3

Centralization, Fragmentation, and Replication in the Genomic Data Commons

Peter Lee

INTRODUCTION

Genomics – the study of organisms' genomes – holds great promise to advance biological knowledge and facilitate the development of new diagnostics and therapeutics. Genomics research has benefited greatly from various policies requiring rapid disclosure of nucleotide sequence data in public databases. Such disclosure has helped create a genomic data commons, a repository of information widely accessible to all members of the scientific community. Notably, this commons operates almost completely outside the strictures of formal intellectual property law through a combination of public funding, agency policy, and communal norms. The genomic data commons has attracted significant scholarly interest because of both its great potential to facilitate biomedical research and its broader lessons about the nature of commons-based production systems (Frischmann et al. 2014; Contreras 2014; Van Overwalle 2014). For instance, recent data release policies by governmental and nongovernmental entities have helped shift the genomic data commons from a more open structure toward a governance regime of selective access and exclusivity. This chapter builds on this rich literature to provide a more granular analysis of the genomic data commons, exploring less appreciated but highly significant challenges of managing existing information in the commons. In so doing, it seeks to shed greater light on the nature of commons in general.

In particular, this chapter focuses on the governance challenges of correcting, updating, and annotating vast amounts of sequence data in the commons. Most legal accounts of the genomic data commons focus on researchers' initial provisioning of data and its use by the scientific community. These accounts implicitly assume that

Peter Lee is Professor of Law at the University of California, Davis, School of Law, Davis, California. Professor Lee would like to thank Janice Ahn, Martin Bobrow, Liza Vertinsky, and workshop participants at Yale Law School, NYU School of Law, and the US Patent and Trademark Office for comments on earlier versions of this chapter. Special thanks to Brett Frischmann, Michael Madison, and Kathy Strandburg for organizing and editing this volume.

the data is largely "correct" and that the most significant governance challenges involve managing access to that data. Delving into the science of genome sequencing, assembly, and annotation, however, this chapter highlights the indeterminate nature of sequence data and related information, thus giving rise to a strong need to correct, complete, and update existing data. This chapter draws on the Institutional Analysis and Development (IAD) methodological framework developed by Elinor Ostrom and refined by Michael Madison, Brett Frischmann, and Katherine Strandburg (Ostrom and Hess 2007; Madison et al. 2010) to examine four approaches for rendering the genomic data commons more accurate and intelligible: third-party biocuration, contributor-centric data management, community-based wikification, and specialized databases and genome browsers. It argues that these approaches reveal deep tensions between centralization and fragmentation of control over data modification within the genomic data commons, a tension that can be mitigated through a strategy of replicating information.

On the one hand, third-party biocuration and contributor-centric data management tend to consolidate and centralize control over data. On the other hand, wiki-based annotation fragments control throughout the community, exploiting the power of peer production and parallel data analysis to modify existing data records. Both centralization and fragmentation have their strengths and weaknesses, and this chapter argues that stakeholders can capture the best of both worlds through exploiting the nonrivalrous nature of information, which can be consumed without diminishing its availability for other uses. In particular, researchers are engaged in a strategy of replicating and reproducing multiple views of sequence data, employing specialized databases and genome browsers that combine centralized, archival data and widespread community input to provide more textured, value-added renderings of genomic information.[1] Among other advantages, this approach has important epistemological implications as it both reflects and reveals that genomic knowledge is the product of social consensus.

This analysis sheds new light on the dynamic structure of the genomic data commons. Within a conventional perspective, the commons represents a repository of land, information, or other assets that is open to a particular community (Smith 2000). Perhaps because of analogies to physical resources, commentators often characterize the commons' constituent resource units – such as fish, oil, or bits of information – as largely static and fixed. While the overall number of units may change, such as when fish are caught or additional bits of information enter the commons, the underlying resource does not change. However, the genomic data commons reveals that communal efforts may change the nature of the constituent resource units themselves. Contributors to the commons not only provide and use

[1] Notably, community-based wikification as well as specialized databases and genome browsers also represent significant forms of user innovation. Users develop knowledge and tools to advance their own research interests and freely share them, thus benefiting the research community at large (Strandburg 2008, 2009; von Hippel 1976).

data; they also fundamentally transform the nature of that data. In this sense, as Frischmann et al. describe, the commons plays an active role in refining and producing knowledge (Frischmann et al. 2014: 11).

Furthermore, although the genomic data commons has been lauded as accelerating research, closer examination complicates the question of how and to what extent it truly operates as a commons. This chapter reveals that the genomic data commons is both less and more of a commons than previously thought. On the one hand, it features a highly centralized data architecture. The efforts of thousands of genomic researchers around the world feed into a consortium of three publicly sponsored databases, which members of the community may not modify directly. On the other hand, it may be more accurate to characterize this knowledge system as a set of commons on top of a commons. On one level, it is an archival data repository emerging from a global community of scientists. On another level, the genomic data commons also encompasses many subcommunities (often organized around model organisms) that develop their own specialized databases and nomenclatures. Additionally, user groups develop infrastructural meta-tools such as genome browsers and freely distribute them throughout the community. In this sense, the genomic data commons represents a nested set of commons (Strandburg et al. 2014: 156). To simply refer to this as a genomic data commons is to miss some of the nuance and complexity of this knowledge management construct.

Finally and relatedly, this chapter highlights the strong role of centralization and standardization in the effective operation of a commons. The commons is often perceived as an open space free of government intervention and insulated from market demands (Dagan and Heller 2001: 555). Indeed, the genomic data commons has been structured quite conscientiously to operate outside the legal and financial influence of patents. However, the genomic data commons reveals that commons-based productivity systems are not simply free-for-alls lacking order or regulation (Rose 1986: 713). Too much control, and the power of parallel processing and peer production goes unrealized. Too little control, and the commons simply dissipates into chaos and entropy. Truly effective commons function at the balance of centralization and fragmentation.

Section 3.1 reviews the history of genomic data release policies and their implications for commons scholarship. Section 3.2 builds on this foundation by exploring the underappreciated challenges of correcting, updating, and annotating genomic data. Applying the IAD framework, this section examines in greater detail the contingent nature of the information at the heart of the genomic data commons. Delving into the science of genome sequencing, assembly, and annotation, this section shows that genomic information is much less determinate and complete than generally perceived and is constantly in need of updating and refinement. Section 3.3 continues to apply the IAD framework, this time focusing on issues of openness, control, and governance. Drawing on the scientific discussion in Section 3.2, it explores third-party biocuration, contributor-centric data management,

community-based wikification, and specialized databases and genome browsers as mechanisms for adding value to existing information. In particular, it explores deep tensions between centralized and fragmented control of data modification in the genomic data commons and attempts to revolve these tensions by exploiting nonrivalry and replication. Section 3.4 considers the deeper implications of these observations for the genomic data commons as well as commons in general.

3.1 THE EVOLUTION OF GENOMIC DATA POLICIES

The genomic data commons represents an illuminating case study of the promise and challenges of commons-based production. As early as the 1970s, molecular biologists and computer scientists recognized the need for a centralized, computerized database for DNA sequence data (Strasser 2008: 537). The National Institutes of Health (NIH) solicited various proposals and ultimately adopted the submission from Los Alamos National Laboratory in significant part because of its open design; the database operators structured it to be accessible through ARPANET and disavowed any proprietary interest in the data (Strasser 2008: 538). In 1982, NIH moved the Los Alamos database to the National Center for Biotechnology Information (NCBI) and renamed it GenBank (Lathe et al. 2008). In the years leading up to the Human Genome Project, the National Research Council recommended that all data generated by the project "be provided in an accessible form to the general research community worldwide" (Nat'l Res. Council 1988: 8). Based largely on its open design, GenBank ultimately became one of the primary databases of the Human Genome Project.

NIH and the Department of Energy (DOE) launched the Human Genome Project in 1990. Initially, researchers participating in the project released sequence data only upon publishing their findings in a peer-reviewed journal, consistent with scientific norms (Contreras 2011: 65; Nat'l Human Genome Res. Inst. 2000). Early on, however, NIH and DOE recognized the importance of sharing data even *before* publication to promote progress and avoid duplicative effort (Nat'l Inst. of Health and Dept. of Energy 1991). Accordingly, in 1992, NIH and DOE adopted a policy requiring publicly funded researchers to deposit sequence data into a public database, such as GenBank, the European Molecular Biology Laboratory (EMBL)–Bank, or the DNA Databank of Japan (DDBJ)[2] within *six months* of data generation, which may be long before publication in a scientific journal (Nat'l Inst. of Health and Dept. of Energy 1991).

