How Do Patents Affect Follow-on Innovation? Evidence from the Human Genome*

Bhaven Sampat
Columbia and NBER

Heidi L. Williams
MIT and NBER

October 31, 2017

Abstract

We investigate whether patents on human genes have affected follow-on scientific research and product development. Using administrative data on successful and unsuccessful patent applications submitted to the US Patent and Trademark Office, we link the exact gene sequences claimed in each application with data measuring follow-on scientific research and commercial investments. Using this data, we document novel evidence of selection into patenting: patented genes appear more valuable—prior to being patented—than non-patented genes. This evidence of selection motivates two quasi-experimental approaches, both of which suggest that on average gene patents have had no quantitatively important effect on follow-on innovation.

*Daron Acemoglu, Josh Angrist, David Autor, Pierre Azoulay, Stefan Bechtold, Nick Bloom, Tim Bresnahan, Joe Doyle, Dan Fetter, Amy Finkelstein, Alberto Galasso, Nancy Gallini, Joshua Gans, Aaron Kesselheim, Pat Kline, Amanda Kowalski, Mark Lemley, Josh Lerner, Petra Moser, Ben Olken, Ariel Pakes, Jim Poterba, Arti Rai, Mark Schankerman, Scott Stern, Mike Whinston, and seminar participants at Analysis Group, Brown, Chicago Booth, Clemson, Dartmouth, Duke, the Federal Reserve Board, Harvard, HBS, MIT, Northwestern Kellogg, the NBER (Law and Economics, Productivity and Public Economics), Stanford, UC-Berkeley Haas, UC-Santa Barbara, the USPTO, and Williams College provided very helpful comments. We are very grateful to Ernie Berndt for help with accessing the Pharmaprojects data; to Osmat Jefferson, the CAMBIA Lens initiative, Lee Fleming, and Guan-Cheng Li for sharing USPTO-related data; and to Joey Anderson, Jeremy Brown, Lizi Chen, Alex Fahey, Cirrus Foroughi, Yunzhi Gao, Grant Graziani, Kelly Peterson, Lauren Russell, Mahnum Shahzad, Sophie Sun, Nicholas Tilipman, Myles Wagner, and Hanwen Xu for excellent research assistance. Research reported in this publication was supported by the National Institute on Aging and the NIH Common Fund, Office of the NIH Director, through Grant U01-AG046708 to the National Bureau of Economic Research (NBER); the content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NBER. This work/research was also funded by the Ewing Marion Kauffman Foundation; the contents of this publication are solely the responsibility of the Grantee. Financial support from NIA Grant Number T32-AG000186 to the NBER, NSF Grant Number 1151497, the NBER Innovation Policy and the Economy program, and the Toulouse Network for Information Technology is also gratefully acknowledged. Contact: bns3@columbia.edu, heidiw@mit.edu.

1 Introduction

Competitive markets may under-incentivize innovation due to the public good nature of new ideas. Intellectual property rights, such as patents, aim to address this under-investment problem by allowing inventors to capture a higher share of the social returns to their research investments. By awarding inventors a temporary right to exclude others from marketing their invention, patents aim to allow inventors to earn quasi-rents—temporarily—as a way to re-coup their research and development costs, thus providing dynamic incentives for investments in new technologies. Dating back at least to analyses such as Nordhaus (1969), optimal patent policy design has traditionally been framed as a trade-off between this benefit of providing incentives for the development of new technologies and the cost of deadweight loss from higher prices during the life of the patent.

Nordhaus-style models of optimal patent policy design have traditionally modeled innovations as isolated discoveries. However, in practice many or most innovations are cumulative, in the sense that a given discovery may be an input into later follow-on discoveries. When innovation is cumulative, optimal patent policy design also depends on how patents on existing technologies affect follow-on innovation. One prominent example in recent policy debates is a patented human gene sequence. Sequenced genetic data is a research input into subsequent follow-on discoveries: by analyzing sequenced genetic data scientists may discover links between genetic variations and diseases, and such knowledge can be applied to commercialize medical technologies such as pharmaceutical treatments and diagnostic tests. If patents on discoveries such as human genes affect follow-on innovation, that effect is a key additional input into optimal patent policy design. In this paper, we investigate whether patents on human genes have affected follow-on scientific research and product development. Our broad goal is to inform whether the Nordhaus-style trade-off between ex ante incentives and deadweight loss is sufficient for optimal patent policy design, or whether—at least in this context—the effects of patents on follow-on innovation need to be considered.

Investigating how patents on existing technologies affect follow-on innovation requires addressing two key challenges. First, in most markets it is difficult or impossible to take a given set of technologies which are claimed as intellectual property in patents and measure follow-on innovation on those technologies. Follow-on innovations legally require a license from the original innovator. But because licenses are very rarely publicly disclosed, constructing appropriate measures of follow-on innovation for any given patented invention is quite difficult. Second, a priori we expect a selection bias problem to arise if inventors are more likely to file for or to obtain patents on technologies that are inherently more valuable. This type of selection raises the concern that any measured differences in follow-on innovation across patented and

non-patented technologies may reflect the selection of which technologies are patented, rather than a causal effect of patents on follow-on innovation.

The contribution of this paper is to construct new data and develop two new quasi-experimental approaches to address these challenges. To address the first challenge—measurement—we take advantage of the fact that US patent applications claiming human genes as intellectual property must disclose the exact DNA sequences claimed in the text of the patent. By applying bioinformatics methods (Jensen and Murray 2005), these DNA sequences can be annotated with gene identifiers, and these gene identifiers can in turn be linked to standard scientific and medical datasets providing measures of follow-on scientific research and product development related to each gene. Specifically, we measure the scientific publications related to each gene as an indicator of scientific research investments, and measure the use of genes in pharmaceutical clinical trials and diagnostic tests as indicators of commercial research investments. Gene patents have been widely interpreted as sufficiently broad that these types of commercial follow-on inventions would require gene patent licenses. For example, then-United States Patent and Trademark Office (USPTO) biotechnology examination unit head John Doll noted in 1998 (Doll 1998): "*...once a [gene] is patented, that patent extends to any use, even those that have not been disclosed in the patent.*" This implies that our measures of follow-on innovation correspond closely to the way follow-on innovation has been defined in the theoretical literature on cumulative innovation, where licenses are required for follow-on innovations developed outside of the firm holding the patent.

Because we observe our measures of follow-on scientific research and product development for all genes, this DNA sequence-based linkage allows us to compare follow-on innovation across patented and non-patented genes. If patents were as good as randomly assigned across genes, this would be sufficient to estimate the causal effect of interest. However, if inventors are more likely to file for and obtain patents on technologies that are more valuable, then this type of simple comparison could instead reflect the selection of which genes are patented. By taking advantage of the fact that we observe our measures of follow-on innovation both before and after gene patents are granted, we are able to document novel evidence of selection into patenting: genes that will be patented in the future are the focus of more scientific research and more commercial investments prior to being patented, relative to genes that will not be patented. This evidence suggests that estimating a causal effect of patents on follow-on innovation requires constructing an empirical strategy to address this selection bias.

To address this second challenge—selection bias—we develop two new quasi-experimental methods for estimating how patents have affected follow-on innovation. First, we present a simple comparison of follow-on innovation across genes claimed in accepted versus rejected patent applications.[1] In our context,

[1]Strictly speaking, patent applications are never formally rejected by the USPTO, only abandoned by applicants (see the

this method is reasonable if—conditional on being included in a patent application—whether a gene is granted a patent is as good as random. Again taking advantage of the fact that we observe our measures of follow-on innovation both before and after gene patents are granted, we document empirically that genes claimed in accepted and rejected patent applications are the focus of similar levels of scientific research and commercial investments prior to the applications being filed, providing evidence for the validity of this empirical approach. Second, we develop a novel instrumental variable for which patent applications are granted patents: the "leniency" of the assigned patent examiner. Patent examiners are charged with a uniform mandate: grant patents to patent-eligible, novel, non-obvious, and useful inventions. However, prior research has documented that in practice this mandate leaves patent examiners a fair amount of discretion (Cockburn et al. 2003; Lemley and Sampat 2010, 2012). We leverage the interaction of this across-examiner heterogeneity with the quasi-random assignment of patent applications to examiners as a source of variation in which patent applications are granted patents. Past qualitative evidence—and new empirical evidence we document—supports the assertion that the assignment of patent applications to examiners is plausibly random conditional on some covariates (such as application year and technology type), suggesting that the leniency of the patent examiner to which a patent application is assigned can provide a valid instrument for whether the patent application is granted a patent.[2]

In contrast with what one would infer from a naïve comparison of follow-on innovation on patented and non-patented genes, both of our quasi-experimental approaches suggest that gene patents have not had quantitatively important effects on either follow-on scientific research or follow-on commercial investments. The estimates from our first quasi-experimental approach—comparing follow-on innovation across genes claimed in successful versus unsuccessful patent applications—document estimates which are economically small and meaningfully precise. While the estimates from our second quasi-experimental approach—using the leniency of the assigned patent examiner as an instrument for which patent applications are granted patents—are less precise, the fact that these two approaches generate similar conclusions provides additional confidence in our estimates.

When interpreting these empirical results through the lens of the well-developed theoretical literature on cumulative innovation (e.g., Green and Scotchmer 1995), several points are relevant to highlight. First, as best we can measure, very little of the follow-on innovation in this market is done by the gene

discussion in Lemley and Sampat 2008). Appendix A provides both a qualitative discussion and some quantitative analyses of this issue. For brevity, we colloquially refer to such applications as "rejected" and discuss the relevance of this point for our measurement and empirical approaches as relevant throughout the text.

[2]Importantly, both of our sources of quasi-experimental variation hold the disclosure of inventions constant: both accepted and rejected inventions are disclosed in the published patent applications that comprise our sample. This implies that our analysis corresponds to a partial equilibrium analysis (holding disclosure constant), whereas in general equilibrium firms may choose to protect their inventions with e.g. trade secrecy or other strategies rather than with patents if patent protection is not available.

patent holders—suggesting that follow-on innovation in this market will often require licensing agreements. Second, essentially all (96%) of the patent applications in our sample are assigned to for-profit firms, and in the vast majority of clinical trials (86%) both the patent assignee and the follow-on innovator are for-profit entities. Taken together, these facts suggest that cross-firm licensing contracts in this market seem to have operated at least somewhat efficiently. Consistent with this interpretation, there has been relatively little patent litigation observed in this market.

Interpreting these empirical results in the context of existing empirical evidence, perhaps most closely related to this paper is work by one of the authors—Williams (2013)—who found that a non-patent form of database protection held by the private firm Celera on their version of the sequenced human genome was associated with large declines in follow-on scientific research and commercial product development, on the order of 30%. Why did Celera's intellectual property cause declines in follow-on innovation, while gene patents did not? Theoretical models tend to analyze stylized characterizations of intellectual property rights in which Celera's intellectual property could reasonably be seen as practically identical to patent protection. However, Celera's intellectual property differed from the gene patents we analyze in one key dimension: while the sequenced genetic data in both the accepted and the rejected patent applications we analyze was disclosed in a way that enabled open access to the data for all prospective follow-on users, Celera's data was disclosed in a much more restrictive way. Importantly, disclosure is not specific to gene patents but rather is a general feature of patents, which are intended to disclose discoveries in a way that other researchers can build on them (Walsh et al. 2003a,b). As we discuss in more detail in Section 6, both the institutional details of our context and a growing academic literature suggest that this difference in disclosure may plausibly account for the different observed effects of Celera's intellectual property and gene patents on follow-on innovation outcomes (Aghion et al. 2008; Furman and Stern 2011; Galasso and Schankerman 2015; Murray et al. 2016).

From a policy perspective, one reason this disclosure feature of the patent system may be important is that it highlights a potential unintended consequence of a recent set of high-profile legal rulings on the case *Association for Molecular Pathology v. Myriad Genetics*.[3] The firm Myriad Genetics was granted patents on human genes correlated with risks of breast and ovarian cancer, and in June 2013 the US Supreme Court unanimously ruled to invalidate a subset of Myriad's gene patent claims, arguing that such patents *"would 'tie up'...[genes] and...inhibit future innovation."* That is, the Court argued that gene patents had sufficiently strong negative effects on follow-on innovation that genes should be ineligible for patent protection. While—consistent with the Court's view—there has been widespread concern that patents on human genes may hinder follow-on innovation, as argued by a 2006 National Academies report and by

[3] Appendix B provides some additional background on this case.

Caulfield et al. (2013) there was no empirical evidence available to either support or refute that assertion prior to this paper. From the disclosure perspective discussed above, underlying this ruling seems to be an assumption that if genes are not patented, they would be placed in the public domain. But at least in the case of Celera, when Celera's patent applications were rejected the firm instead chose to rely on an alternative form of intellectual property which—taken at face value—resulted in both lower private returns and lower social returns to Celera's research investments. This type of counterfactual is critical to assessing the potential welfare consequences of the US Supreme Court's ruling not just for human genes but also for other technologies which have recently been declared to be unpatentable.[4]

Methodologically, our two quasi-experimental approaches build on similar applications in labor economics and public finance. Our comparison of accepted and rejected patent applications builds on Bound (1989) and von Wachter et al.'s (2011) investigations of the disincentive effects of disability insurance on labor supply, and Aizer et al.'s (2016) investigation of the effects of cash transfers on mortality. Likewise, our approach of using examiner "leniency" as a source of variation in which patent applications are granted patents builds on past work investigating the effects of incarceration length using variation across judges (Kling 2006), the effects of foster care using variation across foster care case workers (Doyle 2007, 2008), and the disincentive effects of disability insurance on labor supply using variation across disability insurance examiners (Maestas et al. 2013). Especially given the relatively small number of patent law changes in countries like the US in recent years, these two new sources of quasi-experimental variation will likely provide valuable opportunities to investigate the effects of patents in a variety of other applications.[5]

Section 2 describes our data, and Section 3 documents some descriptive statistics on our data. Section 4 presents estimates from our first quasi-experimental approach, comparing follow-on innovation across genes claimed in successful and unsuccessful patent applications. Section 5 presents estimates from our second quasi-experimental approach, using the leniency of the assigned patent examiner as an instrumental variable for whether the patent application was granted a patent. Section 6 provides some interpretations of our empirical results, and Section 7 concludes.

[4]More specifically, this same line of argument has shaped at least three other recent US Supreme Court rulings which have moved to restrict the set of discoveries eligible for patent protection in other industries (Kesselheim et al. 2013). First, in *Bilski v. Kappos* the Court invalidated patent claims on an investment strategy, announcing it supported a "high enough bar" on patenting abstract ideas that it would not "put a chill on creative endeavor and dynamic change." Second, in *Mayo v. Prometheus*, the Court invalidated patent claims on methods of using genetic variation to guide pharmaceutical dosing, expressing concern that "patent law not inhibit further discovery by improperly tying up the future of laws of nature." Finally, in *Alice Corp v. CLS Bank* the Court invalidated patent claims on software based on similar arguments.

[5]Indeed, a number of papers have already started to apply these methodologies in other contexts. See, for example, Gaule (2015), Feng and Jaravel (2016), Farre-Mensa et al. (2017), and Kline et al. (2017).

2 Data

This section describes our data construction.[6] To fix ideas, we start by describing an example patent application—USPTO patent application 08/483,554—claiming intellectual property over the BRCA1 gene, and describe our measures of follow-on innovation in the context of that example (Section 2.1). We then describe in more detail how we construct our sample of USPTO patent applications claiming intellectual property over human genes (Section 2.2), and our gene-level measures of follow-on scientific research and product development (Section 2.3).

2.1 Example: USPTO Patent Application 08/483,554

On 7 June 1995, Myriad Genetics filed USPTO patent application number 08/483,554: *17Q-Linked Breast and Ovarian Cancer Susceptibility Gene*. This application was subsequently granted a patent on 5 May 1998 (US patent number 5,747,282), and would later become a focus of the US Supreme Court case *Association for Molecular Pathology v. Myriad Genetics*.

This patent claimed intellectual property over an isolated sequence of nucleotide bases (adenine, cytosine, guanine, and thymine), the sequence of which is listed explicitly in the patent (*SEQ ID NO:1: AGC TCG CTG...*). By comparing this sequence of nucleotide bases to the census of human gene sequences, we can uncover that this sequence corresponds to the BRCA1 gene.

The text of the Myriad patent describes how variation in the precise sequence of nucleotide bases in the BRCA1 gene can induce variation in an individual's risk of developing breast and ovarian cancers. For example, women with certain abnormal types of BRCA1 or BRCA2 genes have a 38 to 87 percent lifetime risk of developing breast cancer, relative to about 12 percent in the general population.[7]

Such links between genetic variation and diseases are referred to as genotype-phenotype links. In the case of the BRCA1 gene, scientific papers have investigated links between BRCA1 and breast cancer (181 publications), ovarian cancer (96 publications), and pancreatic cancer (3 publications).[8] Once scientific knowledge of a given genotype-phenotype link has been documented, this knowledge can be applied to develop medical technologies such as pharmaceutical treatments and gene-based diagnostic tests. In the case of the BRCA1 gene, 17 pharmaceutical clinical trials have been conducted that focus on mutations in the BRCA1 gene, and a BRCA1 genetic test is marketed by the firm Myriad Genetics.[9] The goal of

[6] Appendix C describes our data construction in more detail.

[7] These statistics are drawn from GeneReviews, published by the US National Institutes of Health (NIH).

[8] The data is drawn from the NIH Online Mendelian Inheritance in Man (OMIM) database, described below, and is accurate as of 27 July 2017.

[9] The data is drawn from the Citeline Pharmaprojects database (accurate as of 9 July 2012) and the NIH GeneTests.org database (accurate as of 18 September 2012), both described below.

our data construction is to trace these measures of follow-on scientific research and product development for each human gene, and to link the data to records of which human genes have been included in USPTO patent applications and granted patents.

