Making the Cut:
The Rate and Direction of CRISPR Innovation

by

Samantha Zyontz

B.S. The College of William & Mary, 2000

M.S. Kellogg School of Management, Northwestern University, 2006

S.M. Sloan School of Management, Massachusetts Institute of Technology, 2016

SUBMITTED TO THE SLOAN SCHOOL OF MANAGEMENT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN MANAGEMENT

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2019

Signature of Author: Signature redacted

Department of Management
May 3, 2019

Certified by: Signature redacted

Scott Stern
David Sarnoff Professor of Management of Technology
Professor, Technological Innovation, Entrepreneurship, and Strategic Management
Thesis Supervisor

Accepted by: Signature redacted

Ezra Zuckerman Sivan
Alvin J. Siteman (1948) Professor of Entrepreneurship and Strategy
Deputy Dean
Faculty Chair, MIT Sloan PhD Program

Making the Cut:
The Rate and Direction of CRISPR Innovation

by

Samantha Zyontz

Submitted to the Sloan School of Management on May 3, 2019
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Management

Abstract

This dissertation explores, in real time, key institutional factors contributing to the diffusion and impact of a breakthrough technology from its very first days. The studies combine rigorous quantitative empirical methods with a deep understanding of the institutions of a novel setting that allows for a nuanced picture of the actors, institutions, technologies, and rules necessary to make recommendations on policies and strategies for the diffusion of emerging innovations.

The first chapter examines whether the introduction of a breakthrough technology, the CRISPR DNA-editing system, affects the trajectory of a scientific field through project selection and new entry. Using proprietary data from the primary distributor of CRISPR to academic scientists, Addgene, the study shows that the relative proportion of scientists focusing on editing mammalian cells after the introduction of CRISPR increased over their counterparts working in bacteria and other eukaryotes. The shift towards mammalian research may result mostly from entry of new authors.

The second chapter (with Neil Thompson), explores whether characteristics of individual scientists who experiment with CRISPR differ from those who incorporate that experimentation into a new project. Using Addgene data we separately observe both groups by matching CRISPR orders to scientists' publication histories. We find that some characteristics (e.g., proximity to the discoverers) do not impact experimentation but do influence the ability to publish, empirically showing that access to a complex new tool does not automatically translate into the ability to use the tool.

The third chapter builds on the previous two by noting that many new tools require specialized complementary know-how to be applied effectively and delving into how teams form to acquire that know-how. Teams in any research domain face the tradeoff of either acquiring this know-how themselves or working with scarce external tool specialists who also have a choice over domain teams. CRISPR enables identification of external tool specialists on research teams by exploiting natural difficulties of applying the tool across disease domains. External tool specialists appear more often in teams for difficult diseases, especially in subsequent innovations, suggesting that external tool specialists may be more attracted to complex but influential problems.

Thesis Supervisor: Scott Stern
Title: *David Sarnoff Professor of Management of Technology*
Professor, Technological Innovation, Entreprenеurship, and Strategic Management

Acknowledgements

Getting a Ph.D. less about the destination than the journey, and what a long strange trip it's been. If someone had told me in the beginning that I would study one of the biggest breakthroughs in modern genetics or work with people who believe as many as six impossible things before breakfast and make them happen, I would have laughed. Yet, here I am, thankful for my time at MIT Sloan. This was not a solitary trip though, and I owe everything to the amazing individuals who were there along the long, winding, sometimes difficult, but fascinating path. This dissertation is as much theirs as mine.

As a graduate student, I could not ask for a more outstanding advising committee. Professors Scott Stern, Pierre Azoulay, and Jeff Furman cared both professionally and personally, were generous with their time, and encouraged me to push beyond what I thought were my limits to discover a researcher I did not know existed. More importantly, they let me explore my own paths in my own way to become the scholar I am today; for that, I am forever grateful.

Without Scott, I would never have entered the Ph.D. program in the first place. He has been quietly guiding my career in the background after meeting at Northwestern over a decade ago. The consistent support and guidance from a dissertation chair as impressive as Scott has me humbled, grateful, and wanting to push myself harder. If I have seen further, it is because I have been able to stand on his "giant" shoulders, benefiting from his creativity and engagement with my work. There are not enough words to thank him for believing in me all this time. I only hope to become even a fraction of the scholar, teacher, mentor, and loyal supporter that he is.

Pierre's counsel has improved my work immeasurably. He taught me to ask better questions, helped me to identify holes in my arguments, and gave me the tools necessary to find the most rigorous answers, always with a smile. Pierre continuously encouraged me to achieve at the highest levels, and no one was happier when I succeeded. In addition to all of his guidance, Pierre, his wife Andi, and their amazing daughters allowed me to be an "adopted" part of their family. Every time I am welcomed into their home, it is like being with my own family. Their support helped me get through some tough times in the program and I will always be deeply thankful for their kindness.

Jeff was not only an invaluable sounding board but is one of the most giving people I have ever met. He provided insightful advice on how to frame my papers or presentations based on his own experiences (even last minute) and was always ready with a kind and encouraging word. I am grateful for every call we had on his commute and for every coffee at BU. Everyone should be so lucky as to have a Jeff Furman in their career, a cheerleader who builds up his students when they need it, provides critiques if things are going off-course, and creates opportunities for exposure to a larger audience. Meeting Jeff is a true highlight of the last six years, so I can forgive him for being a Flyers fan!

Although they were not officially on my committee, I am indebted to many other faculty members for their encouragement and guidance as I formed my CRISPR work and learned how to become a scholar. Brainstorming new ideas with TIES professors Matt Marx, Olenka Kacperczyk,

and Danielle Li is one of my favorite activities and their insights both shaped and improved my work. Professors Susan Silbey, Roberto Fernandez, Ray Reagans, and Ezra Zuckerman came to my presentations at ESWG and offered excellent advice from different perspectives that is reflected in this dissertation. I'm especially thankful that Susan and Roberto have been behind me the whole way. Professors Rosemarie Ziedonis, Michelle Gittelman, Heidi Williams, Florenta Teodoridis, and Julian Kolev's enthusiasm for my research and advice for navigating graduate school kept me moving forward every day. A special thank you to Professor Mercedes Delgado, who is always my champion. I will treasure the many coffees, lunches, dinners, we had over the years as we worked weekends and late into the evening. Her work ethic and friendship are inspirations.

My faculty advisors and mentors were instrumental throughout my journey, but the CRISPR work itself would have never occurred without the people who introduced me to and taught me about the technology. Many graciously offered their time and expertise, but a few must be mentioned. A chance conversation with Neil Thompson in the Fall of 2013 instigated all the work contained in this dissertation, including our co-authored chapter. Our third colleague, Aditya Kunjapur, not only helped teach us about genetic engineering, but made us aware of CRISPR with an off-hand remark. As an empiricist, I only heard "exogenous shock" when Aditya explained no one had expected CRISPR or knew what it was capable of in 2014. Aditya's scientific expertise and his willingness to review all of the science sections for accuracy made this project possible for me. The project also would not have been nearly as successful without support, enthusiasm, and data from Joanne Kamens, Executive Director of Addgene. Our relationship with Joanne and the Addgenies has been more fruitful than I ever could have hoped. It has been a wonderful experience to watch them grow from a few rooms in Kendall Square to having their name on a building in Watertown in a few short years and I'm glad to have been along for the ride.

This dissertation is also the product of countless hours with the most amazing student colleagues I could have ever asked for. I often say that the best thing about getting a Ph.D. at Sloan is the collaborative, brilliant, helpful, and fun students because my friends have proven this to me over and over again. Without them, this journey would have been much more difficult and a whole lot less entertaining. Some of the people I have to thank the most are in my original cohort: Ankur Chavda, Danny Kim, Minjae Kim, Rebecca Grunberg, James Riley, and Daniel Rock. Together we helped each other through thousands of pages of reading, impossible problem sets, exams, Generals, second year papers, TAs, RAs, good research ideas, not-so-good research ideas, job market papers, the job market, dissertation defenses, and the transition out of graduate school. Interspersed with the hard work were dinners, birthday parties, movies, baking extravaganzas, and all the little celebrations along the way. We were each other's harshest critics and most enthusiastic cheering squads. We pushed each other and rose together – I owe them everything and will be rooting for them as I know they will be for me in this next journey.

The TIES students form a special community that persists even after graduation. Alumni Michael Bikard, Eunhee Sohn, Abhishek Nagaraj, Dan Fehder, Josh Krieger, and Jorge Guzman have spent countless hours commenting on my work, helping me navigate through the program, or just offering me a friendly ear. I am so glad for the kind, helpful, and supportive culture they helped

create and I hope that I have been able to pay some of that kindness forward to the TIES students who came after me. In addition to Ankur and Danny, Hyejun Kim, Caroline Fry, Michael Kearney, Jane Wu, Wes Greenblatt, Soomi Kim, Lindsey Raymond, and Luca Guis not only endured my endless discussions about CRISPR for years (Wes gets a special thank you for further letting me exploit his medical expertise to refine my models and for putting up with my jokes), but they created a place I wanted to be. I will forever be grateful for all the TIES conversations, debates, laughter, and comradery whether it was in an office, at a wedding reception, or just playing bocce ball on the front lawn of Sloan.

The group of fierce and brilliant women scholars I am honored to call my friends Christine Riordan, Brittany Bond, Melissa Staha, Becky Karp, Aruna Ranganathan, Julia DiBenigno, Ceci Zenteno, Kimia Ghobadi, Hye Jin Rho, Ozge Karanfil, Maja Tampe, Maite Tapia, Yuly Fuentes, Mabel Abraham, Emily Truelove, Vanessa Conzon, Jenna Myers, Heather Yang, Summer Jackson, Mahreen Kahn, Duanyi Yang, Carolyn Fu, Tatiana Labuzova, Zanele Munyikwa, and Claire McKenna, helped me navigate my career, provided a sounding board for ideas, and extended their friendship whenever it was needed. These women showed me that there are many ways to be a successful scholar and added so much joy to my journey. I'll fondly remember the monthly Skype calls, research sessions, exercise classes, Sunday dinners, craft nights, festivals, concerts, bridal showers, and baby showers that formed the fabric of my time at Sloan. Christine was a constant light throughout the program. I will dearly miss our exercise sessions, walks around the pond, dinner conversations, concerts, craft nights, gardening parties, and baking extravaganzas. Brittany, my roomie, kept me grounded. Our conversations always provided me with insights I would have never discovered on my own and our dinner and movie nights were always a highlight of my time here. Melissa was my friend even during the worst moments. Our weekly dinners kept me sane and our conversations are endlessly fascinating. I am in awe of her perseverance and curiosity. Here's to many more memorable summer vacations in P-town!

So many other students have spent time talking with me about my work and just making life at Sloan memorable, including Michael Wahlen, Erik Duhaime, Taylor Moulton, Tristan Botelho, Alex Kowalski, Simon Friis, Avi Collis, Ethan Poskanzer, and Will Kimball. Michael was my cheering section and was willing to take the time to listen to all my latest half-baked ideas. Erik always found a way to make me laugh with his endlessly entertaining (and brilliant) ideas and stories. I will truly miss all of my brilliant and inspiring colleagues and friends.

No acknowledgement section would be complete without thanking all the wonderful people went out of their way to help make life at Sloan run smoothly including Natalia Kalas, Tetyana Pecherska, Bella DiMambro, Judy Graham-Robey, Helen Yap, Lisa Barone, and Stephanie Taverna. Natalia deserves my undying thanks for doing the hero's work of keeping me sane during the job market. I'm not sure that I fully comprehend how much she does, but I'm so grateful for her help, understanding, and friendship. Hillary Ross, Davin Schnappauf, and Ollie in the Ph.D. Office were always there to lend a supportive ear (or paw) and to deal with me at my most stressed with kindness and patience.

