J Technol Transfer (2007) 32:133–156
DOI 10.1007/s10961-006-9016-9

The science commons in health research: structure, function, and value

Robert Cook-Deegan

Published online: 7 December 2006

Abstract The ''science commons,'' knowledge that is widely accessible at low or no cost, is a uniquely important input to scientific advance and cumulative technological innovation. It is primarily, although not exclusively, funded by government and nonprofit sources. Much of it is produced at academic research centers, although some academic science is proprietary and some privately funded R&D enters the science commons. Science in general aspires to Mertonian norms of openness, universality, objectivity, and critical inquiry. The science commons diverges from proprietary science primarily in being open and being very broadly available. These features make the science commons particularly valuable for advancing knowledge, for training innovators who will ultimately work in both public and private sectors, and in providing a common stock of knowledge upon which all players—both public and private—can draw readily. Open science plays two important roles that proprietary R&D cannot: it enables practical benefits even in the absence of profitable markets for goods and services, and its lays a shared foundation for subsequent private R&D. The history of genomics in the period 1992–2004, covering two periods when genomic startup firms attracted significant private R&D investment, illustrates these features of how a science commons contributes value. Commercial interest in genomics was intense during this period. Fierce competition between private sector and public sector genomics programs was highly visible. Seemingly anomalous behavior, such as private firms funding ''open science,'' can be explained by unusual business dynamics between established firms wanting to preserve a robust science commons to prevent startup firms from limiting established firms' freedom to operate. Deliberate policies to create and protect a large science commons were pursued by nonprofit and government funders of genomics research, such as the Wellcome Trust and National Institutes of Health. These policies were crucial to keeping genomic data and research tools widely available at low cost.

R. Cook-Deegan (✉)
Center for Genome Ethics, Law & Policy, Institute for Genome Sciences & Policy and Sanford Institute of Public Policy and Duke Medical School, Duke University, 242 North Building, Durham, NC 27708-0141, USA
e-mail: gelp@duke.edu

Keywords Patents · Genomics · Public domain · Open science · Intellectual property · Innovation

JEL Classifications 031 · 032 · 034 · 038

1 Introduction

When Robert Merton wrote about the sociology of science in the 1970s, the central task at hand was explaining how a set of social norms and practices yielded knowledge—what was different about science compared to the humanities and the professions (Merton, 1973). John Ziman and others addressed what makes the methods of science produce ''reliable knowledge'' (Ziman, 1978). This paper addresses a related, but somewhat different aspect of science—how reliable knowledge can be turned into social benefit, using genomics as a case in point. The value of a science commons—a pool of knowledge that is widely available at little or no cost—is the central focus. The zone of intersection between reliable knowledge and useful knowledge falls squarely into what the late Donald Stokes described as ''Pasteur's Quadrant,'' (1997) where research results both contribute insight into the workings of nature and at the same time find practical application.

The value of having knowledge widely and freely (or almost freely) available is particularly salient in Pasteur's Quadrant. Knowledge is more likely to advance, and to be applied, if it is available at little or no expense to a broad array of scientists and innovators. These features of network efficiency are well known in software and other fields characterized by widely distributed cumulative innovation under mantras such as ''to many eyes, every bug is shallow,'' and theoretically described by Benckler (2002).

I shamelessly steal the term ''science commons'' from the new organization of that name that has spun out of the Creative Commons movement. Science Commons is dedicated to ''making it easier for scientists, universities, and industries to use literature, data, and other scientific intellectual property and to share their knowledge with others. Science Commons works within current copyright and patent law to promote legal and technical mechanisms that remove barriers to sharing'' (Science Commons, 2005). (While I endorse their mission, they may not endorse my agenda or analysis. I have no direct connection to the organization, and do not speak for it.)

The main approach in what follows is historical, using background on how the science commons functioned in genomics to illustrate the role of a commons in general. Genomics will be the main topic, occasionally straying into collateral fields of biomedical research such as bioinformatics or molecular and cellular biology when they provide better examples.

There is some fuzziness around the edges of what constitutes a science commons, and how it relates to ''the public domain.'' There can be variants of many terms marching under the banner of open science or public research. ''Open access,'' for example, can mean free access to view information, but not necessarily freedom to use it in all ways without restriction. The information in patents, for example, is openly available, but users may need to get permission or pay fees to use a patented invention (including some basic methods used in science). To some, open science means no one can fence it in. Access to information, say through ''viral'' licensing or copyleft, may be

conditioned on agreeing not to restrict subsequent users. Information may also simply be put into the public domain for any and all subsequent uses, by deposit at a freely available public database, for example. I focus on this last meaning, with information available to all at low or no cost. But again, this may not mean completely unfettered use, as sequence information in GenBank may be covered by patent claims. Jensen and Murray, for example, noted that more than 4,000 human DNA sequences are subject to claims in US patents (Jensen & Murray, 2005). Sometimes there are restrictions on use of information and materials in the science commons, but those restrictions must also impose low or no costs to subsequent users, or else that information has left the science commons (e.g., through subsequent patenting, copyright, or database protection). There is no bright line dividing the science commons from proprietary R&D, and indeed in the case of some sequences, materials, and methods in molecular biology, the expense associated with use of genomic information may depend more on licensing terms and practices than what is or is not patented—and one person's ''reasonable terms'' may be out of reach for some users.

The world of genomics is not a simple world. Some of the data shown below, for example, are drawn from a free public database, the DNA Patent Database at Georgetown University (DNA Patent Database, 2005). That database is, in turn, drawn from the freely available US Patent and Trademark Database (USPTO Database, 2005). But the search engine and database used to generate the DNA Patent Database are derived from an intermediate subscription database that is not widely or freely available, but available to subscribers at several thousand dollars a year, through the Delphion patent database (Delphion Database, 2005). Duke pays a subscription to The Thomson Corporation for its use. This is a major tool for our research, and for us the subscription cost is balanced by ease of use and reliability. We pay for Delphion for its special features (such as corporate trees that track ownership of patents) and because its search results have proven more reliable than several alternatives, including the USPTO's own computers and software available in northern Virginia. We are happy to pay for a proprietary database because it helps us do our work and the price is reasonable, within reach of our nonprofit institution. Delphion does not restrict our use, and does not prevent our creating a free public database. I raise the example of a pay-for-use database sandwiched between two public resource databases to hint that the story will get complicated, and to signal early that this is not a diatribe against for-profit intrusions into research. This is important because many of the points that come out later will seem unfriendly to purveyors of databases. Those objections are not deep-seated rejections of capitalism in science, but rather pragmatic judgments about adverse effects of particular policies.

Innovation is exquisitely sensitive to policies of many kinds. Innovation depends on how much information is produced as well as how widely and easily it is shared. Funding of R&D is a major determinant of how much research is conducted, and thereby how much information is created. Policies governing the science commons—or alternative, more restricted informational spaces—determine how widely and quickly information and materials are distributed. My purpose here is to highlight why the science commons matters. Some reasons are obvious, but some are not so obvious, and some are even counterintuitive.One final conceptual point will be helpful to flag before proceeding into the narrative. There is extensive overlap between ''academic health research'' and the science commons in molecular biology. Academic science is important in many fields, not just the life sciences. In all lines of

scientific and technical work, universities and nonprofit research institutions and government laboratories (''academic research institutions'') play key roles. Everyone is trained in academe, not just academic scientists, but also those working in industry. And within industry, academic training is not just for those doing R&D, but also managers and professionals. Academe is also one place where the norms of Mertonian science have real traction, where the norms of openness, community, mutual criticism, and fair allocation of credit are supposed to be respected, at least as an ideal. In some circumstances, however, academic science is done under strictures of secrecy, or results are made available only at great cost or encumbered by restrictions on use. Such academic science is not part of a science commons. Great science goes on in industry, including or even particularly in the life sciences, but no one expects the norms of openness to prevail in industrial R&D, even if in some circumstances at some times scientists in companies publish in the open literature, present their findings at open scientific conferences, make materials freely available, and contribute data to public databases. When industrial R&D is widely shared openly, results flowing from industrial R&D can become part of the science commons, and there are several instances of this in the stories to follow. In sum, most academic research contributes to the science commons, and some industrial R&D also does so. Most industrial R&D is proprietary, as is some academic research.

The science commons thus does not reduce to academic research. It remains true nonetheless, that most of the science commons—at least in the life sciences—is based on academic research funded by government and nonprofit organizations, and most academic research probably enlarges the science commons, although to my knowledge no one has quantitatively assessed what fraction of the research funded by government and nonprofit organizations remains in the science commons. Policies put in place over the past three decades have raised concerns about how big the science commons will be, and in particular, whether and to what degree government and nonprofit funders and academic research institutions will maintain it. Richard Nelson of Columbia University, in particular, has expressed concerns about intrusions on open science, based on his decades of studying the innovation process as an economist (Nelson, 2006).