2.2 Constructing USPTO Patent Application Sample

Our quasi-experimental approaches require constructing data on the census of published USPTO patent applications that claim intellectual property over human genes—both successful (granted patents) and unsuccessful (not granted patents). Traditionally, unsuccessful USPTO patent applications have not been made publicly available. However, as part of the American Inventors Protection Act of 1999, the vast majority of USPTO patent applications filed on or after 29 November 2000 are published in the public record—regardless of whether they are granted patents—at or before eighteen months after the filing date.[10]

From the census of USPTO patent applications filed on or after 29 November 2000 and published on or before 31 December 2013, we identify the subset of applications claiming intellectual property over genes. To do this, we follow the methodology proposed by Jensen and Murray (2005), which can be briefly summarized as follows.[11] Since the early 1990s DNA sequences have been listed in USPTO patent applications in a standard format, labeled with the text "SEQ ID NO" (sequence identification number).[12] This standardized format allows for DNA sequences to be cleanly extracted from the full text of USPTO published patent applications. Once extracted, standard bioinformatics methods can be used to compare these sequences against the census of human gene DNA sequences in order to annotate each sequence with standard gene identifiers that can in turn be linked to outside databases.[13]

We apply this Jensen and Murray (2005) methodology to construct two samples. First, we construct

[10] For more details, see `http://www.uspto.gov/web/offices/pac/mpep/s1120.html` and the discussion in Lemley and Sampat (2010). Most applications not published eighteen months after filing are instead published sixty months after filing. Some US patent applications opt out of publication: Graham and Hegde (2013) document that around 8% of US applications opted for pre-grant secrecy of patent applications. For the NBER patent technology field "drugs and medical," which includes the most common patent class in our sample, the share was 3.5%.

[11] While we focus on the Jensen and Murray (2005) definition of gene patents, there are many different types of DNA-related patent claims—e.g., protein-encoding sequences, expressed sequence tags (ESTs), single-nucleotide polymorphisms (SNPs), sequence-based claims, and method claims that pertain to specific genes or sequences; see Scherer (2002) and Holman (2012) for more discussion. As one point of comparison, nearly all of the patent applications in our sample are included in the *DNA Patent Database* (`http://dnapatents.georgetown.edu/SearchAlgorithm-Delphion-20030512.htm`), which is constructed using a completely different methodology; we are grateful to Robert Cook-Deegan and Mark Hakkarinen for sharing the patent application numbers included in the *DNA Patent Database*, which enabled this comparison. Our empirical strategies could of course be applied to any empirically implementable definition of gene patent applications, but we are not aware of data on any other alternative definitions.

[12] See `https://www.uspto.gov/web/offices/pac/mpep/s2422.html`. On disclosure of genetic innovations in patent applications more generally, see Berman and Schoenhard (2004).

[13] Following Jensen and Murray (2005), we focus attention on the subset of DNA sequences that are explicitly listed in the claims of the patent applications (excluding DNA sequences that are referenced in the text of the patent application—and hence listed in the SEQ ID NO format—but not explicitly claimed as intellectual property).

a "first stage sample" that aims to include the census of published USPTO patent applications that claim any non-human (e.g. mouse) DNA sequences in their patent claims (Bacon et al. 2006). Second, we construct a "human gene sample" that includes only the subset of published USPTO patent applications that claim human genes (Lee et al. 2007). This human gene sample is the focus of our analysis, given the focus in this paper on investigating how patents on human genes have affected follow-on innovation. However, this sample includes far fewer patent applications than the first stage sample: we have around 1,500 patent applications in the human gene sample, relative to around 14,000 patent applications in the first stage sample. This matters for our second quasi-experimental design, for which it is (as we will discuss more below) econometrically useful to estimate variation in the examiner patent grant propensities in a separate sample of patent applications. Hence, we use the first stage sample to estimate the examiner patent grant propensities, and apply those estimates to the examiners in our human gene sample.

It is worth noting that one limitation of our quasi-experimental approaches is that they can only be implemented over the time period when unsuccessful patent applications were published (that is, applications filed on or after 29 November 2000). Some criticisms of gene patents—such as Heller and Eisenberg (1998)—focus on the subset of gene patents covering expressed sequence tags (ESTs), which Rai (2012) argues were less commonly granted in later years.[14]

2.3 Measuring Follow-on Innovation

We collect data on three measures of gene-level follow-on innovation: scientific publications as a measure of scientific research effort; and two measures of product commercialization: gene-related pharmaceutical research, and gene-based diagnostic tests.

We collect data on the scientific publications related to each gene from the Online Mendelian Inheritance in Man (OMIM) database, which catalogs scientific papers that have documented evidence for links between genetic variation and phenotypes. Using this data, we construct a count of the number of scientific papers published related to each gene—across all phenotypes—in each year.

Because of the long time lags between basic drug discovery and the marketing of new drugs, new approvals of drugs that take advantage of sequenced genetic data are just barely starting to enter the market (Wade 2010). Given these time lags, rather than using drug approvals as a measure of pharmaceutical research, we instead focus on an intermediate measure of drug discovery—namely, drug compounds under development, as disclosed in clinical trials.[15] Specifically, we measure gene-related pharmaceutical clinical

[14]For the purposes of our descriptive analyses we can (and do) apply the Jensen and Murray (2005) methodology to identify granted USPTO patents that claim human genes but were filed prior to 29 November 2000. The data is also drawn from Lee et al. (2007), and includes USPTO patents granted through 2005.

[15]Data on clinical trial investments has also been used as a measure of research effort in prior work, starting with

trials using the Citeline Pharmaprojects database, a privately-managed competitive intelligence database that tracks drug compounds in clinical trials and—critically for our project—assigns gene identifiers to compounds related to genetic variation on specific genes. Using this data, we construct a count of the number of clinical trials related to each gene in each year.

Finally, we collect data on gene-based diagnostic tests from the GeneTests.org database, a genetic testing registry. Some gene-based tests provide individuals with information about disease risk, such as the BRCA tests related to risks of breast and ovarian cancers; other gene-based tests identify individuals as relatively more or less appropriate for a given medical treatment, such as a test which predicts heterogeneity in the side effects of the widely-prescribed blood thinner warfarin. Using the GeneTests.org data, we construct an indicator for whether a gene is included in any gene-based diagnostic test as of 2012. Unfortunately, this data is only available in a cross-section (not a panel).

A priori, the impact of patents on follow-on scientific research could differ from the impact of patents on product development. For example, many have argued that most patented inventions are made available to academic researchers on sufficiently favorable licensing terms that academics are able to continue their research (USPTO 2001). Hence, even if transaction costs hinder licensing agreements for commercial applications we could expect to see no impact of patents on measures of follow-on academic research such as scientific publications. Alternatively, patents may change researchers' incentives of whether to disclose the results of their research through academic publications (Moon 2011), in which case observed differences in scientific publications could be explained by differences in disclosure rather than by differences in the true amount of underlying scientific research. By contrast, we would not expect the product development outcomes we measure to be affected by such disclosure preferences given that the measures we observe are revealed in the natural course of firms commercializing and selling their technologies.

3 Descriptive Statistics

3.1 Summary Statistics

Table 1 presents patent application-level summary statistics on our first stage sample and our human gene sample. Our first stage sample includes 14,476 USPTO patent applications with DNA sequences listed in their claims, while our human gene sample includes 1,545 USPTO patent applications with human genes listed in their claims. As described in Section 2, our sample includes US patent applications filed starting in 2000, given that the requirement for unsuccessful patent applications to be published came into force for patent applications filed on or after 29 November 2000. The first stage sample ends with application

Acemoglu and Linn (2004) and Finkelstein (2004), and including more recently Galasso and Schankerman (2015).

year 2010. The human gene sample ends earlier, with application year 2005, because 2005 is the last year of data used as an input into the data construction done by Lee et al. (2007).[16]

In both samples, about 30 percent of the patent applications had been granted patents by 2010. This patent grant rate tabulated in Table 1 is a so-called simple grant rate that does not account for the fact that US patent applications can spawn closely related "new" applications, such as so-called continuations or divisionals. Carley et al. (2015) use internal USPTO data to calculate simple and "family" (including patents granted to continuations or divisionals) grant rates in the universe of new utility patent applications filed at the USPTO from 1996-2005. In their sample, 55.8% of applications were granted patents directly, and including patents granted to children increases the allowance rate to 71.2%. For the NBER patent technology field "drugs and medical," which includes the most common patent class in our sample, the progenitor allowance rate is 42.8% and the family allowance rate is 60.7%. This measurement issue is relevant to our second empirical strategy—which analyzes application-level data—but not to our first empirical strategy, which analyzes gene-level data, and hence captures all patent applications related to each gene. In practice, this issue does not appear to be quantitatively important in our sample: the family grant rates in our first stage and second stage samples are 32.5% and 27.7%, which are relatively close to the analogous simple grant rates tabulated in Table 1 of 30.4% and 25.6%. As would be expected given these small differences, we show in Appendix Table D.1 that if we implement our second empirical strategy using a family grant rate instead of a simple grant rate that our main conclusions are unchanged.

Appendix Table D.2 tabulates gene-level summary statistics on our follow-on innovation outcomes, for the sample of genes included in our human gene patent application sample (N=15,524). The counts of scientific publications per gene in calendar year 2011 and clinical trials per gene in calendar year 2011 are both right-skewed: the median gene has a value of zero for each of these variables, but the maximum values are 22 and 230, respectively. Around 12% of genes are used in a diagnostic test as of calendar year 2012.

3.2 Comparison of Patented and Non-Patented Human Genes

Our interest in this paper is in comparing follow-on innovation across patented and non-patented genes. In this sub-section, we investigate the selection process in order to analyze which genes are patented. On one hand, we may expect inventors to be more likely to both file for and be granted patents on technologies that are more valuable: inventors may be more willing to pay the time and monetary cost of filing patent applications for inventions that are more valuable, and patent applications claiming intellectual property

[16] Unfortunately, we do not know of a data source which applies the Lee et al. (2007) methodology to later years of data.

over more valuable inventions may also be more likely to "clear the bar" and be granted patents. On the other hand, the USPTO grants patents based on criteria that may not closely correspond with measures of scientific and commercial value (Merges 1988). Hence, it is an empirical question whether inventions with higher levels of scientific and commercial value are more likely to be patented.

There is a remarkable absence of empirical evidence on this question, largely due to a measurement challenge: in most markets, it is very difficult to measure the census of inventions, and to link those inventions to patent records in order to identify which inventions are patented. The only other paper we are aware of which has undertaken such an exercise is Moser (2012), who constructs a dataset of innovations exhibited at world's fairs between 1851 and 1915 and documents that "high quality" (award-winning) exhibits were more likely to be patented, relative to exhibits not receiving awards. Given this dearth of previous estimates, documenting evidence on the selection of inventions into patenting is itself of interest. In addition, if we do observe evidence of selection into patenting, that would imply that any measured differences in follow-on innovation across patented and non-patented genes may in part reflect the selection of which genes are included in patent applications and granted patents, as opposed to an effect of patents on follow-on innovation.

For this exercise, we start with the full sample of human genes (N=26,440). As measured in our data, approximately 30% of human genes have sequences that were explicitly claimed in granted US patents (N=7,975).[17] Figure 1 documents trends in follow-on innovation by year separately for genes that ever receive a patent (triangle-denoted solid blue series), and for genes that never receive a patent (circle-denoted dashed red series). In Figure 1(a) we plot the average log number of scientific publications by year in each year from 1970 to 2012.[18] In Figure 1(b) we plot the average log number of clinical trials by year in each year from 1995 to 2011.[19]

Because by construction our human gene patents are measured starting in the mid-1990s (when the SEQ ID NO notation was introduced), the cleanest test of selection into patenting is a comparison of pre-1990 follow-on innovation across (subsequently) patented and non-patented genes. While the clinical trials data only starts later (in 1995), the scientific publications data is available prior to 1990 and therefore can

[17] As a point of comparison, Jensen and Murray (2005) document that as of 2005, approximately 20 percent of human genes had sequences that were explicitly claimed in granted patents. Because our sample includes patent applications that were granted patents after 2005, we would expect our estimate to be larger.

[18] We focus on the average log number of scientific publications by year because even within a calendar year, the number of publications per human gene is quite right-skewed. The pattern of selection that we document is unchanged if we instead plot the share of genes with at least one scientific publication by year (Appendix Figure D.1 (a)), or the average number of scientific publications by year (Appendix Figure D.1(c)). Here and elsewhere, we add one to our outcome variables before logging them in order to include observations with no observed follow-on innovation.

[19] These years—1995 to 2011—are the only years for which the Pharmaprojects data is available. As with the scientific publications measure, we focus on the average log number of clinical trials by year because this variable is quite right-skewed. Again, the pattern of selection that we document is unchanged if we instead plot the share of genes with at least one clinical trial by year (Appendix Figure D.1(b)), or the average number of clinical trials by year (Appendix Figure D.1(d)).

be used for this comparison. Looking at the data series in Figure 1(a) from 1970 to 1990 provides clear evidence of positive selection: genes that will later receive patents were more scientifically valuable—based on this publications measure—prior to being patented. Moreover, even within the 1970 to 1990 period this positive selection appears not just in the levels of follow-on innovation, but also in the trends: genes that will later receive patents appear to have divergent trends in scientific publications relative to genes that will never receive patents, even before any patents are granted.

These patterns—a level difference, and a divergence in trends—also appear in the clinical trials data presented in Figure 1(b), although it is not possible to cleanly separate selection and "treatment" (that is, any causal effect of patents on subsequent research effort) because that data series starts in 1995. Likewise, if we tabulate the probability that genes are used in diagnostic tests for patented and non-patented genes, 13.6% of patented genes are used in diagnostic tests, compared to 6.2% of non-patented genes.

Taken together, these patterns suggest strong evidence of positive selection: genes that will later receive patents appear more scientifically and commercially valuable prior to being granted patents, relative to genes that will never receive patents. This evidence is important for three reasons. First, this evidence suggests that our measures of value (pre-patent filing scientific publications and clinical trials) are correlated with patenting activity, which supports the idea that these measures can provide a meaningful basis for assessing selection in our two quasi-experimental approaches. Second, this analysis provides novel evidence on the selection of technologies into patenting, in the spirit of Moser (2012). Third, this evidence implies that a simple comparison of follow-on innovation across patented and non-patented genes is unlikely to isolate an unbiased estimate of how gene patents affect follow-on innovation. While a naïve comparison that did not account for selection would conclude based on Figure 1 that patents encourage follow-on innovation, our two quasi-experimental approaches will suggest a different conclusion.

4 Comparison of Accepted and Rejected Patent Applications

Our first quasi-experimental source of variation investigates a simple idea, which is whether genes that were included in unsuccessful patent applications can serve as a valid comparison group for genes that were granted patents.

This approach has previously been applied to other research questions in labor economics and public finance. For a wide range of public programs, a key challenge in identifying the effect of the program is identifying a plausible counterfactual: what would recipients' lives have been like in the absence of the program? Dating back at least to the work of Bound (1989), researchers have proposed constructing this counterfactual by comparing accepted and rejected applicants. The validity of this approach of course

depends on the extent to which accepted and rejected applicants differ on observable and unobservable characteristics, and researchers have hence explored the validity of this approach by comparing accepted and rejected applicants based on observable characteristics fixed at the time of application. For example, Aizer et al.'s (2016) analysis of accepted and rejected applicants for the Mothers' Pension program documented balance tests comparing accepted and rejected applicants on various covariates fixed at the time of application, such as mother's age and family size. If accepted and rejected applicants are similar on observables, then under the assumption that they are also similar on unobservables rejected applicants can provide a plausible counterfactual for the outcomes of accepted applicants in the absence of the program.[20]

We apply an analogous framework to our dataset of gene patent applications, in which some applications are accepted and some are rejected. One nuance is that our outcomes of interest are at the gene level, not at the patent application level. If every patent application included only one gene, and each gene was never included in more than one patent application, this distinction would be irrelevant. But in practice, patent applications can include more than one gene, and a given gene can be included in more than one patent application. To address this divergence from the simplest thought experiment, we divide the sample of genes that are ever included in any patent application into two sub-samples: a sub-sample of genes that is included in at least one accepted patent application, and a sub-sample of genes that is included in at least one patent application but never in an accepted patent application.

Defined in this way, a comparison of follow-on innovation on genes included in accepted and rejected patent applications will be valid if—conditional on being included in a patent application—whether a gene is granted a patent is as good as random. A priori, it is not clear that this offers a valid comparison. The USPTO is responsible for assessing whether patent applications should be granted patents based on five criteria: patent-eligibility (35 U.S.C. §101), novelty (35 U.S.C. §102), non-obviousness (35 U.S.C. §103), usefulness (35 U.S.C. §101), and the text of the application satisfying the disclosure requirement (35 U.S.C. §112). These criteria suggest that patent grants are not quasi-randomly assigned across patent applications.[21] However, because any given patent application in our sample may claim intellectual property rights over more than one gene, and given that our interest is in how patent grants affect gene-level outcomes, what is required for this first quasi-experimental approach to be valid is that conditional on

[20] Note that Bound (1989) and others (including Aizer et al. 2016) argue that in their context they can make a slightly different assumption, which is that rejected applicants are better in terms of counterfactual outcomes than are accepted applicants. For example, in his context of disability insurance Bound (1989) argues that rejected applicants are healthier and more capable of work, and thus that their labor force participation should provide an upper bound for what work could be expected of accepted beneficiaries. In our view, this type of "signing the bias" argument does not have a natural analog in our empirical context.

[21] Although, to foreshadow our second empirical approach, note that if patent applications were randomly assigned to patent examiners and patent examiners—who varied in their leniency—completely determined grant outcomes, that would be one substantive example of why patent grants could be as good as randomly assigned across accepted and rejected applications.

being included in a patent application, whether a gene is granted a patent is as good as randomly assigned. That is, the relevant assumption is that genes that are included in patent applications that are granted patents are comparable to genes that are included in patent applications that are not granted patents. Analogous to the balance tests used in the previous literature to assess the validity of accepted/rejected designs, we will document a comparison of 'accepted' and 'rejected' genes based on covariates fixed at the time of application (in our case, gene-level scientific publications and gene-level clinical trial activity at the time of application). As we will see, 'accepted' and 'rejected' genes look quite similar based on these observables at the time the patent application is filed. Hence, we will conclude that 'rejected' genes can provide a plausible counterfactual for what follow-on innovation on the 'accepted' genes would have looked like in the absence of being granted patents. While this comparison is quite simple, we will see that the resulting estimates are similar to the estimates from our second quasi-experimental source of variation (the examiner leniency variation, presented in Section 5).

