NBER WORKING PAPER SERIES

INTELLECTUAL PROPERTY RIGHTS AND INNOVATION: EVIDENCE FROM THE HUMAN GENOME

Heidi L. Williams

Working Paper 16213
http://www.nber.org/papers/w16213

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2010

I am very grateful to Joe Doyle, Dan Fetter, Matt Gentzkow, Fiona Murray, Eva Ng, Scott Stern, and especially my advisers David Cutler, Amy Finkelstein, and Larry Katz for detailed feedback. Many other colleagues and seminar participants also provided very helpful comments. Several individuals from Celera, the Human Genome Project, and related institutions provided invaluable guidance, including Sam Broder, Peter Hutt, and particularly Mark Adams, David Altshuler, Bob Cook-Deegan, Eric Lander, and Robert Millman. David Robinson provided valuable assistance with the data collection. Financial support from NIA Grant Number T32-AG000186 to the NBER, as well as the Center for American Political Studies at Harvard, is gratefully acknowledged. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Intellectual Property Rights and Innovation: Evidence from the Human Genome
Heidi L. Williams
NBER Working Paper No. 16213
July 2010
JEL No. I10,I18,O3,O34

ABSTRACT

This paper provides empirical evidence on how intellectual property (IP) on a given technology affects subsequent innovation. To shed light on this question, I analyze the sequencing of the human genome by the public Human Genome Project and the private firm Celera, and estimate the impact of Celera's gene-level IP on subsequent scientific research and product development outcomes. Celera's IP applied to genes sequenced first by Celera, and was removed when the public effort re-sequenced those genes. I test whether genes that ever had Celera's IP differ in subsequent innovation, as of 2009, from genes sequenced by the public effort over the same time period, a comparison group that appears balanced on ex ante gene-level observables. A complementary panel analysis traces the effects of removal of Celera's IP on within-gene flow measures of subsequent innovation. Both analyses suggest Celera's IP led to reductions in subsequent scientific research and product development outcomes on the order of 30 percent. Celera's short-term IP thus appears to have had persistent negative effects on subsequent innovation relative to a counterfactual of Celera genes having always been in the public domain.

Heidi L. Williams
National Bureau of Economic Research
1050 Massachusetts Avenue
Cambridge, MA 02138
and NBER
heidiw@mit.edu

It has long been recognized that competitive markets may not provide adequate incentives for innovation (Nelson, 1959; Arrow, 1962). Given the presumed role of innovation in promoting economic growth, academics and policy makers have thus focused attention on the design of institutions to promote innovation. Intellectual property (IP), such as patents and copyrights, is one frequently-used policy lever. IP is designed to create incentives for research and development (R&D) investments by granting inventors exclusive rights to their innovations for a fixed period of time. An important but relatively under-studied question is how IP on a given technology affects subsequent innovation in markets where technological progress is cumulative, in the sense that product development results from several steps of invention and research. In this paper, I provide empirical evidence on this question by analyzing the sequencing of the human genome by the public Human Genome Project and the private firm Celera, and by estimating the impact of Celera's gene-level IP on subsequent scientific research and product development outcomes.

The sign of any effect of IP on subsequent innovation is theoretically ambiguous. I outline a simple conceptual framework that focuses attention on two counteracting effects. Consider two firms: Firm A holds IP on discovery A, and Firm B has an idea for a downstream product B. On one hand, IP on discovery A could discourage R&D on product B if an appropriate licensing agreement cannot be reached. In a classic Coasian framework, Firms A and B could always negotiate appropriate licensing agreements (Green and Scotchmer, 1995). However, transaction costs may arise if, for example, Firm B's research costs are private information. Such transaction costs could cause licensing agreements to break down, potentially discouraging R&D on product B (*e.g.* Bessen (2004)). On the other hand, IP on discovery A could encourage R&D on product B if there is imperfect IP protection in downstream markets, in which case IP on discovery A could increase Firm B's ability to capture rents in the market for product B (*e.g.* Kitch (1977)).

Empirical study of this question has traditionally been hampered by concerns that the presence of IP may often be correlated with other factors, such as the expected commercial potential of a given discovery. The contribution of this paper is to identify an empirical context in which there is variation in IP across a relatively large group of *ex ante* similar technologies, and to trace out the impacts of IP on subsequent scientific research investments and product development outcomes.

Two efforts, the public Human Genome Project and the private firm Celera, aimed to sequence the DNA of the human genome. The two efforts took different approaches to DNA sequencing, inducing differences in which effort first sequenced a given gene. Once sequenced by the public effort, genes were placed in the public domain, with the stated aim to *"...encourage research and development."* If a gene was sequenced first by Celera, the gene was held with Celera's IP, and a variety of institutions paid substantial fees to access Celera's sequencing data even though Celera genes would move into the public domain once re-sequenced by the public effort. Celera's contract law-based (rather than patent-based) IP applied for a maximum of two years, with all Celera genes moving into the public domain by the end of 2003.

From this empirical context, I construct two research designs to test for the effects of Celera's IP on subsequent scientific research and product development outcomes. The first research

design tests whether genes that ever had Celera's IP differ in subsequent innovation, as of 2009, from genes initially sequenced by the public effort. Any observed differences in this cross-section specification could be due to an IP effect, or to non-random selection of genes into Celera's IP on the basis of factors such as expected commercial potential. Historical accounts suggest such selective sequencing was relevant in the early years of the public effort, but was less relevant once the public effort was sequencing at full scale. Consistent with these historical accounts, comparing Celera and non-Celera genes based on *ex ante* observable characteristics provides evidence of selective sequencing by the public effort in the full sample; however, once I limit the non-Celera sample to genes sequenced by the fully-scaled public effort, Celera and non-Celera genes appear balanced on *ex ante* observable characteristics. To further address selection concerns, the second research design is a complementary panel analysis that traces the effects of removal of Celera's IP on within-gene flow measures of subsequent innovation.

My empirical analysis relies on a newly-constructed data set that traces out the distribution of Celera's IP across the human genome over time, linked to gene-level measures of scientific research and product development outcomes. Whereas in most contexts it is not straightforward to trace the path of basic scientific discoveries as they are translated into marketable products, I am able to construct my data at the level of naturally occuring biological molecules that can be precisely identified at various stages of the R&D process. Specifically, I trace cumulative technological progress by collecting data on links between genes and phenotypes, which represent the expression of a gene into a trait such as the presence or absence of a disease. For each gene, I collect data on publications investigating potential genotype-phenotype links, on successfully generated scientific knowledge about genotype-phenotype links, and on the development of gene-based diagnostic tests that are available to consumers.

Both the cross-section and panel specifications suggest Celera's IP led to economically and statistically significant reductions in subsequent scientific research and product development outcomes. Celera genes have had 35 percent fewer publications since 2001 (relative to a mean of 1 publication per gene). Based on two measures of successfully generated scientific knowledge about genotype-phenotype links taken from a US National Institutes of Health database, I estimate a 16 percentage point reduction in the probability of a gene having a known but scientifically uncertain genotype-phenotype link (relative to a mean of 30 percent), and a 2 percentage point reduction in the probability of a gene having a known and scientifically certain genotype-phenotype link (relative to a mean of 4 percent). In terms of product development, Celera genes are 1.5 percentage points less likely to be used in a currently available genetic test (relative to a mean of 3 percent). The panel estimates suggest similarly-sized reductions, on the order of 30 percent.

Taken together, these results suggest Celera's short-term IP had persistent negative effects on subsequent innovation relative to a counterfactual of Celera genes having always been in the public domain. The panel estimates measure a transitory effect of Celera's IP, and suggest that innovation on Celera genes increased after Celera's IP was removed. However, the cross-section estimates measure more persistent effects and suggest that Celera genes have not "caught up"

by the end of my data in 2009 to *ex ante* similar genes that were always in the public domain. One interpretation of these results is as suggestive evidence of increasing returns to R&D. That is, to the extent that existing stocks of scientific knowledge provide ideas and tools that allow future discoveries to be achievable at lower costs, the production of new knowledge may rise more than proportionately with the stock. Celera genes appear to have accumulated lower levels of scientific knowledge during the time they were held with IP, and these temporarily lower levels of innovation may have led the accumulation of new scientific knowledge to be relatively more costly on Celera genes even after Celera's IP was removed.

It is important to note that this analysis is not evaluating the overall welfare effects of Celera's entry into the effort to sequence the human genome. To the extent that Celera's entry spurred faster completion of the public sequencing efforts, Celera's entry likely shifted the overall timing of genome-related innovation earlier, which would have had welfare gains even if Celera's IP in isolation ended up hindering innovation. More generally, the overall welfare effects of IP depend on factors beyond the impact of IP on subsequent innovation, including the provision of dynamic incentives for innovation. Rather, these results suggest that, holding Celera's entry and sequencing efforts constant, an alternative institutional mechanism may have had social benefits relative to Celera's chosen form of IP. For example, under the patent buyout mechanism discussed by Kremer (1998), the public sector (or another entity) could have paid Celera some fee to "buy out" Celera's IP and place Celera genes in the public domain.[1]

The question of how intellectual property affects subsequent innovation will almost certainly have different answers in different contexts. My results are most directly relevant for assessing the role of gene-related IP in realizing the full potential of genetic medicine.[2] Prior to its completion, the sequenced human genome was likened by scientist Walter Gilbert to the Holy Grail (Duenes, 2000), and called by scientist Eric Lander "the 20th century's version of the discovery and consolidation of the periodic table" (Lander, 1996). Yet today, many argue that the medical and scientific advances realized because of the sequencing of the human genome have not fulfilled these grand expectations (Wade, 2009). Although scientific factors are surely important in explaining this fact, these results suggest institutions may also have played an important role. Given that the design of Celera's IP is similar to IP used by other science-based firms in attempts to provide returns to investors, the effects observed in this context may be expected to generalize to firms using similar packages of IP in related markets. Also relevant, although difficult to assess, is whether the intensity of research activity in the Human Genome Project is representative of open access research institutions.

These results join a handful of recent studies in suggesting that open access to scientific materials may encourage cumulative innovation (Furman and Stern, 2010; Murray and Stern,

[1] Kremer (1998) proposes an auction mechanism for determining the price in such a patent buyout. See Kremer and Williams (2010) for further discussion of other alternative mechanisms for rewarding innovation.

[2] Previous survey-based research has provided conflicting evidence on this question. Walsh, Arora and Cohen (2003a,b) present survey evidence suggesting "working solutions" to gene patents (*e.g.* patent infringement) are common, and that gene patents have generally not interfered with innovation on "worthwhile projects." However, Cho et al. (2003) present survey evidence suggesting gene patents may have hampered genetic test development.

2007; Murray et al., 2008).[3] One limitation of these papers is that they are constrained to examine only publication-related outcome variables - focusing on citations to scientific papers as an outcome. Hence, an important contribution of this paper is to trace out whether differences in scientific publications translate into differences in the availability of commercial products. Although these measured changes in the space of products available to consumers clearly have some link to social welfare, measuring the social value of these new medical technologies is difficult due to the potential inefficiencies in health care markets generated by asymmetric information and other factors.

The paper proceeds as follows. Section 1 presents a conceptual framework for the analysis. Section 2 provides a brief scientific background, and describes the public and private sequencing efforts. Section 3 describes the data, and Section 4 presents the empirical framework. Section 5 presents the empirical results, and Section 6 concludes.

1 Conceptual framework

To clarify why the sign of any effect of IP on subsequent innovation is theoretically ambiguous, consider a firm such as Celera that holds a set of upstream technologies (here, genes). If the upstream technologies are in the public domain, any firm can freely develop downstream products, so firms with ideas for downstream products will invest as long as their costs are less than expected profits. Alternatively, if the upstream technologies have IP, firms with ideas for downstream products must obtain a license from the upstream firm.[4]

In this conceptual framework, I focus attention on two factors: whether appropriate licensing agreements can be reached, and the strength of IP in downstream markets.[5] First, in a classic Coasian framework upstream and downstream firms can always negotiate licensing agreements such that R&D on downstream products is not hindered (Coase, 1960; Green and Scotchmer, 1995). However, licensing agreements could break down due to transaction costs, in which case IP on the upstream technologies could deter R&D on downstream products. For example, Bessen (2004) extends the Green and Scotchmer framework to show that if the downstream firm's research costs are private information, the optimal licenses may not be offered, and socially desirable R&D investments may be deterred.[6] Second, if IP protection in downstream markets

[3]Furman and Stern (2010) use a difference-in-difference approach to analyze shifts in biologic materials across institutional settings, and find that deposition of biomaterials in research enhancing institutions increased citations to the scientific discoveries. Murray and Stern (2007) use data on life sciences technologies, and find that patent grants decrease citations to scientific papers on the patented technology, relative to scientific papers on similar non-patented technologies. Finally, Murray et al. (2008) find that the removal of IP restrictions on certain types of genetically engineered mice increased citations to scientific papers on affected mice relative to scientific papers on unaffected mice; Murray et al. (2008) also provide evidence, consistent with the model of Aghion, Dewatripont and Stein (2008), that IP reduces the diversity of scientific experimentation.

[4]In many markets, downstream firms may need to obtain not one but rather multiple licenses; this case is the classic Cournot complements problem (Cournot, 1838), highlighted in the biotechnology context by Heller and Eisenberg (1998) and discussed by Shapiro (2000).

[5]IP may also affect subsequent innovation for other reasons; see, *e.g.*, Arora, Fosfuri and Gambardella (2001), Hellmann (2007), and Merges and Nelson (1990).

[6]Bessen's analysis relates to the classic work of Myerson and Satterthwaite (1983), which highlights that private information may induce inefficiencies in bilateral exchange mechanisms. Gans and Stern (2000) discuss one reason

is otherwise imperfect, IP on the upstream technologies could encourage R&D on downstream products by increasing firms' ability to capture rents in downstream product markets. Perhaps the best-known example of this argument was made by Kitch (1977), whose "prospect theory" of patent rights argued that inventors would not invest in putting their innovation to efficient use unless they obtain exclusive rights to their invention, for fear that their investment will produce upatentable information appropriable by competitors.[7] Intuitively, the relevant trade-off is that with IP downstream firms may lose profits to the upstream firm's licensing fee, whereas without IP downstream firms may lose profits to increased competition. The overall effect of IP on subsequent innovation depends on the relative magnitudes of these two effects.

Imperfect IP protection in downstream markets is relevant for my gene-based diagnostic test outcome, since diagnostic method patents generally provide only weak IP protection to diagnostic test innovators.[8] In contrast, consider the example of Myriad Genetics: Myriad holds patents on two genes with links to increased risks for breast and ovarian cancer (BRCA1 and BRCA2) that grant the firm exclusive monopoly rights to all diagnostic testing related to these genes.[9] The idea that IP might encourage subsequent innovation also underlies public policies such as the US Bayh-Dole Act, which aim to spur the translation of academic discoveries into marketable products by encouraging academics to patent discoveries resulting from federally-funded research.[10]

2 Background: Sequencing of the human genome

This section reviews the scientific background and institutional context necessary to understand the construction of the data and the design of the empirical specifications.[11]

why Celera might have been in an unusually strong bargaining position: namely, because Celera *itself* marketed gene-based diagnostic tests, Celera's threat to engage in imitative R&D during licensing negotiations may have increased its bargaining power.

[7]The underlying concern is that this unpatentable information will allow imitators to enter the downstream market, and that anticipation of the entry of those imitators will prevent innovators from ever making the initial R&D investments in the first place. Since Kitch's classic article, other so-called *ex post* justifications for IP have been made, for example by Landes and Posner (2003). For a recent overview of *ex ante* and *ex post* justifications for IP, see Lemley (2004).

[8]This weak IP protection reflects patent policy as well as technological characteristics of diagnostics, in the sense that the fixed R&D costs facing potential imitators are quite low. In terms of patent policy, once a genotype-phenotype link has been documented, there are frequently many similar but "different enough" ways (from the US Patent and Trademark Office's perspective) to test for the link. Pitcher and Fairchild (2009) discuss some specific examples. In terms of low fixed R&D costs, Cho et al. (2003) note: "...it may only take weeks or months to go from a research finding that a particular genetic variant is associated with a disease to a clinically validated test."

[9]Currently, no BRCA-related diagnostic tests can be conducted outside of Myriad's lab, and no alternative tests related to the BRCA genes can be offered (Schwartz, 2009). The status of Myriad's gene patents is currently in flux after an April 2010 US District Court decision invalidating seven patents related to the BRCA1 and BRCA2 genes.

[10]The type of IP encouraged under Bayh-Dole clearly does not aim to encourage R&D investments on the original scientific discoveries, since the federal government is already itself funding the research on those initial discoveries. For more on the US Bayh-Dole Act, see Mowery et al. (2004).

[11]For more extensive discussions of the public and private sequencing efforts, see Cook-Deegan (1994), Davies (2001), Shreeve (2005), Sulston and Ferry (2002), Venter (2007), and Wade (2001).

2.1 Scientific primer on the human genome

A genome is essentially a set of instructions for creating an organism. In humans, two sets of the human genome are contained inside the nucleus of basically every human cell.[12] Each copy of the human genome is composed of deoxyribonucleic acid (DNA), and contains approximately three billion nucleotide bases - adenine (A), cytosine (C), guanine (G), and thymine (T). DNA sequencing is the process of determining the exact order of these bases in a segment of DNA.

The DNA of the human genome is organized into forty-six chromosomes - twenty-two pairs of autosomes (numbered 1 to 22) together with two sex chromosomes (X and Y chromosomes in males, or two X chromosomes in females). Chromosomes are the cellular carriers of genes, and in total the human genome is currently estimated to include approximately 28,000 genes. With some exceptions, genes encode instructions for generating proteins, which in turn carry out essential functions within the human body. Genes manufacture proteins through a two-step process of transcription and translation. In the transcription process, a messenger ribonucleic acid (mRNA) transcript is generated.[13] In the translation process, the mRNA transcript is used to generate a protein. Genes are able to encode more than one protein through generating more than one mRNA transcript.[14]

Appendix Figure A1 graphically summarizes this scientific background. As suggested by this figure, once a segment of DNA has been sequenced, genes and mRNAs can be identified, as well as the proteins for which they code.[15] Intuitively, the meaningful unit for tracking the sequencing efforts is the mRNA level, since each mRNA encodes exactly one protein (as opposed to genes, which can encode more than one protein), and proteins are what carry out functions within the human body. Reflecting this, the data will track the public and private sequencing efforts (as well as Celera's IP) at the mRNA level.[16]

2.2 The sequencing of the human genome

The public sequencing effort, known as the Human Genome Project (HGP), was first proposed by the US Department of Energy (DOE) in the late 1980s, and later jointly launched between the DOE and the US National Institutes of Health (NIH) in 1990.[17] The public effort was headed by James Watson and later Francis Collins, and originally aimed to be complete by 2005.