However, the demand for even faster, prepublication data release soon led to another policy revision. At a 1996 conference in Bermuda, leaders of the biomedical research community agreed that all DNA sequence assemblies greater than 1 kilobase (kb) should be deposited in publicly accessible databases within *24 hours* after

[2] These databases comprise the central repositories of the International Nucleotide Sequence Database Collaboration (INSDC).

https://doi.org/10.1017/9781316544587.004 Published online by Cambridge University Press

generation (Dept. of Energy 1996).[3] In addition to facilitating rapid scientific advance, the so-called Bermuda Principles also preempted patents on genomic sequences (Contreras 2010: 393; Eisenberg 2000: 72). As Jorge Contreras observes, the Bermuda Principles "represent a significant achievement of private ordering in shaping the practices of an entire industry and establishing a global knowledge resource for the advancement of science" (Contreras 2011: 65). As will be a consistent theme, policymakers played a catalytic role in both recognizing and solidifying community consensus, which ultimately became formalized as agency policy. In 1997, the National Human Genome Research Institute (NHGRI) officially adopted the Bermuda Principles in its funding policy (Nat'l Human Genome Res. Inst. 1997).

Although the Bermuda Principles garnered praise for accelerating collective research, concerns arose that rapid access to sequence data might compromise other interests. For example, researchers often had to release data well before they could publish their findings, thus allowing other scientists to "scoop" them by publishing first (Eisenberg 2006: 1021; Marshall 2002: 1206). To address these concerns as well as complications arising from new sequencing technology, in 2000 NHGRI revised its data release policy (Nat'l Human Genome Res. Inst. 2000). The new policy prohibited users from utilizing public data "for the initial publication of the complete genome sequence assembly or other large-scale analyses." The policy further stated that minor changes in sequence assemblies need not be subject to the 24-hour release policy of the Bermuda Principles and also noted the difficulty of applying the Bermuda Principles to the more recent technological development of "whole genome shotgun sequencing." For this high-throughput sequencing technique, a significant amount of time typically elapsed between generating initial sequence reads and assembling a clean sequence. While there was scientific value to releasing individual sequence reads immediately, such early release put data generators at a potential disadvantage. The policy thus encouraged other scientists to wait to allow data generators to first publish sequence assemblies and large-scale analyses, thus imposing restraints on data users (Contreras 2011: 89).

Soon, however, the balance shifted slightly back toward less constrained use of genomic data. In 2003, a high-profile gathering of sequence producers and users in Fort Lauderdale, Florida, reconsidered existing data release policies. The attendees enthusiastically "reaffirmed"[4] the 1996 Bermuda Principles, and they recommended extending this policy to all sequence data, including raw traces and whole genome shotgun assemblies (The Wellcome Trust 2003: 2). The attendees also agreed that rapid data release policies should apply to so-called community resource projects (CRPs), which are "specifically devised and implemented to create a set of data,

[3] The Bermuda Principles applied to sequence assemblies of 1 kb or greater.

[4] Although the text of the Fort Lauderdale Principles states that it "reaffirm[s] the 1996 Bermuda Principles," the Bermuda Principles require immediate release of sequence assemblies larger than 1 kb, while the Fort Lauderdale Principles require immediate release of sequence assemblies larger than 2 kb.

reagents or other material whose primary utility will be as a resource for the broad scientific community" (The Wellcome Trust 2003: 2). Significantly, the rules eliminated any formal restriction preventing users from publishing whole genome analyses before the sequencers' initial publication of the complete genome (Dennis 2003: 877). Nonetheless, the Fort Lauderdale Principles included hortatory language emphasizing that users should recognize and respect data generators' interest in publishing the first analyses of this data. A few months after the Fort Lauderdale meeting, NHGRI adopted several elements of the Fort Lauderdale Principles in a formal policy statement (Nat'l Human Genome Res. Inst. 2003).[5] Although somewhat mitigated, the Fort Lauderdale Principles reflected an intuition that some proprietary interest in data may be valuable to maintain incentives to conduct research.

In addition to protecting scientific credit, concerns over privacy also led to greater pockets of exclusivity in the genomic data commons. Privacy concerns are particularly relevant for genome-wide association studies (GWAS) that discern how genetic patterns may contribute to disease (Kaye 2012: 417–18). Such studies produce data sets that link individual genotypes to phenotypes, thus raising the specter of associating individual test subjects with a genetic predisposition for a particular disease. Indeed, privacy issues have led researchers to remove large data sets from public databases (Lathe et al. 2008; Kaye 2012: 419). Concerns over privacy, particularly in the context of GWAS, motivated a "second generation" of data release policies (Contreras 2011: 97). Such policy evolution is evident in the Genetic Association Information Network (GAIN), a public-private partnership aimed at elucidating the genetic basis of various diseases (Contreras 2011: 97–99). Researchers participating in GAIN deposit their data in the database of Genotypes and Phenotypes (dbGaP), a database housed at the National Library of Medicine that links sequence and phenotype data for test subjects. As Contreras describes, dbGaP features a two-level system of both open and controlled access to data. Although the general public enjoys open access to summary, nonsensitive data, researchers seeking controlled access must register with the GAIN Data Access Committee (DAC) (Kaye 2012: 418, 424). Among other considerations, researchers must agree to utilize the data only for research purposes and not identify or contact research participants or make intellectual property claims derived directly from the data sets (Genetic Assoc. Info. Network 2010).

This interest in controlling access and protecting privacy is further reflected in broader NIH policies governing GWAS. In 2007, NIH released a policy emphasizing that "[r]apid and broad data access is particularly important for GWAS" (Nat'l Inst. of Health 2007). However, the policy establishes that an NIH DAC will regulate prospective users' access to the data. Among other conditions, users may only use

[5] There are some differences between the Fort Lauderdale Principles and NHGRI policy nominally adopting them. The NHGRI policy diverges from the Fort Lauderdale Principles in holding that sequence traces, including those from whole genome shotgun projects, should be deposited in a trace archive within *one week* of production.

data for approved research, and they must protect confidentiality, follow appropriate data security protections, and not identify individuals from whom data was collected. Additionally, although the policy requires rapid data release for federally funded researchers, it imposes a 12-month publication and presentation "embargo" on users to allow original data generators to publish their findings.

Government agencies and the scientific community have applied these policies to an evermore complex set of sequencing initiatives. Soon after the draft of the human genome was published in 2001, researchers launched the International HapMap Project, an ambitious initiative to create a haplotype map of the human genome that locates sets of statistically associated single-nucleotide polymorphisms (SNPs).[6] Participants in the HapMap Project adopted a data release policy similar to the Fort Lauderdale Principles, and they characterized their initiative as a CRP (Contreras 2011: 92). The 1000 Genomes Project builds off the International HapMap Project and GWAS "to discover, genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations" (The 1000 Genomes Project Consortium 2010). Similarly, NHGRI launched the Encyclopedia of DNA Elements (ENCODE) pilot project in 2003 to "elucidate the biological functions of various genetic elements" (Contreras 2011: 93). In so doing, NHGRI also adopted a data release policy modeled on the Fort Lauderdale Principles and designated ENCODE as a CRP (Contreras 2011: 94).[7] Additionally, NIH is sponsoring an ambitious project to sequence the genomes of the huge numbers of microbes that inhabit the human body (McGuire et al. 2008). NIH designated the data producing elements of the Human Microbiome Project as a CRP, and it has endorsed rapid data disclosure (Nat'l Inst. of Health 2013). NIH recently issued a new policy that requires researchers to deposit large-scale human genomic data in a public repository but embargoes data access for up to six months to allow the depositors to publish related findings (Nat'l Inst. of Health 2014).[8] Increasing complexity, including tensions between open and regulated access to data, will be a consistent theme of ongoing genome sequencing projects.

Not surprisingly, the emergence and evolution of the genomic data commons has attracted significant scholarly attention. In particular, Contreras and Van Overwalle have fruitfully applied a modified version of Ostrom's IAD framework to evaluate this socially constructed commons. These policies reflect a significant shift: "Whereas the initial HGP required the rapid release of genomic data to the public,

[6] An SNP arises when the DNA sequences of two organisms from the same species vary by a single nucleotide.

[7] Efforts to extend rapid-release policies beyond Bermuda and Fort Lauderdale have continued in the context of whole-genome association studies, microarray surveys, epigenomics scans, protein structures, screening of small molecules for biological activity, and functional genomics data (Pennisi 2009: 1000).

[8] Additionally, the new policy requires researchers to include a Genomic Data Sharing Plan in applications for funding and establishes a tiered system in which a Data Access Committee differentiates between data that is available on a controlled or unrestricted basis (Nat'l Inst. of Health 2014).

effecting what might be considered a *public good* in economic terms, later projects added increasingly complex rules governing human subject protection and publication priority" (Contreras 2014: 123). Indeed, these constraints have led Van Overwalle to reject characterizing the genomic data commons as a truly open knowledge commons, instead describing it as a limited "genome *research* commons" (Van Overwalle 2014: 150).