4.1 Graphical Analysis

For this exercise, we start with the sample of human genes included in at least one patent application in our USPTO patent applications sample (N=15,524; 59% of the full sample of 26,440 human genes). Of this sample, 4,858 genes are claimed in a patent application that is subsequently granted a patent (31%), relative to 10,666 genes that are never observed to be subsequently granted a patent (69%).[22] Figure 2(a) illustrates the time pattern of when the patented group receives its (first) patent grant over time. Over half of these genes have received a patent by 2005, and (by construction) all have received a patent by 2010.

Figures 2(b) and 2(c) document trends in follow-on innovation by year. As in Figure 1, we plot the average log number of scientific publications by year in each year from 1970 to 2012 (Figure 2(b)), and the average log number of clinical trials by year in each year from 1995 to 2011 (Figure 2(c)).[23] The solid blue triangle-denoted line represents genes claimed in at least one granted patent, and the dashed red circle-denoted line represents genes claimed in at least one patent application but never in a granted patent.

As a point of comparison, the dashed green square-denoted line represents genes never claimed in a patent application filed after 29 November 2000 (N=10,916; 41% of the full sample of 26,440 human genes).

[22] Note that this 4,858 figure is lower than the 7,975 figure in Section 3, because we here focus only on patents granted on patent applications filed after 29 November 2000 (the date when unsuccessful applications began to be published).

[23] Appendix Figure D.2 documents analogous figures if we instead plot the share of genes with at least one scientific publication or clinical trial by year (Appendix Figure D.2 Panels (a) and (b)) or the average number of scientific publications or clinical trials by year (Appendix Figure D.2 Panels (c) and (d)).

Comparing this group of genes to the other two groups of genes, we see clear evidence of selection into patent filing: genes included in successful and unsuccessful patent applications are much more valuable both scientifically (publications) and commercially (clinical trials) prior to the patent application filing compared to genes that are never claimed in a patent application. In terms of interpreting the results in Section 3.2, the data suggests that the major source of selection into which genes are patented is selection in which genes are included in patent applications (as opposed to which genes are granted patents, conditional on being included in patent applications).

The key comparison of interest in this section is across the first two data series: genes included in successful patent applications, and genes included in unsuccessful patent applications. The vertical line in calendar year 2001 denotes that because this figure focuses on patent applications filed in or after November 2000, all years prior to 2001 can be considered a "pre-period" and used to estimate the selection of genes into patenting based on pre-patent filing measures of follow-on innovation. Strikingly, we see little evidence of selection in pre-2001 levels or trends of our two follow-on innovation measures once we limit the sample to genes included in patent applications. For scientific publications (Figure 2(b)), the two groups follow each other quite closely from 1970 to 1990, and diverge slightly in trends from 1990 to 2000—with genes that will subsequently be included in unsuccessful patent applications having slightly more scientific publications.[24] For clinical trials (Figure 2(c)), the two groups follow each other quite closely in both levels and trends over all available pre-2001 years of data. Taken at face value, the similarity of these two groups in pre-2001 outcomes provides evidence for the validity of this empirical approach. A priori, one might have expected genes that were more scientifically or commercially valuable to have been more likely to receive patents. However, conditional on being included in a patent application, this appears not to be the case.

Looking at the post-2001 time period, we see that although these two groups of genes diverge (by construction) in whether they are claimed in granted patents (Figure 2(a)), we do not see any evidence of a divergence in follow-on innovation outcomes. That is, these figures suggest that gene patents have not had a quantitatively important effect on either follow-on scientific research or on follow-on commercialization.

4.2 Regression Analysis

We quantify the magnitudes of these patterns in a regression framework in Table 2. Because our scientific publication and clinical trial outcomes are quite skewed, a proportional model or binary outcome (measuring "any follow-on innovation") is more econometrically appropriate than modeling the outcome in levels.

[24] Note that this slight divergence is not apparent in the robustness check documented in Appendix Figure D.2(c), where we plot the average number of scientific publications by year.

We focus on the log of follow-on innovation and (separately) an indicator for any follow-on innovation.

Given the absence of strong visual evidence for a difference in follow-on innovation across patented and non-patented genes, our focus here is on what magnitudes of effects can be ruled out by our confidence intervals. Across these specifications, our 95% confidence intervals tend to reject declines or increases in follow-on innovation on the order of more than 5-15%. For brevity, we focus on interpreting the log coefficients. For our measures of follow-on scientific research (publications; Panel A of Table 2) and commercialization (clinical trials; Panel B of Table 2), the 95% confidence intervals can reject declines or increases of more than 2%. For our measure of diagnostic test availability (only measured as a binary indicator; Panel C of Table 2), we estimate that genes receiving patents had a 0.9 percentage point decrease in the likelihood of being included in a diagnostic test as of 2012 relative to genes included in patent applications but not granted patents (statistically significant at the 10% level). Our 95% confidence interval can reject declines of greater than 2 percentage points and reject increases of more than 0.2 percentage points. Relative to a mean of 12%, this confidence interval suggests that we can reject declines in this outcome of greater than 17%.[25]

5 Analyzing Examiner-Level Variation in Patent Grant Propensities

Our second source of quasi-experimental variation constructs an instrumental variables strategy for predicting which patent applications are granted patents. Our key idea is to build on previous research which has established that although patent examiners are charged with a uniform mandate, in practice examiners have a fair amount of discretion, and this discretion appears to translate into substantial variation in the decisions different examiners make on otherwise similar patent applications (Cockburn et al. 2003; Lichtman 2004; Lemley and Sampat 2010, 2012).[26] In the spirit of prior analyses such as Kling (2006), we leverage these patterns in order to use variation in the "leniency" of different patent examiners as a predictor of which patent applications are granted patents.

The exclusion restriction for this instrumental variables approach requires assuming that the examiner only affects follow-on innovation through the likelihood that a gene is patented. As we describe below, the institutional context suggests that the assignment of patent applications to USPTO patent examiners should be effectively random conditional on some covariates (such as application year and technology type).

[25] As a point of comparison, only 3% of genes never included in a patent application are included in a diagnostic test as of 2012.

[26] One of the individuals interviewed by Cockburn et al. (2003) described this variation informally by saying: *"there may be as many patent offices as there are patent examiners."* Similarly, a trade publication written by a former USPTO patent examiner and current patent agent (Wolinsky 2002) described this variation by saying: *"The successful prosecution of a patent application at the USPTO requires not only a novel invention and adequate prosecution skills, but a bit of luck...If you knew the allowance rate of your examiner, you could probably estimate your odds of getting your patent application allowed."*

While the exclusion restriction is inherently untestable, we will document empirically that—consistent with our qualitative description of the institutional context—genes assigned to 'lenient' and 'strict' examiners look similar on observable characteristics fixed at the time of patent application.[27]

To motivate our empirical specification, Section 5.1 provides some qualitative background on the key institutional features underlying our empirical strategy.[28]

5.1 Assignment of Patent Applications to Patent Examiners

A central USPTO office assigns application numbers to incoming patent applications, as well as patent class and subclass codes detailing the type of technology embodied in the application.[29] These class and subclass numbers determine which of approximately 300 so-called Art Units—specialized groups of examiners—will review the application.[30] Within an Art Unit, a supervisory patent examiner assigns the application to a patent examiner for review. While the patent application process up to this point is quite structured, from this point forward substantial discretion is left in the hands of individual examiners, who are responsible for determining which—if any—claims in the application are patentable.

Because no "standard" method for the within-Art Unit assignment of applications to examiners is uniformly applied in all Art Units, Lemley and Sampat (2012) conducted written interviews with roughly two dozen current and former USPTO examiners to inquire about the assignment process. While the results of these interviews suggested that there is not a single "standard" assignment procedure that is uniformly applied in all Art Units, these interviews revealed no evidence of deliberate selection or assignment of applications to examiners on the basis of characteristics of applications other than those observed in standard USPTO datasets (on which we can condition). For example, in some Art Units supervisors reported assigning applications to examiners based on the last digit of the application number; because application numbers are assigned sequentially in the central USPTO office, this assignment system—while not purposefully random—would be functionally equivalent to random assignment for the purposes of this study. In other Art Units, supervisors reported placing the applications on master dockets based on patent classes and subclasses, with examiners specializing in those classes (or subclasses) being automat-

[27] While conditional random assignment of applications to examiners assuages many potential concerns about this exclusion restriction, some additional issues remain. In particular, while we focus on variation in patent grant propensity, examiner heterogeneity may also manifest itself in other ways, such as the breadth of patent grants (in terms of the number or strength of allowed claims) and time lags in grant decisions (Cockburn et al. 2003).

[28] The discussion in Section 5.1 draws heavily on Cockburn et al. (2003), US General Accounting Office (2005), Lemley and Sampat (2012), and Frakes and Wasserman (2017). See Appendix A for more detail on the USPTO patent examination process.

[29] There are currently over 450 patent classes, and more than 150,000 subclasses; see `http://www.uspto.gov/patents/resources/classification/overview.pdf`.

[30] See `http://www.uspto.gov/patents/resources/classification/art/index.jsp`. For the current version of the class/subclass-to-Art Unit concordance, see `http://www.uspto.gov/patents/resources/classification/caau.pdf`. The main Art Units in our sample are from the 1600 group (Biotechnology and Organic Chemistry).

ically assigned the oldest application from the relevant pool when requesting a new application. Our key conclusion from this institutional context is that effective conditional random assignment of applications to examiners—within Art Unit and application year—is plausible. Consistent with this assumption, we will document that patent applications assigned to 'lenient' and 'strict' examiners look similar based on observable characteristics fixed at the time of patent application.

From a practical perspective, it is worth noting that informational barriers limit the extent to which we would expect patent applications to be systematically sorted across examiners in a way that would be problematic for our empirical specifications. Because of the limited attention given to patents prior to their assignment to a specific examiner, and the judgment required to determine the characteristics of a given invention, it seems plausible that informational barriers would impose real constraints on sorting (this argument has been made in more detail by Merges 1999).[31] In particular, the "patentability" of applications is difficult to assess *ex ante*, and there is no evidence that supervisory patent examiners attempt to do so before assigning applications to particular examiners.

5.2 Examiner Leniency Variation

To the best of our knowledge, Kling (2006) was the first paper to leverage this type of variation—in his case, using the random assignment of court cases to judges as an instrument for incarceration length. He adopted the jackknife instrumental variables (JIVE1) approach proposed by Angrist et al. (1999), which predicts the judge effect for each case based on data for all other cases.[32] However, JIVE estimators have been criticized for having undesirable finite sample properties (see, e.g., Davidson and MacKinnon 2006). In our case, a natural alternative is to adopt a two-sample two-stage least squares (TS2SLS) variant of Angrist and Krueger's (1992) two-sample instrumental variable estimator. Specifically, for each patent examiner in our sample, we observe the decisions that examiner makes on non-human gene patent applications, and we can use that separate sample to estimate variation in leniency across examiners.[33] Motivated by the institutional context, we condition on Art Unit-by-application year fixed effects, so that we capture an examiner's patent grant propensity relative to other examiners reviewing applications in that Art Unit in that application year.[34] We use the two-sample two-stage least squares standard error

[31] Merges (1999) also argues that although sorting of patent applications to examiners may be efficient, an additional barrier to such sorting is the strong "all patents are created equal" tradition at the USPTO, which cuts strongly against any mechanism for separating and sorting patents.

[32] Much of the subsequent literature (e.g. Doyle 2007) has approximated JIVE through an informal leave out mean approach.

[33] In practice, our results are similar if we use a leave-one-out-mean approach that computes examiner leniency on the same sample as our outcomes are measured; see Appendix Tables D.3 and D.4.

[34] A separate concern is that if we only observed a small number of applications per examiner, our estimate of the variation in leniency across examiners would be overstated. To address this concern, we limit the first stage sample to examiners that saw at least ten applications (Heckman 1981; Greene 2001). To match our conceptual thought experiment, we also limit the

correction provided by Inoue and Solon (2010), and cluster at the patent application level by extending Pacini and Windmeijer's (2016) generalization to the heteroskedastic case.[35]

5.3 First Stage Estimates

Figure 3 provides a visual representation of our first stage. In our first stage sample, we calculate the mean grant rate for each examiner, residualized by Art Unit-by-application year fixed effects, and relate this measure of examiner "leniency" to patent grant outcomes.[36] Visually, there is a strong relationship.

To quantify this relationship, we estimate the following equation for a patent application i examined by patent examiner j filed in year t assigned to Art Unit a in our first stage sample:

$$\mathbf{1}\left(\text{patent grant}\right)_{ijta} = \alpha + \beta \cdot Z_{ijta} + \Sigma_{ta}\mathbf{1}\left(\text{art unit}\right)_{ta} + \varepsilon_{ijta}$$

where the outcome variable $\mathbf{1}\left(\text{patent grant}\right)_{ijta}$ is an indicator variable equal to one if patent application i was granted a patent, Z_{ijta} is the examiner's mean non-human gene patent grant rate, and $\Sigma_{ta}\mathbf{1}\left(\text{art unit}\right)_{ta}$ are a set of Art Unit-by-application year fixed effects.

In our first stage sample, we estimate a β coefficient in this specification of 0.876, with a standard error of 0.037. This point estimate implies that a 10 percentage point increase in an examiner's average patent grant rate is associated with a 8.8 percentage point increase in the likelihood that a patent application is granted a patent.[37] The F-statistic on the examiner's grant rate is on the order of 500, well above the rule of thumb for weak instruments (Stock et al. 2002).

As a robustness check, in Appendix Table D.5 we replace Art Unit-by-year fixed effects in this specification with Art Unit-by-year-by-class-by-subclass fixed effects, on the subsample for which these finer fixed effects can be estimated. The point estimates from this more stringent specification are very similar to and not statistically distinguishable from the baseline point estimate described above, suggesting that at least in our context, variation in the measured leniency of different examiners is unlikely to be generated

sample to Art Unit-years with at least two examiners.

[35] Conceptually, we would prefer to cluster our standard errors at the gene level. However, only a very small share of the patent applications in our first stage sample can be matched to gene identifiers, making that adjustment infeasible. As a second best, we instead cluster at the patent application level in our two-sample two-stage least squares specifications, and cluster at the gene level in robustness checks where that is feasible, such as in the leave-one-out-mean approach in Appendix Tables D.3 and D.4.

[36] We will describe the overlaid plot of lighter yellow triangles in Section 5.4.

[37] Our patent grant outcome is measured as of 2010 and is censored for patent applications that are still in the process of being examined, but this censoring should be less of a concern for earlier cohorts of gene patent applications. The point estimates on a sub-sample of early cohorts of applications are very similar to our baseline point estimates (results not shown), suggesting that censoring appears to not substantively affect the magnitude of the estimated first stage coefficient. Note that this is likely because the Art Unit-by-application year fixed effects largely account for differences in the probability of patent grant that are mechanically related to time since application. Given this similarity, we retain all cohorts of gene patent applications to retain a larger sample size.

by the systematic sorting of patent applications by classes or subclasses.[38]

5.4 Investigating Selection

In order for examiner leniency to be a valid instrumental variable for the likelihood that a given gene patent application is granted a patent, it must satisfy the exclusion restriction: the instrument can only affect follow-on innovation outcomes through the likelihood that a gene is patented. The institutional details described in Section 5.1 suggest that the assignment of applications to examiners is plausibly random conditional on Art Unit-by-application year fixed effects, lending some a priori credibility to the exclusion restriction. In this section, we empirically assess whether this assumption is reasonable by investigating whether patent applications assigned to 'lenient' and 'strict' examiners look similar on observable characteristics fixed at the time of patent application.

Ideally, we would empirically assess selection using variables that are correlated with the 'patentability' of the application at the time of filing, so that we could test whether applications that appear more patentable tend to be assigned to more lenient examiners. As discussed by Lemley and Sampat (2012), it is difficult to identify variables that measure the 'patent-worthiness' of an invention.[39] A variety of metrics have been proposed as measures of the value of granted patents: forward citations (Trajtenberg 1990), patent renewal behavior (Pakes 1986; Schankerman and Pakes 1986; Bessen 2008), patent ownership reassignments (Serrano 2010), patent litigation (Harhoff et al. 2003), and excess stock return values (Kogan et al. 2017). For our purposes, these measures are not appropriate: we need measures of patent value that are defined for patent applications (not just for granted patents), and also want a measure that is fixed at the time of application (and hence unaffected by subsequent grant decisions). For these reasons, we focus on two value measures which fit these criteria: patent family size and claims count.[40]

Generally stated, a patent "family" is defined as a set of patent applications filed with different patenting authorities (e.g. US, Europe, Japan) that refer to the same invention. The key idea is that if there is a per-country cost of filing for a patent, firms will be more likely to file a patent application in multiple countries if they perceive the patent to have higher private value. Past work starting with Putnam (1996) has documented evidence that patent family size is correlated with other measures of patent value. We define patent family size as the number of unique countries in which the patent application was filed.

[38] Righi and Simcoe (2017) provide a set of formal tests for random assignment of patent applications to examiners, and show that random assignment is rejected in the full sample of patent applications, suggesting that sorting is relevant in at least some technological areas.

[39] In their paper, Lemley and Sampat (2012) show that two observable characteristics fixed at the time of application—the number of pages in the application and the patent family size—are not correlated with a measure of examiner experience (years of employment at the USPTO). That evidence provides some indirect support for our exclusion restriction.

[40] Unfortunately, backward citations-based measures cannot be constructed for patent applications, because published US patent applications do not list backward citations.

We use claims count as an alternative value measure that is fixed at the time of patent application, as proposed by Lanjouw and Schankerman (2001). The key idea underlying this measure is that patents list "claims" over specific pieces of intellectual property, and that patents with more claims may be more valuable.

For our purposes, there are two key empirical questions we want to investigate using these measures. First, do patent family size and/or claims count predict patent grant? While clearly imperfect metrics of patentability, these variables are predictive of patent grants: if we regress an indicator variable for patent grant on these two variables, the p-value from an F test of joint significance is <0.001. Second, is the predicted probability of patent grant—predicted as a function of family size and claims count—correlated with our examiner leniency instrument? If we regress the predicted probability of patent grant (predicted as a function of family size and claims count) on our examiner leniency instrument (residualized by Art Unit-by-application year fixed effects), we estimate a coefficient of 0.013 (standard error 0.003). This relationship is displayed non-parametrically in the plot of lighter yellow triangles in Figure 3: consistent with our economically small regression estimate, there is no visual relationship between the predicted probability of patent grant and our instrument.