In May 1998, a new firm - Celera, led by scientist Craig Venter - was formed, with an intention

[12] The exceptions are egg and sperm cells, each containing one set, and red blood cells, containing no sets.

[13] In recent years the exact definition of a gene has become less clear (see, *e.g.*, Snyder and Gerstein (2003)). For example, two genes can sometimes generate a single, fused mRNA transcript (Parra et al., 2006). My use of the term "gene" will become clear in the context of the data, described in Section 3.2.

[14] In the data, the mean number of known mRNAs per gene is 1.67, the median is 1, and the range is [1,23].

[15] The process of identifying genes and mRNAs from a stretch of human DNA is not always straightforward, and improved methods for identifying genes and mRNAs continue to evolve as of today.

[16] By basing the analysis on mRNA-level data, I focus on those portions of the human genome that generate proteins, avoiding so-called "junk DNA" that does not code for proteins. I do not know of any data sources that would allow measurement of innovation on non-protein coding portions of the human genome.

[17] Roberts (2001) notes that to the DOE the HGP represented a "...logical outgrowth of DOE's mandate to study the effects of radiation on human health;" others (most notably, biologist David Botstein) argued the DOE's effort was a scheme to provide new focus for "unemployed bombmakers."

of sequencing the human genome within three years (Venter et al., 1998). Celera's business model included sales of databases containing sequenced DNA (to pharmaceutical companies, universities, and research institutes) as well as revenues from genes on which Celera obtained intellectual property (Service, 2001). Note that database subscribers paid to access Celera's data even though in expectation the human data would soon be in the public domain. Shreeve (2005) quotes Craig Venter as saying: "*Amgen, Novartis, and now Pharmacia Upjohn have signed up knowing damn well the data was going to be in the public domain in two years anyways. They didn't want to wait for it.*" Although the terms of specific deals were private, Service (2001) reports that pharmaceutical companies were paying between $5 million and $15 million a year, whereas universities and nonprofit research organizations were typically paying between $7,500 and $15,000 for each lab given access to the data.

In September 1998, the public sector announced a revised plan to complete its sequencing efforts by 2003 (Collins et al., 1998); in March 1999 the plan was again revised, aiming to complete a "draft" sequence of the human genome by spring 2000 (Pennisi, 1999). Departing from its previous goal of producing near-perfect sequence, the aim of this draft sequence was to place most of the genome in the public domain as soon as possible. Although Marshall (1998) quotes Francis Collins as claiming this change was not in response to Celera's entry ("*This is not a reaction. It is action.*"), many observers viewed this scale-up as a result of Celera's entry.

The two efforts agreed to jointly publish their draft genomes in 2001, the public effort's draft genome in the journal *Nature* (Lander et al., 2001) and Celera's draft genome in the journal *Science* (Venter et al., 2001). Celera's human genome sequencing effort stopped with this publication, whereas the public effort continued and was declared complete in 2003.

2.3 Sequencing strategies

Given that the empirical strategies will use variation in the timing of when genes were sequenced by Celera and the Human Genome Project, it is relevant to describe each side's stated sequencing strategies - in terms of both structural characteristics and scientific approaches.[18]

In terms of structural characteristics, Celera's human genome sequencing effort was concentrated in one Maryland-based center, and was initiated in September 1999. For the public sector, a number of structural features are relevant. First, the public effort chose to pursue a "map first, sequence later" strategy, focusing first on mapping the general location of genes relative to each other, and only later sequencing the precise order of nucleotide base pairs.[19] Thus, even though the public effort officially commenced in 1990, almost all of the public effort's sequence data was produced over roughly fifteen months, starting in mid-1999 (Lander et al., 2001). Prior to this full-scale sequencing that started in 1999, parts of the public effort were targeting the

[18] Because all humans share the same basic set of genes, the question of "whose genome" was sequenced is not relevant for the analysis. The public effort collected blood or sperm samples from a large number of donors, although only a few samples were processed as DNA resources. Celera used samples from several donors, and Craig Venter later acknowledged that his DNA was among those donors.

[19] One cited motivation for this delay in sequencing was to allow a few years for the development of more efficient and affordable DNA sequencing technologies.

sequencing of some specific genes of medical interest, such as the gene linked to Huntington's disease.[20] Second, the public effort took a "divide and conquer" approach, dividing the genome into separate chromosomes (or pieces of chromosomes) and dividing these among research labs throughout the US and abroad.[21] Shreeve (2005) quotes a head of one major public lab as saying this approach was "stunningly inefficient," in that each lab had to discover and solve the same problems separately, and potentially reflecting this view the public sequencing effort was eventually consolidated into a small group of four labs.[22]

In terms of scientific approaches, the broad scientific methods used by each side share several characteristics, but also some important differences. The primary DNA sequencing technique used by both Celera and the Human Genome Project was first developed by Frederick Sanger and colleagues in 1977.[23] Shortly after, the so-called shotgun sequencing method was introduced, in which DNA is randomly broken up into smaller segments that are then sequenced and re-assembled. Since its introduction, the shotgun method has remained the fundamental method for large-scale genome sequencing (Lander et al., 2001), and is thus itself uncontroversial. However, the two sectors differed in how they applied the shotgun method.

The public Human Genome Project pursued a hierarchical shotgun sequencing approach, which involved generating a set of genome fragments that together covered the genome, separately shotgun sequencing each fragment, and then reassembling. This approach required a relatively larger initial investment (in generating the fragments), but was argued to be easier at the assembly stage since sequenced DNA was local to a known fragment. Celera instead pursued a whole-genome shotgun sequencing approach, which involved shredding the entire genome, sequencing the fragments, and then re-assembling. This approach avoided the initial investment needed under the public approach, but because of the high frequency of repeat sequences on the human genome was argued to be more difficult in the assembly stage given the lack of local information on where sequenced pieces fit.[24] The arguments over the scientific validity of the whole-genome shotgun approach as applied to the human genome grew quite heated, but were

[20] See The Huntington's Disease Collaborative Research Group (1993). Intuitively, DNA sequencing can be done in two ways: first, scientists can sequence a set of nucleotide bases and use that sequence to identify genes and begin studying gene functions; second, scientists can take a gene with suspected function and approximate known location on the genome, and purposefully target finding and sequencing the DNA underlying that gene. Under the first model, we would not expect targeting (since genes are only identified after the DNA itself is sequenced), but under the second model targeting is relevant.

[21] Lander et al. (2001) note that most centers focused on particular chromosomes or, in some cases, larger regions of the genome. Sulston and Ferry (2002) note that the public effort explicitly took steps to avoid letting researchers "cherry-pick" sections of the genome to sequence that were more likely to contain important genes.

[22] The US effort was concentrated in three centers: Richard Gibbs's team at Baylor College of Medicine; Eric Lander's team at the Whitehead Institute for Biomedical Research; and Robert Waterston's team at Washington University in St. Louis. The major international center was John Sulston's team at the Sanger Centre near Cambridge, UK.

[23] See Sanger, Nicklen and Coulson (1977). An alternative technique was independently developed in the same year by Allan Maxam and Walter Gilbert (Maxam and Gilbert, 1977); Gilbert and Sanger shared (together with Paul Berg) the 1980 Nobel Prize in Chemistry for these advances.

[24] The whole-genome shotgun approach was first proposed to be applied to the human genome in 1997 (Weber and Myers, 1997), and immediately came under harsh criticism (Green, 1997) highlighting this and other concerns. In the end, Celera combined its own data with some of the public effort's sequence data (which was publicly available, as described in Section 2.4) in forming its assembly of the human genome.

centered on concerns over gaps in Celera's assembled genome, *not* on the quality of Celera's fragments conditional on having been sequenced (in terms of actual mistakes in the set of sequenced nucleotide bases). Thus, differences in the quality of sequenced DNA do not appear to be a major issue for my empirical analysis.

This institutional context informs the empirical work in several ways. In its early years (specifically, pre-2000), the public effort was sequencing a relatively small number of specific genes of medical interest. Having such genes in the data should imply that genes sequenced first by the public effort were *ex ante* more commercially attractive - which indeed is true in the full sample. Although such targeted sequencing was not irrelevant in later years, from 2000 forward both Celera and the fully-scaled public effort were relying on variants of the "shotgun" DNA sequencing approach that induced some random variation in when specific genes were sequenced. Data limitations prevent me from being able to perfectly separate genes sequenced under what I am referring to as the "public sector effort" from various independent efforts to sequence this relatively small number of specific genes of medical interest. However, limiting non-Celera genes to those sequenced from 2000 forward is a first step towards addressing this form of selection, providing a sample that appears balanced on *ex ante* gene-level observables. Limiting the sample of non-Celera genes to those sequenced from 2000 forward is also attractive in that it focuses on the "risk set" of genes that Celera could have sequenced, removing genes that were sequenced before Celera began its sequencing effort. Finally, selection was not practically feasible on a large scale since the vast majority of genes had unknown functions at the time of sequencing. Section 4 discusses in more detail how my empirical strategies are oriented to address selection issues.

2.4 Intellectual property strategies

For genes sequenced by the public effort, the relevant intellectual property regime is the set of so-called "Bermuda rules." In 1996, the heads of the largest labs involved in the public effort agreed (at a Bermuda-based meeting) to these rules as a set of guidelines for data sequenced under the public effort. The Bermuda Rules applied to all stretches of DNA longer than 1000-2000 nucleotide bases, and required data to be submitted to the public online database GenBank within twenty-four hours of sequencing.[25] The stated goal of the Bermuda rules was that "*...all human genomic DNA sequence information, generated by centers funded for large-scale human sequencing, should be freely available and in the public domain in order to encourage research and development and to maximize its benefit to society.*" Eisenberg (2000) discusses how the Bermuda rules may also have been motivated by a desire to discourage gene patenting by public researchers (as the accelerated timetable made it difficult for grantees to file patent applications before public disclosure) as well as to discourage gene patenting by others (as public disclosure

[25]Marshall (2001a) notes that the Bermuda Rules replaced a US policy that data should be made available within six months, although as discussed in Section 2.3 sequencing efforts did not begin in earnest until the late 1990s, at which time the Bermuda Rules were already in place.

creates so-called prior art that could defeat potential patent claims by other researchers).[26]

When a gene had been sequenced by Celera but not yet sequenced by the public effort, the gene was held with Celera's chosen form of IP. Although the implementation of Celera's IP was tailored to the specifics of this market, at its core the goal of Celera's IP strategy was similar to other forms of IP, in that it aimed to use excludability to provide returns to investors. The details of Celera's IP strategy are described in more detail in Appendix 1, but the key features were restrictions on redistribution of Celera's data (aiming to prevent other commercial firms from directly copying the data for use in either products or product development), and a requirement that individuals wanting to use the data for commercial purposes negotiate a licensing agreement with Celera. Celera's data were disclosed with the 2001 publication of Celera's draft genome in *Science*, in the sense that any individual could view data on the assembled genome through the Celera website, or by obtaining a free data DVD from the company.[27] Academic researchers were free to use the Celera data for non-commercial research purposes. This package of Celera IP comprises the intellectual property "treatment" I focus on in this paper.[28]

2.5 Gene patents

Although not my IP treatment of interest, it is important to clarify the role of gene patents in the analysis. Jensen and Murray (2005) provide a detailed analysis of gene patenting on the human genome as of 2005, estimating that nearly 20 percent of human genes were explicitly claimed under patents as of that date. Although the majority of patents (63 percent) were held by private firms (such as Incyte Pharmaceuticals), 28 percent were held by public institutions.[29] While patents have been an important and controversial form of intellectual property on the human genome, the effects of gene patents are unclear for several reasons. First, what the US Patent and Trademark Office (USPTO) has or has not allowed to be patented has changed dramatically over time (see the discussion in National Academy of Sciences (2006)). For example, reflecting concerns that USPTO patent examiners had become overly lax in their granting of gene patents, in 2001 a set of guidelines was issued that effectively raised the utility standards for gene patents.[30] Second, there has been substantial variation over time in the judicial enforcement of

[26]These rules are formally described in detail by various policy statements, such as the 1996 document by the US National Human Genome Research Institute (NHGRI) that applied to human DNA sequenced under the public effort's pilot sequencing grants. In terms of enforcement, NHGRI grantees were required to adopt this policy as a condition of the grant awards. NHGRI policy statements also explicitly discouraged patenting of large blocks of primary human genomic DNA sequence, and suggested that NHGRI would actively monitor grantee activity to discourage such patenting. Marshall (2001a) notes that US officials made clear at the time that failure to abide by the Bermuda Rules "...could be a black mark in future grant reviews."

[27]Viewing the assembly online or obtaining the data DVD required an agreement to neither commercialize nor distribute the data.

[28]By 2001, it had been announced that the public Human Genome Project aimed to complete its sequencing efforts - finishing what was left unfinished in the 2001 draft genome - by 2003, and in fact met this deadline. In this sense, Celera's IP expiration was similar to a patent expiration: there was a known maximum date of 2003 (similar to a fixed patent expiration date), and some uncertainty within that time frame of the date at which Celera's IP would be removed (similar to the threat of patent litigation as discussed by Jaffe and Lerner (2006), among others). Celera also filed for gene patents, but these patents were generally not granted.

[29]Nine percent of patents were held by "unclassified" institutions.

[30]See USPTO, *Utility Examination Guidelines*, 66 Fed. Reg. 1092 (5 January 2001).

existing patents. For example, the 2001 USPTO guidelines referenced above were later upheld upheld by a Federal Circuit court in the *In re Fisher* case, which changed the enforcement of many existing patents.[31] Finally, there were reports of some researchers filing gene patents themselves (and licensing at zero cost) to prevent commercial firms from obtaining the patents (and presumably licensing at above-zero costs). I instead focus attention on Celera's IP.[32]

3 Data

3.1 Conceptual issues in data construction

Several units of analysis are relevant for the empirical work: mRNAs, genes, and genotype-phenotype links. Before describing the specific data sets used in the analysis, it is worth addressing three questions on how these units of analysis conceptually relate to the data construction. First, what is the appropriate unit of observation for tracking the sequencing efforts? Second, what is the appropriate unit of observation for measuring economically meaningful outcome variables? And third, what is the appropriate unit of analysis for the empirical work?

First, as discussed in Section 2.1, one gene may produce more than one mRNA transcript, in which case the gene encodes instructions for generating more than one protein. This motivates that the mRNA is the meaningful unit for tracking the sequencing efforts, since each mRNA encodes exactly one protein, and proteins are what carry out functions within the human body. Reflecting this, the data will track the public and private sequencing efforts (as well as Celera's IP) at the mRNA level.

Second, the genotype-phenotype level of observation is what is relevant for measuring economically meaningful outcome variables. A genotype-phenotype link is what is relevant to human health since it represents the link between a gene and an observable trait or characteristic, such as the presence or absence of a disease. For example, the known link of the Huntingtin gene to Huntington's disease represents a genotype-phenotype link. Genotype-phenotype links, once identified, can then be used in combination with a sequenced gene to form the basis for genetic tests. One gene can be involved in more than one genotype-phenotype link, and one genotype-phenotype link can involve more than one gene. Genotype-phenotype-level data can be collapsed to the gene level to measure the total volume of innovative activity relevant to a given gene across all genotype-phenotype links in which it is known to be involved.

Third, I use the gene as the level of analysis for the empirical work. Intuitively, a gene is a stable scientific unit, whereas both the number of known mRNAs and the number of known genotype-phenotype links relevant to a given gene are likely functions of the amount of research effort invested in the gene.

[31] See *In re Fisher*, 421 F.3d 1365 (Fed. Cir. 2005). For an overview of gene patent litigation, see Holman (2007, 2008). More recently, in April 2010 a US District Court issued a decision invalidating seven patents related to the BRCA1 and BRCA2 genes discussed above.

[32] I am unaware of any data on gene patents that can be reliably matched to my data. Thus, although both Celera and non-Celera genes were at risk for patenting, I am unable to examine patenting as either an outcome or as a potential mechanism for the observed Celera IP effects.

To summarize, I use mRNA-level data to track the sequencing efforts as well as Celera's IP, aggregate this mRNA-level data to the gene level, and then link this gene-level variation in IP to gene-level measures of the total volume of innovative activity relevant to a given gene across all genotype-phenotype links.

3.2 Data sources

This section provides an overview of the specific data sets used in my analysis.[33]

To track the sequencing efforts as well as Celera's IP, I use mRNA-level data as follows. I track the public sequencing efforts at the mRNA-by-year level for 1999 forward using the US National Institutes of Health's (NIH) RefSeq database, which is used internationally as the standard for genome annotation. Tracking Celera's sequencing effort is less straightforward, requiring a comparison of the Celera data with the public sequencing data at a point in time.[34] Fortuitously for my work, a publication by Istrail et al. (2004) in the *Proceedings of the National Academy of Sciences* journal provides one such snapshot, comparing the Celera data with the NCBI-34 (October 2003) release of the public sequencing data. Using an archived version of the NCBI-34 data together with Istrail et al. (2004)'s analysis, I construct an mRNA-by-year level variable for whether a given mRNA was included in the Celera data but had not yet appeared in the public sequencing data.

My outcome variables are drawn from two NIH databases: the Online Mendelian Inheritance in Man (OMIM) database and the GeneTests.org database. OMIM aims to provide a comprehensive catalog of human genes and genetic phenotypes. From OMIM-assigned classifications, I construct two proxies for the level of "scientific knowledge" about genotype-phenotype links. First, I construct an indicator variable for the existence of a genotype-phenotype link known with some (potentially low) level of scientific certainty (which I refer to as a "known, uncertain phenotype"). Second, I construct an indicator variable for the existence of a genotype-phenotype link known with a higher level of scientific certainty (which I refer to as a "known, certain phenotype"). OMIM records cite published scientific papers relevant for each record, which I collect as an additional outcome variable. OMIM is considered the authoritative database on genotype-phenotype links, and is widely used by genetic researchers as well as physicians (Uhlmann and Guttmacher, 2008).

The GeneTests.org database includes a self-reported, voluntary listing of US and international laboratories offering genetic testing. From GeneTests.org, I construct an indicator for the availability of any genetic test related to a given genotype-phenotype link. GeneTests.org is *not* a comprehensive listing of genetic testing facilities, but is the most common genetic testing directory referenced in literature oriented towards both physicians and patients (Uhlmann and Guttmacher, 2008).

[33] For more details on the data used in my analysis, see Appendix 2. To the best of my knowledge, most of these data sets have not previously been used in the economics literature. The exception of which I am aware is Moon (2008), who uses some variables from the Entrez Gene and OMIM databases in a study of the impact of control rights on decisions over publishing and patenting.

