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

Bhaven Sampat and Heidi L. Williams

November 21, 2018

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

In this appendix, we describe the USPTO patent examination process in more detail.[1]

1.1 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.[2] 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.[3] 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).

1.2 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

[1] The discussion in this appendix draws heavily on Cockburn et al. (2003), Lemley and Sampat (2012), and United States Government Accountability Office (2005), among other sources referenced in the text.

[2] 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`.

[3] 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).

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 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.[4] In the absence of such information, we rely on the interviews in Lemley and Sampat (2012) as a guide to designing our empirical specifications.

1.3 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.[5] 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.

1.4 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.

1.4.1 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 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:

[4] 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.

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

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 1.1. Most applications (around 62 percent) are granted patents, but not insubstantial shares reach the stage of receiving a non-final (12.5 percent) or final (6.5 percent) rejection and never progress further in the process.

Table 1.1: **Furthest Stage Reached in Patent Prosecution Process for USPTO Patent Applications**

	Applications filed in 2000-2010 (N=2,954,249)	Share (percent)
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 1.2. As in Table 1.1,

most applications (around 62 percent) are granted patents, but not insubstantial shares are abandoned after receiving a non-final (12 percent) or final (6 percent) rejection.

Table 1.2: **Final Status in Patent Prosecution Process for USPTO Patent Applications**

	Applications filed in 2000-2010 (N=2,954,249)	Share (percent)
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

1.4.2 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.[6]
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.
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 1.3 documents summary statistics on this "rounds" variable for the full sample of patent applications. Around a quarter (26 percent) of applications start only one decision round; around two thirds (63 percent) start only two or fewer decision rounds.

[6] 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.

Table 1.3: **Number of Decision Rounds for USPTO Patent Applications**

	USPTO applications sample (N=2,954,235)	Share (percent)	Cumul. share (percent)
# rounds started			
	768,371	26.01	26.01
	1,096,279	37.11	63.12
	604,524	20.46	83.58
	254,598	8.62	92.20
	125,944	4.26	96.46
	54,939	1.86	98.32
	26,661	0.90	99.22
	12,167	0.41	99.64
	5,768	0.20	99.83
	2,625	0.09	99.92
	1,264	0.04	99.96
	567	0.02	99.98
	276	0.01	99.99
	121	0.00	100.00
	60	0.00	100.00
	31	0.00	100.00
	15	0.00	100.00
	10	0.00	100.00
	8	0.00	100.00
	3	0.00	100.00
	1	0.00	100.00
	1	0.00	100.00
	1	0.00	100.00
	0	0.00	100.00
	0	0.00	100.00
	1	0.00	100.00

2 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."[7]

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" (p. 12) 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" (p. 11). 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 nonnaturally occurring cDNA patent-eligible.

Numerous legal scholars have argued that the distinction between DNA and cDNA is "puzzling and contradictory" (Burk, 2013, p. 747) given that "both isolated sequences and cDNA...have identical informational content for purposes of protein coding" (Golden, 2013 p. 1345); 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.

[7]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.

3 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.[8] The USPTO states that these application numbers are assigned by the Office of Patent Application Processing (OPAP) immediately after mail has been opened.[9] As suggested by this process, the USPTO and other sources note that application numbers are assigned chronologically.[10] 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.[11]

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).[12] 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.[13] Utility patent numbers are six or seven digits; reissue patents start with "RE" followed by six digits;[14] 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;[15] X-series patents start with "X" followed by seven digits;[16] H documents start with "H" followed by seven digits;[17] and T documents start with "T" followed by seven digits.[18]

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.[19] We parse the full text of these patent applications using a Python script. Each published patent application is

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

[9] 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 theipt 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."

[10] 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.").

[11] 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.

[12] 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.

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

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

[15] 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`.

[16] 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`.

[17] 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`.

[18] 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`.

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

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.[20] 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 are hosted on the USPTO website.[21] 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. These data are hosted on the USPTO website.[22] 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).[23] 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

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

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

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

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

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*.[24] 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 percent were continuations, 20 percent were continuations-in-part, and 40 percent were divisionals (an additional 10 percent 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 privately valuable to be filed in a larger number of jurisdictions.[25] 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.

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

[25] 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.

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).

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.[26] 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.[27] 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.[28]

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 nonhuman protein and DNA sequences appearing in patent claims (Bacon et al., 2006).[29] 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

[26] Available at `https://www.google.com/googlebooks/uspto-patents-grants-text.html`.
[27] PatFT can be accessed at `http://patft.uspto.gov/netahtml/PTO/srchnum.htm`
[28] Available at: `https://www.nber.org/patents/`.
[29] Available at `http://www.patentlens.net/sequence/US_A/nt-inClaims.fsa.gz`.

number and a mRNA sequence number. The patent publication numbers were extracted from the CAMBIA Patent Lens data using a Perl script.[30]

Patome Data

Patome annotates biological sequences in issued patents and published patent applications (Lee et al. 2007).[31] 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.[32] 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.[33] 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.[34] 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 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.

