

**PATENT METRICS:
THE MISMEASURE OF INNOVATION IN THE BIOTECH PATENT DEBATE**

David E. Adelman^{*} and Kathryn L. DeAngelis[†]

The view that biotechnology patenting has reached unsustainable levels is well accepted among many legal scholars. This Article presents the first comprehensive empirical study of biotechnology patents designed to test this hypothesis. Our analysis reveals the striking rise and fall in biotechnology patenting, the surprisingly diffuse pattern of patent ownership, and the consistent influx of new entrants conducting biotechnology research and development. This Article finds little evidence that the rise in biotechnology patenting is adversely affecting innovation. Counting patents, as it turns out, offers few insights on its own. One must also have a measure of the geographic scope of the scientific commons and the distribution of patents within it. These findings lead to a cautionary corollary for patent metrics generally—certain fundamental uncertainties associated with the statistics of innovative success cannot be overcome using even the most sophisticated empirical methods. Ironically, the current enthusiasm for empirical work may have caused academics to reify simple patent metrics over the manifest complexity of innovative processes.

TABLE OF CONTENTS

I.	Introduction.....	1
II.	Searching for Signs of a Biotechnology Anticommons.....	6
A.	The Rise and Fall of Biotechnology Patenting	10
1.	Tipping Points: Lagging PTO Resources and Shifting Patent Terms	13
2.	(Relatively) Low Patent Rates for Genes and Proteins	16
B.	Diffuse Patent Ownership and Small Patent Portfolios	19
C.	Overlooking Scale and Distribution in the Anticommons Debate.....	22
1.	Biotechnology's Uncongested Commons.....	24
2.	The Multiple Dimensions of Biomedical Science.....	26
D.	Towards Reliable Metrics for the Scientific Commons.....	31

^{*} Associate Professor, James E. Rogers College of Law, University of Arizona. This Article greatly benefited from the suggestions, comments, and criticism of Graeme Austin, John Barton, Mark Lemley, Greg Mandel, Rob Merges, Michael Meurer, Marc Miller, Barak Orbach, F.M. Scherer, and Kathy Strandburg.

[†] Graduate, James E. Rogers College of Law, 2005.

III.	The Statistical Insignificance of Patent Metrics as Predictive Tools	33
A.	Skewed Distributions and Uncharacteristic Patents	35
1.	<i>Mismeasures of Biotechnology Patents: Prosecution Times, Number of Claims, and Citation Counts</i>	<i>35</i>
2.	<i>An Empirical Haze in Common</i>	<i>41</i>
B.	The Elusive Search for Metrics of Innovative Success	44
1.	<i>Statistical Significance Without Predictive Power</i>	<i>48</i>
2.	<i>Base Rates and the Slim Tail of Innovative Success</i>	<i>54</i>
IV.	Conclusions	59
V.	Figures.....	61
	Appendix: Data and Methodology	75

I. Introduction

The debate over the dramatic rise in patents issued each year in the United States springs from a deceptively simple question: Do we have too much of a good thing? For many commentators, the numbers speak for themselves.¹ As Jon W. Dudas, Director of the Patent and Trademark Office (“PTO”), remarked recently, “the PTO issued more patents last year [] than it did during the first 40 years of its existence.”² Yet, the economic theory that frames this debate, the tragedy of the (anti)commons, is a matter of relative scale—fifty patents distributed over a narrow field of invention may be grounds for concern whereas fifty patents of analogous scope scattered over a broad field will not.³ Large numbers alone are thus meaningless. One must have a measure of the geographic scope of the relevant scientific commons and the distribution of patents within it to assess the impact of rising patent numbers.

Patent metrics are central to the debate over the 1990s patent bubble. The simplest metric, patent counts, has dominated theorizing about the effects of the extraordinary rise in U.S. patenting on innovation. The most prominent example of this approach, Michael Heller and Rebecca Eisenberg’s beguilingly elegant “antcommons” theory, posits that transaction costs spiral out of control as patent numbers increase.⁴ Other patent metrics, such as the number of claims or citations in a patent, are also important, most notably as indicators of patent value.

Commentators have proposed that patent characteristics be used by the PTO as metrics for

¹ See, e.g., Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 SCIENCE 698, 698 (1998) (arguing that the recent surge in patenting will harm innovation by creating “anti-commons” that threaten innovation by fatally raising the transaction costs of research and development); Rebecca S. Eisenberg, *A Technology Policy Perspective on the NIH Gene Patenting Controversy*, 55 U. PITT. L. REV. 633, 640 (1994).

² *The Patent System Today and Tomorrow* Before Subcommittee on Intellectual Property of the Senate Committee on the Judiciary, 109th Cong. 3 (2005) (statement of John W. Dudas, Under Secretary of Commerce for Intellectual Property and Director of the U.S. Patent and Trademark Office).

³ Carol Rose, *Rethinking Environmental Controls: Management Strategies for Common Resources*, 1991 DUKE L.J. 1, 5-7 (arguing that government-imposed restrictions are not warranted when a commons is uncongested); David E. Adelman, *A Fallacy of the Commons in Biotech Patent Policy*, 20 BERKELEY TECH. L.J. 985, 1021-23 (2005) (arguing that the tragedy of anticommons also disappears if a commons is uncongested).

⁴ Heller, *supra* note 1, at 698.

enhancing the efficiency of PTO reviews and reducing the burden of rising patent application numbers.⁵ Patent metrics thus play an integral role in efforts to diagnose and mitigate the impacts of the recent surge in U.S. patenting.

This Article challenges the widely held belief that the rapid growth in biotechnology patenting over the last decade is impeding innovation. It argues that the misuse of patent metrics has both fostered dire predictions and created unrealistic expectations about the capacity of patent data to guide policy. The Article's focus on biotechnology is motivated by its singular role in the debate—for many commentators, biotechnology is the proverbial canary in the coal mine signaling the need for major reforms.⁶ The unique status of biotechnology follows from the extensive overlap between public and private biotechnology research, the importance of patents to the biotechnology industry, and the high social value accorded to biomedical research.⁷ These traits have dramatized the tensions between the ownership model of patent law and the open-access principles of science.

The centerpiece of the Article is a comprehensive empirical study of biotechnology patenting in the United States.⁸ Contrary to much legal scholarship, the study finds little evidence that the recent growth in biotechnology patenting is threatening innovation. This

⁵ John R. Allison, *et al. Valuable Patents*, 92 GEO. L.J. 435, 439, 464-65 (2004).

⁶ Arti K. Rai & Rebecca S. Eisenberg, *Bayh-Dole Reform and the Progress of Biomedicine*, 66 L. & Contemp. Probs. 289, 291 (2003).

⁷ James Bessen & Michael J. Meurer, *Lessons for Patent Policy From Empirical Research on Patent Litigation*, 9 Lewis & Clark L. Rev. 1, 13 (2005) (observing that biotechnology is uniquely important to universities and is the primary field in which technology transfer is heavily reliant on patents).

⁸ Very few data of any kind were available until the late 1990s when several major studies were initiated, including the following: Bronwyn H. Hall, *et al.*, *The NBER Patent Citations DATA File: Lessons, Insights and Methodological Tools*, Working Paper 8498, available at <http://www.nber.org/papers/w8498>; John R. Allison & Mark A. Lemley, *Who's Patenting What? An Empirical Exploration of Patent Prosecution*, 53 VAND. L. REV. 2099 (2000); Kimberly A. Moore, *Judges, Juries, and Patent Cases—An Empirical Peek Inside the Black Box*, 99 Mich. L. Rev. 365 (2000-01); Jonathan A. Barney, *A Study of Patent Mortality Rates: Using Statistical Survival Analysis to Rate and Value Patent Assets*, 30 AIPLA Q.J. 317 (2002); Jean O. Lanjouw & Mark Shankerman, *Enforcement of Patent Rights in the United States* in PATENTS IN THE KNOWLEDGE-BASED ECONOMY 145 (Wesley M. Cohen & Stephen A. Merrill, eds., 2003); Allison, *supra* note 5, at 435; Josh Lerner, *Patenting in the Shadows of Competition*, 38 J.L. & ECON. 463 (1995).

analysis is based on a dataset comprised of biotechnology patents granted in the U.S. from January 1990 through December 2004, more than 52,000 patents in all.⁹ The years encompassed by the dataset offer a rich context for examining patent policy. The data cover the period of the most dramatic rise in biotechnology patenting, important shifts in PTO policy towards more stringent standards for obtaining patents on genetic sequences, and the dramatic growth followed by a significant retrenching of the biotechnology financial markets during the late 1990s.¹⁰

The study examines several categories of data, including investigations of broad patent trends, patterns of patent ownership, and the distribution of patents across PTO patent subclasses.¹¹ One of the Article's most significant findings is the degree to which ownership of biotechnology patents is diffuse. Even the largest companies, on average, are granted fewer than thirty biotechnology patents per year, and the number of entities obtaining biotechnology patents has consistently increased over the fifteen years covered by the dataset.¹² Interpreting these trends is necessarily impressionistic, but the lack of concentrated control, rising number of patent applications, and the continuous record of new market entrants provide strong evidence that biotechnology patenting is not adversely affecting innovation.

These results are consistent with the growing schism between current theorizing about biotechnology patenting and recent empirical work. The existing empirical studies find few clear signs that the patenting of biotechnology inventions is adversely affecting biomedical

⁹ The biotechnology patent database consists of 52,039 patents that issued between January 1990 and December 2004. These patents were collected from a larger patent database consisting of all patents, more than 950,000 in total, that issued between January 1990 and March 2005. *See infra* Appendix A.

¹⁰ Ashish Arora, *et al.*, MARKETS FOR TECHNOLOGY: THE ECONOMICS OF INNOVATION AND CORPORATE STRATEGY 67 (2001) (observing that equity financing of biotech companies declined significantly between 1997 and 1998); *see infra* Part II.A. (describing the PTO changes to its regulations regarding certain types of biotechnology patents during the 1990s).

¹¹ The PTO has an elaborate system for categorizing patents into 400 broad classes and more than 120,000 subclasses, which evolve and develop over time with technological advances. Hall, *supra* note 8, at 13.

¹² These numbers are tiny in comparison to the massive patent portfolios being amassed in the information technology and electronics industries, where companies like IBM Corporation obtain more than one thousand patents each year. Gideon Parchomovsky & R. Polk Wagner, *Patent Portfolios*, 154 U. Pa. L. Rev. 1, 46-48 (2005) (describing IBM's strategy of obtain upwards of three thousand patents annually).

innovation.¹³ This Article shows that the divergence between data and theory is traceable to theorists' over reliance on patent counts, which prove to be a very weak metric. Heller and Eisenberg's anticommons theory exemplifies this reductivism. They predict that rising patent numbers will create patent anticommons, whose sheer numbers cause the transaction costs for licensing patents to explode.¹⁴ Heller and Eisenberg gloss over the conditions necessary for patent anticommons to emerge, and their vivid metaphors obscure the complexities of interpreting patent-count data. In essence, they offer a one-dimensional model premised on a simple relationship existing between patent counts and transaction costs.

Little reason exists for accepting the highly simplified model implicit in Heller and Eisenberg's theory. Unlike Garret Hardin's uniform agricultural commons for which a simple metric (number of cattle) was available,¹⁵ patent policy must contend with a much more complex environment. Science lacks a unique set of spatial dimensions; its geography is too heterogeneous and multidimensional, and numerous points of reference exist from which to assess the distribution and size of patented enclosures. Further, the scope of patents themselves is both highly variable and exceedingly difficult to quantify. Direct patent counts, as a consequence, are unlikely to be consistently correlated with the fragmentation of the scientific commons on which their theory is predicated.

This Article exposes a distinct set of interpretive challenges for metrics based on patent characteristics. The study data display remarkably broad variances for the number of claims, citations made by or to a patent, and the time spent by PTO examiners prosecuting (*i.e.*,

¹³ See John P. Walsh, Ashish Arora, and Wesley M. Cohen, *Effects of Research Tool Patents and Licensing on Biomedical Innovation*, in PATENTS IN THE KNOWLEDGE-BASED ECONOMY 285 (Wesley M. Cohen and Stephen A. Merrell eds., 2003); John P. Walsh, Charlene Cho, & Wesley M. Cohen, *View From the Bench: Patents and Material Transfers*, 309 SCIENCE 2002, 2002 (2005).

¹⁴ Heller, *supra* note 1, at 698.

¹⁵ Garrett Hardin, *The Tragedy of the Commons*, 162 SCIENCE 1243, 1244 (1968).

reviewing) patent applications. To give just one example, in any given year of our study the number of claims in a patent ranged from one to several hundred. The broad variances we observe effectively rule out using these characteristics to predict other attributes of a patent, such as its technological field, ownership status (*e.g.*, government, corporation), or economic value.

The practical ramifications of this finding are far reaching. Chief among the issues implicated is fashioning an effective response to the dramatic rise in patenting, which has created a large backlog of patents at the PTO.¹⁶ Legal commentators have singled out the substantial majority of patents—some data suggest more than ninety-five percent¹⁷—that have little economic value as a target for economizing PTO resources.¹⁸ Under this scheme, PTO resources would be conserved by subjecting valuable patents to careful scrutiny and all other patents to light reviews. Our analysis shows that this triage scheme is unworkable for the simple reason that valuable patents cannot be identified *ex ante*.

This result has a noteworthy theoretical corollary. The economist F.M. Scherer was one of the first people to show that innovative success is chaotic and that the distribution of valuable inventions defies standard statistical methods.¹⁹ A recent empirical study, “*Valuable Patents*,” challenges, at least implicitly, the empirical work on which Scherer’s findings are based.²⁰ We reexamine the *Valuable Patents* study in light of our own data and conclude that Scherer’s work

¹⁶ Dudas, *supra* note 2, at 2.

¹⁷ Mark A. Lemley, *Rational Ignorance at the Patent Office*, 95 NW. U. L. REV. 1495, 1507 (2001) (estimating that “the total number of patents litigated or licensed for a royalty (as opposed to cross-license) is on the order of five percent of issued patents”).

¹⁸ See, *e.g.*, Allison, *Valuable Patents*, *supra* note 5, at 439, 464-65.

¹⁹ Fredrick M. Scherer, Firm Size, Market Structure, Opportunity, and the Output of Patented Inventions, 55 AM. ECON. REV. 1098, 1098 (1965); Fredrick M. Scherer & Dietmar Harhoff, *Technology Policy for a World of Skew-Distribution Outcomes*, 29 RESEARCH POLICY 559, 563 (2000) (the innovation lottery “implies that it is difficult or impossible to achieve stable mean expectations and hence to hedge against risk by supporting sizeable portfolios of projects”).

²⁰ Allison, *Valuable Patents*, *supra* note 5, at 462 (claim that “if valuable patents can be reliably identified at the time of application, or at least at the time of issue, the lottery theory runs into difficulty. At best, it becomes only a partial explanation”).

ultimately withstands the challenge. Regrettably, this result implies that one cannot avoid the statistical uncertainties exposed by Scherer and other economists.

The Article is divided into two central sections and an appendix that describes the data and analysis in detail. Part II examines the general trends in biotechnology patenting, including patent counts, patent ownership patterns, and the distribution of biotechnology patents across distinct areas of research and development. This analysis finds few tangible signs of the negative impacts presumed to be associated with patent anticommons. Part III assesses the characteristics of biotechnology patents and evaluates our results comparatively with three recent studies of U.S. patents. This analysis reveals the methodological obstacles impeding empirical work on patents—the uncertainties prove to be greater than Scherer estimated. The Article concludes with a short discussion of opportunities for refining current empirical methods.

II. Searching for Signs of a Biotechnology Anticommons

The spectacular rise in patenting since the mid-1980s has inspired a flurry of writing by scholars concerned about its negative impacts.²¹ Critics of the current patent bubble, led by Michael Heller and Rebecca Eisenberg's work, have cultivated a potent metaphor—the anticommons—to illustrate how patents can deter innovation.²² Consistent with the image evoked by the metaphor, these critics argue that expansive patenting fragments the scientific commons such that no single entity has sufficient patent stock to pursue its program of research and development.²³

²¹ Wagner & Parchomovsky, *supra* note 12, at 17-18 (describing the rise in patent applications and intensity in the United States).

²² The standard model for the public commons is an area (*e.g.*, public lands, body of water) that is vulnerable to overexploitation by multiple actors because none of them bears the full impact of poor management. By contrast, a patent anticommons impedes development because narrow patent rights are dispersed among different entities too broadly. See Heller, *supra* note 1, at 698; Eisenberg, *supra* note 1, at 640.

²³ *Id.* Making matters worse, Heller and Eisenberg argue that transaction costs are exacerbated by certain strategic behaviors and cognitive biases. Heller, *supra* note 1, at 698, 700 (Licensing will fail because: transactions

Heller and Eisenberg draw on experience in the biomedical sciences to identify two scenarios in which patents unduly increase the transaction costs of research and development. In the first, the presence of numerous patents owned by different entities places a prohibitive burden on a scientist or company to negotiate licenses to patented technologies.²⁴ In the second, patents on numerous “upstream” technologies, or research tools, “act like tollbooths on the road to product development, adding to the costs and slowing the pace of downstream biomedical innovation.”²⁵ The direct signs that (presumably numerous) localized anticommons are burdening a scientific field will thus be reduced scientific output (*e.g.*, papers, patents, data) and rising patent licensing and equipment costs. Indirect effects may include reduced private sector investment, diminished entry of new scientists and companies, and increased concentration of patent ownership (*i.e.*, patent portfolio races).²⁶

Despite the widespread attention that the anticommons theory has garnered, few empirical studies are available either to confirm or refute its predictions. The best studies have been conducted by social scientists, who have surveyed scientists working in the public and private sectors about the impacts of patents on their work.²⁷ These studies have found little clear evidence that patenting of biotechnology inventions is impeding biomedical innovation.²⁸

costs high, patents matter more in biomedical science (only drugs), lack of substitutes, strategic behavior, cognitive biases; skeptical about patent pools emerging).

²⁴ Heller, *supra* note 1, at 698.

²⁵ *Id.* at 699. Critics are particularly concerned about the negative effects of patents on research tools, which stand to hinder subsequent research dependent on them and could lead to monopoly control of key research tools. See, *e.g.*, Heller, *supra* note 1, at 698; Rai, *supra* note 6, at 295-96.

²⁶ Josh Lerner, *Patenting in the Shadow of Competitors*, 28 J. LAW & ECON. 463, 464-65 (1995) (empirical study claiming to show that firms avoid working in areas of biotechnology research that are the subject of numerous patents); Bronwyn H. Hall & Rosemarie Ham Ziedonis, *The Patent Paradox Revisited: An Empirical Study of Patenting in the U.S. Semiconductor Industry, 1979-1995*, 32 RAND J. ECON. 104, 102, 110, 121 (2001) (describing patent portfolio races as a byproduct of the patent anticommons characteristic of the semiconductor industry).

²⁷ See Walsh, *supra* note 13, at 285; Walsh, *View From the Bench*, *supra* note 13, at 2002.

²⁸ *Id.*

Setting aside often-repeated anecdotal evidence,²⁹ no empirical studies of biotechnology patenting have been conducted to assess whether the anticommons scenarios predicted by Heller and Eisenberg are representative of the biomedical science commons.

The lack of supporting empirical work has both conceptual and practical origins. From the outset of the debate, the intuitive appeal of the anticommons theory has obscured the limited value of patent counts in determining whether an anticommons exists. This perspective marginalizes the denominator—the scope of the field of invention at issue—by focusing attention on seemingly impressive patent statistics. In so doing, the theory avoids the far more difficult problems entailed in defining the relevant fields of invention and identifying accurate metrics for the scope of the patents at issue.

To appreciate the significance of this oversight, it is useful to return Garret Hardin’s seminal theory on the tragedy of the commons. Hardin’s economic parable describes how individual self interest leads inexorably to unsustainable exploitation of resources held in common.³⁰ When applied to intellectual property, however, Hardin’s storyline inverts because intellectual property is an inexhaustible, non-excludable resource. The problem thus becomes inadequate *internalization of positive* externalities and the tragic result *under* exploitation of the resource. The traditional remedy in both cases is privatization.

Hardin’s theory has important shortcomings that derive from his simple model of the commons as a homogenous resource used by several operators for the same purpose.³¹ Subsequent commentators have exposed the limits of his model by describing counter examples

²⁹ See, e.g., Heller, *supra* note 1, at 699 (search of Lexis database that found 100 patents with the term “adrenergic receptor” in their claims); Rai, *supra* note 6, at 293-93 (discussing the Wisconsin patents on an important line of stem cells).

³⁰ Hardin, *supra* note 15, at 1243.

³¹ *Id.*

(e.g., rivers, parks) for which privatization is not the optimal legal regime.³² They show that the physical characteristics of the resource and the ways in which it is exploited, e.g., consumptively versus non-consumptively, have important implications for selecting an efficient legal regime.³³ Above all, these examples demonstrate that policies for managing property, intellectual or otherwise, cannot be assessed in the abstract. One must have a detailed understanding of the characteristics and uses of the underlying resource itself.

Proponents of the anticommons theory either ignore the characteristics of the scientific commons altogether or base their views on questionable assumptions about it, such as that so called upstream patents will inevitably restrict access to essential research tools for which no alternatives exist.³⁴ This neglect causes the anticommons theory to be incomplete, if not misleading, and may account for the divergence between the social science data and the dire predictions of its proponents.³⁵

The subsections that follow employ several complementary methods to assess the potential significance of the anticommons theory for biotechnology research and development. The first subsection explores the interplay of patent statistics, shifts in PTO policies, and changing economic conditions. This analysis affords a number of insights into trends in biotechnology patenting and their relationships to external factors. The second subsection evaluates ownership patterns of biotechnology patents and assesses the implications of the small portfolios found for most entities. The third subsection argues that biotechnology, at this stage

³² Carol Rose, PROPERTY AND PERSUASION 140-45, 185-88 (1994); Carol M. Rose, *The Several Futures of Property: Of Cyberspace and Folk Tales, Emission Trades and Ecosystems*, 83 MINN. L. REV. 129, 150-162 (1998) (discussing cyberspace property regimes and arguing for limited-commons property).

³³ Rose, *supra* 32, at 185-88. They have also provocatively demonstrated that efficiency is maximized for certain “inherently public” resources only when they are held in common, a finding that has been enthusiastically embraced by intellectual property scholars aligned with the open source movement. *Id.* at 140-45.

³⁴ See, e.g., Heller, *supra* note 1, at 700 (suggesting that patents often cover research tools for which alternatives do not exist).

³⁵ Although, proponents do concede the difficulty of determining when patents warranted (Rai-Eisenberg, 300; 303) (acknowledge the uncertainty that exists).

of its development, is generally uncongested, but finds that determining whether localized anticommons exist raises irreducibly complex problems for studies based on patent counts. The anticommons theory is shown to be compelling in theory but elusive in practice.

A. *The Rise and Fall of Biotechnology Patenting*

At the broadest level, we find that the number of biotechnology patents issued per year peaked at 5977 patents in 1998 and then declined to 4324 patents (a twenty-nine percent drop) by 2004, *see* Figure 1. The same basic progression—peak in the late 1990s followed by a flattening or significant decline in the number of patents issued—is mirrored in virtually all of the data. This trend is observed for each the five biotechnology subfields we defined, *see* Figure 2; individual PTO subclasses drawn from each of the biotechnology subfields, *see* Figure 3; collectively the thirty PTO subclasses with the highest numbers of patents, *see* Figure 4; the four groupings of biotechnology companies we created, *see* Figure 5; and the three categories of assignees (*i.e.*, the federal government, universities, and corporations), *see* Figure 6.³⁶

The other major trends we observe concern assignee types and differences between biotechnology subfields. As one would anticipate, corporate ownership of biotechnology patents dominates, accounting for an average of eighty percent of the patents issued from 1990 through 2004. However, this overall average obscures the growth, in absolute and relative terms, of university and government patenting over this period.³⁷ These gains represent about a ten-fold

³⁶ See Part I of the Appendix for a detailed explanation of each of these data categories.

³⁷ In 1990, university and government patenting accounted for fifteen percent of biotechnology patents, but by 1994 they accounted for twenty percent, and this has remained at that level ever since. Our results differ somewhat from the results of Walsh, *et al.*, who found that the university share of the patents issued in three key biomedical utility classes increased from eight to twenty-five percent between the early 1970s and the mid-1990s. Walsh, *supra* note 13, at 295.

increase in patents issued to universities and the federal government between 1990 and 1998-99, although the peak numbers were followed by a decline of more than twenty-five percent.³⁸

Contrary to our expectations, the division of patents between the five biotechnology subfields is similar for corporations, universities, and the federal government, *see* Figure 7. Corporations patent more in the proteins subfield and less in the nucleotides and immunological subfields, but the differences are nominal.³⁹ Patenting of genetically modified organisms is the one area of substantial divergence. Universities and the federal government received fully twenty-nine percent of the patents, but the absolute numbers are quite low.⁴⁰ We anticipated that corporations would receive a higher proportion of patents covering measuring and testing processes, as one might expect corporations to focus on applied work, but no such differentiation between basic and applied patenting is observable in our data. This finding adds credence to concerns that universities are very actively patenting biotechnology research tools.

