by NIH.[74] 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.

Figure 1

Organizational Schematic of the Genome Commons

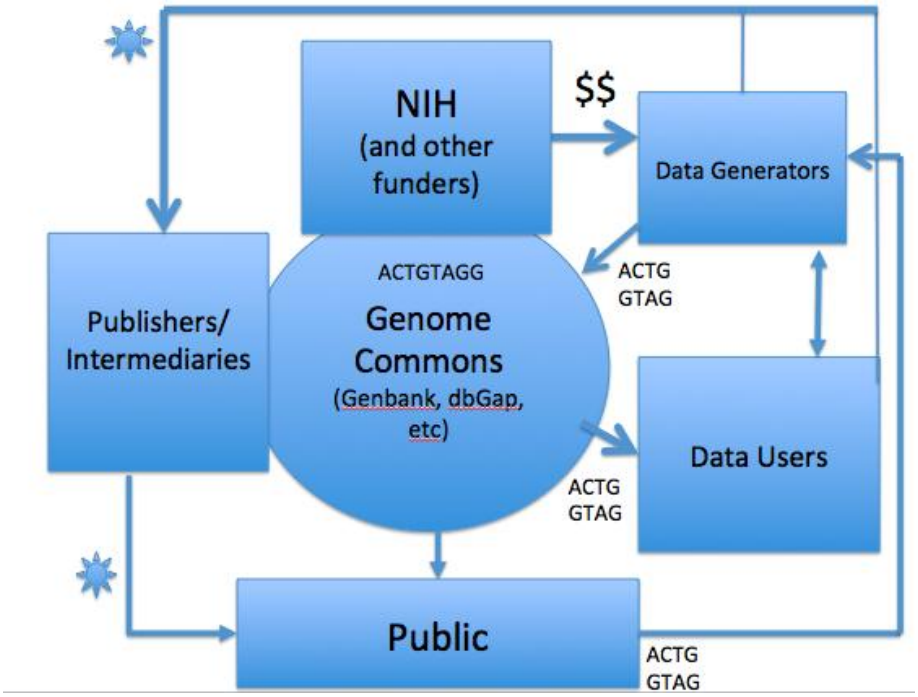


Figure 1 depicts the organizational structure of the genome commons and its associated stakeholder groups. NIH and other funding organizations support research by data generators that creates the common resource, using biological samples collected from members of the public. The resource is used by data generators and data users, who are often collaborators or even identical, and, to a lesser degree, the general public. Data generators and data users release results through publishers (typically in the form of scientific articles), which enrich the understanding of the public. The resource itself is housed within large databases maintained by government funding agencies (e.g., NIH's National Library of Medicine) and by some journals.

C. SECOND GENERATION PUBLIC DATA RELEASE POLICIES

In the years following the Ft. Lauderdale meeting, numerous large-scale genomic research projects were launched with increasingly sophisticated requirements regarding data

[74] *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].

release. These policies implement their requirements through contractual mechanisms that are more tailored and comprehensive than the broad policy statements of the HGP era. Moreover, increasingly sophisticated database technologies have enabled the provision of differentiated levels of data access, the screening of user applications for data access, and improved tracking of data access and users.

1. Genetic Association Information Network (GAIN). The Genetic Association Information Network (GAIN) was established in 2006 by the Foundation for the National Institutes of Health (FNIH), the NIH and several corporations.[75] GAIN's purpose was to conduct GWA studies of the genetic basis for six common diseases. Data generators in the GAIN program were required to sign an agreement calling for immediate release of data generated by the project.[76] Over the course of the three-year project, approximately 18,000 human DNA samples were genotyped and the resulting data was deposited in dbGaP.[77] 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 were required to be approved by the GAIN Data Access Committee.[78] Once approved, they were required to agree to keep the data secure, use it only for approved research purposes, refrain from patenting the data or conclusions drawn directly from it, acknowledge data generators, and refrain from attempting to identify individual study participants.[79] Perhaps most importantly, the GAIN policy is the first genomic data release policy to introduce a temporal restriction on the *users* of released data. That is, in order to secure a period of exclusive use for data generators, data users are prohibited from publishing and making presentations based on GAIN data for a specified embargo period.[80] 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.

2. The NIH GWAS Policy. In response to the growing number of GWA studies being funded by NIH and the large amount of genomic data generated by such studies, in August 2007 NIH released a new policy regarding the generation, protection and sharing of data generated by federally-funded GWA studies.[81] The NIH GWAS Policy requires that researchers submit descriptive information about each GWA study for inclusion in the "open access" portion of dbGaP. Grantees are also "strongly encouraged" to submit study results, including phenotypic, exposure and genotypic data, for inclusion in the "controlled access" portion of the database "as soon as quality control procedures have been completed."