Taken together, the analysis in this section provides indirect support of our exclusion restriction in the following sense: we find no evidence that applications which appear more likely to be patented based on covariates that are fixed at the time of filing are differentially assigned to more lenient examiners. Hence, the variation in grant rates across examiners appears to reflect differences in the decisions made on ex ante similar applications.

5.5 Instrumental Variables Estimates

Table 3 documents our instrumental variable estimates, relating patenting (instrumented by examiner leniency) to follow-on innovation outcomes.[41]

For our measures of follow-on scientific research (publications; Panel A of Table 3) and commercialization (clinical trials; Panel B of Table 3), the 95% confidence intervals on our log estimates can reject declines of more than 9%; the 95% confidence intervals on our binary versions of these outcomes ("any publication" or "any clinical trial") are much less precise. For our measure of diagnostic test availability (Panel C of Table 3), the 95% confidence interval suggests that—relative to a mean of around 9%—we can reject declines in this outcome of greater than 42% and reject increases of greater than 11%.

The estimates from our first quasi-experimental approach—comparing follow-on innovation across

[41] Appendix Table D.6 documents corresponding ordinary-least-squares (OLS) estimates for this specification, for comparison.

genes claimed in successful versus unsuccessful patent applications—are more precise than these estimates. Our confidence in interpreting those estimates as causal is strengthened by the fact that the examiner leniency instrument generates similar results, albeit which are much less precise. From an economic perspective, the estimates from our first approach can cleanly reject the effect sizes documented in the prior literature on how non-patent forms of intellectual property affect follow-on innovation (Williams 2013; Murray et al. 2016), whereas the estimates from our second approach can only sometimes reject these effect sizes.

6 Discussion

6.1 Connecting Theory to Data

A well-developed theoretical literature has generated predictions for how patents will affect follow-on innovation under different conditions. For example, if a patent holder can develop all possible follow-on inventions herself, then all socially desirable follow-on inventions will be developed. However, as stressed by Scotchmer (2004), the patent holder may not know all potential opportunities for follow-on innovation; in Scotchmer's language, ideas for follow-on innovations may be "scarce." The data suggests this is the relevant case in our context: as best we can measure, very little of the follow-on research in our sample is done by the patent assignee.[42] Tracing the identity of follow-on innovators is most feasible for our clinical trials data, which records the name of the organization in charge of the clinical trial.[43] We matched clinical trial organizations with assignee names in the patent application data, and found that among all observed clinical trials in the human gene sample these entities were the same in only 0.1% of cases.[44] This suggests the most relevant case in our setting is one in which follow-on innovations require cross-firm licensing agreements.

If ex ante licensing agreements can be signed prior to follow-on inventors sinking their R&D invest-

[42]Note that documenting this type of heterogeneity requires some relatively unusual data construction efforts because published patent applications are not required to list the assignee of the application, which is missing for about 50 percent of published applications. Based on discussions with individuals at the USPTO, we developed an approach which allows us to fill in missing assignee names in most but not all cases; details of this approach are included in Appendix C.

[43]Tracing the identity of follow-on innovators is much less feasible for our other two outcomes: in the case of scientific publications, publicly available data on scientific publications (in PubMed) only records the institutional affiliation of the first author (when the last would arguably be most relevant), and our diagnostic tests data does not provide information on the institution offering the test.

[44]Specifically, we matched using the Levenshtein distance divided by the average length of the two strings. Pairs with low scores and pairs with at least one word in common were manually checked for false negative matches. Because companies could conceivable operate clinical trials through subsidiaries, we purchased data from ReferenceUSA on the parent companies of each firm in our clinical trials and patent application data in order to classify firms that share the same parent company as belonging to the same "company family." In practice, our results in this section are essentially identical whether or not we account for company families using this data.

ments, all socially desirable follow-on inventions will still be developed.[45] However, if there are impediments to ex ante licensing—such as asymmetric information—then follow-on innovation will be lower than socially desirable due to the problem of how to divide the profit across firms (Scotchmer 1991; Green and Scotchmer 1995). This profit division problem could be exacerbated if transaction costs hinder cross-firm licensing agreements (Heller and Eisenberg 1998; Anand and Khanna 2000; Bessen 2004). One natural conjecture is that bargaining failures may be less common in cases where one or both parties is a non-profit entity such as a university. We investigated this possibility in our data by manually inspecting each clinical trial organization and patent application assignee in order to classify each as a for-profit or not-for-profit. In essentially all cases (96%), patent applications are assigned to for-profit firms. Moreover, in nearly all cases both the assignee and the clinical trial organization are for-profit entities (86% of our sample). Unfortunately, this means we lack sufficient variation in our sample to explore heterogeneity in the for-profit status of assignees or follow-on innovators.[46]

Interpreted through the lens of this theoretical literature, the fact that follow-on innovations in this market will often require cross-firm licensing agreements (given the low rates of follow-on innovation by patent holders themselves), together with the fact that we empirically find no evidence that patents induce economically meaningful reductions in follow-on innovation, suggests that licensing contracts in this market operated at least somewhat efficiently.

Importantly, this interpretation depends critically on the assumption that gene patents are sufficiently broad that the types of follow-on inventions we measure would require gene patent licenses. In our view, legal and policy writings on this topic clearly support this assumption. In a now-classic paper, Heller and Eisenberg (1998) documented that more than 100 issued US patents included the term 'adrenergic receptor' in the claim language, and pointed to this as an example of the complex biomedical patent landscape. Consistent with this concern, in a comment on the Heller-Eisenberg paper published in the same issue of *Science*, the then-USPTO biotechnology examination unit head John Doll argued that gene patents would be interpreted quite broadly by the USPTO (Doll 1998). For example, Doll wrote: "*...once a [gene] is patented, that patent extends to any use, even those that have not been disclosed in the patent.*" This interpretation strongly supports the assumption that the follow-on innovations we measure would require gene patent licenses.

[45] Also important to mention is the strand of research dating back at least to Kitch (1977), which has argued—in contrast with this Scotchmer-style literature—that patents may facilitate investment and technology transfers across firms (Arora 1995; Arora et al. 2001; Kieff 2001; Gans et al. 2002, 2008), increasing incentives for follow-on research and commercialization.

[46] Galasso and Schankerman (2015) also explore heterogeneity based on the size of the patent assignee. Again, in practice, this dimension of heterogeneity is not meaningful in our sample of human gene patents: 88% of the applications in our sample are assigned to one of four firms which would all be reasonably classified as small biotechnology firms over our period of study (Genentech, Corixa, Incyte, and Millenium).

However, one can find writings which take the opposite position. For example, in a second comment on the Heller-Eisenberg paper published in the same issue of *Science*, Seide and MacLeod (1998) argued that based on a cursory patent clearance review, at most only a small number of licenses might be required in the adrenergic receptor case. A similar argument has been made in the context of gene patents by Holman (2012): although a large number of human genes are claimed by US patents, he argues that a reading of the claims suggests that few licenses would be required for common types of follow-on innovation. An anecdote is perhaps helpful in illustrating a concrete example of why gene patent licenses may not always be needed. Pollack (2001) reported in the *New York Times* that pharmaceutical firm Bristol Myers abandoned research on more than fifty cancer-related proteins due to conflicts with gene patent holders. While interpreted by many as evidence of gene patents deterring follow-on innovation, in fact this example was reported as part of a licensing agreement Bristol Myers brokered with another firm (Athersys) for a method enabling the use of protein targets without infringing gene patents. That is, in this case, despite Bristol Myers reporting frictions in licensing markets for gene patents, the follow-on research they wished to pursue was able to occur.

One might think that the question of whether follow-on innovations require licenses could be resolved simply by reading the text of the claims of the gene patents in our sample. However, patent breadth is determined not only by the text of the patent, but also by courts' interpretations of these claims. Because very few gene patents have been litigated in court, we have little data on how broadly these patents would be interpreted in practice.

We want to stress two key take-aways from this discussion of patent breadth. First, from a theoretical perspective narrow patents should not deter follow-on innovation, so to the extent that this alternative interpretation is correct our analysis would be most useful in highlighting that documentation of biomedical patenting frequency (as in Jensen and Murray 2005) should not necessarily be interpreted as evidence of a problem; analyses such as the one in this paper are needed in order to investigate whether patents are sufficiently broad to actually be affecting real economic behavior and outcomes. Second, while these two interpretations are interesting and important to consider from an academic perspective, distinguishing between them is not required from a policy perspective: the fact that gene patents do not appear to have hindered follow-on innovation is sufficient to inform both the Nordhaus-style theoretical question of interest in this market, and the policy-relevant question at the basis of the recent US Supreme Court *AMP v. Myriad* ruling.

6.2 Interpreting Our Estimates: All Intellectual Property Is Not Alike

Our empirical estimates suggest that gene patents have not had quantitatively important effects on follow-on scientific research nor follow-on product development. This conclusion stands in contrast with a June 2013 US Supreme Court decision which unanimously ruled that human genes should not be patentable because such patents *"would 'tie up'...[genes] and...inhibit future innovation."* While the Court cited no empirical evidence on this point, and a broad range of commenters (e.g., National Academy of Sciences 2006; Caulfield et al. 2013) argued that there was essentially no empirical evidence available to either support or refute that assertion, we here give some context for why the previous literature may have been—perhaps erroneously—interpreted as supporting this view.

Perhaps most closely related to this paper is previous work by one of the authors—Williams (2013)—who analyzed a non-patent form of database protection held by the private firm Celera on their version of the sequenced human genome. Celera attempted (but largely failed) to obtain patents on its sequenced genetic data, most likely because Celera's sequenced genes alone were not sufficient to meet the USPTO's utility requirement in the absence of specific knowledge about the functions of those genes in the human body. When they were unable to obtain patents, Celera hired a well-known intellectual property lawyer and asked him to design an alternative form of intellectual property that would be the next best available alternative means of capturing a return on their investment in sequencing the human genome. This lawyer designed a contract-law based form of intellectual property that had several key features. First, Celera's data was "disclosed" in 2001 (Venter et al. 2001), in the sense that any individual could view data on the assembled genome through Celera's website or by obtaining a data DVD from the company. Second, viewing or obtaining Celera's data required agreeing not to redistribute the data. These restrictions on redistribution allowed Celera to grant academic researchers free access to Celera's data for non-commercial research, and to sell its data to larger institutions such as pharmaceutical companies. Although 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 paying between $7,500 and $15,000 for each lab given access to the data. Importantly, these restrictions on redistribution also meant that Celera's data was not disclosed in the main open-access databases (e.g. GenBank) used by researchers over this period. Third, any researcher wanting to use Celera's data for commercial purposes was required to negotiate a licensing agreement with Celera. Williams (2013) documents that this non-patent form of intellectual property on human genes was associated with large declines in follow-on scientific research and commercial product development—on the order of 30%—relative to genes that were sequenced at the same time by the publicly-funded Human Genome Project, which placed

all of their sequenced genes in public open-access databases such as GenBank.

Why did Celera's intellectual property cause declines in follow-on innovation, while gene patents did not?[47] Theoretical models tend to analyze stylized characterizations of intellectual property rights in which Celera's contract law-based form of intellectual property could reasonably be seen as practically identical to patent protection: both intended to provide an incentive for the discovery of an invention, and both required licensing agreements for any follow-on commercial applications developed by inventors outside of the firm. Seen from this perspective, the contrast in impacts on follow-on innovation between Celera's intellectual property and gene patents is perhaps surprising.

However, viewed through an alternative lens—what one might call the "disclosure" lens—this contrast is less surprising and perhaps even expected. The patent system requires inventors to disclose the claimed invention. Although theoretical models of intellectual property have tended to pay relatively little attention to the disclosure requirement, in our view this is a critical contrast between Celera's intellectual property and gene patents. For gene patent applications, the patent system's disclosure requirement induced explicit reporting of the claimed DNA sequence in a way that enabled open access to the sequence data underlying both accepted and rejected applications for all prospective follow-on users. In contrast, as detailed above Celera's data was disclosed in a much more restrictive way.

Both the institutional details of our context and a growing academic literature suggest that this difference in disclosure may plausibly account for the different observed effects of Celera's intellectual property and gene patents on follow-on innovation outcomes. Over the time period of our study, all DNA sequences claimed in patent applications submitted to the USPTO were disclosed in the main open-access databases (e.g. GenBank) used by researchers.[48] This is a concrete example illustrating the broader idea that patents, by nature of requiring disclosure of discoveries, generally retain open access to research materials, a point made by Walsh et al. (2003a,b). In contrast, the no-redistribution restriction on Celera's sequenced data meant that their data was not added to the key open-access databases that were widely used by researchers. A natural hypothesis is that this lack of open disclosure led to fewer scientific researchers analyzing Celera's data, and that that in turn led to fewer commercializable discoveries. Consistent with this hypothesis, both the theoretical work of Aghion et al. (2008) and other sources of empirical evidence (Furman and Stern 2011; Murray et al. 2016) suggest that limiting access to research materials may discourage follow-on research. Also consistent with this hypothesis, and with our empirical results, Galasso and Schankerman (2015) find no evidence that patent invalidations—which should induce

[47] Notably, some commenters such as Joseph Stiglitz argued that gene patents were likely to deter follow-on innovation on the basis of the Williams (2013) estimates; see Marshall (2013).

[48] As we discuss below, this was not true for rejected patent applications filed before 29 November 2000, but is true for all patent applications in our sample as all were filed after that date.

a change in licensing requirements but no change in disclosure—increase patent citations (their measure of follow-on innovation) in the technology field that includes most of our gene patents.[49]

Read together with the previous literature, our findings provide indirect support for the idea that patents effectively disclose information, at least in the context of gene patents (see Ouellette 2012 for a related discussion). More precisely, our findings support the idea that patent *applications*—both accepted and rejected, over the time period of our study—effectively disclose information. However, one limitation of our analysis is that we cannot speak to whether the effect of a patent application being accepted versus rejected may have been different for patent applications filed before 29 November 2000, when rejected applications were not required to be publicly disclosed. Under that earlier regime, patent applicants had the option—if their applications were rejected—of pursuing alternative forms of intellectual property protection on their inventions. In the case of Celera, when their patent applications were rejected and *not* required to be disclosed, the firm chose to rely on an alternative form of intellectual property, which—taken at face value—resulted in both lower private returns and lower social returns to their research investment. That is, the lower levels of follow-on innovation on Celera's genes suggest that both Celera and the social planner may have preferred to have granted Celera's gene patent applications rather than rejecting them.

From a policy perspective, these findings point to a potential unintended consequence of the US Supreme Court's rulings that human genes as well as several other types of technologies—such as business methods and diagnostic tests—are no longer eligible for patent protection. These rulings seem to be based on the idea that if these inventions are not patented, they will be placed in the public domain. However, the alternative to patent protection may instead be non-patent intellectual property that could generate worse outcomes than allowing patent protection. As a concrete example, in response to the *Mayo v. Prometheus* ruling which rendered many medical diagnostic methods unpatentable, the law firm Goodwin Proctor issued an "alert" arguing that the ruling called for "a renewed look at trade secret law for protection of diagnostics innovations as an alternative to patents."[50] This type of counterfactual is critical to assessing the potential welfare implications of the US Supreme Court's decisions on many types of technologies which have recently been declared to be unpatentable.

[49]Specifically, the most common patent class in our sample is 435 (Chemistry: molecular biology and microbiology), which is included in the NBER patent technology field "drugs and medical"; Galasso and Schankerman do not find evidence that patent invalidations increase follow-on citations in that subcategory.

[50]See http://www.goodwinlaw.com/publications/2012/03/lessons-from-mayo-v-prometheus-assessing-patentability-and-obtaining-patent-protection.

7 Conclusion

The contribution of this paper is to investigate whether patents on human genes have affected follow-on scientific research and product development. Using administrative data on successful and unsuccessful patent applications submitted to the US Patent and Trademark Office (USPTO), we link the exact gene sequences claimed in each patent application with data measuring gene-related scientific research (publications) and commercial investments (clinical development). Building on this data, we develop two new sources of quasi-experimental variation: first, a simple comparison of follow-on innovation across genes claimed in successful versus unsuccessful patent applications; and second, use of the "leniency" of the assigned patent examiner as an instrumental variable for whether the patent application was granted a patent. Both approaches suggest that—on average—gene patents have not had quantitatively important effects on follow-on innovation.

This empirical evidence speaks against two existing views. First, there has been widespread concern that patents on human genes may hinder follow-on innovation. For example, in the recent *Association for Molecular Pathology v. Myriad Genetics* case, the US Supreme Court invalidated patent claims on genomic DNA, arguing that such patents *"would 'tie up' the use of such tools and thereby inhibit future innovation premised upon them."* Our empirical estimates do not provide support for patents inducing economically meaningful reductions in follow-on innovation in the context of human genes. Second, dating back at least to the academic work of Kitch (1977), many have argued that patents on basic discoveries play an important role in facilitating subsequent investment and commercialization. Our empirical estimates do not provide support for patents spurring follow-on innovation in the context of human genes.

Taken together with the prior literature, our evidence suggests two conclusions. First, for the case of human genes, the traditional patent trade-off of ex ante incentives versus deadweight loss may be sufficient to analyze optimal patent policy design, because any effects of patents on follow-on innovation appear to be quantitatively small. Second, our evidence, together with the evidence from Williams (2013) on how a non-patent form of intellectual property on the human genome affected follow-on innovation, suggests a somewhat nuanced conclusion: while patent protection on human genes does not appear to have hindered follow-on innovation, an alternative non-patent form of intellectual property—which was used by a private firm after its gene patent applications were largely unsuccessful in obtaining patent grants—induced substantial declines in follow-on scientific research and product development. This pattern of evidence suggests that changes to patent policy must carefully consider what strategies firms will use to protect their discoveries in the absence of patents, and that an understanding of the relative costs and benefits of patent protection compared to those outside options is needed in order to evaluate the welfare effects of

patent policy changes.