From the perspective of commons scholarship, this transition reflects an evolution from a largely open structure to a more tightly regulated "governance" model in which various actors manage access to data. Scholars have noted that governance regimes incur high information costs as decision makers must determine what parties have access to what resources under what circumstances (Smith 2002: S453). Many stakeholders are involved: government agencies that fund projects and define data policies, data generators who "populate" the genomic data commons, users who are expected to delay certain publications, and centralized data access committees that screen access to sensitive information. Far from a simple model in which data generators produce vast amounts of information for the public at large, the genomic data commons has assumed much greater complexity in its rules and operation. This chapter continues this theme of complexity by examining the challenges of correcting, updating, and annotating the data residing in the genomic commons. To do so, it must first examine in greater depth the kinds of information at the heart of this commons.

3.2 RESOURCE ATTRIBUTES: GENOME SEQUENCING, ASSEMBLY, AND ANNOTATION

As Frischmann et al. observe, an important characteristic of any commons is the resource that is the subject of communal pooling and sharing (Frischmann et al. 2014: 24). This section examines the information – or, more precisely, the multiple kinds of information – at the heart of the genomic data commons. Commentators generally describe this resource as "genomic data," but that phrase includes a multiplicity of types of information, including raw sequence reads, sequence assemblies, and annotated genomes as well as phenotypic and demographic data related to research subjects. To elucidate some of this complexity, this section delves into the science of genome sequencing, assembly, and annotation. While commons scholarship to date has focused largely on "upstream" functions such as sequencing and related analyses, this chapter looks more closely at the "downstream" functions of sequence assembly and genome annotation. Ultimately, these "value-added" functions represent additional vantage points from which to evaluate the structure, governance, and operation of the genomic data commons.

Traditional accounts of the Human Genome Project and its progeny focus on genome *sequencing*, an activity that is more complex than commonly perceived. At a conceptual level, sequencing involves determining the nucleotide sequence of

https://doi.org/10.1017/9781316544587.004 Published online by Cambridge University Press

particular segments of DNA. Thus, it is popularly understood that the Human Genome Project determined the nucleotide sequence for the human genome, and subsequent initiatives aim to determine the nucleotide sequence of various model organisms and other entities of interest. Mainstream conceptions of sequencing imply a high degree of determinism; one might conceive of scientists taking DNA samples and simply sequencing chromosomes from one end to the other, thus revealing a definitive series of As, Ts, Cs, and Gs that constitutes an organism's genome.

For a variety of technical reasons, however, DNA sequencing is far from determinate. For instance, researchers cannot sequence extremely long strips of DNA in one fell swoop. Conventional sequencing utilizing the chain-termination method (also known as Sanger sequencing) (Price 2012: 1619; Sanger et al. 1977) is rather limited and can only directly sequence relatively short nucleotide fragments (up to 1000 nucleotides long) in a single reaction (Nelson et al. 2011: 7.0.3). Chromosomes and other strips of DNA of interest, however, can be hundreds of thousands of nucleotides long. Scientists have utilized a technique known as "primer walking" to painstakingly sequence contiguous fragments one at a time until they can combine them to ascertain a sequence of interest.

In particular, "whole genome shotgun sequencing" has accentuated the indeterminate nature of genome sequencing and the extent to which researchers *assemble* sequences rather than simply discover them (Salzberg 2007). Whole genome shotgun sequencing utilizes the basic Sanger method but transcends some of its limitations by utilizing massive parallel processing. With the advent of commercial automated sequencing machines in the 1980s, this method of "high-throughput" sequencing became the norm (Nelson et al. 2011: 7.0.2). In this technique, researchers first purify the DNA of interest and then shear it into a huge number of small fragments.[9] Researchers then clone and sequence these small fragments in parallel, thus greatly accelerating the pace of sequencing. These sequenced fragments are called "traces,"[10] and fitting traces together into the natural genomic sequence is not a straightforward task. Researchers utilize complex software programs to analyze sequence overlaps and assemble these fragments into what is believed to be their natural order. Assembly can proceed *de novo*, similar to fitting the pieces of a jigsaw puzzle based on their edges, or based on a reference sequence that provides a guide (Price 2012: 1620). These assemblies consist of contiguous DNA sequences (contigs) held together by scaffolds.

[9] Sequencing traditionally used DNA purified from a pool of cells, which created more "noise." The current trend is to perform single-cell sequencing, particularly for RNA (interview with Dr. Janice Ahn, March 3, 2015).

[10] NIH's 2003 policy (following the Fort Lauderdale Principles) requires depositing traces from whole genome shotgun projects in an archive such as the NCBI Trace Repository or Ensembl Trace Server within one week of production (Nat'l Human Genome Res. Inst. 2003).

A crucial knowledge-enhancing step following sequencing and assembly is genome *annotation*. Such annotation involves "mapping" the location and function of genes on (what is perceived to be) a particular segment of DNA (Claverie 2000: 12).[11] Starting with the raw DNA sequences, annotation involves "adding the layers of analysis and interpretation necessary to extract its biological significance and place it into the context of our understanding of biological processes" (Stein 2001: 493). To this end, researchers might start by utilizing gene-finding software such as GlimmerM, which identifies sequences that are characteristic of protein-coding strands of nucleotides. This initial analysis provides the basis for a series of programs called Basic Local Alignment Search Tool (BLAST), a bioinformatics algorithm that enables researchers to compare a query sequence with a database of known sequences to correlated proteins (Benson et al. 2014: D36). Researchers utilize BLAST in cross-species comparisons to identify new genes. For example, a researcher who has discovered a new mouse gene may use BLAST to compare that nucleotide sequence with a human genome database to determine if humans have a similar gene. Once genes have been identified and mapped, researchers attempt to relate them to biological processes through functional annotation (Stein 2001: 500). Within this process, researchers often compare proteins (as well as messenger RNA) between species to determine process-level annotation (Stein 2001: 499). Ultimately, "customized annotation programs are used to decide what name and function to assign to each protein, leading to the final annotated genome" (Salzberg 2007: 102.2 fig. 1).

Genome sequencing, assembly, and annotation are far from precise sciences, and there is uncertainty and indeterminacy in each step. Whole genome shotgun sequencing generates enormous amounts of data, and the likelihood of errors is relatively high (Pennisi 2009: 1000). For example, sequencing techniques can introduce contaminants that get sequenced along with the DNA of interest (Pennisi 1999: 447). As mentioned, assembly is an indeterminate exercise that fits traces together based on probabilities, not certainties. Annotation may also produce errors. Pseudogenes may complicate gene-finding efforts, gene fragments may be contaminated, and complementary DNA (cDNA) sequences often contain repetitive sequences, which may cause incorrect genomic matches (Stein 2001: 495). Annotation often proceeds on the "draft" form of a genome, which may be problematic if a gene "runs off" the end of a contig; in such cases, the annotation protocol may assign the same gene to two different locations (Salzberg 2007: 102.3). Additionally, the accuracy of BLAST analysis depends on the quality of the annotation software as well as

[11] This section includes a basic overview of genome annotation. For a more comprehensive discussion, see Stein 2001. Genome annotation can be split into various subtasks. For example, some commentators distinguish between nucleotide-level, protein-level, and process-level annotation. Others distinguish between structural annotation, which encompasses identifying genomic elements, and functional annotation, which involves attaching biological information to genomic elements. Researchers utilize comprehensive data "warehouses" such as Microbiogenomics and querying modules such as GenoQuery to aid in functional annotation (Lemoine et al. 2008).

how up to date its reference databases of known sequences are (Salzberg 2007: 102.2). Given that annotation software finds genes based on probabilistic algorithms, even the best programs can produce errors, thus accentuating the need for manual oversight and verification by biologists with relevant expertise in the species in question. Not surprisingly, many researchers complain of flaws in GenBank annotations, and inaccuracies regarding gene structure and function "plague" GenBank and related databases (Bidartondo et al. 2008: 1616; Bridge et al. 2003: 44; Claverie 2000: 12; Lathe et al. 2008; Nilsson et al. 2006: e59; Pennisi 2008: 1598; Salzberg 2007: 102.1, 102.3).

This cursory examination of sequencing, assembly, and annotation sheds new light on the information at the heart of the genomic data commons. First, it illustrates that genomic information varies along a value-added continuum from raw sequence data to assemblies to annotated genomes identifying the location and function of specific genes. Second, it illustrates that these resources are not necessarily determinate or complete. Although lists of As, Ts, Cs, and Gs may suggest that genomic sequences are precise and apprehensible, many sequence assemblies and annotations are incorrect and incomplete. Put simply, "GenBank is full of mistakes" (Pennisi 1999: 448). Correcting and updating information, and thus adding value to existing data records, remain critical challenges to maintaining the integrity of the genomic data commons. The question of how to accomplish these functions, moreover, involves underappreciated governance challenges and sheds new light on the deep structure of the genomic commons, a topic to which this chapter now turns.