[34] There was essentially one "version" of the Celera data - namely, the 2001 draft genome (Venter et al., 2001).

For two of my outcome variables I am able to construct gene-by-year measures for use in the panel specification. First, I use paper publication dates to construct the number of publications by gene by year. Second, and less straightforward, I construct the first date each "known, uncertain phenotype" link appears in OMIM. I observe this latter measure with error, but expect this error to be uncorrelated with Celera's IP.[35]

Finally, as discussed in Section 3.1, an important issue is how I aggregate my mRNA-level data and collapse my genotype-phenotype-level data to construct the gene-level data used in the analysis. First, I aggregate my mRNA-level Celera IP variable to a gene-level indicator for whether all known mRNAs on a gene were initially sequenced by Celera. Other gene-level definitions of the Celera IP variable, such as the share of known mRNAs that were Celera mRNAs, or an indicator for whether any mRNA on the gene was a Celera mRNA, are identical for the majority of genes that have only one known mRNA, and as expected generally yield similar results. Second, I collapse the genotype-phenotype-level measure of publications by summing the total number of publications related to a gene across all genotype-phenotype links. Finally, I collapse the binary genotype-phenotype-level indicators for scientific knowledge and genetic test availability by taking the maximum value for each gene across all genotype-phenotype links, thus generating variables representing (for example) whether a gene is used in any currently available diagnostic test.

3.3 An example

To clarify the data construction, I briefly discuss one example. The mRNA transcript with RefSeq identification number *NM_032753.3* first appeared in the RefSeq database in 2001, and based on the analysis of Istrail et al. (2004) was never held with Celera's IP. This is the only known mRNA encoded by the RAX2 gene, located on chromosome 19.

Looking in OMIM, the RAX2 gene is included in two genotype-phenotype entries, both of which were documented in 2006 (based on a 2004 publication in the journal *Human Molecular Genetics*) and are classified by OMIM as being scientifically certain. First, the RAX2 gene is listed in OMIM entry +610362 for a link to age-related macular degeneration, a medical condition arising in older adults that destroys the type of central vision needed for common tasks such as driving, facial recognition, and reading. Second, the RAX2 gene is listed in OMIM entry #610381 for cone-rod dystrophy, an eye disease tending to cause vision loss, sensitivity to bright lights, and poor color vision.

Looking in GeneTests.org, a genetic test for RAX2's link to age-related macular degeneration is available at several testing facilities (including some academic medical centers as well as the Nichols Institute of the for-profit firm Quest Diagnostics). There are no such listings for genetic tests for RAX2's link to cone-rod dystrophy.[36] The results of a genetic test for RAX2's

[35] See Appendix 2 for details on this measurement error.

[36] A non-exhaustive internet search revealed that a genetic test for RAX2's link to age-related macular degeneration is also available from at least one testing facility not listed in the GeneTests.org directory (namely, the firm 23andMe), consistent with the note in Section 3.2 that GeneTests.org is *not* a comprehensive listing of genetic testing facilities. At least in this case, despite not being a comprehensive directory, GeneTests.org appears to be

link to age-related macular degeneration are likely valuable to consumers in part because several preventive health behaviors can reduce an individual's risk of developing age-related macular degeneration, including dietary adjustments and a specific combination of vitamin supplements.[37]

Whereas in most contexts it is not straightforward to trace the path of basic scientific discoveries as they transition from lab to market, as this example clarifies I am able to construct my data at the level of naturally occuring biological molecules that can be precisely identified at various stages of the R&D process. Moreover, the outcomes used in the analysis are drawn from the same data sets used by scientific researchers and medical professionals - providing comfort that I am capturing scientifically and economically relevant outcomes.

Finally, an important question is whether the outcome variables are measuring real differences in the amount of scientific research being conducted, or measuring differences in the amount of scientific research that is being disclosed. If academic and public researchers face higher incentives to disclose the results of their research than do private researchers, and if Celera's IP induced an increase in the share of research done by private researchers, then observed differences in my scientific publication and scientific knowledge outcomes could in part be explained by differences in disclosure. However, the product development outcome - diagnostic test availability - should be invariant with respect to disclosure preferences of researchers that could affect the other outcome variables.[38] In addition, disclosure itself has social value, and to the extent that IP induces reductions in disclosure this effect is also relevant in measuring the effects of IP.

4 Empirical framework

To motivate the design of the empirical specifications, this section presents some descriptive statistics and analyses attempting to understand selection of genes into Celera IP. I then describe the empirical specifications, with a focus on attempting to address selection issues.

4.1 Descriptive statistics

Table 1 presents descriptive statistics on the Celera IP treatment variable, outcome variables, and covariates for the gene-level data. Of the approximately 46,000 currently known mRNA transcripts on the human genome, 3,062 were sequenced only by Celera as of 2001. Aggregating this IP variable to the gene level, of the 27,882 currently known genes on the human genome, 1,682 genes were held (that is, all mRNAs on the gene were held) with Celera IP for some amount of time. As reflected in Panel A of Table 1, this implies that the mean of the Celera IP treatment variable is approximately 6 percent.

sufficient to accurately capture the availability of a genetic test. I did not find any non-GeneTests.org testing facilities offering tests for RAX2's link to cone-rod dystrophy, although such facilities may of course exist.

[37] See `http://www.nei.nih.gov/health/maculardegen/armd_facts.asp`.

[38] The exceptions to this statement are a few firms, such as the firm 23andMe, which do not list their genetic tests in the GeneTests.org directory. However, to the extent that such companies offer tests based on publicly available research - as suggested in a recent article by Ng et al. (2009) - my diagnostic test outcome should be sufficient to capture the availability of a diagnostic test.

As discussed in Section 2.3, Celera's human genome sequencing efforts commenced in September 1999, and its draft human genome was disclosed in 2001. Unfortunately, I do not observe the timing of when specific genes were sequenced within this time frame. In the absence of such data, I label all Celera genes as being disclosed in 2001. Although Celera scientists and a few "early subscriber" firms had access to some unknown number of intermediate data updates prior to 2001, my reading of the historical accounts of Celera's sequencing effort suggest the release of Celera's draft genome in 2001 represented the release date for the majority of the data. To the extent that some Celera genes are mis-coded as having a 2001 disclosure date instead of a true, earlier disclosure date (such as late 1999 or 2000), this should positively bias the estimated Celera IP effect, working against the negative effect we will observe in the data.[39] All Celera genes were in the public domain by 2003, implying the maximum time a gene was treated with Celera IP is two years. On average, genes had their first mRNA disclosed in 2002 (see Panel C of Table 1), with a range from 1999 to 2009.[40]

I collect several sets of gene-level covariates to assess the presence and magnitude of selection into Celera IP. Intuitively, I would like to measure gene characteristics that were observable to scientists at the time of sequencing and may have been used to target the sequencing of specific genes of medical or commercial interest. Based on my reading of historical accounts of the efforts to sequence the human genome, two main factors seem relevant.

First, scientists may have targeted their sequencing efforts based on scientific knowledge that a specific disease has a genetic basis. For example, scientists have long known that Huntington's disease has a genetic basis, and likely searched for genes related to Huntington's disease more than genes related to conditions that were less well-understood. I proxy for this type of *ex ante* attractiveness of a gene using count variables for the number of scientific publications related to the gene in years 1970 and later.[41] In the benchmark set of controls, I include eight such variables for publications in each year from 1970 to 1977, because 1977 was the year in which DNA sequencing technologies were first developed, and thus differences in average gene-year publications post-1977 between Celera and non-Celera genes likely in part reflect increases in scientific publications that occur as a result of some non-Celera genes being sequenced. When I limit the sample to genes sequenced (for example) in or after 2000, I show results including these variables for 1970 through 1999.[42]

Second, scientists may have targeted their sequencing efforts based on a gene's (*ex ante* known) approximate location on the genome. For example, certain chromosomes (such as chromosome 19) were estimated to be more "gene-rich" than others, and scientists may in turn have targeted the sequencing of such chromosomes. As discussed in Appendix 2, I collect detailed variables on both types of gene location descriptors used by geneticists (namely, cytogenetic

[39] Consistent with this expected positive bias, if I code one "2000/2001" disclosure date variable for all Celera and non-Celera genes disclosed in either 2000 or 2001, my estimated negative effects of the Celera IP variable tend to increase in magnitude; see Appendix Table A9.

[40] Although some genes were sequenced prior to 1999, 1999 is the first year coded in the RefSeq database.

[41] There are relatively few publications in the data prior to 1970.

[42] In comparing Celera and non-Celera genes based on these covariates, or including these variables in the regressions, I stop in 1999 because as noted above some Celera genes were sequenced in 2000.

location and molecular location). However, as reflected in Panel D of Table 1, many genes are missing data on these covariates: 37 percent of genes are missing at least one cytogenetic location variable, and 6 percent of genes are missing at least one molecular location variable. As one descriptive analysis of these gene location variables, Appendix Figure A2 graphically presents the distribution of genes across chromosomes.[43]

Moving on to examine my outcome variables, Panel B of Table 1 presents summary statistics on the four outcome variables. First, in measuring scientific publications as an outcome, I focus on publications from 2001 to 2009. This avoids (as opposed to using "total publications" as an outcome variable) using an outcome variable that includes the 1970-1977 publication covariates, and also focuses on publications from a time period when all Celera genes had been sequenced. On average, genes have had 2 publications over this time period, with a relatively large standard deviation.[44] Second, 45 percent of genes have at least one known, uncertain phenotype link.[45] Third, a much lower (as expected) share - 8 percent - of genes have at least one known, certain phenotype link. Finally, 6 percent of genes are used in at least one currently available genetic test.

4.2 Analyzing selection into Celera IP treatment

In this section, I examine differences in gene-level observable variables across Celera and non-Celera genes, attempting to better understand the selection effects suggested by the qualitative discussion in Section 2.3.

Table 2 shows the outcome variables and covariates cut by the Celera IP treatment variable, presenting the mean values for non-Celera and Celera genes and the p-value of the difference in means, for three different groups of non-Celera genes: non-Celera genes sequenced in all years, in 2001, and in or after 2000. As motivated by the institutional details discussed in Section 2.3, the latter two samples of non-Celera genes attempt to isolate genes sequenced under the fully-scaled public sector sequencing effort, for which I expect less selection.

Panel A suggests large differences in innovation outcomes across Celera and non-Celera genes in the full sample (Columns (2) and (3)), with non-Celera genes having higher means on each outcome variable. These differences are generally smaller but still persist when I focus on non-

[43] As discussed by Scherer (2008), in terms of the number of nucleotide bases the autosomes (that is, chromosomes 1 to 22) are generally numbered according to size, from largest to smallest; on this scale, the X chromosome would generally lie between chromosome 7 and chromosome 8, and the Y chromosome would generally lie between chromosome 20 and chromosome 21. However, as is clear from Appendix Figure A2 and consistent with other analyses such as those by Scherer (2008), there is no such monotonic relationship in terms of the number of genes across chromosomes.

[44] Panel (a) of Appendix Figure A3 shows the number of total gene-year publications for all genes, by year, for 1970 to 2008; I exclude 2009 from this figure given the truncation of the data. Flow publications peaked by this measure in 2003, although it is likely that some of the post-2003 decline is due to time lags in the addition of scientific publications to the OMIM database. In the panel specifications using the gene-year level data, the inclusion of year fixed effects will remove any year-specific shocks to the overall level of publications that are common across genes, such as time lags in updating of the OMIM database.

[45] Panel (b) of Appendix Figure A3 shows the total number of genes that have at least one such known, uncertain phenotype link by year. I retain the 1970-2008 scale on the x-axis of this graph, even though I only observe this variable from 1986 forward, for comparability to the trend in Panel (a) of Appendix Figure A3.

Celera genes sequenced in 2001 (Columns (4) and (5)). When I examine non-Celera genes sequenced in or after 2000 (Columns (6) and (7)) - that is, including some genes sequenced in more recent years - these differences disappear, with Celera genes having slightly *higher* mean innovation outcomes. These higher levels of innovation outcomes for Celera genes relative to genes publicly sequenced in or after 2000 are likely in part due to Celera genes having been sequenced earlier and thus having been "at risk" for research for a longer period of time.[46]

Of course, these raw differences in mean outcomes for Celera and non-Celera genes may in part reflect non-random selection of genes into Celera's IP. To shed light on the selection of genes into Celera's IP, we can examine whether Celera and non-Celera genes look comparable based on *ex ante* characteristics that were fixed at the time the gene was sequenced. Looking at the covariates in Panel B of Table 2, we see substantial differences in mean pre-2000 publications across non-Celera and Celera genes in the full sample (Columns (2) and (3)), as expected from the discussion in Section 2.3. Selection appears reduced but still substantial when I focus on non-Celera genes sequenced in 2001 (Columns (4) and (5)), suggesting that conditional on fixed effects for year of disclosure selection issues will be a concern in my cross-section specification. When I examine non-Celera genes sequenced from 2000 forward (Columns (6) and (7)), as motivated by the discussion in Section 2.3, Celera and non-Celera genes now look balanced in mean pre-2000 publications. The differences in individual years are generally not statistically significant, with the exception of 1999.[47] In an ordinary-least-squares (OLS) model predicting an indicator variable for Celera IP treatment as a function of these count variables for publications in each year from 1970-1999 for the 2000 forward subsample, the p-value from an F-test for their joint significance is 0.177.

Panel C of Table 2 suggests Celera genes are much less likely to have missing data on cytogenetic and molecular location information. Because missing data on these location variables is an outcome of the amount of research effort invested in a given gene, I do not include these variables nor indicators for missing data on these variables in the main empirical specification. As one descriptive analysis, a two-sample Kolmogorov-Smirnov test for equality of the distributions of Celera and non-Celera genes across chromosomes does not reject that the two distributions are equal ($p = 0.100$).[48]

Appendix Table A1 presents one additional set of descriptive statistics, limiting the sample to Celera genes, and examining differences in the outcome variables and covariates cut by whether the Celera gene was re-sequenced by the public effort in 2002 or in 2003.[49] The "treatment" in this sub-sample is thus being held with Celera IP for one additional year. I discuss mean

[46] As is clear from Appendix Table A2, earlier dates of sequencing are strongly positively correlated with the outcome variables. This patterns likely reflects a combination of selection (that is, that more valuable genes were more likely to have been sequenced in earlier years) and that genes sequenced in earlier years have higher levels of innovation as of 2009 because they have been at risk for research for a longer period of time.

[47] I expect that the difference in 1999 likely arises because some genes coded in my data as having been sequenced in 2000 may have been sequenced in 1999.

[48] Consistent with this lack of observed differences in the distribution of Celera and non-Celera genes across chromosomes, when I limit the sample to genes with non-missing location data in Appendix Table A7, including this more detailed set of control variables as covariates does not substantially alter the estimated coefficients.

[49] As discussed in Section 4.1, all Celera genes had been re-sequenced by the public effort by the end of 2003.

differences in the outcome variables across these treatment and control groups in Section 5.3. Here, I simply highlight that these treatment and control groups appear balanced on *ex ante* gene-level covariates. In an OLS model predicting an indicator variable for a gene being re-sequenced by the public sector in 2003 as a function of the count variables for publications in each year from 1970-1999, the p-value from an F-test for their joint significance is 0.169.

In summary, using data on observable gene characteristics that scientists could have used to target their sequencing efforts, I find evidence consistent with selection based on these observables in the full sample, with the public sector having been more likely to sequence genes that were *ex ante* more commercially attractive. When I limit my sample to genes sequenced in the years when the public effort was operating at scale (namely, 2000 forward), Celera and non-Celera genes appear balanced on *ex ante* gene-level observables, which motivates my focus on this sub-sample of data in the main analysis. Finally, when I limit my sample to Celera genes and look at Celera genes re-sequenced by the public effort in 2002 versus 2003, the two groups of genes appear balanced on *ex ante* gene-level observables.

4.3 Cross-section specification

In the cross-section specification, for gene g, I estimate the following:

$$(outcome)_g = \beta(celera)_g + \lambda'(covariates)_g + \epsilon_g$$

The coefficient on the "*celera*" variable is the main estimate of interest. I focus attention on two sets of covariates. First, I include a set of indicators variables for the first year the sequence for any mRNA on the gene was disclosed, to control for variation in innovation outcomes across genes that is a function of the year in which genes were sequenced.[50] In the language of age-time-cohort effects, these indicator variables control for cohort effects in the sense of accounting for variation due to the year a gene was "born" (here, sequenced). Second, I include a set of eight count variables for the number of publications on each gene in each year from 1970 to 1977, to control for the *ex ante* attractiveness of a gene for medical or commercial purposes. In samples restricted to genes sequenced after 2000, I show robustness checks that include these publications variables for years through 1999.

My publications outcome variable naturally lends itself to count data regression models; I show results from pseudo-maximum likelihood Poisson models for this outcome.[51] For the

[50] Disclosure is defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4).

[51] The Poisson model is generally preferred to alternative count data models, such as the negative binomial model, because the Poisson model is more robust to distributional misspecification (Cameron and Trivedi, 1998; Wooldridge, 2002). As long as the conditional mean is correctly specified, maximum likelihood estimation of the Poisson model will be consistent even if the data generating process is misspecified. Valid statistical inference in the Poisson maximum likelihood model requires assuming equality of the conditional mean and variance (the equidispersion property), but the Poisson pseudo-maximum likelihood model relaxes the equidispersion assumption, and will be consistent and offer valid statistical inference as long as the conditional mean is correctly specified. Results from a conditional fixed-effects pseudo-maximum likelihood Poisson model gave very similar results.

binary scientific knowledge and product development outcome variables, I show results from ordinary-least-squares (OLS) models, and in robustness checks also report marginal effects from probit models. For all models, I report heteroskedasticity robust standard errors.

The clear question arising with this specification is whether Celera IP was as good as randomly assigned across genes, conditional on the included covariates. As discussed in Section 4.2, when I limit my sample to genes sequenced in the years when the public effort was operating at scale (namely, 2000 forward), Celera and non-Celera genes appear balanced on *ex ante* gene-level observables. However, this sample limitation will not fully address selection concerns, since *conditional* on year of disclosure selection issues are still relevant. Given this, I address selection concerns in several additional ways. First, I show results from several propensity score specifications. Second, I condition on a broader set of publication measures, through 1999 (as opposed to the main specification, which as noted above controls for publications in 1970-1977). Third, I limit the sample to genes with non-missing data on cytogenetic and molecular location, replicate the results on this sample, and test whether the results are sensitive to conditioning on these detailed location covariates. Fourth, I use a different, complementary panel research design (described in Section 4.4) as an additional method of addressing these selection concerns. Finally, I present some descriptive results limiting the sample to Celera genes, relying only on variation in how long Celera genes were held with IP (that is, one or two years).