2 Genomics: public and private science in a fishbowl

Genomics became the grounds for a vigorous, sometimes even vicious, fight over what should or should not be in the public domain, and under what conditions. Many of the fights were over preserving the science commons. How much genomic data should be in the science commons has been a matter of explicit policy-making in government, nonprofits, academic institutions, and private firms since 1992 or 1993, when the commercial promise of genomics became apparent, and private funding for genomics in for-profit companies began to accelerate.

Several features of genomics make it an interesting field to study as an instance of the science commons. It is clearly derived from a scientific project that was initially conceived as a public works project—to construct maps and derive a reference sequence of the human genome and other genomes. The original intent of the Human Genome Project was to produce information and tools to make that information useful and valuable. Some commercial uses were foreseen from the beginning, but the main focus was on producing public data of permanent scientific value.

The genome revolution began in the mid-1980s. It caught a wave of enthusiasm for ''the new biotechnology'' that had become both scientifically hot and also a darling on Wall Street. Cetus was founded in 1971 and turned to recombinant DNA techniques soon after they were discovered (Stanley Cohen, co-inventor of recombinant DNA, joined the Cetus Board in 1975). Genentech was founded in 1976. Those companies went public with high-profile stock offerings in 1980 and 1981, raising sums that startled the markets (Smith Hughes, 2001).

The origins of the Human Genome Project were not in commercial biotechnology, however, but in publicly funded science. The ideas behind the Human Genome Project began to appear in 1985, while the embers of biotechnology were still warm but too distant from this particular part of molecular biology to catch fire. (And not for want of trying. Walter Gilbert tried to start Genome Corp. in 1987, for example, and had to resign from a National Research Council study as a consequence.) Scientists conceived a grand idea and focused on the scientific value of having a reference human genomic sequence (Cook-Deegan, 1994).[1] Commercial interest lagged for several years, until in 1991 a conflict over patenting short sequence tags derived from human genes blew up into a major controversy, and created commercial interest in human genomic sequencing.

J. Craig Venter, a scientist in NIH's intramural (government laboratory) research program, started using automated DNA sequencing machines rapidly to identify sequences unique to human genes. A Genentech lawyer, Max Hensley, contacted the NIH technology licensing lawyer, Reid Adler, who in turn contacted Venter about filing a patent application on his method and the resulting DNA sequences. The method was eventually given over to the public domain through a statutory registration of invention, but the patent application for the sequences themselves continued through the patent examination process. That 1991 patent application generated tremendous controversy until 1994, when NIH Director Harold Varmus decided to abandon the patents, following the advice of patent scholars Rebecca Eisenberg and Robert Merges (1995).

The controversy at NIH paradoxically induced interest in commercial biotechnology circles. In 1991 and 1992, noise over ''Darth Venter's'' turn to the dark side (by patenting DNA sequences from gene fragments) attracted the attention of scientist Randall Scott at Incyte in California, and Incyte began to focus on DNA sequencing of human genes. Through 1994, several other companies—including Human Genome Sciences, Mercator Genetics, Genset, Myriad Genetics, Millennium Pharmaceuticals, Genome Therapeutics (renamed from Collaborate Research), Hyseq, and Sequenom—were formed around the idea of mapping and/or the sequencing the human genome, or turned from other pursuits to those ends.

One company illustrates the public-science origins of private genomics in particular: Human Genome Sciences. Wallace Steinberg, a former Johnson & Johnson executive who had started several biotech companies after leaving J&J, decided to meet Venter, having read about him amidst the patenting controversy. He talked Venter into leaving NIH to form a nonprofit research unit, eventually named The Institute for Genomic Research (TIGR), by promising Venter $70 million ($85 million by the time the deal was done) (Cook-Deegan, 1994). That was enough to build a larger sequencing and sequence-analysis facility than existed anywhere

[1] Robert Sinsheimer, Renato Dulbecco, and Charles DeLisi each independently proposed the idea of sequencing the human genome.

else at the time. Human Genome Sciences Inc. (HGSI), was formed as a for-profit corporation that would own the patent rights to TIGR's results as well as pursuing its own research leads. There were also plans to form additional companies, Industrial Genome Sciences and Plant Genome Sciences, to exploit different opportunities deriving from high-throughput sequencing and other genomic technologies. Steinberg tapped William Haseltine to become chief executive at HGSI. Haseltine had been involved in several previous Steinberg startup firms. Haseltine also had his roots in academic science, most notably from his work on HIV/AIDS at Harvard.

Genomics startups were a subgroup of biotech startups. The first boomlet in genomics startups in the early 1990s paralleled a significant increase in pharmaceutical R&D among established pharma and biotech firms that started in the early 1980s. The 1980s marked an intensification of competition among pharma companies based on R&D. A pharmaceutical R&D arms race of sorts began in the early 1980s, and during that decade, firms delved ever more deeply into molecular and cellular biology to bolster their ''absorptive capacity'' for drug discovery (Cockburn & Henderson, 1998; Fabrizio, 2005, unpublished data), recognizing the importance of rapid and effective use of public domain science to their business plans. Not all companies did this with equal success. Indeed, their ability to tap public science was one of the indicators of firms' success in pharmaceuticals (Fabrizio, 2004).

By historical happenstance, this birth of genomics out of publicly funded science took place as patent rights were being expanded and strengthened, by a combination of changes in legislation, in court decisions, and in patent offices. In academia, the major change was the Bayh-Dole Act of 1980, which gave grantees and contractors rights—and indeed a mandate—to seek patents on federally funded research results. Mowery and his coauthors review this history and some of its consequences in their book, *Ivory Tower and Industrial Innovation*, which combines economic empiricism, historical research, and policy analysis (Mowery et al., 2004).

Genomics, because of its timing as well as its commercial relevance—foreseeable to some immediately, to others a few years after its launch—took root as a field in American academe under the new Bayh-Dole regime. Academic institutions began to patent much more frequently after 1980, and genomics is one of the areas where this effect was pronounced. Moreover, patent rights were being expanded and strengthened in many areas of American law, including biotechnology. The Court of Appeals for the Federal Circuit (CAFC) was formed in 1982. It was designed to handle appeals of certain cases, including appeals of federal district court decisions about patent litigation. The CAFC quickly established itself as a generally pro-patent court. It and the patent office expanded the kinds of inventions that could be patented (including software and business methods, for example) and tended to strengthen the hand of patent-holders relative to those contesting patent rights (Jaffe & Lerner, 2005). In the hands of the CAFC, more territory could be enclosed and patent fences generally got higher. These developments had a particularly strong impact on areas of rapid innovation, including both ''wet lab'' biotechnology and bioinformatics, fields directly relevant to genomics.

Three other factors are not related to changes in policy, but nonetheless make genomics a useful field to study for this policy history: (1) the story was compressed into a decade, so its narrative is shorter and crisper, (2) there was intense media coverage, producing an ample public record of events, and (3) the patentable inventions arising from genomics can be tracked because it is possible to identify relevant patents. Patents resulting from genomics R&D almost always make patent

claims that use terms distinctive to DNA and RNA, which can be used to create a searchable patent database mapping to genomics research.[2]

3 Work enabled by a science commons

A science commons can supply information needed to achieve social benefit that for-profit markets in goods and services may fail to achieve. Moreover, even in markets well served by the profit motive, a science commons can in some circumstances improve efficiency, when many disparate firms can draw on a common pool of knowledge and data, rather than having to construct the same information firm-by-firm at substantial cost because of duplication. One theoretical rationale for this effect has been set forth by Benkler (2002). The cases arising in genomics suggest that network theory may have some practical applications in the real world of science and its application. I will illustrate three social goals that can benefit from a robust scientific commons in genomics: advancing science, improving public health, and creating a shared foundation for productively diverse forms of industrial R&D and commercialization. But first, some historical background.

4 Public and private genomics in mortal combat

The beginning of the Human Genome Project was marked by conflict between scientists who thought it was a poor use of resources versus those who thought it was a useful and efficient way to spend public research dollars. By broadening the project to include maps, tools, and organisms in addition to the human, most scientists came around to support the Project. A 1988 report of the National Research Council reported that consensus (National Research Council, 1998). That did not eliminate all conflict, however, because the question of which federal agency should play the larger role remained unresolved, and both the National Institutes of Health and the Department of Energy assumed active roles, in a roughly 2–1 ratio of funding. And even as the rival agencies in the US settled into a generally amicable cooperative framework, other nations began to engage in genomics R&D.