3.3 MAKING SENSE OF THE GENOMIC DATA COMMONS

Controlling access to data to safeguard scientific authorship and privacy is far from the only governance challenge facing the genomic data commons. The objective of correcting, updating, and adding greater intelligence to existing data, thus facilitating the transition from data to information to knowledge (Machlup 1983: 641), raises difficult questions of *who* should perform these functions and *how*. As we will see, a variety of players, including biocurators, data submitters, and the research community at large, have various strengths and weaknesses in managing this data. A background complicating factor, of course, is the sheer amount of data in the commons and its rapid expansion. To address these challenges, this section draws on the IAD framework derived from Ostrom and Madison et al. to examine additional dimensions of the genomic data commons, namely its degree of openness as well as the "rules in use" of community participants (Madison et al. 2010: 694–703; Van Overwalle 2014: 138). This analysis reveals that the genomic data commons features an ongoing tension between openness and control and has developed a variety of rules and institutions to govern knowledge-enhancing functions.

As Madison et al. (2010: 694) observe, "Commons regimes are defined both by the *degree of openness and control* that they exhibit with respect to contributors, users, and

https://doi.org/10.1017/9781316544587.004 Published online by Cambridge University Press

resources, and by the *assignment of control, or custody of the power to administer access*." Openness, of course, has many dimensions. Most accounts of the genomic data commons focus on *access* and *use* of data. Thus, data policies have evolved from a largely open system to one in which concerns over scientific credit and privacy have constrained researchers' ability to use data. This section, however, focuses on another type of openness, one that deals with the deeper architecture of the genomic data commons. In particular, it observes a deep tension between centralization and fragmentation of *control* over data, including the ability to modify existing data records. This tension is endemic to many commons-based production systems, such as Wikipedia. Regarding governance and rules in use, Madison et al. (2010: 701) emphasize the importance of "the identities and roles of . . . institutions and how their functions relate to the pool and its members." A variety of formal and informal institutions, from government agencies to loosely connected communities to temporary, project-specific groups, play important roles in correcting, updating, and annotating genomic data.

Although the open nature of the genomic data commons has greatly accelerated biomedical research, there are ways in which it is quite closed and centralized. This tension is evident in the very genesis of the Human Genome Project, in which decentralized teams of scientists working around the globe deposited nucleotide sequences in a highly centralized system of databases. The tension between centralization and fragmentation is particularly germane to efforts to render the genomic data commons more accurate, complete, and intelligible. These functions span basic (but important) biocuration functions such as standardizing data and nomenclature (Standardizing Data 2008) as well as higher valued-added processes such as genome annotation. This chapter now examines several approaches that are currently playing a prominent role in enhancing the accuracy, richness, and intelligibility of genomic data (Figure 3.1). It contrasts more centralized models of control, such as third-party biocuration and contributor-centric data management, with highly fragmented approaches, such as community-based wikification. Finally, it turns to specialized databases and genome browsers to illustrate how nonrivalry and replication may allow researchers to obtain the best of both approaches.

Third-Party Biocuration

At the centralized end of the spectrum is third-party biocuration (Howe et al. 2008: 47; Salimi and Vita 2006). Biocuration, which is performed both by professional scientists and automated processes, involves collecting, validating, annotating, and organizing biological information. Biocuration entails several functions, including "extracting, tagging with controlled vocabularies, and representing data from the literature" (Howe et al. 2008: 47). Such standardization renders information more coherent, intelligible, and useful given that "[t]he value of data is only as good as its annotation and accessibility: it must be properly curated and stored in machine-

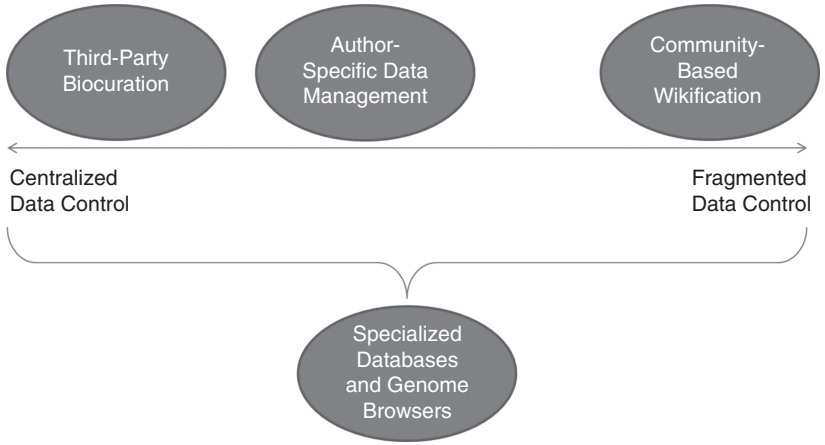

FIGURE 3.1 Continuum of data modification governance models.

readable form in public databases" (Standardizing Data 2008: 1123). Centralized, professional biocuration can help establish a single standard annotation for the genomes of model organisms, thus greatly facilitating research in those fields. Although biocuration generally focuses on such basic functions, it sometimes ventures into higher-level data analysis. For instance, one of the potential functions of centralized biocurators is to recompute genome annotations using the most up-to-date software and reference databases (Salzberg 2007: 102.4).

One set of stakeholders that is well positioned to perform biocuration is database operators themselves. For example, GenBank standardizes data formats (Standardizing Data 2008: 1123)[12] and assigns unique accession numbers to sequences and annotations upon intake. These accession numbers are shared with EMBL-Bank and DDBJ and represent "the most efficient and reliable way to cite a sequence record in publications" (Benson et al. 2014: D33–34). More ambitiously, NCBI maintains and actively curates the Reference Sequence (RefSeq) database, which links information in nucleotide and protein databases. RefSeq identifies a single "best" sequence for each protein-coding region of DNA and dynamically culls and updates the best information available on each segment of DNA (Pennisi 1999: 450). Among other functions, NCBI staff members coordinate with other database operators to maximize consistency between databases (Pruitt et al. 2012: D132). Additionally, journals also perform biocuration by requiring that contributing authors submit data to the appropriate public database in standardized formats (Walsh et al. 2003: 329). Interestingly, the operators of DNA databases originally helped convince journal editors to make electronic data submission a condition for publication (Strasser 2008: 538).

[12] For instance, NCBI checks for proper strandedness and chimeric sequences for submissions of prokaryotic ribosomal RNA data (Benson et al. 2014: D32).

https://doi.org/10.1017/9781316544587.004 Published online by Cambridge University Press

Although it adds significant value, third-party biocuration is somewhat limited in its ability to enhance the knowledge content of the genomic data commons. Almost by definition, biocuration focuses on lower-level functions such as standardization of data formats rather than more complex functions such as genome annotation. Furthermore, even if biocurators perform annotation, their distance from the original data and lack of familiarity with particular species may compromise outcomes. Biocuration may involve automated processes, and centralized recomputation of genomic annotations may overwrite manually annotated genomes that are even more accurate (Salzberg 2007: 102.4). Another limitation of third-party biocuration is inherent in its centralized nature: such efforts do not take advantage of the enormous community of users who can help correct and add value to existing data records. There has been increasing interest in harnessing commons-like mechanisms to take advantage of user-generated correction and annotation of genomic data.

3.3.1 *Contributor-Centric Data Management*

Beyond third-party biocuration, a slightly less centralized model of data management involves original data contributors exercising exclusive control over particular data records. Of course, in many ways, it makes logical sense for a data generator to take the lead in curating and annotating "her" data; she has intimate, perhaps tacit (Lee 2012), knowledge of the data and has personal and professional incentives to ensure its quality and completeness. Contributors can do much to enhance the value of existing data. They can label their data with standardized tags to facilitate subsequent studies[13] as well as recompute raw sequence data with newer software and more up-to-date reference databases to yield newer (and presumably more accurate) assemblies. Similarly, newer software and reference databases can produce more accurate annotations.

Notably, a centralized, contributor-centric approach to data modification is hardwired in the structure of GenBank. The rules and structure of that database establish that a scientist who contributes a sequence is the designated "author" of that entry, and only she can update the data record. Although data is free for *use* by members of the community (subject to constraints imposed by various policies discussed earlier), they may not modify it. The guiding metaphor for GenBank is that of a library: authors contribute books, which others can check out and access, but only original authors can revise the text of their volumes with new information. Interestingly, in a system built on openness and designed to avoid intellectual property claims, this centralized structure establishes something of a proprietary interest in the data on the part of the contributing researcher. Indeed, this sense of ownership and exclusive

[13] These tags include the National Center for Biotechnology Information (NCBI) Taxon IDs, the Gene Ontology (GO) IDs, and Enzyme Commission (EC) numbers (Howe et al. 2008: 48).

control over modifying data may help motivate widespread contributions to GenBank (Salzberg 2007: 102.3).