Beyond the people I've met during my Ph.D., there is a whole other cheering squad that never wavered in their support. The Dinner Gang, my chosen family, Sarah Balcom and Dave Benson, Jenn Dial and Mac Musselwhite, Red and Tris Walker-Buckton, Jess Bonzo and Tim Plymette (who we miss dearly), Shaq and Ivy Dastur, Anne Seville and Rob Flax, and Scott Crabbs and Emily Gerhold have been with me since our undergraduate days. They have watched me become the person I am today, cheering the good times, and being supportive in the bad. Many have their Ph.Ds., so they know. I am so lucky to have them and couldn't imagine life without them. My long-time friends Preeti Advani and Jen Moore, Elisabeth Deaton, Christina Waugh Gadrinab, Rachel and Stephen Spence, and T Gregory and Adam Stewart have also been there whenever I needed to get away. Some of my deepest gratitude belongs to Lynne Kiesling and Matt Coffey. Lynne was my first economics professor at William and Mary and her infectious enthusiasm for the field and for life in general forever changed my path. We are still friends today and I like to say this is all her fault. She should be proud.

Finally, I never would have made it to this point in my journey without the love and encouragement of my family. My mom, Marlene Zyontz; my dad, Larry Zyontz; my stepmother, Christine Zyontz; and my aunt and uncle Wendy Rothman and Jeff Zyontz have seen all of my successes and failures, but have always supported my decisions and let me talk through my concerns and fears. At this point, they probably know my dissertation better than I do. I'm grateful they let me make my own way since I left home for college but always let me know that they would be there for me when I needed them. I hope I make them proud as the first "Dr. Zyontz." My younger brothers Nathan, and his wife Jen Schmied-Zyontz, Kevin, and Steven are constant sources of pride as they find their own ways. I know they will be successful in whatever paths they choose and I'll be there to cheer them on too. And thanks, Kev, for doing that CRISPR experiment with me in the living room! To my amazing little sister Liann, I hope my success shows you that as a woman it is possible to reach for and achieve whatever goal you set for yourself. Nevertheless, you can persist. Thank you all for everything. I truly could not have done this without you. I love YOU more!

Samantha Zyontz
May 3, 2019
Cambridge, Massachusetts

Table of Contents

Exploring Institutions of Emerging Technologies

For technological progress, innovators not only need knowledge in a field of study, but also require tools (e.g., Rosenberg 1982, 1994, 2009; Nelson 1981, 2003; David 1990; Bresnahan and Trajtenberg 1995; Rosenberg and Trajtenberg 2004). The interplay between field knowledge and new tools leads to an ever-expanding base from which innovation emerges (Cohen and Levinthal 1989; Mokyr 2002; Wuchty et al. 2007). Part of the reason tools are important for innovation is that tools often embed much of the know-how necessary to use them. Thus, access to a tool can lower research costs as well as entry barriers to innovators (e.g., Furman and Stern 2011; Williams 2013; Murray et al. 2016; Teodoridis 2018). However, to use research tools, adopters must have both access and the ability to apply the tool in a field (Teece 1986; Scotchmer 1991; Weitzman 1996; Fleming 2001). The ability to apply a research tool is more difficult to develop when the tool is first introduced, however. Little is known about how an emerging technology impacts the rate and direction of innovation in its earliest days. This gap in understanding may reflect the fact that the data necessary for robust causal analysis, a hallmark of the research tools and innovation literature, is difficult to obtain.

One alternative method that allows the researcher to gain important insights in areas that have data limitations is institutional analysis. As a method of inquiry, institutional analysis appears in a number of disciplines but often takes on different meanings. For example, institutional analysis can refer to the study of social norms, rules, and constraints that shape interactions and decision making (e.g., North 1990, von Schmoller 1904, Ostrom 1990) or it can be broadened to include the study of formal institutions like agencies, boards, or other rule-making bodies (e.g., Komesar 2001, Coase 1984, Hughes 1939, Scott 2008). As noted by Cole (2013), the word "'institution' seems to have nearly as many definitions as definers" in the various literatures. Although definitions vary, institutional analysis commonly provides a way to begin answering the question "How do fallible humans come together, create communities and organizations, and make decisions and rules in order to sustain a resource or achieve a desired outcome?" (Ostrom and Hess 2011). Such institutional analyses can be applied to many different situations at various levels of decision-making, making the method adaptable to not only questions of policies, but to managerial strategies as well.

In order to derive effective implications for changes in rules and strategies, studies of institutions delve deeply into the details of the actors, institutions, technologies, and rules of a particular setting. This allows policymakers and knowledge managers to make systematic comparisons among strategies to encourage certain outcomes, including increased innovation, rather than relying on naïve ideas about appropriate rules to apply. Gaining a nuanced picture of underlying institutions also provides a way to develop meaningful questions and answers that might otherwise be hidden from the researcher's view.

This dissertation contributes to the research tools and innovation literature mentioned above, but is also influenced heavily by Elinor Ostrom's institutional analysis frameworks(Ostrom 1990, Ostrom and Hess 2011, Ostrom 2011). Ostrom highlights the importance of understanding the details behind an ongoing industry or an emerging one before recommending new institutions or strategies. As such, the essays presented here combine rigorous quantitative empirical methods common to the research tools and innovation literature with a deep understanding of the institutions of a new setting as suggested by Ostrom. They represent the beginning of a larger study, in real time, of key underlying factors contributing to the diffusion and impact of a breakthrough technology from its very first days. Not only will the information derived from these studies provide a solid base from which to make recommendations on policies and strategies for the diffusion of innovations, but it will also provide an in-depth understanding of variations in a new setting that can be applied to a number of important questions in innovation management and economics.

CRISPR is one of the most important breakthroughs in modern genetics and its speed of diffusion makes it an ideal setting to study the role of emerging tools on the rate and direction of innovation in affected fields. Starting in June 2012, separate research teams at UC-Berkeley (Jinek, et al. 2012), MIT (Cong et al. 2013), and Harvard (Mali et al. 2013) introduced CRISPR, a system that could edit the DNA of almost any organism more easily, accurately, quickly, and cheaply than previously available gene editing methods. Its discovery was unexpected, making CRISPR a shock to scientists engaged in gene editing. The tool quickly reduced the cost of research and introduced new gene editing applications not previously possible. For example, CRISPR has been used to modify crops for blight resistance (Wang et al. 2014), to create "malaria-proof" mosquitoes that are genetically unable to transmit malaria (Gantz et al. 2015), cure HIV in a mouse, introduce promising treatments for genetic diseases like muscular dystrophy, and to edit genes in human embryos (Ma et al. 2017) among other breakthroughs (Stockton 2017). CRISPR has already proven to be one of the most important medical discoveries this century, winning the prestigious Kavli Prize in Neuroscience and often speculated as a future Nobel Prize winning discovery.

Interest in the tool exploded because of its accuracy, flexibility, and relative ease of use (Pennisi, 2013; Regalado, 2014). From June 2012 through March 2019, almost 13,000 CRISPR-related articles were published worldwide. Further, due in part to the successful CRISPR experiments described, there have also been many commercial activities surrounding CRISPR. Over 5,000 United States Patent and Trademark Office (USPTO) patent applications and grants have been published by March 2019. Further, funding for venture backed firms licensed to use CRISPR technology has soared (Ledford 2015). The top biotech firms founded on CRISPR technology, Caribou Biosciences (Berkeley, CA), Editas Medicine (Cambridge, MA), CRISPR Therapeutics (Basel, Switzerland), and Intellia Therapeutics (Cambridge, MA) collectively raised initial funding of more than $150 million. The last three all had IPOs in 2016, each currently with market capitalizations of over $1 billion.

The three chapters in this dissertation are the first to systematically explore the adoption and diffusion of this breakthrough technology. They provide broader insights on how innovative teams adopt important research tools in the very beginning by focusing on the communities and individuals experimenting with the CRISPR tool as well as those able to overcome adoption barriers to innovate with CRISPR.

The first chapter, "Technological Breakthroughs, Entry, and the Direction of Scientific Progress: Evidence from CRISPR/Cas9," examines how the introduction of a breakthrough technology like CRISPR affects the trajectory of a scientific field through project selection and new entry. The main hypothesis is that the new technology's impact on the direction of gene editing research depends on its relative value across different organisms. Using proprietary data from the primary distributor of the CRISPR tool to academic scientists worldwide, Addgene, the paper shows that the relative proportion of scientists focusing on editing mammalian cells after the introduction of CRISPR increased over their counterparts working in bacteria and other eukaryotes in the first three years. Interviews with academic scientists who use gene editing techniques support this finding, noting that CRISPR represented a greater improvement over alternative tools for mammalian gene editing. In contrast, bacterial cells have natural editing functions scientists have effectively learned to manipulate, making CRISPR more of a curiosity than a necessity for bacterial labs. The data further suggest that the shift towards mammalian gene editing research may result mostly from new entry, but not from incumbent productivity.

The second chapter, co-authored with Neil Thompson, "Who Tries (and Who Succeeds) in Staying at the Forefront of Science: Evidence from the DNA-editing technology, CRISPR," delves into a question raised in the first chapter: why doesn't CRISPR seem to increase research productivity for incumbent mammalian researchers? To do that, the article explores the characteristics of scientists who experiment with CRISPR (order CRISPR from Addgene) and who successfully incorporate that experimentation with CRISPR into a project (publish a paper on CRISPR conditional on ordering CRISPR). Generally, it is difficult to empirically separate individuals who experiment with and those who successfully adopt a tool because failures are difficult to observe. Using Addgene data it is possible to separately observe both groups by matching CRISPR orders to publication histories for individual scientists in Web of Science. We developed a new algorithm to match 57,000 CRISPR orders to 1.3 million papers in Web of Science from 2012 – 2015 by scientist. The study suggests that there are important differences between those who experiment and those who are able to produce new research with the tool. For example, some characteristics traditionally associated with adoption (e.g. proximity to the discoverers) have negligible impacts on experimentation but do influence an author's ability to turn experimentation into a paper after controlling for other factors, empirically showing that access to a new complex tool does not automatically translate into the ability to successfully use the tool.

The third chapter, "Running with (CRISPR) Scissors: Tool Adoption and Team Assembly," builds on the first two chapters by recognizing that access to a new tool does not automatically translate into the ability to use it in an innovation in all research domains. Research tools are essential inputs to technological progress. Yet many new tools require specialized complementary know-how to be applied effectively. Teams in any research domain face the tradeoff of either acquiring this know-how themselves or working with external tool specialists, individuals with tool know-how independent of a domain. These specialists are scarce early on and can choose domain teams to create many applications for the tool or to focus on complicated problems. Ex ante it is unclear where the match between domain teams and external tool specialists dominates. The introduction of the DNA-editing tool CRISPR enables identification of external tool specialists on research teams by exploiting natural difficulties of applying CRISPR across disease domains. Teams have a higher share of external tool specialists in difficult diseases, especially for subsequent innovations. This suggests that external tool specialists and domain teams match more often to solve complex but influential problems. As more tools like Artificial Intelligence emerge, research teams will have to also weigh the importance of their possible solutions when considering how best to attract and collaborate with external tool specialists.