The observed rise and fall in the number of biotechnology patents issued is consistent with the anticommons theory. One could interpret the falloff in patents issued after 1999 as a drop in innovative output brought about by the fragmenting effects of thousands of patents and growing patent tolls on research and development. The turnaround in 1999 thus could represent a dramatic tipping point beyond which spiraling licensing costs outweighed the incentives patents provide.

³⁸ Universities obtained fewer than one hundred biotechnology patents in 1990, reached a high of 1101 patents in 1999, followed by decline to 782 patents through 2004. The federal government was issued 19 biotechnology patents in 1990, the number peaked at 159 in 1998, and leveled off at 118 by 2004.

³⁹ The proportions are as follows: (1) measuring and testing processes – corporations 80%, universities and the government 20%; (2) protein and polypeptide sequences – corporations 84%, universities and the government 16%; (3) nucleotide sequences – corporations 76%, universities and the government 24%; and (4) immunochemical inventions – corporations 75%, universities and the government 25%.

⁴⁰ This difference could be attributable to the early stage of the technology or possibly to the adverse politics surrounding genetically modified organisms. As described in the Appendix, we generally excluded agricultural biotechnology patents and specifically patents on genetically modified plants. *See infra* note 218.

A closer examination of the data quickly reveals the deficiencies of this simple storyline. For one, the number of applications received (as opposed to issued) by the PTO for biotechnology patents rose substantially post-1999.⁴¹ This observation on its own suggests that the anticommons theory cannot be accepted before eliminating several alternative explanations.

Changing economic conditions are an obvious factor to evaluate. The most dramatic decline over the fifteen-year period of the study took place in 2001,⁴² but this follows the peak in biotechnology patenting by more than a year, *see* Figure 1. The 1998 slump in equity financing of biotechnology companies is more promising,⁴³ as it aligns with the peak in biotechnology patenting. It also runs into trouble, though. A narrow market contraction cannot explain the concurrent drops in the number of biotechnology patents issued to universities and the federal government.⁴⁴ More importantly, research and development (R&D) funding in the private and public sectors continued to grow after 1998. The largest increases in R&D funding among public companies occurred after 1999,⁴⁵ and federal funding for the life sciences climbed steadily through 2004.⁴⁶ These trends are at odds with economic studies linking growth in R&D

⁴¹ After 1999, the number of biotechnology patent applications filed with the PTO increased by about forty percent. Dudas, *supra* note 2, at 2. Note that we are assuming here that the PTO's definition of what falls within the scope of biotechnology is reasonably consistent with our definition. This is not unreasonable given that we checked our subclasses against assigned to the PTO's biotechnology art unit.

⁴² Robert Garvin, *US Pace of Job Growth Slows Questions are Raised on Recovery's Strength*, BOSTON GLOBE, July 3, 2004, at A1 (stating that "the nation has regained about 1.5 million of the 2.7 million jobs lost in the economic downturn that began in early 2001.").

⁴³ The biotechnology equity financing market was \$8 billion in 1997 and declined to just \$5.5 billion in 1998. Arora, *supra* note 10, at 67. Econometrics analyses have found that the number of patents issued by the PTO is sensitive to business cycles. Zvi Griliches, *Patent Statistics as Economic Indicators*, 28 J. ECON. LIT. 1661, 1693 (1990); Josh Lerner & Robert P. Merges, *The Control of technology Alliances: An Empirical Analysis of the Biotechnology Industry*, 46 J. IND. ECON. 125, 126-7 (1998).

⁴⁴ In addition, given the more than two-year delay between the filing and issue dates of a patent, one would expect a significant lag-time to exist between declining economic conditions and changes in the number of patents granted.

⁴⁵ In the private sector, public companies increased their biotechnology R&D by about \$3.7 billion dollars (from \$7 to 10.7 billion) between 1994 and 1999, but their R&D funding was up almost another \$10 billion by 2002, and appears to have leveled off at close to \$20 billion annually. Stacy Lawrence, *Biotech Drug Market Steadily Expands*, 23 NATURE BIOTECH 1446 (2005).

⁴⁶ It is difficult to obtain the levels of federal R&D funding for biotechnology specifically. However, data reveal are available on federal funding for the life sciences and on funding for the National Institutes of Health

funding to rises in the number of patent issued.⁴⁷ Faltering economics conditions are therefore unable to account for the late-1990s decline in biotechnology patenting.⁴⁸

1. Tipping Points: Lagging PTO Resources and Shifting Patent Terms

The PTO's decision to strengthen the utility requirements in 1999 stands out as a legal reform that could explain the dramatic leveling off of biotechnology patenting.⁴⁹ This rule change reversed the PTO's 1995 decision to liberalize its utility guidelines, which led to much freer granting of patents on DNA and polypeptide sequences.⁵⁰ Both the relaxing of the utility doctrine in 1995 and the PTO's reversal in 1999 correlate well with the dramatic rise in biotechnology patenting that began in 1994 and the leveling off (or decline) that occurred in 1998-99.⁵¹ The obvious defect in this explanation is that the changes in the utility doctrine target

(NIH), which is the primary source of government funding biotechnology R&D. Funding for the life sciences increased from \$15.4 billion in 1999 to \$29.3 billion in 2004, and NIH funding rose from \$13.0 billion in 1999 to \$26.9 billion in 2004. National Science Foundation, *Federal Funds for Research and Development: Fiscal Years 2002, 2003, and 2004* 251, 311 (Feb. 2005) available at <<http://www.nsf.gov/statistics/nsf05307/pdfstart.htm>>. Detailed accounts of the NIH budget show funding for biotechnology specifically increasing during this period. NIH, *Estimates of Funding for Various Diseases, Conditions, Research Areas* available at <<http://www.nih.gov/news/fundingresearchareas.htm>>.

⁴⁷ Griliches, *supra* note 43, at 1673, 1684 (stating that a strong relationship between R&D levels and patent numbers). For biotechnology, the subsidy is estimated to be twenty-five percent, that for every four dollars spent on R&D, on average, one would expect one patent to issue. Jean O. Lanjouw & Mark Schankerman, *Characteristics of Patent Litigation: A Window on Competition*, 32 RAND J. ECON. 129, 130 (2001).

⁴⁸ No obvious evidence exists of scientific factors contributing to the decline in the number of patents issued. For one, the number of patent applications continued to increase through this period. Dudas, *supra* note 2, at 2. More importantly, many major successes of the Human Genome Project occurred in the period after 1998 and important developments have continued since. See Lesley Roberts, *et al.*, *Timeline: A History of the Human Genome Project*, 291 SCIENCE 1195, 1195 (2001) (describing the important developments in the biological sciences up through 2003). It is, however, notable that the difficulties of translating biotechnology methods into products have become increasingly apparent over time. See Robert F. Service, *Surviving the Blockbuster Syndrome*, 303 SCIENCE 1797, 1799 (2004) ("The plain truth is that many of the most dramatic scientific advances that have recently been made in the lab have not transformed medicine.").

⁴⁹ Heightened utility standards first promulgated in an interim form in 1999 and later finalized in 2001. See Revised Interim Utility Examination Guidelines, 64 Fed. Reg. 71440 (1999); Utility Examination Guidelines, 66 Fed. Reg. 1092 (2001).

⁵⁰ See Utility Examination Guidelines, 60 Fed. Reg. 36263, 36264 (1995) (requiring the utility of an invention to be "credible" or "well established"). The rule change was motivated by patents on DNA sequences that were being sought when no knowledge of their actual function was known. *Id.* The primary effect of the rule was to preclude patents on DNA or protein sequences lacking any well-defined function or utility. *Id.*

⁵¹ Data on specific PTO subclasses that cover protein sequences, polypeptides, DNA/RNA fragments, and nucleotide sequences follow a similar path of leveling off or decline, see Figure 2. Interestingly, we do not observe a perfect alignment with the shift in policy. In at least one important subclass, large polypeptides, the decline

a subset, albeit an important subset, of biotechnology inventions.⁵² Yet, the same trend is observed for biotechnology inventions that do not directly claim isolated nucleotide or polypeptide sequences, *see* Figure 2.

On its face, the narrow scope of the change in the utility guidelines suggests that it cannot possibly explain the rise and fall in biotechnology patenting that occurred during the 1990s. However, if the PTO's move first to liberalize and then to strengthen its utility guidelines represented a general policy to relax standards followed by a renewed vigilance in reviewing biotechnology patent applications, it could explain, at least partially, the late-1990s inflection. Moreover, if these changes did in fact reflect a broader shift in PTO policy, more stringent review of biotechnology patents would likely impact the average time for patent prosecutions and the rate of patent application denials, both of which are measurable.

We do in fact observe a significant increase (about a year) in the average prosecution time during this period.⁵³ Evidence also exists that denials of biotechnology patents increased after 1999—the number of biotechnology patent applications filed with the PTO increased by about forty percent while the number of biotechnology patents issued declined by almost thirty percent.⁵⁴ The striking temporal correlation and apparent rise in the number of application denials for biotechnology patents support an inference that the shift in PTO policy contributed to the observed decline in the number of biotechnology patents granted.

actually precedes the change in PTO policy, and in another, polynucleotide sequences, the decline does not occur until 2002. Both of these deviations suggest that other factors must also be important, such scientific or economic factors that altered the value of such patents. The delayed impact on nucleotide sequences patents is particularly anomalous, though, given that once the PTO policy went into effect, it presumably affected all pending patents.

⁵² This change in PTO guidelines is also confounded by the 1997 case *Regents of the University of California v. Eli Lilly & Co.*, 119 F.3d 1559 (Fed. Cir. 1997), which substantially narrowed the scope of patents on proteins and nucleotide sequences. *Id.* at 1562. This substantial narrowing of scope would have two counterbalancing effects. On one hand, it necessitated inventors obtaining more patents, as a single patent could no longer cover a broad class of related compounds; on the other, it lowered the effective value of an given protein or nucleotide patent.

⁵³ *See infra* Part II.A.

⁵⁴ Dudas, *supra* note 2, at 2.

Two transient factors, however, appear to be much more important influences on the declining patent numbers. First, the change in June 1995 from the seventeen-year patent term to the twenty-year term caused a large jump in biotechnology patent applications—4602 applications were filed in 1994, 7626 in 1995, and 4045 in 1996.⁵⁵ This discontinuity magnified the rise in patents issued during the late 1990s and, as a one-time infusion of applications, contributed to the leveling off and decline observed post-1998.⁵⁶

Second, the falloff in biotechnology patents issued likely reflects a saturation of examiner resources. In other words, the PTO cannot process more biotechnology patents, such that the rate-limiting step in issuing patents is no longer inventive output but the PTO itself. Tellingly, this predicament is not unprecedented. In the 1970s, the leveling off and slight decline observed at that time in the total number of patents issued was traced back to inadequate PTO resources.⁵⁷ Circumstantial evidence also supports this theory now. The existence of a resource shortfall is consistent with recent calls from the PTO for more resources and growing alarm about the backlog of patents waiting for review.⁵⁸

Several factors undoubtedly account for the decline in biotechnology patents issued each year. Although our data cannot resolve the key causative factor unequivocally, we suspect that the shortage in PTO resources is, as in the 1970s, the dominant factor. Even ignoring the much larger total number of biotechnology patent applications filed with the PTO, the highest number

⁵⁵ The rush to file before the shift to the twenty-year term is predictable given the extended development process, particularly the rigorous regulatory hurdles, that occurs after a patent application for most biotechnology inventions is filed. Because of this, the seventeen-year term, which is measured from the patent issue date, typically affords patentees a term that is effectively longer, particularly if the patentee takes measures to prolong the patent review process.

⁵⁶ The jump in 1995 patent applications also aligns well with the peak in issued patents that occurred in 1998, consistent with the two- to three-year time for patent reviews by the PTO. The number of patent applications begins to rise again starting in 1997 and hits a second peak of 6972 applications in 1999. This is probably not a real peak, as our data omit patents that issued after 2004 and aggregate statistics from the PTO indicate that patent applications continued to rise throughout the period covered by our study. Dudas, *supra* note 2, at 2.

⁵⁷ Griliches, *supra* note 43, at 1691 (discussing the gradual realization that the perceived draught in innovation during the 1970s was attributable to a lack of PTO resources).

⁵⁸ Dudas, *supra* note 2, at 2.

of biotechnology patents issued in a year, about 5900, is far lower than the numbers of biotechnology patents filed in 1995, 1999, and 2000 that also ultimately issued. These data suggest that the PTO's maximum review capacity lags current application rates by hundreds of patents, which, consistent with our data, would effect all categories of biotechnology patents. This explanation has the virtue of being simple and, particularly given that biotechnology patent applications have continued to rise, it is also more plausible than the scenario predicted by the anticommons theory.

2. (Relatively) Low Patent Rates for Genes and Proteins

The anticommons debate has fixated on gene patents. Yet, the biotechnology subfield with by far the largest number of patents is measuring and testing processes. This subfield accounts for almost fifty percent of the biotechnology patents granted from 1990 through 2004, and it has the single most-populated PTO subclass (approximately eleven percent of the total), “measuring and testing processes involving nucleic acids.”⁵⁹ Patents specifically directed at nucleotide sequences and genetically modified organisms account for only nine percent and three percent, respectively, of the total.⁶⁰ At the mid-level, patents on protein sequences and immunological inventions account for about twenty-six percent and twelve percent, respectively.

The specific types of inventions that dominate biotechnology patenting include the following:

- Measuring & testing devices (e.g., involving nucleic acids, viruses, bacteria, enzymes, microorganisms)

⁵⁹ More than 6100 patents were granted in PTO subclass 435/6 between 1990 and 2004, and from 1998 through 2004 an average of approximately 600 patents were granted in this subclass each year.

⁶⁰ These numbers do not reflect all patents that cover genetic or protein sequences. Most importantly, patents on measuring and testing processes will sometimes cover genetic and protein sequences. Patents on genetic testing devices, for example, will often cover specific genetic probes. While significant, we believe that these patents are unlikely to cover the most valuable genetic and protein sequences, for it would make no sense for a patentee to limit her patent protection to the use of the sequence in a device and not patent it separately.

- Methods for making proteins & polypeptides⁶¹ (e.g., microorganisms, molecular vectors)
- Polypeptides of various lengths and protein sequences (e.g., enzymes)
- Polynucleotides of various lengths and DNA/RNA fragments⁶²
- Polymerase chain reaction (PCR)-related methods⁶³
- Immunological testing methods (e.g., monoclonal antibody-based methods)

Patents on complete gene sequences are notably absent from this list. Moreover, while gene fragments are among the more highly patented biomolecules, the absolute number of patents issued per year recently has been, on average, fewer than 800 patents.⁶⁴ A similar trend is observed for patents on proteins sequences. Almost 160 patents directed at protein sequences were issued in 1998, although this number has since declined to about 120 patents per year.⁶⁵ To put these numbers in perspective, genetic markers number in the millions,⁶⁶ and scientists estimate that there are likely about one million human proteins.⁶⁷ The number of potential

⁶¹ Proteins are made up of “peptides” and are themselves polypeptides. Bernard R. Glick & Jack J. Pasternak, *MOLECULAR BIOLOGY* 23-26 (1998). For our purposes, a polypeptide is simply a short peptide sequence. *Id.*

⁶² Nucleotides are the building blocks of DNA (*i.e.*, genes) and RNA, which is a chemical cousin of DNA. The primary function of RNA is as a molecular messenger between genes and the cellular constituents that transcribe genes into proteins. *Id.* at 19-22.

⁶³ PCR is the canonical process used for exponentially reproducing DNA sequences. It allows initially small samples of DNA to be increased to much larger samples that can then be analyzed and manipulated. Bruce Alberts, *et al.*, *MOLECULAR BIOLOGY OF THE CELL* 316-17 (3rd ed., 1994).

⁶⁴ This estimate includes the subclass “measuring and testing processes involving nucleic acids,” which is the most clearly relevant and largest subclass that appears to claim nucleotide sequences as part of a broader invention.

⁶⁵ These numbers are under inclusive because they are limited to the PTO subclasses directed specifically at polypeptide (*i.e.*, protein) and nucleotide (*i.e.*, DNA) sequences. We know that a significant number of patents covering polypeptide and nucleotide sequences are categorized in other subclasses, most notably the subclasses for measuring and testing devices. We nevertheless believe that these are the key subclasses for protein and nucleotide sequences that have broad biological importance, as they are sequences that are patented independently of a device, and thus represent the best measure of patent trends on these critically important biomolecules.

⁶⁶ A genetic marker is a genetic variant of a gene that is used to identify a chromosomal region, much as a postal code is used to identify an area in a city or locality. Genetic markers are distinctive (unique) in much the same way a number is; further, millions of genetic markers have been identified, allowing scientists to develop a relatively high-resolution “map” of the human genome. Eric S. Lander & Nicholas J. Schork, *Genetic Dissection of Complex Traits*, 265 *SCIENCE* 2037, 2037 (1994); Christopher S. Carlson, *Additional SNPs and Linkage-Disequilibrium Analyses are Necessary for Whole-Genome Association Studies in Humans*, 33 *Nat. Genetics* 518, 518 (2003) (estimating that there are more than 2.7 million genetic markers currently available).

⁶⁷ Christopher P. Austin, *et al.*, *NIH Molecular Libraries Initiative*, 306 *SCIENCE* 1138, 1138 (2004) (observing that scientists estimate that there are 20,000 to 25,000 genes in the human genome and that they code for probably more than one million proteins).

polypeptide and protein sequences is certain to be far greater. In this light, the number of gene and protein patents currently being issued appears to be much less threatening.⁶⁸

The dominance of the measuring and testing processes subfield is a constant throughout the fifteen-year period of our study. By contrast, patents on protein and polypeptide sequences experienced almost a fifty-percent drop in their relative share while patents on genetically modified organisms, nucleotide sequences, and immunological processes and compounds almost tripled their share of biotechnology patents. This relative drop in the number of protein and polypeptide patents could reflect a shift in research and development priorities. During this period, dramatic technological advances were achieved in genome mapping and the Human Genome project, which directed substantial funds and scientific talent towards resolving the nucleotide sequence of the human genome, was launched and completed.⁶⁹

Overall, these data contradict fears that upstream patents on gene and protein sequences are privatizing the human genome and proteome. The relatively low numbers of patents on genetic and protein sequences suggest that worries about excessive patenting of genes and proteins maybe overblown. While interpreting the stabilization in the number of patents granted on genetic and protein sequences is difficult, given that limited PTO resources appear to be a major factor, there are other signs—particularly recent announcements of large databases of gene- and protein-sequence data being dedicated to the public domain—that speculative patenting in this area is not wildly proliferating.⁷⁰ More studies and consistent monitoring

⁶⁸ Adelman, *supra* note 3, at 1021-23 (arguing that biotechnology patenting is relatively uncongested and biotechnology science unbounded at this point in its development).

⁶⁹ Francis S. Collins, *et al.*, *The Human Genome Project: Lessons From Large-Scale Biology*, 300 SCIENCE 286, 286 (discussing the scale of human and economic resources that went into the human genome project).

⁷⁰ Recent decisions to dedicate genetic data to the public domain add further evidence against the potential for speculative patenting of genetic and protein sequences to inhibit biotechnology research and development. *See, e.g.*, Andrew Pollock, *Celera to Quit Selling Genome Information*, N.Y. Times, April 27, 2005, at C2.

undoubtedly will be needed before this less-ominous view is likely to be accepted, but the existing trends provide grounds for hope rather than cynicism.

B. Diffuse Patent Ownership and Small Patent Portfolios

One of the most striking features of our data is the large number of entities that own biotechnology patents.⁷¹ Figure 8, which displays the number of assignees over time, mirrors the trend—rapid rise during the mid-1990s followed by a stabilization or decline—observed in the number of biotechnology patents issued each year. Almost identical trends are observed for each of the five biotechnology subfields, Figure 9a, and the data for assignee corporations and universities, Figure 9b. These trends prove that biotechnology patents are spread broadly across an expanding number of patent owners.⁷²

The distribution of patents over a large number of assignees is also reflected in the low averages for the number of patents received annually per assignee.⁷³ The average total number of patents obtained for all three types of assignees is about twelve, or fewer than one per year, and almost fifty percent of assignees obtained no more than twenty-five patents over the fifteen-

⁷¹ Determination of the number of distinct assignees was complicated by the multiple variants of entities' names in the PTO database caused by typos in the filings, name changes over the years, *etc.* For example, Smithkline Beecham appears both as SmithKline Beecham and Smithkline Beecham, which would be treated in our analysis as two different assignees. These differences belied straightforward searches. For the small studies, such as the top thirty assignees, we were able to obtain precise numbers by making sure that we covered every variant of a company's name. For the larger studies, such as the overall average and averages for specific classes of assignees or biotechnology subfields, we identified assignees by the first two words in their name. This approach errs towards indicating a *higher* average than actually exists for these estimates, and therefore provides a conservative estimate, meaning higher average per assignee than actually exists.

⁷² Our data do not take into account circumstances in which the assignee of a patent is a subsidiary off another company that also is the assignee of biotechnology patents. Large biotechnology companies or large pharmaceutical companies active in the biotech area, in some cases may have multiple subsidiaries that appear independently in our database. A review of the top thirty biotechnology companies (based on patents received) found that most biotechnology companies do not have subsidiaries. However, the large pharmaceutical companies represented in our database do have subsidiaries or controlling interests in biotechnology companies (*e.g.*, Novartis has a significant stake in Genentech). In these cases, our data will undercount the number of patents controlled by the parent company.

⁷³ Because of the inconsistencies in the names provided in PTO data we received, this number excludes 2971 patents (*i.e.*, about half of percent of our data) from entities listed as "institutes" or "foundations." Including the patents with uncorrected names has a negligible effect on the overall average. Accordingly, given the small number of patents at issue, we decided not to go through the laborious process of ensuring that the names were all correct.

year period. Even among the top thirty patent owners (excluding the federal government and University of California), assignees obtained on average just 440 patents over the fifteen years of the study, or about twenty-nine per year, and they account for twenty-eight percent of the total patents issued. These numbers are tiny in comparison to information technology and electronics industries, where the largest companies (*e.g.*, IBM Corporation, Canon Kabushiki Kaisha, Samsung Electronics Corporation) regularly obtain more than one thousand patents each year.⁷⁴

We disaggregated the data to evaluate the trends for the university and corporate assignee classes and biotechnology subfields. The highest annual average observed, about 4.5 patents in one year, for corporations occurred in 1998, and the current average is only about three patents per year. The ten corporations with the largest biotechnology patent portfolios obtained an average of just twenty-seven patents per year and have total biotechnology patent portfolios that average 409 patents.⁷⁵ Interestingly, the numbers for our large, successful biotechnology companies are significantly lower, with the average annual rate being about fourteen patents and portfolio sizes averaging 264 patents. We observe a slightly higher concentration of patent ownership among top universities, with the ten largest universities averaging about thirty-one

⁷⁴ See Parchomovsky, *supra* note 12, at 46-48; Press Release, USPTO, USPTO Releases Annual List of Top 10 Organizations Receiving Most U.S. Patents (Jan. 11, 2005), <http://www.uspto.gov/web/offices/com/speeches/05-03.htm>; Press Release, USPTO, USPTO Releases Annual List of Top 10 Organizations Receiving Most U.S. Patents (Jan. 12, 2004), <http://www.uspto.gov/web/offices/com/speeches/04-01.htm>; Press Release, USPTO, USPTO Releases Annual List of Top 10 Organizations Receiving Most U.S. Patents (Jan. 13, 2003), <http://www.uspto.gov/web/offices/com/speeches/03-01.htm>; Press Release, USPTO, USPTO Releases Annual List of 10 Organizations Receiving Most Patents (Jan. 10, 2002), <http://www.uspto.gov/web/offices/com/speeches/02-01.htm>; Press Release, USPTO, USPTO Releases Annual List of Top 10 Organizations Receiving Most U.S. Patents (Jan. 10, 2001), <http://www.uspto.gov/web/offices/com/speeches/01-02.htm>; Press Release, USPTO, IBM Repeats at Top of PTO's Annual List of 10 Organizations Receiving Most Patents (Jan. 11, 2000), <http://www.uspto.gov/web/offices/com/speeches/00-03.htm>.

⁷⁵ We also evaluated the ten pharmaceutical companies (*i.e.*, not solely biotechnology focused) with the largest patent portfolios. The averages for them are almost identical to the biotechnology companies, with the average portfolio size being 384 patents and the average number of patents obtained in a year about twenty-six. Restricting the sample to ten of the largest strictly biotech companies, we find that they account for just 2.9-10 percent (mean 6.6 percent) of the total number of patents in any given year. Similarly, our class of mid-level companies accounted for between 0.2-6.7% or an average 2.6 percent. These data do not preclude instances in which companies obtain a large number of patents in a single year. Genentech holds the single-year record in our dataset with 120 patents issued to it in 1998.

patents per year, although if the two largest multi-campus systems, California and Texas, are removed, the average drops to twenty-three patents per year.⁷⁶

Only the top few assignees have patent portfolios for the fifteen-year period that are quite large. The federal government and the University of California have, by far, the largest numbers with 1322 and 1287 patents, respectively; however, neither is a single entity in the common or practical sense of this term. Consistent with these findings, the average fifteen-year patent portfolio size for corporations and universities are about nine and twenty-four patents, respectively. Similarly, the fifteen-year averages for each biotechnology subfield varies from nine, for measuring and testing process patents, to about three, for genetically engineered organisms.