Although data contributors are well positioned to manage "their" data records, GenBank's centralized approach features some limitations. Empirically, data contributors seldom revise their records; in the busy life of academic science, researchers submitting sequence data often simply move on to other projects (Pennisi 1999: 448, 2008: 1598). Furthermore, although disclosing tacit information may greatly benefit the research community, data generators have little incentive to do so since such disclosure does not increase the value of existing data (Howe et al. 2008: 48). In a broader sense, strict control of data records by original contributors prevents other members of the community from updating and adding value to such data, a phenomenon to which this chapter now turns.

3.3.2 *Community-Based Wikification*

At the opposite end of the spectrum from third-party biocuration and contributor-centric data management are approaches to data management based on fragmentation of control. In particular, community-based wikification represents a very different model for correcting and annotating genomic data. As the name implies, this model involves wide swaths of independent researchers annotating and re-annotating existing sequences in a manner analogous to peer editing of Wikipedia (Pennisi 2008: 1598; Salzberg 2007: 102.5). Unlike centralized or contributor-centric data management systems, wikification is highly decentralized and requires data modification capabilities to be open to the broader user community (Madison et al. 2010: 694–96). Community-based wikification holds great promise for enhancing the knowledge content of genomic data. After all, subsequent users of GenBank "often learn more about the data than the initial depositors or curation staff" (Chu et al. 2008: 1289). As noted, although data generators are well positioned to correct and augment their data, they seldom do so. Community-based wikification enables huge numbers of subsequent data users to incorporate their experience and knowledge to improve data records. In many ways, wikification reflects Eric Raymond's observation that "[g]iven enough eyeballs, all bugs are shallow" (Raymond 1999: 30).

This distributed model of wikification has already facilitated valuable community-based genome annotation. For example, communal efforts to annotate genomic data for the Daphnia Genomics Consortium, the International Glossina Genomics Initiative, and PortEco (a hub for researchers studying *E. coli*) have greatly enhanced the value of these communal resources (Chu et al. 2008: 1289; Howe et al. 2008: 48). Indeed, community-based annotation has gained significant support (Hubbard and Birney 2000: 825), and researcher-edited wiki-type websites have proliferated (Waldrop 2008: 22). Commentators speculate that open-content databases modeled on Wikipedia might render GenBank, EMBL-Bank, and DDBJ obsolete in the same way that Wikipedia has done for the *Encyclopedia Britannica* (Strasser 2008: 538).

Not surprisingly, researchers have integrated wiki-based gene annotation directly into Wikipedia. The Gene Wiki project utilizes software to automatically generate thousands of Wikipedia stubs for human genes, thus inviting communal updating of such pages (Howe et al. 2008: 48; Huss et al. 2008; Waldrop 2008: 24). Its founders explicitly sought to exploit the "long tail" of incorporating small contributions from large numbers of contributors (Huss et al. 2008: 1398). Early results indicate that automated creation of "stubs" has "roughly doubled the amount of mammalian gene annotation activity in Wikipedia" (Huss et al. 2008: 1400). Although an important resource, the organizers of Gene Wiki recognize its uniqueness and limitations. Gene Wiki pages utilize an unstructured format of free text, images, and diagrams rather than the more structured (and easier-to-analyze) organization of gene portals. Gene Wiki organizers, moreover, recognize that it is not meant to substitute for more authoritative sources like gene portals and model organism databases, nor is it intended to be a source for citation in traditional peer-reviewed articles (Huss et al. 2008: 1401). Indeed, many researchers do not place citation-level confidence in wiki pages because of their open, decentralized nature.

In subtle ways, community-based wikification has even crystallized into more formal, government-run databases. For instance, NCBI maintains a Third Party Annotation (TPA) database, which includes sequences, assemblies, and annotations based on information already contained in an INSDC database. The TPA database represents a location where parties other than the contributor of the primary sequence record can submit data and analyses building off the primary source material (Benson et al. 2014: D36). In some ways, the TPA operates as a bridge between GenBank and RefSeq, allowing third parties to re-annotate sequences already existing in public databases. However, TPA records are distinct from original data records.

Though a potentially powerful source of annotation and curation, wikification has raised concerns about the degree to which it fragments control over genomic data. Indeed, one of the reasons why NCBI developed the TPA database is that it has resisted direct modification of GenBank records by third parties. Though scientists have argued for direct access to GenBank data (Salzberg 2007: 102.3), NCBI has consistently opposed wikification as undermining a structure where data records "belong" to the original contributor (Pennisi 2008: 1598). Along these lines, some worry that providing editorial access to GenBank records may "quickly destroy the archival function of GenBank, as original entries would be erased over time" (Salzberg 2007: 102.3). Interestingly, although GenBank maintains that only contributors themselves can modify sequence data or annotations, it provides an email address for feedback from the user community, and "all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank" (Benson et al. 2014: D34).

As with other instances of peer production, wikification of genome correction and annotation faces challenges of motivation and accuracy (Benkler 2006).

Community-based wikification relies on broad-based participation and voluntary contributions, which may not necessarily materialize (Waldrop 2008: 23; Mons et al. 2008: 89.8). For instance, researchers rarely submit records to the TPA database, presumably because of lack of incentive (Standardizing Data 2008: 1123). Recall that in the evolution of data release policies concerns over scientific credit motivated policy reforms to limit "scooping" by data users, thus maintaining incentives for original data generators to perform research and publish their findings. A similar motivational challenge faces wikification, and providing scientific credit for third-party annotation can accelerate such activity (Stein 2001: 502; Waldrop 2008: 25). Indeed, technical measures can address this concern. Thus, for instance, WikiGenes tracks – and allocates credit for – every contribution made (Waldrop 2008: 23, 25).[14] Even in a data commons, "attribution of scholarly contributions can be tracked and acknowledged" (It's Not About the Data 2012: 111). In a broader sense, cultural and institutional norms may need to change so that community-based contributions provide credit in university tenure and promotion decisions. In addition to challenges of motivation, the open and anonymous nature of wikification raises concerns about accuracy. Here, however, decentralization can not only help generate information but can also help ensure its quality: "[a] sizable population of readers then serves simultaneously as consumers, reviewers, and editors of content" (Huss et al. 2008: 1401). Thus, concerns that wikification leads to overlapping, duplicative work (Claverie 2000: 12) may be overstated, as such duplication and rechecking can ensure the accuracy of community-based annotations.

More subtly, certain aspects of genome sequencing, assembly, and annotation may more naturally lend themselves to wikification than others. The initial phases of annotation, which involve finding genes, mapping variations, and locating landmarks, lend themselves to brute force automation (Stein 2001: 500–01). Interpreting their functional roles, however, may be better suited for centralized biocuration by specialists in the field. Biocurators, perhaps working for model organism databases, may be well positioned to systematically catalog and classify genomes, correcting the mistakes of computational annotation (Stein 2001: 501). At this stage, community-based annotation may also be helpful, and some research communities have even organized "jamborees" to annotate parts of model organism genomes in a short period of time (Stein 2001: 501).

Ultimately, community-based wikification represents a powerful, decentralized model of knowledge production and verification that shows promise within genomics as well as other fields. For instance, SNPedia is a wiki resource wherein users contribute information relating to the functional consequences of human genetic information (Cariaso and Lennon 2011). Moving to proteomics, PDBWiki is a system for annotating protein structures deposited in the Protein Data Bank (PDB) (Stehr et al. 2010). This

[14] WikiGenes, which is separate from Gene Wiki, is an online platform by which scientists can contribute to particular research topics, including but not limited to genes.

database operates in parallel to central archives, and PDB is synchronized with PDBWiki every week (Stehr et al. 2010: 2). Similarly, WikiProteins is a "web-based, interactive and semantically supported workspace based on Wiki pages" that allows community members to annotate protein-related data (Mons et al. 2008: 89.2). Moving beyond annotation, wikification can help organize scientific knowledge more broadly. For example, the WikiProject Molecular and Cellular Biology initiative is a community of Wikipedia users committed to organizing and improving scientifically relevant articles; it hosts several subcommunities, including the RNA WikiProject (Daub et al. 2008: 1–2). At an even more distributed level, the general public could be involved in annotating genomic (and other scientific) data (Howe et al. 2008: 49).[15] Ultimately, "[c]ommunity data curation promises to be a solution to the problem of coping with the increasing size and complexity of biological data. The challenge is to make use of the 'wisdom of the many' without compromising the advantages of central, trusted and manually curated databases" (Stehr et al. 2010: 6).