The 1991 controversy over gene-tagging sequences erupted while the Genome Project was getting underway. As that controversy died down, an even more public conflict over sequencing the entire genome exploded in 1998, pitting a private company against the public sector genome project. The battleground for both these two conflicts was the science commons.

TIGR and HGSI were formed in 1992. The heads of the two private organizations, Venter (TIGR) and Haseltine (HGSI), never sang close harmony, despite their supposed corporate matrimony. As TIGR moved away from human gene sequencing and into microbial sequencing, including proof of principle that whole-genome

[2] This is not true of patent collections based on ''gene patents,'' which flag patents containing sequence data (amino-acid or nucleic acid sequences from peptides and nucleic acid structures). The main reason is that many DNA-based patents claim methods, algorithms, or compositions other than DNA or RNA sequences. The algorithm for selecting patents into the DNA Patent Database is available at the site, http://dnapatents.georgetown.edu/SearchAlgorithm-Delphion-20030512.htm (accessed 2 April 2005). All patent studies have fuzzy edges, but the database is a tool that captures patents roughly corresponding to genetics and genomics.

shotgun sequencing could work, the noise from the TIGR-HGSI conflict got downright cacophonous. TIGR's scientific interests hardly coincided with what HGSI would want from its R&D partner, and two alpha-males confined in close corporate space found themselves in frequent conflict. In 1997, TIGR and HGSI severed their ties, with TIGR foregoing rights to future payments and HGSI foregoing rights to future TIGR discoveries (TIGR, 1997). It was a divorce made in heaven.

Venter became a free agent, heading up a free-standing nonprofit research institute, until Michael Hunkapiller of Applied Biosystems approached him with another Big Idea (Shreeve, 2004). In discussions with its parent company (then Perkin-Elmer Cetus, which became Applera), Applied Biosystems had begun to think seriously about sequencing the human genome with private funds. It would be a high-profile use of a promising new DNA sequencing instrument that was much faster and more scalable than existing sequencers. The question was whether the methods TIGR had used on smaller genomes, such as bacteria, could work on the human genome, and produce a final sequence faster than the public genome project. If so, a company could charge both for access to the data, and for access to informatic tools to mine the data. In order to charge users, the company would need a truly impressive bioinformatic capacity, and great tools for analyzing sequence data. If a private company decided to sequence the genome, it might even kick up a market for sequencing instruments, including Applied Biosystems machines, among the publicly funded laboratories doing DNA sequencing, who would buy the same machines to compete with the new genomic sequencing company.

In May 1998, Craig Venter became the head of a company, later named Celera Genomics, which would carry out the sequencing and pull together the computing infrastructure to assemble it into a reference sequence, and then begin to interpret the sequence information. Celera's 1998 establishment inaugurated another boom in genomics startups, this one entailing many more companies and much more money than the 1992–1994 boomlet.

The ''private genome project'' idea gathered steam and was announced through a sophisticated media roll-out strategy. The initial kernel of the media snowball was an exclusive to Nicholas Wade of the *New York Times* in May 1998 (Wade, 1998). Thus began a privately financed scientific effort at Celera running into the hundreds of millions of dollars that competed head to head with the publicly financed Human Genome Project. The drama played out over 3 years and became the biggest story in science, and one of the most visible general interest stories of its period.

The story is often told as a race, competition between Venter at Celera and the public Human Genome Project whose most conspicuous spokesmen were Francis Collins in the United States and Sir John Sulston in the United Kingdom. Collins was director of the National Human Genome Research Institute at NIH, and Sulston directed the Sanger Institute affiliated with the University of Cambridge and funded mainly by the Wellcome Trust of London (with additional funding from the UK Medical Research Council). The usual narrative strategy was to use the metaphor of a race, but in fact there were not just two human genome projects running in parallel, there were many.

A consortium of laboratories funded by government agencies and nonprofit organizations in North America, Europe, and Japan constituted the ''public genome project.'' Sulston emerged as the champion of that faction, emphasizing open science, rapid sharing of data and materials, and a passionate appeal to refrain from patenting bits of the human genome except when they could foreseeably induce

investment in developing end-products such as therapeutic proteins. Sulston was the leader and rhetorical warrior for Open Science.

Sulston's model for the human genome project was the biology of the worm (Ankeny, 2001)—a close-knit community of scientists who studied nematodes, and had made immense scientific progress in a hub-and-spoke model of biology. Two central laboratories—one at the University of Cambridge and another at Washington University in Saint Louis—did high-tech, whiz-bang, expensive mapping and sequencing projects on the worm genome. Those hubs shared data quickly and widely with the spokes—a vibrant network of smaller laboratories throughout the world. Sulston wrote *The Common Thread* with Georgina Ferry to tell the genome story from his point of view (Sulston & Ferry, 2002). His was the public works model of genomics, with public funding producing a valuable scientific resource.

The Wellcome Trust was the crucial nonprofit funder that supported this open science model. Michael Morgan from Wellcome Trust believed fervently in open science, and wanted the ''public'' genome project to succeed. The Sanger Institute was the Trust's foremost research institution, and John Sulston its most visible scientist. NIH and Francis Collins, as a government organization and employee, respectively, had to be more cautious in their rhetoric—and were, most of the time.

The Wellcome Trust sponsored a Bermuda meeting of the major sequencing centers throughout the world in 1996 (despite the exotic sound of it, the weather was miserable, it was off-season, and the site was chosen deliberately to be neutral, not in the USA or Europe). One theme of the meeting was how to make sequence data widely available, modeled on the worm science world. A set of ''Bermuda Rules'' emerged from the meeting, mandating daily disclosure of DNA sequence data. The pledge to rapidly share data was linked to a plea not to patent DNA, unless a gene or DNA sequence had been studied further to show its function or practical utility. That kind of functional biology was not the business of the publicly funded DNA sequencing centers, so it was in effect a ''no patents'' policy for the sequencing centers. Venter was present at the beginning of the Bermuda meeting, in his pre-Celera days as head of TIGR, but fittingly he left early and was gone by the time the Bermuda Rules were agreed. This left him room to later repudiate them.

The Wellcome Trust played another important role in 1998, soon after Venter announced his intention to sequence the genome at a new startup company. Wellcome reacted to the announcement of the new company by proposing to do a faster, better public genome sequence by increasing its commitment to fund genomic sequencing through the public project. Wellcome's move, in turn, bolstered funding from the US government, UK government, and other government and nonprofit funders of the public genome project.

In addition to the upstart startup Celera and the public genome project, the private firms HGSI and Incyte were in effect conducting a different kind of genome project in parallel—call it a Human Genes Project. Their strategy centered on human *gene* sequences—DNA coding for protein products. Both companies had been sequencing genes for 5–6 years before Celera was even formed, and had been sending in patent applications the whole time. Incyte worked with a group of pharmaceutical company subscribers. HGSI had one main client—SmithKline Beecham (which later merged with Glaxo Wellcome to become Glaxo SmithKline).

The business strategies of Incyte and HGSI were both initially based on sequencing human genes. Their styles were quite different, however. Randall Scott at Incyte was part of a scientific network with many links to the public genome

project. Indeed, at times Incyte was contemplated as a partner in the public project (Shreeve, 2004). HGSI had some academic and industrial collaborations beyond SmithKline Beecham, but far fewer than Incyte. And Haseltine's relation to the public genome project was as an outsider. Scott of Incyte was in the public genome family, or at least ate some meals with them; Haseltine was never welcome at the table.

Haseltine reinforced his role as troublemaker for the public genome project when he wrote a 1998 editorial to the *New York Times* arguing Congress should pull the plug on the public project because it was already being done ''without tax money'' by his company and others (Haseltine, 1998). Haseltine argued government funds for DNA sequencing would be better spent on smaller projects in individual laboratories to understand gene sequences. There were two problems with this argument. First, it assumed almost all the value of sequencing came from gene sequences, whereas molecular genetics has become focused on many regulatory processes that happen at the RNA and DNA level and are never translated into protein. It seems a safe bet that a lot of biology would never be approachable if we only got protein-coding sequences. Haseltine is an excellent scientist and knew this full well, although for commercial purposes, he could certainly make a good case that the most rapid returns were likely to come from coding sequences. This argument conflated commercial value with scientific value, but as an argument about public support for science, it is simply wrong, as the complexity of gene regulation is becoming obvious, and the importance of DNA sequences in addition to protein-coding regions is becoming apparent. Gene-based strategies made eminently good sense when hunting for drug targets, because drugs are designed to interact with proteins that are secreted outside of cells, that bind DNA, or that extend outward from the surface of cells. But as a tool to understand biology, the entire sequence was a much more powerful tool than just protein-coding regions.

