How do patents affect cumulative innovation?
Evidence from the assignment of gene patent applications to patent examiners
Bhaven Sampat and Heidi Williams

Do patents on existing technologies hinder subsequent innovation? Consider a patent on the sequenced data for a human gene. This data is a research input into the subsequent development of so-called 'downstream' products –scientists need to study the sequenced genetic data in order to learn about links between genetic variations and diseases, and can then apply that knowledge to develop drugs and gene-based diagnostic tests. On one hand, patents on genetic sequences may discourage innovation if innovation takes place in multiple firms and transaction costs hinder the negotiation of cross-firm contracts. For example, in 2001 pharmaceutical firm Bristol Myers reported to the *New York Times* that it had abandoned research on more than 50 cancer-related proteins because of conflicts with gene patent holders (Pollack 2001). On the other hand, survey evidence has suggested that researchers adopt 'working solutions' to patents such that patents do not interfere with cumulative innovation (Cohen and Walsh 2008). A 2006 National Academies report concluded there is currently little empirical evidence to either support or refute the assertion that gene patents are impeding medical research. This issue is central to current policy discussions as legal appeals continue following a April 2010 US District Court decision to invalidate seven gene patents held by biotechnology firm Myriad Genetics.

Two main empirical challenges arise in investigating how gene patents affect subsequent innovation. A first challenge is to disentangle 'selection' effects from 'treatment' effects: we need a source of variation in patenting across genes that is not correlated with factors such as a genes' inherent commercial potential. A second challenge is to develop metrics that provide consistent ways of tracking cumulative innovation on both patented *and non-patented* genes. The traditional outcome variable social scientists have used to track cumulative innovation is patent citations, but in this context we need to be able to track cumulative innovation on technologies that *are not granted patents*.

This research project aims to assess whether gene patents have imposed real impediments to scientific research and product development using newly collected data and a novel research design based on the quasi-random assignment of patent applications to patent examiners at the US Patent and Trademark Office (USPTO). In addition to providing a unique setting in which quasi-experimental variation can be used to identify the effects of gene patents on subsequent innovation, this project will develop a new research design that can subsequently be applied in other contexts.

A novel research design: Leveraging heterogeneity across patent examiners

Patent examiners are charged with a uniform mandate: grant patents to novel, non-obvious, useful inventions. However, in practice this mandate leaves patent examiners a fair amount of discretion. Several previous papers have documented tremendous heterogeneity across USPTO patent examiners in general (that is, not specifically focusing on gene patents; see Cockburn, Kortum, and Stern 2003; Lemley and Sampat 2008, 2010, forthcoming). Our idea is to leverage the interaction of this heterogeneity with the quasi-random assignment of patent applications to patent examiners as a source of quasi-experimental variation.[1] Intuitively, if two otherwise similar patent applications are submitted to the USPTO, and one is assigned to a 'lenient' examiner while the other is assigned to a 'strict' examiner, then the patent application assigned to the lenient examiner is more likely to be granted. As long as the assignment of applications to examiners is plausibly unrelated to the underlying characteristics of the patent applications (an assumption we discuss in more detail below), the 'leniency' of the examiner to which a given patent application is assigned can be used as a source of quasi-experimental variation in patenting.

Concretely, this empirical design will start with a sample of patent applications, and test whether patent examiner 'fixed effects' have explanatory power in which of these applications are granted. If there is heterogeneity across patent examiners in a sample of gene patent applications, then we can use the patent examiner fixed effects as instrumental variables for whether a given patent application is granted. That is, we can leverage this variation to ask how patent grants affect subsequent innovation – such as the use of genes in genetic research, and in gene-based drugs and diagnostic tests.

In order to apply this instrumental variables approach, two conditions must be met. First, patent examiner fixed effects must be predictive of patent grants; as detailed below, our preliminary analyses suggest this is the case. Second, applications must be 'as good as randomly assigned' across examiners. Selection would threaten the validity of this research design if different types of patent examiners were systematically assigned relatively more or relatively less patentable applications. Lemley and Sampat (forthcoming) present qualitative evidence suggesting that – conditional on observables such as the technological classification of the patent – such selection is not a problem for this research design. For example, some supervisors assign applications to examiners based on the last digit of the application serial number; because application serial numbers are assigned sequentially in the central USPTO, this assignment system – while not purposefully random – is not subject to manipulation by applications or supervisory examiners.

Measurement: Defining gene patents and tracking cumulative innovation

The empirical analysis will start with a sample of 'gene patent' applications filed in or after 2001 (the first year that unsuccessful patent applications were published). This itself represents a measurement challenge, because the decision of which DNA-related patents should appropriately be considered to be 'gene patents' is controversial in any given application.