3.3.3 *Specialized Databases and Genome Browsers*

A promising method for harmonizing the perceived trade-offs of centralized data management and community-based wikification is to pursue both simultaneously. This approach manifests in several forms. First, parallel databases have emerged alongside GenBank that draw from GenBank's data while also providing flexibility for community-based annotation and curation. In this fashion, GenBank remains the "archival" resource, but more specialized, value-added databases – which are also widely accessible within research communities – build off its contents.[16] Second, "genome browsers" have emerged that find, aggregate, filter, and present all information about a particular DNA sequence, thus offering the user both original, archival data as well as alternate views and additional layers of annotation. Perhaps the best approach to balancing centralization and fragmentation is to embrace both approaches through exploiting replication and the nonrivalry of data. In this manner, data represents an infrastructural resource that facilitates many downstream uses and is not itself subject to scarcity once created (Frischmann 2012: 62; OECD 2015: 178).

Based partly on the constraints of GenBank, various communities have developed specialized databases that replicate archival data while augmenting it with additional input and modification. Databases organized around model organisms have become particularly important (Salzberg 2007: 102.5). Resources such as FlyBase for the fruit fly and TAIR for *Arabidopsis* both incorporate GenBank data as well as clean up mistakes in sequence information (Pennisi 2008: 1599). The Daphnia

[15] However, some members of the scientific community oppose such a movement because of a lack of quality control and standardization (interview with Dr. Janice Ahn, March 3, 2015).

[16] It should be noted that Entrez, the retrieval system for accessing data in GenBank, links to other informational sources, such as academic articles in PubMed and PubMed Central that discuss particular sequences (Benson et al. 2014: D36).

Genomics Consortium performs a similar function for its community members (Lathe et al. 2008). Notably, many of these specialized databases incorporate wiki-based annotation (Galperin and Fernandez-Suarez 2012: D4). Hundreds of species- and taxa-specific genome databases exist that integrate archival and value-added data to serve specific research needs (Lathe et al. 2008). Even NCBI, which has resisted community-based annotation and modification of data records in GenBank, has embraced a system of parallel databases. As mentioned, NCBI maintains both GenBank, an "archival" resource, and RefSeq, an actively curated database that combines original data with value-added information (Pennisi 2008: 1599). Ultimately, an "ecosystem of databases" has emerged with replication, synchronization, and cross-linking of data (Galperin and Fernandez-Suarez 2012: D6). These resources take advantage of the inherent nonrivalry of information by copying and modifying archival data while leaving the original data unchanged.

Another resource that exploits accretive and overlapping information is genome browsers, sometimes referred to as gene portals (Furey 2006: 266). Such browsers "repackage genome and gene annotation data sets from GenBank and other subject-specific databases to provide a genomic context for individual genome features, such as genes or disease loci" (Lathe et al. 2008). In essence, they collect, aggregate, and display desired data from multiple sources, including central databases such as GenBank, wikified community pages, and specialized databases (Furey 2006: 266; Hubbard and Birney 2000: 825). This information often integrates value-added information, such as the location of clones from bacterial artificial chromosomes (BAC) libraries, expressed sequence tags (ESTs), sequence-tagged site (STS) markers from genetic maps, and boundaries of cytogenetic bands, all of which aid in mapping genomes (Furey 2006: 266). Among other virtues, genome browsers allow researchers to select only data sources that they seek to view (Hubbard and Birney 2000: 825). Such data aggregation and visualization can greatly enhance data analysis (Cline and Kent 2009: 153).

Just as genome browsers provide a multiplicity of views, there is a multiplicity of genome browsers that differ in their presentation, content, and functionality (Furey 2006: 266–69). Prominent examples include UCSC's genome browser, EBI's Ensembl, and NCBI's MapViewer (Lathe et al. 2008). The multitude of options allows researchers to pick the particular tool that best suits their needs (Furey 2006: 269). There are now more than 3000 distinct genomic resources, tools, and databases publicly available on the Internet (Lathe 2008).

Contrary to conventional views, genome browsers more realistically depict the messiness and indeterminateness of genome sequencing, assembly, and annotation. A common format is to present several parallel "tracks" showing various depictions of the same nucleotide sequence from different sources, often with different standards of evidence and confidence intervals (Cline and Kent 2009: 153). As Cline and Kent (2009: 153) observe, "Virtually any genomic data can be erroneous, and one should be wary of data suggested by only a single observation." Genome browsers

more faithfully represent the indeterminacy of genomics, aggregating multiple views to aid the researcher in her particular pursuit. While the guiding metaphor for GenBank may be a library that lends out archival resources, the guiding metaphor for genome browsers may be something like BitTorrent, wherein a user, perhaps with the aid of a central registry, can obtain, aggregate, and filter any and all information related to a particular genome sequence.[17] In some ways, this system resembles whole genome shotgun sequencing as a mechanism for building genomic knowledge. Decentralized, massively parallel efforts to process data – including community-based wikification – are aggregated and assembled, thus rendering the genomes of various organisms more intelligible.

A system of "replicative" specialized databases and genome browsers seeks to combine the virtues of centralization and fragmentation. They maintain the archival status of contributor-centric and professionally curated databases such as GenBank, thus shoring up authorial incentives to submit to such databases because other members of the community cannot modify data records directly. However, these resources also take advantage of wikification by giving users access to both original data sets and community-based annotation. In so doing, they exploit the inherent nonrivalry of information to powerful effect; specialized databases can "consume" and modify all of the data in GenBank without depleting or altering the original data records at all.

Drawing on the theme of fragmentation versus centralization, however, this multiplicity of data resources gives rise to a strong need for standardization. The specialized nature of community-specific databases is both a feature and a bug. Because such databases are tailored to the unique nomenclature and practices of biological subfields, they may represent the same information differently in different contexts. As a result of fragmentation, organism-specific communities and databases often utilize their own unique nomenclature for genes and their products (The Gene Ontology Consortium 2000: 25). Data is not completely consistent between species and research projects, thus making cross-species comparisons difficult (Lathe et al. 2008). This proliferation of alternate names undermines database interoperability and complicates the process of finding related genes and gene products across different species (The Gene Ontology Consortium 2000: 25). Ironically, while standardization within communities has helped make information more internally coherent (Field and Sansone 2006: 85), it has exacerbated incommensurabilities between communities, as one group's standards may be incompatible with those of another (Field and Sansone 2006: 90).

Genomics thus faces the significant challenge of adopting a standardized classification system that can account for myriad biological functions across diverse species while maintaining the particularity to distinguish subtle nuances within a given

[17] I am indebted to Jonathan Eisen for this metaphor (interview with Dr. Jonathan Eisen, UC Davis Genome Center, March 17, 2014).

species (Stein 2001: 500). Various user communities have led standardization efforts to help bridge these gaps, thus bringing some degree of centralization to the fragmented realm of community-specific nomenclature, practices, and databases. In a sense, a metacommons has emerged to mediate among different communities utilizing different standards and nomenclatures. For example, the Generic Model Organism Database (GMOD) has collaboratively developed a standardized set of tools and database schema (Lathe et al. 2008).

Perhaps most ambitiously, several model organism databases formed the Gene Ontology (GO) consortium to standardize annotation conventions across all species (Contreras 2013: 116).[18] The GO aims to "produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism" (The Gene Ontology Consortium 2000: 26). Within this structure, "[d]ata can be annotated to varying levels depending on the amount and completeness of available information" (The Gene Ontology Consortium 2000: 27). While GO may sacrifice some species-specific conventions, it ultimately serves the broader purpose of standardizing the accumulation of knowledge and facilitating cross-species comparisons. Particularly amid such multiplicity, standardization and centralization remain highly valuable.

At a more fundamental level, the multiplicity of databases and the power of genome browsers to aggregate all relevant information on a particular nucleotide sequence hold important epistemological implications for genomic knowledge. Neat rows of As, Ts, Cs, and Gs, suggest that the genome of any organism is completely knowable, an objective fact just waiting to be discovered. However, genome sequencing, assembly, and annotation are imprecise sciences, and different methodologies, technologies, equipment, and techniques can yield different results. Even the same or related gene may be represented quite differently in different databases or systems of nomenclature. Genome browsers are sensitive to this multiplicity by providing a wide array of views of the "same" physical phenomenon. And, of course, human subjectivity informs both the various processes that create data as well as the perspectives that interpret it. Just as Thomas Kuhn postulated that scientific theory reflects more social consensus than objective fact (Kuhn 1962), perhaps "true" genomic knowledge only arises from gathering, comparing, and reconciling a multiplicity of competing perspectives. The genomic data commons certainly *holds* knowledge in the sense that it encompasses an enormous repository of data. But it also *produces* knowledge by allowing social consensus to coalesce around particular epistemological paradigms.