Looking forward, the analysis of CRISPR will be expanded to include topics in intellectual property (IP) and innovation, academic entrepreneurship, and regulation and product development, among others. For example, the next paper in the series will explore rules governing cumulative innovations in this setting by asking what is the effect of fragmented versus consolidated IP rights on innovation under uncertainty? Innovations cumulatively build on older ideas (Scotchmer 1991) but IP on early stage inventions has mixed innovation outcomes. Patents on inventions controlled by a dominant owner can have a chilling effect on the amount and breadth of follow-on innovation (e.g., Murray and Stern 2007; Williams 2013; Murray et al. 2016). However, fragmented IP ownership has also been shown to encourage more patenting so as to mitigate possible hold-up concerns (Ziedonis 2004). Therefore, IP strategies and the amount of follow-on innovation can depend on how fragmented the core IP ownership rights are. Cumulative innovation is also affected by disputes over IP. When there are early disputes over core IP, it is unclear whether the associated uncertainty has a chilling effect on follow-on innovation and firm formation (e.g., Bessen 2014; Tucker 2014; Cohen et al. 2018) or whether innovation moves forward regardless (e.g., Sampat and Williams 2019). As such, follow-on innovation and IP strategies may change entirely under uncertain but fragmented IP ownership.

References

Bessen, J. 2014. "The Evidence Is In: Patent Trolls Do Hurt Innovation." *Harvard Business Review.*

Bresnahan, T. F., and M. Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65(1): 83-108.

Coase, R. H. 1984. "The New Institutional Economics." *Journal of Institutional and Theoretical Economics* 140: 229-231.

Cohen, L., U.G. Gurun, and S.D. Kominers. 2018. "Patent Trolls: Evidence from Targeted Firms." Harvard Business School Finance Working Paper No. 15-002.

Cohen, W. M. and D.A. Levinthal. 1989. "Innovation and learning: the two faces of R & D." *The Economic Journal* 99(397): 569-596.

Cole, D. H. 2013. "The Varieties of Comparative Institutional Analysis." *Wisconsin Law Review* 2: 383-409.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. 2013. "Multiplex genome engineering using CRISPR/Cas systems." *Science* 339(6121): 819-823.

David, P. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80(2): 355-36.

Fleming, L. 2001. "Recombinant Uncertainty in Technological Search." *Management Science* 47: 117-132.

Furman, J., and S. Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production." *American Economic Review* 101(5): 1933-1963.

Gantz, V., N. Jasinskiene, O. Tatarenkova, A. Fazekas, V. Macias, E. Bier, and A. James. 2015. "Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi." *PNAS* 112(49): E6736–E6743.

Hughes, E.C. 1939. "Institutions" in *An Outline of the Principles of Sociology*, R.E. Park ed. Barnes & Noble, Inc.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." *Science* 337(6096): 816-21.

Komesar, N. K. 2001. *Law's Limits: The Rule of Law and the Supply and Demand of Rights.* Cambridge University Press.

Ledford, H. 2015. "CRISPR: The disruptor." *Nature* 522(7554): 20-24.

Ma, H., N. Marti-Gutierrez, S. W. Park, J. Wu, Y. Lee, K. Suzuki, A. Koski, D. Ji, T. Hayama, R. Ahmed, H. Darby, C. Dyken, Y. Li, E. Kang, A. R. Park, D. Kim, S. T. Kim, J. Gong, Y. Gu, X. Xu, D. Battaglia, S. Krieg, D. Lee, D. Wu, D. Wolf, S. Heitner, J. C. Belmonte, P. Amato, J. S. Kim, S. Kaul, and S. Mitalipov. 2017. "Correction of a pathogenic gene mutation in human embryos." *Nature* 548(7668): 413-419.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. "RNA-guided human genome engineering via Cas9." *Science* 339(6121): 823-6.

Mokyr, J. 2002. *Gifts of Athena: Historical Origins of the Knowledge Economy.* Princeton University Press.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy* 8(1): 212-252.

Murray, F. and S. Stern 2007. "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior and Organization* 63: 648-687.

Nelson, R.R. 1981. "Research on productivity growth and productivity differences: Dead ends and new departures." *Journal of Economic Literature* 19(3): 1029–1064.

Nelson, R.R. 2003. "On the uneven evolution of human know-how." *Research Policy* 32(6): 909–922.

North, D.C. 1990. *Institutions, Institutional Change, and Economic Performance.* Cambridge University Press.

Ostrom, E. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.

Ostrom, E. 2011. "Background on the Institutional Analysis and Development Framework." *The Policy Studies Journal* 39(1): 7-27.

Ostrom, E. and C. Hess. 2011. "A Framework for Analyzing the Knowledge Commons" in *Understanding Knowledge as a Commons*, C. Hess and E. Ostrom eds. MIT Press.

Pennisi, E. 2013. "The CRISPR craze." *Science* 341(6148): 833-836.

Regalado, A. 2014. "Who owns the biggest biotech discovery of the century?" *MIT Technology Review* (December 4).

Rosenberg, N. 1982. *Inside the Black Box: Technology and Economics.* Cambridge University Press.

Rosenberg, N. 1994. *Exploring the Black Box: Technology, Economics, and History.* Cambridge University Press.

Rosenberg, N. 2009. "Some critical episodes in the progress of medical innovation: An Anglo-American perspective." *Research Policy* 3(2): 234–242.

Rosenberg, N., and M. Trajtenberg. 2004. "A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States." *Journal of Economic History* 64(1): 61-99.

Sampat, B., and H. L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109 (1): 203-36.

Scotchmer, S. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5(1): 29-41.

Scott, W. R. 2008. *Institutions and Organizations: Ideas and Interests* (3rd ed.). SAGE Publications, Inc.

Stockton, N. 2017. "CRISPR kills HIV and eats Zia 'like Pac-man'. It's Next Target? Cancer." *WIRED.* (May 26).

Teece, D. J. 1986. "Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy." *Research Policy* 15(6): 285-305.

Teodoridis, F. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64(8): 3469-3970.

Thompson, N. and S. Zyontz 2017. "Who tries (and who succeeds) in staying at the forefront of science: Evidence from the DNA-editing technology, CRISPR." Working paper.

Tucker, C. 2014. "Patent Trolls and Technology Diffusion: The Case of Medical Imaging." Working paper.

Von Schmoller, G. 1904. *Grundriss Der Allgemeinen Volkswirtschaftslehre* as translated in Furubotn, E. and R. Richter 1998. *Institutions and Economic Theory: The Contribution of the New Institutional Economics.* University of Michigan Press.

Wang, Y., X. Cheng, Q. Shan, Y. Zhang, J. Liu, C. Gao, and J. L. Qiu. 2014. "Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew." *Nature Biotechnology* 32(9): 947–951.

Weitzman, M. 1996. "Hybridizing Growth Theory." *American Economic Review* 86: 207-212.

Williams, H. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy* 121(1): 1-27.

Wuchty, S., B. Jones, and B. Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316(5827): 1036-9.

Ziedonis, R. H. 2004. "Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms." *Management Science* 50(6): 804-820.

Zyontz, S. 2016. "Technological breakthroughs, entry, and the direct of scientific progress: Evidence from CRISPR/Cas9." Working paper.

Zyontz, S. 2019. "Running with (CRISPR) Scissors: Tool Adoption and Team Assembly." Working paper.

Chapter 1

Technological Breakthroughs, Entry, and the Direction of Scientific Progress: Evidence from CRISPR/Cas9

1 Introduction

In June 2012, Professors Jennifer Doudna and Emmamuelle Charpentier at the University of California, Berkeley published a paper demonstrating a technique for efficiently, accurately, and simply editing DNA in bacteria (Jinek, et al. 2012). At the time the paper received only a little notice despite the fact that the technique, CRISPR/Cas9, would become one of the most important discoveries in genetic engineering to date. In a recent Nature article, a geneticist from Cornell with 30 years' experience in the field, John Schimenti noted, "I've seen two huge developments since I've been in science: CRISPR and PCR... CRISPR is impacting the life sciences in so many ways" (Ledford 2015). However, it was not until Professor Feng Zhang at MIT proved six months later (Cong, et al. 2013) that CRISPR/Cas9 could be used to edit mammalian DNA that interest in the technique exploded. CRISPR/Cas9 has already been used to create blight resistant crops (Wang et al. 2014) and "malaria proof" mosquitoes (Gantz et al. 2015).

Although CRISPR/Cas9 has been shown to work in almost any organism, anecdotally much of the excitement surrounding the technique seems to be in mammalian organisms. For example, members of Feng Zhang's lab have created a "Cas9 mouse" (Platt et al. 2014) that can be modified to model lung cancer, something that was incredibly difficult to do prior to CRISPR/Cas9. Also, less than three years after CRISPR/Cas9's introduction, the tool was used to edit a human embryo for the first time (Liang et al. 2015). As such experiments continue, genetic engineers believe CRISPR/Cas9 has the potential to advance gene therapies to someday eliminate certain genetic diseases.

Based on the qualitative evidence, it is reasonable to hypothesize that CRISPR/Cas9 caused a disproportionate increase in mammalian genetic engineering research. Further support for this hypothesis is that, unlike in bacteria, the previous genome editing techniques for mammalian cells were inadequate, so CRISPR/Cas9 represents a substantial advance for mammalian researchers. Although, this was not obvious ex ante as CRISPR/Cas9 is part of the bacterial immune system and originally was discovered as a system to edit bacterial DNA. Therefore, it is possible that the breakthrough innovation, CRISPR/Cas9, caused a shift the direction of genetic engineering research towards the sub-field with the highest relative value for the tool due to researchers' responses to the new innovation.

Historical research focusing on how science progresses and changes paths generally finds that the process is not straightforward, requiring the interplay between scientific discoveries and implementable tools (Rosenberg 1990; Nelson and Rosenberg 1993; Brooks 1994; Mokyr 2002). More recent empirical studies have shown that access to new innovations, including scientific research tools, likely has a large positive impact on the rate of scientific progress (e.g. Griliches 1957, 1958; David 1990; Bresnahan and Trajtenberg 1995; Rosenberg and Trajtenberg 2004; Azoulay, et al. 2009; Murray et al 2016; Furman and Stern 2011; Williams 2013; Teodoridis 2018). Likewise, empirical work has highlighted varying effects of access to innovations on the direction of scientific progress (e.g., Moser 2005; Azoulay, et al. 2009; Budish, Roin, and Williams 2015).

This paper contributes to the current literature on influences to the direction of scientific progress by studying how researchers respond to an innovative technological breakthrough and by uncovering mechanisms that lead to rapid changes due to the discovery in an important new setting. Using genetic engineering as a novel setting, and by exploiting an exogenous technological shock that may have lowered the barriers to entry disproportionately in mammalian research, it may be possible to better understand whether scientists are more influenced by technology shocks that lower barriers to entry or by past behaviors. The paper also introduces two novel datasets: one that provides measures of academic researcher experimentation from the biological resource center, Addgene, and another that assesses researcher productivity through author publication histories.

Using the data from Addgene, first I find that the technology shock of the CRISPR/Cas9 genome editing system to the field of genetic engineering caused a shift in the direction of experimentation towards mammalian organisms. As the introduction of CRISPR/Cas9 lowered the costs of working with mammalian cells, it is reasonable to expect an increase in experimentation in that sub-field over bacterial or other eukaryotic (e.g. plants and fruit flies) experimentation.