[1] Analogous research designs have been applied to foster care via studying foster care case workers (Doyle 2007, 2008) and disability insurance work disincentives via studying disability insurance examiners (Maestas, Mullen, and Strand 2010).

For example, there are many different types of DNA-related patents – *e.g.* protein-encoding sequences, ESTs, SNPs, sequenced-based claims, and method claims that pertain to specific genes or sequences. In the short-term, we will follow the work of Jensen and Murray (2005) to identify gene patent applications as those with DNA sequences listed in them (as catalogued in the Cambia PatentLens database). These patent applications can be linked to the USPTO PAIR data in order to obtain variables such as the patent examiner to which the application was assigned, the date the application was filed, the Art Unit to which the application was assigned, and the subsequent grant outcome (*e.g.* the patent grant date, if granted). In the longer-term, we will explore whether alternative samples of gene patent applications may be more policy-relevant.

From an empirical perspective, human gene patent applications offer a unique opportunity to consistently measure cumulative innovation. Because the gene sequence itself is explicitly listed in the patent application over the time period of our analysis, we can match each patent application to the human genes for which it is relevant. In particular, this matching can be done using 'gene ID numbers' (specifically, RefSeq and Entrez Gene ID numbers) in a way that allows us to reliably track gene-related cumulative innovation on gene patent applications regardless of whether they are granted patents. As implemented in previous research (Williams 2010), Entrez Gene IDs are linkable to the Online Mendelian Inheritance in Man (OMIM) database, which catalogs scientific papers that have documented evidence for links between genes and diseases; this 'paper trail' provides a consistent measure of publications relevant to a given gene. Entrez Gene IDs can also be linked to product market databases: the GeneTests.org database provides information on the use of genes in gene-based diagnostic tests, and the Pharmaprojects database provides information on drug compounds in clinical trials that relate to specific genetic variations. This data construction is novel, and requires the use of bioinformatics methods to implement the matching.

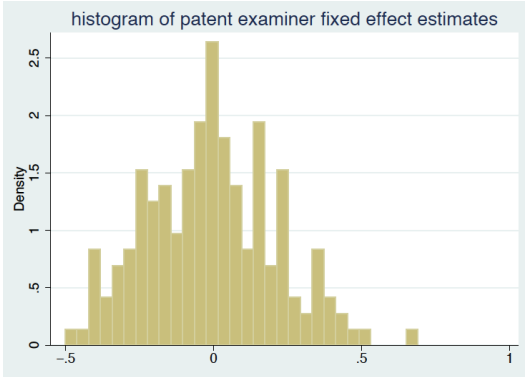
Preliminary results

Thus far, we have completed a preliminary test of the research design on a subset of our gene patent sample – specifically, checking whether patent examiner fixed effects have explanatory power in which patent applications are granted. On a sub-sample of the gene patent applications that was most easily amenable to analysis (n = 4,357), the mean patent grant rate was 30%. Approximately 470 patent examiners appear in this sample. A rule of thumb for the number of observations that is needed per group to allow for meaningful estimates of fixed effects is eight observations per group (Heckman 1981; Greene 2001). We thus restrict to patent examiners with at least ten applications; note that in the full sample of data this restriction will likely be less binding. In this sub-sample of data this restriction reduces our number of examiners to 181 (n for this sub-sample is 3,447). On average, each of these remaining examiners reviews twenty-one patent applications.

In our preliminary analysis we estimate the following regression using patent application-level data for a patent application i examined by patent examiner e, filed in year t, in Art Unit a:

$$(0/1: =1 \text{ if a patent granted for application } i)_i = \delta_e + \gamma_t + \varphi_a + \varepsilon_i$$

where the δ_e is a fixed effect for the patent examiner to which the patent application was assigned, γ_t is a fixed effects for the year the patent application was filed, φ_a is a fixed effect for the Art Unit to which the patent application was assigned (needed for the conditional random assignment assumption), and ε_i is an error term. We cluster the standard errors by patent examiner. Although we will eventually implement the first stage analysis in a number of different ways, the basic test is a joint F-test on the patent examiner fixed effects – which yielded an F-statistic of 759. Almost 70 percent of the individual fixed effects had t-statistics over the conventional threshold for statistical significance (p=0.05). Below is a graph illustrating the empirical distribution of the estimated patent examiner fixed effects. We will implement a number of refinements to this rough initial analysis, but these preliminary results are suggestive that this research design will yield informative results that will form the basis for a first paper from this project.



Requested funding for this proposal

We are applying for research funds to support both data acquisition and research assistance support. The major anticipated data acquisition cost is the cost of digitizing some of the USPTO PAIR data; although much of this data has been digitized by Google, that effort is still in progress and does not cover all of the patent applications in our sample. Our estimated cost of digitizing the remaining data is $7,000. Research assistance support will cover an undergraduate student to help with writing scripts to parse and clean the datasets used in this analysis.