The even deeper flaw in Haseltine's argument was about access to genomic information. Here he crossed the line from perhaps inadvertently disingenuous to deliberately misleading. The fact that genes were being sequenced by companies did no one but those companies any good if the sequences were not public. Academic scientists could, of course, approach HGSI or Incyte or other companies to collaborate, getting access to their data, but relatively few did so. The reason was that such collaboration came with strings—or ropes, or even cables. The constraints were patent rights—that is, exclusive property rights that were routinely being granted for full-length genes by the US Patent and Trademark Office. Collaboration with HGSI or Incyte meant nondisclosure agreements, publication review, and rights on resulting intellectual property. Sometimes this made sense, but it was not terribly attractive for those mainly interested in advancing science. A central condition of collaboration was control of information and constraints on open sharing of data. It made sense in a business context, but as a public works project in science, it made none. And to argue that proprietary gene sequencing was a substitute for public funding of genomic sequencing was ridiculous. Scientists could of course wait for patents to issue from privately sequenced genes, but that was not really a practical option because of the many-year delay. Perhaps scientists could hope HGSI and Incyte would publish the sequences voluntarily someday, but the companies would do that only when patents issued, or if it suited their business needs. The companies did publish, but only very selectively. To academic scientists in the field, waiting for companies to do the work would be surrendering to the competition in any event.

Pharmaceutical companies working with the Incyte and HGSI played from power—money and the ability to generate the data themselves if need be with their huge R&D war chests. Small academic laboratories were on the other end of the power curve, with relatively little leverage. Academic laboratories had a much better alternative, to scan the public GenBank for genes of interest, at no cost and with no strings attached. GenBank and other databases received sequences from thousands of laboratories throughout the world, as well as (eventually) the output of major DNA sequencing centers. Incyte and HGSI drew regularly on GenBank data, but company gene sequence data made their way back to public sequence databases only when a patent issued, or if the company chose to publish an article in the scientific literature.[3] In effect, company projects built on the foundation laid by the public genome project and drew regularly upon its data, but only occasionally contributed data back to public databases. This was sensible business practice, but it was misleading for Haseltine to imply that leaving the genome projects to companies and small laboratories would produce a genome project with the desired features of the concerted public project.

One forthright way to make Haseltine's case would have been to indeed allow the private gene sequencing firms to proceed, but for the government and nonprofit funders to pay to make the data public. There are two reasons Haseltine may have chosen not to take his arguments to this logical conclusion. First, the offer would likely have been refused by the companies, because their business plan was precisely to keep sequence data proprietary until they could be patented. Government procurement of the data would have vitiated this business plan, and turned the companies into contractors. The second reason was price. It would have been embarrassingly high, and certainly would have undercut the argument the work could be done ''with no tax dollars.'' But pushed to its conclusion, Haseltine's line of argument could have made a clean case—it might have made sense for the government to buy this particular genomic real estate and dedicate it to the science commons, if the private sector could produce the data faster and cheaper. Haseltine started from the premise that human gene sequence data were valuable—and who could argue with that?

From 1998 until February 2001, when *Nature* and *Science* published rival articles containing draft reference sequences of the human genome prepared by the public genome project (Lander et al., 2001) and by Celera (Venter et al., 2001), there were in effect two competing projects focused on sequencing the entire human genome, and in parallel also several other ''genome projects'' focused on expressed sequences and bits and pieces of the genome of interest to research communities in both public and private sectors. In addition to the two companies sequencing human genes, many other companies were mapping and sequencing parts of the human genome. And thousands of laboratories were contributing sequencing and mapping information to databases and to scientific publications. By the time the initial genomic sequence publications came out, the ratio of private to public

[3] After November 1999, a US patent application was published 18 months after being filed, if the applicant sought patent rights in any country with an 18-month publication rule. The companies did also publish occasionally in the scientific literature. HGSI, for example, listed 70 publications with one or more HGSI authors as of March 2005 on its website. These publications sometimes included sequence data, and if so, would make their way to GenBank, often before publication of the corresponding patent or patent application.

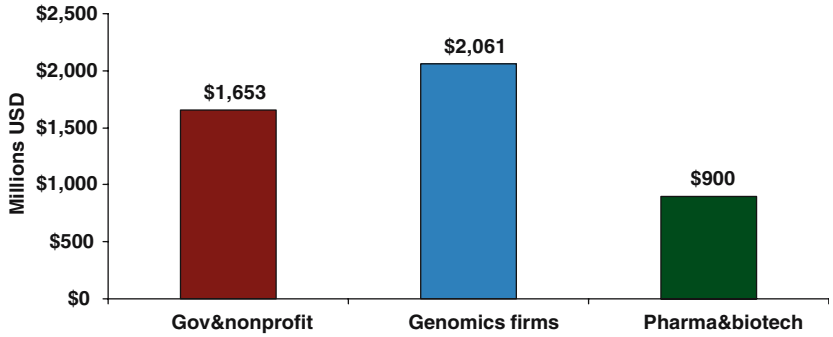

Fig. 1 Genomics Research Funding, 2000, *Source*: World Survey of Funding for Genomics Research, Stanford University, 2001 (unpublished data from Robert Cook–Deegan, Amber Johnson, and Carmie Chan, Stanford-in-Washington Program, based on a survey of over 200 funders)

funding appeared to be roughly two private dollars for every one government or nonprofit dollar (see Fig. 1).[4]

In 2001, the financial genome bubble burst. At the end of 2000, 74 publicly traded firms were valued at $94 billion, of which the largest 15 accounted for approximately $50 billion. By the end of 2002, those 15 firms' market value had dropped to $10 billion, but their reported R&D expenditures nonetheless climbed from $1 to $1.7 billion (Kaufman, Johnson & Cook-Deegan, 2004, unpublished data).

These data make three simple points: First, the private sector has invested heavily in genomics, but those investments are made in expectation of financial return. That is quite different from the public and nonprofit funding of genomics, which is mainly intended to produce public goods—knowledge and materials that are widely available to advance knowledge and combat disease. Second, private R&D investment is a powerful complement to the public and nonprofit funding. Private R&D follows public R&D in time, it draws on the science commons but does not necessarily contribute back to it. If successful, private R&D investment can create wealth and jobs as well as the social benefit from developing goods and services that would otherwise not be produced. This benefit is real, but it is distinct from the social value of the science commons.

Genomics also provides several examples of private funding to augment the science commons, such as the SNP Consortium, and the Merck funding to Washington University to fund gene sequencing (Cook-Deegan & McCormack, 2001). And third, and most to the point for policy purposes, it would be foolhardy to generalize from the happy circumstances when private R&D expands the science commons—to expect private R&D to substitute for the science commons except in unusual circumstances, usually related to the grounds of competition among firms in a particular industrial

[4] In a snapshot taken of year 2000 genomics research funding, ~70 nonprofit and government funders provided an estimated $1.6–1.7 billion; 74 publicly traded firms dedicated wholly to or including genomics research as a major function reported over $2 billion in R&D expenditures; and projecting 3–5% of R&D in major pharmaceutical firms was for genomics (based on survey responses and rough informal estimates of pharma R&D managers), established pharma firms were spending $800 million to $1 billion in genomics research. [World Survey of Funding for Genomics Research, Stanford-in-Washington program http://www.stanford.edu/class/siw198q/websites/genomics/entry.htm (accessed 2 April 2005)].

sector. Private industrial R&D will sometimes find it useful to contribute to the science commons, but expecting industry to do so always and consistently would be foolhardy.

5 Applications in public health: when markets fail

To see why having a healthy science commons matters, we move away from genomics to make a general point about health research. Murphy and Topel estimated that the gains in life expectancy from medical research 1970 to 1990 were staggering—in the range of $2.8 trillion per year ($1.5 trillion of this from cardiovascular disease reduction alone) (Murphy & Topel, 1999). Many of the health benefits of discovering new information about health and disease come not from drugs or vaccines or medical services, but from individuals acting on information. Cutler and Kadiyala attributed 2/3 of the health gains in cardiovascular disease reduction to effects of ''public information,'' such as stopping or reducing tobacco use, changing diet, getting more exercise, or monitoring one's blood pressure. The second largest determinant was technological change, such as introduction of new drugs and services, followed by increasing cigarette taxes to reduce tobacco use (Cutler & Kadiyala, 2001). The estimated return on investment in medical treatment was 4–1, but on the ''public information'' it was 30–1.