The average patenting rates for the top assignees provide an important calibration point for understanding the implications of our findings. More than a factor of thirty-five separates the aggregate average from that of the top thirty assignees. But, as we have already noted, even the top assignees obtained modest numbers of patents per year and maintained patent portfolios typically in the range of several hundred patents.⁷⁷ Further, no single entity owns more than a couple percent of the biotechnology patents issued over the past fifteen years. The low aggregate and disaggregated averages we calculate expose the low ownership density of biotechnology

⁷⁶ The top ten universities include University of California, University of Texas, Johns Hopkins University, Harvard University, Columbia University, Washington University, Stanford University, University of Pennsylvania, New York University, Rockefeller University, Thomas Jefferson Medical Center, Yale University, Baylor University, and Duke University. They account for about seven percent (3.8 percent without the two large university systems) of the patents in our database, and the federal government accounts for almost three percent of the total patents issued.

⁷⁷ This point is critical to interpreting our data given the uncertainties involved in deriving the aggregate patent ownership averages. While uncertainties are unavoidable for the broad averages, including those based on assignee type and biotechnology subclass, no such uncertainties exist for the top thirty assignees, for which we have been able to obtain precise data. Accordingly, the highest values we calculate are not subject to any uncertainties and provide an unambiguous upper bound on patent ownership densities.

patents. The modest sizes of current biotechnology patent portfolios suggest that no single entity has the patent capital necessary to dominate biotechnology research and development.

The small patent portfolio sizes also provide indirect evidence that patent anticommons are uncommon in the biotechnology sector.⁷⁸ Where patent anticommons are pervasive, such as the semiconductor industry, they have caused portfolio sizes to balloon.⁷⁹ One would anticipate a similar phenomena to be reflected in our data. Other factors, of course, could be influencing portfolio sizes and large portfolios may not be an inevitable byproduct of patent anticommons.⁸⁰ Nevertheless, an inconsistency exists between the low portfolio sizes observed here and the patenting trends generally associated with patent anticommons that, at the very least, ought to be explained if the anticommons theory were to be accepted.

C. Overlooking Scale and Distribution in the Anticommons Debate

Studies of general patent trends illuminate the aggregate effects of rising biotechnology patent numbers. The question at the heart of this debate—are patents deterring innovation?—ultimately turns on the size and distribution of the scientific spaces enclosed by biotechnology patents. This subsection focuses on these two elements. Drawing on an earlier paper by Adelman,⁸¹ we argue that patents occupy a small region of the biomedical science commons and show that this general lack of congestion mitigates against patent anticommons emerging.⁸²

⁷⁸ Interestingly, the small patent portfolios found in the biotechnology industry appears to be an exception to the predictions made by Gideon Parchomovsky and Polk Wagner, who have argued that patent portfolios are going to be inexorably larger because power synergies that derive from developing a robust patent portfolio. Parchomovsky, *supra* note 12, at 9-10.

⁷⁹ Hall & Ziedonis, *supra* note 26, at 102, 110, 121 (observing that intensive patenting is the norm in the semiconductor industry and often encompass hundreds if not thousands of patents).

⁸⁰ However, to the extent that there are defining properties that distinguish distinct technological fields—semiconductor technologies are incremental and highly interlocked through generic chip fabrication methods—these differences may be central to explaining the absence of patent anticommons. For example, it could be that the research nodes (*e.g.*, specific genetic factors or biochemical processes) for biotechnology are much smaller and more diffuse.

⁸¹ Adelman, *supra* note 3.

⁸² Heller, *supra* note 1, 698; Eisenberg, *supra* note 1, at 640.

The discussion then turns to the distribution of biotechnology patents. It is here that our analysis focuses on the structure of the scientific commons. We begin by examining the heterogeneous and multidimensional character of biomedical science using the PTO's classification system. This analysis shows that no single analytical framework can provide a definitive picture of biotechnology patenting or the distribution of patents within the biomedical science commons. Common objections to the arbitrariness of the PTO classification system⁸³ therefore do not necessarily derive solely from limitations that are unique to the PTO's approach classifying patents.⁸⁴

A point on methodology is worth making here. Two primary approaches exist for analyzing patent data—one narrow and finely grained, the other broad and based on general patent characteristics. Several excellent studies have been conducted using the narrow, finely-grained approach; in the legal literature, most notably by John R. Allison and various co-authors.⁸⁵ Our study adopts the broad approach that is often found in the economics literature, which sacrifices detail for completeness. We opted for this approach because completeness is central to our objective of assessing the potential value of patent counts as a metric for the prevalence of anticommons problems.

⁸³ Allison & Lemley, *Who's Patenting What?*, *supra* note 8, at 2114 (asserting that the PTO's classification of patents often not reliable and, in any case, improperly groups technologies together with disparate characteristics); Griliches, *supra* note 43, at 1667 (describing the inconsistencies in the PTO's classification system).

⁸⁴ In fact, after a detailed review of the 704 subclasses encompassed by our database, we concluded that the PTO's classification system, while admittedly arbitrary in many respects, often had to grapple with classification judgments for which there were no easy or obviously correct solutions.

⁸⁵ See, e.g., *Id.*; Allison, *Valuable Patents*, *supra* note 5; Hall & Ziedonis, *supra* note 26.

1. *Biotechnology's Uncongested Commons*

Proponents of the anticommons theory presume, as they must, that the commons for biomedical science is strictly finite and congested.⁸⁶ Yet, a characteristic of biomedical science that stands out is its unbounded scope. Biotechnology methods have produced vast quantities of genetic data, but scientists have not been able to keep up with the explosion of new information.⁸⁷ The opportunities for biotechnology consequently far exceed the capacities of the scientific community.⁸⁸ Research scientists' accounts affirm this view through their observations that, in the great majority of cases, patents can be avoided by undertaking parallel lines of research.⁸⁹ It is this disparity between resources and opportunities that makes biomedical science an unbounded and uncongested resource.⁹⁰

The complexity of human biology is in large part responsible for the open-ended nature of biotechnology science. Many biological systems have built-in redundancies that protect

⁸⁶ See, e.g., Heller, *supra* note 1, at 698 (describing the tragedy of the anticommons as occurring when “multiple owners each have a right to exclude others from a *scarce* resource and no one has the effective privilege of use”).

⁸⁷ The scientific barriers are exemplified by the declining rate of new drug development over the past decade, despite increased spending (and patenting) by the public and private sectors. See Food & Drug Administration, INNOVATION OR STAGNATION: CHALLENGE AND OPPORTUNITY ON THE CRITICAL PATH TO NEW MEDICAL PRODUCTS 2 (March 2004) <<http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.pdf>>. Service, *supra* note 48, at 1799 (“The plain truth is that many of the most dramatic scientific advances that have recently been made in the lab have not transformed medicine.”); Richard S. Cooper & Bruce M. Psaty, *Genomics and Medicine: Distraction, Incremental Progress, or the Dawn of a New Age?*, 138 ANN. INTERNAL MED. 576, 577 (2003) (“To date, both [gene expression] studies and genome-wide scans have identified only weak and inconsistent genetic signals . . .” for common diseases such as cardiovascular disease and cancer).

⁸⁸ Jenifer Couzin, *NIH Dives into Drug Discovery*, 302 SCIENCE 218, 219 (2003) (describing how drug companies have “barely dipped their toes into the ‘data dump’ that is the human genome sequence”); Walsh, *supra* note 13, at 304-5 (“we have more targets than we have chemists to work on them”). See also Allison Abbott, *Geneticists Prepare for Deluge of Mutant Mice*, 432 NATURE 541, 541 (2004) (describing how there will be a “glut of new mouse strains” for use as experimental models of human disease in the next five to ten years).

⁸⁹ Adelman, *supra* note 3, at 1003 (observing that few, if any, important medical problems are not wide open with regard to opportunities for research and development); Walsh, *supra* note 13, at 304-5, 324 (relating the response of a scientist respond observing that few, if any, common diseases will have only a single drug target and one commented that “we have more [drug] targets than [personnel needed] to work on them”); Walsh, *View From the Bench*, *supra* note 13, at 2002.

⁹⁰ In other contexts, particularly environmental regulation, property theorists have recognized that commons problems do not emerge until a commons becomes “congested,” that is the number of users rises beyond the point of sustainable exploitation of the resource. See Rose, *supra* note 3, at 5-7. The distinction I make here is a simple variation on this basic insight, with the proliferation of patents restricting access to intellectual resources taking the place of the mounting numbers of resource extractors in the typical tragedy of the commons scenario.

against failures of specific processes, and this redundancy is more prevalent the more important the process.⁹¹ Further, diversity is found in the huge range of genetic variants scientists are discovering and the multigenic nature of common diseases.⁹² This complexity belies an atomistic, gene-by-gene analysis of disease processes.⁹³ Common diseases will, as a consequence, be associated with multiple pathways or molecules, implying that most important diseases will have numerous potential drug targets.⁹⁴ Thus, by both affording numerous opportunities for research and a variety of treatment options, the complexity of biological processes reduces the potential for anticommons problems to emerge.

The unboundedness of biomedical science transforms the anti-commons debate. Garret Hardin's original storyline once again illustrates why relative capacity is critically important. Under the traditional commons scenario, individual self interest inexorably leads to overexploitation. An obvious, though often ignored, assumption implicit in Hardin's tale is that individual actions must be capable of collectively overexploiting a resource. If overexploitation is impossible, the tragedy of the commons disappears.⁹⁵ The anticommons theory is subject to the same logic—a large scientific commons with relatively small regions of patenting will afford numerous opportunities for scientists to circumvent potential blocking patents and, in effect, make patent anticommons less problematic and less likely to arise.⁹⁶

⁹¹ For example, DNA repair processes that are central to preventing the onset of diseases like cancer often integrate several parallel functions. Adelman, *supra* note 3, at Part II.B.

⁹² *Id.*

⁹³ *Id.*

⁹⁴ Walsh, *supra* note 13, at 324; Kenneth M. Weiss & Joseph D. Terwilliger, *How Many Diseases Does It Take to Map a Gene With SNPs?*, 26 NATURE GENETICS 151, 153 (2000) (discussing how common disease traits will be associated with many potential candidate genes that scientists will have to investigate). Even relatively simple diseases like malaria reveal complex underlying genetic susceptibilities that offer several potential targets. Adelman, *supra* note 3, at 1010.

⁹⁵ *Id.* at 1021; Rose, *supra* note 3, at 5-7.

⁹⁶ Adelman, *supra* note 3, at 1021-22. In fact, one might reasonably conclude that the dispersal of research activity that follows from extensive patenting could be a positive outcome. *Id.*; see also

This argument requires an important qualification. Biomedical science has numerous sources of heterogeneity, including external market factors, fundamental biological characteristics, and the vagaries of research priorities. At any given time, isolated areas of dense patenting are therefore bound to exist. Our argument does not, and cannot, rule out localized areas of congestion, such as those that could emerge around powerful technologies (e.g., genetic replication methods) or areas with large market potential (e.g., stem cells, diabetes, monoclonal antibodies).⁹⁷ It demonstrates only that the ambient conditions of biotechnology patenting mitigate against patent anticommons being a pervasive problem. Detecting isolated regions of dense patenting requires either narrow studies or a mapping of the distribution of biotechnology patents in the scientific commons, which is the subject of the subsection that follows.

2. The Multiple Dimensions of Biomedical Science

The current focus of the academic debate on patent counts ignores the challenges of identifying reliable metrics for areas of intense patenting.⁹⁸ Unlike patent ownership patterns, for which an obvious base unit exists (the assignee), the very novelty and ill-defined boundaries of most scientific disciplines defy simple conventions. As a consequence, defining the scope of specific fields of invention will generally be far from trivial—one need only consider the Byzantine PTO classification system to appreciate the difficulty of this task.⁹⁹ Typically these obstacles are dealt with by omitting the denominator—the commons itself—from the discussion altogether, which amounts to treating ignorance as a virtue.

The root question is not simply one of classification, but one also of gauging the potential (*i.e.*, currently unrealized) scope of the specific discipline or area of technological development.

⁹⁷ *Id.* at 1023.

⁹⁸ This is not to deny that simple cases will exist, such as the thousands of patents governing virtually every aspect of current semiconductor technologies. Hall & Ziedonis, *supra* note 26, at 104.

⁹⁹ As mentioned above, the PTO classification system has 400 classes and more than 120,000 subclasses, and no obvious rules for systematically adding new classes or subclasses. Hall, *supra* note 8, at 13.

The complexity and rapidly developing nature of biotechnology heighten this uncertainty and complicate the interpretation of our data. For example, at first blush, it would seem that patents are heavily concentrated in certain subfields and subclasses. About three-quarters of the biotechnology patents issued over the past fifteen years fall into our methods (forty-eight percent) and protein sequence (twenty-seven percent) subfields. Similarly, approximately fifty percent of the patents in our study fall into just thirty of the more than 700 PTO subclasses represented in our biotechnology patent database.¹⁰⁰

Yet, in absolute terms, only thirty percent of the PTO subclasses covered by our study have more than one hundred patents, and only about one percent have more than 1000 patents.¹⁰¹ Further, if one considers the top thirty PTO subclasses, on average two-thirds of them increased by fewer than fifty patents per year. In fact, for the vast majority of PTO subclasses, fewer than 100 patents issued within them over the fifteen-year period of our study.¹⁰² Like so many features of inventive activity, the distribution of patents across the PTO subclasses is highly skewed—most subclasses have small populations, while a few super subclasses are vast.

Of the 704 PTO subclasses covered by our study, 113 contain more than one hundred patents. The challenges involved in interpreting these data are illuminated by Table 4, which provides a list of the thirty PTO subclasses with the largest numbers of patents granted between 1990 and 2004. While the PTO classification system lacks the bizarre qualities of the infamous

¹⁰⁰ The top thirty subclasses started at thirty-seven percent of the total in 1990 and gradually rose to about fifty percent of the total by 1996, where they have remained relatively constant.

¹⁰¹ The distribution of patents in PTO subclasses is as follows: two subclasses with more than 2000 patents, six subclasses had 1000-1999 patents, eight have more than 500-999, twenty-seven had 250-499 patents, eighty-five have 100-249 patents, seventy-five have 50-99 patents, ninety-six have 25-49 patents, 165 have 10-24 patents, 103 have 5-9 patents, and 136 have 1-4 patents.

¹⁰² It is important to recognize, however, that the PTO classifications are confounded by the fact that most, if not all, biotechnology inventions can be categorized in more than one PTO subclass, as demonstrated by the PTO's conventional use of a primary and, often numerous, secondary subclasses for each invention. Indeed, the specific subclass into which an invention is placed can be little more than happenstance or, perhaps worse, may be driven by the patentee's strategic decisions regarding the art unit or examiner she would like to review her patent.

Chinese Encyclopedia “The *Celestial Emporium of Benevolent Knowledge*,” in which animals are divided into classes that include “those that belong to the Emperor,” “embalmed ones,” “fabulous ones,” “those that from a long way off look like flies,” and “stray dogs,” to name just a few,¹⁰³ it does not lack for seemingly arbitrary and obviously overlapping categories.

Table 4: The Top Thirty PTO Subclasses for Biotechnology Patents

Subclass	Description of Subclass
435/6	Measuring and Testing (M&T) Processes involving nucleic acids
435/69.1	Recombinant DNA techniques included in methods of making a protein or polypeptide
514/12	Designated Organic Active Ingredient (“DOAI”) containing 25 or more peptide repeating units in a known peptide chain structure
530/350	Protein Sequences having more than 100 amino acid residues
435/71	Microorganisms that make polypeptides or proteins
435/5	M&T Process Involving a virus or bacteriophage
514/2	DOAI -- Peptide containing (e.g., protein, peptones, fibrinogen, etc.)
514 44	DOAI -- Polynucleotide (e.g., RNA, DNA, etc.)
435/320.1	Genetic-engineering vectors, per se
536/23.1	DNA or RNA fragments or modified forms thereof (e.g., genes, etc.)
435/172.3	Defunct category ¹⁰⁴
435/325	Animal cell, per se
536/23.5	DNA or RNA fragments or modified forms thereof that Encode animal polypeptides
435 4	M&T processes involving enzymes or microorganisms
435/252.3	Microorganism transformants (e.g., containing rDNA, vector, foreign or exogenous gene)
436/518	Analytical or immunological testing involving an insoluble carrier for immobilizing immunochemicals
435/7.21	M&T process involving microorganisms with a cell-membrane bound antigen, receptor, antibody, or lysate and animal cell
530/324	Polypeptides having more than 25 amino acid residues
435/194	Enzyme transferases that transfer phosphorus containing groups (e.g., kinases)
514/18	3 or 4 peptide repeating units in known peptide chain
435/240.2	Defunct category
435/91.2	Polynucleotide acellular exponential or geometric amplification (e.g., PCR)
514/19	2 peptide repeating units in known peptide chain
435/29	M&T processes involving microorganisms
514/8	DOAI – Glycoproteins
435/7.23	M&T processes involving animal tumor cell or cancer cell
435/193	Enzyme -- transferases other than ribonucleases
424 851	DOAI – Lymphokines
435/69.7	rDNA techniques included in making fusion proteins or polypeptides
514/11	DOAI -- Monocyclic Peptides (e.g., protein, peptones, fibrinogen, etc.)

¹⁰³ Michel Foucault, *THE ORDER OF THINGS: AN ARCHEOLOGY OF THE HUMAN SCIENCES* XV (1971).

¹⁰⁴ The PTO periodically abandons subclasses as it revises its classification system. These two categories, 435/172.3 & 240.2, were abandoned during the 1990s and the patents in them were reassigned to new categories. The process of resolving these reassignments can be very time consuming in some cases. Accordingly, for a relatively small number of patents, including those in these two categories, we did not attempt to identify the new categories into which they were relocated.

Even this short list of subclasses illustrates the erratic scope of the PTO classification scheme. Almost half of the subclasses describe specific classes of biomolecules, particularly proteins, polypeptides, nucleotides (i.e., DNA and RNA sequences) of certain lengths or having certain characteristics. The scope of many of these molecular classes is enormous and, equally importantly, they are not unique to any particular areas of biotechnology. For example, should issuance of 1462 patents on “protein sequences having more than 100 amino acid residues” over the past fifteen years be troubling when scientists estimate that there are approximately one million human proteins and most of the proteins in this subclass will be biologically unrelated?¹⁰⁵ One could reasonably make the case that we should be more worried about 300-plus patents issued on glycoproteins, which encompass a much narrower, although still broad and important, class of compounds.

The two largest PTO subclasses in our study cover biotechnology methods and their essential materials, such as biomolecules used in genetic engineering, techniques for manipulating DNA, and isolated animal cells and microorganisms. These subclasses are, if anything, more amorphous and potentially broader than those for classes of biomolecules (e.g., polypeptide sequences of a specific length). The subclass “Measuring and testing processes involving nucleic acids,” which dwarfs all other subclasses, is vast and can include anything from commercial tests (e.g., genetic tests for infants, DNA tests used in law enforcement), to genetic tests for microorganisms, to cutting-edge micro-array technologies used in monitoring gene expression levels in living organisms. This single subclass encompasses a vast area of technological development that is applicable to a huge range of biological problems.

It is possible that insights into localized conditions may be gained through case studies of well-defined areas of technological development. We attempted two preliminary studies along

¹⁰⁵ Austin, *supra* note 67, at 1138.

these lines, focusing on patents related to diabetes and HIV Aids research, each of which is at a different stage of research and development. The data are somewhat informative insofar as we found that patenting in these areas encompassed a relatively broad range of technologies, as measured by the number of PTO subclasses in which the patents issued. We observe a close correlation between number of patents issued annually and the number of PTO subclasses they cover, which is evidence that their distribution is diffuse. While these results are enticing, they demonstrate the limits of relying on technology categories and generic patent data. Even knowing the field of invention, it is difficult to assess the significance of patent numbers in each of the PTO subclasses without more detailed information on the specific patents at issue.

The preceding examples illustrate the variable scope of PTO subclasses and the poor basis patent counts provide for making inferences about patent densities.¹⁰⁶ An obvious question raised by these problems is whether it is possible to do better than the PTO. If one simply wishes to improve the logic and rigor of the PTO's classification system, the answer is probably yes. For example, an approach that more systematically draws on established functional taxonomies could be implemented.¹⁰⁷ However, if the objective is to establish a system of categories with carefully calibrated scope, the answer is almost certainly no. The novelty of inventions and evolving, multidimensional nature of science preclude such a grid-like approach, and multiple

¹⁰⁶ It also appears that the PTO classification system, which in effect maps the scientific landscape of biotechnology, does not align well with the structure of biotechnology research and development. This mismatch is plainly evident in the PTO's classification of certain biomolecules (e.g., proteins, polypeptides) by length, as opposed to the area of research (e.g., neurological diseases, specific types of cancer, diabetes) to which they are relevant. Accordingly, patents from multiple PTO subclasses often will map onto a specific area research and development, suggesting that patent densities of specific PTO subclasses really tell you very little about what you care about, namely, the density of patenting in a specific area of research and development.

¹⁰⁷ Even a superficial review of the PTO classes and subclasses reveals that the PTO draws on established scientific nomenclature, but it does so inconsistently because new classes and subclasses are created in a largely reactive fashion. Hall, *supra* note 8, at 13. Just to give one example, eliminating categories based on the size of biomolecules, such as proteins and DNA, would aid efforts to undertake distributional studies.

frameworks (*e.g.*, biotechnology patents could be classified by biochemical pathway, organ, or disease) will always exist for classifying inventions.

Judging whether the volume of patenting is good, bad, or indifferent requires concrete estimates of the distribution and scope of the patents at issue. Our analysis demonstrates the difficulties of interpreting patent-count data given the impossibility of constructing a synoptic quantitative framework for analyzing them. Unsurprisingly, biomedical science bears little resemblance to Hardin's vision of the public commons, with its relatively simple spatial metrics and model of competition. With this added complexity, assessing specific regions in the scientific commons for potential congestion will require detailed analysis of related groups of patents that is clearly beyond the capacities of high-level patent studies and will be far from trivial for even localized areas of research and develop. In the absence of quantitative measures for patent density or distribution, the analysis will largely turn on how patents are categorized, which, as we have seen, will be subject to significant uncertainties and debate.

D. Towards Reliable Metrics for the Scientific Commons

Using patent counts as a metric for gauging patent policy offers important insights when used in conjunction with other information. Our broad surveys of biotechnology patenting provide a surprisingly coherent picture. The lack of concentrated ownership, continuous growth in patent applications, rising research and development expenditures, and consistent infusion of new patent owners all suggest that biotechnology patenting is not adversely affecting innovation. These findings are consistent with the existing social science survey data, which find little clear evidence that patents are interfering with scientists' research work.¹⁰⁸ They are also consistent with our characterization of biotechnology as a still-emerging field in which research

¹⁰⁸ *Id.* at 289 (finding that there is "little evidence of routine breakdowns in negotiations over rights, although research tool patents are observed to impose a range of social costs and there is some restriction of access.").

opportunities far outstrip current scientific resources. Taken together, these observations suggest that anticommons problems ought to be the exception rather than the rule for biotechnology research and development.

Reliance on patent counts runs into trouble when it becomes the sole metric. To date, the debate over the growth in biotechnology patenting has been conducted as though large numbers alone signal the imminent emergence of patent anticommons.¹⁰⁹ In order for the current debate to progress, those concerned that innovation is threatened by expansive patenting will have to move beyond this one-dimensional model. To do so, however, will require commentators to jettison the rhetorically potent metaphors that have framed the debate over biotechnology patenting over the last decade.

It will also, we hope, force academics and policymakers to develop new, more accurate strategies for assessing the impacts of patents on innovation. This could entail, for example, conducting detailed studies using random samples of patents taken from carefully specified research domains. Identifying areas for such focused studies could be based on a combination of factors, including scientific importance, economic potential, and prevalence of patenting in related PTO subclasses. The broad metrics utilized in this study, such as trends in the numbers and types of entities obtaining patents, also provide important information about the dynamics of patenting. Analogizing to analytical strategies used in the ecological sciences,¹¹⁰ which must contend with similar types of complexity, these dynamical measures may prove to be more reliable metrics than simple patent counts. The concluding section of the paper returns to these questions and provides some initial suggestions for future empirical work.

¹⁰⁹ See, e.g., Heller, *supra* note 1, at 699 (using patent counts to suggest that a patent anticommons exists based solely on a general search of patents with “adrenergic receptor” in their claim language).

¹¹⁰ See, e.g., Simon A. Levin, *The Problem of Pattern and Scale in Ecology*, 73 *ECOLOGY* 1943, 1945 (1992) (discussing the value of aggregate metrics that provide measures of the dynamics of ecosystem processes).