Next, to delve more deeply into underlying mechanisms that may have driven the shift towards mammalian research in genetic engineering due to CRISPR/Cas9, I use publication histories from authors that eventually adopt CRISPR/Cas9 to test for a few mechanisms. First, CRISPR/Cas9 could be making incumbent mammalian researchers more productive and cause an increase in the rate of paper production. Second, it could be the case that researchers previously in the bacterial sub-field shift their focus to the mammalian sub-field as it becomes easier to work with mammalian organisms. Finally, as new researchers enter the field of genetic engineering, they could be entering the mammalian sub-field in greater proportions after the introduction of CRISPR/Cas9.

After testing these three mechanisms, it appears that the increase in mammalian research may be due to the type of researchers entering the field after the introduction of CRISPR/Cas9. Given CRISPR/Cas9's rapid rise in popularity, and the particular emphasis on mammalian research, new researchers may be entering genetic engineering with a focus on mammalian research in higher proportion after the introduction of CRISPR/Cas9.

There is not support for increased productivity or switching between sub-field as possible mechanisms, however. Given that there is strong qualitative evidence that CRISPR/Cas9 increased mammalian researcher productivity, the lack of empirical support for is surprising. However, it could be possible that learning to use a new tool could cause a temporary delay in productivity as incumbent researchers move up the learning curve. Some labs abandoned work in old editing techniques due to the power of this new method (Pennisi 2013; Ledford 2015) and might have to take a temporary step backwards before reaping productivity benefits. At least initially, there may also be an opportunity cost to adopting CRISPR/Cas9 if authors are substituting CRISPR/Cas9 projects for others.

The lack of evidence for immediate author sub-field switching is less surprising. Researchers may face additional switching costs that outweigh the cost savings from using the new technology or they may have a strong preference for their initial organism. Barriers that could discourage scientists from shifting sub-fields include high costs to adoption (e.g., Mokyr, 2002; Murray and Stern, 2007; Furman and Stern, 2011; Williams, 2013) and entrenchment in previous work behaviors (e.g., Siggelkow, 2001; Henderson and Clark, 1990). Therefore, a technological change can lower costs to working with mammalian organisms, but whether scientists shift direction depends on whether they can see a direct value in the new idea and how set they are in past working practices. Because CRISPR/Cas9 is only three years old, more data is necessary before it is possible to see whether the effects are due to a delay or to entrenched working practices.

Although there are important limitations to the current analyses, the paper provides a foundation for an exciting set of new research on how innovations are used in scientific communities and their influence on future directions of science. For example, Addgene has recently provided the data to match researcher experimentation with publication outputs, which will allow for studies on the translation between scientific experimentation and academic publication.

The remainder of this paper is organized as follows: Section 2 discusses the historical and empirical literature on how access to innovations influences the rate and direction of scientific progress as well as proposes empirical questions that add to the literature. Section 3 provides background on the field of genetic engineering, explains the science of CRISPR/Cas9, and discusses the popular reaction to the discovery. Section 4 describes the data sources used in this paper. Section 5 discusses the analysis and findings. Section 6 concludes and provides a path for future research.

2 Influences on Rate and Direction of Technological Progress

Historically, science (and human knowledge in general) has been thought to progress through a combination of knowledge that answers the question "what" and knowledge that answers the question "how." Mokyr (2002) calls this "propositional" and "prescriptive" knowledge respectively. An early view proposed progress as a linear process moving from propositional to prescriptive knowledge (Bush 1945). However, this "linear model" has been challenged by the idea that science, technology, and utilization develop along different but parallel paths (Allen 1984) and that the interconnections are complex and vary by industry (Rosenberg 1990; Nelson and Rosenberg 1993; Brooks 1994).

More recent research recognizes that it is possible for propositional knowledge to affect prescriptive knowledge but that the reverse is possible as well (Allen 1984; Mokyr 1990 and 2002; Brooks 1994; Stokes 1997). As an example of propositional knowledge contributing to prescriptive knowledge, Brooks (1994) lays out a number of contributions of science to technology including: new technological ideas, tools, techniques, instrumentation, human skills, technology assessment, and development strategies. But technological advances can influence the direction of science as well. Mokyr (2002) explained that before 1800 most technological progress was in inventions where people knew that something worked, but not why. Mokyr uses chemistry as an example of an area where practitioners knew that compounds had certain properties but had little idea of the underlying chemical structures. However, because of this prescriptive knowledge it was possible to use the working compounds as a basis for asking "why?" The resulting propositional knowledge then lead to a fundamental understanding of chemistry and contributed to the development of previously undiscovered compounds. Allen (1984) provides another example from electronics where electron tube technology faced an upper limit of useable frequencies, which forced a return to the underlying physics to be able to expand the number of channels for the increasing demand for radio communications. Therefore, propositional and prescriptive knowledge are mutually reinforcing and together shape the rate and direction of innovation.

Stokes (1997) takes this a step further and explained that ideas come in several different forms depending on the inspiration for the research. He calls ideas that come from a quest for fundamental understanding with no considerations of use "pure basic research" and ideas that come from a desire to use them, but no fundamental quest for understanding "pure applied research." His primary contribution is to then identify ideas that represent both a fundamental quest for understanding and a desire for use, calling them "use-inspired basic research" or "Pasteur's Quadrant" named after the research of Louis Pasteur that both expanded scientific understanding and provided useful applications. The ideas in Pasteur's Quadrant could be scientific research tools that embody new propositional knowledge and provide prescriptive knowledge to generate future scientific advances.

2.1 Empirical Studies on Innovations and the Rate and Direction of Technological Progress

The existing empirical literature builds on this historical base to show how access to innovations causes changes to the rate and direction of technological and scientific progress. In general these studies find that providing access to previously unavailable innovations encourages an increase in the rate of follow-on innovation. Although there are fewer studies that focus on the role of innovations on the direction of future ideas, a number of factors have been shown to influence the direction of technological progress.

2.1.1 Access to Innovations and the Rate of Follow-on Research

Some of the earliest empirical literature on how new inventions affect the rate of technological progress focuses on General Purpose Technologies (GPTs). GPTs are considered "enabling technologies," that encourage economic growth in a large range of downstream sectors (e.g., semiconductors or hybrid corn) (Bresnahan and Trajtenberg 1995; Griliches 1957, 1958; Rosenberg and Trajtenberg 2004; David 1990). For example, Rosenberg and Trajtenberg (2004) looked at the dynamics of GPTs through the history of the Corliss steam engine and found that it had a positive long-term impact on growth through productivity gains.

A second set of empirical research shows how providing access to research tools increases the rate of scientific progress. In Furman and Stern (2011), the authors use difference-in-differences models that look at treatment and control groups and timing differences to understand the effect of a deposit of a biological material into a biological resource center (BRC) on future citations to the related paper. They find a post-deposit citation increase of about 57-135 percent depending on the specification due to the increased access to the material through the BRC. Murray et al. (2016) also shows positive effects to the follow-on use of research mice (the Oncomouse, a key research tool) after restrictions to access were lifted. Finally, Teodoridis (2018) focused on a tool that reduced the cost of conducting motion-sensing research (Microsoft Kinect) and provided greater access to knowledge in that field. In general her findings suggest that the tool increased research in motion-sensing, which is in line with much of the other literature.

Another set of papers focuses on the reverse – what happens to the rate of future innovation when access to scientific tools is hindered by a form of intellectual property? Much of this literature finds that early stage intellectual property rights (IPRs) on cumulative basic research does seem to hinder the rate of follow-on innovation. Murray and Stern (2007) and Williams (2013) are some of the key papers that causally prove this point. Both papers show that property rights on early stage innovations reduce follow-on research by 20-40 percent. Murray and Stern (2007) use "patent-paper pairs" to show through a differences-in-differences identification strategy how new forward citations to scientific papers decrease after the invention described received a patent. Williams (2013) used

the differential timing on the sequencing of the human genome by private firm Celera and the public Human Genome project to disentangle the effects of intellectual property on future research on the human genome and subsequent commercial product development

The empirical results on the effects of restrictions have been mixed, however. For example, Azoulay, Ding, and Stuart (2009) find that individual inventors are more likely to produce more research of the same quality after getting a patent. This may seem like a difference between the aggregate result of patenting and the individual result of patenting, but updated studies by Williams and Murray and Stern have now found either no effects or even positive effects on follow-on research due to intellectual property (Sampat and Williams 2019; Fehder, Murray, and Stern 2014).

Finally, the market for ideas literature also suggests a positive role for IPRs in the rate of scientific progress. By design, patents create "fences" around the innovative area claimed in the grant. As with all fences, the delineation of property lines can help to exclude competitors from the space, but they also serve to define the space. By having a clearer idea of the property rights, owners can not only choose who to include, but those interested in being included also know where the boundaries of the rights are. This makes transacting over the property right much easier since both parties know approximately what rights are covered. This factor makes the "market for ideas" possible (Arora, Fosfuri, and Gambardella 2001; Gans and Stern 2003; Gans, Hsu, and Stern 2002). As has been shown by Gans and Stern with Hsu (2002, 2003, and 2008), the market for ideas can actually promote cooperation between innovating and established firms. The cooperation through patent licensing can expand the amount of innovation in the market rather than old ideas being subject to the gale of creative destruction described by Schumpeter (1942). Part of the reason such a market for ideas works is because incumbents have complementary assets such as sales forces that smaller, more innovative firms do not. These complementary assets could cause the innovative firm to lose its first mover advantage (Teece 1986) and so they may be better off licensing or forming an alliance with the firm with the complementary assets (Aghion and Tirole 1994; Lerner and Merges 1998; Lerner and Malmendier 2010). The market for ideas could help to diffuse the innovations much faster rather than just create competition between ideas.

2.1.2 Influences on the Direction of Technological Progress

A number of papers in Section 2.1.1 also discuss possible influences on the direction of technological progress. For example, Murray et al. (2016) found that the new access to the Oncomouse in comparison to non-restricted mice lead to a very diverse set of research paths that seem to encourage more horizontal exploration. So that access to tools lead to both more and diverse research streams.

Teodoridis (2018) focuses on the organization of knowledge that springs from Microsoft Kinect's ability to lower the cost of doing research in the area. She finds that such tools bring into

the field an array of different researchers she calls generalists and specialists. She concludes that it is the generalists that have the largest influence on follow-on knowledge creation since they are able to connect specialists to new research opportunities opened by the new technology.

Finally, some papers have explored the different directions of innovation that IPRs incentivize. For example Azoulay et al. (2009) find that individuals that get patents are more likely to shift their research towards commercialization and Budish, Roin, and Williams (2015) find that functional differences in patent lengths shift research in diagnostics for cancer towards solutions that take less time to develop. However, the absence of a patent system may also affect the direction of scientific research. The best evidence comes from Moser (2005) who studied innovations from the World Fair. She finds that there is not less innovation for countries without the patent system but those innovations tended to be best suited to trade secrecy. This suggests that without a patent, system innovators publicly disclose inventions that cannot easily be reverse engineered or for which they have some other comparative advantage. This could indeed limit the types of innovations that would otherwise become available under a patent system such as pharmaceuticals that can be easily reversed engineered.

2.1.3 Empirical Questions

In many of the empirical settings mentioned above, the innovations are incremental with gradual adoption. Although as Mokyr (1990) points out, technological change is a punctuated process so that changes can happen in short abrupt leaps rather than gradual process that consistently builds on itself like evolutionary biology. Considering this, the current paper adds to the empirical literature on the rate and direction of scientific research by studying how science progresses when confronted with an abrupt leap embodied in a breakthrough innovation that is rapidly adopted throughout a scientific field. Specifically, it seeks to answer two related questions:

1. Can an innovative technological breakthrough influence the direction of research in an emerging field?
2. If so, how do researchers initially respond to the innovation in order to generate rapid change?