4.4 Panel specification

In the panel specification, for gene-year gy, I estimate the following:

$$(outcome)_{gy} = \delta_g + \gamma_y + \beta(celera)_{gy} + \epsilon_{gy}$$

The "*celera*" variable is now an indicator for whether all mRNAs on gene g were sequenced only by Celera as of that year.[52] This "*celera*" variable now varies within genes over time, and a transition from 1 to 0 in this variable represents the removal of Celera's IP from a given gene. Year fixed effects control for year-specific shocks that are common across genes, such as (for example) annual changes in the level of research funding available from public sector agencies. Gene fixed effects control for time-invariant differences across genes, such as a gene's inherent commercial potential. In the language of age-time-cohort effects, I control for time effects with the year fixed effects, and cohort effects with the gene fixed effects (which take out variation due to the year in which a gene was sequenced). For all outcome variables, I show results from OLS models and report heteroskedasticity robust standard errors clustered at the gene level.[53]

As discussed in Section 2.3, Celera's human genome sequencing efforts commenced in September 1999, and its draft human genome was disclosed in 2001. Unfortunately, I do not observe

[52] In the data, 62 percent of Celera genes were resequenced by the public sector in 2002, and the remaining 38 percent in 2003. My understanding is that the date when a Celera gene would be resequenced by the public sector was not predictable in advance, within the general timeframe of expecting all Celera genes would be in the public domain by the stated deadline of 2003.

[53] For the publications outcome, results from a conditional fixed-effects pseudo-maximum likelihood Poisson model gave very similar results.

the timing of when specific genes were sequenced within this time frame. In the absence of such data, I limit my panel specification to include the years 2001-2009 since prior to 2001 I do not know whether or not Celera genes had yet been sequenced. This sample limitation focuses on the "experiment" in which Celera genes have been sequenced, but vary in IP status over time.

By including gene fixed effects, this panel approach allows me to control for time-invariant differences across genes, such as a gene's inherent commercial potential. However, this approach has several limitations. First, this approach is only feasible for the two outcome variables I observe in a panel (that is, not for the "known, certain phenotype link" and diagnostic test availability outcome variables). Second, any observed differences in this specification could in theory be driven by short-term shifts in the timing of when research takes place that may or may not have persistent effects on welfare. In practice, I do not observe clear "bunching" of publications that would be predicted by stories in which researchers strategically wait until IP is removed to publish scientific papers. In addition, the cross-section specification addresses this concern through testing for longer-run, persistent impacts on innovation outcomes. Finally, to the extent that there are increasing returns to R&D, and non-Celera genes have higher levels of publications than Celera genes during the time period when Celera's IP is active, the implicit parallel trends assumption underlying this specification is less plausible.[54] For all of these reasons, I focus attention on the cross-section specification and rely on the panel specification primarily as a robustness check.

A natural question is whether the panel specification can provide an informal check on the validity of the cross-section specification. To the extent that the gene-level covariates are adequately proxying for differences in "potential innovation" across genes, we would like the panel estimates to be similar if I replace the gene fixed effects with the (time-invariant) gene-level covariates. Although not a formal test of the identification assumption underlying the cross-section specification, this informal test can offer suggestive evidence on how effective the cross-section gene-level covariates are in controlling for gene-specific variation in innovation.

Finally, I also present results from a "timing" panel specification that provides an event study-type graph. Specifically, I estimate the following:

$$(outcome)_{gy} = \delta_g + \gamma_y + \Sigma_z \beta_z (celera)_g * 1(z) + \epsilon_{gy}$$

Here, I define the years z relative to a "zero" relative year that marks the last year the gene was held with Celera IP.

[54] This type of increasing returns to R&D should positively bias the panel estimate of the effect of Celera IP. If non-Celera genes have higher levels of publications than Celera genes during the time when Celera's IP is active, increasing returns would imply non-Celera genes would have larger increases in publications in subsequent years, relative to Celera genes, which would bias the estimate towards finding that the "*celera*" indicator variable has a positive effect. As discussed in Section 5.2, I instead find a negative effect, which nonetheless may be biased towards zero.

5 Empirical results

5.1 Cross-section results

Table 3 presents the main results from my cross-section specification, for the sample of genes sequenced in and after 2000. Column (1) includes indicator variables for the year of disclosure, and Column (2) adds eight count variables for the number of publications in each year from 1970 to 1977.

Panel A of Table 3 reports estimates from quasi-maximum likelihood Poisson models for the publications outcome. Focusing on the estimate in Column (2) suggests Celera genes had 35 percent fewer publications from 2001 to 2009, relative to non-Celera genes.[55] Despite not observing mean differences across Celera and non-Celera genes in 1970-1977 publications in this sample in Table 2, adding these variables as covariates does affect my point estimates - highlighting that conditional on year of disclosure, selection issues are still relevant in this cross-section specification.[56]

Panels B, C, and D in Table 3 report analogous results from ordinary-least-squares (OLS) models for the three additional dependent variables. The estimates in Panel B of Table 3 suggest a 16 percentage point reduction in the probability of a gene having a known, uncertain phenotype link, relative to a mean of 30 percent. The estimates in Panel C of Table 3 suggest a 2 percentage point reduction in the probability of a gene having a known, certain phenotype link, relative to a mean of 4 percent. Turning to product development, the estimates in Panel D of Table 3 suggest a 1.5 percentage point reduction in the probability of a gene being used in any currently available diagnostic test, relative to a mean of 3 percent. As in Panel A of Table 3, the addition of controls for pre-existing scientific knowledge does have some effect on the point estimates of interest, but this change is relatively small.[57]

Of course, a lingering concern is whether unobserved gene characteristics could bias these cross-section estimates, a concern I address in a series of robustness checks. First, Appendix Table A5 presents results from several propensity score specifications, which condition on observables in alternative ways. Appendix Table A4 reports marginal effects from a probit model which predicts the Celera IP indicator as a function of the count variables for the number of publications in each year from 1970 to 1999. Appendix Figure A4 plots the distributions of this predicted probability of Celera IP treatment for Celera and non-Celera genes, and shows a clear overlap in these two distributions. Appendix Table A5 then uses this predicted probability of

[55] A Poisson estimate of β_i on a binary independent variable can be interpreted as an $(e^{\beta_i} - 1) \cdot 100$ percent change in the dependent variable, given a change from 0 to 1 in the independent variable (Cameron and Trivedi, 1998).

[56] Appendix Table A2 reports estimated coefficients on the covariates included in Column (2) of Table 3. As expected, genes with earlier dates of sequence disclosure tend to be associated with higher levels of my innovation outcome variables, as do genes with higher levels of 1970-1977 publications.

[57] Appendix Table A3 reports marginal effects from probit models for these three binary outcome variables. The point estimates are generally similar, but slightly smaller, suggesting a 10 percentage point reduction in the probability of a gene having a known, uncertain phenotype link; a 1 percentage point reduction in a gene having a known, certain phenotype link; and a 1 percentage point reduction in the probability of a gene being used in any currently available diagnostic test.

Celera IP treatment in two propensity score specifications: Columns (1) and (2) use the propensity score to construct inverse probability weights, and Columns (3) and (4) break the data into blocks based on the propensity score, and includes fixed effects for each block as covariates (following Dehejia and Wahba (1999)). In general the point estimates are quite similar, both across alternative propensity score specifications and relative to the main estimates presented in Table 3.

Second, Appendix Table A6 presents results analogous to those in Table 3, conditioning on additional later years of publication variables, through 1999. This robustness check addresses the possibility that the benchmark set of 1970-1977 publication variables may contain less information than the full set of publication variables through 1999. Empirically, results conditioning on these later years of publications are very similar to the results in Column (2) of Table 3.[58]

Third, I limit the sample to genes with non-missing data on the detailed cytogenetic and molecular location variables (N = 13,871), replicate the main results from Table 3 on this sub-sample, and examine robustness to conditioning on these additional locational covariates. This robustness check addresses the possibility that scientists may have targeted their sequencing efforts based on a gene's (*ex ante* known) approximate location on the genome. Columns (1) and (2) in Appendix Table A7 suggest that replicating the main results on this sub-sample of data gives point estimates similar to those in Table 3. Column (3) adds the detailed cytogenetic and molecular location covariates, which do not substantively alter the estimated magnitudes of the results.

For completeness, Appendix Table A8 presents results analogous to those in Table 3 for the full sample of genes and for the sub-sample of genes sequenced in 2001. In the full sample of data (Columns (1) and (2) of Appendix Table A8), I find similar point estimates to those in Table 3, consistent with the covariates addressing selection relatively well even in the full sample. The estimates limiting the sample to genes sequenced in 2001 (Column (3)) are also quite similar to the estimates in Table 3.

In summary, consistent with the main results in Table 3, these robustness checks offer additional evidence that Celera's IP has had negative impacts of economically meaningful size on both scientific research and product development outcomes, but there are of course still lingering concerns that these effects could be driven by non-random selection of genes into Celera's IP.[59] In Section 5.2, I present results from a complementary panel analysis as one way of addressing these selection concerns. In Section 5.3, I present results that limit the sample to Celera genes, and rely only on variation in the amount of time a given gene was held with Celera's IP. Another check on the potential impact of unobserved gene characteristics on the cross-section estimates is

[58] I can add these additional, later years of publication variables in this specification because I am limiting the sample to genes sequenced in or after 2000. My main results focus on the 1970-1977 publication variables for comparability with my estimates from the full sample of genes, for which I do not include post-1977 covariates as controls (for reasons discussed in Section 4.1).

[59] These average differences in innovation outcomes are not inconsistent with a model in which Celera genes were developed into products conditional on having high expected commercial value, whereas non-Celera genes were developed into products regardless of commercial value. In the absence of data on the commercial or social value of the gene-based diagnostic tests, I am unable to test for such effects.

to apply the methodology developed by Altonji, Elder and Taber (2005) and Murphy and Topel (1990) to bound the amount of selection on unobservables relative to selection on observables that would be required to completely explain the estimated effect of Celera IP. Applying this method to the diagnostic test outcome in Column (2) of Table 3, for example, I estimate a ratio of 1.8 using this method. Altonji, Elder and Taber (2005) argue that the ratio of selection on unobservables relative to selection on observables is likely to be less than one, suggesting part of the observed negative effect of Celera IP is likely real based on this approach.

5.2 Panel results

Table 4 presents the main results from the panel specification, for the sample of genes sequenced in or after 2000. Columns (1) and (2) of Table 4 are analogous to the cross-section specifications from Table 3: both control for year fixed effects, Column (1) includes indicator variables for the year of disclosure, and Column (2) adds eight count variables for the number of publications in each year from 1970 to 1977. Column (3) retains the year fixed effects but replaces the time-invariant covariates with gene fixed effects.

Panel A of Table 4 reports estimates from OLS models for the gene-year level publications outcome. As in the cross-section specification, the set of 1970-1977 publication variables do affect the estimate of the effect of Celera IP. In addition, replacing the time-invariant covariates with gene fixed effects does further reduce the magnitude of the estimate of the effect of Celera IP. That said, the magnitudes of the coefficients in Columns (2) and (3) are broadly similar, which I interpret as suggestive evidence that the cross-section controls are at least somewhat effective in controlling for gene-specific variation in the publications outcome. In terms of magnitudes, the coefficient in Column (3) in Panel A of Table 4 suggests Celera's IP was associated with 0.05 fewer publications per year, relative to a mean of 0.12 publications per gene-year.

Panel B of Table 4 reports analogous estimates for the gene-year level indicator variable for a gene having any known but uncertain phenotype link. The coefficient in Column (3) suggests Celera IP was associated with a 7 percentage point reduction in the probability that a gene had a known, uncertain phenotype link, relative to a mean of 22 percent.

Figure 1 presents graphical versions of the "timing" panel specification. On the x axes are years z relative to a "zero" relative year that marks the last year the gene was held with Celera IP (that is, year 1 marks the first year the gene was in the public domain). The dotted lines show 95 percent confidence intervals.

Panel A of Figure 1 presents results for the gene-year level publications outcome. These estimates suggest that in the first year a gene enters the public domain ($t = 1$, on the graph), there is a discrete level shift in the flow of publications related to that gene, which remains relatively constant through the end of my data. Although visually the levels of the estimated coefficients are somewhat higher in the first few years after Celera's IP was removed relative to later years, the increase in publications is persistent through the end of my sample, suggesting the positive coefficient observed in the panel specification is not simply driven by a short-term increase in publications.

Panel B of Figure 1 presents results for the gene-year level indicator for a gene having any known but uncertain phenotype link. This outcome increases in the first year a gene enters the public domain ($t = 1$, on the graph), and continues to increase through the end of my data.

For completeness, Appendix Table A10 presents results analogous to those in Table 4 for the full sample of genes. In this full sample, I find point estimates generally similar in magnitude to those in Table 4.

Given that Celera genes were held with Celera's IP for a maximum of two years, and that we observe relative increases in each of the two gene-year panel outcome variables after Celera genes moved into the public domain, a natural question is why this short-term form of IP might have had the persistent negative effects we observed in the cross-section results (Table 3). Perhaps the most natural story is that the relative costs of doing research on Celera genes must have been higher even after their IP was removed, which could be true for several reasons. First, this may be interpreted as suggestive evidence of increasing returns to R&D. That is, to the extent that existing stocks of scientific knowledge provide ideas and tools that allow future discoveries to be achievable at lower costs, the production of new knowledge may rise more than proportionately with the stock. The results of the panel specification suggest Celera genes accumulated lower levels of scientific knowledge during the time they were held with IP, and it could be that these temporarily lower levels of publications led the accumulation of new scientific knowledge to be relatively more costly on Celera genes even after Celera's IP was removed. Second, while increasing returns to R&D is a natural story given its prominence in the economics literature (Aghion and Howitt, 1992; Romer, 1990), other factors could also have increased the costs of doing research on Celera genes even after their IP was removed. For example, scientists could in theory have been more likely to invest in research on new genes that were in the public domain during the peak of the sequencing efforts in 2000-2001, relative to later years.

5.3 Focusing on Celera genes

Figure 2 presents results from an additional descriptive analysis. I here limit the sample to include *only* Celera genes, and rely solely on variation in how long these genes were held with Celera's IP - that is, whether the Celera gene was re-sequenced by the public effort in 2002 (N = 1,047, which I refer to as "*public in 2002*") or in 2003 (N = 635, which I refer to as "*public in 2003*"). The summary statistics in Appendix Table A1 suggest that the year in which Celera genes were re-sequenced by the public effort cannot be predicted with gene-level observables.

Figure 2 presents means by year for the two panel outcome variables for each of the "*public in 2002*" and "*public in 2003*" groups. As expected from the fact that Celera genes re-sequenced in 2002 and 2003 look balanced on *ex ante* gene-level observables, the mean levels of both outcome variables are quite similar across the two groups in 2001, when both sets of genes were held with Celera IP. Panel A shows that, comfortingly, Celera genes re-sequenced in 2002 saw a relative uptick in publications in that year, while Celera genes re-sequenced in 2003 show a similar uptick in 2003.[60] Panel B similarly shows that Celera genes re-sequenced in 2002 saw a relative increase

[60]The difference in means in 2002 is statistically significant at the 10 percent level; mean differences in other

in the probability of having a known, uncertain genotype-phenotype link in 2002.

Perhaps the most striking feature of Panel B is that the difference between the "*public in 2002*" and "*public in 2003*" samples appears to grow over time. Rather than the "*public in 2003*" group catching up with their "*public in 2002*" counterparts one year later, the "*public in 2003*" group has persistently lower levels of this outcome variable over time, with differences that become larger and more strongly statistically significant in later years - which, again, may be interpreted as suggestive evidence of increasing returns to R&D.[61]

5.4 Potential substitution of R&D from Celera to non-Celera genes

These results provide evidence that Celera genes have lower scientific research and product development outcomes relative to non-Celera genes. In theory, this could reflect a decrease in total innovation on all genes, or could at least in part reflect the substitution of innovative effort away from Celera genes and towards non-Celera genes. That is, the observed relative decrease in scientific research and product development outcomes for Celera genes could be consistent with a *zero* net change in total innovation on all genes, if the relative decrease were completely explained by the substitution of effort away from Celera genes towards non-Celera genes. I focus attention on three aspects of this type of substitution. First, is it likely - *a priori* - that such substitution was important? Second, within the context of my research designs how would such substitution bias the estimation of whether Celera's IP reduced research on Celera genes? And third, in the extreme case in which the entire relative difference were explained by such substitution, would we still care about the measured effects from a welfare perspective?

First, is it likely that such substitution was important? *A priori*, this depends on whether the number of researchers conducting gene-related research should be considered relatively fixed or relatively flexible. In the case of academics, a relatively fixed supply of researchers in the short run seems likely. However, private firms may have otherwise been working in alternative product markets, implying a relatively flexible supply of private researchers.

Second, in a very narrow sense, to the extent that I wish to measure the reduction in research on Celera genes arising due to Celera's IP, this type of substitution would lead me to over-estimate that reduction in my current research design. As an example, consider the gene-level publications outcome variable in the cross-section specification. Assume that if no genes had IP, each gene would have n publications, and that Celera IP reduces the number of publications on Celera genes to $n-x$. If there is no substitution, then the cross-section difference in publications between Celera and non-Celera genes equals $-x$. If each publication that is deterred on a Celera gene accrues to a non-Celera gene, then the cross-section difference in publications between Celera and non-Celera genes equals $-2x$. This suggests that in this simple model in which the entire relative difference in outcomes between Celera and non-Celera genes is driven by substitution, substitution could be inflating the estimated coefficients by a factor of 2. More

years are not statistically significant.

[61] The difference in means is statistically significant in 2003 (at the 10 percent level), 2006 (at the 10 percent level), 2007 (at the 5 percent level), and 2008 (at the 5 percent level).

generally, in any given model with assumptions about the elasticity of the supply of researchers with respect to the number of projects, one can bound the extent to which substitution would be inflating the magnitudes of my estimates; in general, this type of substitution would not alter the sign of the estimated coefficients, but could affect the magnitudes of the estimates.