III. The Statistical Insignificance of Patent Metrics as Predictive Tools

Only the most churlish cynic could fail to be excited by the wave of empirical studies animating patent scholarship today. Recent studies have transformed our understanding of patent trends in the U.S. and abroad.¹¹¹ Exciting new work combining patent citation data and network theory is being conducted to analyze the relative importance of patents and their relationship to each other.¹¹² More provocatively, a 2004 study by John R. Allison, *et al.*, *Valuable Patents*, purports to have demonstrated that it is possible to identify valuable patents reliably using certain patent characteristics, such as citations received by a patent.¹¹³

On the other hand, new empirical work should not be embraced uncritically. The rising interest in patent metrics among legal scholars follows decades of work by economists,¹¹⁴ who have long viewed patents as a unique, though difficult-to-interpret, source of information on national innovative output.¹¹⁵ Their experience suggests that patent metrics, and particularly patent characteristics, must be analyzed with great care. Economists' efforts have often been frustrated by the highly skewed (*i.e.*, non-bell shaped) distribution of patent characteristics.¹¹⁶ The vast majority of patents, for example, have very low economic value, whereas a small number of patents (the skewed tail of the distribution) account for a disproportionate share of the

¹¹¹ See *supra* n. 8. We have learned that patent litigation grew by almost ten-fold between 1978 and 1999 and gained new understanding about the important class of patents being litigated. Jean Lanjouw & Mark Shankerman, *Enforcement of Patent Rights in the United States* in *Patents in the Knowledge-Based Economy* 146 (Wesley M. Cohen & Stephen A. Merrill, eds., 2003) (also noting that the per-patent rate of law suits has not increased because the number of patents filed rose at a proportionate rate).

¹¹² See Gábor Csárdi, *et al.*, *Modeling Innovation by A Kinetic Description of the Patent Citation System* (submitted for publication) (describing the application of network theory to the study of patent citations).

¹¹³ Allison, *Valuable Patents*, *supra* note 5, at 438, 451, 462.

¹¹⁴ Griliches, *supra* note 43, at 1661-62, 1669-70; Hall, *supra* note 8, at 1-3; Jean O. Lanjouw, *et al.*, *How to Count Patents and Value Intellectual Property: The Uses of Patent Renewal and Application Data*, 46 J. INDUS. ECON. 405, 413 (1998).

¹¹⁵ Lanjouw, *supra* note 114, at 413 (“one of the longest lasting debates in the history of economic measurement has been whether the noise and the biases in patent count measures can be made small enough to make patent counts useful measures of innovative output”); Griliches, *supra* note 43, at 1670; Brownyn H. Hall, *et al.*, *Market Value and Patent Citations*, 36 RAND 16, 16 (2005).

¹¹⁶ A distribution is characterized as skewed because it is asymmetric about the most probable value of the metric, that is the tail of the distribution is substantially longer on one side relative to the other.

wealth generated by innovation.¹¹⁷ This distributional feature of patent data, as we will show, creates significant analytical challenges.¹¹⁸

It is also important to recognize the potential pitfalls of patent studies. Unlike atoms, which have perfectly fungible characteristics, patents resemble living organisms, with their highly variable characteristics that can evolve over time. The value of patented inventions, for example, changes as technological fields develop—advances can lead to improved methods that replace it or synergistic technologies may arise that make it more valuable. While statistical analyses can capture broad trends and important relationships, these numbers represent composite pictures made up of thousands of distinct inventions. The broad, highly skewed distributions found for most patent characteristics evidence this underlying diversity and undermine efforts to use statistical methods to construct predictive models for classifying individual patents (*e.g.*, based on economic value, assignee type, or industry).

The discussion that follows begins by presenting our data and examining its limiting features. This analysis focuses on the statistical trends in three distinct patent characteristics: number of claims, the length of PTO patent prosecutions, and the number of citations made in and received by a patent.¹¹⁹ This work is unique in that it focuses on the statistical limits of evaluating individual patents, as opposed to samples or populations of them. Drawing on these results, we reevaluate the findings of several recent empirical studies, most notably the studies

¹¹⁷ Griliches, *supra* note 114, at 1679-80 1686; Mark Schankerman, *How Valuable is Patent Protection? Estimates by Technology Field*, 29 RAND J. ECON. 77, 79 (1998) (“The distribution of the private value of patent rights is sharply skewed in all technology fields”).

¹¹⁸ Griliches, *supra* note 114, at 1661-62, 1666.

¹¹⁹ The analysis of citation trends utilizes data generated by Bronwyn H. Hall, Adam Jaffe, and Manuel Trajtenberg (herein after “the Hall Study”). Hall, *supra* note 8. The recent wave of empirical work on U.S. patents owes a great deal to their pioneering work, which created the NBER Patent Citations Data File. *Id.* The Hall Study collected every patent issued between January 1963 and December 1999 into a single database and established a new standard for empirical work in the field, particularly with regard to the massive quantity of information that the authors compiled on patent citations. *Id.* Several subsequent studies have utilized the NBER database, including Allison, *Valuable Patents*, *supra* note 5, and Moore, *supra* note 8.

reported in *Valuable Patents* and a recent paper by John R. Allison and Mark A. Lemley.

Contrary to the conclusions of these studies, we find no support for the view that broad statistical regularities in patent metrics have practical utility for predicting patent value.

A. *Skewed Distributions and Uncharacteristic Patents*

Our data suggest that few generalizations can be made about the characteristics of biotechnology patents. With the exception of the recent increases in prosecution times, the data display weak trends, virtually all of which are tiny relative to the broad distribution of the data.¹²⁰ Further, median values for the three patent characteristics differ only slightly between our biotechnology subfields and assignee types. In short, we find little evidence that the characteristics of biotechnology patents vary meaningfully over time or between assignee types and biotechnology subfields.

1. *Mismeasures of Biotechnology Patents: Prosecution Times, Number of Claims, and Citation Counts*

The data on patent prosecution times offer some limited insights.¹²¹ We find that the median patent prosecution time diminished between 1990 and 1994 from 2.7 to 2.1 years and then began to climb (most dramatically since 2002), such that by 2004 the median prosecution

¹²⁰ Interpreting the data is complicated by their broad, often-skewed distributions. In most data analysis, the standard measures used are means and standard deviations, the latter being the average squared difference of the data from the mean. However, means and standard deviations are not informative when the distribution of the data is not “bell-shaped,” typically referred to as “normal distribution,” and the variance of the data is not constant for all values of the data (*i.e.*, the variance must be independent of the magnitude of the statistic). Michael O. Finkelstein & Bruce Levin, STATISTICS FOR LAWYERS: STATISTICS FOR SOCIAL SCIENCE AND PUBLIC POLICY 4-5, 22 (2001).

¹²¹ The distributions of these data are relatively well-behaved for purposes of the statistical analysis, meaning they adequately, though imperfectly, meet the criteria for using standard statistical averages and standard deviations. However, because the distribution are still somewhat skew, and we want to present the data in a consistent format, we have chosen to present the data using “box plots.” A box plot provides direct measures of the center of each distribution, the spread of the center of the distribution, and the extremes of each distribution, and it uses statistics that are insensitive to the skew of the data—medians and quartile plots. Finkelstein, *supra* note 120, at 4-5. Box plots also include averages, but here the difference between the mean and median values is the important measure because it provides a measure of the skew of the distribution. John A. Rice, MATHEMATICAL STATISTICS AND DATA ANALYSIS 372-73 (1995).

time for a biotechnology patent was 3.2 years, *see* Figure 10a.¹²² This rise, which constitutes a forty-three-percent increase in median prosecution time from the low in 1994, is consistent with recent warnings from the PTO that prosecution times and backlogs for patents involving complex technologies have increased substantially.¹²³ Its significance is also borne out by statistical simulations we conducted to obtain a measure of the year-to-year random variation of the mean prosecution time.¹²⁴

Aggregate statistics alone do not provide a complete account of the data. If one is concerned about the implications of the data for individual patents—such as for purposes of discriminating between patents based on prosecution time—one must take into account the exceptionally broad range of prosecution times.¹²⁵ The median values and the quartiles presented in Figure 10a are the relevant factors for this analysis. By these measures, the recent rise in patent prosecution time clearly falls within the box plots for each year, which demarcate the range containing fifty percent of the data. This overlap implies that the observed median values for each year all have at least a fifty-fifty chance of being observed in one year as another. Simply stated, the range of likely prosecution times for a patent are indistinguishable from year-

¹²² *See* Figure 14. The corresponding mean values are 2.8 for 1990, 2.4 for 1994, and 3.7 for 2004, which show less of a drop and more of rise, suggesting the influence of the skew of the data.

¹²³ Dudas, *supra* note 2, at 2. It is nevertheless worth noting that while the number of biotechnology patents grew by 750 percent, the median prosecution time increased by less than forty-three percent. Although dramatic—about a year—in absolute terms, this rise is nevertheless a testament to the PTO’s efforts to keep abreast of the rapid advances that occurred during this period.

¹²⁴ We used the means in this case to estimate the year-to-year variation of the data. Despite the skew of our population data, the robustness of the data set allows an assumption of normality for the annual patent population averages. This assumption was confirmed using a technique called bootstrapping, which involves assessing the distribution of repeated samples of sample size *n* (in our case *n* = 300) collected from within the populations tested for normality. Bootstrapping confirmed that our 30 samples of just 300 patents were normally distributed, therefore allowing us to assume a normal distribution for the means of our yearly populations, for which the minimum sample size is about 700 patents. We calculated the expected variability, formally referred to as “standard error,” to range from 15.5 days in 1990 to 7.0 days in 2004 for samples taken from each of the years covered by our data. The very small standard errors indicate that the aggregate trend, as expressed in the mean value, is clearly beyond expected random fluctuations of the annual averages.

¹²⁵ This is a critical distinction from a statistical standpoint. If the question you are seeking to answer is, what is the prosecution time likely to be if I submitted a new application for a patent, you must use the median and the quartile plots to give you a sense of the likely bounds within which your patent is likely to fall. Note, however, that if the data were not skew, you would use the mean and the standard deviation. Finkelstein, *supra* note 120, at 4-5.

to-year over the fifteen years of our study. The broad dispersion of the data, in effect, create a statistical haze that makes it impossible to discern any trend at all.

The small differences in median patent prosecution times between the three assignee types and the five biotechnology subfields lead to similar findings, *see* Figures 10b & c. For the three assignee groups, the differences in median prosecution time is largest between corporations and government, with the former having a mean prosecution time of 3.0 years and the latter 3.5 years. By contrast, the upper and lower bounds of the fifty-percent interquartiles in the box plots span 1.5 years and 1.7 years for corporations and government, respectively. Similarly, the largest difference in the median patent prosecution times between biotechnology subfields is only 0.7 years (2.7 years for polypeptide and protein patents versus 3.4 years for immunological and genetically modified organism patents), whereas the smallest fifty-percent interquartile span is 1.5 years. In both cases, the differences fall well within the broad distribution of the data.

These results confirm our intuitions about the dynamics of the patent prosecution process. At base, there is little reason to believe that patent prosecution time is sensitive to technical differences between biotechnology subfields. First, while patent review may stretch over more than two years, the actual time spent on patent applications by examiners averages out to be just eighteen hours.¹²⁶ In other words, because much of the delay derives from external factors, one would expect the total time for prosecuting a patent to correlate weakly with the actual time a PTO examiner and a patentee's representative(s) spend on the prosecution process. As recent appeals for greater resources suggest,¹²⁷ PTO backlogs caused by limited resources, more complex patents, and rising number of applications have a much more direct effect on time for patent review than differences between technologies.

¹²⁶ Lemley, *supra* note 17, at 1500.

¹²⁷ Dudas, *supra* note 2, at 2.

Second, while backlogs may vary between technology fields because of differences in the numbers of patents filed and the complexity of the technology, the influence of these factors will be obscured by the diverse characteristics of inventions within a given field.¹²⁸ It is obvious, for example, that some biotechnology inventions are complex, while others are elegantly simple. Similarly, many patents involve follow-on inventions that represent minor variations on a basic theme within an established area of research and development, whereas others are truly novel and groundbreaking. The range of inventions, whether characterized by novelty, complexity, or predictability, is reflected in the broad variability of the patent prosecution times for each of our biotechnology subfields, see Figures 10c. The underlying diversity of inventions within a field thus blurs the slight association that might exist between patent prosecution time and substantive differences between fields of research and development.

The analysis of our data for number of claims and patent citations mirrors that for prosecution time data.¹²⁹ The overall trends are straightforward enough. The median number of claims per patent decreases slightly until 1994, increases through 2000, and then levels off through 2004, Figure 11a. Similar to the prosecution-time data, the aggregate values suggest a slight trend towards increasing numbers of claims, ranging from a low in the median of nine claims to a high of fourteen claims. The modest nature of this trend is highlighted by the fact that the number of claims in a patent can exceed 400. As above, when considered on a patent-by-patent basis, these changes once again fall well within the fifty-percent interquartile of the

¹²⁸ Polk Wagner, *Of Patents and Path Dependency: A Comment on Burk and Lemley*, 15 Berkeley Tech. L.J. 1341, 1342-43 (making a similar point in an argument about the importance of “micro-specificity” when interpreting modern patent jurisprudence).

¹²⁹ The distributions for these data deviate dramatically from normality. They are highly asymmetrical because they are bounded on the low end by zero or one and they have a substantial tail at the high end that extends out to very large numbers of either claims or citations. As already note, this skew creates precludes us from using means or standard deviations. Finkelstein, *supra* note 120, at 4-5, 22.

box plots and thus are meaningless for predictive purposes.¹³⁰ Nor do we observe any meaningful differences in median claim numbers between biotechnology subfields or assignee types, *see* Figures 11b & c.

These results also bear out our intuitions. No grounds exist to anticipate that the number of claims in a patent correlates in a systematic way with its issue date or area of invention. Valuable inventions, which may justify more expansive (and therefore expensive) claims, arise in all areas of biotechnology research and development. Further, claiming strategies are as variable as the inventions they are designed to protect, which range from multifaceted to streamlined, from complex to simple, and from paradigm-shifting to cumulative. These differences transcend variations over time and between subfields; cumulative to transformative work occurs across time and between specific areas of research and development. Our data provide quantitative support for this view. The variation in the number of claims is almost as broad within discrete patent subpopulations as it is for biotechnology as a whole, suggesting that differences within subfields or years are likely to overwhelm differences between them.¹³¹

These observations expose another closely related point. Commentators have become increasingly prone to categorizing certain fields of invention as being dominated by specific types of inventive activity. Biotechnology has been characterized, for example, as being

¹³⁰ The actual medians for each of the assignee categories are twelve for corporations, thirteen for universities, and eleven for the federal government, whereas the fifty-percent interquartile for each is seven to twenty-one claims, six to twenty-one claims, and six to eighteen claims, respectively. In the tech subfields, the medians range from ten, for protein sequences, to fourteen, for measuring and testing processes, while the narrowest interquartile ranges are seven to twenty claims, for genetically modified organisms, and five to eighteen claims protein sequences.

¹³¹ The five largest PTO subclasses, for example, have medians and fifty-percent interquartiles as follows: (1) the median is 16 and the interquartile range 9-27 for subclass 435/6; (2) the median is 12 and the interquartile range 7-19 for subclass 435/7.1; (3) the median is 14 and the interquartile range 7-22 for subclass 435/5; (4) the median is 11 and the interquartile range 6-18 for subclass 514/12; and (5) the median is 6 and the interquartile range 3-12 for subclass 530/350.

“research oriented” and distinguished by its discrete technological leaps.¹³² We find that almost fifty percent of biotechnology patents cover new measuring and testing processes,¹³³ many of which are more analogous to devices than to drugs, which typically provide the model for biotechnology research and development.¹³⁴ While these sorts of generalizations can be useful, they risk distorting legal policies by propagating a one-dimensional image of biotechnology that marginalizes important innovative work.¹³⁵ Our data show that the potential for misunderstanding is particularly high for biotechnology.

The citation data are the least illuminating of the study.¹³⁶ We find that the number of citations made in biotechnology patents exhibit no trends whatsoever. Almost across the board, whether by year, assignee type, or biotechnology subfield, the median number of citations is two; indeed, only a few instances exist in which it deviates to one or three citations. Yet, the range in the number of citations made is quite broad and apparently becoming broader. The data on citations received are much more limited because of the poor temporal overlap between our data and the citation data we integrated from the Hall Study.¹³⁷ The median number of citations

¹³² Dan L. Burk & Mark A. Lemley, *Policy Levers in Patent Law*, 89 VA. L. REV. 1575, 1667-79 (2003); Dan L. Burk & Mark A. Lemley, *Biotechnology’s Uncertainty Principle*, 54 CASE W. RES. L. REV. 691, 732-34 (2004).

¹³³ See *infra* Part II.A.

¹³⁴ Instrument development is distinct from drug development. Discovering novel drug therapies, which by their nature are rare, involves extensive trial-and-error research and development and entails often stratospheric costs. Burk, *Policy Levers*, *supra* note 132, at 1581-82 (noting that drug development can take more than a decade and cost hundreds of millions of dollars). By contrast, measuring and testing processes often involve more standardized science, incremental advances, and substantially lower costs.

¹³⁵ In fact, even drug development, which is often portrayed as the archetype for this type of research and development, is a mixture of discontinuous advanced followed by cumulative follow-on research and development. See, e.g., Geoffrey Duyk, *Attrition and Translation*, 302 Science 603, 604 (2003) (describing how economic pressures are driving drug companies towards “fast follow-on” strategies or “me too” drugs that involve minimal additional research because the new drugs are slight variants of existing products); Stephen S. Hall, *The Drug Lords*, N.Y. Times, Nov. 14, 2004, §7 (Book Rev.), at 8.

¹³⁶ Note that given the absence of trends, we have chosen not to include any figures in the Article for the citation data; direct description of our results seems more the adequate under the circumstances.

¹³⁷ The citation data from the Hall Study end in 1999, and this poor overlap is compounded by the time-lag in citations received by patents—typically it takes several years before citations to a patent begin to accumulate.

received was highest (four citations) for patents that issued in 1990 and 1991, after which it diminishes monotonically.¹³⁸

Citation data are arguably most interesting for what they could reveal about successful assignees and possible differences between the public and private sectors. Intuitively one would expect the number of citations a patent receives to correlate with the value or, perhaps less significantly, the scope of the invention, and the most successful assignees to own more patents that receive a greater number of citations.¹³⁹ The data display no difference between the median value for our “big ten” (*i.e.*, the largest and most successful) biotechnology companies and that for all other entities. This result must be treated skeptically, as it could be an artifact of the short time period covered or of the citations in biotechnology patents tending to be to non-patent sources.¹⁴⁰ Alternatively, it may be that successful companies also fail often or perhaps that the existing measures of patent value are too narrow.¹⁴¹ In any event, our citation data for biotechnology patents prove to be decidedly mute.

2. *An Empirical Haze in Common*

Two recent studies of U.S. patent trends, the Hall Study and an earlier study by Allison and Lemley, purport to reach opposite conclusions about the insights that can be gleaned from patent characteristics.¹⁴² On closer inspection, however, it becomes apparent that the authors

¹³⁸ This trend is just what one would expect, as the older patents have had the longest time to accumulate citations. We also observe no differences between the medians for either the three categories of assignees or the five biotechnology subclasses.

¹³⁹ Hall, *supra* note 8, at 15-16, 23-26.

¹⁴⁰ Allison & Lemley, *supra* note 8, at 2131 (finding that citation data for a small sample of biotechnology patents suggest that most citations are to non-patent sources).

¹⁴¹ For example, researchers may learn a great deal from their failures or, as some authors have argued, the value of a patent may be its role in a portfolio rather than on its own. Parchomovsky, *supra* note 12, at 8-10.

¹⁴² The Hall Study provides a comprehensive overview of every patent issued between 1963 and 1999, whereas the Allison-Lemley study is a more detailed analysis of a one-thousand-patent sample. Hall, *supra* note 8, at 8-9; Allison & Lemley, *supra* note 8, at 2108-09. For example, beyond the basic information collected in the Hall Study, Allison and Lemley provide much more precise information on the technology area of each patent, the number and type of prior applications on which each patent bases its priority date, the nationality of the inventor(s) and assignee, the number and type of prior art references cited in the patent, and the number and types of claims. *Id.*

gloss over several significant limitations of their findings.¹⁴³ This is particularly true of certain aspects of the Allison-Lemley study. For example, the authors investigate average patent prosecution times for the fourteen technology fields defined in their study and from these results claim that the “patent prosecution system [] spends much more time and attention on some sorts of patents than others.”¹⁴⁴

The basis for this conclusion is a regression analysis,¹⁴⁵ which reveals that the association between technology field and prosecution time is statistically significant (i.e., greater than zero¹⁴⁶) for five of the fourteen technology fields.¹⁴⁷ But the authors fail to report the magnitude of these associations, which is essential to interpreting the practical significance of their findings. This omission is equivalent to saying that today’s temperature will not be zero without giving any indication of how far from zero it is predicted to deviate.¹⁴⁸ Their data thus do not disclose whether much more time is actually spent on certain classes of patents.

¹⁴³ The data in the Hall Study undoubtedly provide important insights into aggregate patent trends. The primary shortcoming of this study is its failure to examine statistical variance. Possibly because it is population study—as opposed to discrete samples of patents—the authors do not discuss the distribution of their data. Yet, similar to our data, all of their standard deviations are large, hovering around one year, and dwarf the differences in their annual averages for prosecution times. at Table 1. As a consequence, none of the putative trends is meaningful from the standpoint of using them as predictors for individual patents.

¹⁴⁴ *Id.* at 2102.

¹⁴⁵ Linear regression is an alternative mode of statistical analysis that is commonly used to resolve trends in sample data. A primary virtue of regression analysis is that, unlike statistical significant testing, it assesses both statistical significance and the *magnitude* of the association.

¹⁴⁶ In regression analysis, statistical significance implies that one should reject the hypothesis that no association exists (i.e., the correlation coefficient, β , equals zero) between the independent variable (e.g., number of claims) and the probability that a patent will be litigated. In other words, if β were zero, there would be a very low probability (typically less than five percent) of obtaining the observed β value. Finkelstein, *supra* note 120, at 375-76. The standard for statistical significance in science is a ninety-five confidence limit, or ± 1.96 standard deviations from the mean. *Id.* at 120-21, 166-67, 171; Rice, *supra* note 121, at 303.

¹⁴⁷ Allison & Lemley, *supra* note 8, at 2126-27 (finding statistically significant results for chemical pharmaceutical, medical devices, biotech patents, software, and chemical inventions). The authors use “Poisson regression,” which is a variant of standard least-squares linear regression. This type of regression applies to data that fit a Poisson distribution, which is simply a variant of the normal distribution used when the dependent variable is discrete (i.e., a count) rather than continuous. Finkelstein, *supra* note 120, at 141, 351, 472-73. The method employs a logarithmic transformation of the data, such that the natural logarithm of the dependent variable (here prosecution time) becomes a linear function, specifically $\alpha + \beta x$, of the independent variable “x” (here patent issue date), where α and β are both constants. *Id.*

¹⁴⁸ A finding of statistical significance says nothing about whether the magnitude of the difference is large or small. This result may initially appear to be contradictory—how can we find the association statistically significant

Allison and Lemley single out their citation data as being particularly compelling.¹⁴⁹ They base this conclusion on the purportedly divergent results among their fourteen technology categories. Their data disclose that the overall median number of references for the fourteen technology categories is ten references¹⁵⁰ and that the range of median values for the technology fields varies from eight references, for semiconductor and communications-related inventions, to seventeen references, for biotechnology inventions.¹⁵¹ However, all but three of the technology fields are clustered around the median of ten references, and each of the divergent categories has a small sample size.¹⁵² Accordingly, far from showing dramatic differences, the broad variances of their data make it difficult to infer that these differences are real, and where the differences appear noteworthy, the data are too thin for one to have confidence in the result.¹⁵³

and of nominal magnitude at the same time? This apparent conflict disappears once one realizes that measures of statistical significance are independent of the estimated magnitude of the association. Gerd Gigerenzer & David J. Murray, COGNITION AS INTUITIVE STATISTICS 24 (1987). Statistical significance indicates only that the likelihood of *no* association existing is very low. Finkelstein, *supra* note 120, at 120-21; Ian Hacking, An Introduction to Probability and Inductive Logic 213-15, 222-23 (2001). Further, where, as here, the sample of population size is very large, it is relatively easy to find statistical significance because the likelihood of finding statistical significance increases with sample or population size. Finkelstein, *supra* note 120, at 188-89; Gigerenzer, *supra* note 125, at 15 (observing that “one can always increase the sample size to a level where the difference will be significant. Consequently, the null [hypothesis] can always be disproved if the sample size *n* is large enough.”). Accordingly, if one is uncertain of the magnitude or nature of the effect, findings of statistical significance may have no physical or practical meaning. *Id.*

¹⁴⁹ Allison & Lemley, *supra* note 8, at 2130. Allison and Lemley rightly observe that prior art references are “the best proxy we have in these data for the quality and thoroughness of a patent examination.” *Id.* at 2130.

¹⁵⁰ *Id.* at 2149 (the minimum number of references in a single patent is zero and the largest number is 163). The mean is fifteen references, which, with the large maximum number of references and fixed minimum of zero references, evidences a highly skewed distribution. Finkelstein, *supra* note 120, at 4-5.

¹⁵¹ *Id.* at 2160.