---

[75] The GAIN Collaborative Research Group, *New models of collaboration in genome-wide association studies: the Genetic Association Information Network*, 39 NATURE GENETICS 1045 (2007).
[76] The GAIN Collaborative Research Group, *supra* note 75, at 1048 (Box 1).
[77] 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).
[78] Gain Collaborative Research Group, *supra* note 75, at 1049.
[79] *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*].
[80] GAIN Collaborative Research Group, *supra* note 75, at 1049.
[81] 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].

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.[82] To address this concern, the NIH GWAS Policy requires that GWAS data be de-identified in accordance with applicable regulatory guidelines. 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, and then only under a signed Data Use Certification that contains stringent protective clauses.

The NIH GWAS Policy addresses the publication priority concerns of data generators by announcing an "expectation" that users of GWAS data refrain from publishing or presenting their analyses and conclusions during an "exclusivity" period of up to twelve months from the date that the data set is first made available (Fig. 1). The agency also expresses a "hope" 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."[83] Regarding patents, the GWAS policy states 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." However, in an effort to show some support for patent seekers, the GWAS Policy also "encourages patenting of technology suitable for subsequent private investment that may lead to the development of products that address public needs."

3. The ENCODE Project. In 2007 NIH launched the ENCODE and modENCODE projects to identify functional genomic elements in humans and two simpler model organisms.[84] The ENCODE data release policy[85] designates the project as a "community resource project", but 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. The ENCODE Policy distinguishes between published and unpublished data, verified and unverified data, and offers several examples of the data use implications for different types of studies. The length and complexity of the policy evidences the agency's and the participants' desire for clear guidelines and the avoidance of misunderstandings regarding the release of data, as the diversity of participants, organisms and data types has expanded dramatically beyond those originally considered by the framers of the Bermuda Principles.

4. 1000 Genomes. The 1000 Genomes Project is an international cooperative effort to improve understanding of human genetic variation by sequencing the genomes of approximately 2000 individuals from diverse populations.[86] Much of the sequencing work for the project has been funded by NIH. As noted above, the project has generated unprecedented quantities of data and has given rise to novel approaches to data handling and management. The project's genomic data is release to the public through Amazon Web Services with very few

[82] Pilar N. Ossorio, *Bodies of Data: Genomic Data and Bioscience Data Sharing*, 78 Social Res. 907, 915-19 (2011).
[83] *NIH GWAS Policy*, *supra* note 81, at 49296.
[84] *See* Susan E. Celniker et al., *Unlocking the Secrets of the Genome*, 459 NATURE 927 (2009).
[85] 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"].
[86] http://www.biomedcentral.com/content/pdf/gm124.pdf

restrictions. The project is classified as a community resource project and cites the Ft. Lauderdale Principles. Though no formal embargo requirements are imposed on the data, the project's data release policy states an expectation that data generators be allowed to "make the first presentations and to publish the first paper with global analyses of the data".[87] Other guidance regarding the order of publication for different types of analyses is provided, though the binding nature of these guidelines is questionable.

5. The Human Microbiome Project. The Human Microbiome Project (HMP) is a large-scale community resource project initiated in 2007 and fully launched in 2010 that will identify and sequence the genomes of many of the microorganisms inhabiting the human body.[88] While much HMP data is subject to rapid Bermuda-like disclosure requirements, investigators are permitted to withhold certain other data from the public for a period of severeal months.[89] This hold-back period is intended to permit HMP researchers to analyze and prepare publications on their data before it is released to competing researchers. The reasons that researchers, who are driven by intense competitive pressure to publish and claim credit for discoveries, have pushed for such hold-back periods is clear. However, it also appears, at least in the case of HMP, that NIH has not vigorously advanced the patent deterrent arguments that previously motivated policy decisions during the HGP and its immediate aftermath. Whether the experience of the HMP indicates a new direction for NIH, or simply a minor deviation from its overall policy mission, is not clear.

D. PRIVATE SECTOR INITIATIVES.

In addition to the HGP and other public sector sequencing efforts described above, a number of private sector projects have 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. While it is certainly the case that many private-sector research efforts have been undertaken within the highly proprietary environments of pharmaceutical and biotechnology companies, the existence of privately-funded activities that contribute to the public genome commons suggests that the common resource structure established by NIH and its publicly-funded projects has taken hold as an accepted mode of organizing genomic research, even in the private sector.