Cutler and Kadiyala's result cannot be generalized, because smoking is a very large risk factor that is *sui generis*, and cardiovascular disease has proven far more malleable to many kinds of interventions than nonlung cancer and other chronic diseases. The path from scientific understanding of cause to prevention of cancer, diabetes, arthritis, and Alzheimer's disease, among others, appears far less linear, and so the value of public information about risk is correspondingly less powerful and has less impact on health outcomes. Few if any risk factors will ever be found to rival tobacco use as predictors of poor health. But the finding that information can have value irrespective of being translated into products and services in a paying market is nonetheless important. Even if public information will not be quite as powerful in reducing other chronic diseases as it has been for cardiovascular disease, the vector is likely to point in the same direction. We cannot say that public information will always prove more powerful than information channeled into new drugs, vaccines, biologics, devices, and medical services sold for profit in the health care system. Where there are public health benefits from public research results, however—and the probability there will be no such public health effects of genomics seems vanishingly small—the health science commons is essential, because it alone can supply the public information benefits. Both words in ''public information'' do a lot of work. We need new information that arises from science, but to capture social benefits based on that knowledge itself, we also need it to be public.

Genomics provides several other examples of how public information is valuable. The 2002 report from the World Health Organization, *Genomics and World Health* gave the example of fosmidomycin (Advisory Committee on Health Research, 2002). This drug is currently being testing to treat malaria in Africa (Missinou et al., 2002). That use came to light as a consequence of sequencing the genome of the malaria parasite, and noticing a metabolic pathway not previously known to exist. The compound fosmidomycin was known to inhibit the pathway, and had been

developed as a treatment for urinary tract infections. When the new possible use to treat malaria was revealed, fosmidomycin was pulled off the shelf and moved into clinical trials against malaria. This is a treatment that may never turn a profit for any company, but the social returns could be enormous if fosmidomycin works, because so many millions of people are infected with malaria. If not fosmidomycin, then perhaps other findings will lead to prevention or treatment of malaria, enabled by now having the full genomic sequence available for host, pathogen, and mosquito vector (Gardner et al., 2002; Holt et al., 2002). Making the information about these organisms available worldwide is essential to accruing the benefits of research. There is only a weak world market for drugs to treat malaria because it is largely an affliction in resource-poor populations. The usual profit motives of the intellectual property system cannot create incentives where there is no prospect of profit to pull products through an expensive discovery and testing process. But networks of nonprofit organizations, such as the Malaria Vaccine Initiative, the Global Fund, the WHO essential medicines program, and other sources of ''public'' capital might nonetheless be capable of discovering and developing new treatments despite the unlikely prospect of commercial profit.[5] In theory, public funds might induce a sufficient incentive to motivate profit-driven investments for diseases of poor people living in poor countries, but it is not true now, and betting that money will be found could prove wrong. Having a scientific commons with information relevant to vaccines and treatments at least offers an alternative pathway. Many of the scientists most motivated to study such diseases work in resource-poor countries; they do not have rich resources, but they do have strong motivation, as well as computers and access to public databases.

Another case example is SARS. Strains of the coronavirus that causes SARS were identified and sequenced within a month by at least three laboratories in Asia, Canada, and the United States. That sequence information was shared widely, and a ''chip'' to detect the virus was available for research and possible clinical use just a few months later. Making progress with such alacrity requires strong norms of open science, with obvious social benefit.[6]

Many of the infectious diseases that plague mankind have long eluded measures to combat them. In many cases, this is because they are difficult to grow in tissue culture, and therefore research progress is slow. With new technology, the genomes of hundreds of ''nasty bugs'' have been fully sequenced, giving scientists an entirely new tool to develop drugs, vaccines, and control measures. It is far from clear that this will tilt the battle decisively in favor of humans over schistosomes, trypanosomes, plasmodia, bacteria, viruses, and other organisms that maim and kill humans by the billions, but it is a new line of attack. In the case of organisms on the Select Agent list of ''bioterror'' bugs, there is now extensive research underway to develop

[5] My colleagues Anthony So, Arti Rai, Jerry Reichman, Henry Grabowski, and others at Duke have joined many others from around the world to seek creative ways to ensure that essential drugs and vaccines are developed.

[6] This story of sharing sequence information is commingled with a potential intellectual property story that could be complicated. At least three of the institutions that did the sequencing have applied for patents, and interference proceedings could be complex, as they are in different countries and on different strains that might need to be cross-licensed for many practical applications. A patent pool could emerge, or a monster interference proceeding to sort out the questions of inventorship. The legal costs could exceed the costs of deriving the sequence itself. See E. Richard Gold 2003. SARS Genome Patent: Symptom or Disease? *The Lancet* (361): 2002–2003.

preventive and treatment measures. For most infectious agents that afflict those in the poorest parts of the world, however, the prospect of profit will not create a demand-pull for innovation that could improve billions of human lives, unless indirect incentives such as prizes or guaranteed payments for effective remedies by third parties serve as surrogates for paying markets.

Unlike the ''public information'' case described above, however, here the market failure has a different cause. It is not due to the fact that the research results are ''public goods,'' but because the potential users are deeply impoverished, and the economic incentives for drug development in advanced economies do not prevail. The failure here is one of profound inequality and distributive injustice. Nongovernment organizations around the globe, including major funders such as the Gates Foundation, the TB Alliance, the Global Fund, and others, are attempting to use philanthropy, government funding, and creative networking to address this form of market failure. Their efforts depend critically on access to scientific and technical information at low or no cost. Success on this front thus depends critically on the health of the scientific commons.

Another likely use of genomic information will be newborn screening, as more diseases are characterized, linked to possible intervention, and incorporated into routine testing. This must be done with care to avoid harms and false positives, but as knowledge accumulates, the list of conditions that can be treated will lengthen, and costs of testing should drop. Any benefits from newborn screening are unlikely to arise from strong profit motives, however, as most testing is done by state-funded laboratories in the United States and government public health programs in most other countries. The dollar amounts are small. [Two-thirds of US states spent between $20 and $40 per infant for all screening in 2002, and no state spent more than $61 (US General Accounting Office, 2003)]. This is far less than most single genetic tests, or even routine medical laboratory tests. Newborn screening is now, and will likely continue to be, a public health service (Newborn Screening Steering Committee, 2005). Any shift to DNA-based testing, or addition of tests beyond the current testing regimes, will face very serious cost constraints, and advances are unlikely to result from prospect of ample profits in this market.

6 Public inputs to private science

Even if we were to stipulate that the ''public information'' impact of health research might be less important in the future than it has been in the past, does it diminish the role and importance of the science commons? In this section, the focus is not on social benefits foregone for lack of a robust commons. Instead, the argument shifts to efficiency gains to private R&D that follow from being able to draw upon the commons.

Several lines of research corroborate the intuition that a pool of public information and materials must surely ''raise all ships'' to the benefit of each. The case is likely to be stronger in health research than in other lines of research, just because of the well known deep mutualism between public and private R&D in health research.

The late Edwin Mansfield's surveys of industrial leaders clearly showed that executives in firms believed their lines of business-related R&D depended on academic research, and pharmaceuticals to a greater degree than any other sector he characterized (Mansfield, 1995). Narin and colleagues have repeatedly shown how

industrial publications cite academic research, and patents related to pharmaceuticals and biotechnology cite academic research far more heavily than most other kinds of inventions (Narin & Olivastro, 1992). When Steve McCormack and I read through the more than 1,000 DNA-based US patents issued 1980–1993, we found that 42%were assigned to universities (14% to private; 9% to public), nonprofit institutions (13%), or government (six percent), compared to less than three percent academic ownership of patents overall (McCormack & Cook-Deegan, 1996–1997, unpublished data). This is a tenfold enrichment of academic involvement in life sciences compared to most other kinds of invention.

The 1997 survey of the Association of University Technology Managers was the last year for which the questions made it possible to analyze life sciences separately from physical sciences. That year, life sciences accounted for 70 percent of licenses and 87% of income (Massing, 1998).[7] Industries closest to health research depend on academe, and academic institutions are more heavily involved in technology transfer activities related to the life sciences. If we were looking for a place where public science matters to industry, life sciences would be a good place to start.

What is really going on? Beyond the special role of academic institutions as the training grounds for both technical and nontechnical workers in the knowledge economy, academic institutions also play a unique role in creating and sustaining the science commons. It is worth noting that the studies above generally focus on academic R&D, not specifically on the science commons, or only ''open science.'' Recall that universities and nonprofit research centers do not always practice open science, and some elements in the commons come from private industry R&D. Academic research institutions are nonetheless the main stewards of the science commons. While we cannot be completely sure, it is quite likely that the main explanation for the importance of academic research is that it is open, producing data and materials available to all.