References

Acemoglu, Daron and Joshua Linn, "Market size in innovation: Theory and evidence from the pharmaceutical industry," *Quarterly Journal of Economics*, 2004, *119* (3), 1049–1090.

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

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney, "The long-run impact of cash transfers to poor families," *American Economic Review*, 2016, *106* (4), 935–71.

Anand, Bharat and Tarun Khanna, "The structure of licensing contracts," *Journal of Industrial Economics*, 2000, *48* (1), 103–135.

Angrist, Joshua and Alan Krueger, "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples," *Journal of the American Statistical Association*, 1992, *87* (418), 328–336.

_ , Guido Imbens, and Alan Krueger, "Jackknife instrumental variables estimation," *Journal of Applied Econometrics*, 1999, *14* (1), 57–67.

Arora, Ashish, "Licensing tacit knowledge: Intellectual property rights and the market for know-how," *Economics of Innovation and New Technology*, 1995, *4* (1), 41–60.

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

Bacon, Neil, Doug Ashton, Richard Jefferson, and Marie Connett, "Biological sequences named and claimed in US patents and patent applications: CAMBIA Patent Lens OS4 Initiative," 2006. `http://www.patentlens.net` (last accessed 2 January 2012).

Berman, Richard and Amy Schoenhard, "The level of disclosure necessary for patent protection of genetic innovations," *Nature Biotechnology*, 2004, *22* (10), 1307–1308.

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

_ , "The value of U.S. patents by owner and patent characteristics," *Research Policy*, 2008, *37* (5), 932–945.

Bound, John, "The health and earnings of rejected disability insurance applicants," *American Economic Review*, 1989, *79* (3), 482–503.

Burk, Dan, "Are human genes patentable," *International Review of Intellectual Property and Competition Law*, 2013, *44* (7), 747–749.

Carley, Michael, Deepak Hegde, and Alan Marco, "What is the probability of receiving a U.S. patent?," *Yale Journal of Law and Technology*, 2015, *17* (1).

Caulfield, Timothy, Subhashini Chandrasekharan, Yann Joly, and Robert Cook-Deegan, "Harm, hype and evidence: ELSI research and policy guidance," *Genome Medicine*, 2013, *5* (21).

Cockburn, Iain, Samuel Kortum, and Scott Stern, "Are all patent examiners equal? Examiners, patent characteristics, and litigation outcomes," in Wesley Cohen and Stephen Merrill, eds., *Patents in the Knowledge-Based Economy*, National Academies Press, 2003.

Davidson, Russell and James MacKinnon, "The case against JIVE," *Journal of Applied Econometrics*, 2006, *21* (6), 827–833.

Doll, John, "The patenting of DNA," *Science*, 1998, *280* (5364), 689–690.

Doyle, Joseph, "Child protection and child outcomes: Measuring the effects of foster care," *American Economic Review*, 2007, *97* (5), 1583–1610.

_ , "Child protection and adult crime: Using investigator assignment to estimate causal effects of foster care," *Journal of Political Economy*, 2008, *116* (4), 746–770.

Farre-Mensa, Joan, Deepak Hegde, and Alexander Ljunqvist, "What is a patent worth? Evidence from the U.S. patent 'Lottery'," 2017. unpublished Harvard Business School mimeo.

Feng, Josh and Xavier Jaravel, "Who feeds the trolls? Patent trolls and the patent examination process," 2016. unpublished Harvard mimeo.

Finkelstein, Amy, "Static and dynamic effects of health policy," *Quarterly Journal of Economics*, 2004, *119* (2), 527–567.

Frakes, Michael and Melissa Wasserman, "Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from micro-level application data," *Review of Economics and Statistics*, 2017, *99* (3), 550–563.

Furman, Jeffrey and Scott Stern, "Climbing atop the shoulders of giants: The impact of institutions on cumulative research," *American Economic Review*, 2011, *101* (5), 1933–1963.

Galasso, Alberto and Mark Schankerman, "Patents and cumulative innovation: Causal evidence from the courts," *Quarterly Journal of Economics*, 2015, *130* (1), 317–369.

Gans, Joshua, David Hsu, and Scott Stern, "When does start-up innovation spur the gale of creative destruction?," *RAND Journal of Economics*, 2002, *33* (4), 571–586.

_ , _ , **and** _ , "The impact of uncertain intellectual property rights on the market for ideas: Evidence from patent grant delays," *Management Science*, 2008, *54* (5), 982–997.

Gaule, Patrick, "Patents and the success of venture-capital backed startups: Using examiner assignment to estimate causal effects," 2015. unpublished CERGE-EI mimeo.

Golden, John and William Sage, "Are human genes patentable? The Supreme Court says yes and no," *Health Affairs*, 2013, *32* (8), 1343–1345.

Graham, Stuart and Deepak Hegde, "Do inventors value secrecy in patenting? Evidence from the American Inventor's Protection Act of 1999," 2013. unpublished USPTO mimeo.

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

Greene, William, "Estimating econometric models with fixed effects," 2001. unpublished mimeo.

Hall, Bronwyn, Adam Jaffe, and Manuel Trajtenberg, "The NBER U.S. patent citations data file: Lessons, insights, and methodological tools," 2001. NBER working paper.

Harhoff, Dietmar, Frederic Scherer, and Katrin Vopel, "Citations, family size, opposition, and the value of patent rights," *Research Policy*, 2003, *32* (8), 1343–1363.

Harrison, Charlotte, "Isolated DNA patent ban creates muddy waters for biomarkers and natural products," *Nature Reviews Drug Discovery*, 2013, *12*, 570–571.

Heckman, James, "The incidental parameter problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process," in Charles Manski and Daniel McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, 1981.

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

Holman, Christopher, "Debunking the myth that whole-genome sequencing infringes thousands of gene patents," *Nature Biotechnology*, 2012, *30* (3), 240–244.

Inoue, Atsushi and Gary Solon, "Two-sample instrumental variables estimators," *Review of Economics and Statistics*, 2010, *92* (3), 557–561.

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

Kesselheim, Aaron, Robert Cook-Deegan, David Winickoff, and Michelle Mello, "Gene patenting - The Supreme Court finally speaks," *New England Journal of Medicine*, 2013, *369* (9), 869–875.

Kieff, Scott, "Property rights and property rules for commercializing inventions," *Minnesota Law Review*, 2001, *85*, 697–754.

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

Kline, Patrick, Neviana Petkova, Heidi Williams, and Owen Zidar, "Who profts from patents? Rent-sharing at innovative firms," 2017.

Kling, Jeffrey, "Incarceration length, employment, and earnings," *American Economic Review*, 2006, *96* (3), 863–876.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru, and Noah Stoffman, "Technological innovation, resource allocation, and growth," *Quarterly Journal of Economics*, 2017, *132* (2), 665–712.

Lanjouw, Jean and Mark Schankerman, "Characteristics of patent litigation: A window on competition," *RAND Journal of Economics*, 2001, *32* (1), 129–151.

Lee, Byungwook, Taehyung Kim, Seon-Kyu Kim, Kwang H. Lee, and Doheon Lee, "Patome: A database server for biological sequence annotation and analysis in issued patents and published patent applications," *Nucleic Acids Research*, 2007, *35* (Database issue), D47–D50.

Lemley, Mark and Bhaven Sampat, "Is the patent office a rubber stamp?," *Emory Law Journal*, 2008, *58*, 181–203.

_ **and** _ , "Examining patent examination," *Stanford Technology Law Review*, 2010, (4), 1–11.

_ **and** _ , "Examiner characteristics and patent office outcomes," *Review of Economics and Statistics*, 2012, *94*, 817–827.

Lichtman, Doug, "Rethinking prosecution history estoppel," *University of Chicago Law Review*, 2004, *71* (1), 151–182.

Maestas, Nicole, Kathleen Mullen, and Alexander Strand, "Does disability insurance receipt discourage work? Using examiner assignment to estimate causal effects of SSDI receipt," *American Economic Review*, 2013, *103* (5), 1797–1829.

Marshall, Eliot, "Lock up the genome, lock down research," *Science*, 2013, *342* (6154), 72–73.

Merges, Robert, "Commercial success and patent standards: Economic perspectives on innovation," *California Law Review*, 1988, *76*, 803–876.

_ , "As many as six impossible patents before breakfast: Property rights for business concepts and patent system reform," *Berkeley Technology Law Journal*, 1999, *14*, 577–615.

Moon, Seongwuk, "How does the management of research impact the disclosure of knowledge? Evidence from scientific publications and patenting behavior," *Economics of Innovation and New Technology*, 2011, *20* (1), 1–32.

Moser, Petra, "Innovation without patents: Evidence from World's Fairs," *Journal of Law and Economics*, 2012, *55* (1), 43–74.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern, "Of mice and academics: Examining the effect of openness on innovation," *American Economic Journal: Economic Policy*, 2016, *8* (1), 212–52.

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

Nordhaus, William, *Invention, Growth, and Welfare: A Theoretical Treatment of Technological Change*, MIT Press, 1969.

Ouellette, Lisa Larrimore, "Do patents disclose useful information?," *Harvard Journal of Law and Technology*, 2012, *25* (2), 546–607.

Pacini, David and Frank Windmeijer, "Robust inference for the two-sample 2SLS estimator," *Economic Letters*, 2016, *146*, 50–54.

Pakes, Ariel, "Patents as options: Some estimates of the value of holding European patent stocks," *Econometrica*, 1986, *54* (4), 755–784.

Pollack, Andrew, "Bristol-Myers and Athersys make deal on gene patents," *New York Times*, 2001, *8 January*.

Putnam, Jonathan, "The value of international patent protection," 1996. Yale PhD dissertation.

Rai, Arti, "Patent validity across the executive branch: Ex ante foundations for policy development," *Duke Law Journal*, 2012, *61*, 1237–1281.

_ **and Robert Cook-Deegan**, "Moving beyond 'isolated' gene patents," *Science*, 2013, *341*, 137–138.

Righi, Cesare and Timothy Simcoe, "Patent Examiner Specialization," 2017. NBER working paper.

Schankerman, Mark and Ariel Pakes, "Estimates of the value of patent rights in European countries during the post-1950 period," *Economic Journal*, 1986, *96* (384), 1052–1076.

Scherer, Frederic, "The economics of human gene patents," *Academic Medicine*, 2002, *77* (12), 1348–1367.

Scotchmer, Suzanne, "Standing on the shoulders of giants: Cumulative research and the patent law," *Journal of Economic Perspectives*, 1991, *5* (1), 29–41.

_ , *Innovation and Incentives*, MIT Press, 2004.

Seide, Rochelle and Janet MacLeod, "Comment on Heller and Eisenberg," 1998. ScienceOnline: `http://www.sciencemag.org/feature/data/980465/seide.dtl`.

Serrano, Carlos, "The dynamics of the transfer and renewal of patents," *RAND Journal of Economics*, 2010, *41* (4), 686–708.

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

Stock, James, Jonathan Wright, and Motohiro Yogo, "A survey of weak instruments and weak identification in generalized method of moments," *Journal of Business and Economic Statistics*, 2002, *20* (4), 518–529.

Trajtenberg, Manuel, *Economic Analysis of Product Innovation: The Case of CT Scanners*, Harvard University Press, 1990.

United States Supreme Court, "Association for Molecular Pathology et al. v. Myriad Genetics Inc. et al.," 2013. 12-398.

US General Accounting Office (GAO), "Intellectual property: USPTO has made progress in hiring examiners, but challenges to retention remain," 2005.

USPTO, "Utility examination guidelines," *Federal Register*, 2001, *66* (4), 1092–1099.

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

von Wachter, Till, Jae Song, and Joyce Manchester, "Trends in employment and earnings of allowed and rejected applicants to the Social Security Disability Insurance program," *American Economic Review*, 2011, *101* (7), 3308–3329.

Wade, Nicholas, "A decade later, genetic map yields few new cures," *New York Times*, 2010, *12 June*.

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

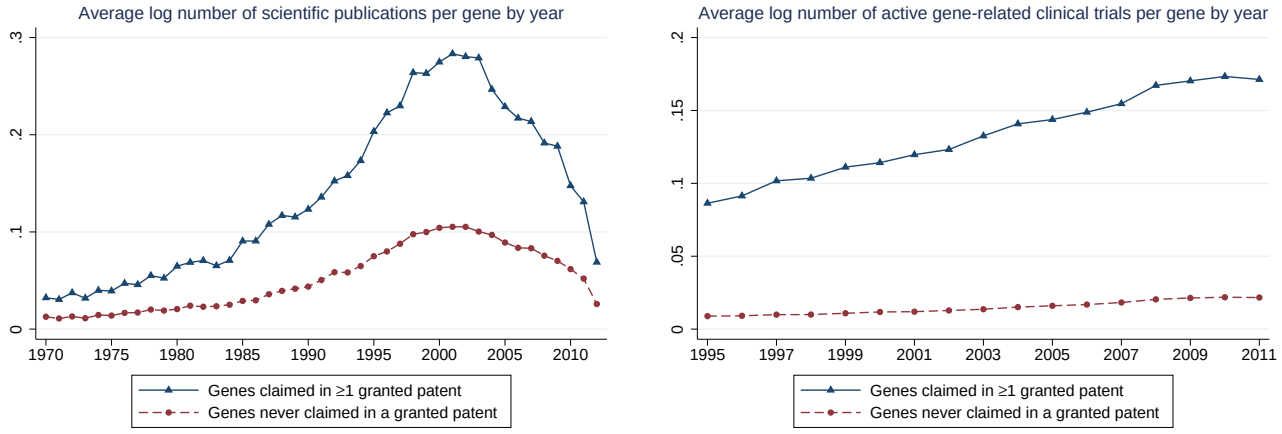
_ , _ , **and** _ , "Working through the patent problem," *Science*, 2003, *299* (5609), 1021.

Williams, Heidi, "Intellectual property rights and innovation: Evidence from the human genome," *Journal of Political Economy*, 2013, *121* (1), 1–27.

_ , "How do patents affect research investments?," *Annual Review of Economics*, 2017, *9*, 441–469.

Wolinsky, Scott, "An inside look at the patent examination process," *The Metropolitan Corporate Counsel*, 2002, *10* (9), 18.

Figure 1: **Follow-on Innovation on Patented and Non-Patented Human Genes**

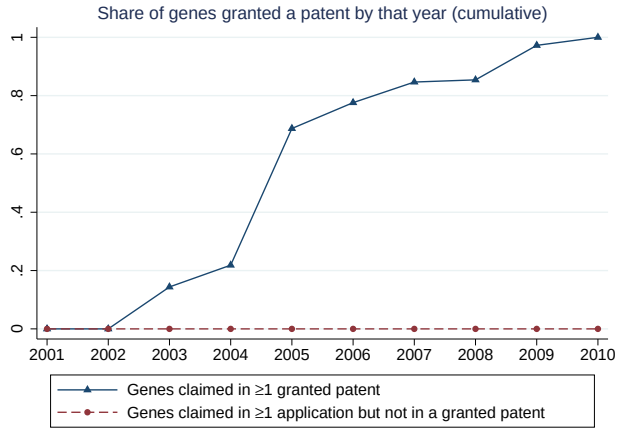


(a) Gene-Level Scientific Publications

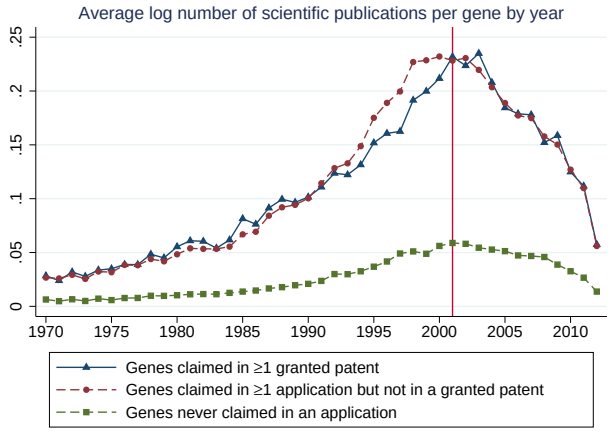
(b) Gene-Level Clinical Trials

Notes: This figure plots trends in follow-on innovation by year separately for genes that ever receive a patent and for genes that never receive a patent. The figure is constructed from gene-level data. Panel (a) uses gene-level scientific publications as a measure of follow-on innovation, and plots the average log number of scientific publications by year in each year from 1970 to 2012. Panel (b) uses gene-level clinical trials as a measure of follow-on innovation, and plots the average log number of clinical trials by year in each year from 1995 to 2011.

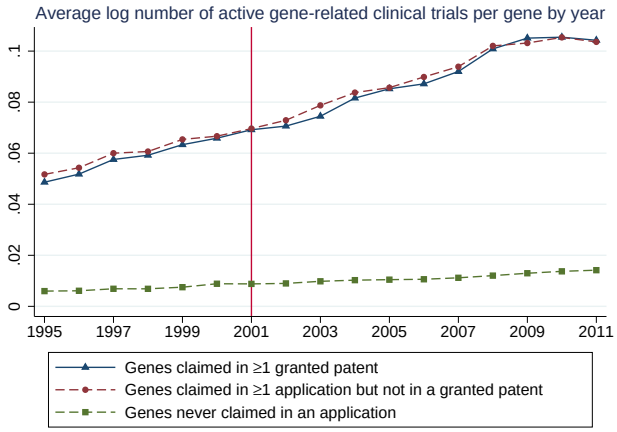
Figure 2: **Patents and Follow-on Innovation on Human Genes Claimed in Accepted/Rejected Patent Applications**



(a) Share of Genes Receiving a Patent Grant



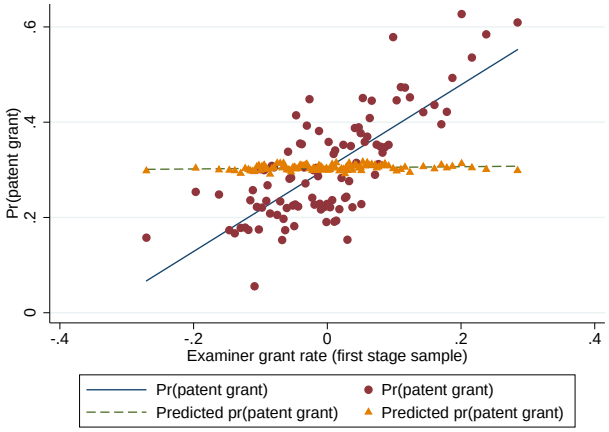
(b) Gene-Level Scientific Publications



(c) Gene-Level Clinical Trials

Notes: This figure plots trends in patenting and follow-on innovation by year separately for three groups of genes: genes claimed in at least one granted patent; genes claimed in at least one patent application but never in a granted patent; and (in Panels (b) and (c)) genes never claimed in a patent application. The figure is constructed from gene-level data. Panel (a) documents the share of genes receiving a patent grant by year; by construction, this is zero for the circle-denoted red dashed line in all years and reaches one for the triangle-denoted blue line in 2010; the intermediate years simply illustrate the time path of patent grants between 2001 and 2010 for the triangle-denoted blue line. Panel (b) uses gene-level scientific publications as a measure of follow-on innovation and plots the average log number of scientific publications by year in each year from 1970 to 2012. Panel (c) uses gene-level clinical trials as a measure of follow-on innovation and plots the average log number of clinical trials by year in each year from 1995 to 2011. The vertical line in the calendar year 2001 in Panels (b) and (c) denotes that, because this figure focuses on patents that were filed in or after November 2000, all years prior to 2001 can be considered a pre-period and used to estimate the selection of genes into patenting based on pre-patent filing measures of scientific research (publications) and commercialization (clinical trials).