In order to address these questions, novel datasets that provide measures of researcher experimentation and productivity are used. Further, genetic engineering is introduced as an important new setting to study the impact of a breakthrough innovation, the genome editing system CRISPR/Cas9. The innovation is firmly in Pasteur's Quadrant (Stokes 1997) and has been called the "biggest game changer to hit biology since PCR" (Ledford 2015). It has the potential to make possible everything from blight resistant crops to targeted genetic drug therapies.

3 Empirical Context

This section describes the innovative environment in genetic engineering and provides a primer on the CRISPR/Cas9 genome editing system. The second part of the section discusses the scientific and popular reaction to the CRISPR/Cas9 discovery.

3.1 Scientific Primer in Genetic Engineering Innovations

Genetic engineering lends itself to answering the empirical questions posed in Section 2. Through genetic engineering, researchers can modify an organism's genome to produce or remove traits that otherwise occur naturally. Two characteristics of genetic engineering research make it a relevant object of analysis: the field naturally divides into research sub-fields and the field faces rapid innovative changes.

Genetic engineering sub-fields emerge as the result of the type of organism scientists use in their research. Labs often specialize in organisms such as bacteria (e.g., E. coli), mammals (e.g., mice or humans), or other eukaryotes (e.g., yeast or plants). This specialization occurs for a number of reasons since organisms have different genetic properties. First, the costs of bringing together and maintaining a lab of research scientists with appropriate skill sets can be high, so focusing on organisms that require similar knowhow across the members of a research team makes the research process more cost effective. Second, different organisms require different biosafety levels as regulated by the Centers for Disease Control (CDC). Each biosafety level requires the lab to have progressively more safety equipment for the researchers and more safety features built into the lab space (CDC 2015). Thus it will cost labs more to switch from working with low biosafety level organisms like non-infectious E. coli to higher biosafety level ones like humans and other mammals, so each has an incentive to specialize in organisms at its biosafety level. Third, research scientists tend to continue research they did as graduate and post-doctoral students and tend to work with the same organisms due to preference, comfort level, or fit with the type of research conducted. Finally, not all tools and biological materials in genetic engineering can be used across different organisms. Thus E. coli labs might be expected to focus mainly on innovations specific to E. coli.

Although there are many organisms that fit within each broad category, the biological properties of organisms within each group are more similar than between groups. The remainder of the paper divides genetic engineering into the sub-fields of bacteria, mammals, and other eukaryotes.

Genetic engineering is also a useful setting for analyzing the empirical questions of this paper because it has seen a number of innovations that rapidly changed the field. The rate and nature of these recent changes makes it relatively easier to observe researchers responses to new innovations than in slower moving fields. As an example, one set of innovations increased the productivity and

decreased the price of DNA sequencing and synthesis (referred to as "Carlson Curves" by The Economist), see Figures 1a and 1b.

[Figures 1a and 1b here]

Over time, the drop in sequencing (reading DNA) and synthesis (writing DNA) costs made it generally possible for labs to access many different genetic parts that they could not previously afford. The decrease in the cost of sequencing DNA base pairs (bp) has been faster than Moore's Law, going from \$1.11/bp in 2004 to \$0.00005/bp in 2014, although there has been evidence of a slow-down in cost decreases over the last few years.[1] In synthesis, GenScript, one of the largest and earliest entrants to the gene synthesis market, dropped the cost of synthesizing a DNA sequence 1000 base pairs (bp) in length from \$2.35/bp in 2004 to \$0.23/bp in 2014.[2] Generally these cost shocks applied to all sub-fields.

A second set of innovations made it increasingly easier to cut DNA fragments in very specific locations, allowing for precise editing of genomes in organisms that were historically difficult to manipulate. Generally, these DNA editing technologies all have a mechanism for targeting a specific DNA sequence and making cuts. However, earlier technologies (Zinc Finger Nucleases and TALENs) were rather limited in their applications, whereas the most recent technology, CRISPR/Cas9, is a universal cutting technology. For example, in the early 2000s, most genome editing was accomplished through a technique known as Zinc Finger Nucleases (ZFNs). The ZFNs technique required the researcher to create a unique enzyme that could recognize and cut each individual DNA sequence of interest. However, the original ZFNs could only recognize naturally occurring DNA sequences, which was a limitation for genetic engineers wishing to study modified DNA sequences that became possible with ever cheaper synthesis. The specificity of ZFNs may have encouraged labs to focus on specific kinds of cells in their experiments rather than amassing the materials and knowhow necessary to use the broad array of individual ZFNs available. Starting in late 2009, many labs began editing genomes using Transcription Activator-Like Effector Nucleases (TALENs) (Moscou and Bogdanove, 2009; Boch, et al., 2009). TALENs are similar to ZFNs in that they can be used to recognize and cut out specific DNA sequences and require a new enzyme to be created to recognize and cut each unique DNA sequence. However, desired new TALENs are much easier to engineer than ZFNs because researchers have solved the code for how TALENs recognize different DNA sequences. An individual TALEN can be used in multiple cell types if the desired DNA sequence exists in each, but the specificity of the method likely limited most labs to specific types of organisms. Although the limiting set is larger than the limiting set for the original ZFNs.

---

[1] Source: NHI National Human Genome Research Institute, available at: http://www.genome.gov/sequencingcosts/.

[2] Source: GenScript, available at: http://www.genscript.com/gene_synthesis.html?src=pullmenu. Historical prices were obtained using the April snapshots captured by the WayBack Machine Internet Archive (web.archive.org).

TALENs were chosen as "Method of the Year" in 2011 by Nature Methods (Method, 2012) but in a true shock to the field, a new method appeared only months later.

In June 2012, CRISPR/Cas9 was discovered to be a highly efficient cutting tool for genome editing. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are distinctive sequences of DNA that were first noticed in bacteria as early as 1987. The CRISPRs, which consist of a repeated DNA sequence appearing between unique DNA sequences, were largely a curiosity until 2007. That year, CRISPRs' role in the bacterial immune system became well understood (Barrangou et al, 2007). Researchers discovered that bacteria had a system that could recognize and cut out viruses it had previously fended off. When the virus attacks again, the CRISPR recognizes it and brings along an enzyme (usually called Cas9) to cut the viral DNA, rendering the virus harmless. Breakthroughs in 2012 and 2013 when researchers discovered the CRISPR/Cas9 system could be used to find and cut more than viral DNA sequences through the use of guide RNA (gRNA) (Jinek et al. 2012, Cong et al. 2013, Mali et al. 2013).

Part of CRISPR/Cas9's attractiveness is that unlike ZFNs and TALENs, the CRISPR/Cas9 system does not require the researcher to find individual CRISPR enzymes to cut each DNA sequence. Instead the gRNA acts as a genetic GPS device that can find the programmed DNA sequence in an organism and the Cas9 enzyme can cut the targeted DNA in almost any organism. Because the gRNA can recognize longer sequences of DNA than ZFNs or TALENs, the CRISPR/Cas9 system is also more accurate in finding the correct target DNA. For example, it is the difference between telling a word processor to find "absolutely" rather than "abs" which will find the word "absolutely," but also "absolute," and "abstract." The system can also be programmed to carry with it a set of replacement DNA, so that genetic engineers can use CRISPR/Cas9 as a genetic version of a word processor's find and replace function to cut out a certain segment of DNA and replace it with the DNA sequence they want to study. In addition, a researcher can use CRISPR/Cas9 to make multiple cuts at the same time in the same cell, something that was not possible with the previous editing techniques.

The shock of CRISPR/Cas9 drastically reduced the costs of using organisms that had been historically difficult to work with in experiments since gRNAs are very simple and cheap to order and easily enter a variety of organisms. The response to this new technology was almost immediate given its accuracy, flexibility, and relative ease of use (Pennisi, 2013; Regalado, 2014).

3.2 Tracking the Adoption of CRISPR/Cas9

CRISPR/Cas9 was first introduced as a proof of concept for easy genome editing in bacteria by Professors Jennifer Doudna and Emmanuelle Charpentier at the University of California, Berkeley in June 2012 (Jinek, et al., 2012). Less than six months later, in January 2013, MIT Professor Feng Zhang and his collaborators published a breakthrough paper describing the use of

CRISPR/Cas9 as a viable easy-to-use general-purpose genome editing system for mammalian cells, including human cell lines (Cong et al., 2013).[3] The work of Zhang and his co-authors at MIT and the Broad Institute (and the related work of George Church and his colleagues at Harvard Medical School) lead to significant interest in the role that CRISPR/Cas9 can play in life sciences innovation. Between June 2012 and November 2015, there have been almost 2,000 individual publications mentioning CRISPR based on a search of the Web of Science, which is an average rate of approximately 48 papers a month since Doudna's first article.

As emphasized by Jennifer Doudna in a February 2015 JAMA editorial, "This discovery has triggered a veritable revolution as laboratories worldwide have begun to introduce or correct mutations in cells and organisms with a level of ease and efficiency not previously possible." (Doudna, 2015). For example, CRISPR/Cas9 has already been used to modify crops to be blight resistant (Wang et al. 2014) and to create "malaria-proof" mosquitoes that are genetically unable to be affected by or transmit malaria (Gantz et al. 2015). The introduction of CRISPR/Cas9 will be especially useful in medical applications since it will more easily allow researchers to build new mouse and human cell disease models with very specific sets of mutations on which to test new drugs. As a recent example, members of Feng Zhang's lab have created a "Cas9 mouse" (Platt et al. 2014) that can be modified to model lung cancer. Before CRISPR/Cas9 creating such a mouse model for lung cancer would have taken many people and a decade of time. Creating this model took one person four months (Specter 2015). As further evidence of the potential and speed of CRISPR/Cas9, start-ups based on CRISPR/Cas9 technology have raised more than $80 million in financing (Regalado 2014; Ledford 2015) and scientists attempted to edit the genome of human embryos using CRISPR/Cas9 for the first time in April 2015 (Liang et al. 2015).

Interviews with several experts in genetic engineering indicate that the impact of CRISPR/Cas9 did make working with mammals and other eukaryotes much easier than in the past. Although the original Doudna lab paper showing CRISPR/Cas9 as a proof of concept for bacteria was published in June 2012, it was not until the Zhang lab paper on the usefulness of CRISPR/Cas9 for mammalian cells was published that interest seemed to increase. A Google Trends analysis of the term "CRISPR" provides some evidence that searches for the term increased after the publication of the Zhang article (Figure 2).

[Figure 2 here]

Part of the reason for this interest from mammalian researchers is the biological differences in bacteria versus eukaryotes that make editing bacteria genes far simpler. Unlike mammals and

[3] An earlier paper (Jinek et al., 2012) describes the use of CRISPR/Cas9 for non-mammalian cells and a simultaneous paper (Mali et al, 2013) independently confirms results found by Zhang and his collaborators and tests CRISPR/Cas9 efficiency against previous genome editing techniques.

other eukaryotes, bacteria cells have no nucleus that contains the DNA. Therefore, the process of editing bacterial DNA can be as simple as inserting modified bacterial DNA through the cell membrane. This can be done by simply applying an electric shock to the bacteria cells, which opens holes in the cell membrane through which new DNA can be introduced in a process called electrophoresis. Because of the relatively simple process that exists for engineering bacterial cells, techniques like CRISPR/Cas9 are not usually necessary to perform experiments. Eukaryotic cells are far more complicated and cannot be edited using the same electrophoresis technique. Instead, more complicated techniques such as ZFNs, TALENs, or CRISPR/Cas9 are required to edit eukaryotic cells. Therefore, researchers in mammalian organisms had a greater incentive to experiment with CRISPR/Cas9.