The most direct line of evidence for this comes from the Carnegie-Mellon Survey of industrial R&D managers. Cohen, Nelson, and Walsh conclude that ''public research has a substantial impact on industrial R&D in a few industries, particularly pharmaceuticals,'' and

> ''The most important channels for accessing public research appear to be the public and personal channels (such as publications, conferences, and informal interactions), rather than, say, licenses or cooperative ventures. Finally, we find that large firms are more likely to use public research than small firms, with the exception that start-up firms also make particular use of public research, especially in pharmaceuticals (Cohen et al., 2002)''.

This certainly corroborates the stories of genomics startup companies, including companies like Celera, depending heavily on their recent past in academic research, and their ongoing collaborations with (and sometimes customers and markets in) academic research. And it confirms the role of large firms in preferring to draw inputs from a science commons, rather than having to collect atomized, individually expensive fragments of proprietary technologies and data.

[7] The AUTM survey has continued in subsequent years, but the questions separating life sciences from physical sciences have been dropped. See Figs. 4 and 5, page 8 of the 1997 survey report (Massing, 1998).

The history of genomics provides many examples of this, but two are particularly famous. One salient example is the decision in the period 1988–1991 by the National Institutes of Health not to sequence human genes (i.e., protein-coding regions), but instead to focus on systematically mapping and sequencing the entire genome (Cook-Deegan, 2003). That decision opened the way for private firms Human Genome Sciences and Incyte to fill the void, attracting private capital to do what the public sector had chosen not to do. Because it fell victim to the law of unintended effects, NIH's decision not to pursue cDNA sequencing, however well-intended and understandable, was a mistake in retrospect.

The story behind that decision is mainly about the sociology of science, not a theory of the science commons, but it is instructive nonetheless. The decision not to sequence protein-coding regions was initially about fairness between big labs and small ones, not about commercial prospects. As the genome project took shape, the importance of maps of humans and various ''model organisms'' was apparent. What kinds of maps deserved substantial funding and concerted effort remained, however, a matter of ongoing dispute. One of the bones of contention was a ''gene map'' based on cDNA technology—that is, making DNA copies of the messenger RNA translated into protein within cells. Construction of cDNA libraries was standard fare, and remains a seminal technology in efforts to study expression of many genes through microarray technologies.

One question left open during the early debates about the Human Genome Project, 1987–1991, was whether the genome project would include ''gene'' sequencing—to start sequencing efforts with DNA known to code for protein, and therefore certain to provide codes for most of the important building blocks of cells, while also providing targets for drug development. A technical means to isolate the RNA that is translated into proteins was readily available. DNA could be made from such RNA molecules. This was called cDNA technology, ''complementary'' to the messenger RNA that is exported from the nucleus of the cell to its cytoplasm to be translated into protein. In fact, one could take it a step further and look for genes coding proteins likely to be of particular biological significance—and focus on just those cDNAs coding for secreted proteins and peptides (such as hormones or neurotransmitters), for receptor or transporter molecules extending outside the cell (with many trans-membrane domains), or proteins that bind DNA (with ''zinc fingers), etc. These ''functional motifs'' could be predicted, if imperfectly, from DNA sequence data. One logical strategy to start the DNA sequencing program was to sequence cDNAs of particular interest first, then other cDNAs, and then turn to ''genomic'' DNA between genes. (DNA between genes would still be of interest because such sequences were likely to house regulatory signals for turning genes on and off, and affecting the timing of gene expression, as well as structures involved in cell division and the 3-dimensional shape of DNA in cells.)

At one of the first public discussions of the human genome project, at Cold Spring Harbor in June 1986, Walter Gilbert responded to one attack on the idea of sequencing the genome by noting, ''of course you would start by sequencing the cDNAs'' (Gilbert, 1986). When the congressional Office of Technology Assessment presented a plausible budget for funding the genome project, it included a cDNA sequencing component (US Congress, Office of Technology Assessment, 1988). The Department of Energy did pursue some cDNA sequencing, but NIH's genome program did not. It was a matter of some discussion, but in the end it was largely James Watson's call, as director of the relevant NIH center. Several arguments were

made against cDNA sequencing. First, it was already going to happen, since incentives to find genes were strong with funding from other NIH institutes, but incentives for individual labs to produce whole genomic sequence data were entirely dependent on ''genome project'' funding. Another, related, argument was about the sociology of science—if big sequencing centers did cDNA sequencing, they would inevitably also be at least tempted to pause to characterize particularly interesting genes, and turn to the fascinating biology sure to follow. There were two problems with this: (1) it would distract them from the major task at hand of deriving a complete reference sequence of the entire genome, and (2) it would give them an unfair advantage over the thousands of smaller laboratories lacking the DNA sequencing firepower.

It was the NIH decision not to fund cDNA sequencing that left the door open to Incyte and HGSI to follow human cDNA sequencing with private funding, because in the absence of a big public effort, the low-hanging fruit of the genome was there to be plucked, sequenced, and shipped off with claims to the patent office.

When Incyte and HGSI began to go down this path, those who saw genes as increasingly important inputs to their R&D efforts—particularly large pharmaceutical companies—got concerned, for two reasons. One was that the US Patent and Trademark Office was obviously patent-friendly, industry-oriented, and seemingly tone-deaf to the concerns of scientists about enclosing the public domain. Patents would issue. And if patents were granted, then any firm making, using or selling a gene or gene fragment could be hit up for a piece of the action by the company that first sequenced it. Incyte and HGSI were clearly capable of filing patent applications on hundreds of thousands of gene tags, and thousands of full-length genes. Moreover, the small genomic startups had a running start on large pharmaceutical firms, the plodding Apatosaurus's of the biotech Jurassic.

Merck decided to take action (Williamson, 1999).[8] It stepped forward to fund a public domain sequencing effort, starting with gene fragments and moving on to full-length cDNAs. The work was to be done at Washington University in Saint Louis, home of one of the largest public genome sequencing facilities, and the data were to be moved quickly into the public domain.

Merck funded the work through a nonprofit arm and had no privileged access to the data. Here was a large company funding data to flow into the science commons where it would be freely available to all. Why would it do this? Four reasons suggest themselves: (1) it poisoned the well for Incyte, HGSI, and other startup firms, creating an open, academic competitor (albeit funded by industry) to shut the window on securing exclusive property rights on genes, and thus limiting the number of genes that would be have to be licensed; (2) it built good will with scientists, vital collaborators in Merck's drug discovery efforts; (3) it was great PR; and (4) it took

[8] An excerpt from the press statement upon the first data release explained some details: ''The Merck Gene Index is a broad collaborative effort, coordinated by Dr. Alan Williamson, Vice president, Research Strategy Worldwide, and Keith O. Elliston, Associate Director, Bioinformatics, of the Merck Research Laboratories. Dr. Greg Lennon's laboratory at the Lawrence Livermore National Laboratory (Livermore, California) has been supplying arrayed cDNA clones to Dr. Robert Waterston's laboratory (the Genome Sequencing Center) at the Washington University School of Medicine (St. Louis, Missouri) for sequencing. The sequence data are being submitted to the Expressed Sequence Tag (EST) division of GenBank on a regular basis for immediate distribution. [GenBank, built and distributed by the National Center for Biotechnology Information (NCBI) is a central repository of publicly-available gene sequence information, widely known and heavily used by researchers in government, academe, and industry]''.

advantage of nonprofit funding. If Merck paid for it as corporate R&D, it could deduct the R&D as an expense, but would also have to justify public domain science at stockholder expense. Through a nonprofit arm, Merck funded great science, burnished Merck's image, and enhanced Merck's future freedom to operate cleanly, without having to appropriate any returns on an ''investment.''

The SNP Consortium story started 5 years later, but followed the same general outline, with an added level of sophistication. During the late 1990s, it became apparent that there were many single-base-pair differences in DNA sequence among individuals. These were dubbed Single Nucleotide Polymorphisms, or SNPs, because of molecular biologists' penchant for impenetrable polysyllabic neologism (IPN) and three-letter acronyms (TLAs).