Figure 3: **Probability of Patent Grant by Examiner Leniency**



Notes: The figure relates our examiner leniency measure, residualized by Art Unit-by-application year fixed effects, to two variables: (1) the patent grant rate and (2) the predicted patent grant rate, where we predict patent grant as a function of our two measures of patent value fixed at the time of application (patent family size and claims count). All measures are constructed in our first stage sample (N=14,476).

Table 1: **Patent Application-Level Summary Statistics**

	Mean	Standard deviation	Minimum	Maximum	Number of observations
Panel A: First stage sample					
Application year	2005	2.715	2000	2010	14,476
0/1, patent granted as of 2010	0.3043	0.4601	0	1	14,476
Panel B: Human gene sample					
Application year	2002	0.692	2000	2005	1,545
0/1, patent granted as of 2010	0.2557	0.4364	0	1	1,545

Notes: This table shows summary statistics for our patent application-level data in each of our two samples: Panel A for the first stage sample of patent applications, and Panel B for the human gene sample of patent applications.

Table 2: **Patents and Follow-on Innovation on Human Genes Claimed in Accepted/Rejected Patent Applications: Regression Estimates**

	Log of follow-on innovation in 2011/2012 (1)	Any follow-on innovation in 2011/2012 (2)
Panel A: Scientific publications		
Patent granted	0.0019	-0.0014
	(0.0060)	(0.0054)
Mean of dependent variable	0.1104	0.1094
Number of observations	15,524	15,524
Panel B: Clinical trials		
Patent granted	0.0006	-0.0015
	(0.0080)	(0.0043)
Mean of dependent variable	0.1038	0.0659
Number of observations	15,524	15,524
Panel C: Diagnostic test		
Patent granted	-	-0.0092
	-	(0.0056)
Mean of dependent variable	-	0.1199
Number of observations	-	15,524

Notes: This table estimates differences in follow-on innovation on genes claimed in at least one granted patent relative to genes claimed in at least one patent application but never in a granted patent. The sample for these regressions is constructed from gene-level data, and includes genes claimed in at least one patent application in our USPTO human gene patent application sample (N=15,524). Each coefficient is from a separate regression. Estimates are from ordinary-least-squares models. Heteroskedasticity robust standard errors.

Table 3: **Patents and Follow-on Innovation on Human Genes by Examiner Leniency: Instrumental Variables Estimates**

	Log of follow-on innovation in 2011/2012 (1)	Any follow-on innovation in 2011/2012 (2)
Panel A: Scientific publications		
Patent granted (instrumented)	-0.0230	-0.0187
	(0.0102)	(0.0089)
Mean of dependent variable	0.0798	0.0888
Number of observations	293,652	293,652
Panel B: Clinical trials		
Patent granted (instrumented)	-0.0488	-0.0293
	(0.0209)	(0.0118)
Mean of dependent variable	0.0690	0.0500
Number of observations	293,652	293,652
Panel C: Diagnostic test		
Patent granted (instrumented)	-	-0.0141
	-	(0.0123)
Mean of dependent variable	-	0.0918
Number of observations	-	293,652

Notes: This table presents instrumental variable estimates, relating follow-on innovation to whether a patent application was granted a patent, instrumented by our examiner leniency instrument. The sample for these regressions is constructed from application-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=293,652). All regressions include Art Unit-by-application year fixed effects. Each coefficient is from a separate regression. Standard errors are clustered at the patent application level (Inoue and Solon 2010; Pacini and Windmeijer 2016).

A Appendix: Background on the USPTO Patent Examination Process (for Online Publication)

In this appendix, we describe the USPTO patent examination process in more detail.[51]

A.1 Overview of the USPTO Patent Examination Process

The USPTO is responsible for determining whether inventions claimed in patent applications qualify for patentability. The uniform mandate for patentability is that inventions are patent-eligible (35 U.S.C. §101), novel (35 U.S.C. §102), non-obvious (35 U.S.C. §103), useful (35 U.S.C. §101), and the text of the application satisfies the disclosure requirement (35 U.S.C. §112).

Patent applications include a written description of the invention (the "specification"), declarations of what the application covers ("claims"), and a description of so-called prior art—ideas embodied in prior patents, prior publications, and other sources—that is known to the inventor and relevant to patentability. Once a patent application is received, as long as it satisfies a series of pre-examination formalities the Office of Patent Application Processing will assign it an application number, as well as a patent class and subclass.[52] These classifications, in turn, determine—based on a concordance between classes/subclasses and Art Units—the Art Unit to which the application is assigned, where Art Units are specialized groups of patent examiners that work on related subject matter.[53] Once assigned to an Art Unit, a supervisory patent examiner (SPE) then refines the patent classification if it is incorrect. In some cases, this means the application needs to be re-assigned to another Art Unit, though that is thought to be rare. The SPE then assigns the application to a patent examiner for review (via a process described in more detail below).

A.2 Within-Art Unit Assignment of Patent Applications to Patent Examiners

The process outlined above clarifies that the assignment of patent examiners to applications is a function of at least two factors: first, the Art Unit to which the application is assigned; and second, the year the application is filed, given that the group of examiners in an Art Unit will vary over time. In this section, we discuss how—within an Art Unit in a given application year—patent applications are assigned to patent examiners.

The USPTO does not publish rules regarding the assignment of applications within Art Units to particular examiners. Given this absence of formal written rules, Lemley and Sampat (2012) conducted written interviews with roughly two dozen current and former patent examiners and supervisory patent examiners to inquire about the assignment process. While the results of these interviews suggested that there is not a single "standard" assignment procedure that is uniformly applied in all Art Units, these interviews revealed no evidence of deliberate selection or assignment of applications to examiners on the basis of characteristics of applications other than those observed in standard USPTO datasets. In some Art Units, supervisors reported assigning applications to examiners based on the last digit of the application number; because application numbers are assigned sequentially in the central USPTO office, this assignment system—while not purposefully random—would be functionally equivalent to random

[51] The discussion in this appendix draws heavily on Cockburn et al. (2003), Lemley and Sampat (2012), and US General Accounting Office (2005), among other sources referenced in the text.

[52] There are currently over 450 patent classes; the most common class in our sample is 435 (Chemistry: molecular biology and microbiology). There are currently more than 150,000 subclasses. For more details, see `http://www.uspto.gov/patents/resources/classification/overview.pdf`.

[53] There are over 300 Art Units; see `http://www.uspto.gov/patents/resources/classification/art/index.jsp`. For the current version of the class/subclass-to-Art Unit concordance, see `http://www.uspto.gov/patents/resources/classification/caau.pdf`. The main Art Units in our sample are from the 1600 group (Biotechnology and Organic Chemistry).

assignment for the purposes of this study. In other Art Units, supervisors reported placing the applications on master dockets based on patent classes and subclasses, with examiners specializing in those classes (or subclasses) being automatically assigned the oldest application from the relevant pool when requesting a new application. Consistent with what we would expect given these types of assignment mechanisms, Lemley and Sampat (2012) present empirical evidence that observable characteristics of patent applications are uncorrelated with characteristics such as the experience of the examiner to which the application was assigned. Unfortunately, we do not have information on the specific set of assignment processes used by the Art Units most common in our sample over the relevant time period.[54] In the absence of such information, we rely on the interviews in Lemley and Sampat (2012) as a guide to designing our empirical specifications.

A.3 Overview of the Patent Prosecution Process

While the patent application process described above is quite structured, from this point forward substantial discretion is left in the hands of individual examiners.[55] An initial decision on whether a given patent application meets the standards for patentability is made by the assigned examiner. If the examiner issues a so-called "initial allowance" of the application, the inventor can be granted a patent. In most cases, the examiner's initial decision is instead a so-called "non-final rejection." However, in practice patent applications cannot be rejected by the USPTO, only abandoned by applicants (Lemley and Sampat 2008). Hence a rejection is essentially an invitation for the applicant to submit a revised patent application that, for example, eliminates one or more claims or changes the text of some claims to be narrower in scope. For non-final rejections, applicants have a fixed length of time (usually six months) during which to revise the application. After receiving the applicant's response, the examiner can then allow the application, negotiate minor changes, or send a second rejection.

The patent prosecution process—a phrase used to refer to the interaction between the USPTO and the patent applicant or her representative (such as a lawyer)—can involve several rounds of "rejection" and revision, and in this sense can best be conceptualized as an iterative process between the applicant and the examiner, rather than as a one-time decision by the examiner. Applicants presumably choose between "revising and resubmitting" or abandoning a rejected patent application by weighing the relevant costs and benefits: if a revision that accommodates the examiner's criticisms would result in a patent that—even if granted—would be too narrow to provide much economic value to the applicant, we would expect the application to be abandoned by the applicant.

A.4 Descriptive Statistics on the Patent Prosecution Process

While hopefully the conceptual description of the patent prosecution process as described above is clear, measuring this process in practice is quite complicated. The USPTO PAIR "transactions" data lists a record of every correspondence between applicants (or their lawyer, on their behalf) and the USPTO, but the data is provided in raw form that does not naturally correspond to economically meaningful events such as initial decisions, final rejections, applicant responses, etc. To attempt to shed some light on the patent prosecution process, we here document some descriptive statistics summarizing our best effort to quantify the prosecution process using this data.

A.4.1 Descriptive Statistics on Latest Stage Reached in Patent Prosecution Process

In this sub-section, we develop two sets of descriptive statistics: first, statistics on the furthest stage that a patent application progressed to in the prosecution process; and second, statistics on the final status for

[54] We have tried, unsuccessfully, to track down individuals who were supervisory patent examiners in the Art Units most common in our sample over the relevant time period.

[55] The description of the patent prosecution process in this section draws in part from Williams (2017).

each patent application as of the end of our data. Final statuses are measured as of the end of our PAIR transactions data (26 January 2015). As an input to each, we first generate a categorical variable from the PAIR transactions data coding the following:

1. Non-final rejection: event code "CTNF"

2. Response to non-final rejection: event code "A..."

3. Final rejection: event code "CTFR"

4. Response to final rejection: "A.NE," "ACPA," "N/AP," or "RCEX"

Creating this categorical variable required a number of assumptions:

1. We recode all final rejection responses that occur prior to a final rejection as non-final rejection responses, under the assumption that these were clerical errors.

2. Two or more consecutive rejections or responses of the same type are recoded so that each appears only once (e.g., if an application has, say, two non-final rejection responses after a given non-final rejection, then we treat this application as if it had only a single non-final rejection response after that non-final rejection).

3. We count appeals as responses only if they follow final rejections. Hence, we drop all instances of a notice of appeals (i.e. event code "N/AP"), continuations (i.e. event code "ACPA"), or requests for time extensions (i.e. event code "RCEX") that do not follow a final rejection.

4. Except for applications that are initially accepted, we drop all observations where the first event code in the transaction history is not "CTNF." Our goal here is to prevent instances where a non-final rejection response (event code "A...") happens prior to the first non-final rejection due to a restriction requirement from a parent application.

5. We treat any events other than a patent grant after the first final rejection and/or first final response as irrelevant. That is, we treat all applications that restart the rejection-and-response process after receiving and/or responding to a final rejection as simply having reached those stages, unless they are granted a patent.

Given this categorical variable, for each patent application number in the PAIR transactions data we retain information on the final (maximum) stage reached for each application, among the four categories constructed above (non-final rejection, response to non-final rejection, final rejection, and response to non-final rejection). We then merge this information to our full list of patent application numbers, and add in information on three additional application stages reported as 'disposals': abandonment (ABN), pending (PEND), and grants (ISS).

Given this data, we are then able to construct our two variables of interest. First, we construct a variable recording the furthest stage that a patent application progressed to in the prosecution process; summary statistics on this variable for the full sample of patent applications are shown in Table A.1. Most applications (around 62%) are granted patents, but not insubstantial shares reach the stage of receiving a non-final (12.5%) or final (6.5%) rejection and never progress further in the process.

Table A.1: **Furthest Stage Reached in Patent Prosecution Process for USPTO Patent Applications**

	Applications filed in 2000-2010 (N=2,954,249)	Share (%)
No rejections or responses	92,341	3.13
Non-final rejection	370,436	12.54
Non-final rejection response	57,568	1.95
Final rejection	192,786	6.53
Final rejection response	406,792	13.77
Granted	1,834,326	62.09

Second, we construct a variable recording the final status for each patent application as of the end of our data; summary statistics on this variable for the full sample of patent applications are shown in Table A.2. As in Table A.1, most applications (around 62%) are granted patents, but not insubstantial shares are abandoned after receiving a non-final (12%) or final (6%) rejection.

Table A.2: **Final Status in Patent Prosecution Process for USPTO Patent Applications**

	Applications filed in 2000-2010 (N=2,954,249)	Share (%)
Abandoned after no rejections or responses	83,185	2.82
Abandoned after non-final rejection	363,068	12.29
Abandoned after non-final rejection response	47,960	1.62
Abandoned after final rejection	185,871	6.29
Abandoned after final rejection response	277,377	9.39
Pending	162,462	5.50
Granted	1,834,326	62.09

A.4.2 Descriptive Statistics on Decision-and-Response Rounds

To quantify a different aspect of the patent prosecution process, we separately analyzed the number of decision-and-response rounds that each patent application entered by the end of our PAIR transactions data (26 January 2015).

Our methodology for constructing this data is as follows:

1. Measure the initial decision on each application, which we designate to be the start of round 1.[56]

2. If an allowance (event code "MN/=." or "N/=.") is made in round 1, then we denote the application as having completed its examination process. If a rejection (event code "CTNF," "MCTNF," "CTFR," or "MCTFR") is made in round 1, then we search for a response after that rejection.

[56] If an application has no observed allowances, rejections, or responses, we infer from the event code "IEXX" that although the application was received by the USPTO, no decision was ever rendered by the office.

3. If no response is found to said rejection, then no new round starts. If we do find a response to a rejection (event code "A...," "A.NE," or "A.QU") then round 2 begins with the date of that response, and we search for the next decision occurring after that response.

4. If an allowance (event code "MN/=." or "N/=.") is made in round 2, then we denote the application as having completed its examination process. If a rejection (event code "CTNF," "MCTNF," "CTFR," or "MCTFR") is made in round 2, then we search for a response after that rejection.

5. If no response is found to said rejection, then no new round starts. If we do find a response to a rejection (event code "A...," "A.NE," or "A.QU") then round 3 begins with the date of that response, and we search for the next decision occurring after that response.

6. We repeat steps 4 and 5 until the last observed decision without a response is made, or until the applicant responds but no decision is made by the USPTO.

We apply the methodology above to all patent applications included in the PAIR transactions data, except for 14 applications with filing dates that are later than the date of their earliest observed decision (under the assumption that these were clerical errors). Table A.3 documents summary statistics on this "rounds" variable for the full sample of patent applications. Around a quarter (26%) of applications start only one decision round; around two thirds (63%) start only two or fewer decision rounds.

Table A.3: **Number of Decision Rounds for USPTO Patent Applications**

	USPTO applications sample (N=2,954,235)	Share (%)	Cumul. share (%)
# rounds started			
1	768,371	26.01	26.01
2	1,096,279	37.11	63.12
3	604,524	20.46	83.58
4	254,598	8.62	92.20
5	125,944	4.26	96.46
6	54,939	1.86	98.32
7	26,661	0.90	99.22
8	12,167	0.41	99.64
9	5,768	0.20	99.83
10	2,625	0.09	99.92
11	1,264	0.04	99.96
12	567	0.02	99.98
13	276	0.01	99.99
14	121	0.00	100.00
15	60	0.00	100.00
16	31	0.00	100.00
17	15	0.00	100.00
18	10	0.00	100.00
19	8	0.00	100.00
20	3	0.00	100.00
21	1	0.00	100.00
22	1	0.00	100.00
23	1	0.00	100.00
24	0	0.00	100.00
25	0	0.00	100.00
26	1	0.00	100.00

B Appendix: Additional Background on *AMP v. Myriad* Case (for Online Publication)

This appendix provides some additional background information on the recent *AMP v. Myriad* case.