1. The Merck Gene Index. Beginning in 1994, pharmaceutical giant Merck initiated a project to identify and publicly release a large number of expressed sequence tags (ESTs).[90] By 1998, the so-called Merck Gene Index included more than 800,000 ESTs, which were also released through GenBank.[91] A full analysis of motivations fueling industrial scientific research

[87] http://www.1000genomes.org/data/
[88] http://ukpmc.ac.uk/articles/PMC2940224/reload=0;jsessionid=uXLFA9IFUXx9aRFjxOZX.0
[89] *HMP Data Release and Resource Sharing Guidelines for Human Microbiome Project Data Production Grants*, NIH COMMON FUND, http://commonfund.nih.gov/hmp/datareleaseguidelines.asp.
[90] *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].
[91] DON TAPSCOTT & ANTHONY D. WILLIAMS, WIKINOMICS: HOW MASS COLLABORATION CHANGES EVERYTHING 166 (2006).

is beyond the scope of this chapter. However, in the case of the Merck Gene Index, it is generally believed that Merck chose to release these potentially valuable assets through a combination of philanthropic intent and corporate self-interest (i.e., preempting patenting of ESTs by biotech companies, several of which had already announced business plans that involved the patenting and licensing of ESTs and other genetic information).[92] To achieve these goals, Merck found it most expedient to release EST information directly to the public domain, without material restrictions, much as the HGP would do. The development of an organizational structure to oversee the use of the Merck Gene Index does not seem to have been necessary for Merck to achieve its immediate goals.

2. *The SNP Consortium.* A related but distinct approach was adopted by the SNP Consortium. This non-profit entity was formed in 1999 by a group of pharmaceutical companies and the Wellcome Trust to identify and map genetic markers known as "single nucleotide polymorphisms" (SNPs).[93] Responding at least in part to the threat that biotechnology companies might identify and patent these SNP markers first, the SNP Consortium made all SNPs that it identified available to the public.[94] It also sought to deter patenting of SNPs by third parties using an innovative "protective patenting" strategy that has been cited as a model of the private industry's potential to contribute to the public genome commons.[95] In short, the consortium's approach was to file U.S. patent applications covering SNPs that it discovered, and then contribute these applications to the public domain prior to issuance. 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.

3. International SAE Consortium. Since the completion of the SNP Consortium project, several other privately-funded research collaborations have adopted similar data release models and have made large quantities of genomic data publicly accessible. One of these is the International SAE Consortium (SAEC), a group of pharmaceutical and healthcare companies organized in 2007 to fund research seeking to identify DNA markers associated with serious drug side effects (adverse events).[96] The SAEC seeks to minimize patent encumbrances on the genetic markers and associations that it identifies via a "protective" patent strategy modeled on that of the SNP Consortium. Like the other policies discussed in this section, the SAEC imposes various security, research 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.[97] Unlike previous private sector efforts such as the Merck Gene Index and SNP Consortium, SAEC has created a managed common resource subject to rules that echo those of contemporary NIH-funded projects. The fact that this commons-based organizational structure has emerged in the private sector is particularly noteworthy, as the

[92] Marshall, *supra* note 48. TAPSCOTT & WILLIAMS, *supra* note 91; Arti Kaur Rai, *Regulating Scientific Research: Intellectual Property Rights and the Norms of Science*, 94 NW. U. L. REV. 77, 134 (1999–2000).
[93] Arthur Holden, *The SNP Consortium: Summary of a Private Consortium Effort to Develop an Applied Map of the Human Genome*, 32 BIOTECHNIQUES 22 (2002).
[94] *See, e.g.*, Holden, *supra* note 93, at 26. TAPSCOTT & WILLIAMS, *supra* note **91**, at 168.
[95] *See, e.g.*, Marshall, *supra* note 48, at 1192; Cook-Deegan & McCormack, *supra* note 51.
[96] *iSAEC's Background and Organizational Structure* INT'L SAE CONSORTIUM http://www.saeconsortium.org/ (last accessed Oct. 28, 2010).
[97] Int'l SAE Consortium Ltd., DATA RELEASE AND INTELLECTUAL PROPERTY POLICY (last amended Nov. 5, 2009) (on file with author).

research funded by SAEC is more akin to hypothesis-driven research than a community resource project, a fact that might push the project toward a more proprietary model, even in the government-funded realm. As such, the SAEC approach is a testament to the unusually strong shared community norms within the genomics research community, even among scientists working in the private sector.