Single Nucleotide Polymorphisms could be used as DNA markers, to trace inheritance, to look for associations with diseases or traits, and to study population differences. They were valuable research tools. Many genomic firms, including Celera, began to signal they were finding SNPs and filing patent applications. Given the uncertainty about what the patent office would allow to be claimed in patents, it seemed possible patents on SNPs would be granted, meaning anyone using patented SNPs would need to get a license. This raised the prospect of needing to get licenses on hundreds or even thousands of SNP sequences from some unknown (but potentially large) number of patent owners. The Court of Appeals for the Federal Circuit had instructed the patent office that the ''nonobvious'' criterion for DNA sequence was met by any new DNA sequence, so ''obvious'' did not mean ''obvious how to find it'' but ''sequence determined and in hand.'' The patent office was signaling it might permit patents for any plausible utility, demonstrated or not, and related to biological function or not (Doll, 1998). SNPs might be patentable. This was just the kind of nightmare that Michael Heller and Rebecca Eisenberg had speculated might arise in their classic 1998 article on the ''anticommons''—situations when too many exclusive rights upstream needed to be assembled, thus thwarting the development of final products, such as drugs, vaccines, biologics, or instruments (Heller & Eisenberg, 1998).

This threat awakened some companies and scientific institutions to forge an alliance to defeat patent rights in SNPs (The SNP Consortium, 2005; Holden, 2002; Thorisson & Stein, 2003). The SNP Consortium was founded in 1999 to first discover SNPs, file patent applications, map and characterize the SNPs, and then finally abandon the patent applications. The expense and paperwork of this elaborate dance were intended to ensure SNPs landed in the public domain unfettered by patent rights. It was deemed necessary as a defensive strategy to ensure that consortium members would have standing as inventors should disputes arise about priority for related inventions (in patent parlance, interference proceedings, the administrative procedure to determine the real first inventor). Here a group of private firms of various sizes found common cause in defeating patents on research tools. They valued their freedom to operate highly and the threat of patenting sufficiently to pay for a complicated, expensive procedure to enlarge the public domain.

Again, what in the world was going on? Private firms that dearly loved patents for their own products were working together with academic institutions to defeat patents? One interpretation might be that the public sector failed to support lines of research with a strong need for a science commons sufficiently. But members of the public genome project were well aware of the need for unfettered access to SNPs and were as worried about the problem as the private firms that wanted to use SNPs

in their research. The issue here was the presence of many different kinds of genomics firms, some of which saw an opportunity to create and sell access to SNP research tools. It was no accident that this episode played out during the genomics bubble years, 1998–2001, when seemingly any startup with ''omics'' in its name could raise millions in private placements and months later (before any products hit the market) tens of millions through Initial Public Offerings of stock. It was conceivable that a company could raise private capital to find SNPs based on a possible paying market to use them in research. The public sector was simply not going to be able to mount a systematic SNP initiative fast enough and large enough to compete, and other companies wanted to avoid having to deal with the SNP upstart firms (yes, Celera was one of the firms with an interest in SNPs).

One interpretation of this story is that ''the market,'' some market somewhere, solved the problem. The wonder of capitalism worked its magic by creating public domain resources at private expense to forestall the undue private appropriation of rents from research tools. OK, maybe so. The explanation is as complex as the sentence that contains it. And it is clearly true that private firms funded public domain science. Does it generalize? Can we learn to relax, and assume that excesses of the patent system will be compensated by enlightened capitalists guarding their long-term best interests and future freedom to operate? The Merck Gene Index and SNP Consortium show the answer is ''sometimes yes.'' The nagging worry is that sometimes the answer may be no.

7 The science commons and economic efficiency: costs of data access

A final historical pastiche before closing out the arguments. Consider again the prospect of an alternative universe in which free access to data about the medical literature and scientific data we take for granted in health research might instead be constrained by exclusive proprietary rights. If the history and geography had been different and database firms had turned their attention to genomics just a bit sooner, the story might have been quite different. As it was, the early algorithms for interpreting DNA sequence—such as the BLAST and Smith–Waterman algorithms—were developed by individuals committed to open science. In more recent years, patents have begun to issue on bioinformatic methods relevant to genomics. In some cases, these patents confer incentives to support ''products'' marketed by firms, with service teams and development teams to improve their quality.[9] How this story will play out remains to be seen, but the ideas of ''open genomics'' are being tested in the real world along-side more proprietary models.

Databases themselves could become a focus of concern. The early years of the human genome project were marked by many decisions about the disposition of crucial databases. Human genetic disease and variation was lovingly cataloged by a team surrounding its founder, Victor McKusick of Johns Hopkins University, in Online Mendelian Inheritance in Man (OMIM). Many databases were established to retain data on human genetic maps of various types, and similar databases for other organisms. DNA sequence data were collected primarily by a trio of databases in the United States, Europe, and Japan, and these shared data among themselves. There

[9] My Duke colleague Arti Rai is working on how ideas of ''open source'' in software might be applied (or not) to genomics. ''Open source genomics'' is the subject of her 5-year project.

was, in effect, just one major, central DNA sequence database beginning in the early 1980s. Creating and coordinating these databases, including the sequence databases, was its own titanic struggle (Smith, 1990), but the battle was waged with only glancing concern for commercial potential. The databases contain many errors (Pennissi, 1999), and creating financial incentives sufficient to encourage careful curation and maintenance is one reason to support proprietary rights in making databases. But that step should not be taken lightly, and now we have a decade-long experiment in the real world to inform such decisions, with strong protection in Europe and only copyright and contractual protections for databases in the United States.

How different it might have been had the genome project begun in Europe, just a decade later, when the European Community saw fit to create a new exclusive right in databases as an incentive for companies to create and maintain valuable data. The impacts of this new form of intellectual property have received particular attention from the scientific community. Scientists have become concerned that rights could hinder research. The landmark report on the topic was the *Bits of Power* report from the National Research Council (1997), which has led to a line of further work. Much of the most advanced work has focused on weather, remote imaging and other huge and complex data sets. There may be cause for worry, and not just for scientists, but for the innovation system as a whole. Exclusive property rights create friction and inefficiency. It may be that free access to data generated at government and non-profit expense is far more efficient, and a more powerful prime for the economic engine, than allowing every incremental advance to form the basis for rent-seeking.

In the patent-happy United States that moves toward ever-longer copyright and protects creative works with the Digital Millennium Copyright Act, there is an anomaly. Data generated at government expense and published by the government cannot be copyrighted, and are thus freely available to anyone who wants to use them. American government agencies are generous suppliers of data to which others add value. It turns out that when it comes to data about the weather, it is the Europeans who are Scrooge, charging for access. And yet US businesses that provide weather information to various kinds of users have flourished, and the US market for such information is vastly larger than in Europe, despite the nearly equal size of the economies of the European Union and the United States. An analysis by Peter Weiss of the National Weather Service concludes, ‘‘The primary reason for the European weather risk management and commercial meteorology markets lag so far behind the US is the restrictive data policies of a number of European national meteorological services’’ (Weiss, 2002).

Given that genomic databases and most health research databases are publicly administered and protect strong norms of open sharing, concern over database protections could prove a sideshow. Perhaps it is silly to think that DNA sequence might have been housed in a proprietary database owned by Reed Elsevier, Springer, or Thomson. But some databases do straddle nonprofit and for-profit worlds, and if a strong US database right were created, the rules of the game could change. SwissProt, a database with information about proteins of interest in molecular biology, has been the subject of dispute, both about how to fund it, and about its pricing and access policies driven by trying to ensure its long-term financial survival. The analogies between weather and DNA sequence data are not exact, but careful thinking about policies bearing on health research data, including genomic

data, is crucial, because the creation of a US database right similar to the European counterpart remains a distinct possibility.

8 Conclusion: deliberate policies preserved a healthy science commons in genomics

The various genome projects, both public and private, pursued quite disparate policies about sharing of data and materials. Proprietary technologies and data were created, mainly by private startup firms, and they contributed to the pace and success of the Human Genome Project. Deliberate policies of funding organizations, especially the Wellcome Trust and National Human Genome Research Institute and other funders of the ''public genome project'' created and preserved a large and important science commons of genomic data and technologies for analyzing DNA structure and function. Agreements such as the Bermuda Rules, privately funded initiatives such as the Merck Gene Index, and public-private hybrids such as the SNP Consortium were deliberately designed to promote broad access to data and materials.

Genome projects spanned a full range of openness, from rapid open access under the Bermuda Rules, to subscription-based access to genomic data and analytical tools at moderate cost (e.g., Celera), to highly proprietary gene-sequencing with public disclosure mainly limited to patents as they were granted and published (Human Genome Sciences and Incyte). The practical ''public information'' benefits from having information widely and inexpensively available, such as public health advances from new knowledge about health risk, reinforce the benefits for science, where a broad network of investigators can draw on masses of information. The value of the science commons is not an argument against the private R&D. It is, however, a powerful argument for the need to support open science and a healthy science commons upon which both public and private science can draw. Without explicit policies to foster the science commons, this valuable pool of knowledge would have been shallower, and a less productive fountain of social benefits.