The private firm Myriad Genetics was granted patent rights on human genes correlated with risks of breast and ovarian cancer. In 2009, the American Civil Liberties Union (ACLU) and the Public Patent Foundation filed suit against Myriad, arguing that many of Myriad's patent claims were invalid on the basis that DNA should not be patentable. One technical detail that is critical to understanding the *AMP v. Myriad* case is that two types of nucleotide sequences were at issue: naturally occurring genomic DNA (gDNA), and complementary or cDNA, which is produced in a laboratory using gDNA as a template. After a series of lower court decisions, in June 2013 the US Supreme Court issued a unanimous ruling drawing a distinction between these two types of sequences: "A naturally occurring DNA segment is a product of nature and not patent eligible...but cDNA is patent eligible because it is not naturally occurring."[57]

The question of whether DNA is patent eligible may at first blush seem very far removed from the economics of gene patents. Yet in fact, the US Supreme Court decision was made in part on the basis of the research question examined in this paper: whether patents on human genes would impede follow-on innovation. A brief background on patent eligibility is helpful in clarifying this point. The patent eligibility criteria set out in the US Code (35 U.S.C. §101) has long been interpreted to exclude laws of nature, natural phenomena, and abstract ideas from patent eligibility. The *AMP v. Myriad* decision followed this precedent, arguing that *"[g]roundbreaking, innovative, or even brilliant"* discoveries of natural phenomena should be patent-ineligible, because patents *"would 'tie up' the use of such tools and thereby inhibit future innovation premised upon them."* As discussed by Rai and Cook-Deegan (2013), the Court decision essentially aimed to draw a line between patent-eligible and patent-ineligible discoveries based on the "delicate balance" between patents prospectively creating incentives for innovation and patent claims blocking follow-on innovation. In the end, the Court drew this line by ruling naturally occurring DNA patent-ineligible, and non-naturally occurring cDNA patent-eligible.

Numerous legal scholars have argued that the distinction between DNA and cDNA is "puzzling and contradictory" (Burk 2013) given that "both isolated sequences and cDNA...have identical informational content for purposes of protein coding" (Golden and Sage 2013); in interviews, patent attorneys expressed similar confusion (Harrison 2013). A recent analysis of gene patent claims by Holman (2012) concluded that most human gene patents claimed cDNA, and would thus be unaffected by the Court ruling.

[57] The earlier decisions were a 2010 ruling by the US District Court for the Southern District of New York (see `http://www.pubpat.org/assets/files/brca/brcasjgranted.pdf`) and a 2011 ruling by the US Court of Appeals for the Federal Circuit (see `https://www.aclu.org/files/assets/10-1406.pdf`); a subsequent re-hearing of the case by the US Court of Appeals at the request of the US Supreme Court did not substantively change this decision.

C Appendix: Data Construction (for Online Publication)

This appendix describes our data construction in more detail. A brief background on application, publication, and patent numbers is useful before describing our data from the United States Patent and Trademark Office (USPTO).

Application numbers: The USPTO assigns patent applications application numbers, which consist of a series code and a serial number.[58] The USPTO states that these application numbers are assigned by the Office of Patent Application Processing (OPAP) immediately after mail has been opened.[59] As suggested by this process, the USPTO and other sources note that application numbers are assigned chronologically.[60] While application serial numbers are six digits, the use and length of application series codes has changed over time: in recent years series codes are two digits, but previously these codes were one digit and historically series codes were not used.[61]

Publication numbers: Traditionally, unsuccessful patent applications were not published by the USPTO. However, as part of the American Inventors Protection Act of 1999, the vast majority of patent applications filed in the US on or after 29 November 2000 are published eighteen months after the filing date. There are two exceptions. First, applications granted or abandoned before eighteen months do not appear in this sample unless the applicant chooses to ask for early publication. Lemley and Sampat (2008) estimate that about 17 percent of patents are granted before eighteen months, of which about half (46 percent) are published pre-patent grant. Second, applications pending more than eighteen months can "opt out" of publication if they do not have corresponding foreign applications, or if they have corresponding foreign applications but also have priority dates pre-dating the effective date of the law requiring publication (Lemley and Sampat 2008).[62] If the patent application is published, then the USPTO assigns the application a publication number of the form USYEARXXXXXXX: a 2-digit country code, always US; followed by a 4-digit year (denoting year of publication); followed by a 7-digit identifier.

Patent numbers: Applications that are granted patents are assigned patent numbers. The number of characters in the patent number varies by the type of patent.[63] Utility patent numbers are six or seven digits; reissue patents start with "RE" followed by six digits;[64] plant patents start with "PP" followed by six digits; design patents start with "D" followed by seven digits; additions of improvement patents start with "AI" followed by six digits;[65] X-series patents start with "X" followed by seven digits;[66] H documents

[58] For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm`.

[59] See `http://www.uspto.gov/web/offices/pac/mpep/s503.html`: *"Application numbers consisting of a series code and a serial number are assigned by the Office of Patent Application Processing (OPAP) immediately after mail has been opened. If an application is filed using the Office's electronic filing system, EFS-Web provides an Acknowledgement Receipt that contains a time and date stamp, an application number and a confirmation number."*

[60] See `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/filingyr.htm` (*"In general, patent application serial numbers are assigned chronologically to patent applications filed at the U.S. Patent and Trademark Office."*) and `http://www.thomsonfilehistories.com/docs/RESOURCES_Series_Codes.pdf` (*"US patent applications consist of a 2-digit series code and a 6-digit application serial that is assigned chronologically as they are received at the USPTO."*).

[61] Note that design applications, provisional applications, and reexamination (*ex parte* and *inter partes*) applications are assigned different series codes; reissue patent application numbers follow the utility and design application structures. See `http://www.uspto.gov/web/offices/pac/mpep/s503.html` for details on these series codes.

[62] For more details, see `http://www.uspto.gov/web/offices/pac/mpep/s1120.html` and the discussion in Lemley and Sampat (2010). Most applications not published eighteen months after filing are instead published sixty months after filing.

[63] For more details, see `http://www.uspto.gov/patents/process/file/efs/guidance/infopatnum.jsp`.

[64] For more details on reissue patents, see `http://www.uspto.gov/web/offices/pac/mpep/s1401.html`.

[65] Addition of improvement patents were issued between 1838 and 1861 and covered an inventor's improvement on his or her own patented device. For more details, see §901.04 of `http://www.uspto.gov/web/offices/pac/mpep/s901.html`.

[66] X-series patents were issued between 1790 and 1836. For more details, see §901.04 of `http://www.uspto.gov/web/offices/pac/mpep/s901.html`.

start with "H" followed by seven digits;[67] and T documents start with "T" followed by seven digits.[68]

Data on USPTO Published Patent Applications

USPTO Full-Text Published Patent Applications

Google currently hosts bulk XML downloads of US patent applications published between 15 March 2001 to March 2015.[69] We parse the full text of these patent applications using a Python script. Each published patent application is associated with a publication number.

Using the filing dates listed on the published patent applications, we restrict our sample to applications filed on or after 29 November 2000, the date when "rejected" (abandoned) applications were required to be published. We also use the kind codes listed on the published patent applications to exclude corrected applications as well as subsequent publications of applications.

The full-text published patent applications also provide one measure of patent value: claims count. We use claims count as one proxy for the ex ante value of a patent application that is fixed at the time of patent application, as proposed by Lanjouw and Schankerman (2001). The key idea here is that patents list "claims" over specific pieces of intellectual property, so that patents with more claims may be more valuable. Past work has documented mixed empirical evidence on whether that is a valid assumption based on correlations of claims counts with other value measures.

USPTO Patent Document Pre-Grant Authority Files

We exclude from our analysis a very small number (1,025) of published patent applications that are "withdrawn" applications, which tend to be inconsistently reported across the various datasets used in our analysis. We use the Pre-Grant Authority files made available by the USPTO to exclude these withdrawn applications from our sample. The Pre-Grant Authority files are made available as part of the USPTO's Patent Document Authority Files, and contain listings of all US published applications beginning 15 March 2001.[70] Our versions of these files were downloaded on 24 March 2014 and are up to date as of February 2014. Each published patent application in this data is associated with a publication number.

USPTO PAIR Data

The USPTO Patent Application Information Retrieval (PAIR) data records information about the status of patent applications as they are reviewed by the USPTO, and provides a unique set of insights into many aspects of the patent prosecution process. For example, the PAIR data records examiner names as well as actions by both the applicant and the USPTO on each application. The PAIR data is hosted on the USPTO website.[71] Each application in this data is associated with an application number.

USPTO Patent Assignment Data

The USPTO Patent Assignment data records assignment transactions. These are legal transfers of all or part of the right, title, and interest in a patent or application from an existing owner to a recipient. This

[67] H documents are part of the statutory invention registration series. For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/issudate.pdf`.

[68] T documents are part of the defensive publication series. For more details, see `http://www.uspto.gov/web/offices/ac/ido/oeip/taf/issudate.pdf`.

[69] Available at `http://www.google.com/googlebooks/uspto-patents-applications-text.html`.

[70] Available at: `http://www.uspto.gov/patents/process/search/authority/index.jsp`.

[71] Available at: `https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-examination-research-dataset-public-pair`.

data is hosted on the USPTO website.[72] Each transaction is associated with a patent number, application number, and/or publication number wherever applicable.

We use the USPTO Patent Assignment Dataset to fill in—where possible—missing assignee names in the full-text published patent applications data. Based on conversations with individuals at the USPTO, we infer initial patent application assignments as follows. For each transaction associated with a patent or patent application we define the date of the transaction to be the latest execution date of a given transaction (one transaction can have multiple execution dates if, for example, there are multiple assignees).[73] Once we assign this unique date to each transaction, we then select the earliest transaction. We then infer initial transactions only in cases where assignee names are missing in the published application, and in those cases we fill in the assignee names from the initial assignment included in the USPTO Patent Assignment Dataset.

Thomson Innovation Data on Published USPTO Patent Applications

We use the Thomson Innovation data as an additional source of data on patent applications. Specifically, the Thomson Innovation database provides information on a second measure of patent value: patent "family size." We use patent family size as a second proxy for the ex ante value of a patent application that is fixed at the time of patent application, as developed in Putnam (1996). A patent family is a group of related patents covering the same invention. Conceptually, this includes two types of patents: first, within-country family members include continuations, continuations-in-part, and divisionals; and second, foreign family members include patent applications covering the same technology in other jurisdictions. We here briefly describe each group of patents to motivate our family size measure:

- *Within-country patent families.* Within a country, patent families may include continuations, continuations-in-part, and divisionals. Because our focus is on US patent applications, we focus here on describing within-country patent families only for the US. This description summarizes material in the USPTO's *Manual of Patent Examining Procedure*.[74] A "continuation" is a subsequent application covering an invention that has already been claimed in a prior application (the "parent" application). A "continuation-in-part" is an application filed which repeats some portion of the parent application but also adds in new material not previously disclosed. A divisional application arises when an applicant divides claims in a parent application into separate patent applications. Taken together, the use of continuations, continuations-in-part, and divisionals imply that more than one patent can issue from a single original patent application. Lemley and Sampat (2008) document that among utility patent applications filed in January 2001 and published by April 2006 (a sample of 9,960 applications), 2,016 "children" (continuations, continuations-in-part, or divisionals) had been filed by April 2006: around 30% were continuations, 20% were continuations-in-part, and 40% were divisionals (an additional 10% were of indeterminable types).

- *Foreign patent families.* Patent protection is jurisdiction-specific, in the sense that a patent grant in a particular jurisdiction provides the patent assignee with a right to exclude others from making, using, offering for sale, selling, or importing the invention into that jurisdiction during the life of the patent (subject to the payment of renewal fees). Hence, for any given patent application, applicants must choose how many jurisdictions to file patent applications in, and given that there is a per-jurisdiction cost of filing we would expect patents that are perceived by the applicant as more

[72] Available at: `https://www.uspto.gov/learning-and-resources/electronic-data-products/patent-assignment-dataset`.

[73] See page 12 of the USPTO Patent Assignment data documentation: `https://www.uspto.gov/sites/default/files/documents/USPTO_Patents_Assignment_Dataset_WP.pdf`.

[74] Available at `http://www.uspto.gov/web/offices/pac/mpep/s201.html`.

privately valuable to be filed in a larger number of jurisdictions.[75] The first patent application is referred to as the priority application, and the filing date of the first application is referred to as the priority date; while the priority application can be filed in any jurisdiction, Putnam (1996) argues that the transaction costs involved with foreign filings (e.g. translation of the application) generally imply that domestic filing is cheaper than foreign filing, and that most priority applications are filed in the inventor's home country. Under the Paris Convention for the Protection of Industrial Property (signed in 1883), all additional filings beyond the priority application that wish to claim priority to the priority application must occur within one year of the priority date. Putnam (1996) argues that most foreign applications—if filed—are filed near the one-year anniversary of the home country filing.

- *Commonly used measures of patent family size.* The term patent family can be used to describe different constructs: a patent family can be defined to include only documents sharing exactly the same priority or combination of priorities, or as all documents having at least one common priority, or as all documents that can be directly or indirectly linked via a priority document. There are three commonly-used measures of family size: Espacenet (produced by the European Patent Office), the Derwent World Patents Index (DWPI; produced by Thomson Reuters), and INPADOC (produced by the European Patent Office). Researchers tend to rely on these measures because collecting data from individual non-USPTO patent authorities would be quite time-consuming.

    1. Espacenet uses a 'simple' patent family definition which defines a patent family as all documents with exactly the same priority or combination of priorities. This family is constructed based on data covering around 90 countries and patenting authorities.

    2. DWPI uses a similar patent family definition which defines a patent family as all documents with exactly the same priority or combination of priorities, but also includes non-convention equivalents (e.g. applications filed beyond the 12 months defined by the Paris Convention). This family is constructed based on data covering around 50 patent authorities and defensive publications (international technology disclosures and research disclosures). Continuations and divisionals are not included in the DWPI family definition.

    3. INPADOC defines a patent family more broadly, defining an 'extended' patent family as all documents that can be directly or indirectly linked via a priority document even if they lack a common priority. This family is constructed based on the same data as the Espacenet measure.

- *Our measure of patent family size.* For the purpose of our study, we would like to use the general concept of family size to develop a proxy for patent value that is fixed at the time the patent application is filed. Given this objective, it is clear that we should exclude continuations, continuations-in-part, and divisionals from our family size measure: these applications arise—by construction—after the filing date of the original patent application. In addition—and potentially more concerning in our specific context—the propensity for applications to develop continuations, continuations-in-part, or divisionals may differ across examiners, and hence could be affected by the examiner. We define patent family size as the number of unique countries in which the patent application was filed (as measured in the DWPI patent family).

[75] Multi-national routes such as applications filed with the European Patent Office or Patent Cooperation Treaty applications are intermediate steps towards filings in specific jurisdictions.

Data on USPTO Granted Patents

USPTO Full-Text Granted Patents

Google currently hosts bulk XML downloads of US patent grants published between 1 January 1976 and 31 December 2014.[76] We parse the full text of these patent grants using a Python script.

While patent number is a unique identifier of patent grants, there are twenty-one patent numbers in this data which correspond to patents granted in 1987 that each appear twice with different grant dates. Checking these patent numbers on the USPTO's online Patent Full Text (PatFT) database reveals that, in each of these cases, the duplicated patent number with the earlier grant date is correct.[77] Accordingly, we drop the twenty-one observations with the later grant dates.

NBER Technology Category Data

Hall et al. (2001) constructed a linkage of US patents granted between January 1963 and December 1999 with the Compustat data. As part of that work, the authors constructed technology categories to describe the broad content area of different patents, based on aggregations of the patent technology class and subclass variables. From their work, we draw a crosswalk between United States Patent Classification values and the technology categories as defined by these authors. The NBER hosts this data on its website.[78]

Data on DNA-Related USPTO Published Patent Applications

CAMBIA Patent Lens Data

The CAMBIA Lens database provides a list of published USPTO patent applications associated with human and non-human protein and DNA sequences appearing in patent claims (Bacon et al. 2006).[79] This data construction was supported by the Ministry of Foreign Affairs of Norway through the International Rice Research Institute for CAMBIA's Patent Lens (the OS4 Initiative: Open Source, Open Science, Open Society, *Orzya sativa*).

Over the time period relevant for our analysis, US patent applications list DNA sequences in patent claims with a canonical 'sequence listing' label, e.g. SEQ ID NO:1 followed by the relevant DNA sequence. The CAMBIA Patent Lens data construction parses patent claims text for lists of SEQ ID NOs to determine which sequences are referenced in the claims. Importantly, CAMBIA makes available several versions of their data; following Jensen and Murray (2005), we focus on the dataset of nucleotide sequences (as opposed to amino acid sequences), and on the 'in-claims' subsample (as opposed to a larger dataset which includes DNA sequences listed in the text of the patent application, but not explicitly listed in the patent claims).

The CAMBIA Patent Lens data is updated over time; our version is current as of 8 February 2012. The level of observation is a patent-mRNA pair indexed by a publication/sequence number that combines the patent publication number and a mRNA sequence number. The patent publication numbers were extracted from the CAMBIA Patent Lens data using a Perl script.[80]

---

[76] Available at `https://www.google.com/googlebooks/uspto-patents-grants-text.html`.

[77] PatFT can be accessed at `http://patft.uspto.gov/netahtml/PTO/srchnum.htm`

[78] Available at: `https://www.nber.org/patents/`.

[79] Available at `http://www.patentlens.net/sequence/US_A/nt-inClaims.fsa.gz`.

[80] We are very grateful to Mohan Ramanujan for help extracting this data from FASTA format. This Perl script is available on request.

Patome Data

Patome annotates biological sequences in issued patents and published patent applications (Lee et al. 2007).[81] This data construction was supported by the Korean Ministry of Science and Technology (MOST).

Although the full Patome dataset contains issued patents and published patent applications from several jurisdictions—including Japan and Europe—in this paper we focus on the subsample of US published patent applications and granted patents. As in the CAMBIA Patent Lens data, the Patome data construction parses patent application texts for lists of SEQ ID NOs to determine which sequences are referenced in patent applications. Following the methodology pioneered by Jensen and Murray (2005), BLAST (Basic Local Alignment Search Tool) searches are used to compare listed sequences against a census of potential matches in order to identify regions of similarity. Using these BLAST searches, the DNA sequences are annotated with mRNA and gene identifiers (RefSeq and Entrez Gene numbers).

The Patome data includes some patent applications which do not correspond to the definition of human gene patents proposed by Jensen and Murray (2005); to follow the Jensen-Murray definition, we impose some additional sample restrictions. First, the Patome data include sequences which appear in the text of patent applications but are not explicitly listed in patent claims; to follow the Jensen and Murray (2005) definition of gene patents, we exclude observations that do not appear in the patent claims.[82] Second, following Jensen and Murray (2005) we limit the sample to BLAST matches with an E-value of exactly zero; the goal of this conservative E-value is to prevent spurious matches. Finally, following Jensen and Murray (2005) we limit the sample to disclosed sequences that are at least 150 nucleotides in length; the motivation of this restriction is that this is the average length of one human exon and yet still small enough to capture EST sequences.