[18] As Field and Sansone describe, "An ontology is an explicit formal representation of the knowledge in a subject area, which includes controlled vocabularies referring to the concepts and logical statements that describe what the concepts are and how they can or cannot be related to each other" (Field and Sansone 2006: 87).

3.4 BROADER IMPLICATIONS FOR THE GENOMIC DATA COMMONS AND COMMONS IN GENERAL

In a variety of ways, this chapter sheds new light on the genomic data commons. Applying the IAD framework, it reveals the indeterminate nature of the shared and pooled resource at the heart of the commons – genomic data and related information. Popular views of gene sequencing envision an orderly and determinate process of discerning the As, Ts, Cs, and Gs that make up an organism's genome. However, genome sequencing, assembly, and annotation are probabilistic sciences, and data records in GenBank are rife with errors and incompleteness. These technological limitations exacerbate tensions between centralization and fragmentation of control over data. Data contributors and professional biocurators can add significant value to existing data, but their efforts are necessarily limited. The research community represents an enormous resource for annotating and updating genomic data, but disaggregated control over information threatens the "purity" and provenance of archival data records. In various contexts, the genomic data commons features centralized control over data modification, wide community participation in data modification, and information replication as a structural strategy for harmonizing these competing approaches.

More subtly, this chapter also sheds new light on the concept of the commons itself. The commons is often understood as a repository of ontologically determinate resources. Although the commons may grow or shrink in size, each individual resource unit has a particular identity that is largely fixed. However, at least in a knowledge commons, members of the community can change the fundamental nature of these constituent elements, as when a user corrects or annotates an existing data record. Thus, the community is responsible for not only provisioning and extracting resources but also fundamentally changing the character of those resources along the way. Far from being a passive repository, the commons is a teeming, dynamic entity continually subject to human intervention and manipulation. Derived from the Human Genome Project, the genomic data commons is an indelibly *human* commons not only in the DNA sequences that constitute its subject matter but also in the social and communal processes that modify data and produce knowledge.

This chapter also reveals that the genomic data commons is both less and more of a commons than previously appreciated. On one level, the genomic data commons certainly reflects a social construct where wide access to pooled, shared resources – sequence data and related information – has greatly accelerated biomedical research. Through a complex ecosystem of public funding, agency policy, enabling infrastructure, and communal norms, a practice has emerged in which researchers routinely submit highly valuable sequence and related information to public databases well before publication. As scholars have fruitfully explored, the genomic data commons has evolved from a largely open structure to a more complex governance regime that

constrains the use of genomic data to preserve scientific credit and research subject privacy. Notwithstanding these complexities, the genomic data commons reflects the power and potential for widely shared resources to advance productivity.

Upon closer inspection, however, some aspects of the genomic data commons are highly centralized and do not function as a commons at all. Although the Human Genome Project involved large numbers of researchers around the world, it required data release into a synchronized network of three centralized databases. While GenBank receives data from many parties and provides access to the entire scientific community, certain aspects of data control are highly centralized; it is designed so that only original data contributors can modify "their" data records. NCBI has consistently resisted community-based wikification of GenBank, thus preserving the archival nature of data records rather than making them truly open. The evolution of data release policies has limited how researchers can use sequence data, and the structure of GenBank flatly prohibits directly modifying such data, even when it is incorrect. In this fashion, the structure of GenBank propagates a property-like interest on the part of data generators, which seems contrary to the character of a true commons.

At the same time, however, data users have built multiple commons on top of a commons with a proliferation of peer-based wikification, specialized databases, and genome browsers. Harnessing the power of distributed, parallel processing, researchers around the world are adding greater intelligence to existing data records in the form of wiki-based correction, updating, and annotation, and they are making such value-added knowledge widely available to others. Model organism communities are creating their own commons, crafting openly accessible, specialized databases to serve their needs. And innovative users are creating genome browsers and freely distributing them to aid in aggregating and representing enormous amounts of data. This study of the genomic data commons reveals that it actually represents a collection of multiple, overlapping commons operating at several levels.

Along related lines, these multiple commons also reflect a high degree of user innovation (Strandburg 2008, 2009). Users have coordinated formal wikification initiatives such as Gene Wiki as well as more informal annotation "jamborees." Model organism communities have developed specialized databases, combining archival as well as more recent data to advance their research. In doing so, they have created a communal resource with significant spillovers. Similarly, users have developed genome browsers to serve their own needs for enhanced data visualization and analysis, which are now available for all to use. Such user innovations have significantly enhanced the value of the underlying sequence information in the genomic data commons.

Finally, this chapter reveals the enduring importance of standardization and centralization to unleash the creative power of a commons. The commons is lauded for its openness and fluidity, but an effective commons is not a free-for-all where anything goes. Just as a physical commons may benefit from centralized governance

(Rose 1986: 719), the genomic data commons requires some degree of standardization and centralization to function effectively. Community-based wikification represents a powerful resource for annotating genomes, but it requires centralized coordination. Furthermore, the proliferation of model organism databases provides a high degree of specialization for particular communities, but it creates a strong need for an overarching, standardized system for referring to genomic elements to facilitate cross-species comparisons and interoperability. Balance between the competing forces of centralization and fragmentation is critical to effective operation of this commons.

CONCLUSION

The genomic data commons has rightly attracted significant attention both as a formidable resource for advancing biomedical research as well as an exemplar of commons-based management systems. This chapter has built upon prior studies by further applying the IAD framework to assess underappreciated facets of this knowledge commons. In particular, it has focused on annotating, curating, and rendering intelligible the vast amounts of genomic information produced around the world. In some ways, the genomic data commons is highly centralized, with institutionalized data curation and exclusive authorial rights in GenBank data records. In other ways, the genomic data commons is actually a metacommons in which communal annotation and curation help enhance the value of communally accessible data. In an innovative fashion, users have developed specialized databases and genome browsers to draw from and build upon centralized resources, thus exploiting the inherent nonrivalry of information to render genomic data more intelligible. These dynamics have enormous practical impact, for it is only by transforming data to information to knowledge that the promise of the Human Genome Project and its progeny will be fully realized. More broadly, these dynamics reflect that massively fragmented productivity must be subject to some level of centralized coordination to maximize the creative potential of a commons.

REFERENCES

Benkler, Yochai, *The Wealth of Networks* (Yale University Press 2006).
Benson, Dennis A. et al., GenBank, 42 *Nucleic Acids Res.* D32 (2014).
Bidartondo, M. I. et al., Preserving Accuracy in GenBank, 319 *Science* 1616 (2008).
Bridge, Paul D. et al., On the Unreliability of Published DNA Sequences, 160 *New Phytologist* 43 (2003).
Cariaso, Michael, and Greg Lennon, SNPedia: A Wiki Supporting Personal Genome Annotation, Interpretation, and Analysis, 40 *Nucleic Acids Res.* D1308 (2011).
Chu, James C. et al., The Emerging World of Wikis, 320 *Science* 1289 (2008).

Claverie, Jean-Michel, Do We Need a Huge New Centre to Annotate the Human Genome? 403 *Nature* 12 (2000).
Cline, Melissa S. and W. James Kent, Understanding Genome Browsing, 2 *Nature Biotechnology* 153 (2009).
Contreras, Jorge L., Prepublication Data Release, Latency, and Genome Commons, 329 *Science* 393 (2010).
Contreras, Jorge L., Bermuda's Legacy: Policy, Patents and the Design of the Genome Commons, 12 *Minnesota J.L. Sci. & Tech.* 1 (2011).
Contreras, Jorge L., Technical Standards and Bioinformatics, in *Bioinformatics Law: Legal Issues for Computational Biology in the Post-Genome Era* (Jorge L. Contreras and A. James Cuticchia eds., ABA Book Publishing 2013).
Contreras, Jorge L., Constructing the Genome Commons, *in Governing Knowledge Commons* (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).
Dagan, Hanoch and Michael A. Heller, The Liberal Commons, 110 *Yale L.J.* 549 (2001).
Daub, Jennifer et al., The RNA WikiProject: Community Annotation of RNA Families, 14 *RNA* 1 (2008).
Dennis, Carina, Draft Guidelines Ease Restrictions on Use of Genome Sequence Data, 421 *Nature* 877 (2003).
Dept. of Energy, Summary of Principles Agreed Upon at the First International Strategy Meeting on Human Genome Sequencing (Bermuda, 25–28 February 1996) as Reported by HUGO, http://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml
Eisenberg, Rebecca S., Genomics in the Public Domain: Strategy and Policy, 1 *Nature Rev. Genetics* 70 (2000).
Eisenberg, Rebecca S., Patents and Data-Sharing in Public Science, 15 *Indus. & Corp. Change* 1013 (2006).
Field, Dawn and Susanna-Assunta Sansone, A Special Issue on Data Standards, 10 *OMICS* 84 (2006).
Frischmann, Brett M., *Infrastructure: The Social Value of Shared Resources* (Oxford University Press 2012).
Frischmann, Brett M. et al., Governing Knowledge Commons, in *Governing Knowledge Commons* (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).
Furey, Terrence S., Comparison of Human (and Other) Genome Browsers, 2 *Hum. Genomics* 266 (2006).
Galperin, Michael Y. and Xose M. Fernandez-Suarez, The 2012 Nucleic Acids Research Database Issue and the Online Molecular Database Collection, 40 *Nucleic Acids Res.* D1 (2012).
Genetic Assoc. Info. Network (GAIN), Data Use Certification Agreement (March 1, 2010), https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?view_pdf&stacc=phs000016.v1.p1.
Howe, Doug et al., The Future of Biocuration, 455 *Nature* 47 (2008).