This qualitative evidence suggests that the technology shock of CRISPR/Cas9 should have had a differential effect on genetic engineering sub-fields. Specifically, it should provide a natural experiment that enables a test of the hypothesis that CRISPR/Cas9 caused an increase in mammalian experimentation as compared to the other eukaryote and bacteria sub-fields. Multiple underlying mechanisms could lead to such an increase. CRISPR/Cas9 could affect the both the rate and direction of research in genetic engineering by making current researchers more productive, inducing new researchers to enter the field, or by encouraging researchers in bacteria to shift their research focus. Alternatively, the technology shock could have little to no effect on the rate and direction of new research due to entrenchment of specific standards within labs or strong barriers to entry in certain sub-fields. Either outcome is interesting and would help enrich our understanding of the role of scientific tools in encouraging research in a specific sub-field.

4 Data Construction

To explore the reaction of researchers to the introduction of CRISPR/Cas9, this paper relies on two different data sources. The first is a novel dataset provided by Addgene, one of the most popular biological materials repositories in genetic engineering (Kahl and Endy 2013). Addgene facilitates the donation and transfer of biological materials between academic researchers for new experiments. The second is a dataset of the publication histories of every last author that published a paper referencing one of the original three papers on CRISPR/Cas9 as a technique for genome editing from Doudna, Zhang, or Church. The first dataset is a proxy for the level of experimentation in genetic engineering. The second dataset represents the actual productivity of researchers using CRISPR/Cas9.

4.1 Addgene Plasmid Database

Addgene, a non-profit company, was founded in 2004, by Melina Fan, Kenneth Fan, and Benjie Chen for scientists to easily store, search for, and share plasmids for use in biological research (Fan et al., 2005). Plasmids are small molecules of DNA that are used in genetic cloning to introduce new genes into a host cell. Each plasmid has a variety of defining characteristics, including the type of host cell it is best suited for in experiments. The organization is a repository that helps facilitate the exchange of genetic material between laboratories by storing plasmids and their associated cloning data donated from academic labs all over the world. Addgene performs quality control on all donated plasmids and can ship validated materials to academic laboratories all over the world for use in scientific research. Although their plasmid donation repository is publicly available on their website (http://www.addgene.org), Addgene's plasmid order history has never before been released outside the company.

For this project, Addgene provided the full, daily order history from September 2004 through October 2014. Included in the history are the plasmids in each order; an ID for each requesting lab; and categorizations for whether the plasmid is for use in bacterial, mammalian, or other eukaryotic organisms.[4] Since CRISPR/Cas9 can be physically manifested in a set of plasmids, each designed for use within a certain type of cell, the Addgene data also identifies whether the plasmid is for use in CRISPR/Cas9 experiments.[5] Unfortunately, due to confidentiality concerns, Addgene was not able to divulge the actual identities of the ordering labs. This poses some limitations on the ability to control for lab characteristics until the plasmid data can be matched to actual papers. Addgene has recently granted access to the identification data, so future articles will be able to overcome this limitation.

Plasmids are relatively inexpensive to purchase from Addgene at $65 per plasmid, so Addgene provides a low barrier to experimentation in genetic engineering. Although labs can make their own plasmids or attempt to get them directly from the originating lab, it is often easier to order directly from Addgene. Therefore, labs order from Addgene when they want to experiment with existing plasmids. The Addgene order data thus provides a proxy for the amount of experimental behavior from genetic engineering labs and a leading indicator for shifting interests in the field.

---

[4] Plasmids ordered from Addgene can be used to indicate the type of research ordering labs focus on by looking at whether the plasmid is categorized for bacteria, mammal, or other eukaryotic expression. It is important to note that the ordered plasmid expression is a proxy for the type of cell it will be used in. It is not necessarily the case that the lab eventually used the ordered plasmid. However, plasmid cell expression is a reasonable proxy for the type of research done in the ordering lab since it is generally not useful for a lab to order a plasmid it cannot work with.

[5] The original authors of the CRISPR/Cas9 papers donated all of their CRISPR plasmids to Addgene at the time of the papers' publications.

The initial database contains 435,926 individual plasmids ordered in 171,352 orders by 48,857 unique labs in 82 countries over the ten-year period. The orders represent 20,780 unique plasmids. The growth rate of Addgene orders has been increasing over time as shown in Figure 3. The drop in 2014 is solely from the fact that it is not a full year of data (the dashed bar shows the estimated total for 2014).

[Figure 3 here]

The increases in orders have ranged from 16 to almost 30 percent since 2008 except for the 37 percent increase from 2012 to 2013. At least half of that increase was due to new orders of plasmids embodying the CRISPR/Cas9 system (Figure 4). Therefore, CRISPR/Cas9 likely contributed to an increased experimentation rate in genetic engineering as measured by orders to Addgene.

[Figure 4 here]

Given that CRISPR/Cas9 has a low cost of adoption and in general was seen to lower the barriers to entry into genetic engineering, it is not surprising that there was an increase in experimental activity after CRISPR/Cas9 was introduced. Ex ante, though, the effect on different sub-fields of genetic engineering was not obvious. It is possible that CRISPR/Cas9 had an equal effect on all sub-fields, bacterial research could have had an increase in experimentation since this was the area where CRISPR/Cas9 first emerged as a proof of concept, or mammalian research could have seen the increase since this was the area that biologically benefited most from the CRISPR/Cas9 tool. Given that CRISPR/Cas9 made it easier to work in mammalian cells more than any other, it could be reasonably hypothesized that CRISPR/Cas9 caused a greater increase in mammalian experimentation than in any other sub-field. The summary statistics for the overall Addgene order database are presented in Table 1.

[Table 1 here]

4.2 Publication History Database for CRISPR/Cas9 Authors

To the extent that CRISPR/Cas9 has a differential effect on mammalian research in genetic engineering, this effect could be due to a number of underlying behaviors of the researchers. Understanding the mechanisms behind a change in the direction of research can help better explain the effect of a new technology shock on a rapidly changing area of research. Unfortunately, the

Addgene data does not currently provide identifying characteristics of the ordering labs,[6] but it is possible to understand which researchers contributed to the increase in mammalian research after CRISPR/Cas9 was introduced using the publication histories of the authors that eventually write an article that cites to one of the three original CRISPR/Cas9 papers. The implicit assumption here is that the labs that order CRISPR/Cas9 from Addgene are similar to those that eventually publish papers in CRISPR/Cas9. Although the groups that experiment and those that eventually publish are likely not identical, it is reasonable to assume that there is overlap. Data on the author publication histories is available from Scopus.

Scopus, owned by Elsevier, is an abstract and citation database for peer-reviewed articles that covers more than 21,800 journals in science, technology, medicine, economics, other social sciences, and the humanities (Scopus 2014). Data from Scopus was collected by first identifying every paper that cited to one of the three original CRISPR/Cas9 papers from June 2012 through June 2015. The last listed author on each paper[7] was then searched by Scopus' Author Identifier to obtain the author's entire publication history. The Author Identifier is a unique ID number Scopus assigns to authors to disambiguate common names and combine different formats of the same name. Author papers are grouped by an algorithm that matches on author affiliation, subject area, co-author names, and dates of publication citations (Scopus 2015). Scopus was chosen over other citation databases such as PubMed for this analysis because it was necessary to be able to search abstracts and keywords into 2015 to obtain as much data as possible after the introduction of CRISPR/Cas9 given its very recent discovery.

There has been an incredible amount of research done using CRISPR/Cas9 since its introduction in June 2012. There are 1,475 unique articles that cite to at least one of the original papers between June 2012 and June 2015. These papers were written by 1,085 unique last authors. The initial dataset of author-paper pairs contains almost 108,000 papers published from 1959 through June 2015. Each record contains the disambiguated author name, his or her co-authors, the paper title, the paper journal, the year published, author affiliations, the abstract, and Scopus-assigned keywords identifying the topics of the articles.

Unlike the Addgene plasmids, papers can have more than one research focus. Therefore, it was necessary to devise a set of measures to determine which organisms a particular lab may work on over time. First, lists of keywords indicative of research in bacterial, mammalian, or other eukaryotic organisms were developed in conjunction with an expert in genetic engineering (see Appendix A for the list of keywords). Then the proportion of sub-field specific keywords was used to determine the relative focus of the paper. For example, a paper with five keywords in either the Scopus keywords or paper abstract: "mouse" "mammal" "human" "coli" "eukaryot" would be

[6] This is true for the dataset used in this paper. Addgene has since granted access to the identification data that will allow labs' ordering histories to be linked to their publication histories. Future research will not have this limitation.

[7] In scientific papers, the last author is usually the person responsible for the article and the lab.

classified as 60 percent mammalian, 20 percent bacterial, and 20 percent other eukaryotic focused. It is important to note that the keywords are a proxy for the type of organisms a lab uses. Although, if such keywords are used in the abstract and Scopus keywords, then there is a high likelihood that the lab conducts experiments pertaining to these sub-fields.

Because some authors historically publish more than one paper a year, the data was aggregated to an author-year panel. For each sub-field, the proportions were averaged to obtain the percent mammalian, bacterial, and other eukaryotic focus respectively that a given author had in his or her papers that year. In addition to the yearly measures of focus, it was possible to classify each author as initially a mammalian, bacterial, or other eukaryotic researcher. Due to abstract limitations in the database prior to 2000, the initial author focus was determined by taking the average of the sub-field percent focus from 2000-2006 and then assigning the author to the sub-field where the average percent was greater than 60 percent. For example, if an author had an average percent focus of 60 percent mammalian, 20 percent bacteria, and 20 percent other eukaryote from 2000-2006, then the author was classified as initially mammalian focused. The year 2006 was chosen as the end point since 2007 was the first year where the biology of CRISPRs garnered attention from genetic engineers. Since 2000-2006 is used to determine the initial focus of the lab, all regressions in the paper use the remaining panel from 2007 through June 2015 unless otherwise noted. The summary statistics for the entire dataset from 2007-June 2015 are displayed in Table 2.

[Table 2 here]

5 Empirical Results

In general, when a new tool is introduced that makes working in a field easier, previous research has shown that the overall rate of productivity increases. If this is true for productivity in terms of forward citations, it should also be true in experimentation. As shown in Section 4.1, this appears to be the case when CRISPR/Cas9 was introduced as a tool for genetic engineering. Perhaps less obvious is whether CRISPR/Cas9 had an influence in shifting the direction of experimentation in the field. If there is a larger increase in experimentation within a specific sub-field then it will be useful to the understanding of such a rapidly changing industry to explore the underlying mechanisms that lead to this increase. Section 5.1 explores the hypothesis that mammalian experimentation benefits more from the introduction of a new tool than other genetic engineering sub-fields. Section 5.2 delves deeper into the underlying mechanisms that may lead to such an increase.

5.1 CRISPR/Cas9's Effect on the Direction of Genetic Engineering Research

Since CRISPR/Cas9 could be used to make experiments on all types of organisms more efficient, it may be the case that all sub-fields benefited similarly from the new tool. Alternatively, some sub-fields could have benefited more from the technological shock. For example, since mammalian research biologically did not have easy tools to work with but bacterial research had a set of standard effective techniques, it is reasonable to hypothesize that experimentation in mammalian research should increase more after the introduction of CRISPR/Cas9 than bacterial or other eukaryotic experimentation.