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

[31] 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 dataset 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`.

[32] 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.

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

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

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.[35] 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

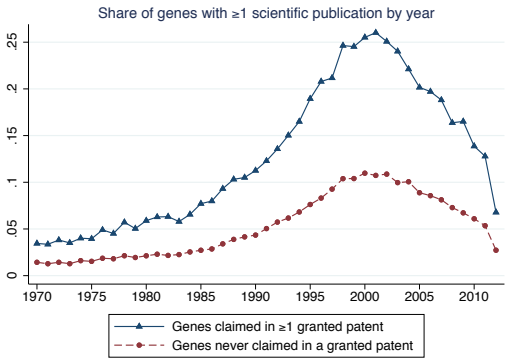
NCBI-generated crosswalks are used to link mRNA- and RNA-level RefSeq accession/version numbers to Entrez Gene ID numbers.[36] We also use an NCBI-generated crosswalk that links discontinued Entrez Gene ID numbers to current Entrez Gene ID numbers, which is useful for linking Pharmaprojects observations that list discontinued Entrez Gene ID numbers to our data.[37]

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

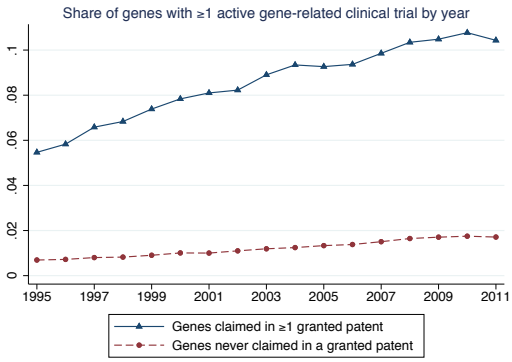
4 Appendix: Additional Results (for Online Publication)

Figure 4.1: **Follow-on Innovation on Patented and Non-Patented Human Genes: Robustness**

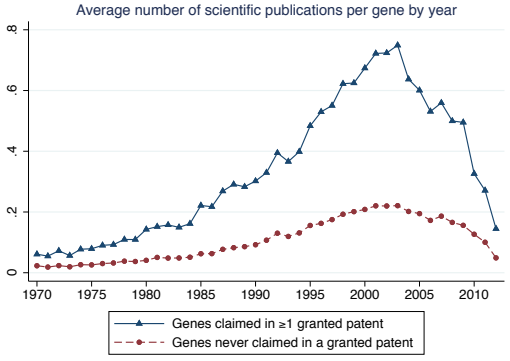
(a) Gene-Level Scientific Publications: Binary



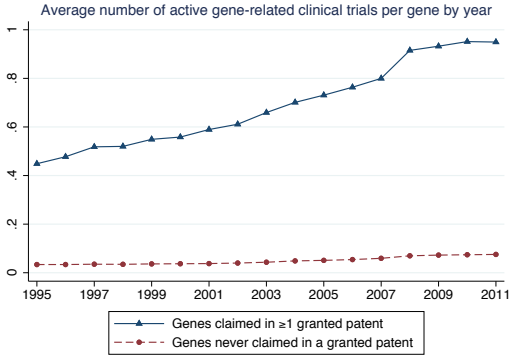
(b) Gene-Level Clinical Trials: Binary



(c) Gene-Level Scientific Publications: Count



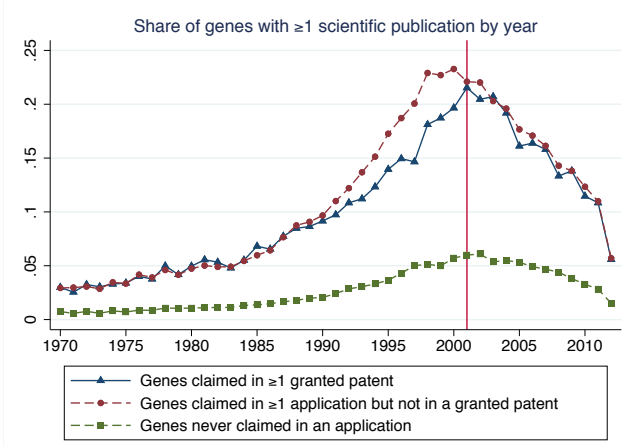
(d) Gene-Level Clinical Trials: Count



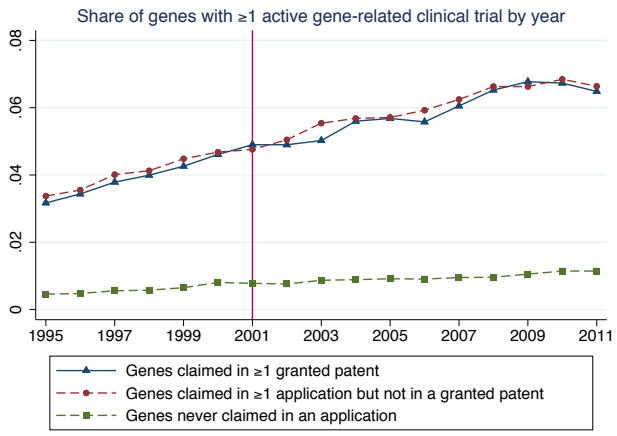
Notes: This figure plots trends in follow-on innovation by year separately for genes that ever receive a patent, and for genes that never receive a patent. The figure is constructed from gene-level data. The left-hand side panels use gene-level scientific publications as a measure of follow-on innovation, 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 4.2: **Patents and Follow-on Innovation on Human Genes Claimed in Accepted/Rejected Patent Applications: Robustness**