Hubbard, Tim and Ewan Birney, Open Annotation Offers a Democratic Solution to Genome Sequencing, 403 *Nature* 825 (2000).
Huss, Jon W. III et al., A Gene Wiki for Community Annotation of Gene Function, 6 *PLoS Biology* 1398 (2008).
It's Not About the Data, 2 *Nature Genetics* 111 (2012).
Kaye, Jane, The Tension between Data Sharing and the Protection of Privacy in Genomics Research, 13 *Ann. Rev. Genomics & Hum. Genetics* 415 (2012).
Kuhn, Thomas, *The Structure of Scientific Revolutions* (University of Chicago Press 1962).
Lathe, Warren C. III et al., Genomic Data Resources: Challenges and Promises, 2 *Nature Educ.* 1 (2008).
Lee, Peter, Transcending the Tacit Dimension, 100 *California L. Rev.* 1503 (2012).
Lemoine, Frederic et al., GenoQuery: A New Querying Module for Functional Annotation in a Genomic Warehouse, 24 *Bioinformatics* i322 (2008).
Machlup, Fritz, Semantic Quirks in Studies of Information, in *The Study of Information: Interdisciplinary Messages* (Fritz Machlup and Una Mansfield eds., Wiley 1983).
Madison, Michael J. et al., Constructing Commons in the Cultural Environment, 95 *Cornell L. Rev.* 657 (2010).
Marshall, Eliot, DNA Sequencer Protests Being Scooped with His Own Data, 295 *Science* 1206 (2002).
McGuire, Amy L. et al., Ethical, Legal, and Social Considerations in Conducting the Human Microbiome Project, 18 *Genome Res.* 1861 (2008).
Mons, Barend et al., Calling on a Million Minds for Community Annotation in Wiki Proteins, 9 *Genome Biology* R. 89 (2008).
Nat'l Human Genome Res. Inst., Current NHGRI Policy for Release and Data Deposit of Sequence Data (March 7, 1997), https://www.genome.gov/10000910#old.
Nat'l Human Genome Res. Inst., NHGRI Policy for Release and Database Deposition of Sequence Data (Dec. 21, 2000), www.genome.gov/pfv.cfm?pageID=10000910
Nat'l Human Genome Res. Inst., Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects (Feb. 2003), www.genome.gov/pfv.cfm?pageID=10506537
Nat'l Inst. of Health, Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS) (Aug. 28, 2007), http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07–088.html
Nat'l Inst. of Health, HMP Data Release and Resource Sharing Guidelines for Human Microbiome Project Data Production Grants (last updated Aug 27, 2013), http://commonfund.nih.gov/hmp/datareleaseguidelines
Nat'l Inst. of Health, Final NIH Genomic Data Sharing Policy, 79 *Fed. Reg.* 51345 (Aug. 28, 2014).
Nat'l Inst. of Health and Dept. of Energy, Guidelines for Access to Mapping and Sequence Data and Material Resources (1991), www.genome.gov/10000925.
Nat'l Res. Council, Mapping and Sequencing the Human Genome (1988).

Nelson, F. Kenneth et al., Introduction and Historical Overview of DNA Sequencing, *Current Protocols in Molecular Biology*, 7.0.1 (Oct. 2011).
Nilsson, R. Henrik et al., Taxonomic Reliability of DNA Sequences in Public Sequence Databases: A Fungal Perspective, 1 *PLoS One* e59 (2006).
OECD, Data-Driven Innovation: Big Data for Growth and Well-Being (2015).
Ostrom, Elinor and Charlotte Hess, A Framework for Analyzing the Knowledge Commons, in *Understanding Knowledge as a Commons: From Theory to Practice* (Charlotte Hess and Elinor Ostrom eds., MIT Press 2007).
Pennisi, Elizabeth, Keeping Genome Databases Clean and Up to Date, 289 *Science* 447 (1999).
Pennisi, Elizabeth, Proposal to "Wikify" GenBank Meets Stiff Resistance, 319 *Science* 1598 (2008).
Pennisi, Elizabeth, Group Calls for Rapid Release of More Genomics Data, 324 *Science* 1000 (2009).
Price, W. Nicholson II, Unblocked Future: Why Gene Patents Won't Hinder Whole Genome Sequencing and Personalized Medicine, 33 *Cardozo L. Rev.* 1601 (2012).
Pruitt, Kim D. et al., NCBI Reference Sequences (RefSeq): Current Status, New Features and Genome Annotation Policy, 40 *Nucleic Acids Res.* D130 (2012).
Raymond, Eric S., *The Cathedral and the Bazaar* (O'Reilly Media, Inc. 1999).
Rose, Carol, The Comedy of the Commons: Custom, Commerce, and Inherently Public Property, 53 *U. Chicago L. Rev.* 711 (1986).
Salimi, Nima, and Randi Vita, The Biocurator: Connecting and Enhancing Scientific Data, 2 *PLoS Computational Biology* 1190 (2006).
Salzberg, Steven L., Genome Re-annotation: A Wiki Solution, 8 *Genome Biology* 102 (2007).
Sanger, F. et al., DNA Sequencing with Chain-Terminating Inhibitors, 74 *Proc. Nat'l Acad. Sci.* 5463 (1977).
Smith, Henry E., Semicommon Property Rights and Scattering in the Open Fields, 29 *J. Leg. Stud.* 131 (2000).
Smith, Henry E., Exclusion versus Governance: Two Strategies for Delineating Property Rights, 31 *J. Leg. Stud.* S453 (2002).
Standardizing Data (Editorial), 10 *Nature Cell Biology* 1123 (2008).
Stehr, Henning et al., PDBWiki: Added Value through Community Annotation of the Protein Data Bank, 1 *Database* (2010).
Stein, Lincoln, Genome Annotation: From Sequence to Biology, 2 *Nature Rev. Genetics* 493 (2001).
Strandburg, Katherine J., Users as Innovators: Implications for Patent Doctrine, 79 *U. Colo. L. Rev.* 467 (2008).
Strandburg, Katherine J., User Innovator Community Norms: At the Boundary between Academic and Industry Research, 77 *Fordham L. Rev.* 2237 (2009).
Strandburg, Katherine J. et al., The Rare Diseases Clinical Research Network and the Urea Cycle Disorders Consortium as Nested Knowledge Commons, in *Governing Knowledge Commons* (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).

Strasser, Bruno J., GenBank – Natural History in the 21st Century? 322 *Science* 537 (2008).

The 1000 Genomes Project Consortium, A Map of Human Genome Variation from Population-Scale Sequencing, 467 *Nature* 1061 (2010).

The Gene Ontology Consortium, Gene Ontology: Tool for the Unification of Biology, 24 *Nature Genetics* 24 (2000).

The Wellcome Trust, Sharing Data from Large-Scale Biological Research Projects: A System of Tripartite Responsibility 2 (2003), www.genome.gov/pages/research/wellcomereport0303.pdf

Van Overwalle, Geertrui, Governing Genomic Data: Plea for an "Open Commons," in *Governing Knowledge Commons* (Brett M. Frischmann, Michael J. Madison, and Katherine J. Strandburg eds., Oxford University Press 2014).

von Hippel, Eric, The Dominant Role of Users in the Scientific Instrument Innovation Process, 5 *Res. Policy* 212 (1976).

Waldrop, Mitch, Wikiomics, 455 *Nature* 22 (2008).

Walsh, John P. et al., Effects of Research Tool Patents and Licensing on Biomedical Innovation, in *Patents in the Knowledge-Based Economy* (Wesley M. Cohen and Stephen A. Merrill eds., National Academies Press 2003).