¹⁵² *Id.* Similar trends exist when the references are restricted to U.S. patents alone—all but four of the technology areas have medians of five to eight references. *Id.* at 2158. Given the skew of their data, Allison and Lemley erroneously focus on differences between the mean values, which, unlike the medians, vary from a low of nine references for semiconductors to highs of twenty-six and twenty-four references for medical devices and biotechnology inventions, respectively. *Id.* The distribution for the number of references in medical device patents is the most extreme example of this—it has a ten-reference difference between its mean and median values and the highest extreme value of all (163 references). *Id.* at 2160.

¹⁵³ Allison and Lemley do find one notable difference between the technology fields: pharmaceutical and biotechnology patents cite far more non-patent prior art references than the other technology areas. *Id.* at 2131, 2158-59. The authors also find that patents are the primary references cited in patents, which they rightly challenge given the importance of non-patent references in many technology fields. *Id.* at 2120-21.

The Allison-Lemley study displays the same limitations we observe in our own data. The most consistent features are the large variance of the data and the nominal differences observed between classes of patents—differences between classes are much less than the differences within them. This disparity limits the insights these studies can provide. As we argue above, intuitive grounds exist for expecting these results, but a more fundamental reason may also exist. It may be that few “general characteristics” of patents exist, and this variability may be particularly true of rare high-value patents. The analytical challenges created by this go beyond the descriptive studies analyzed here. They also hamper predictive studies, like the *Valuable Patents* study discussed below, that seek to identify useful associations between specific characteristics of a patent and its economic value or likelihood of being litigated.

B. The Elusive Search for Metrics of Innovative Success

Metrics for predicting patent value are receiving significant attention from economists and lawyers following the construction of new databases.¹⁵⁴ Economists have released pioneering studies correlating patent renewal statistics and citation rates to patent value, and they were the first to use litigated patents to determine whether valuable patents have distinct characteristics.¹⁵⁵ Studies by lawyers soon followed with the creation of major databases on

¹⁵⁴ Griliches, *supra* note , at 1669-70 (classic review article describing the need to evaluate patent quality to use patents as indicators of innovative output); Dietmar Harhoff, *et al.*, *Citations, Family Size, Opposition and the Value of Patent Rights*, 32 RESEARCH POL. 1343, 1343-44 (2003) (describing new study that correlates various patent characteristics with patent quality); Jean O. Lanjouw & Mark Schankerman, *Patent Quality and Research Productivity: Measuring Innovation With Multiple Indicators*, 114 ECON. J. 441, 442 (2004) (study that develops “a composite index of patent quality using multiple characteristics of patents”); Lanjouw, *supra* note 47, at 129 (study finding that “the frequency of legal disputes is strongly correlated with a variety of characteristics of innovations and their owners”); Lanjouw, *supra* note 114, 407 (study evaluating the use of patent application and renewal data as indicators of patent value); Brownwyn H. Hall, *et al.*, *Market Value and Patent Citations*, 36 RAND J. ECON. 16, 16-17 (study assessing the correlation between citations to a patent and its economic value); Jean O. Lanjouw, *Patent Protection in the Shadow of Infringement: Simulation Estimations of Patent Value*, 65 REV. ECON. STUDIES 671, 672-73 (1998) (simulation model for patent value using data on patent renewals and statutory legal fees from Germany); Schankerman, *supra* note 117, at 78-79 (study providing estimates of the private value of patents based on patent renewal decisions).

¹⁵⁵ See, e.g., Schankerman, *supra* note 117, at 78-79; Lanjouw, *supra* note 47, at 129, 144-45; Lanjouw, *supra* note 8, at 157-58, 164.

litigated patents and comparative work on litigated and non-litigated patents that exploits the new data.¹⁵⁶ This work has generated important insights but, as we will show, metrics based on patent characteristics have severe limitations that have gone unrecognized.

Among legal scholars, this empirical work is closely associated with concerns about the PTO's growing backlog of patents. Legal commentators have suggested that PTO resources could be economized by limiting reviews of the substantial majority of patents that have little or no economic value.¹⁵⁷ Their reasoning is simple. Because the slim tail of high-value patents accounts disproportionately for the success of the U.S. patent system, valuable patents ought to be subject to careful review by the PTO, whereas the masses of so called valueless patents ought to be reviewed lightly.¹⁵⁸

This triage strategy raises practical and theoretical problems. At the practical level, it requires that valuable patents be identifiable *ex ante*. We show that the broad variance of patent characteristics makes this exceedingly difficult. At the theoretical level, it implicates the work of Harvard economist F.M. Scherer, who revealed that innovative success is chaotic and resistant to statistical methods.¹⁵⁹ Scherer's work has not dampened optimism. To the contrary, proponents of a patent triage system, particularly the authors of *Valuable Patents*, all but ignore the teachings of Scherer's studies.¹⁶⁰ We reassess the empirical bases of this approach to determine whether it succeeds in contravention of Scherer's work.¹⁶¹

¹⁵⁶ Moore, *supra* note , at 367-68; Kimberly Moore; Allison, *Valuable Patents*, *supra* note 5, at 436-38.

¹⁵⁷ See, e.g., Allison, *Valuable Patents*, *supra* note 5, at 439, 464-65. One recent study suggests that more than ninety-five percent of all patents issued are neither litigated nor licensed for a royalty and thus presumably have low economic value. Lemley, *supra* note 17, at 1507.

¹⁵⁸ Allison, *Valuable Patents*, *supra* note 5, at 464-65.

¹⁵⁹ Scherer & Harhoff, *supra* note 19, at 563.

¹⁶⁰ Allison, *Valuable Patents*, *supra* note 5, at 462.

¹⁶¹ James Bessen and Michael Meurer are also critical of this study, but their worry rests on the likely selection bias that derives from using litigated patents as a surrogate for valuable patents. Bessen, *supra* note 7, at 4-6.

Valuable Patents begins with two provocative statistics: patentees spend a total of \$4.43 billion per year obtaining patents, but ninety-nine percent of U.S. patents are never enforced through litigation.¹⁶² The authors do not find the imbalance between dollars spent on prosecuting patents and the lack of enforcement to be problematic. Instead, they posit that “some patents are intrinsically more valuable than others.”¹⁶³ For them, the important issue is identifying the distinctive characteristics of valuable patents.¹⁶⁴ Their approach is elegant—they use litigation as a proxy for value on the premise that a patent will not be litigated unless it has significant economic or strategic value to the patentee.¹⁶⁵ Litigation status provides the authors with a reliable criterion for collecting a sample of valuable litigated patents to compare against otherwise valueless non-litigated patents.¹⁶⁶

The paper contains two independent empirical studies, one large and one small. The large study uses the Hall Study dataset (*i.e.*, every patent that issued between 1963 and 1999) as the putative population of valueless patents.¹⁶⁷ Their sample of valuable patents is drawn from patent suits that terminated during the period January 1999 to December 2000.¹⁶⁸ The smaller, and more detailed, second study uses the thousand patents collected by Allison and Lemley

¹⁶² *Id.* at 435.

¹⁶³ *Id.* at 436-37.

¹⁶⁴ *Id.* at 437.

¹⁶⁵ The authors cite a number of sources to support this premise. *Id.* at 441. Moreover, given that only a small fraction of all patents are believed to be valuable, recent estimates suggest less than five percent, the fact that some valuable patents will be included in the non-litigated population is unlikely impair their results noticeably. Lemley & Lemley, *supra* note 8, at 1507 (estimating that “the total number of patents litigated or licensed for a royalty (as opposed to cross-license) is on the order of five percent of issued patents”).

¹⁶⁶ As the authors put it, “[o]ur study allows us to evaluate the efficacy of these measures by determining how well they predict the likelihood of litigation. Because litigated patents are a representative subset of valuable patents, measures offered as predictors of value should, if accurate, also predict litigation. Some, but not all, of the value measures used by economists pass this test.” *Id.* at 456-60.

¹⁶⁷ Allison, Valuable Patents, *supra* note 5, at 437.

¹⁶⁸ *Id.* at 437 (the sample consists of 4,247 cases involving at total of 6,861 patents). To ensure a fair comparison, the authors actually weighted the NBER data, such that, “if 1% of the litigated patents were issued in a given year, our study gives issued patents for that year 1% of the total weight” for calculation of parameters like means and median used in their comparisons. *Id.* at 445.

(discussed above) as the putatively valueless patent sample.¹⁶⁹ The valuable patents are drawn from the same sample of law suits used in the first study.¹⁷⁰ For each study, the authors independently evaluate the characteristics of litigated and non-litigated patents to determine whether statistically significant differences exist between the two sets of patents.

The authors claim that their “data conclusively demonstrate that valuable patents differ in substantial ways from ordinary patents both at the time the applications are filed and during their prosecution. This suggests that valuable patents can be identified before-hand, at least in the aggregate.”¹⁷¹ The authors find that litigated patents have numerous stable characteristics,¹⁷² although they single out number of claims, prior citations made, and citations received as “unambiguously strong predictors of patent litigation.”¹⁷³ To clarify the bases of their findings, we reexamine the data on these three characteristics and patent prosecution times.¹⁷⁴

The analytical limits we identified in the preceding subsection apply equally here.¹⁷⁵ The authors focus exclusively on statistical significance. Yet, as already noted, statistical significance does not have a direct bearing on the strength of association between, say, patent value and number of patent claims; it only demonstrates that it is very unlikely that *no*

¹⁶⁹ *Id.* at 437, 446. The additional variables in the second study include the following: (1) prior art citations to foreign patents and no-patent prior art, (2) nationality of assignee, (3) assignee status as “small entity”, and (4) a refined fourteen-category assessment of the area of technology. *Id.* at 446.

¹⁷⁰ *Id.* at 437. Only the 300 law suits for which the patents at issue were granted between mid-1996 and mid-1998 are included. *Id.* This constraint is imposed to ensure that the litigated patents were issued during the same years that the non-litigated patents were issued. *Id.*

¹⁷¹ *Id.* at 438. The authors also conclude that “valuable patents differ in substantial ways from ordinary patents both at the time the applications are filed and during their prosecution.” *Id.*

¹⁷² *Id.*

¹⁷³ *Id.* at 451 (“The more claims, prior art citations, and citations received a patent has, the more likely it is to be litigated.”).

¹⁷⁴ The authors identify several additional “significant predictors of litigation,” including the number of patent applications associated with a patent, the length of time for patent prosecution, and patent age at time of litigation. *Id.* at 456-60.

¹⁷⁵ This consistency should come as no surprise given that this paper uses data collected in both the NBER and Allison-Lemley studies and utilizes the same types of statistical methods. *Id.* at 437.

association exists at all.¹⁷⁶ Weak associations, though statistically significant, will have no predictive value for either individual patents or samples of modest size; their only value will be in estimating the relative proportion of valuable patents in a large sample.

1. *Statistical Significance Without Predictive Power*

Virtually all of the authors' statistical analyses are conveyed in two tables at the end of the paper.¹⁷⁷ In Tables 1 and 2 below, their findings are presented to ensure that the results of the two studies are readily comparable. For our purposes, two of the factors in the Tables are of particular interest: (1) "p-values," which typically must be less than five percent (0.05) for the result to be statistically significant, and (2) "unit increases," which represent the percentage change for each unit increase in a descriptive variable (e.g., number of patent claims).¹⁷⁸

Table 1: Large Study Statistics

	p-value	Odds Ratio	Unit Increase	Not Litigated ¹⁷⁹	Litigated	Probability Diff. ¹⁸⁰
No. Claims	<0.0001	1.014	1.4%	13	20	10%
Pros. Time	<0.0001	1.024	2.4%	2.8 ¹⁸¹	4.1	3%
Cites Received	<0.0001	1.128	13%	4.3	12.2	159%
Cites Made	<0.0001	1.096	9.6%	8.4	14.2	70%
U.S. Corp.	<0.0001	0.733	-27%	0	1	-27%
Foreign-Corp	<0.0001	0.523	-47%	0	1	-47%
U.S. Individual	0.0003	1.450	45%	0	1	45%
U.S. Gov't	<0.0001	0.072	-93%	0	1	-93%
Chemical	<.0001	0.663	-34%	0	1	-34%

¹⁷⁶ Indeed, because the number of patents in the two studies is quite large, even weak associations will be found to be statistically significant. See note 148.

¹⁷⁷ *Id.* at 478-79. The regression analysis is the same in both, although the data are presented somewhat differently.

¹⁷⁸ For example, if a unit increase for the number of claims in a patent is two percent, a patent with one more claim than another, say fourteen claims versus thirteen, has a two percent greater probability of being litigated.

¹⁷⁹ Binary categories, such as whether an invention is mechanical or not, are categorized as "zero" (not litigated) or "one" (litigated). Non-binary categories, such as number of claims or citations, are categorized by the mean number of claims for litigated and not-litigated patents.

¹⁸⁰ This column provides an estimate of the increase in the probability that a patent will be litigated between the mean values of non-litigated patents and litigated patents. Small differences indicate that the characteristic at issue does not provide an effective means of discriminating between valuable litigated and non-litigated patents.

¹⁸¹ Mean values for the prosecution times were not provided for the large study; we use the means observed in the sample study as surrogates.

Computers & Communications	<.0001	1.673	67%	0	1	67%
Drugs & Med.	<.0001	1.841	84%	0	1	84%
Electrical & Electronics	0.0078	0.876	-12%	0	1	-12%
Other Tech.	<.0001	1.344	43%	0	1	43%

Table 2: Sample Study Statistics

	p-value	Odds Ratio	Unit Increase	Not Litigated	Litigated	Probability Diff.
Ind.-claims	0.0060	1.095	9.5%	2.8	4.4	16%
Dep.-Claims	0.0807	1.013	1.3%	12.1	21	12%
Pros. Time ¹⁸²	0.0962	1.108	11%	2.8	4.1	14%
Cites Rec.	0.0001	1.162	16%	1.3	3.4	37%
Cites US-Pat	0.1703	1.008	0.8%	10.3	22.8	10%
Cites For-Pat	0.0228	1.021	2%	2.4	8.3	13%
Total Apps.	0.2421	1.162	16%	1.5	2.6	16%
Foreign Invtr.	0.0001	0.21	-79%	0	1	-79%
Individual	0.0001	2.821	182%	0	1	182%
Small Bus.	0.0001	3.162	216%	0	1	216%
Electronics	0.0015	2.616	162%	0	1	162%
Med. Devices	0.0521	0.526	-47%	0	1	-47%
Chemistry	0.0088	2.364	136%	0	1	136%
Mechanics	0.0001	4.818	382%	0	1	382%
Auto Related	0.0194	0.326	-67%	0	1	-67%

Average Time for PTO Patent Prosecution. The association between prosecution time and probability of litigation is marginal. For their large and small studies, we calculate a three- and fourteen-percent difference, respectively, between non-litigated and litigated patents.¹⁸³ The large study exemplifies the limited value of statistical significance for large samples, which because of their high statistical power cause even nominal differences (here three percent) to be found statistically significant. The smaller, less-powerful study identifies a larger, though still modest, association between prosecution time and probability of litigation. However, the estimated value, though larger, is not statistically significant, meaning that zero association

¹⁸² As noted earlier for the Allison-Lemley study, patent prosecution time is measured from date of earliest application, whereas we measure prosecution time only as of the filing date of the most recent patent application. *Id.* at 459.

¹⁸³ More precisely, litigated patents are only three (large study) and fourteen (small study) percent more likely to have been prosecuted for average prosecution time for litigated patents than non-litigated patents. In other words, at the peak of the distribution for litigated patents, litigated patents are only slightly more likely than non-litigated patents to have been prosecuted for this length of time.

cannot be rejected. In either case, the slight differences in patent-review times are plainly inadequate for predicting whether a patent is likely to be litigated.

Average Number of claims. This characteristic is also a poor predictor of whether a patent is likely to be litigated. While the association is statistically significant, its magnitude is small—just a 1.4 percent increase per claim for the large study and 9.5 percent per independent claim (only 1.3 percent per dependent claim) in the sample study. The weakness of the association is evident in the small difference in the probability of litigation—less than twenty percent—between the average number of claims for non-litigated and litigated patent samples. Moreover, the author’s use of averages, rather than medians, will likely over estimate the differences between litigated and non-litigated patents.¹⁸⁴ These weak associations simply reflect the fact that counting claims does not tell us much about the nature of a patent.

Average Number of Citations Made and Received. The data on citations made by patents are, at best, ambiguous. The large study suggests quite a strong association, about a seventy percent difference between the average values for litigated and non-litigated patents. By contrast, the small study finds the differences to be nominal, only about ten percent, and not statistically significant. The large differences between the two studies are indicative of the instability of these estimates and grounds for caution when assessing whether the number of citations made by a patent is a reliable metric.¹⁸⁵

¹⁸⁴ We infer that the distributions are skewed from the Hall Study and Allison-Lemley data, both discussed above, both of which are integrated into this study, and we have no reason to believe that the data for the litigated patents, which are subsets of the other two studies, will be any different. Further, the standard error of a regression coefficient, which they use to calculate statistical significance, can underestimate the variance of the coefficient. Finkelstein, *supra* note 120, at 475-76. Accordingly, the authors’ findings of statistical significance, particularly where they are close calls, could be inaccurate.

¹⁸⁵ In their discussion, the authors log transform their data with no explanation whatsoever and then fail to explain the differences between the log-transform results for statistical significance and the results reported in their logistic analysis in the tabulated data. Allison, Valuable Patents, *supra* note 5, at 454.

Both studies reveal a consistent and sizeable association between average number of citations to a patent and its litigation status. The large study, because of its much higher average,¹⁸⁶ demonstrates that the number of citations received is well correlated with the litigation status of a patent. Litigated patents on average receive 1.6 times more citations.¹⁸⁷ The large study thus provides solid grounds for a significant association existing between the number of citations a patent receives and its litigation status.

Status of Owner and Technology Areas. It is worth commenting briefly on owner status and the differences observed between technology areas. Both studies reveal strong positive associations if the patent owner is an individual, and the sample study shows a strong association when the owner is a small business.¹⁸⁸ Similarly, strong associations are evident for certain technology fields, particularly mechanics (small study) and computers and communications (large study).¹⁸⁹ Several jarring differences between the two studies nevertheless exist.¹⁹⁰ For example, the small study finds strong *positive* associations for electronics and chemistry, whereas the large study predicts modest *negative* associations for similar fields.¹⁹¹ The strength

¹⁸⁶ The two studies generate divergent estimates of the difference in average number of citations received per patent between litigate and non-litigated patents. However, this disparity is almost certainly attributable to the fact that the small study is limited to patents granted relatively recently, whereas the large study is not—the small study is limited to patents that issued between 1996 and 1998, whereas the large study includes patents that date back to 1963. *Id.* at 445.

¹⁸⁷ It is nevertheless important to reiterate that the authors' use of averages, as opposed to medians, causes the difference between non-litigated and litigated patents to be over estimated.

¹⁸⁸ The significance of the large study data is strengthened by the strong negative associations for U.S. and foreign corporations and the U.S government. We have no sense, other than the differences between the two studies, of the variance of these estimates; we only know that they are statistically significant.

¹⁸⁹ The authors analyze their data using two distinct methods. For example, they claim "striking variations by industry," emphasizing med-devices, computer-related, software, electronics, and mechanics as being more likely to be litigated and chemistry, automotive-related, semi-conductor inventions as being much less likely to be litigated. *Id.* at 474-75. Yet, the regression analysis of the sample study finds that the associations with semiconductor, computer-related, and chemical inventions are not statistically significant. *Id.* at 479.

¹⁹⁰ The large variability between studies may also reflect a large variance in the regression coefficients, which may not be accurately represented in the standard error used to calculate statistical significance. Finkelstein, *supra* note 120, at 475-76.

¹⁹¹ The authors are in fact dismissive of the level of association between technology areas and probability of litigation in the large study because of the roughness of its classification scheme, despite the fact that the magnitudes of the associations are some of the largest for any of the characteristics evaluated in the large study. *Id.* at 466, 471-76.

of the association for individual inventors is also markedly different (more than a factor of three) between the two studies.¹⁹² As above, these differences suggest that one should be cautious about reading the strength of these associations too literally.

Our reanalysis clearly affirms two of the authors' central conclusions: the strong association between citations received per patent and the status of the patent owner. It also finds solid evidence for concluding that mechanical patents are more likely to be litigated than the average patent. This interpretation of the *Valuable Patents* data departs from that of the authors with respect to the three standard patent characteristics, which we find to be poor predictors of whether a patent is likely to be litigated.

If our analysis were to stop here, some prospect, albeit limited, would remain for prioritizing the PTO's patent prosecution process. Unfortunately, as the next subsection will show, a more fundamental problem also exists.¹⁹³ The small fraction of litigated patents makes it exceedingly difficult to exploit differences that may exist between non-litigated and litigated patents. Other than extreme values (*e.g.*, patents with fifty or more citations), the distribution of non-litigated patents will overwhelm that of the litigated patents. This finding implies that, for the vast majority of patents, even patent characteristics strongly associated with economic value will have only nominal predictive power.

If one assumes that conflicting results between the studies cancel each other and suspends judgment on technologies likely to have low sample sizes (*i.e.*, medical devices, auto-related patents), the only technology field with a firmly established, strong association is mechanics.

¹⁹² Some of these divergences may derive from the low sample sizes for the small study. Sample-size problems will be particularly at issue for the 300 litigated patents, which the author's divide into fourteen technology fields, implying that most categories will have fewer than thirty patents (a standard minimum number used by statisticians). Finkelstein, *supra* note 120, at 4-5..

¹⁹³ It is also worth noting that citations received are not manifested until after a patent issues and cannot be assessed for a still longer period of time. In fact, rather predicting value, the association with citations received may run in the opposite direction—the success of a patented invention and the ensuing recognition that the patent on it is valuable may be what alerts other inventors to cite the patent, as opposed to strict substantive judgments.

2. *Base Rates and the Slim Tail of Innovative Success*

The economist Joseph Schumpeter once famously speculated that a small number of inventions account for a disproportionate share of the total profits from inventive activities.¹⁹⁴ F.M. Scherer was among the first economists to provide a sound empirical basis for Schumpeter's prediction.¹⁹⁵ In addition to confirming Schumpeter's hypothesis, Scherer showed that "it is very hard or maybe even impossible to secure stable average profit returns by pursuing portfolio strategies, e.g., pooling many research and development projects into a portfolio."¹⁹⁶ In other words, the pooling of risks permitted by standard statistical methods is ineffectual in the face of the variability that is characteristic of innovative success.

In this tail-wags-dog universe, statistical measures, such as averages and standard deviations, are often erratic and unreliable. This failure arises because statistical analyses are driven by the center of a sample distribution, not its periphery. Consistency—and predictability—emerge because the large number of small, random events cancel out each other, exposing the systematic influences on the population as a whole. Patents defy this basic model because a small number of extremely valuable patents are disproportionately important to aggregate innovative success; it is the aberrant patents that make the difference.

Scherer's results are at odds with recent empirical work on valuable patents—if general descriptive statistics fail, statistical analyses of specific patent characteristics will fail for the same reason. Scherer likens the market for innovation to a sweepstakes, "the innovation lottery," that randomly bestows huge prizes on a very small number of winners.¹⁹⁷ In *Valuable Patents*,

¹⁹⁴ Joseph A. Schumpeter, CAPITALISM, SOCIALISM, AND DEMOCRACY 73-74 (1942).

¹⁹⁵ Scherer & Harhoff, *supra* note 19, at 563; Fredrick M. Scherer, *The Innovation Lottery in EXPANDING THE BOUNDARIES OF INTELLECTUAL PROPERTY: INNOVATION POLICY OF THE KNOWLEDGE SOCIETY* 3, 4 (Rochelle C. Dreyfuss, et al., eds. 2001).

¹⁹⁶ *Id.* at 7.

¹⁹⁷ *Id.* at 15-16.

the authors make their objections to Scherer's lottery theory explicit: "if valuable patents can be reliably identified . . . the lottery theory runs into difficulty. At best, it becomes only a partial explanation."¹⁹⁸ The paper in fact turns on Scherer being wrong, and the authors acknowledge that PTO policy cannot be rationally calibrated and targeted towards certain classes or types of patents if Scherer's findings are accurate.¹⁹⁹

The findings of statistical significance presented in *Valuable Patents* are interpreted as refutations of Scherer's work.²⁰⁰ These proofs fall short of the mark. All that the paper establishes is that weak associations between patent value (*i.e.*, litigation status) and certain patent characteristics are unlikely to be the result of random variation. Their findings—unlike strong, consistent predictors of patent value—are not incompatible with Scherer's empirical findings. To the contrary, Scherer's work implies only that statistical measures, such as average values, are unstable over time and unreliable in the long run. Evidence of slight, often uneven associations from two samples of patents cannot refute Scherer's findings. Moreover, absent adequate disproof that inventive processes are chaotic, one would be foolish to rely on their results as trustworthy long-term estimates of the associations they purport to identify.

The authors' characterization of the results in *Valuable Patents* is also somewhat misleading. They fail to acknowledge the practical problems inherent in their approach. As we have pointed out several times, a finding of statistical significance is not proof that an association has predictive power.²⁰¹ This oversight is ultimately secondary, though, as a much more acute

¹⁹⁸ Allison, *Valuable Patents*, *supra* note 5, at 462 (citations omitted).

¹⁹⁹ *Id.*

²⁰⁰ *Id.* at 463-65.