To determine whether CRISPR/Cas9 differentially affected experimentation in genetic engineering sub-fields, all plasmids ordered were categorized by expression type (i.e., mammalian, bacterial, and other eukaryotic) and aggregated by quarter. If CRISPR/Cas9 had a differential effect on the bacterial, mammalian, or other eukaryotic sub-fields, a change in the popularity of plasmids in that sub-field should be observable. Given that CRISPR/Cas9 may be relatively more useful in working with more complex mammalian cells, it would be reasonable to expect to see an increase in the use of mammalian plasmids over bacterial or other eukaryotic plasmids after the introduction of CRISPR/Cas9. A graph of these trends shows that although orders of mammalian plasmids consistently grew faster than bacterial or other eukaryotic plasmids over time, the growth of mammalian plasmid orders was even faster after the introduction of CRISPR/Cas9 (Figure 5). This difference is statistically significant, as can be shown by a decomposition of trends analysis.

[Figure 5 here]

The basic specification to test whether CRISPR/Cas9 affected plasmid orders for mammalian experimentation more than in the other sub-fields is an aggregate decomposition of trends analysis where β_5 is the incremental effect of CRISPR/Cas9 on mammalian plasmid orders:

$$PlasmidOrders_t = \beta_0 + \beta_1 t + \beta_2 t_{post} * Post + \beta_3 Mammal + \beta_4 Mammal * t + \beta_5 Mammal * t_{post} * Post + \beta_6 Other + \beta_7 Other * t + \beta_8 Other * t_{post} * Post + \varepsilon_t$$

$PlasmidOrders_t$ = the number of plasmids ordered in quarter t
t = Count of quarters starting in 2004 Q3
t_{post} = Count of quarters after CRISPR/Cas9, starting in 2012 Q3
$Post$ = 1 if the quarter is after the CRISPR/Cas9 shock
$Mammal$ = 1 if the plasmid is for use in mammalian experiments
$Other$ = 1 if the plasmid is for use in other eukaryotic experiments

In this model, $\beta_0, \beta_3, \beta_6$ represent the regression constant for bacterial, mammalian, and other eukaryotic plasmid orders respectively. Then $\beta_1, \beta_4, \beta_7$ represent the slopes of the overall ordering trends for bacterial, mammalian, and other eukaryotic plasmids respectively. The remaining coefficients, $\beta_2, \beta_5, \beta_8$ are then the incremental increase per quarter in the ordering trends for bacterial, mammalian, and other eukaryotic plasmids respectively after the introduction of CRISPR/Cas9. If mammalian experimentation increased more due to CRISPR/Cas9, then the mammalian post-CRISPR/Cas9 incremental trend should be highly significant and larger than those post-CRISPR/Cas9 trends for the other sub-fields.

Table 3 shows that after controlling for the ordering trends in bacterial and other eukaryotic plasmids, mammalian plasmid orders increase by 338 orders per quarter after CRISPR/Cas9. The result is significant at the 0.001 level and has a larger magnitude than the other sub-fields. The 338 orders represents a 95 percent increase over the total mammalian trend. This result provides evidence that CRISPR/Cas9 had a larger effect on mammalian experimentation than on the other sub-fields.

[Table 3 here]

Another way to evaluate whether CRISPR/Cas9 had a differential effect on mammalian experimentation is to look at the incremental contribution plasmids that are both for mammalian and CRISPR/Cas9 experimentation make to total plasmid orders as compared to the other sub-fields. To do this, it is possible to exploit the panel nature of the data to look at the effect of a plasmid being mammalian and CRISPR/Cas9 on total plasmid orders while controlling for sub-field characteristics as well as year and plasmid cohort effects. The basic OLS panel model specification for this test can be written:

$$PlasmidOrders_{it} = \alpha_0 + \alpha_1 Mammal_i * CRISPR_i + \alpha_2 Mammal_i + \alpha_3 CRISPR_i + \alpha_4 Other_i * CRISPR_i + \alpha_5 Other_i + \alpha_6 X_{it} + \delta_t + \delta_{t-entry} + \varepsilon_{it}$$

$PlasmidOrders_{it}$	= the number of plasmid i ordered in quarter t
$Mammal_i$	= 1 if the plasmid i is for use in mammalian experiments
$Other_i$	= 1 if the plasmid i is for use in other eukaryotic experiments
$CRISPR_i$	= 1 if the plasmid i is for use in CRISPR/Cas9 experiments
$Mammal_i$*$CRISPR_i$	= 1 if the plasmid i is for use in mammalian and CRISPR/Cas9 experiments
$Other_i$*$CRISPR_i$	= 1 if the plasmid i is for use in other eukaryotic and CRISPR/Cas9 experiments
X_{it}	= 1 if the plasmid i was ordered by a lab new to Addgene in quarter t
$\delta_t, \delta_{t-entry}, \varepsilon_{it}$	= year effects, age effects, and error term respectively

In this specification, if CRISPR/Cas9 had a larger effect on mammalian experimentation, it would be reasonable to expect that plasmids characterized as for both mammalian and CRISPR/Cas9 experiments would have a positive and significant effect on total plasmid orders after controlling for the other sub-fields as well as year and plasmid age effects.[8] As shown in Table 4, being a mammalian CRISPR/Cas9 plasmid has a significant positive effect on total plasmid orders in the full panel dataset.

[Table 4 here]

In Table 4, Model 1, the OLS specification suggests that being a mammalian CRISPR/Cas9 plasmid contributes an additional 19 orders per quarter over bacterial plasmids controlling for the factors mentioned above. However, because the plasmid order data is skewed to the right and contains a large number of zeros, Table 4 Model 2 shows the results of a zero-inflated Poisson model using the same controls. Here, the result is less significant, but still positive at the 5 percent level. The incident rate ratio suggests that being a mammalian and CRISPR/Cas9 plasmid increases the rate of orders by 1.6 times over bacterial plasmids. Again, this result suggests that CRISPR/Cas9 encouraged more experimentation in mammalian organisms than in other sub-fields.

Finally, as a robustness check, the results found in Tables 3 and 4 could be due to an imbalance of the number of CRISPR/Cas9 plasmids available for each sub-field. Although it is true that more unique mammalian CRISPR/Cas9 plasmids are available than bacterial or other eukaryotic CRISPR/Cas9 plasmids, the number of orders per unique plasmid is higher for those that are for mammalian and CRISPR/Cas9 experimentation (Table 5). Since labs do not have to order a plasmid more than once, this helps support the evidence that there is more experimentation with mammalian CRISPR/Cas9 plasmids than with other sub-fields.

[Table 5 here]

The novel plasmid order data provided by Addgene does suggest that the introduction of CRISPR/Cas9 had a larger differential effect on the mammalian sub-field of genetic engineering. Given that the change occurred immediately, it will be informative to better understand how the sub-field reacted to such a major shock.

[8] It is not possible to include plasmid fixed effects in these models since the variables of interest do not vary over time and would drop out in such a specification.

5.2 Underlying Mechanisms for the Shift to Mammalian Research

CRISPR/Cas9 could be influencing mammalian genetic engineering research over bacterial genetic engineering research in a few ways. The first is that CRISPR/Cas9 could be making previously existing mammalian researchers more productive and cause an increase in the rate of paper production. Second, as new researchers enter the field of genetic engineering, they could be entering the mammalian sub-field in greater numbers after the introduction of CRISPR/Cas9. Finally, it could be the case that researchers previously in the bacterial sub-field could be shifting their focus to the mammalian sub-field as it becomes easier to work with mammalian organisms.

As mentioned in Section 4.2, it is not possible to assess the different mechanisms described above using the Addgene data, so the following analyses will rely on the author history dataset collected from Scopus. First, to shift datasets, it is important to show a similar increase in mammalian research among the researchers that adopt CRISPR/Cas9. As with the trends in the Addgene data, the increase in CRISPR/Cas9 research comes not from the sub-field that first developed the technique, but from the mammalian sub-field that benefited most from the more valuable tool.

Figure 6 shows the number of CRISPR/Cas9 papers that contain mammalian keywords over time as compared to the number of papers that contain bacterial or other eukaryotic keywords. In general, there is a swift increase in papers with mammalian keywords after the introduction of CRISPR/Cas9 for mammals that is not replicated for those with bacterial keywords.

[Figure 6 here]

Although it is not possible to prove here that the labs experimenting in CRISPR/Cas9 on Addgene are the same that eventually publish using the tool, the similarity of mammalian and bacterial research trends between the two provide some comfort in using publishing data to explore the underlying mechanisms behind the increase in mammalian research. The remainder of this section evaluates each possible mechanism in turn.

5.2.1 Incumbent Researcher Productivity

Some of the qualitative evidence suggests that CRISPR/Cas9 was thought to reduce the amount of time research takes, especially with mammalian organisms. For example, consider the creation of the "Cas9 mouse" which allows one person to create an animal model in a few months rather than a decade (Specter 2015). If the technology shock made the mammalian sub-field of genetic engineering more productive than the bacterial sub-field, one would expect to see a significant increase in the percent change of papers for mammalian researchers after CRISPR/Cas9

was introduced.[9] To test for this effect, it is possible to exploit the panel nature of the author history database to control for sub-field, year effects, author effects, and author experience effects. The basic OLS panel model specification for this test can be written:

$$PctChangePapers_{jt} = \gamma_0 + \gamma_1 Mammal_{jt} * PostCRISPR_t + \gamma_2 Mammal_{jt} + \gamma_3 Other_{jt} * PostCRISPR_t + \gamma_4 Other_{jt} + \delta_j + \delta_t + \delta_{t-entry} + \varepsilon_{jt}$$

$PctChangePapers_{jt}$	= the percent change in papers for author j in year t
$Mammal_{jt}$	= 1 if the author j is initially classified as having a mammalian focus OR the percent mammalian focus of author j in year t
$Other_{jt}$	= 1 if the author j is initially classified as having an other eukaryotic focus OR the percent other eukaryotic focus of author j in year t
$PostCRISPR_t$	= 1 if year t is after the introduction of CRISPR/Cas9 (here, after 2012)
$\delta_j, \delta_t, \delta_{t-entry}, \varepsilon_{jt}$	= author effects, year effects, author experience effects, and error term

As described in Section 4.2, the regression uses yearly data from 2007 – 2014 to calculate the percent change in papers by author. For regressions that use the author's initial focus, that measure was determined using data from 2000 – 2006. All included authors published before and after the introduction of CRISPR/Cas9. The coefficient of interest in this model is γ_1 which is the increase in the percent change of papers due to focusing on mammalian research after CRISPR/Cas9 as compared to bacteria holding all else constant.

Table 6 shows the results of regressing each author's percent change in papers by year on whether the papers were published after CRISPR/Cas9 and if they focused on mammalian research. In general across all models, the results are insignificant and small, which does not allow us to reject the null hypothesis of no increase in productivity for mammalian focused authors. Models 1 and 2 in Table 6 show OLS results without author fixed effects and using the author's initial focus. Adding author fixed effects (as in Model 3) brings the coefficient of interest practically to zero. If instead of strictly assigning an author an initial sub-field from a pre-period, the author's sub-field focus was allowed to vary over time, the result is still insignificant, but at least positive for authors with higher mammalian after CRISPR/Cas9 (Table 6, Model 4).