Science is not just about creating knowledge, it is also about making it widely available and making it useful. Deliberate policies to promote open access and low-cost use enable some social benefits that profit-driven R&D cannot. Private genomics is a laudable complement to public genomics. Public genomics is, however, a creature of deliberate policies, not just to fund the science but also to ensure that the results are shared. It is not a system that can be left to mindless self-assembly or politics as usual. Without an expansive science commons, many benefits would be lost and private genomics would be vastly less productive and valuable.

Acknowledgments Suparna Salil and Ilse Wiechers edited and helped format citations and references. Work was funded by The Duke Endowment, and by the National Human Genome Research Institute and the US Department Energy, through grant P50-HG003391.

References

Advisory Committee on Health Research, World Health Organization (2002). *Genomics and World Health*. Geneva, CH: World Health Organization.

Ankeny, R. (2001). The natural history of *Caenorhabditis elegans* research. *Nature Revies Genetetics*, *2*, 474–479.

Benkler, Y. (2002). Coase's penguin, or linux and the nature of the firm. *Yale Law Journal*, *112*, 369–446.
Cockburn, I., & Henderson, R.M. (1998). Absorptive capacity, coauthoring behavior, and the organization of research in drug discovery. *Journal of Industrial Economics*, *46*, 157–82.
Cohen, W.M., Nelson, R.R., et al. (2002). Links and impacts: The influence of public research on industrial R&D. *Management Science 48*, 1–23.
Cook-Deegan, R. (1994). The Gene Wars: Science, Politics, and the Human Genome. New York: WW Norton.
Cook-Deegan, R. (2003). The colossus of codes. In J.R. Inglis, J. Sambrook, & J.A. Witkowski (Eds.), *Inspiring Science: Jim Watson and the Age of DNA* (pp. 387–395). Cold Spring Harbor, NY: Cold Spring Harbor Press.
Cook-Deegan, R. M., & McCormack, S. J. (2001). Patents, secrecy, and DNA. *Science*, *293*, 217. See especially accompanying online materials at http://www.sciencemag.org/cgi/content/full/293/5528/217/DC1. Accessed 30 Octobar 2005.
Cutler, D. M., & Kadiyala, S. (2001). *The return to biomedical research: Treatment and behavioral effects, Working paper*. Retrieved from http://post.economics.harvard.edu/faculty/dcutler/papers/cutler_kadiyala_for_topel.pdf. Accessed 3 April 2005.
Delphion Database (2005). *A division of the Thomson Corporation*. Retrieved from http://www.delphion.com. Accessed 3 April 2005.
DNA Patent Database (2005). *Georgetown University*. Retrieved from http://dnapatents.georgetown.edu. Accessed 2 April 2005.
Doll, J.J. (1998). The patenting of DNA. *Science*, *280*, 689–690.
Eisenberg, R.S., & Merges, R.P. (1995). Opinion letter as to the patentability of certain inventions Associated with the identification of partial cDNA sequences. *American Intellectual Property Law Association Quarterly Journal 23*, 3–51.
Gardner, M.J, et al. (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*, *41*, 498–511.
Gilbert, W. (1986). Response to questions at impromptu session on the human genome project at the molecular biology of homo sapiens symposium on quantitative biology. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory. (Captured on tape by C. Thomas Caskey, transcribed and archived by Robert Cook-Deegan at the Human Genome Archive, National Reference Center on Bioethics Literature, Georgetown University).
Haseltine, W. A. (1998). Life by design, gene mapping, without tax money. *New York Times*. New York, A33.
Heller, M.A., & Eisenberg, R.S. (1998). Can patents deter innovation? The anticommons in biomedical research. *Science*, *280*, 698–701.
Holden, A.L., (2002). The SNP consortium: Summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques*, (June Supplement), 32, S22–S26.
Holt, R.A., et al. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, *298*, 129–149.
Jaffe, A. B., & Lerner, J. (2005). *Innovation and Its Discontents*, Princeton: Princeton University Press. See especially Chapter IV, which lays out evidence for the ''pro-patent'' position of CAFC.
Jensen, K., & Murray, F. (2005). Intellectual property landscape of the human genome. *Science*, *310*, 239–240.
Lander, E.S., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*, 860–921.
Mansfield, E. (1995). Academic research underlying industrial innovations: Sources, characteristics, and financing. *Rev Econ Stat*, *77*, 55–65.
Massing, D. E. (Ed.). (1998). *AUTM Licensing Survey: FY 1997*. Norwalk, CT: Association of University Technology Managers.
Merton, R.K. (1973). *The Sociology of Science*. Chicago: University of Chicago Press.
Missinou, M.A., et al. (2002). Fosmidomycin for Malaria. *Lancet*, *360*, 1941–1942.
Mowery, D.C., Nelson, R.R., et al. (2004). *Ivory tower and Industrial Innovation, University-Industry Technology Transfer Before and After the Bayh-Dole Act,* Stanford: Stanford Business Press.
Murphy, K. M., & Topel, R. (1999). *The Economic Value of Medical Research*. Retrieved from http://www.laskerfoundation.org/reports/pdf/economicvalue.pdf. Accessed 3 April 2005.
Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. *Research Policy*, *21*, 237–249.

National Research Council (1997). *Bits of Power: Issues in Global Access to Scientific Data*. Washington, DC: National Academy Press.
National Research Council (1998). *Mapping and Sequencing the Human Genome*. Washington, DC: National Academy Press.
Nelson, R. R. (2006). *The Market Economy and the Scientific Commons, lecture to the University of Michigan Law School, 26 January 2006*. Retrieved from http://www.law.umich.edu/CentersAndPrograms/olin/papers/Winter%202006/nelson.pdf. Accessed 22 March 2006.
Newborn Screening Steering Committee, Maternal and Child Health Bureau, Resources and Services Administration, US Department of Health and Human Services (2005). *Newborn Screening: Toward a Uniform Panel and System*. Retrieved from ftp://ftp.hrsa.gov/mchb/genetics/screeningdraftforcomment.pdf. Accessed 3 April 2005.
Pennissi, E. (1999). Keeping genome databases clean and up to date. *Science*, *286*, 447–450.
Science Commons (2005). Executive director, John Wilbanks, with headquarters based at the massachusetts institute of technology. Retrieved from www.sciencecommons.org. Accessed 2 April 2005.
Shreeve, J. (2004). *The Genome War; How Craig Venter Tried to Capture the Code of Life and Save the World*. New York: Knopf.
Smith, T.F. (1990). The history of the genetic sequence databases. *Genomics*, *6*, 701–707.
Smith Hughes, S. (2001). Making dollars out of DNA: The first major patent in biotechnology and the commercialization of molecular biology, 1974–1980. *Isis*, *92*, 541–575.
Stokes, D.E. (1997). *Pasteur's quadrant: Basic science and technological innovation*. Washington, DC: Brookings Institution Press.
Sulston, J., & Ferry, G. (2002). The common thread: A story of science, politics, ethics, and the human genome. Washington, DC: National academy press.
The Institute for Genomic Research (TIGR) (1997). *TIGR/HGS funding relationship reaches early conclusion*. Retrieved from http://www.tigr.org/news/pr_06_24h_97.shtml. Accessed 23 March 2006.
The SNP Consortium (2005). Retrieved from http://snp/cshl.org/about/. Accessed 3 April 2005.
Thorisson, G.A., & Stein, L.D. (2003). The SNP consortium website: Past, present and future. *Nucleic Acids Research*, *31*, 124–127.
US Congress, Office of Technology Assessment (1988). *Mapping Our Genes: The Genome Projects—How Big? How Fast*? OTA-BA-373. Washington, DC: US Government Printing Office. 182–183 (Table B-1).
US General Accounting Office (2003). *Newborn Screening: Characteristics of State Programs*. Washington DC: US General Accounting Office. GAO-03–449, 14ff, Appendix V.
US Patent and Trademark Patent Database (2005). Retrieved from http://www.uspto.gov. Accessed 2 April 2005.
Venter, J.C., et al. (2001). The sequence of the human genome. *Science*, *291*, 1304–1351.
Wade, N. (1998). Scientist's plan: Map all DNA within 3 years: New York Times A1, New York.
Weiss, P. (2002). *Borders in cyberspace: Conflicting public sector information policies and their economic impacts*. Washington DC: National Weather Service, National Oceanic and Atmospheric Administration, US Department of Commerce.
Williamson, A.R. (1999). The merck gene index project. *Drug Discoverys Today, 4*, 115–22.
Ziman, J. (1978). *Reliable knowledge: An exploration of the grounds for belief in science*. New York: Cambridge University Press.