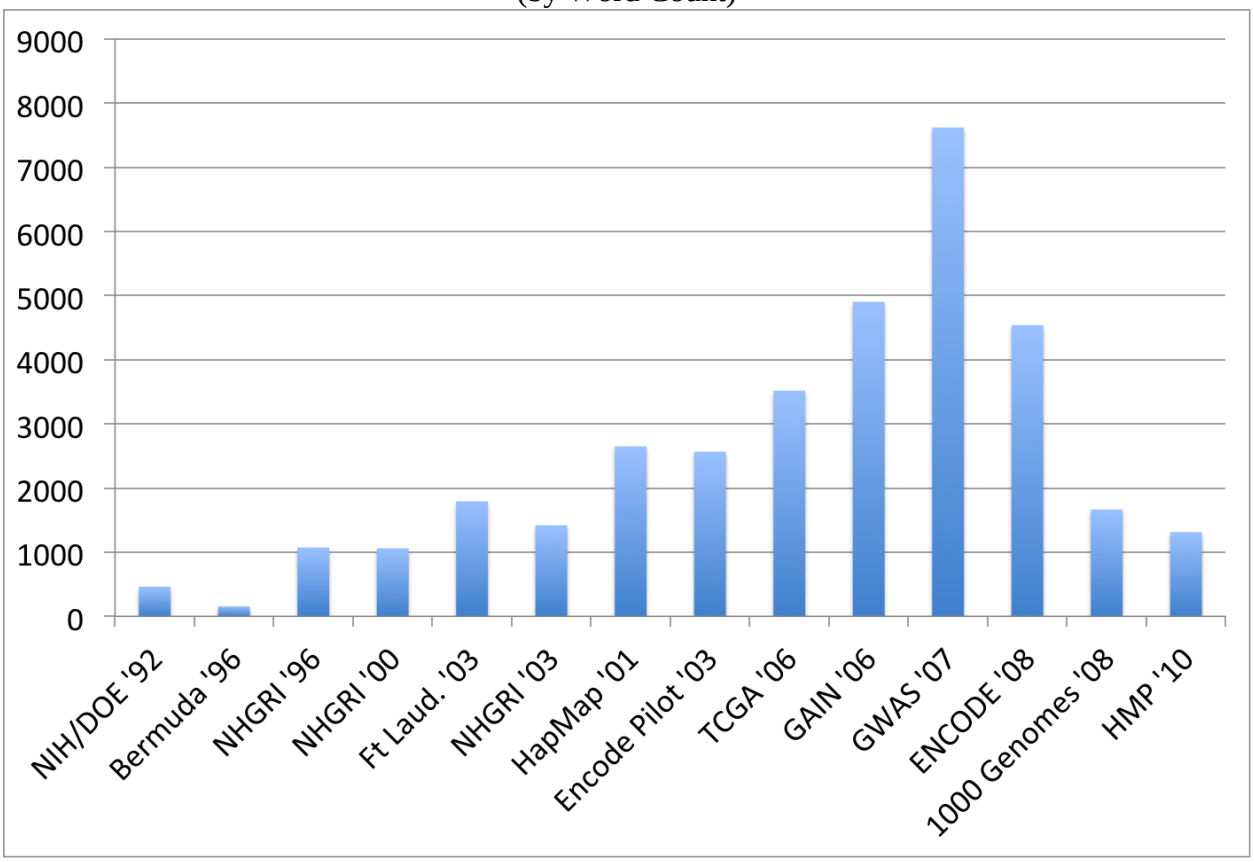
4. *Personal Genome Project*. An interesting recent addition to the genome commons is the Harvard-led Personal Genome Project (PGP).[98] The PGP, launched in 2008 to significant press coverage, solicits volunteers to submit tissue samples and accompanying phenotypic data. Researchers are then authorized to analyze the submitted samples and publish any resulting genomic information on the PGP web site. All such data is released without restriction under the "CC0" Creative Commons copyright waiver.[99] The PGP approach differs markedly from that of the government and privately-funded projects described above, in that it dispenses entirely with any attempt to restrict the use of its genomic data. PGP requires its contributors to waive all privacy-related rights when contributing their tissue samples to the project, and gives no preference to use of the data by researchers of any kind. As such, the PGP policy returns to the broad "public domain" character of the Bermuda Principles and may signal, at least among some researchers, a re-conception of genomic data as a public good.