²⁰¹ The authors acknowledge that identifying valuable patents at the time of application or issue may be possible only "in the aggregate." *Id.* at 438. However, this is akin to claiming you have come up with a security system for detecting terrorists in airports, but the system cannot identify specific terrorists and only provides rough probabilities of finding a terrorist in the general population.

obstacle exists. The process of identifying valuable patents presents a standard “base rate” problem because they constitute such a small minority of the patents issued.²⁰²

Making predictions with low base rates is akin to looking for a rare brand of needle, which may or may not have distinctive attributes, in a needle stack. A slightly different example will illustrate this point. Assume you are attempting to identify a small subset of valuable objects from a much larger collection based solely on their size and color. Low base rates, essentially the disparity in numbers between the valuable subset and the full collection, imply that the only valuable objects from the subset capable of being identified unequivocally will be those with colors and sizes that are very rarely found in the broader collection.²⁰³ By the same logic, if the distributions of sizes and colors in the small subset overlap significantly with those of the broader collection, few if any objects in the subset will be reliably identifiable.

The citation data from the large study reported in *Valuable Patents* illustrate the low base-rate problem well. Recall that the averages for number of citations received by the non-litigated and litigated patents were 4.3 and 12.2, respectively.²⁰⁴ We also found that patents receiving 12.2 citations are about 1.6 times more likely to be litigated than those that receive 4.3 citations.²⁰⁵ Using this regression analysis, we calculate that litigated patents will constitute approximately 1.4 percent of all patents with twelve claims, whereas more than ninety-eight

²⁰² Base rate refers to the proportions, or rates of occurrence, of the two classes of patents—the small number of valuable patents and the large number of valueless patents. As discussed earlier, it is well established that only a small fraction of the patents issued are valuable. Mark Lemley, *supra* note 17, at 1507 (estimating that the proportion of valuable patents is less than five percent of the total issued).

²⁰³ Because the numbers are so much greater in the collection, the numbers of objects with even relatively rare colors and sizes (the tails of these distributions) in the collection will overwhelm the numbers in the subset—even if the objects with those colors and sizes make up a much larger fraction of the small subset. For example, assume there are 1000 objects in the collection (*i.e.*, excluding the subset) and fifty in the subset. Assume further that two percent of the objects in the collection (20) are red and that forty percent of the objects in the subset (20) are red. Despite the disparity in percentages, red would provide a very poor basis for identifying the objects in the subset, as any given red object would be equally likely (a fifty-percent chance) to come from either the collection or the subset.

²⁰⁴ Allison, *Valuable Patents*, *supra* note 5, at 453-54.

²⁰⁵ *Infra* Part II.C.

percent will be valueless.²⁰⁶ Even among patents with fifty claims, litigated patents will constitute only about twenty-five percent of the total, and only beyond fifty-five claims are litigated patents projected to exceed the number of non-litigated patents. The implications of this example are clear—even if reliable, the weak associations identified in *Valuable Patents* will not, as a practical matter, be able to predict whether an individual patent is valuable.

The base-rate problem and the highly skewed distributions of most patent metrics create formidable obstacles to effective empirical studies of U.S. patents. These statistical barriers demonstrate that studies of general patent characteristics will offer few insights into the nature of individual patents or even fairly large ensembles of patents. The hope that patent policy, and particularly PTO review of patent applications, can be rationalized and calibrated using simple patent metrics is undone by implication. Instead, the value of most empirical work on patents will be limited to the broad aggregate trends that it reveals.

It is important to recognize that the analytical barriers identified here stem from the phenomena beings studied, not primarily from methodological shortcomings. The complex array of factors that drive innovation are not reducible to a few variables. Inventive success is simply a hard phenomenon to study given our current level of knowledge and the subject matter's inherent complexity, that is the inherent heterogeneity of patents (and inventions), the non-linearities of innovative success, and the dynamic nature of science.

The current excitement about using empirical methods has arguably obscured these basic limits. Given the void of data that we typically confront, this oversight is understandable—in the

²⁰⁶ The reason for this is actually quite straightforward. Even though valuable patents may be much more likely to have certain characteristics (e.g., large numbers of citations) than valueless patents, the far greater number of valueless patents more than offsets the differences in relative probabilities. Assume that you have 105 patents and that only ten are valuable. Assume further that valuable patents are five times more likely to have a characteristic than valueless patents. If valuable patents have a twenty-percent chance of having the characteristic, then valueless patents have only have a four-percent chance. However, in absolute numbers, four valueless patents will have the characteristic while only one valuable patent will have it.

realm of innovation policy, “patent statistics loom up as a mirage of wonderful plentitude and objectivity.”²⁰⁷ High hopes alone, however, cannot overcome the innate unpredictability of innovation and the highly skewed distribution of its benefits. Ironically, it may be a naïve enthusiasm for the rigor of empirical work that has caused academics to reify simple statistical metrics over the self-evident complexity of inventive processes.

Empirical studies are as challenging as they are alluring. It is therefore with some caution that we close by suggesting possible directions for new empirical work. Our analysis shows the difficulties entailed in interpreting patent data and the value of drawing on multiple types of data to construct a composite picture of patenting in a field. These findings suggest the need for more integrative studies that bring together patent data, economic studies, independent information on innovative output (e.g., publication data), and knowledge about the nature of research and development in the relevant field. The deficiencies we have identified also demonstrate the importance of expanding the methods available to study the interplay between patents and innovation. To this end, recent work using network theory and citation data is particularly important and promising.²⁰⁸

At a more fundamental level, this study has caused us to reassess whether the current focus on patents is ignoring other important opportunities for empirical work. The core issue is, of course, fostering innovation. Yet, as economists have long recognized, patents are a weak, though still valuable, measure of innovative output. The strategy of this paper rests, in part, on utilizing complementary metrics, such as research and development funding levels and numbers of entities entering the field, that are readily measurable, less subject to the severe interpretive challenges of patent-count data, and arguably more directly representative of trends innovation

²⁰⁷ Griliches, *supra* note 43, at 1661.

²⁰⁸ See Csárdi, *supra* note 112.

markets and networks. Importantly, this approach shifts the focus from aggregate innovative output to the dynamics of competition and support between individuals and institutions, whether public or private, engaged in research and development.

In one sense, this type of approach is unremarkable. Legal scholars and economists, for obvious reasons, have long been concerned about the interplay between patents and innovation, such as the impacts of patents on important follow-on research derivative of a keystone patent. In another sense, it differs fundamentally from a patents-oriented model by shifting the focus to the dynamics of innovation itself.²⁰⁹ For example, rather than asking questions about the general characteristics of valuable inventions, which are subject to large variances, we might focus instead on determining the characteristics of vibrant innovation markets, such as the balance between large and small companies or the number of entities in a field, and how these market characteristics evolve as a technology matures. Above all, this kind of strategy has the potential to allow us to identify more informative indices that are less likely to be subject to the chaotic behavior and uncertainties that are characteristic of patent metrics.

IV. Conclusions

The question motivating this Article—do we have too much of a good thing?—turns on the distribution of patents across distinct areas of research and development. Our data reveal the striking rise and fall in biotechnology patenting and the surprisingly diffuse and expanding

²⁰⁹ Ecology and evolutionary biology, which involve the study of natural processes of innovation and competition, may have a lot to teach us about these dynamics. Like innovation markets, ecosystems must balance the opposing processes of natural selection (*i.e.*, competition that ensures resources are used efficiently) and species diversification (*i.e.*, natural innovation), which is essential to its long-term resiliency and adaptability to environmental change. This balance is maintained by, among other things, through the existence of ecological niches, which spare less fit species from competition that they are destined to lose. Although the relationship is not a simple one, the balance between dominant and niche species, as well as of overall species diversity, are important indicators of ecosystem productivity and health. Similar indicators should exist for innovation markets and networks. Indeed, one can imagine how patents can preserve such a balance, by giving small players the space to operate, or undermine it, by giving already-dominant entities rights that allow them to dominate a field of invention completely. Drawing on work in these fields may aid legal scholars in identifying more viable indices.

patterns of patent ownership. We conclude that the lack of concentrated control, rising number of patent applications, and the continuous influx of new patent owners suggest that overall biotechnology innovation is not being impaired by the growth in patents issued each year.

Our analysis also reveals the many pitfalls of seeking to resolve this question at a synoptic level using simple metrics. In this sense, both the advocates of the anticommons theory and enthusiasts of patent characteristics err by oversimplifying the multi-dimensional character of patent dynamics. This Article has shown that identifying putative areas of dense patenting based on patent counts alone tells you very little about whether anticommons problems are likely to exist. Similarly, patent metrics based on general patent characteristics have only nominal predictive value as applied to individual patents. These observations reveal the risks of overly reductive theories and empirical methods, and their close parallels with each other.

The analytical obstacles we have described place a premium on developing new, creative approaches to studying U.S. patents. Complementary sources of information beyond legal sources can no doubt aid in filling some of these gaps. However, patents offer a unique source of data, particularly given the strong incentives companies have to withhold information, and they represent one of the most direct links between legal policies and innovation. It is therefore worth thinking carefully about how patents can be used more effectively in empirical work and, perhaps more ambitiously, about how rules governing patent content could be redesigned to facilitate research on and monitoring of patent trends.

V. Figures

Figure 1- Number of Biotechnology Patents Issued Yearly, 1990-2004 (n=52,039)

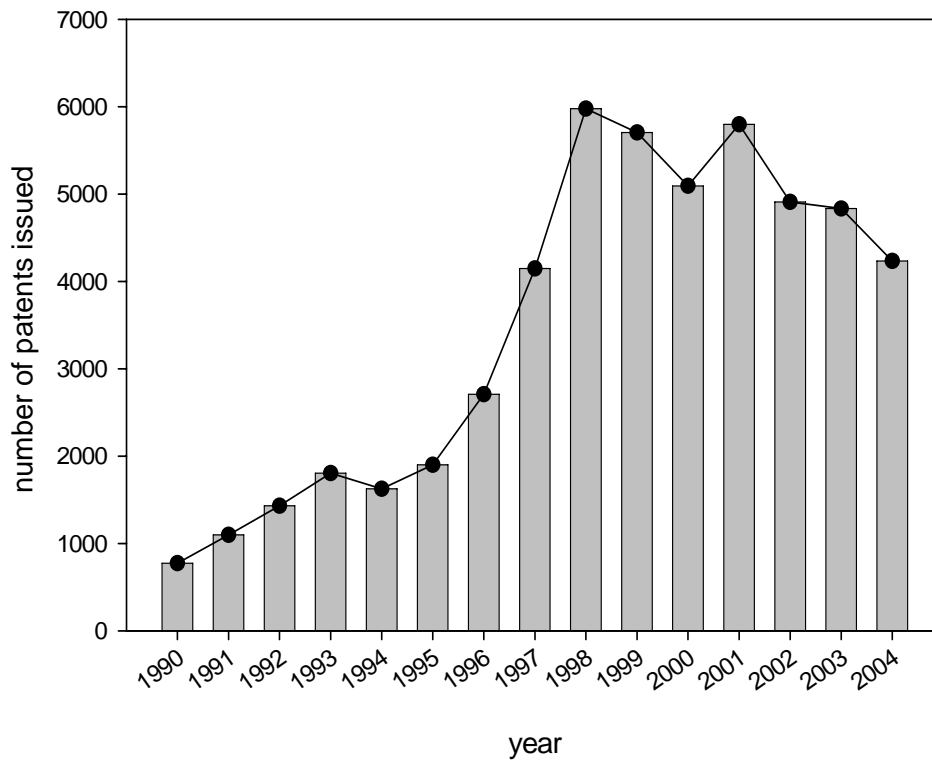


Figure 2- Number of Biotechnology Patents Issued per Year,
by Technology Group (n=52,039)

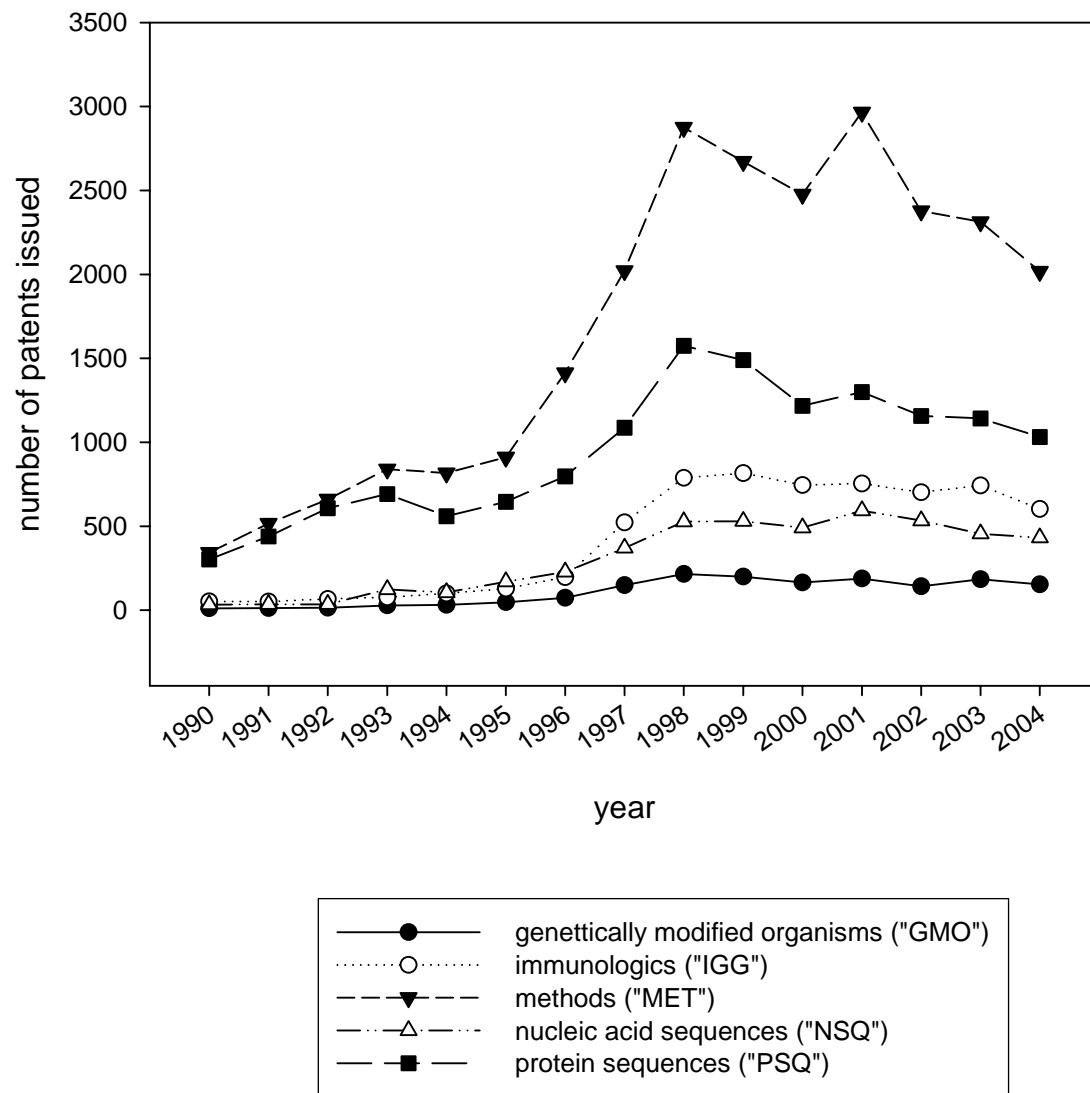


Figure 3- Number of Patents Issued in the Top USPTO Biotechnology Sublasses, by Year (n=52,039)

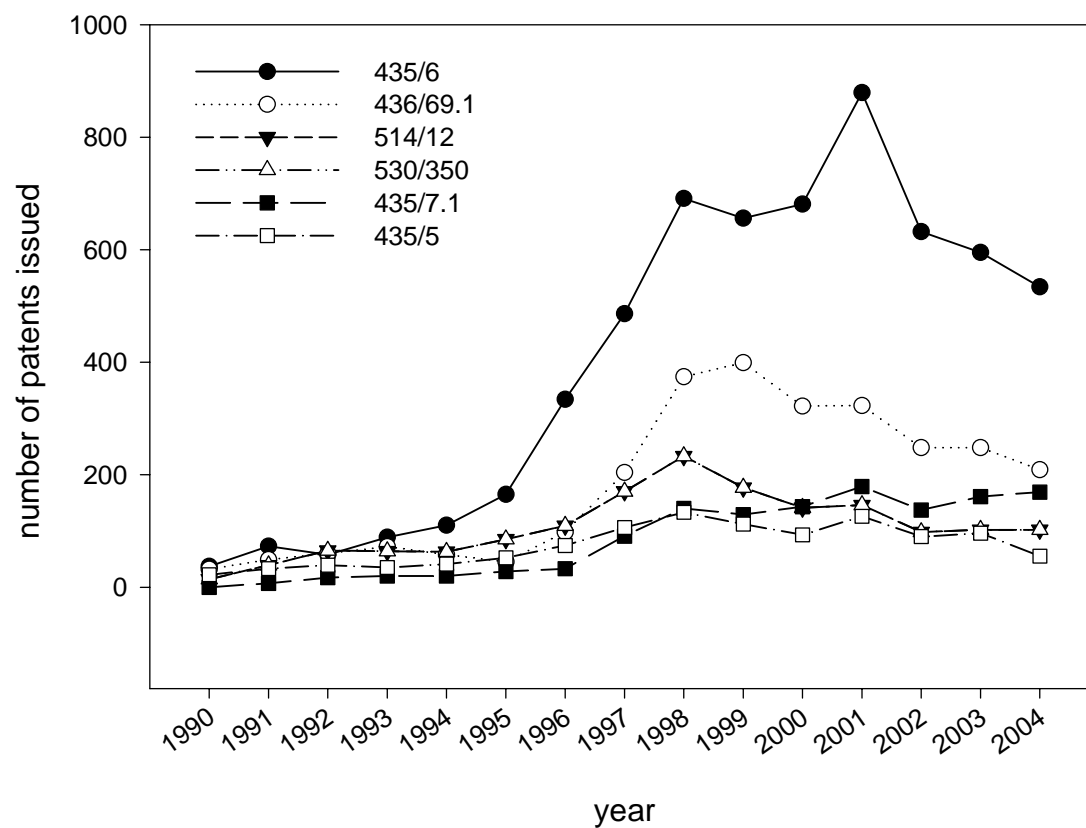


Figure 4- Patents Issued in the Top 30 Biotechnology Subclasses, 1990-2004 (n=52,039)

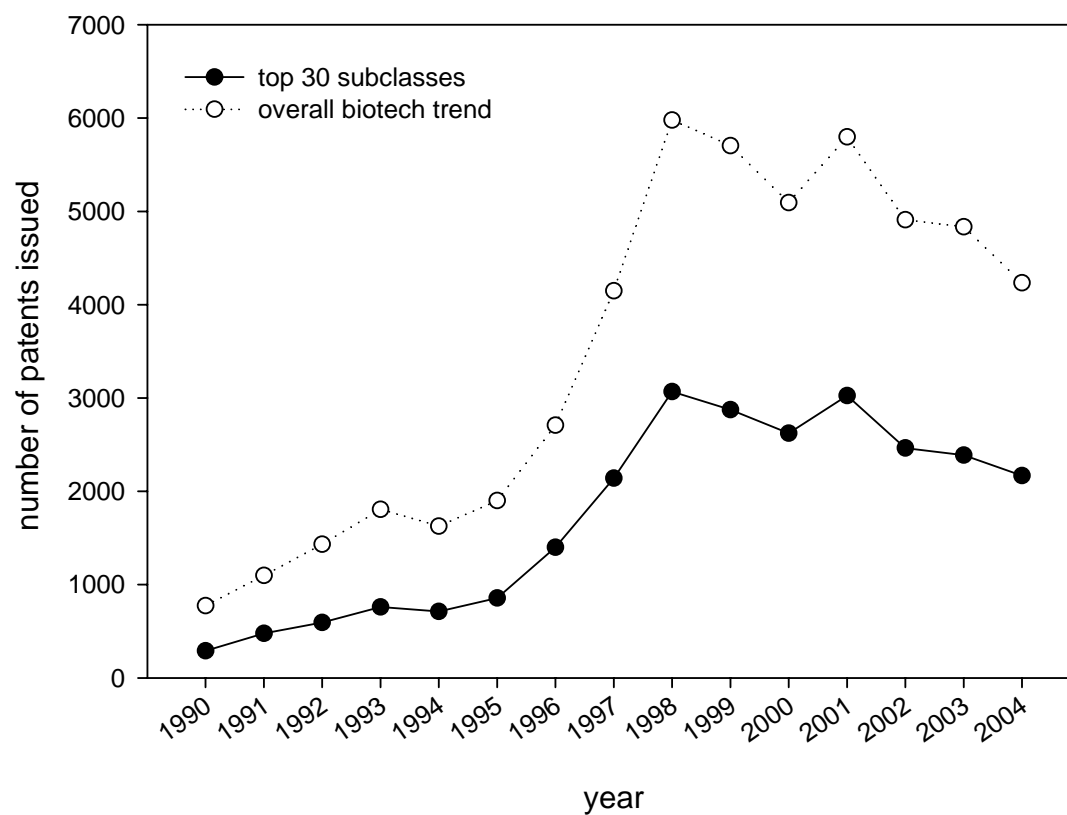


Figure 5- Biotech Patents Issued to Four Classes of Companies
Yearly, 1990-2004

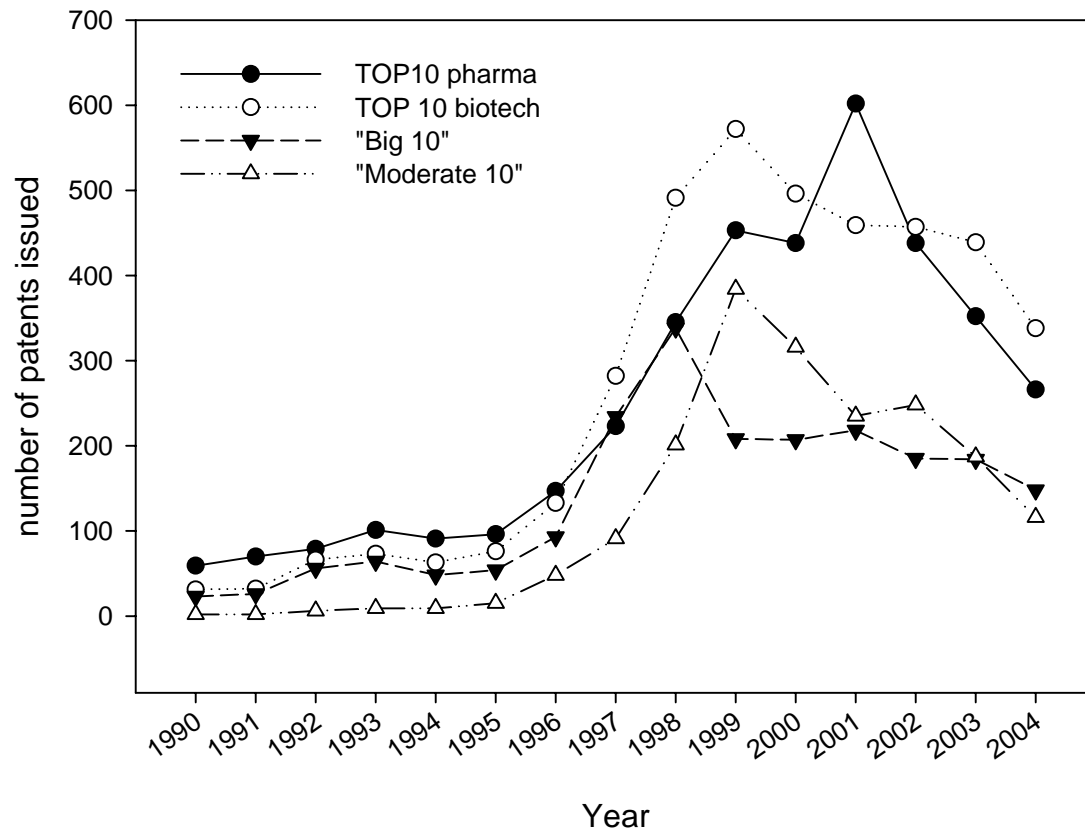


Figure 6- Biotechnology Patents Issued to University, Corporate, and Government Assignees Yearly, 1990-2004