As in Jensen and Murray (2005), many patents claim multiple variants of the same gene (that is, multiple mRNA sequences corresponding to the same gene). Following their methodology, we focus on variation in patenting across human genes.

Data Measuring Innovation on the Human Genome

Gene-Level Measures of Scientific Publications: OMIM Data

We collect our measure of scientific research from the Online Mendelian Inheritance in Man (OMIM) database, a catalog of Mendelian traits and disorders. We use the full-text OMIM data and extract our variables of interest using a Python script.[83] One gene can be included in more than one OMIM record, and one OMIM record can involve more than one gene. We tally the total number of publications related to each gene in each year across all OMIM records related to that gene.

Gene-Level Data on Drug Development: Pharmaprojects Data

We collect data on gene-related drug development from the Pharmaprojects data.[84] According to the company Citeline, which compiles and sells the Pharmaprojects data, "There is continual two-way communication between Pharmaprojects staff and their contacts in the pharmaceutical and biotechnology

[81] The Patome data appears to no longer be available at the URL from which we obtained it; we would be happy to provide our copy of this data upon request. The same data can also be accessed through the Wayback Machine Internet archive service here: `https://web.archive.org/web/20120317030458/http://verdi.kobic.re.kr/patome_int/data/pat_anno.tar.gz`.

[82] Specifically, we merge the list of patent-mRNA numbers in the Patome data to the CAMBIA Patent Lens data, and drop observations that appear in Patome but not in the CAMBIA data; because our version of the CAMBIA data includes only patent-RNA observations listed in patent claims, this allows us to exclude Patome observations which are not explicitly listed in the claims of the patent applications.

[83] Available at `http://omim.org/downloads`.

[84] Pharmaprojects data is available for purchase through Citeline or the Pharmaprojects website: `http://www.pharmaprojects.com`.

industries; both to gather new data and importantly, to verify information obtained from other sources." Citeline employees gather drug information from company websites, reports, and press releases. Annually, every company covered in the database verifies information related to drugs in the development pipeline.

Pharmaprojects annotates a subset of the clinical trials in their data as related to specific Entrez Gene ID numbers. We construct a count of the number of clinical trials in which each gene is used as the basis for a pharmaceutical treatment compound in clinical trials in each year.

Gene-Level Data on Diagnostic Tests: GeneTests.org Data

We collect our measure of genes being used in diagnostic tests from the GeneTests.org database.[85] This data 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 of 1988, which requires laboratories to meet quality control and proficiency testing standards; there are no such requirements for non-US-based laboratories.

We use the GeneTests.org data as of 18 September 2012, which lists OMIM numbers for which there is any genetic test available in the Genetests.org directory. As with the OMIM data above, one gene can appear in more than one OMIM record, and one OMIM record can involve more than one gene. We construct an indicator for whether a given gene is used in any diagnostic test as of 2012.

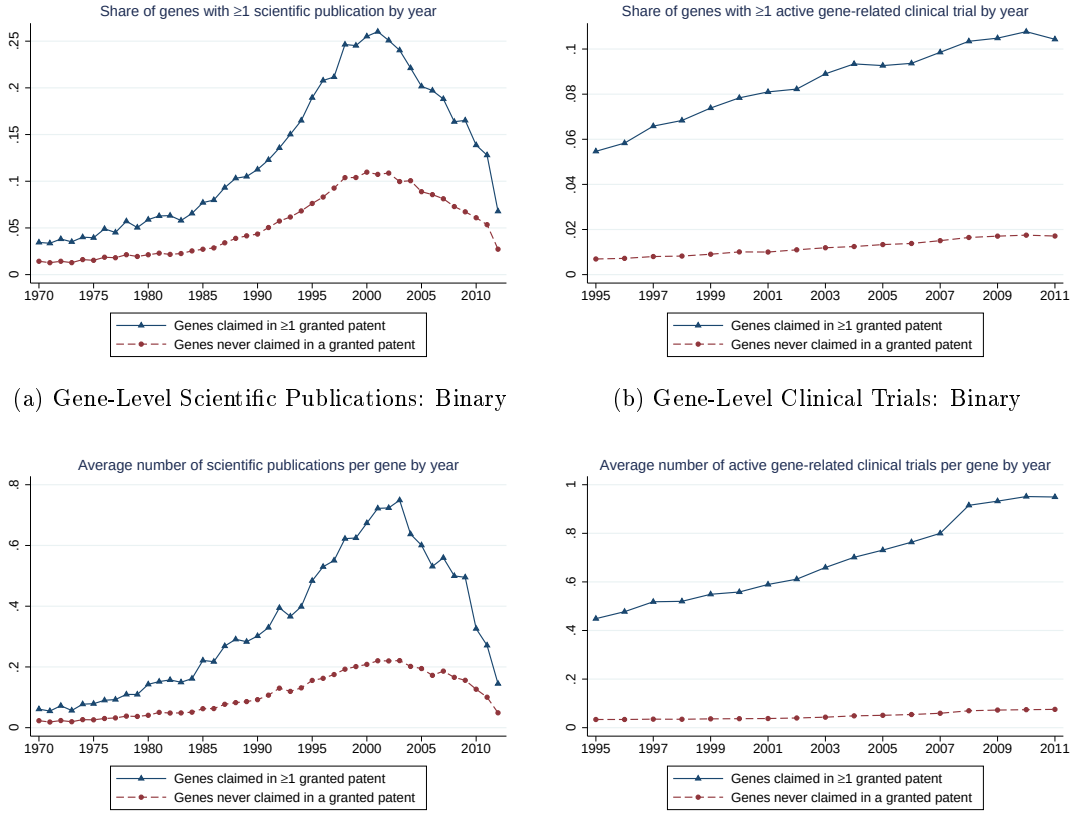
Human Genome-Related Crosswalks

NCBI-generated crosswalks are used to link mRNA- and RNA-level RefSeq accession/version numbers to Entrez Gene ID numbers.[86] We also use an NCBI-generated crosswalk that links discontinued Entrez Gene ID numbers to current Entrez Gene ID numbers, which is useful for linking Pharmaprojects observations that list discontinued Entrez Gene ID numbers to our data.[87]

[85] Available at `ftp://ftp.ncbi.nih.gov/pub/GeneTests/disease_OMIM.txt`.
[86] Available at `ftp://ftp.ncbi.nih.gov/refseq/release/release-catalog/archive/release54.accession2geneid.gz`.
[87] Available at `ftp://ftp.ncbi.nih.gov/gene/DATA/gene_history.gz`.

D Appendix: Additional Results (for Online Publication)

Figure D.1: **Follow-on Innovation on Patented and Non-Patented Human Genes: Robustness**



(a) Gene-Level Scientific Publications: Binary

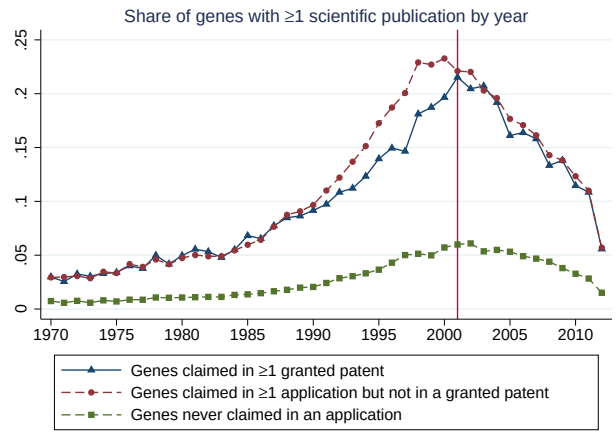
(b) Gene-Level Clinical Trials: Binary

(c) Gene-Level Scientific Publications: Count

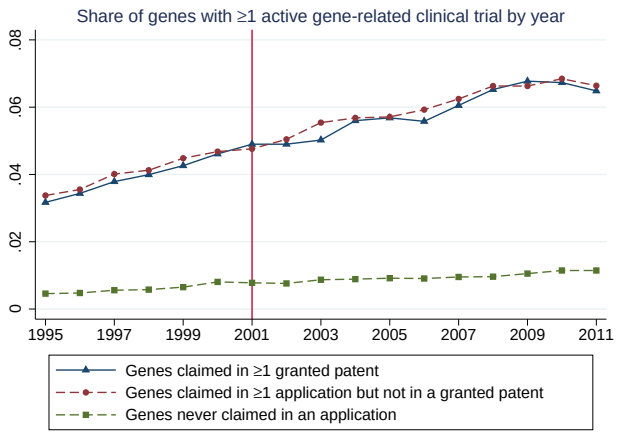
(d) Gene-Level Clinical Trials: Count

Notes: This figure plots trends in follow-on innovation by year separately for genes that ever receive a patent, and for genes that never receive a patent. The figure is constructed from gene-level data. The left-hand side panels use gene-level scientific publications as a measure of follow-on innovation, and the right-hand-side panels use gene-level clinical trials as a measure of follow-on innovation. The first row of figures plots the average of indicator variables for any follow-on innovation by year from 1970 to 2012; the second row of figures plots the average number of each follow-on measure by year from 1995 to 2011.

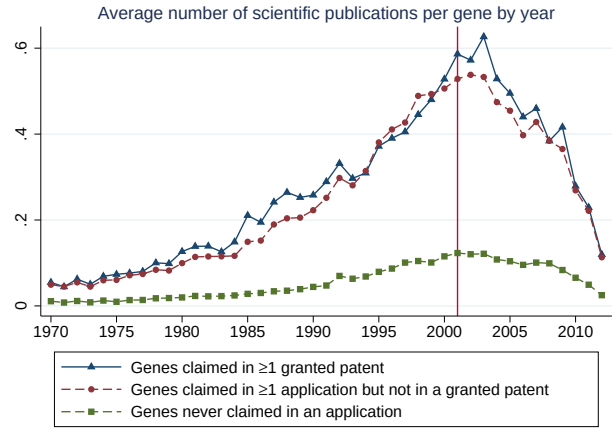
Figure D.2: **Patents and Follow-on Innovation on Human Genes Claimed in Accepted/Rejected Patent Applications: Robustness**



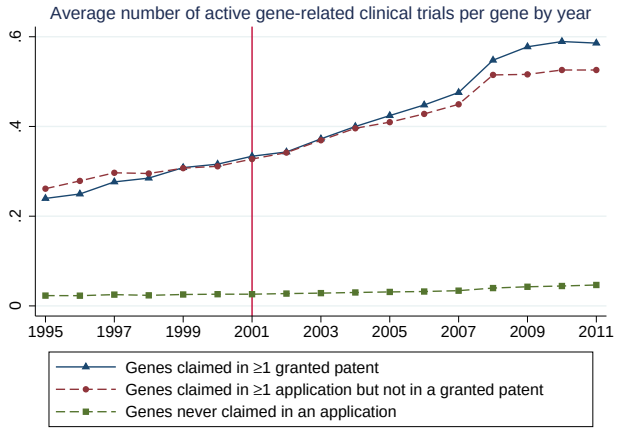
(a) Gene-Level Scientific Publications: Binary



(b) Gene-Gevel Clinical Trials: Binary



(c) Gene-Level Scientific Publications: Count



(d) Gene-Level Clinical Trials: Count

Notes: This figure plots trends in patenting and follow-on innovation by year separately for three groups of genes: genes claimed in at least one granted patent; genes claimed in at least one patent application but never in a granted patent; and genes never claimed in a patent application. The figure is constructed from gene-level data. The left-hand side panels use gene-level scientific publications as a measure of follow-on innovation, and the right-hand-side panels use gene-level clinical trials as a measure of follow-on innovation. The first row of figures plots the average of indicator variables for any follow-on innovation by year, and the second row of figures plots the average number of each follow-on measure by year. The vertical line in the calendar year 2001 denotes that, because this figure focuses on patents that were filed in or after November 2000, all years prior to 2001 can be considered a pre-period and used to estimate the selection of genes into patenting based on pre-patent filing measures of scientific research (publications) and commercialization (clinical trials).

Table D.1: **Robustness to Family Grant Rates: Instrumental Variables Estimates**

	Log of follow-on innovation in 2011/2012 (1)	Any follow-on innovation in 2011/2012 (2)
Panel A: Scientific publications		
Patent granted (instrumented)	-0.0228	-0.0185
	(0.0101)	(0.0088)
Mean of dependent variable	0.0798	0.0888
Number of observations	293,652	293,652
Panel B: Clinical trials		
Patent granted (instrumented)	-0.0483	-0.0290
	(0.0207)	(0.0117)
Mean of dependent variable	0.0690	0.0500
Number of observations	293,652	293,652
Panel C: Diagnostic test		
Patent granted (instrumented)	-	-0.0140
	-	(0.0122)
Mean of dependent variable	-	0.0918
Number of observations	-	293,652

Notes: This table presents instrumental variable estimates, relating follow-on innovation to whether a patent application or any of its child applications were granted a patent, instrumented by our examiner leniency instrument. The sample for these regressions is constructed from application-by-gene-level data, and includes patent applications that claim at least one human gene in our USPTO patent application sample (N=293,652). All regressions include Art Unit-by-application year fixed effects. Each coefficient is from a separate regression. Standard errors are clustered at the patent application level (Inoue and Solon 2010; Pacini and Windmeijer 2016).

Table D.2: **Gene-Level Summary Statistics**

	Mean	Median	Standard deviation	Minimum	Maximum	Number of observations
Scientific publications	0.2238	0	0.8770	0	22	15,524
$\mathbf{1}$(Scientific publications > 0)	0.1094	0	0.3122	0	1	15,524
Clinical trials	0.5446	0	5.1620	0	230	15,524
$\mathbf{1}$(Clinical trials > 0)	0.0659	0	0.2481	0	1	15,524
$\mathbf{1}$(Diagnostic tests > 0)	0.1199	0	0.3249	0	1	15,524

Notes: This table shows summary statistics for our gene-level outcome variables.

Table D.3: **Robustness of Examiner Leniency Estimates: LOOM First Stage Estimates**

	Patent granted
LOOM examiner grant rate	0.6125
	(0.0127)
Mean of dependent variable	0.2515
Number of observations	212,569

Notes: This table estimates the first stage of a patent grant on the leave-one-out examiner grant rate and Art Unit-by-application year fixed effects. The leave-one-out examiner grant rate is the number of applications claiming at least one human gene granted by the examiner other than the given application divided by the total number of applications claiming at least one human gene examined by the examiner minus one for the given application. We only include applications which have examiners with at least 10 examined applications other than the focal patent application. The sample for these regressions is constructed from application-by-gene-level data, and includes patent application-gene-level observations in our human gene sample (N=293,652). Gene-clustered standard errors.

Table D.4: **Robustness of Examiner Leniency Estimates: LOOM Instrumental Variables Estimates**

	Log of follow-on innovation in 2011/2012 (1)	Any follow-on innovation in 2011/2012 (2)
Panel A: Scientific publications		
Patent granted (instrumented)	0.0206	0.0166
	(0.0226)	(0.0262)
Mean of dependent variable	0.0713	0.0821
Number of observations	212,569	212,569
Panel B: Clinical trials		
Patent granted (instrumented)	0.0091	0.0091
	(0.0294)	(0.0186)
Mean of dependent variable	0.0594	0.0445
Number of observations	212,569	212,569
Panel C: Diagnostic test		
Patent granted (instrumented)	-	-0.0426
	-	(0.0284)
Mean of dependent variable	-	0.0836
Number of observations	-	212,569

Notes: This table presents instrumental variable estimates, relating follow-on innovation to whether a patent application was granted a patent, instrumented by our examiner leave-one-out leniency instrument. The leave-one-out examiner grant rate is the number of applications claiming at least one human gene granted by the examiner other than the given application divided by the total number of applications claiming at least one human gene examined by the examiner minus one for the given application. We only include applications that have examiners with at least 10 examined applications other than the focal patent application. The sample for these regressions is constructed from patent application-gene-level data, and includes patent application-by-gene-level observations in our human gene sample (N=293,652). All regressions include Art Unit-by-application year fixed effects. Each coefficient is from a separate regression. Gene-clustered standard errors.

Table D.5: **Robustness of Examiner Leniency Estimates**

Fixed effects included:	Art Unit - by - Application Year	Art Unit - by - Application Year	Art Unit - by - Application Year - by - Class - by - Subclass
	(1)	(2)	(3)
0/1, =1 if patent granted			
Examiner leniency	0.8757	0.8715	0.8316
	(0.0368)	(0.0534)	(0.0538)
Number of observations	14,476	6,747	6,747

Notes: This table presents robustness checks relating the probability of patent grant to examiners' mean non-human gene patent grant rate. Column (1) documents estimates that condition on Art Unit-by-application year fixed effects. Column (3) replaces the Art Unit-by-application year fixed effects with Art Unit-by-application year-by-class-by-subclass fixed effects, estimated on the subsample of data meeting our sample restrictions. For ease of comparability, Column (2) documents estimates that condition on Art Unit-by-application year fixed effects but use the same sample as in Column (3). The sample for these regressions is constructed from patent application-gene-level data, and includes patent application observations in our non-human gene sample (N=14,476).

Table D.6: **Follow-on Innovation on Human Genes: OLS Regression Estimates**

	Log of follow-on innovation in 2011/2012 (1)	Any follow-on innovation in 2011/2012 (2)
Panel A: Scientific publications		
Patent granted	-0.0007	0.0005
	(0.0031)	(0.0036)
Mean of dependent variable	0.0798	0.0888
Number of observations	293,652	293,652
Panel B: Clinical trials		
Patent granted	0.0009	0.0008
	(0.0042)	(0.0027)
Mean of dependent variable	0.0690	0.0500
Number of observations	293,652	293,652
Panel C: Diagnostic test		
Patent granted	-	-0.0062
	-	(0.0036)
Mean of dependent variable	-	0.0918
Number of observations	-	293,652

Notes: This table presents ordinary least squares estimates, relating follow-on innovation to whether a patent application was granted a patent. Each coefficient is from a separate regression. The sample for these regressions is constructed from application-by-gene-level data, and includes patent applications that claim at least one human gene in our USPTO patent application sample (N=293,652). All regressions include Art Unit-by-application year fixed effects. Gene-clustered standard errors.