V. RULES AND COMPLEXITY IN THE GENOME COMMONS

The genome commons has experienced rapid and unanticipated growth over the past two decades. The rise of this valuable public resource has seen an accompanying growth of complexity in the formal rules governing the commons. Whereas the initial HGP required the rapid release of genomic data to the public, effecting what might be considered a *public good* in economic terms, later projects added increasingly complex rules governing human subject privacy and data generator publication priority. *Figure 2* illustrates the relative complexity of the formal policies governing the major U.S. genomic research projects from the HGP through current projects, comparing their data release policies on the basis of a straightforward word count algorithm.

[98] www.personalgenomes.org/mission.html
[99] http://creativecommons.org/publicdomain/zero/1.0/legalcode

Figure 2
Genomic Policy Complexity
(by Word Count)



There are several possible explanations for the trends shown in *Figure 2*. Despite the groundbreaking significance and lasting influence of the Bermuda Principles, they 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. By the mid-2000s, both the sophistication of genomic studies and the range of researchers participating in genomic science had broadened significantly. Thus, the rules governing the commons grew, both in terms of length and complexity.

But *Figure 2* also suggests that NIH's 2007 GWAS policy may represent a "peak" in the expansion of the rules governing the genome commons. The post-2007 policies represented in the figure are significantly less complex (or at least less verbose) than the 2007 GWAS policy (though still much more complex than the original Bermuda Principles). This effect may simply be a result of growing public and scientific familiarity with the concepts set out in data release policies and a concomitant reduction in the need for explanation of various concepts. But it may also represent a lessening community interest in specifying the details surrounding data release. Unlike the carefully, and cautiously, drafted policies adopted by NIH in the early 2000s, the later

policies say little if anything about patents. Could this absence indicate a waning concern by the agency with the patenting of genomic information? Or might it indicate, instead, a growing sense within the agency that patent law has stabilized to a degree that further agency pronouncements and guidance are not necessary?

NIH is in the process of considering yet further revisions to its institutional data release policies and collecting feedback from various stakeholder groups.[100] Though the results 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. It is thus important that stakeholders with an interest in the future structure of the ever-expanding genome commons participate in such deliberations.

VI. THE GENOME COMMONS AS A CONSTRUCTED 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."[101]

Scientific research has not typically been viewed as a form of cultural production. In fact, even Madison, Frischmann and Strandburg point to scientific research as an area adhering to the traditional functionalist view of "IP rights and government subsidies".[102] The results of research – general scientific knowledge – is often characterized as a public good in economic terms. But this 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.[103] An analysis of the genome commons supports this view, both as to researchers in the public sector, and also to a degree within the private sector.

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 governmental grants in the U.S. and a major philanthropy in the UK, 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 organizational structure or dynamics observed in the genome commons.

---

[100] 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.
[101] Cultural Commons, *supra* note x, at 669.
[102] *Id*. at 665, 666
[103] *See, e.g.*, Merton, *supra* note x, and Rai, *supra* note x. 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).

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, as well as the ethical complications arising from the increasing richness of the data unearthed by genomic research. As the HGP neared conclusion, it became evident that a purely public domain/public goods structure for the vast pool of genomic information being produced by researchers worldwide would simply not suffice. Instead, a set of negotiated compromises emerged from the multistakeholder interactions of government, researchers and the public. These interactions resulted in sophisticated rules-in-use, both formal and norms-based, to govern the commons. By the late 2000s, the genome commons less resembled the public domain than a managed common resource studied by Ostrom and her colleagues. Thus, while some may seek to characterize genomic data as an unrestricted public good, the vast majority of genomic data exists within a complex and formalized structure of rules, a structure that both enables and encourages the growth of the resource and its continuing use in further research.

In this sense, the genome commons can, and should, be viewed as a cultural commons of the type conceptualized by Madison, Frischmann and Strandburg. As such, it is crucial to study and understand its complex rule structure. Failing to do so, and too quickly embracing either a simple public good/public domain model or a highly proprietary, protectionist model, would run counter to the carefully negotiated compromises achieved over the past two decades. Complex considerations of human subject protection, data generator priority, patent deterrence and scientific advancement all contributed to the current policy landscape of the genome commons. Failing to appreciate the structural rules implemented to address these issues, or seeking to dispense with them in favor of a more broadly "open" public goods models such as those advanced in Public Genome Project and, to a lesser degree, in later NIH-funded projects, could have adverse consequences. In particular the elimination of rules regulating human subject protection could limit the willingness of individuals to participate in genomic research, and the elimination of data generator priorities could weaken the incentives of data generating scientists. Each of these effects could negatively impact the growth of the commons itself. Thus, policy makers today and in the future should carefully consider the rules structure that has evolved to govern this invaluable global resource and take heed before introducing changes likely to upset its delicate and well-functioning balance of interests.