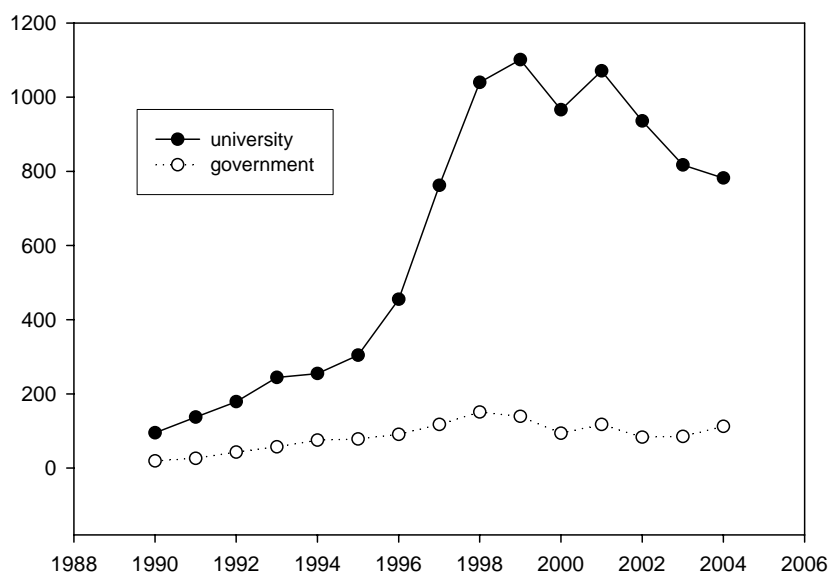
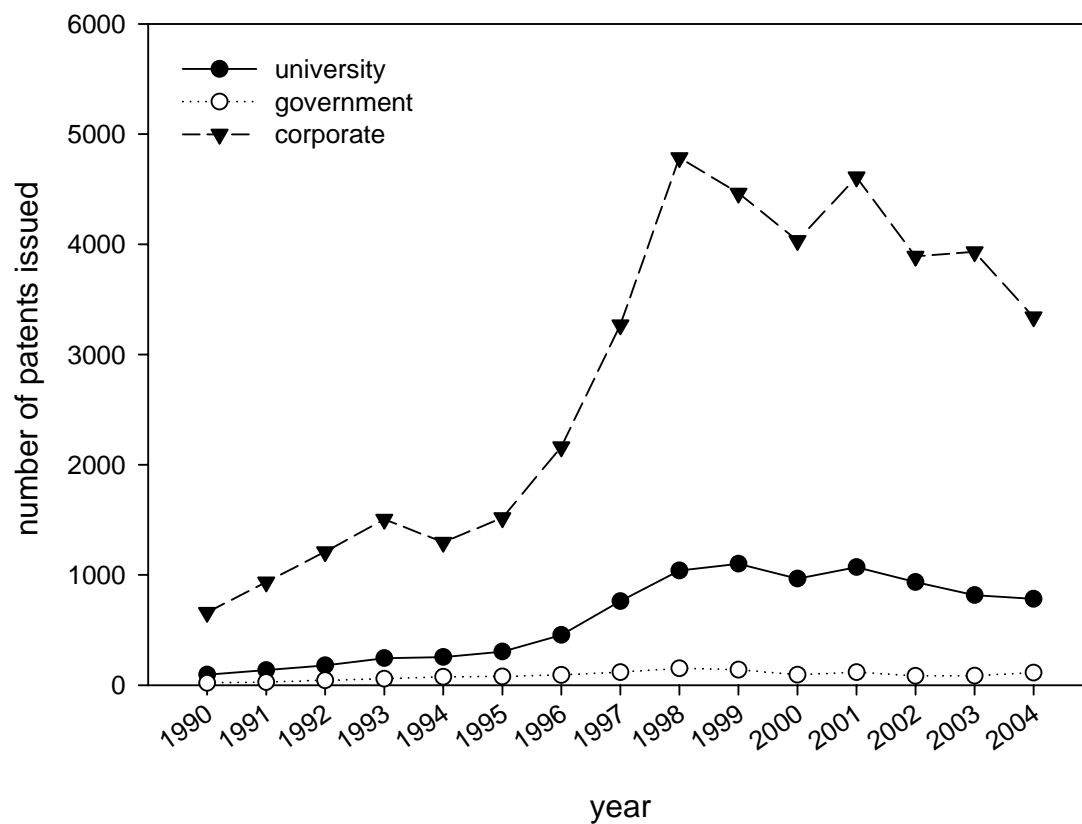


Figure 7- Assignments of Biotech Patents to Corporations, Universities and Government by Technology Group, 1990-2004

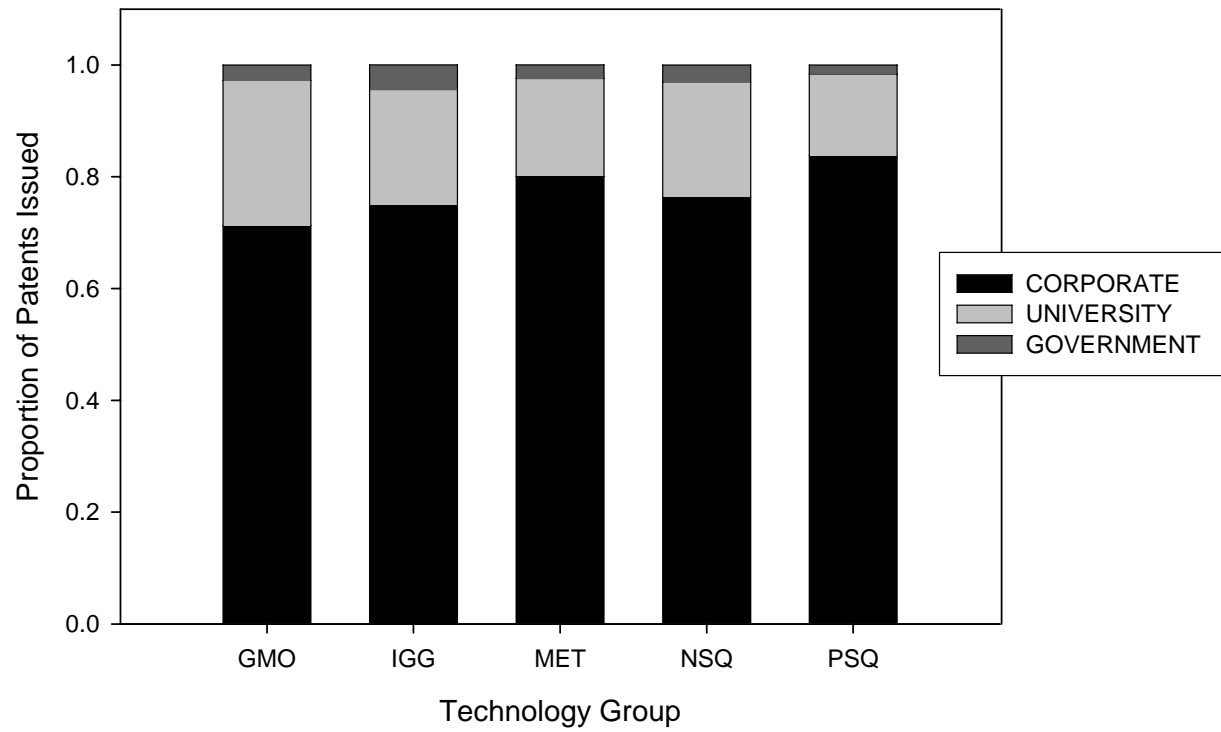


Figure 8- Number of Assignees of Biotechnology Patents Yearly, 1990-2004

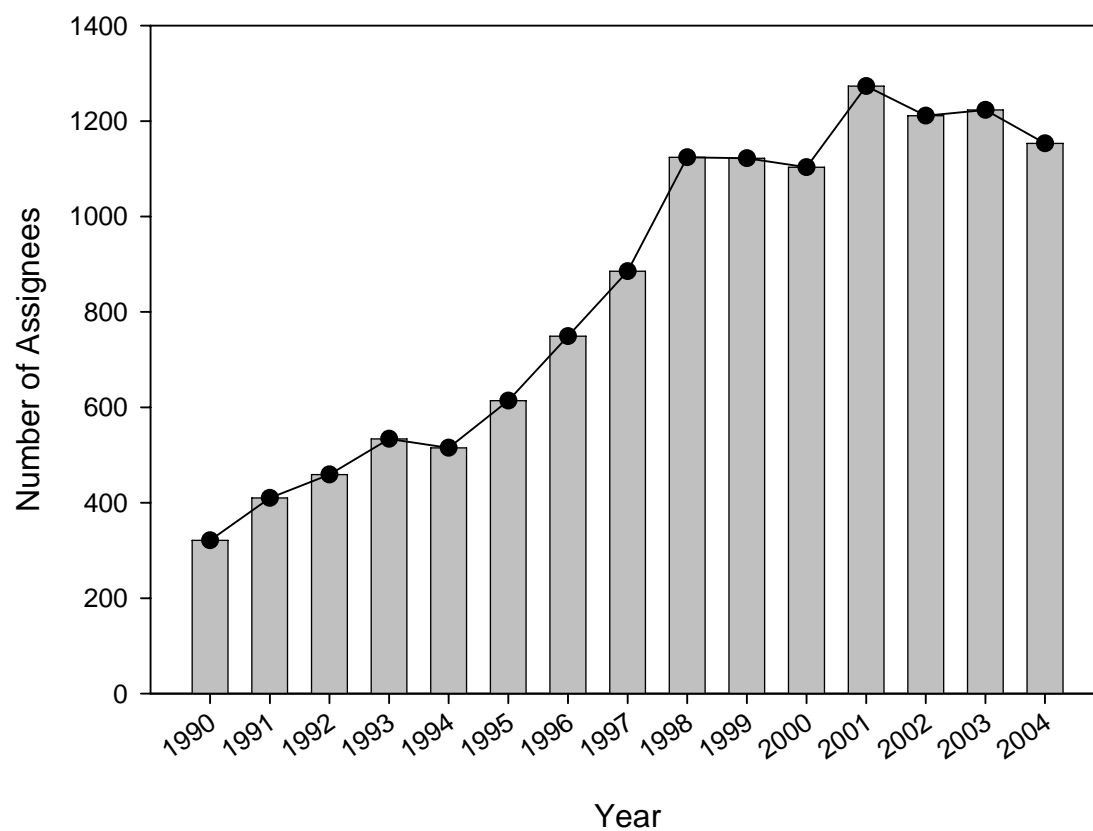
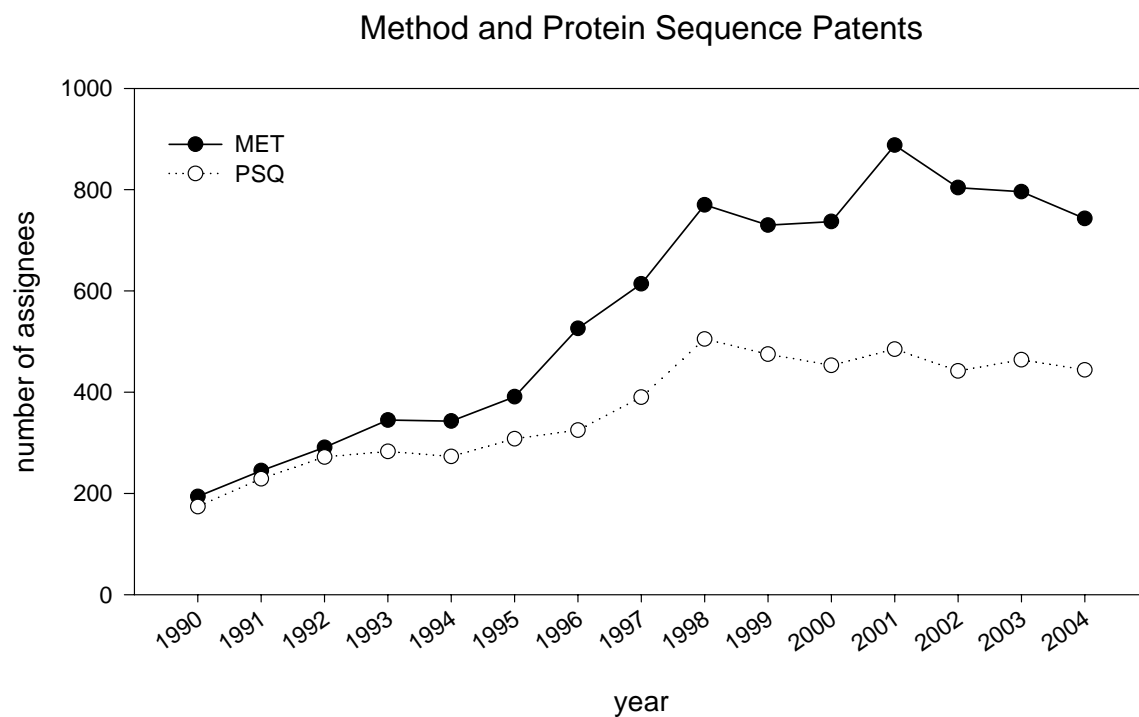


Figure 9a- Number of Assignees by Technology Groups Yearly, 1990-2004



Genetically Modified Organisms, Immunological, and Nucleotide Sequence Patents

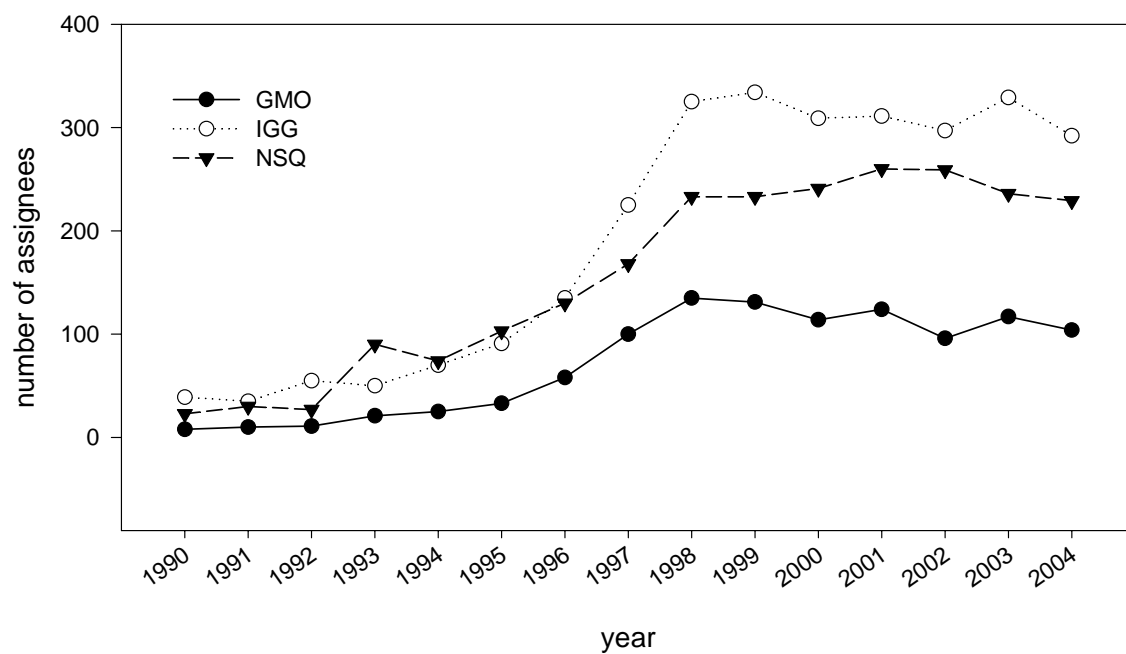


Figure 9b- Number of University and Corporate Assignees Yearly, 1990-2004

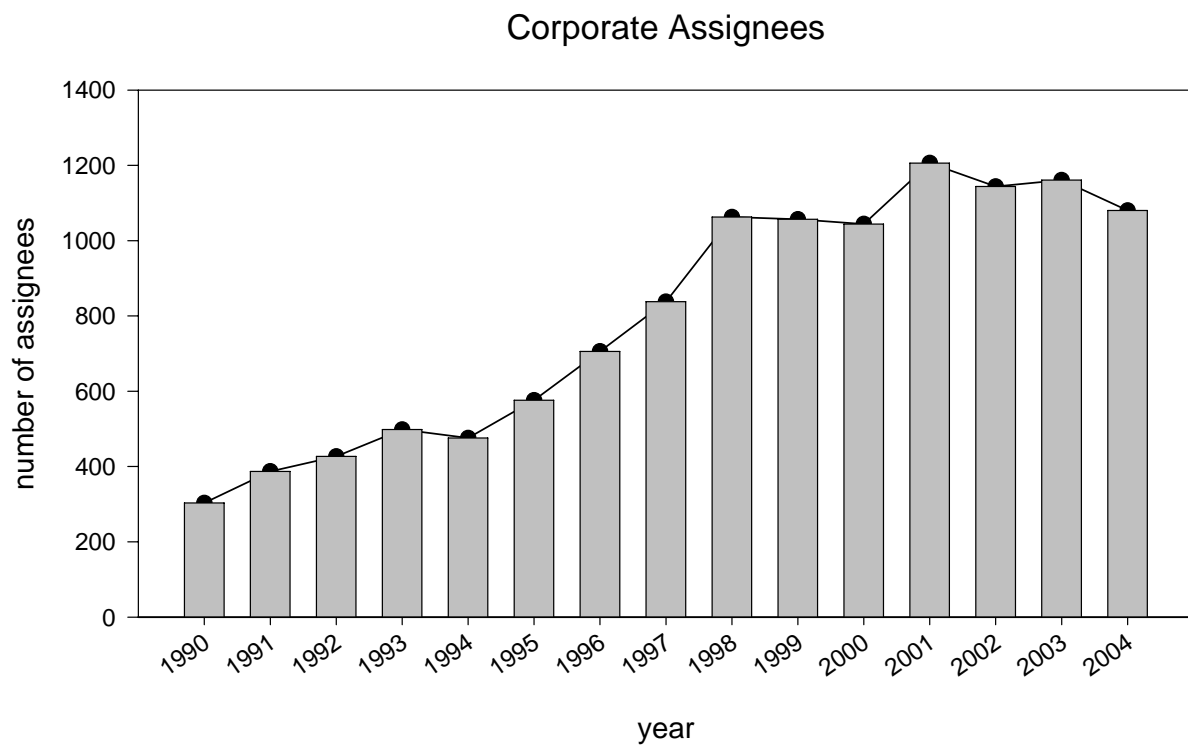
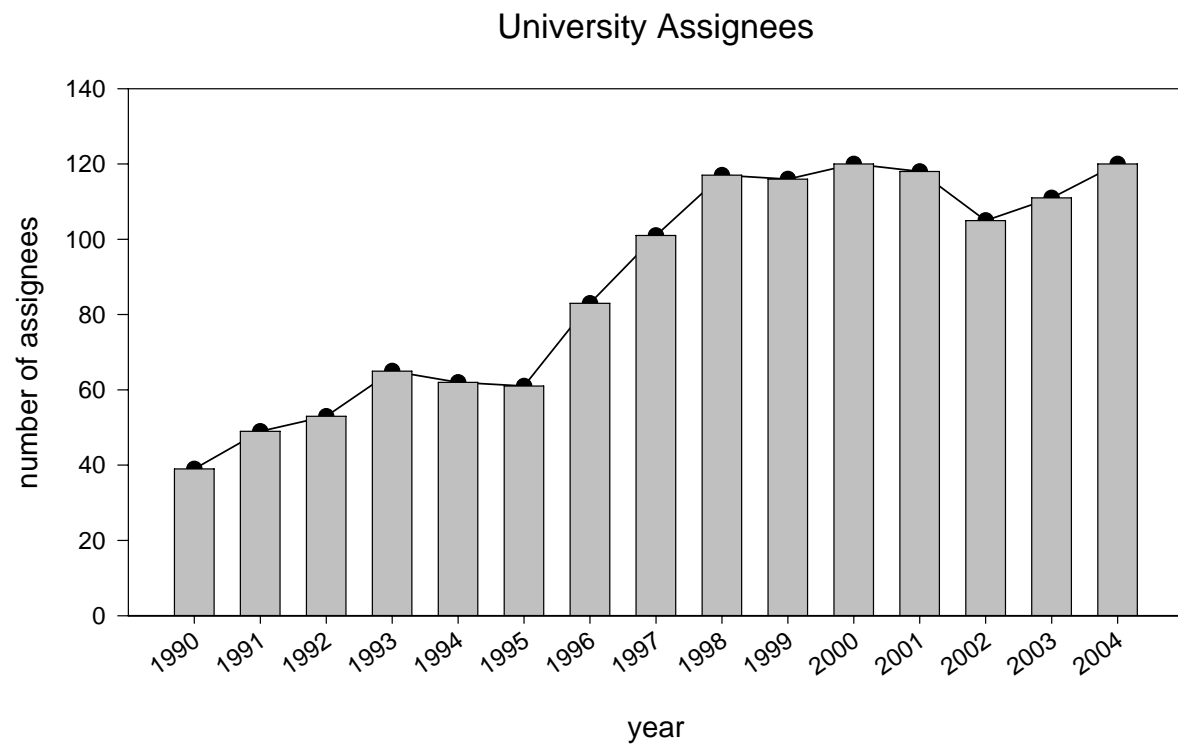


Figure 10a- Prosecution Time, Yearly, 1990 - 2004
trimmed to include 99% of values

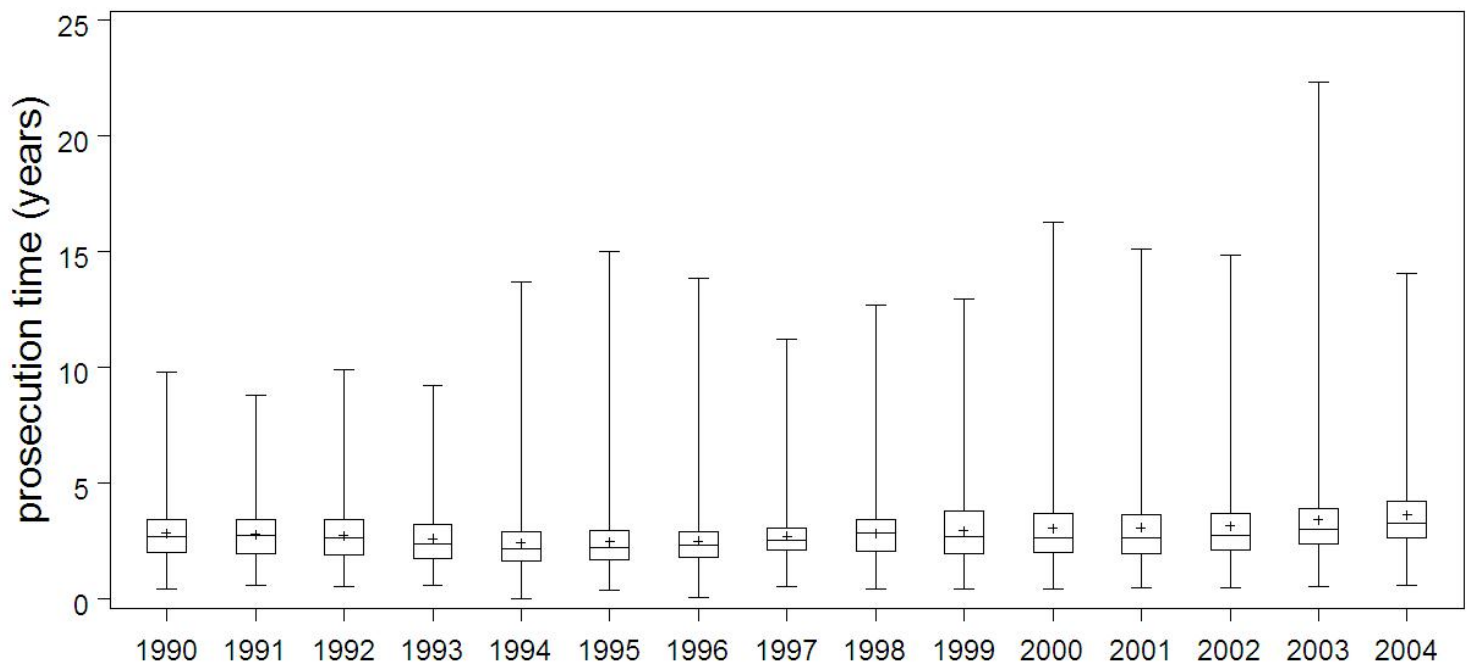


Figure 10b- Prosecution Time by Assignee Group 1990 - 2004
trimmed to include 99% of values