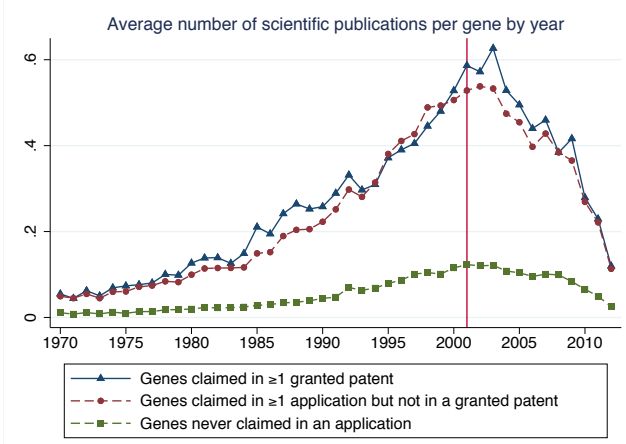
(a) Gene-Level Scientific Publications: Binary



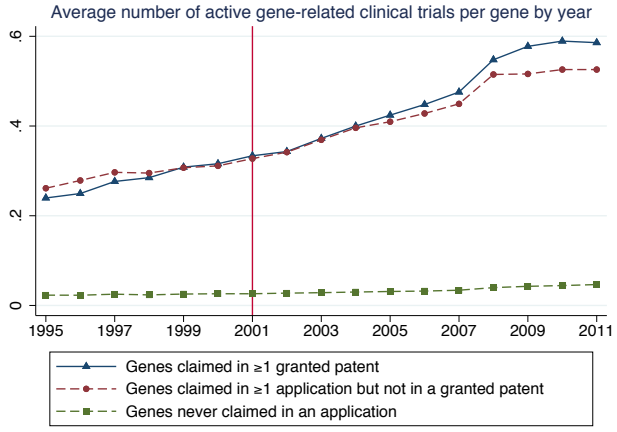
(b) Gene-Level Clinical Trials: Binary



(c) Gene-Level Scientific Publications: Count



(d) Gene-Level Clinical Trials: Count



Notes: This figure plots trends in 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 4.1: **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 4.2: **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 4.3: **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 4.4: **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 4.5: **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 4.6: **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.

References

[1] Neil Bacon and Doug Ashton and Richard Jefferson and Marie Connett, "Biological sequences named and claimed in US patents and patent applications: CAMBIA Patent Lens OS4 Initiative" (2006).

[2] Dan Burk, "Are human genes patentable", International Review of Intellectual Property and Competition Law 44, 7 (2013), pp. 747-749.

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

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

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

[6] Charlotte Harrison, "Isolated DNA patent ban creates muddy waters for biomarkers and natural products", Nature Reviews Drug Discovery 12 (2013), pp. 570-571.

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

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

[9] Kyle Jensen and Fiona Murray, "Intellectual property landscape of the human genome", Science 310, 5746 (2005), pp. 239-240.

[10] Jean Lanjouw and Mark Schankerman, "Characteristics of patent litigation: A window on competition", RAND Journal of Economics 32, 1 (2001), pp. 129-151.

[11] Byungwook Lee and Taehyung Kim and Seon-Kyu Kim and 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 35, Database issue (2007), pp. D47-D50.

[12] Mark Lemley and Bhaven Sampat, "Is the patent office a rubber stamp?", Emory Law Journal 58 (2008), pp. 181-203.

[13] Mark Lemley and Bhaven Sampat, "Examiner characteristics and patent office outcomes", Review of Economics and Statistics 94 (2012), pp. 817-827.

[14] David Pacini and Frank Windmeijer, "Robust inference for the two-sample 2SLS estimator", Economic Letters 146 (2016), pp. 50-54.

[15] Jonathan Putnam, "The value of international patent protection" (1996).

[16] Arti Rai and Robert Cook-Deegan, "Moving beyond 'isolated' gene patents", Science 341 (2013), pp. 137-138.

[17] US Government Accountability Office (GAO), "Intellectual property: USPTO has made progress in hiring examiners, but challenges to retention remain" (2005).

[18] Heidi Williams, "How do patents affect research investments?", Annual Review of Economics 9 (2017), pp. 441-469.