Third, a stronger question is this: if substitution were explaining the entire relative difference in innovation on Celera and non-Celera genes, would we still care about the measured effects from a welfare perspective? In the genome context, substitution of research effort across genes isn't obviously costly exactly because all genes were plausibly similar *ex ante*. However, in other markets the technologies held with IP tend to be the most commercially valuable technologies; in those markets, we would care about research effort being substituted away from more socially valuable technologies towards less socially valuable technologies - suggesting we do care about measuring a substitution effect. The key issue I cannot address formally in my data is whether substitution of research effort across genes is "similar to" substitution of research effort across other technologies. At first glance, the *ex ante* similarity of genes might suggest we expect substitution across genes to be very different from substitution across other technologies. However, the fact that many institutions were willing to pay large sums of money to access the Celera data when the public data was freely available provides evidence that the two sets of genes must not have been viewed as perfect substitutes; I argue that one reason for this imperfect substitutability may have been that institutions expected downstream markets to be less competitive on Celera genes relative to non-Celera genes (as discussed in Section 1). Thinking about substitution in other markets, surveys by Walsh, Cho and Cohen (2005) and Walsh, Cohen and Cho (2007) suggest that substitution is relevant; they present evidence that restricted access to tangible research inputs (including information, data, and software) appear to shift scientists' research project choices. Taken together, this suggests some degree of substitution is relevant both in the genome context and in other markets, but in this paper I am unable to formally assess the degree of this similarity.

6 Conclusions

Intellectual property (IP) is a widely-used policy lever for promoting innovation, yet relatively little is known about how IP on a given technology affects subsequent innovation. The sequencing of the human genome provides a particularly useful empirical context in which to shed light on this question, as the simultaneous sequencing efforts of the public Human Genome Project and the private firm Celera generated variation in IP across a relatively large group of *ex ante* similar technologies (namely, genes). Across a variety of empirical analyses, I find robust evidence that the package of short-term IP used by Celera has been associated with reductions on the order of 30 percent in subsequent gene-level scientific research and product development outcomes.

A natural question is how these observed negative impacts of IP on innovation translate into impacts on social welfare. One contribution of this paper is to trace out the impacts of IP on not only scientific research (the focus of prior studies) but also on product development. Although

changes in the space of products available to consumers clearly has some link to social welfare, in health care markets the social value of new medical technologies is difficult to measure due to the potential inefficiencies introduced by asymmetric information and other factors. Some gene-related diagnostic tests are likely very high-value, such as a genetic test currently under development that could improve doctors' ability to provide patients with appropriate doses of warfarin, a widely-used blood thinner. On the other hand, many have raised concerns that broad genetic testing for common, chronic diseases may be counterproductive in the sense of leading patients to receive low-value treatments (*e.g.* Welch (2004)). The introduction of new genetic tests may also have broader impacts on insurance markets, as recently analyzed by Oster et al. (2009), introducing additional complications in estimating the social value of gene-based diagnostic technologies.

Celera's short-term IP, which lasted a maximum of two years, appears to have had persistent negative effects on subsequent scientific research and product development relative to a counterfactual of Celera genes having always been in the public domain. These results shed light on one important part of the evidence needed to evaluate broader questions about the design of IP systems. Of course, the overall welfare effects of IP depend on factors beyond the impact of IP on subsequent innovation, including the provision of dynamic incentives for innovation.[62] From a policy perspective, these results suggest that, holding Celera's entry and sequencing efforts constant, an alternative institutional mechanism - such as the patent buyout mechanism discussed by Kremer (1998) - may have had social benefits relative to the package of IP used by Celera.

[62]For recent discussions of the overall costs and benefits of IP systems, see Bessen and Meurer (2008), Boldrin and Levine (2008), and Jaffe and Lerner (2006).

References

Aghion, Philippe and Peter Howitt, "A model of growth through creative destruction," *Econometrica*, 1992, *60* (2), 323–351.

_ , Mathias Dewatripont, and Jeremy Stein, "Academic freedom, private-sector focus, and the process of innovation," *RAND Journal of Economics*, 2008, *39* (3), 617–635.

Altonji, Joseph, Todd Elder, and Christopher Taber, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 2005, *113* (1), 151–184.

Arora, Ashish, Andrea Fosfuri, and Alfonso Gambardella, *Markets for Technology: The Economics of Innovation and Corporate Strategy*, MIT Press, 2001.

Arrow, Kenneth, "Economic welfare and the allocation of resources for invention," in Richard Nelson, ed., *The Rate and Direction of Inventive Activity*, Princeton University Press, 1962.

Bessen, James, "Holdup and licensing of cumulative innovations with private information," *Economics Letters*, 2004, *82* (3), 321–326.

_ and Michael Meurer, *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*, Princeton University Press, 2008.

Boldrin, Michele and David K. Levine, *Against Intellectual Monopoly*, Cambridge University Press, 2008.

Cameron, Colin and Pravin Trivedi, *Regression Analysis of Count Data*, Cambridge University Press, 1998.

Cho, Mildred, Samantha Illangasekare, Meredith Weaver, Debra Leonard, and Jon Merz, "Effects of patents and licenses on the provision of clinical genetic testing services," *Journal of Molecular Diagnostics*, 2003, *5* (1), 3–8.

Coase, Ronald, "The problem of social cost," *Journal of Law and Economics*, 1960, *3* (1), 1–44.

Collins, Francis, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, the members of the DOE, and NIH planning groups, "New goals for the US Human Genome Project: 1998-2003," *Science*, 1998, *282* (5389), 682–689.

Cook-Deegan, Robert, *The Gene Wars: Science, Politics, and the Human Genome*, W. W. Norton & Company, 1994.

Cournot, Augustin, *Researches into the Mathematical Principles of the Theory of Wealth*, The MacMillan Company, 1838.

Davies, Kevin, *Cracking the Genome: Inside the Race to Unlock Human DNA*, Johns Hopkins University Press, 2001.

Dehejia, Rajeev and Sadek Wahba, "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 1999, *94* (448), 1053–1062.

Duenes, Steve, "Journey to the genome," *New York Times*, 2000, *27 June.*

Eisenberg, Rebecca, "Genomics in the public domain: Strategy and policy," *Nature Reviews Genetics*, 2000, *1* (1), 70–74.

Furman, Jeffrey and Scott Stern, "Climbing atop the shoulders of giants: The impact of institutions on cumulative research," 2010. unpublished Boston University mimeo.

Gans, Joshua and Scott Stern, "Incumbancy and R&D incentives: Licensing the gale of creative destruction," *Journal of Economics & Management Strategy*, 2000, *9* (4), 485–511.

Green, Jerry and Suzanne Scotchmer, "On the division of profit in sequential innovation," *RAND Journal of Economics*, 1995, *26* (1), 20–33.

Green, Philip, "Against a whole-genome shotgun," *Genome Research*, 1997, *7*, 410–417.

Heller, Michael and Rebecca Eisenberg, "Can patents deter innovation? The anticommons in biomedical research," *Science*, 1998, *280* (5364), 698–701.

Hellmann, Thomas, "The role of patents for bridging the science to market gap," *Journal of Economic Behavior and Organization*, 2007, *63* (4), 624–647.

Holman, Christopher, "The impact of human gene patents on innovation and access: A survey of human gene patent litigation," *University of Missouri-Kansas City Law Review*, 2007, *76*, 295–361.

_ , "Trends in human gene patent litigation," *Science*, 2008, *322* (5899), 198–199.

Istrail, Sorin et al., "Whole-genome shotgun assembly and comparison of human genome assemblies," *Proceedings of the National Academy of Sciences*, 2004, *101* (7), 1916–1921.

Jaffe, Adam and Josh Lerner, *Innovation and Its Discontents: How Our Broken Patent System is Endangering Innovation and Progress, and What to Do About It*, Princeton University Press, 2006.

Jensen, Kyle and Fiona Murray, "Intellectual property landscape of the human genome," *Science*, 2005, *310* (5746), 239–240.

Kitch, Edmund, "The nature and function of the patent system," *Journal of Law and Economics*, 1977, *20* (2), 265–290.

Kremer, Michael, "Patent buyouts: A mechanism for encouraging innovation," *Quarterly Journal of Economics*, 1998, *113* (4), 1137–1167.

_ **and Heidi Williams**, "Incentivizing innovation: Adding to the toolkit," in Josh Lerner and Scott Stern, eds., *Innovation Policy and the Economy Volume 10*, University of Chicago Press, 2010, pp. 1–17.

Lander, Eric, "The new genomics: Global views of biology," *Science*, 1996, *274* (5287), 536–539.

_ **et al.**, "Initial sequencing and analysis of the human genome," *Nature*, 2001, *409* (6822), 860–921.

Landes, William and Richard Posner, "Indefinitely renewable copyright," *University of Chicago Law Review*, 2003, *70* (2), 471–518.

Lemley, Mark, "Ex ante versus ex post justifications for intellectual property," *University of Chicago Law Review*, 2004, *71* (1), 129–149.

Maglott, Donna, Jim Ostell, Kim Pruitt, and Tatiana Tatusova, "Entrez Gene: Gene-centered information at NCBI," *Nucleic Acids Research*, 2005, *33* (Database issue), D54–D58.

Marshall, Eliot, "NIH to produce a 'working draft' of the genome by 2001," *Science*, 1998, *281* (5384), 1774–1775.

_ , "Bermuda Rules: Community spirit, with teeth," *Science*, 2001, *291* (5507), 1192.

_ , "Celera and *Science* spell out data access provisions," *Science*, 2001, *291* (5507), 1191.

Maxam, Allan and Walter Gilbert, "A new method for sequencing DNA," *Proceedings of the National Academy of Sciences*, 1977, *74* (2), 560–564.

McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD), "Online Mendelian Inheritance in Man, OMIM (TM)," 2009. `http://www.ncbi.nlm.nih.gov/omim/`.

Merges, Robert and Richard Nelson, "On the complex economics of patent scope," *Columbia Law Review*, 1990, *90* (4), 839–916.

Moon, Seongwuk, "How does the management of research impact the disclosure of knowledge? Evidence from scientific publications and patenting behavior," 2008. unpublished KDI School of Public Policy and Management mimeo.

Mowery, David, Richard Nelson, Bhaven Sampat, and Arvids Ziedonis, *Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After the Bayh-Dole Act in the United States*, Stanford University Press, 2004.

Murphy, Kevin and Robert Topel, "Efficiency wages reconsidered: Theory and evidence," in Yoram Weiss and Robert Topel, eds., *Advances in the Theory and Measurement of Unemployment*, St. Martin's Press, 1990.

Murray, Fiona and Scott Stern, "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis," *Journal of Economic Behavior and Organization*, 2007, *356* (23), 2341–2343.

_ , Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern, "Of mice and academics: Examining the effect of openness on innovation," 2008. unpublished MIT mimeo.

Myerson, Roger and Mark Satterthwaite, "Efficient mechanisms for bilateral trading," *Journal of Economic Theory*, 1983, *29* (2), 265–281.

National Academy of Sciences, *Reaping the Benefits of Genomic and Proteomic Research: Intellectual Property Rights, Innovation, and Public Health*, National Academies Press, 2006.

Nelson, Richard, "The simple economics of basic scientific research," *Journal of Political Economy*, 1959, *67* (3), 297–306.

Ng, Pauline, Sarah Murray, Samuel Levy, and J. Craig Venter, "An agenda for personalized medicine," *Nature*, 2009, *461* (7265), 724–726.

Oster, Emily, Ira Shoulson, Kimberly Quaid, and E. Ray Dorsey, "Genetic adverse selection: Evidence from long-term care insurance and Huntington disease," 2009. NBER working paper #15326.

Parra, Genís et al., "Tandem chimerism as a means to increase protein complexity in the human genome," *Genome Research*, 2006, *16* (1), 37–44.

Pennisi, Elizabeth, "Human genome: Academic sequencers challenge Celera in a sprint to the finish," *Science*, 1999, *283* (5409), 1822–1823.

Pitcher, Edmund and Brian Fairchild, "Legal affairs: Enforceable diagnostic method patents," *Genetic Engineering & Biotechnology News*, 2009, *29* (7).

Pruitt, Kim, Tatiana Tatusova, and Donna Maglott, "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts, and proteins," *Nucleic Acids Research*, 2007, *35* (Database issue), D61–D65.

Roberts, Leslie, "Controversial from the start," *Science*, 2001, *291* (5507), 1182–1188.

Romer, Paul, "Endogenous technological change," *Journal of Political Economy*, 1990, *98* (5 (Part 2)), S71–S102.

Sanger, Frederick, Steven Nicklen, and Alan Coulson, "DNA sequencing with chain-terminating inhibitors," *Proceedings of the National Academy of Sciences*, 1977, *74* (12), 5463–5467.

Scherer, Stewart, *A Short Guide to the Human Genome*, Cold Spring Harbor Laboratory Press, 2008.

Schwartz, John, "Cancer patients challenge the patenting of a gene," *New York Times*, 2009, *12 May*.

Science Online, "Accessing the Celera human genome sequence data," 2001. `http://www.sciencemag.org/feature/data/announcement/gsp.dtl`.

Service, Robert, "Can data banks tally profits?," *Science*, 2001, *291* (5507), 1203.

Shapiro, Carl, "Navigating the patent thicket: Cross licenses, patent pools, and standard setting," in Adam Jaffe, Josh Lerner, and Scott Stern, eds., *Innovation Policy and the Economy Volume 1*, MIT Press, 2000.

Shreeve, James, *The Genome War: How Craig Venter Tried to Capture the Code of Life and Save the World*, Ballantine Books, 2005.

Snyder, Michael and Mark Gerstein, "Defining genes in the genomics era," *Science*, 2003, *300* (5617), 258–260.

Sulston, John and Georgina Ferry, *The Common Thread: Science, Politics, Ethics, and the Human Genome*, Corgi Books, 2002.

The Huntington's Disease Collaborative Research Group, "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes," *Cell*, 1993, *72* (6), 971–983.

Uhlmann, Wendy and Alan Guttmacher, "Key internet genetics resources for the clinician," *Journal of the American Medical Association*, 2008, *299* (11), 1356–1358.

University of Washington, Seattle, "GeneTests: Medical Genetics Information Resource (database online), Copyright," 2009. `http://www.genetests.org`.

US National Human Genome Research Institute (NHGRI), US National Institutes of Health (NIH), "NHGRI policy regarding intellectual property of human genomic sequence: Policy on availability and patenting of human genomic DNA sequence produced by NHGRI pilot projects (funded under RFA HG-95-005)," 1996. `http://www.genome.gov/10000926`.

Venter, J. Craig, *A Life Decoded: My Genome, My Life*, Viking Adult, 2007.

_ et al., "The sequence of the human genome," *Science*, 2001, *291* (5507), 1304–1351.

_ , Mark Adams, Granger Sutton, Anthony Kerlavage, Hamilton Smith, and Michael Hunkapiller, "Shotgun sequencing of the human genome," *Science*, 1998, *280* (5369), 1540–1542.

Wade, Nicholas, *Life Script: How the Human Genome Discoveries Will Transform Medicine and Enhance Your Health*, Simon & Schuster, 2001.

_ , "Genes show limited value in predicting diseases," *New York Times*, 2009, *15 April*.

Walsh, John, Ashish Arora, and Wesley Cohen, "Research tool patenting and licensing and biomedical innovation," in Wesley Cohen and Stephen Merrill, eds., *Patents in the Knowledge-Based Economy*, National Academy Press, 2003.

_ , _ , and _ , "Working through the patent problem," *Science*, 2003, *299* (5609), 1021.

_ , Charlene Cho, and Wesley Cohen, "View from the bench: Patents and material transfers," *Science*, 2005, *309* (5743), 2002–2003.

_ , **Wesley Cohen, and Charlene Cho**, "Where excludability matters: Material versus intellectual property in academic biomedical research," *Research Policy*, 2007, *36* (8), 1184–1203.

Weber, James and Eugene Myers, "Human whole-genome shotgun sequencing," *Genome Research*, 1997, *7*, 401–409.

Welch, H. Gilbert, *Should I Be Tested for Cancer?*, University of California Press, 2004.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, 2002.

Table 1: Summary Statistics for Gene-Level Data

	mean	standard deviation	minimum	maximum
Panel A: Celera intellectual property (IP)				
0/1, Celera gene	0.060	0.238	0	1
Panel B: Outcome variables				
publications in 2001-2009	2.197	9.133	0	231
0/1, known, uncertain phenotype	0.453	0.498	0	1
0/1, known, certain phenotype	0.081	0.273	0	1
0/1, used in any diagnostic test	0.060	0.238	0	1
Panel C: Main covariates				
year first mRNA disclosed	2002.962	3.551	1999	2009
publications in 1970	0.032	0.323	0	18
publications in 1971	0.027	0.262	0	18
publications in 1972	0.036	0.349	0	26
publications in 1973	0.029	0.301	0	26
publications in 1974	0.037	0.362	0	25
publications in 1975	0.039	0.412	0	35
publications in 1976	0.045	0.395	0	28
publications in 1977	0.047	0.454	0	39
publications in 1978	0.056	0.464	0	30
publications in 1979	0.054	0.460	0	28
publications in 1980	0.066	0.547	0	33
publications in 1981	0.073	0.595	0	42
publications in 1982	0.074	0.577	0	34
publications in 1983	0.075	0.613	0	42
publications in 1984	0.076	0.619	0	33
publications in 1985	0.101	0.763	0	49
publications in 1986	0.099	0.745	0	38
publications in 1987	0.120	0.823	0	39
publications in 1988	0.133	0.899	0	44
publications in 1989	0.133	0.946	0	57
publications in 1990	0.139	0.936	0	57
publications in 1991	0.158	0.968	0	46
publications in 1992	0.189	1.177	0	57
publications in 1993	0.176	0.990	0	32
publications in 1994	0.190	0.962	0	31
publications in 1995	0.232	1.125	0	31
publications in 1996	0.244	1.119	0	34
publications in 1997	0.258	1.158	0	33
publications in 1998	0.283	1.157	0	35
publications in 1999	0.289	1.188	0	32
Panel D: Additional covariates				
0/1, missing cytogenetic location	0.370	0.483	0	1
0/1, missing molecular location	0.059	0.235	0	1
$N = 27{,}882$				

Notes: Gene-level observations. Note that the mean year of disclosure is affected by truncation since a disclosure year of 1999 represents a gene sequenced in or before 1999 (because 1999 is the earliest year any observations appear in the RefSeq database). See text and Appendix 2 for more detailed data and variable descriptions.