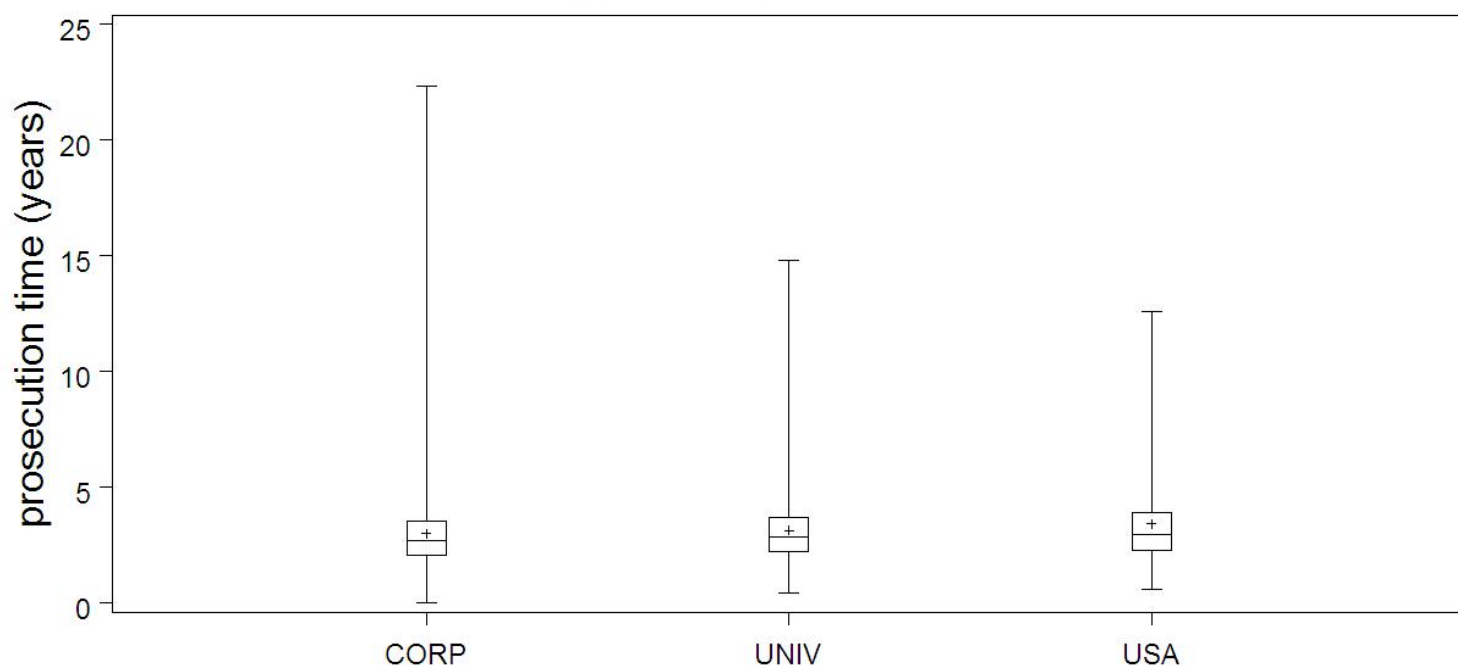


Figure 10c- Prosecution Time by Technology Group 1990 - 2004
trimmed to include 99% of values

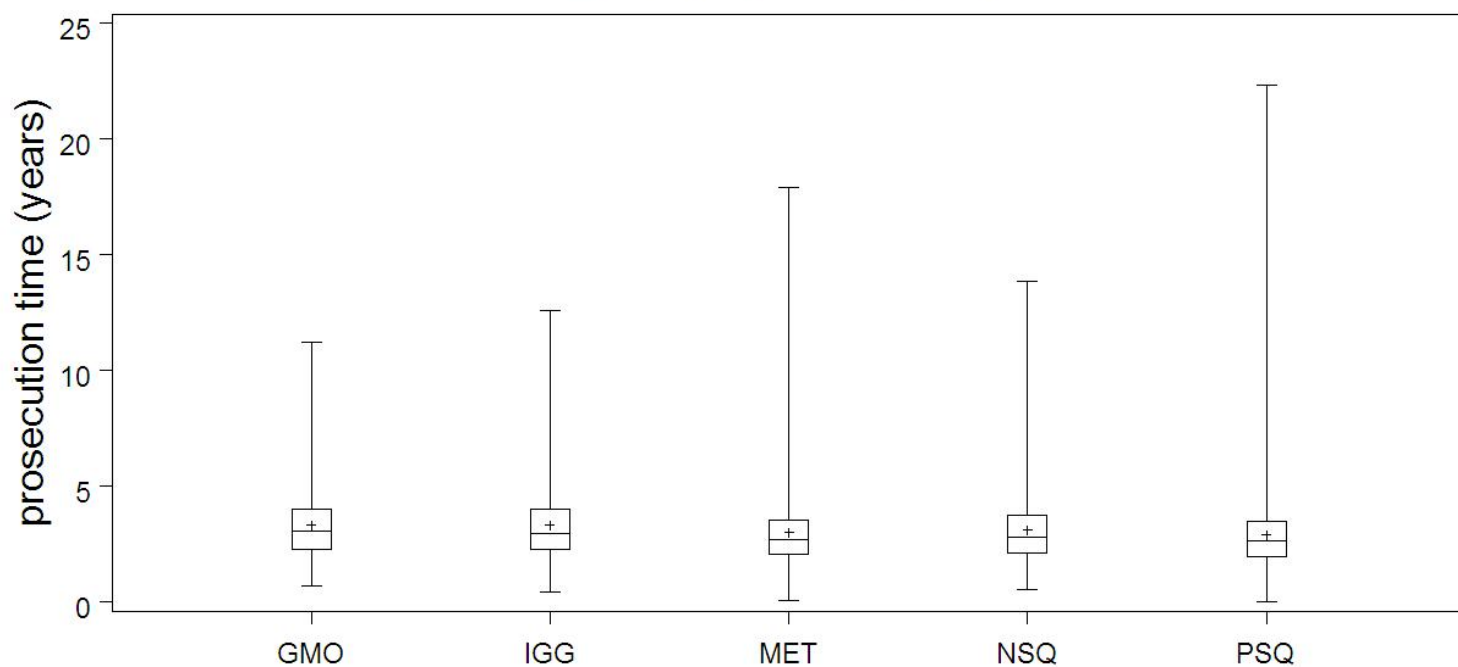


Figure 11a- Number of Claims per Patent, Yearly, 1990 - 2004
trimmed to include 99% of values

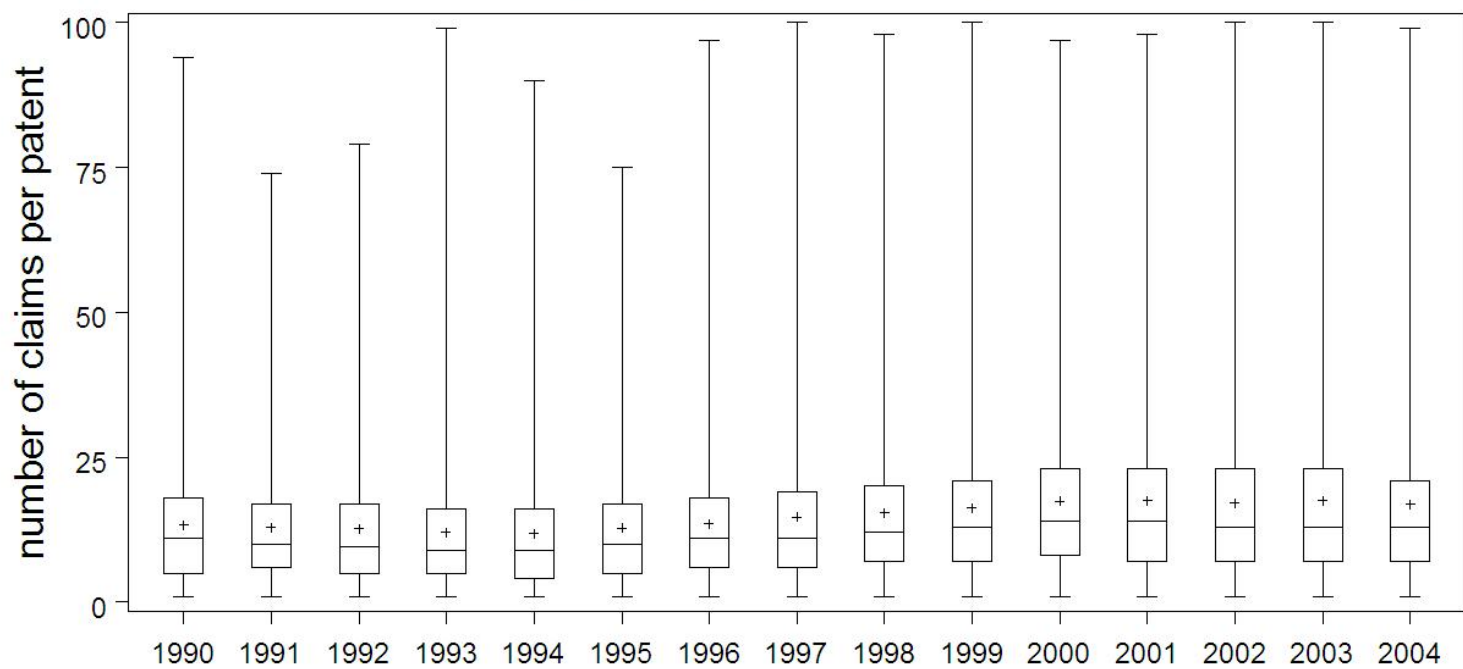


Figure 11b- Number of Claims per Patent by Assignee Group 1990 - 2004
trimmed to include 99% of values

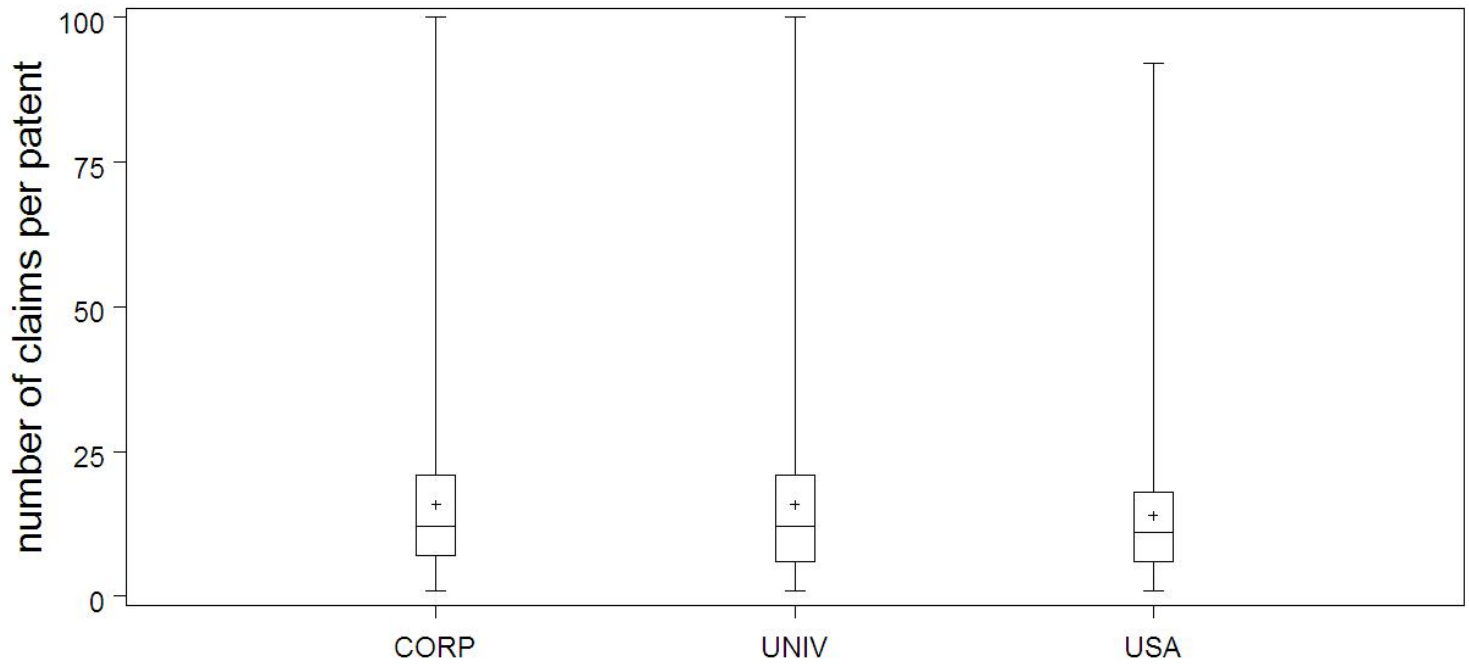
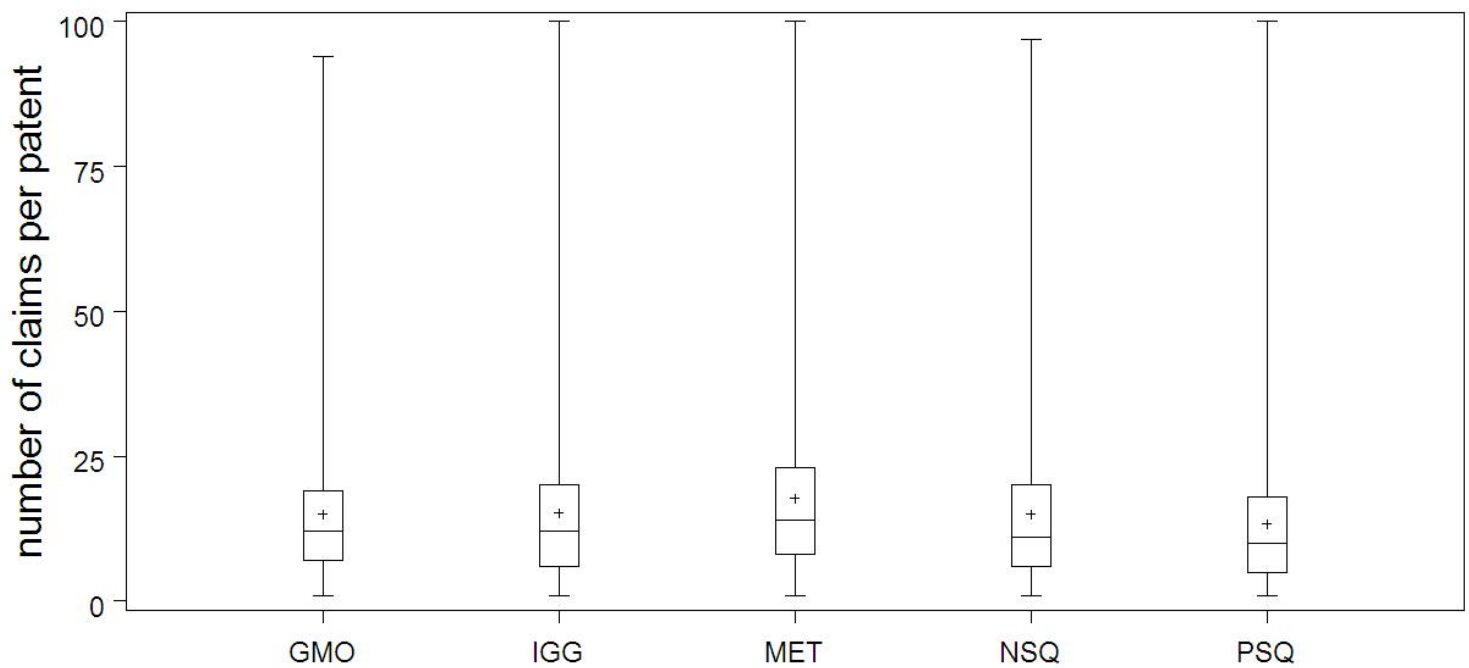


Figure 11c- Number of Claims per Patent by Technology Group 1990 - 2004
trimmed to include 99% of values



Appendix: Data and Methodology

The biotechnology patent database consists of 52,039 patents that issued between January 1990 and December 2004. These patents were collected from a larger patent database consisting of all patents, more than 950,000 in total, that issued between January 1990 and March 2005. These data were obtained directly from the PTO, which has electronic files dating back to at least 1963.²¹⁰ Once the data for all patents issued from January 1990 and March 2005 were collected, they were converted to a consistent format for analysis.²¹¹

The primary issue to resolve in constructing the database was the criteria we would use to determine whether to include or exclude a patent. As other commentators have acknowledged in similar contexts,²¹² the process of categorizing inventions into subject areas is part art and part science that entails difficult judgment calls, and one is unavoidably confronted with inventions that defy simple categorizing.²¹³ Making these decisions proved to be particularly difficult in areas where pharmaceutical research and development closely parallels, or overlaps, biotechnology work. We therefore focused our attention on ensuring that our data were not confounded by patents on inventions associated with traditional pharmaceutical research.

We adopted a two-stage approach: first, we constructed an unambiguously over-inclusive database of biotechnology-related patents using general PTO classes; second, we pared down

²¹⁰ The raw data were transferred to us in a large number of smaller, sometimes incompatible databases, which had to be converted to a “.dat” format to allow them to be combined and manipulated. The PTO provided the data for 1990-1996 on CD in a single database. Week-by-week data for 1996-2005 were downloaded from the .ftp site operated by the PTO and converted from either .xml or .sgml format to .dat or .gbk format using PERL scripts publicly available from the PTO.

²¹¹ This was no small undertaking because the PTO databases varied over this period and often were not inter-compatible. All raw data files were converted to text files, and these files were then concatenated into four major data tables, with each table linked by the patent number for each patent. The four data tables are (1) Patent Information (“Patinfo”), (2) Primary Class (“Class prim”), (3) “Class,” and (4) Assignee Name (“AsnNam”). A number of minor modifications were also made to compile the data properly. In particular, a few entries in each file were corrupted and deleted (<1%) and spacing within columns in each file were adjusted to be uniform (*i.e.*, from a variable format to an eighty-character fixed-length format).

²¹² Allison, Valuable Patents, *supra* note 5, at 2109.

²¹³ Line-drawing is complicated further by the byzantine PTO classification scheme, which includes about 400 main classes and more than 120,000 subclasses. Hall, *supra* note 8, at 13.

this over-broad dataset using several criteria. The first broad database was drawn from forty-nine PTO classes and contained a total of 89,619 patents.²¹⁴ For construction of the second, narrower database we employed three complementary strategies. We examined the PTO subclasses in which well-established biotechnology companies were obtaining patents,²¹⁵ reviewed the subclasses that the PTO treats as biotechnology fields, and undertook our own independent assessment of potentially relevant PTO subclasses to determine their relevance. The process was an iterative one in which we started with the first strategy and then compared these results against the PTO classification system. Where there were differences, or we had our own reservations about specific subclasses, we undertook our own detailed analysis.

After several rounds of this process, our analysis converged on a final database comprised of patents whose primary PTO classification falls under one of 704 PTO subclasses.²¹⁶ As general rule, we sought to be inclusive as possible, but where a subclass might include a few biotechnology inventions but otherwise be dominated by inventions in other fields, we opted to exclude those subclasses.²¹⁷ Most differences between the subclasses included in our database and the PTO's biotechnology art unit-based designation of biotechnology subclasses derived from our efforts to avoid overlap with pharmaceutical inventions and our decision to exclude agricultural biotechnology inventions from the database.²¹⁸

²¹⁴ The specific PTO classes included were 023, 047, 071, 111, 117, 127, 128, 201, 202, 203, 204, 205, 210, 260, 422, 423, 424, 435, 436, 504, 514, 516, 518, 530, 532, 534, 536, 540, 544, 546, 548, 549, 552, 554, 556, 558, 560, 562, 564, 568, 570, 585, 600, 604, 606, 607, 800, 930, 987. These were subsequently winnowed down to just the following six classes: 424, 435, 514, 530, 536, and 800.

²¹⁵ The companies used for these purposes included the following: Amgen, Genentech, Gilead, Chiron, Genzyme, Biogen, IDEC, Millenium, Celgene, and Medimmune.

²¹⁶ To provide an admittedly low baseline, the NBER study includes only two classes, 435 and 800, which account for less than fifty percent of the patents in our database. Hall, *supra* note 8, at Appendix 1.

²¹⁷ We also had to contend with "orphan" subclasses, which were operational during the early part of the study and abandoned by the end of the study. To the extent that we could, we re-assigned patents to the new subclasses, but this was a laborious process and often not worth the effort given the relatively small number of patents involved.

²¹⁸ Two other related areas of invention that we were careful to avoid were organic compounds and surgical procedures, both of which were mixed into the classes covering biotechnology inventions. We decided to exclude

The database contains general information on each patent issued, and can be easily updated to include patent data as it is added to the PTO website. The specific categories of information include the following: patent number, application number, patent title, issue date, assignee, primary subclass, other subclasses, art unit, number of claims, number of figures, primary examiner, assistant examiner, and search classes. We also augmented the data we received from the PTO by integrating data on patents issued from January 1990 through December 1999 from a database, “The NBER Patent Citations DATA File,” compiled by several researchers associated with the National Bureau of Economic Research (“NBER”).²¹⁹ These data gave us two additional categories: (1) number of citations made by each patent granted during the 1990s, and (2) number of citations received by each patent during the 1990s.

We organized the data into discrete data tables to run comparative analyses. The first of these was structured around distinct, albeit still somewhat artificial, biotechnology subfields. Drawing on intermediate categories evident in the PTO classification system, we divided the dataset into five distinct areas of biotechnology research and development: (1) measuring and testing processes (“MET”),²²⁰ (2) polypeptide (i.e., short protein subsequences) and protein sequences (“PSQ”), (3) nucleotide (i.e., DNA, gene) sequences (“NSQ”), (4) immunological processes and compounds (“IGG”),²²¹ and (5) genetically modified organisms (“GMO”). These categories were chosen because they aligned reasonably well with the patents in the database, because they made sense scientifically, and because of their importance in the field of

agricultural biotechnology inventions because of the divide—political, policy, and economic—that exists between the agricultural and biomedical biotechnology sectors.

²¹⁹ Hall, *supra* note 8.

²²⁰ Examples of these inventions include genetic test kits, gene chips, micro-array technology, and polymerase chain reaction technologies.

²²¹ The most important class of inventions in this area is monoclonal antibodies, but it also includes many other related technologies as well.

biotechnology research and development.²²² Each category is treated as exclusive, although some inventions could be placed in more than one subfield. While this ordering could bias our results, we believe that, given the large number of patents studied, its effects are likely marginal.

We created one final subdivision of the data for our analysis. Patents were categorized according to their ownership status, that is whether the assignee of the patent was the federal government, a university, or a corporation.²²³ In addition, to obtain a sense of possible substructures within the corporate category, which is by far the largest one, we created four groups of companies: (1) ten large biotechnology companies based on an examination of revenue, profits, and number of employees;²²⁴ (2) ten, somewhat randomly chosen, mid-range biotechnology companies, virtually all of which were operating at a loss;²²⁵ (3) the ten biotechnology companies with the largest patent portfolios;²²⁶ and (4) the ten pharmaceutical companies (*i.e.*, not solely biotechnology) with the largest patent portfolios.²²⁷ These groupings are included in comparative analyses.

We conducted three central analyses of the data.²²⁸ First, similar to a number of existing studies, we examined trends for specific characteristics of the patents, these included the number

²²² Nevertheless, despite our best efforts, categories do unavoidably overlap—for example, specific polypeptide or nucleotide sequences may be claimed as probes used in measuring or testing inventions.

²²³ We do not distinguish between foreign and U.S.-based assignees because the PTO data we obtained did not include this information. In addition, given the limits of the data base, we assume that all biotechnology patents are assigned, which we believe is reasonable given the high costs and resources required to do biotechnology research.

²²⁴ The group of large, successful companies includes Amgen, Genentech, Biogen IDEC, Genzyme, Serono, Chiron, Gilead, Affymetrix, Medimmune, and Cephalon. All of these companies are profitable, with net profits in 2004 ranging from \$25 million to \$2.36 billion.

²²⁵ The mid-range biotechnology companies include Progenico Pharmaceuticals, Isis Pharmaceutical, Incyte Corp., Human Genome Sciences, Xoma, Ltd., Intermune, Amyline Pharmaceutical, Inc., Myriad Genetics, Alkermes, Inc., and United Therapeutics. Only one of these companies, United Therapeutic, is currently reporting a net profit, although all reported sales in 2004 of from about \$10 to \$150 million.

²²⁶ The biotechnology companies with the ten largest patent portfolio include Genentech, Incyte Corp., Isis Pharmaceutical, Chiron, Human Genome Sciences, Amgen, Zymogenetics, Immunex, Applera, and Corixa.

²²⁷ The pharmaceutical companies with the ten largest patent portfolio include Smithkline, Merck, General Hospital Corporation, Eli Lilly, Aventis, Schering, Millennium, Novo Nordisk, Dupont, and Roche.

²²⁸ All of the analysis was conducted in SAS v. 9.1, licensed to the University of Arizona.

of claims, the length of patent prosecution, and the number of citations made and received.²²⁹

Second, we analyzed trends based on assignee type (i.e., corporations, universities, federal government) and assignee size, as well as the thirty assignees with the largest number of patents.

Third, we evaluated patent trends based on the five biotechnology subfields listed above—measuring and testing processes, protein sequences, nucleotide sequences, immunological processes and compounds, and genetically modified organisms. All of the data were evaluated to determine trends over time, and several studies were conducted to examine the interplay between three patent characteristics (*i.e.*, number of claims, citations, length of patent prosecution).

The analyses of general trends were supplemented by two studies designed to evaluate the distribution of biotech patenting among patent owners and across distinct areas of research and development. The analysis ranged in scope from studies of discrete areas of research and development (e.g., diabetes research) to the full complement of biotechnology patents in the database. Patent ownership patterns were using our data on assignees. The distribution of patents across different subject areas was evaluated using PTO subclasses. In addition, we utilized a variety of statistical methods, particularly standard hypothesis testing and linear regression, to examine the results of three existing studies of U.S. patents. The details of these methods and their implications are integrated into our discussion of the data.

²²⁹ We chose not to study the other patent characteristics available in the database for three primary reasons. First, studying certain variables, such as search classes, would have added nothing to the analysis because examiners follow relatively rigid PTO rules that are determined by the art unit in which the examiner works. Second, we simply were not interested in studying some of the data (e.g., PTO examiner performance or statistics). Third, some data (e.g., other subclasses listed in each patent) simply would have been very difficult to study in a meaningful way without substantial additional work and the benefits we perceived were relatively small.