Table 2: Differences Across Celera and non-Celera Genes in Gene-Level Data

non-Celera genes sequenced in:	(1) -	(2) all	(3) all	(4) 2001	(5) 2001	(6) ≥2000	(7) ≥2000
	Celera mean	mean	p-value	mean	p-value	mean	p-value
Panel A: Outcome variables							
publications in 2001-2009	1.239	2.258	[0.000]	2.116	[0.000]	1.083	[0.250]
0/1, known, uncertain phenotype	0.401	0.456	[0.000]	0.563	[0.000]	0.301	[0.000]
0/1, known, certain phenotype	0.046	0.083	[0.000]	0.073	[0.000]	0.038	[0.126]
0/1, used in any diagnostic test	0.030	0.062	[0.000]	0.054	[0.000]	0.027	[0.430]
Panel B: Main covariates							
year first mRNA disclosed	2001.000	2003.088	[0.000]	2001	-	2004.318	[0.000]
publications in 1970	0.008	0.034	[0.002]	0.021	[0.022]	0.011	[0.536]
publications in 1971	0.005	0.028	[0.000]	0.019	[0.007]	0.009	[0.224]
publications in 1972	0.004	0.038	[0.000]	0.016	[0.020]	0.010	[0.103]
publications in 1973	0.009	0.030	[0.005]	0.017	[0.100]	0.009	[0.996]
publications in 1974	0.008	0.039	[0.000]	0.014	[0.182]	0.011	[0.441]
publications in 1975	0.007	0.041	[0.001]	0.013	[0.166]	0.011	[0.355]
publications in 1976	0.014	0.047	[0.001]	0.029	[0.025]	0.015	[0.799]
publications in 1977	0.010	0.049	[0.001]	0.023	[0.039]	0.015	[0.320]
publications in 1978	0.017	0.058	[0.000]	0.029	[0.071]	0.018	[0.818]
publications in 1979	0.024	0.056	[0.005]	0.026	[0.747]	0.016	[0.142]
publications in 1980	0.015	0.069	[0.000]	0.029	[0.054]	0.020	[0.494]
publications in 1981	0.018	0.077	[0.000]	0.034	[0.081]	0.020	[0.755]
publications in 1982	0.018	0.077	[0.000]	0.041	[0.027]	0.022	[0.634]
publications in 1983	0.020	0.079	[0.000]	0.042	[0.080]	0.021	[0.837]
publications in 1984	0.027	0.079	[0.001]	0.030	[0.784]	0.019	[0.185]
publications in 1985	0.028	0.106	[0.000]	0.042	[0.219]	0.028	[0.996]
publications in 1986	0.020	0.104	[0.000]	0.037	[0.063]	0.026	[0.499]
publications in 1987	0.030	0.126	[0.000]	0.049	[0.097]	0.029	[0.979]
publications in 1988	0.040	0.139	[0.000]	0.058	[0.199]	0.036	[0.671]
publications in 1989	0.039	0.139	[0.000]	0.048	[0.397]	0.034	[0.626]
publications in 1990	0.027	0.146	[0.000]	0.056	[0.006]	0.036	[0.359]
publications in 1991	0.034	0.165	[0.000]	0.063	[0.041]	0.041	[0.516]
publications in 1992	0.035	0.198	[0.000]	0.073	[0.002]	0.048	[0.239]
publications in 1993	0.042	0.185	[0.000]	0.063	[0.043]	0.044	[0.817]
publications in 1994	0.037	0.200	[0.000]	0.088	[0.000]	0.055	[0.119]
publications in 1995	0.046	0.243	[0.000]	0.100	[0.000]	0.061	[0.189]
publications in 1996	0.061	0.256	[0.000]	0.103	[0.008]	0.069	[0.536]
publications in 1997	0.061	0.271	[0.000]	0.105	[0.003]	0.074	[0.335]
publications in 1998	0.072	0.297	[0.000]	0.128	[0.000]	0.087	[0.263]
publications in 1999	0.086	0.302	[0.000]	0.157	[0.000]	0.116	[0.046]
Panel C: Additional covariates							
0/1, missing cytogenetic location	0.196	0.381	[0.000]	0.305	[0.000]	0.326	[0.000]
0/1, missing molecular location	0.021	0.061	[0.000]	0.021	[0.979]	0.076	[0.000]
N	1,682	26,200		2,851		20,142	

Notes: This table shows covariate means for Celera genes (Column 1) together with covariate means for non-Celera genes (Column 2), non-Celera genes sequenced in 2001 (Column 4), and non-Celera genes sequenced in or after 2000 (Column 6). Also shown are p-values for the differences in means between Celera genes and non-Celera genes (Column 3), non-Celera genes sequenced in 2001 (Column 5), and non-Celera genes sequenced in or after 2000 (Column 7). Gene-level observations. In an ordinary-least-squares model predicting *"celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001, as a function of the count variables for publications in each year from 1970-1999, the p-value from an F-test is 0.000 for the full sample of non-Celera genes; 0.033 for the sample of non-Celera genes sequenced in 2001; and 0.177 for the sample of non-Celera genes sequenced in or after 2000. Note that the mean year of disclosure for non-Celera genes in Column (2) is affected by truncation since a disclosure year of 1999 represents a gene sequenced in or before 1999 (because 1999 is the earliest year any observations appear in the RefSeq database). See text and Appendix 2 for more detailed data and variable descriptions.

Table 3: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000

	(1)	(2)
Panel A: publications in 2001-2009		
mean = 1.095		
celera	-0.535	-0.432
	(0.117)***	(0.112)***
Panel B: 0/1, known, uncertain phenotype		
mean = 0.309		
celera	-0.162	-0.158
	(0.015)***	(0.015)***
Panel C: 0/1, known, certain phenotype		
mean = 0.039		
celera	-0.027	-0.018
	(0.007)***	(0.006)***
Panel D: 0/1, used in any diagnostic test		
mean = 0.027		
celera	-0.023	-0.015
	(0.006)***	(0.005)***
indicator variables for year of disclosure	yes	yes
number of publications in each year 1970-77	no	yes
N	21,824	21,824

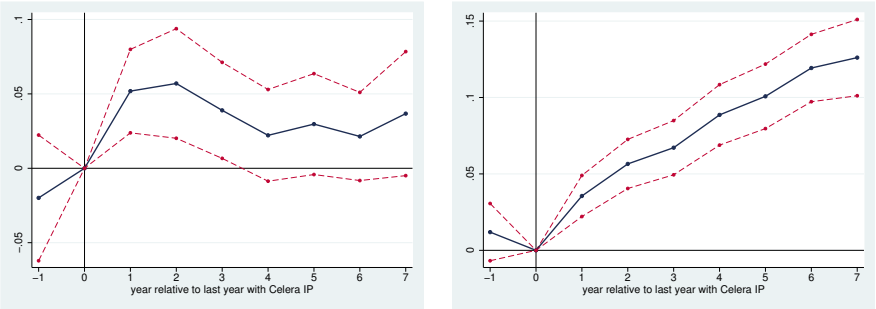
Notes: Gene-level observations. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. Sample includes all genes sequenced in or after 2000 (N = 21,824). Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Table 4: Panel Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000

	(1)	(2)	(3)
Panel A: gene-year publications			
mean = 0.122			
celera	-0.112	-0.084	-0.052
	(0.017)***	(0.014)***	(0.010)***
Panel B: 0/1, known, uncertain phenotype			
mean = 0.223			
celera	-0.151	-0.148	-0.068
	(0.009)***	(0.009)***	(0.008)***
year fixed effects	yes	yes	yes
indicator variables for year of disclosure	yes	yes	-
number of publications in each year 1970-77	no	yes	-
gene fixed effects	no	no	yes
N	196,416	196,416	196,416

Notes: Gene-year-level observations. All estimates are from ordinary-least-squares (OLS) models. As discussed in Section 2.3, Celera's human genome sequencing efforts commenced in September 1999, and its draft human genome was disclosed in 2001. Unfortunately, I do not observe the timing of when specific genes were sequenced within this time frame. In the absence of such data, I limit my panel specification to include the years 2001-2009 since prior to 2001 I do not know whether or not Celera genes had yet been sequenced. The sample includes all gene-years from 2001 to 2009 for genes sequenced in or after 2000 (21,824 genes, for 9 years, implies N = 196,416 total gene-year observations). Robust standard errors, clustered at the gene level, shown in parentheses. *: $p< 0.10$; **: $p< 0.05$; ***: $p< 0.01$. *"Celera"*: 0/1, =1 if all mRNAs on the gene were sequenced only by Celera in that year. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Figure 1: Panel Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000

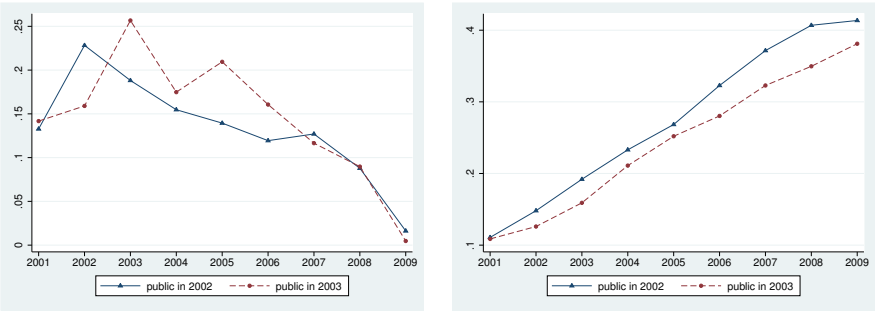


(a) Outcome variable: Gene-year publication count

(b) Outcome variable: Indicator for a gene having any known/uncertain phenotype link in that year

Notes: These figures plot coefficients (and 95 percent confidence intervals) from the panel timing specification, as described in Section 4.4. On the x axes are years z relative to a "zero" relative year that marks the last year the gene was held with Celera IP (that is, year 1 marks the first year the gene was in the public domain). As in the specifications in Table 4, this specification is based on gene-year level observations, the coefficients are estimates from ordinary-least-squares (OLS) models, the sample includes all gene-years from 2001 to 2009 for genes sequenced in or after 2000, and the standard errors are robust and clustered at the gene level. See text and Appendix 2 for more detailed data and variable descriptions.

Figure 2: Average Innovation Outcomes for Celera Genes by Year, by Year of Re-sequencing by the Public Effort



(a) Outcome variable: Gene-year publication count

(b) Outcome variable: Indicator for a gene having any known/uncertain phenotype link in that year

Notes: Sample includes all Celera genes (that is, genes for which all mRNAs on the gene were initially sequenced only by Celera as of 2001). These figures show means by year for the two gene-year outcome variables: gene-year publications, and a gene-year indicator for whether a gene has any known, uncertain phenotype link. Means are shown separately for Celera genes that were re-sequenced by the public effort in 2002 (N = 1,047) and for Celera genes that were re-sequenced by the public effort in 2003 (N = 635). Using the notation that *: $p<$ 0.10; **: $p<$ 0.05; and ***: $p<$ 0.01, the p-value of tests for differences in means are statistically significant in Panel (a) in 2002 (*), and in Panel (b) in 2003 (*), 2006 (*), 2007 (**), and 2008 (**). As in Appendix Figure A3, Panel (a) suggests flow publications peaked by this measure in 2003, although it is likely that some of the post-2003 decline is due to time lags in the addition of scientific publications to the OMIM database. See text and Appendix 2 for more detailed data and variable descriptions.

Appendix 1: Celera's intellectual property strategy

This appendix describes in additional detail Celera's chosen intellectual property (IP) strategy.

Celera's chosen form of IP included a variety of components, summarized in the data access agreement that accompanied Celera's *Science* publication (Science Online, 2001):[63]

- Academic users may access the sequence, do searches, download segments up to one megabase per week, publish their results, and seek intellectual property protection by agreeing that the data will be used for research purposes and will not be redistributed.
- Academic users whose research requires longer stretches of sequence, up to and including the whole genome, will be sent an electronic copy of the Celera data if they submit a statement, with a co-signature by an institutional representative, that the data will be used for research purposes and will not be redistributed.
- There are no reach-through provisions or restrictions on publication of the researcher's results.
- Redistribution of the Celera sequence data is prohibited. However, Celera will deposit sequence data into GenBank on behalf of authors if such deposition is required for publication of research results.
- Commercial users may access the data for validation and verification purposes only upon executing a Material Transfer Agreement. Alternatively, they may subscribe for a fee, or seek a license from Celera to use the data for other purposes.
- *Science* will keep a copy of the database in escrow, to insure that there will be no changes in the ability of the public to have full access to the data. Details are contained in the escrow agreement executed between *Science* and Celera.

As discussed by Marshall (2001b), the key features of Celera's IP strategy were restrictions on redistribution of Celera's data (aiming to prevent other commercial firms from directly copying the data for use in either products or product development), and a requirement that individuals wanting to use the data for commercial purposes negotiate a licensing agreement with Celera. Celera's data were disclosed with the 2001 publication of Celera's draft genome in *Science*, in the sense that any individual could view data on the assembled genome through the Celera website, or by obtaining a data DVD from the company.[64] Academic researchers were free to use the Celera data for non-commercial research purposes.

In terms of the formal legal basis for Celera's IP, in personal correspondence Robert Millman - then-Chief IP Counsel at Celera from 1999-2002 - clarified that Celera viewed the information as copyrighted material (the firm formally filed for copyright protection), and that the license included with the DVD was by nature a so-called shrink wrap license (which has legal basis in contract law).[65]

[63] The agreement included in Celera's data DVD gives some alternate formal language: "*...you are authorized to use the data solely for non-commercial research purposes and only if you qualify as an academic user as defined in the public access agreement. Except as specifically authorized in the public access agreement, any and all other uses of the data are strictly prohibited and all other rights in the data are reserved by Celera.*"

[64] Viewing the assembly or obtaining the data DVD required an agreement to neither commercialize nor distribute the data.

[65] I am very grateful to Robert Millman for several discussions on Celera's IP strategy, as well as to Mike Meurer and Ben Roin for discussions on these legal topics, but of course none of them is responsible for any errors in my descriptions.

Appendix 2: Data

This appendix describes in additional detail the data sets used in my analysis.

Public sequencing data

I track the public sequencing efforts at the mRNA-by-year level from 1999 forward using the online US National Institutes of Health's (NIH) RefSeq database.[66] The RefSeq database is maintained by the National Center for Biotechnology Information (NCBI), a division of the US NIH's National Library of Medicine (NLM). As described on its website, the RefSeq (Reference Sequence) database *"...aims to provide a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins."*

Each RefSeq record represents a naturally occurring molecule from one organism, and is identified by a distinct RefSeq accession-version number (*e.g.* $NM\_000646.1$) that can be used to match RefSeq records with other databases. As noted above, RefSeq records are available for several types of molecules, including genomic DNA, transcripts, and proteins; the relevant molecule for a given RefSeq record is identifiable through the two prefix letters on the RefSeq number.[67] RefSeq records are available for many different organisms, including eukaryoktes, bacteria, and viruses; the relevant organism for a given RefSeq record is identifiable through the taxonomic ID number.[68] I focus on the human messenger RNA (mRNA) RefSeq records.

I use RefSeq release 34, which incorporates data available as of 6 March 2009. The catalog for RefSeq release 34 gives a list of accession/version numbers included in that database.[69] For each RefSeq accession/version number corresponding to a human mRNA transcript, I query (via a Python script) the online Sequence Revision History website to determine the date at which that record first appeared in the RefSeq database.[70]

It is important to note that the public sequencing efforts could be tracked in at least two other ways: using GenBank, another NCBI online database, or using genome assemblies. It is worth clarifying why I chose to track the public sequencing efforts through the RefSeq database, and what the advantages and disadvantages of these data are relative to the GenBank or genome assembly data.

GenBank is the "original" database to which individual laboratories submitted data under the Bermuda rules of the public sequencing effort, and in that sense is the most accurate measure of when a given section of DNA was sequenced by the public effort. Unfortunately, several characteristics of the GenBank data complicate its usefulness for this analysis. As described on the US Department of Energy website, GenBank is an "archival" database, containing records created by individual scientists.[71] Because of this, GenBank may contain hundreds of records documenting the same mRNA transcript. Unfortunately, no identification numbers exist that can link a GenBank record for a given mRNA transcript either to other GenBank records for the same mRNA transcript, or to other databases. Moreover, because there is no independent review system for sequence data submitted to GenBank, the data may contain errors. The RefSeq database was created specifically to overcome these shortcomings of the GenBank database that

[66] Available at `http://www.ncbi.nlm.nih.gov/RefSeq/`. See also Pruitt, Tatusova and Maglott (2007).

[67] The prefix letters for mRNA records are NM, NR, XM, and XR; see `ftp://ftp.ncbi.nih.gov/refseq/release/release-notes/archive/RefSeq-release34.txt`.

[68] The taxonomic ID number for humans is 9606; see `http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Undef&name=Homo+sapiens&lvl=0&srchmode=1`.

[69] Available at `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/RefSeq-release34.catalog.gz`.

[70] Available at `http://www.ncbi.nlm.nih.gov/entrez/sutils/girevhist.cgi`. I am very grateful to David Robinson for assistance in writing this script, which is available upon request.

[71] See `http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/sequence.shtml`.

complicated its use by researchers in many contexts. Many RefSeq records are derived from GenBank records, but RefSeq aims to provide non-redundant records that identify molecules by unique identification numbers, and that undergo a review process to screen for problems such as sequencing errors. RefSeq also includes some data not submitted to GenBank but available elsewhere (such as in published papers). The US Department of Energy website cited above notes, *"Since RefSeq records undergo a review process that screens for problems such as sequencing errors and vector contamination, RefSeq records are good sources of sequence information."* Although I have no systematic way of comparing dates of accession to GenBank with dates of accession to RefSeq, based on some hand-checks it appeared that (as expected) sequences appearing in RefSeq at earlier dates tended to be based on sequences that appeared in GenBank at earlier dates, with a relatively short lag. In sum, I rely on RefSeq records rather than GenBank records because RefSeq records identify unique mRNA observations with identification numbers that can be reliably matched to other databases, because scientists appear to rely on the RefSeq database as a source of sequencing data, and because based on some hand-checks dates of accession to RefSeq appear correlated with dates of accession to GenBank.

The second alternative would be to track the inclusion of mRNAs in NCBI's genome assemblies, which were released approximately annually over my time period of interest. Using the date an mRNA was first included in an NCBI genome assembly as the measure of the date of public sequencing could be more appropriate than the RefSeq measure if scientists primarily relied on the genome assemblies rather than the underlying mRNA transcript-level data. In practice, for the one assembly that I can easily compare these two measures they appear to be quite similar. Specifically, comparing the mRNA transcripts included in the NCBI-34 genome assembly from 2003 (described below in more detail) with the set of mRNA transcripts included in the RefSeq data as of 2003 suggests a relatively close correspondence: no mRNA transcripts were included in NCBI-34 but not included in the RefSeq data, and approximately 1,206 mRNA transcripts were included in RefSeq but not included in the NCBI-34 assembly (relative to 27,348 mRNA transcripts included in both datasets).[72] Relying on the RefSeq records rather than the genome assembly data is also preferable because the latter would require me to run analyses to compare various versions of the human genome assemblies, a task that is feasible but requires a relatively high level of scientific expertise. In sum, I rely on RefSeq records rather than comparisons of genome assemblies for computational ease, and because one comparison of the two measures suggested a close correspondence.

Celera sequencing data

For the private sector effort, there was essentially only one "version" of data, which I refer to as the Celera data. Comparing the Celera data with the public sequence data at a given point in time itself requires a non-trivial scientific analysis. Fortuitously for this work, a 2004 publication (Istrail et al., 2004) performed just such a comparison, and based on this analysis I am able to construct an mRNA-by-year level variable for whether a given mRNA transcript was included in the Celera data but had not yet appeared in the public sequencing data.

Specifically, Istrail et al. (2004) compare the Celera whole genome shotgun assembly (WGSA) as of December 2001 with the NCBI-34 (Build 34, October 2003) release of the public sector human genome assembly. Table 6 in Istrail et al. (2004) gives a list of RefSeq numbers for which the RefSeq mapping was longer in WGSA relative to NCBI-34, and Table 7 in Istrail et al. (2004) gives an analogous list of RefSeq numbers for which the RefSeq mapping was longer in

[72] One reason why an mRNA transcript may be included in the RefSeq data but not in the NCBI-34 assembly is if the transcript was sequenced but it was not clear where the transcript "fit" in terms of its location on the full human genome assembly.

NCBI-34 relative to WGSA.

I obtained an archived version of the mRNA transcripts included in NCBI-34 from the NCBI website (downloaded 27 April 2009), and used a Python script to extract the RefSeq numbers for each mRNA transcript in this data.[73] Three RefSeq IDs in this list were duplicates, and I drop one of each duplicate set. Matching this list to the RefSeq release 34 data described above, some records are included in NCBI-34 but not in RefSeq release 34 (largely "suspended" records), and some records are included in RefSeq release 34 but not in NCBI-34 (as expected, since RefSeq release 34 is a more recent dataset). I discard records in either NCBI-34 or in WGSA that are linked to mRNA transcripts listed in RefSeq release 34 as "suspended" records.

Table 6 of Istrail et al. (2004) lists RefSeq numbers for which the RefSeq mapping was longer in WGSA relative to NCBI-34, but this measure of length can be a fraction less than one – which would imply that a given mRNA transcript was partially but not entirely included in the NCBI-34 data. To be conservative, I define an mRNA transcript as being in the public domain if any part of the transcript was in the public domain according to the analysis of Istrail et al. (2004). Substantively, this means that I consider all RefSeq numbers listed in Table 6 of Istrail et al. (2004) to be in the public domain if any fraction of the transcript was in NCBI-34. Only four RefSeq numbers listed in Table 6 of Istrail et al. (2004) are listed as having been completely absent from the NCBI-34 data, and all four of these RefSeq numbers are listed in the RefSeq release 34 data as "suspended" records. Thus, for the purposes of my analysis there are no RefSeq numbers that were in the WGSA data but not in NCBI-34.

I construct an mRNA-by-year level variable for whether a given mRNA transcript was included in the Celera data but had not yet appeared in the public sequencing data as of 2001 as follows. Let A represent the RefSeq numbers in NCBI-34 but not in WGSA; let B represent the RefSeq numbers in both NCBI-34 and in WGSA; and let C represent the RefSeq numbers in WGSA but not in NCBI-34. Table 7 in Istrail et al. (2004) gives me the set A, and as noted above by my definition the set C has no elements. Together with the full NCBI-34 dataset described above, I can thus construct B as (NCBI-34) minus A. Some elements of B were in the set B as of 2001, whereas other elements of B were sequenced by the public effort sometime after 2001 and before the October 2003 NCBI-34 release. Because I wish to identify those mRNA transcripts that were only included in the Celera version of the human genome as of December 2001, I want to subtract off those elements of B that were added to the public database after December 2001. At the mRNA-year level, I thus create a 0/1 Celera variable, equaling one for observations in the following set:

$$B - (b \in B \mid b \text{ first appearing in RefSeq after December 2001}) + C$$

OMIM database: Publications and scientific knowledge outcome variables

I draw several gene-level outcome variables from the Online Mendelian Inheritance in Man (MIM, or OMIM), database.[74]

A paper version of MIM was initially created in the 1960s by Dr. Victor McKusick as a catalog of Mendelian traits and disorders ("Mendelian" here refers to the transmission of inherited characteristics through genes, named after Gregor Mendel, frequently referred to as the "father of genetics"). Twelve paper editions were published between 1966 and 1998. The online version, OMIM, was created in 1985 by a collaboration between the National Library

[73] Available at `ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/ARCHIVE/BUILD.34.1/RNA/rna.gbk.gz`. This script is available upon request.

[74] Available at `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim`. See also McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD).

of Medicine and the William H. Welch Medical Library at Johns Hopkins, and first became available on the internet in 1987. OMIM is currently authored and edited at the McKusick-Nathans Institute of Genetic Medicine at the Johns Hopkins University School of Medicine.

As described on its website, OMIM aims to provide a *"comprehensive, authoritative, and timely compendium of human genes and genetic phenotypes"* (a phenotype is an observable characteristic or trait of an organism). OMIM is updated daily, and is intended for use by physicians and other professionals concerned with genetic disorders, as well as genetics researchers.

OMIM includes six types of records:

- Genes of known sequence (indicated with an asterisk * preceding the MIM number);
- Descriptive entries, usually of phenotypes, that do not represent a unique locus on the human genome (indicated with a number symbol # preceding the MIM number);
- Descriptions of a gene of known sequence and phenotype (indicated with a plus sign + preceding the MIM number);
- Descriptions of a confirmed mendelian phenotype for which the underlying molecular basis is not known (indicated with a percent sign % preceding the MIM number);
- Descriptions of phenotypes with a suspected but unconfirmed mendelian basis, or with separateness from a phenotype in another OMIM entry that is unclear (indicated with the lack of a symbol preceding the MIM number);
- Removed records (indicated with a caret symbol ^ preceding the MIM number).

I create a "known, uncertain phenotype" indicator variable for whether a gene appears in any of these types of records, as a proxy for the gene being thought to be related to a given phenotype with some (potentially low) level of scientific certainty. I create a "known, certain phenotype" indicator variable for a gene appearing in either the second or the third type of OMIM records listed above, as a proxy for the gene being thought to be related to a given phenotype with a higher level of scientific certainty. OMIM records cite published scientific papers relevant for each record, which I collect as an additional outcome variable.

These OMIM outcome variables are collected in a cross-section (in 2009), but for two of the outcome variables, I am able to construct gene-by-year measures for use in the panel specification. First, I use paper publication dates to construct the number of publications by gene by year. Second, and less straightforward, I construct the first date each "known, uncertain phenotype" link appears in OMIM. I observe this latter measure with error, but expect this error to be uncorrelated with the Celera IP treatment variable. Specifically, the measurement error arises because OMIM includes entries of some phenotypes with unknown genotypes, some of which transition to become entries of phenotypes with known genotypes over time, and I do not observe these transition dates but rather observe the initial date any part of the entry appeared in OMIM. For example, Huntington's disease was known to be a genetic disease prior to the sequencing of the Huntingtin gene, and my measurement of this date would likely capture the first date Huntington's disease was included in the OMIM database rather than the date when the sequenced gene allowed the full genotype-phenotype link to be listed in OMIM.

Each OMIM record includes a distinct MIM number (*e.g.* +611082), which can be used to match OMIM records with other databases. One gene can be included in more than one OMIM record, and one OMIM record can involve more than one gene; I collapse the OMIM data to the gene level. For the measures of genotype-phenotype links, I take the maximum of indicator variables by gene, and for the publications measure I sum the total number of publications relevant to that gene from all OMIM entries.

I use the full-text OMIM version of 19 April 2009 and extract, via a Python script, the outcome variables described above for each OMIM record in this text file.[75]

GeneTests.org database: Diagnostic test availability outcome variable

I draw a gene-level indicator for the availability of any genetic test related to that gene from the US NIH's GeneTests.org online database.[76]

As described on its website, GeneTests.org includes a laboratory directory that is a self-reported, voluntary listing of US and international laboratories offering in-house molecular genetic testing, specialized cytogenetic testing, and biochemical testing for inherited disorders. US-based laboratories listed in GeneTests.org must be certified under the Clinical Laboratory Improvement Amendment (CLIA) of 1988, which requires laboratories to meet quality control and proficiency testing standards; there are no such requirements for non-US-based laboratories.

The GeneTests.org website clarifies several types of information *not* included in its laboratory directory, including genetic testing on the diagnosis and/or monitoring of solid tumors, hematologic malignancies, infectious diseases, and forensic testing.

As described on its website, GeneTests.org aims to provide *"current, authoritative information on genetic testing and its use in diagnosis, management, and genetic counseling"* to promote *"the appropriate use of genetic services in patient care and personal decision making."* Originally based at the University of Washington in Seattle, GeneTests.org has been funded by a series of federal grants and is currently hosted at the US National Institutes of Health's National Center for Biotechnology Information (NCBI).

I use the GeneTests.org data as of 27 May 2009, which lists OMIM numbers for which there is any genetic test available in the GeneTests.org directory.[77] As with the OMIM data described above, one gene can be included in more than one OMIM record, and one OMIM record can involve more than one gene; I collapse the GeneTests.org data to the gene level, taking the maximum of this indicator variable by gene.

Gene-level covariates: Cytogenetic and molecular location variables

I draw several gene-level variables describing the location of a particular gene on the human genome from the US NIH's Entrez Gene database.[78]

Geneticists use two types of variables to describe a gene's location on the human genome: cytogenetic location and molecular location.[79] Cytogenetic variables take forms such as $19q13.4$. For this example, 19 represents the chromosome on which the gene is located (1-22, X, or Y). The letter q represents the arm of the chromosome on which the gene is located; each chromosome is divided into two arms based on the location of a narrowing called the centromere - a shorter arm (p) and a longer arm (q). The numbers after the arm letter describe the position of the gene on the p or q arm, usually designated by two digits (representing a region and a band) and sometimes followed by a decimal point and one or more additional digits (representing

[75] The current full-text OMIM version is available at `ftp://ftp.ncbi.nih.gov/repository/OMIM/omim.txt.Z`. The 19 April 2009 version I use in the analysis is available upon request. I am very grateful to David Robinson for assistance in writing this script, which is available upon request.

[76] Available at `http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests`. See also University of Washington, Seattle (2009).

[77] The current GeneTests.org data is available at `ftp://ftp.ncbi.nih.gov/pub/GeneTests/DiseaseOMIM.txt`. The 27 May 2009 version I use in the analysis is available on request.

[78] Available at `ftp://ftp.ncbi.nih.gov/gene/`. See also Maglott et al. (2005).

[79] The data description in this section draws heavily on the discussion in `http://ghr.nlm.nih.gov/handbook/howgeneswork/genelocation`.

sub-bands). These numbers increase with distance from the centromere.

Molecular location variables are in a sense more precise than cytogenetic location variables in that they describe a gene's location in terms of base pairs. For example, according to the NIH's National Center for Biotechnology Information (NCBI) database, the APOE gene on chromosome 19 begins with base pair 50,100,901 and ends with base pair 50,104,488. Together, these variables tell us both the precise position of the gene and the size of the gene (3,588 base pairs). However, different databases often present slightly different values for these variables.

I use two Entrez Gene files from 18 June 2009: the $gene2refseq$ file and the $gene\_info$ file.[80]

From the $gene2refseq$ file, I extract continuous variables for the start and end base pairs of the gene on the genomic accession (as well as indicator variables for uncertain start and end base pair data) and for the orientation of the gene on the genomic accession (plus and minus, as well as an indicator variable for uncertain orientation data). The $gene2refseq$ observations are at the mRNA-level (identified by RefSeq accession/version numbers), but can include more than one observation for a given mRNA. I collapse this data to the gene level, taking the mean of each variable over all available observations.

From the $gene\_info$ file, I extract indicator variables for the chromosome on which the gene is located (1-22, X, Y, and an indicator for uncertain chromosome data), indicator variables for the arm of the chromosome on which the gene is located (p, q, and an indicator for uncertain arm data), and continuous variables for the region, band, and subband position of the gene on the relevant arm (as well as indicator variables for uncertain region, band, or subband data).[81]

Other gene-level covariates: Disclosure dates

Using data already described above, I construct an additional set of gene-level covariates that *a priori* are likely to affect the amount of research conducted on a given gene: namely, indicator variables for the year sequence data for the gene was first disclosed.

Intuitively, genes sequenced earlier have been "at risk" for research based on the sequenced data for a longer period of time, which we would expect to affect the total amount of research observed as of 2009. I define the date of sequence data disclosure as the minimum of (1) the first year I observe the sequence data in the RefSeq database; and (2) 2001, if the sequence data was included in the Celera data (since the Celera data was publicly disclosed, as discussed in Section 2.4). Note that this minimum is taken over all mRNA transcripts for each gene, so measures the *earliest* date at which sequence data for any mRNA transcript on each gene was disclosed. I chose to use this disclosure date because of a concern that disclosure dates for other mRNA transcripts on a gene may be endogenous to the Celera IP treatment variable of interest. That said, the disclosure date for a gene is unique for the majority of genes, since they produce only one known mRNA transcript.

RefSeq-to-gene and gene-to-OMIM crosswalks

I use NCBI-generated crosswalks to map RefSeq accession/version numbers to Entrez Gene ID numbers and to match Entrez Gene ID numbers to OMIM numbers.[82]

[80] The current versions of these two databases are available at `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2refseq.gz` and `ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz`. The 18 June 2009 versions I use in the analysis are available upon request.

[81] I made eight hand-corrections to the chromosome variable based on redundant information provided in the map location variable, and one hand-correction to the region variable - changing a zero region value (which only appeared once in the data) to an uncertain region value.

[82] Available at `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/release34.accession2geneid.gz` and at `ftp://ftp.ncbi.nih.gov/gene/DATA/mim2gene`, respectively.

Appendix 3: Additional tables and figures

Table A1: Differences Across Celera Genes by Year of Re-sequencing in Gene-Level Data: Sample of Celera Genes

	public in 2002 mean	*public in 2003* mean	*p*-value of difference
Panel A: Outcome variables			
publications in 2001-2009	1.194	1.313	[0.644]
0/1, known, uncertain phenotype	0.414	0.381	[0.188]
0/1, known, certain phenotype	0.053	0.033	[0.052]
0/1, used in any diagnostic test	0.032	0.027	[0.509]
Panel B: Main covariates			
year first mRNA disclosed	2001.000	2001.000	-
publications in 1970	0.010	0.005	[0.314]
publications in 1971	0.006	0.005	[0.784]
publications in 1972	0.006	0.002	[0.347]
publications in 1973	0.010	0.006	[0.562]
publications in 1974	0.008	0.008	[0.968]
publications in 1975	0.009	0.003	[0.218]
publications in 1976	0.010	0.019	[0.284]
publications in 1977	0.009	0.011	[0.823]
publications in 1978	0.015	0.019	[0.669]
publications in 1979	0.016	0.036	[0.135]
publications in 1980	0.013	0.019	[0.560]
publications in 1981	0.014	0.025	[0.284]
publications in 1982	0.017	0.020	[0.750]
publications in 1983	0.013	0.030	[0.156]
publications in 1984	0.021	0.038	[0.252]
publications in 1985	0.026	0.033	[0.642]
publications in 1986	0.017	0.025	[0.500]
publications in 1987	0.024	0.039	[0.412]
publications in 1988	0.028	0.060	[0.242]
publications in 1989	0.019	0.071	[0.009]
publications in 1990	0.017	0.044	[0.046]
publications in 1991	0.023	0.052	[0.167]
publications in 1992	0.028	0.047	[0.220]
publications in 1993	0.034	0.053	[0.189]
publications in 1994	0.039	0.035	[0.756]
publications in 1995	0.050	0.040	[0.529]
publications in 1996	0.065	0.055	[0.679]
publications in 1997	0.052	0.077	[0.188]
publications in 1998	0.061	0.090	[0.162]
publications in 1999	0.087	0.083	[0.892]
Panel C: Additional covariates			
0/1, missing cytogenetic location	0.203	0.184	[0.337]
0/1, missing molecular location	0.018	0.025	[0.326]
N	1,047	635	

Notes: Gene-level observations. Sample includes all Celera genes (that is, genes for which all mRNAs on the gene were initially sequenced only by Celera as of 2001). The first column includes Celera genes for which the first mRNA re-sequenced by the public effort was re-sequenced in 2002 (N = 1,047), and the second column includes Celera genes for which the first mRNA re-sequenced by the public effort was re-sequenced in 2003 (N = 635). In an ordinary-least-squares model predicting *"public in 2003"*: 0/1, =1 if the first mRNA re-sequenced by the public effort was re-sequenced in 2003, as a function of the count variables for publications in each year from 1970-1999, the p-value from an F-test is 0.169. See text and Appendix 2 for more detailed data and variable descriptions.

Table A2: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Coefficients on Covariates

outcome variable:	(1) publications 2001-2009	(2) uncertain phenotype	(3) certain phenotype	(4) used in any diagnostic test
Covariates				
disclosed in 2000	1.401	0.439	0.061	0.038
	(0.228)***	(0.018)***	(0.006)***	(0.005)***
disclosed in 2001	1.199	0.313	0.046	0.031
	(0.227)***	(0.019)***	(0.006)***	(0.005)***
disclosed in 2002	1.019	0.275	0.047	0.034
	(0.242)***	(0.024)***	(0.009)***	(0.008)***
disclosed in 2003	0.877	0.190	0.038	0.019
	(0.249)***	(0.024)***	(0.008)***	(0.007)***
disclosed in 2004	-0.142	-0.007	0.011	0.002
	(0.247)	(0.020)	(0.006)*	(0.005)
disclosed in 2005	-	-	-	-
	-	-	-	-
disclosed in 2006	-1.460	-0.197	-0.009	-0.008
	(0.325)***	(0.017)***	(0.004)*	(0.004)**
disclosed in 2007	-0.463	-0.060	-0.006	-0.002
	(0.277)*	(0.021)***	(0.006)	(0.005)
disclosed in 2008	-3.086	-0.230	-0.012	-0.009
	(0.438)***	(0.016)***	(0.004)***	(0.004)**
disclosed in 2009	-1.947	-0.167	-0.010	-0.008
	(0.360)***	(0.020)***	(0.005)**	(0.004)*
publications in 1970	0.591	0.048	0.097	0.129
	(0.117)***	(0.020)**	(0.034)***	(0.033)***
publications in 1971	0.210	0.089	0.228	0.210
	(0.134)	(0.024)***	(0.049)***	(0.045)***
publications in 1972	0.191	-0.003	0.071	0.081
	(0.152)	(0.021)	(0.043)*	(0.037)**
publications in 1973	-0.159	0.058	0.082	0.071
	(0.141)	(0.026)**	(0.049)*	(0.046)
publications in 1974	-0.386	0.026	0.063	0.020
	(0.174)**	(0.023)	(0.044)	(0.040)
publications in 1975	0.289	0.050	0.043	0.005
	(0.103)***	(0.019)***	(0.038)	(0.037)
publications in 1976	0.347	0.059	0.111	0.066
	(0.101)***	(0.015)***	(0.030)***	(0.029)**
publications in 1977	0.092	0.017	0.038	0.049
	(0.085)	(0.016)	(0.034)	(0.032)
N	21,824	21,824	21,824	21,824

Notes: This table shows the coefficients on the covariates included in Column (2) of Table 3; see the notes of Table 3 for details on this specification. Omitted year of disclosure is 2005.

Table A3: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Probit Models

	(1)	(2)
Panel A: 0/1, known, uncertain phenotype		
mean = 0.309		
celera	-0.101	-0.094
	(0.008)***	(0.008)***
Panel B: 0/1, known, certain phenotype		
mean = 0.039		
celera	-0.007	-0.004
	(0.002)***	(0.002)**
Panel C: 0/1, used in any diagnostic test		
mean = 0.027		
celera	-0.006	-0.004
	(0.001)***	(0.001)***
indicator variables for year of disclosure	yes	yes
number of publications in each year 1970-77	no	yes
N	21,824	21,824

Notes: Gene-level observations. Reported coefficients are marginal effects from probit models. Sample includes all genes sequenced in or after 2000 (N = 21,824). Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Table A4: Selection into Celera IP: Sample of Genes Sequenced in or after 2000

Celera IP treatment	
mean = 0.060	
publications in 1970	-0.009
	(0.021)
publications in 1971	-0.037
	(0.026)
publications in 1972	-0.034
	(0.024)
publications in 1973	0.033
	(0.022)
publications in 1974	0.004
	(0.020)
publications in 1975	-0.023
	(0.021)
publications in 1976	0.006
	(0.017)
publications in 1977	-0.018
	(0.020)
publications in 1978	0.007
	(0.015)
publications in 1979	0.036
	(0.014)***
publications in 1980	-0.007
	(0.014)
publications in 1981	-0.002
	(0.015)
publications in 1982	0.005
	(0.013)
publications in 1983	-0.005
	(0.011)
publications in 1984	0.029
	(0.011)**
publications in 1985	0.004
	(0.011)
publications in 1986	-0.016
	(0.013)
publications in 1987	-0.010
	(0.011)
publications in 1988	0.013
	(0.009)
publications in 1989	0.018
	(0.010)*
publications in 1990	-0.009
	(0.011)
publications in 1991	-0.015
	(0.010)
publications in 1992	-0.009
	(0.009)
publications in 1993	0.014
	(0.008)*
publications in 1994	-0.011
	(0.008)
publications in 1995	-0.003
	(0.007)
publications in 1996	0.007
	(0.006)
publications in 1997	0.001
	(0.007)
publications in 1998	-0.001
	(0.006)
publications in 1999	-0.009
	(0.005)
N	21,824

Notes: Gene-level observations. The dependent variable is *"celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. Coefficients are marginal effects from probit models. Sample includes all genes sequenced in or after 2000 (N = 21,824). Robust standard errors shown in parentheses. *: $p< 0.10$; **: $p< 0.05$; ***: $p< 0.01$. See text and Appendix 2 for more detailed data and variable descriptions.

Table A5: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Propensity Score Models

	(1)	(2)	(3)	(4)
Panel A: publications in 2001-2009				
mean = 1.095				
celera	-0.324	-0.297	-0.422	-0.384
	(0.241)	(0.118)**	(0.106)***	(0.103)***
Panel B: 0/1, known, uncertain phenotype				
mean = 0.309				
celera	-0.157	-0.153	-0.155	-0.155
	(0.015)***	(0.015)***	(0.015)***	(0.015)***
Panel C: 0/1, known, certain phenotype				
mean = 0.039				
celera	-0.021	-0.014	-0.016	-0.014
	(0.009)**	(0.007)**	(0.006)***	(0.006)**
Panel D: 0/1, used in any diagnostic test				
mean = 0.027				
celera	-0.018	-0.012	-0.014	-0.012
	(0.008)**	(0.006)**	(0.005)***	(0.005)**
inverse probability weighting	yes	yes	no	no
blocking	no	no	yes	yes
indicator variables for year of disclosure	yes	yes	yes	yes
number of publications in each year 1970-77	no	yes	no	yes
N	21,824	21,824	21,766	21,766

Notes: Gene-level observations. Appendix Table A4 reports marginal effects from a probit model in which the dependent variable is *"celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001, predicted as a function of the count variables for the number of publications in each year from 1970 to 1999. This table uses the predicted probability of Celera IP treatment from that model in two propensity score specifications: Columns (1) and (2) use the propensity score to construct inverse probability weights, and Columns (3) and (4) break the data into blocks based on the propensity score, and includes fixed effects for each block as covariates. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. Sample includes all genes sequenced in or after 2000 (N = 21,824); following Dehejia and Wahba (1999), Columns (3) and (4) drop non-Celera genes with a predicted probability of treatment less than the minimum or greater than the maximum predicted probability of treatment among Celera genes, hence the smaller sample size (N = 21,766). Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Table A6: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Additional Publication Controls

	(1)	(2)	(3)	(4)
Panel A: publications in 2001-2009				
mean = 1.095				
celera	-0.432	-0.523	-0.456	-0.418
	(0.112)***	(0.104)***	(0.100)***	(0.104)***
Panel B: 0/1, known, uncertain phenotype				
mean = 0.309				
celera	-0.158	-0.160	-0.151	-0.151
	(0.015)***	(0.015)***	(0.015)***	(0.015)***
Panel C: 0/1, known, certain phenotype				
mean = 0.039				
celera	-0.018	-0.022	-0.014	-0.012
	(0.006)***	(0.006)***	(0.006)**	(0.006)**
Panel D: 0/1, used in any diagnostic test				
mean = 0.027				
celera	-0.015	-0.019	-0.012	-0.011
	(0.005)***	(0.005)***	(0.005)**	(0.005)**
indicator variables for year of disclosure	yes	yes	yes	yes
number of publications in each year 1970-77	yes	no	no	yes
number of publications in each year 1980-89	no	yes	no	yes
number of publications in each year 1990-99	no	no	yes	yes
N	21,824	21,824	21,824	21,824

Notes: Gene-level observations. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. Sample includes all genes sequenced in or after 2000 (N = 21,824). Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. *Number of publications in each year 1980-89*: ten count variables for the number of publications in each year from 1980 to 1989. *Number of publications in each year 1990-99*: ten count variables for the number of publications in each year from 1990 to 1999. See text and Appendix 2 for more detailed data and variable descriptions.

Table A7: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Additional Location Covariates

	(1)	(2)	(3)
Panel A: publications in 2001-2009			
mean = 0.944			
celera	-0.557	-0.502	-0.448
	(0.132)***	(0.125)***	(0.127)***
Panel B: 0/1, known, uncertain phenotype			
mean = 0.292			
celera	-0.138	-0.134	-0.125
	(0.018)***	(0.018)***	(0.018)***
Panel C: 0/1, known, certain phenotype			
mean = 0.036			
celera	-0.027	-0.019	-0.014
	(0.008)***	(0.007)***	(0.007)**
Panel D: 0/1, used in any diagnostic test			
mean = 0.025			
celera	-0.023	-0.015	-0.012
	(0.007)***	(0.006)**	(0.006)**
indicator variables for year of disclosure	yes	yes	yes
number of publications in each year 1970-77	no	yes	yes
detailed cytogenetic & molecular covariates	no	no	yes
N	13,871	13,871	13,871

Notes: Gene-level observations. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. Sample includes all genes with non-missing data on all cytogenetic and molecular location variables sequenced in or after 2000 (N = 13,871). Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. *Detailed cytogenetic & molecular covariates*: 0/1 indicator variables for the chromosome (1-22, X, or Y) and arm (p or q) on which a gene is located; continuous variables for region, band, subband, start base pair, and end base pair; and 0/1 indicator variables for the orientation of the gene on the genome assembly (plus or minus). See text and Appendix 2 for more detailed data and variable descriptions.

Table A8: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Alternative Comparison Samples

sample includes non-Celera genes sequenced in:	(1) all	(2) all	(3) 2001
Panel A: publications in 2001-2009			
full sample mean = 2.197			
2001 sample mean = 1.791			
celera	-0.535	-0.517	-0.354
	(0.117)***	(0.114)***	(0.103)***
Panel B: 0/1, known, uncertain phenotype			
full sample mean = 0.453			
2001 sample mean = 0.503			
celera	-0.162	-0.161	-0.157
	(0.015)***	(0.015)***	(0.015)***
Panel C: 0/1, known, certain phenotype			
full sample mean = 0.081			
2001 sample mean = 0.063			
celera	-0.027	-0.022	-0.018
	(0.007)***	(0.007)***	(0.006)***
Panel D: 0/1, used in any diagnostic test			
full sample mean = 0.060			
2001 sample mean = 0.045			
celera	-0.023	-0.019	-0.015
	(0.006)***	(0.006)***	(0.005)***
indicator variables for year of disclosure	yes	yes	-
number of publications in each year 1970-77	no	yes	yes
N	27,882	27,882	4,533

Notes: Gene-level observations. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. Sample includes all genes (N = 27,882) in Columns (1) and (2), and all genes sequenced in 2001 (N = 4,533) in Column (3). I do not show estimates for the sample of all genes sequenced in 2001 without the publication covariates, because these estimates are identical to those in Column (1) since all Celera genes were sequenced in 2001. Robust standard errors shown in parentheses. *: $p<$ 0.10; **: $p<$ 0.05; ***: $p<$ 0.01. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Table A9: Cross-Section Estimates of the Impact of Celera IP on Innovation Outcomes: Sample of Genes Sequenced in or after 2000, Combined 2000/2001 Variable

	(1)	(2)
Panel A: publications in 2001-2009		
mean = 1.095		
celera	-0.717	-0.555
	(0.109)***	(0.105)***
Panel B: 0/1, known, uncertain phenotype		
mean = 0.309		
celera	-0.239	-0.232
	(0.013)***	(0.013)***
Panel C: 0/1, known, certain phenotype		
mean = 0.039		
celera	-0.041	-0.027
	(0.006)***	(0.005)***
Panel D: 0/1, used in any diagnostic test		
mean = 0.027		
celera	-0.032	-0.020
	(0.005)***	(0.005)***
indicator variables for year of disclosure	yes	yes
number of publications in each year 1970-77	no	yes
N	21,824	21,824

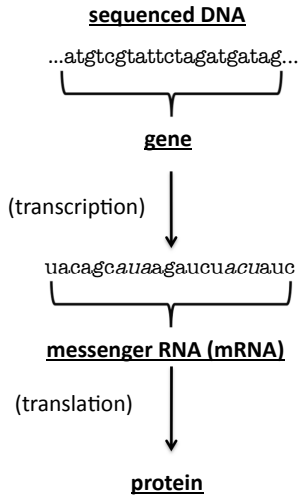
Notes: Gene-level observations. Estimates in Panel A are from quasi-maximum likelihood Poisson models; estimates in Panels B-D are from ordinary-least-squares (OLS) models. The year indicator variables for 2000 and 2001 are combined in this table into one variable, to account for uncertainty over the exact date of sequencing for Celera genes in those years. Sample includes all genes sequenced in or after 2000 (N = 21,824). Robust standard errors shown in parentheses. *: $p<0.10$; **: $p<0.05$; ***: $p<0.01$. *"Celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Table A10: Panel Estimates of the Impact of Celera IP on Innovation Outcomes: Full Sample of Genes

	(1)	(2)	(3)
Panel A: gene-year publications			
mean = 0.244			
celera	-0.160	-0.145	-0.109
	(0.017)***	(0.015)***	(0.011)***
Panel B: 0/1, known, uncertain phenotype			
mean = 0.381			
celera	-0.163	-0.162	-0.083
	(0.009)***	(0.009)***	(0.008)***
year fixed effects	yes	yes	yes
indicator variables for year of disclosure	yes	yes	no
number of publications in each year 1970-77	no	yes	no
gene fixed effects	no	no	yes
N	250,938	250,938	250,938

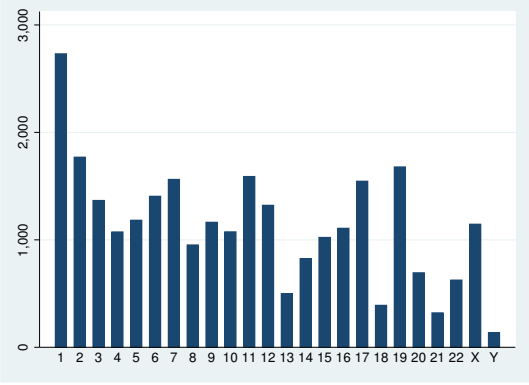
Notes: Gene-year-level observations. All estimates are from ordinary-least-squares (OLS) models. As discussed in Section 2.3, Celera's human genome sequencing efforts commenced in September 1999, and its draft human genome was disclosed in 2001. Unfortunately, I do not observe the timing of when specific genes were sequenced within this time frame. In the absence of such data, I limit my panel specification to include the years 2001-2009 since prior to 2001 I do not know whether or not Celera genes had yet been sequenced. The sample includes all gene-years from 2001 to 2009 (27,882 genes, for 9 years, implies $N = 250{,}938$ total gene-year observations). Robust standard errors, clustered at the gene level, shown in parentheses. *: $p< 0.10$; **: $p< 0.05$; ***: $p< 0.01$. *"Celera"*: 0/1, =1 if all mRNAs on the gene were sequenced only by Celera in that year. *Indicator variables for year of disclosure*: 0/1 indicator variables for the first year the sequence for any mRNA on the gene was disclosed, defined as the minimum of: (1) the first year any mRNA for the gene appears in the RefSeq database; and (2) 2001, if the mRNA was included only in the Celera data as of 2001 (since the Celera data was publicly disclosed in 2001, as discussed in Section 2.4). *Number of publications in each year 1970-77*: eight count variables for the number of publications in each year from 1970 to 1977. See text and Appendix 2 for more detailed data and variable descriptions.

Figure A1: Overview of Scientific Background on the Sequencing of the Human Genome



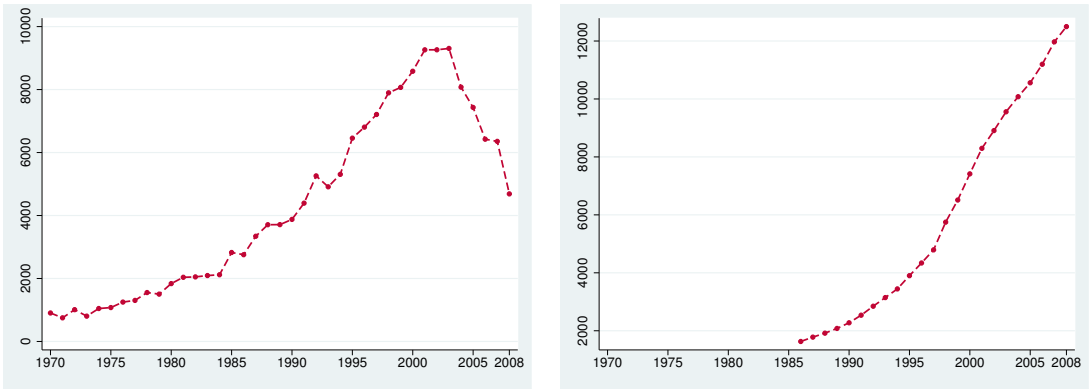
Notes: This figure summarizes the scientific overview discussed in Section 2.1. Sequenced DNA refers to the exact order of nucleotide bases (adenine, cytosine, guanine, and thymine) in a given stretch of DNA. Genes can be identified from a given segment of sequenced DNA. Genes manufacture proteins through a two-step process of transcription and translation. In the transcription process, a messenger ribonucleic acid (mRNA) transcript is generated. A mRNA transcript is complementary to DNA (that is, pairing adenine with thymine, and cytosine with guanine), except that uracil is substituted for thymine (hence, u is substituted for t in the figure). In addition, some portions of code (italicized, in the figure) may be removed from the complementary mRNA code relative to the DNA code. In the translation process, the mRNA transcript is used to generate a protein; genes are able to encode more than one protein through generating more than one mRNA transcript. Proteins in turn carry out functions in the human body.

Figure A2: Distribution of Genes Across Chromosomes: Full Sample of Genes



Notes: This figure shows the frequency distribution of genes across human chromosomes (as discussed in Section 4.1). See text and Appendix 2 for more detailed data and variable descriptions.

Figure A3: Summary Statistics for Gene-Year Level Data: Full Sample of Genes

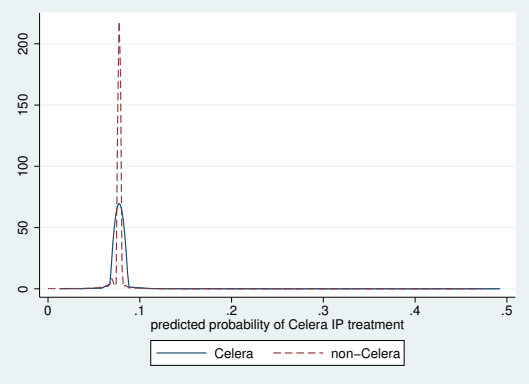


(a) Number of Total Flow Gene-Year Publications across All Genes, by Year

(b) Cumulative Number of Genes with any Known/Uncertain Phenotype Link, by Year

Notes: These figures show aggregate summary statistics by year for the two gene-year outcome variables: gene-year publications, and a gene-year indicator for whether a gene has any known, uncertain phenotype link. As discussed in Section 4.1, Panel (a) suggests flow publications peaked by this measure in 2003, although it is likely that some of the post-2003 decline is due to time lags in the addition of scientific publications to the OMIM database. In the panel specifications using the gene-year level data, the inclusion of year fixed effects will remove any year-specific shocks to the overall level of research that are common across genes, such as time lags in updating of the OMIM database. See text and Appendix 2 for more detailed data and variable descriptions.

Figure A4: Distribution of Predicted Probability of Celera IP Treatment, for Celera and non-Celera Genes Sequenced in or after 2000



Notes: This figure shows the distribution of the predicted probability of Celera IP treatment, for Celera and non-Celera genes, as estimated on gene-level data in Appendix Table A4. Appendix Table A4 reports marginal effects from a probit model in which the dependent variable is *"celera"*: 0/1, =1 if all mRNAs on the gene were initially sequenced only by Celera as of 2001, predicted as a function of the count variables for the number of publications in each year from 1970 to 1999. See text and Appendix 2 for more detailed data and variable descriptions.