[Table 6 here]

Given the results of Table 6, there is not clear evidence that CRISPR/Cas9 increased the productivity of any author in the mammalian sub-field over those in the bacterial sub-field. The

[9] Percent change in year over year papers was used as the productivity measure to account for differences in the levels of papers from mammalian and bacterial authors.

lack of an increased productivity finding is somewhat surprising given the qualitative evidence. However, there could be a few explanations for mammalian authors who adopt CRISPR/Cas9 not becoming relatively more productive that are not accounted for in the above regression. First, incumbent authors often used older techniques like ZFNs and TALENs, so learning to use a new tool could cause a temporary delay in productivity as they move up the learning curve. This would be especially true if there is some tacit knowledge necessary for implementing CRISPR/Cas9 (von Hippel 1994, Ledford 2015). Second, authors in this space may be constrained by the number of papers it is possible to write in a given year, and so may be switching from non-CRISPR/Cas9 projects to CRISPR/Cas9 projects. This would keep the average growth in papers the same and might suggest that at least initially there is an opportunity cost to adopting CRISPR/Cas9.

5.2.2 Incumbent Researcher Focus Shifts

Second, the increase in mammalian research after CRISPR/Cas9 could be due to incumbent authors in the bacterial sub-field switching to the mammalian sub-field. As mentioned previously, improvements in genome editing methods like CRISPR/Cas9, have made it relatively easier to work with mammalian organisms. According to interviews with genetic engineers and experts at Addgene, authors most likely to be influenced to shift research focus by CRISPR/Cas9 are those that began by conducting research with organisms for which the technique was not biologically necessary (i.e., bacteria). One of the reasons a bacteria focused author may switch their focus to mammalian organisms after CRISPR/Cas9 becomes available is that the research they do is applicable to higher order organisms. However, they may have initially chosen to work in bacteria because it had readily available genome editing techniques. The discovery of CRISPR/Cas9 would make it easier for such authors to switch to a higher order organism that more closely matches their research interests. Mammalian researchers can (and do) switch focus to bacteria, but since there were already working editing techniques for bacteria the discovery of CRISPR/Cas9 gave them no additional incentive to switch.

Specifically, the effect of the exogenous CRISPR/Cas9 shock on focus switching should be greatest for authors that would have the most difficult time adopting a new type of organism absent the new tool. The treatment group for this is authors that started with research in bacteria. A concern in this setting is trying to construct an appropriate control group when the entire field is growing so quickly. The initial control group used in this study is authors that started with research in mammalian organisms prior to the discovery of CRISPR/Cas9. The outcome of interest is whether the author chose to focus on a different sub-field from his or her initial research focus.

From the Scopus publication data, it is possible to ask Does the adoption of an exogenous technology shock such as CRISPR/Cas9 induce authors that worked in easier to edit organisms to shift their research focus more than authors that worked in historically difficult to edit organisms?

The simplest difference-in-difference estimator, θ_1, for an increased change in focus by traditionally bacterial researchers due to the CRISPR/Cas9 shock would be:[10]

$$ShiftFocus_{jt} = \theta_0 + \theta_1 Bacteria_j * PostCRISPR_t + \theta_2 Bacteria_j + \theta_3 PostCRISPR_t + \delta_j + \delta_t + \varepsilon_{jt}$$

$ShiftFocus_{jt}$	= 1 if author j in year t (t = 2007 – 2015) focused mainly on a different topic than the author initially had. The initial focus of the author is determined based on the sub-field with 60 percent or more focus on average from 2000 - 2006.
$Bacteria_j$	= 1 if the author j was initially defined as bacteria focused. The Control group is all authors defined as initially mammalian focused. All authors published before and after the shock.
$PostCRISPR_t$	= 1 if the year t is 2013 or after (after the CRISPR/Cas9 shock).
$\delta_t; \delta_j; \varepsilon_{jt}$	= year fixed effects, author fixed effects, and an error term.

Therefore, this analysis will look at whether an author's adoption of CRISPR/Cas9 increased the likelihood of a bacteria lab shifting focus to mammalian research more so than a mammalian lab shifting focus to bacteria research. All else equal, CRISPR/Cas9 should increase the likelihood of bacteria labs shifting focus to mammalian research since it lowered the barriers to entry for mammalian research but less so for bacterial research. In other words, if this is true, the treatment effect, θ_1, should be positive and significant.

A more flexible form of the above model allows for an interaction effect for each year to better understand the year over year trend:

$$ShiftFocus_{jt} = \theta_0 + \sum_{t=2007}^{2015} \theta_t Bacteria_j * PostCRISPR_t + \theta_2 Bacteria_j + \theta_3 PostCRISPR_t + \delta_j + \delta_t + \varepsilon_{jt}$$

Figure 7 plots the year over year treatment effects in Table 7, Model 4. The graph shows that there is no real pre-trend, but that bacterial authors are no more likely to switch focus than mammalian authors after the introduction of CRISPR/Cas9 either. In general, this analysis suggests that the increase in mammalian research after CRISPR/Cas9 is likely not due to bacterial authors shifting their research focus to mammalian organisms. A possible reason for this result may be other costs not accounted for by the change in technology making the cost of adoption still too high (e.g., Mokyr 2002; Murray and Stern 2007; Furman and Stern 2011; Williams 2013). For example,

[10] The coefficients on Bacteria and PostCRISPR would be subsumed in models containing author and year fixed effects.

bacterial laboratories may need to be upgraded to comply with higher biosafety levels before working with mammalian organisms. Alternatively, authors may prefer the type of organism they initially chose and are entrenched in previous work behaviors (e.g., Siggelkow 2001; Henderson and Clark 1990). Unfortunately, distinguishing between these mechanisms is not yet possible given the short length of time since the discovery of CRISPR/Cas9. This analysis may have a different result when more years of data are available.

[Figure 7 and Table 7 here]

5.2.3 Research Focus of New Entrants

A final avenue for mammalian research to increase after CRISPR/Cas9 could be the introduction of brand-new researchers to genetic engineering that choose to enter the mammalian sub-field in greater proportion after the introduction of CRISPR/Cas9. Authors in the publication history dataset could make a first appearance in any year from 1959 to 2015. As the data for analyses is restricted to the years 2007 forward, Figure 8 charts the number of authors that first appeared in the dataset between 2007 and 2015 (estimated) by year. The figure shows that 52 percent of new authors in that period entered after CRISPR/Cas9 was introduced. This suggests that CRISPR/Cas9 may have influenced new entry into genetic engineering.

[Figure 8 here]

To test whether a higher proportion of new entrants came into the mammalian sub-field after CRISPR/Cas9, it is possible to measure the effect of entering after CRISPR/Cas9 on the types of papers entrants write in their first year. The basic OLS model specification for this test can be written:

$$Focus_{jt} = \gamma_0 + \gamma_1 PostCRISPR_t + \varepsilon_{jt}$$

$Focus_{jt}$ = 1 if the percent focus for entrant j in year t is 60% or more (for mammal, bacteria, or other eukaryote)
OR the percent focus for entrant j in year t (for mammal, bacteria, or other eukaryote)

$PostCRISPR_t$ = 1 if year t is after the introduction of CRISPR/Cas9 (here, after 2012)

This analysis only considers new entrants in the year that they enter and limits the years to 2011 – 2015. 2011 was chosen as the starting date for the model since that range likely does not contain new entrants due to ZFNs or TALENs. If new entrants were more likely to focus on mammalian

research after CRISPR/Cas9, then the coefficient on *PostCRISPR* should be significant, positive, and higher when the dependent variable is measuring mammalian focus.

Table 8 shows the results from regressing the type author focus on the post CRISPR/Cas9 introduction from 2011 through 2015. Models 1-3 use the more flexible percentage focus measure and Models 4-6 use the binary focus measure. As expected, regardless of the focus measure used, post-CRISPR/Cas9 entry has a larger positive effect on mammalian focus than on the other sub-fields. The results are significant, but only at the 5 percent level due to the small number of entrants in this conditional set. From Table 8, there is reasonable evidence to suggest that CRISPR/Cas9 encouraged new entrants to come into genetic engineering through the mammalian sub-field more so than before.

[Table 8 here]

6 Concluding Remarks and Future Research

This paper serves to provide new data on factors that affect the direction of research by basic scientists by observing researchers' responses to a breakthrough technology and rapid innovation in an emerging field. It contributes to the current literature by showing how a scientific breakthrough can shift research towards where the innovation's relative value at the margin is highest. It also introduces an important new setting and novel datasets that provide measures of researcher experimentation and productivity.

Based on the analyses presented, the exogenous shock of the CRISPR/Cas9 genome editing system to the field of genetic engineering caused a shift in the direction of experimentation towards mammalian organisms. This result was not obvious ex-ante as the system was originally discovered in bacteria and was only later proven to be applicable to experiments in mammalian organisms. The relative increase in mammalian experimentation over other sub-fields like bacteria and other eukaryotes can be explained by the fact that, biologically, CRISPR/Cas9 was far more beneficial to mammalian researchers than bacterial researchers. By their nature, it is relatively easy to insert new DNA into bacteria since they do not have a nucleus. Mammalian cells are far more complex and, previous to CRISPR/Cas9, the genome editing tools available all had limitations. The introduction of CRISPR/Cas9 lowered the costs of working with mammalian cells, and so it would be reasonable to expect an increase in experimentation in that sub-field over others.

To delve more deeply into underlying mechanisms that may drive the shift towards mammalian research in genetic engineering due to CRISPR/Cas9, the paper uses publication histories for authors that eventually publish a paper citing to one of the original three CRISPR/Cas9 articles by Doudna (Berkeley), Zhang (MIT), and Church (Harvard). A relative increase in mammalian research could occur through a few mechanisms. First, CRISPR/Cas9 could be making

incumbent mammalian researchers more productive and cause an increase in the rate of paper production. Second, as new researchers enter the field of genetic engineering, they could be entering the mammalian sub-field in greater proportions after the introduction of CRISPR/Cas9. Finally, it could be the case that researchers previously in the bacterial sub-field could be shifting their focus to the mammalian sub-field as it becomes easier to work with mammalian organisms.

After testing these three mechanisms, it appears that the increase may be due to the type of researchers entering the field after the introduction of CRISPR/Cas9. There is no evidence to support increased productivity or switching as possible mechanisms, however. Given that there is strong qualitative evidence that CRISPR/Cas9 increased mammalian researcher productivity, the lack of evidence for increased researcher productivity is surprising. However, it could be possible that learning to use a new tool could cause a temporary delay in productivity as incumbent researchers move up the learning curve. Alternatively, at least initially, there may be an opportunity cost to adopting CRISPR/Cas9 if authors are substituting CRISPR/Cas9 projects for others. The lack of evidence for immediate author sub-field switching is less surprising. Authors may face additional switching costs that outweigh the cost savings from the new technology or they may have a strong preference for the organism they began with. Since CRISPR/Cas9 is only three years old, more data may be necessary before it is possible to see delayed effects.

The largest limitation in this paper is the necessary assumption that researchers experimenting in CRISPR/Cas9 are also the ones publishing. Although it is a reasonable assumption that there is overlap between the two communities, there may be researchers that experiment with materials from different sub-fields but never publish there. It is also possible that a number of authors experiment in CRISPR/Cas9 but ultimately never publish. Given the data available at the time of this analysis, it is not possible to directly link authors who order from Addgene to their publication histories. However, Addgene has recently provided access to the identifying information for the ordering authors, which will make it possible to match experimental inputs to eventual research outputs and rectify the major limitation of this paper. As such, the analyses presented here lay the groundwork for subsequent studies on new innovations' effects on the rate and direction of scientific research in this rapidly changing setting.

As a short-term application, the new identification data provided by Addgene will allow for studies on the transition from scientific experimentation to academic publication. Since Addgene plasmid orders can be a proxy for inputs to experimentation in basic science, the order database can be matched to author publications and characteristics to determine how often and how quickly different experimental inputs translate to published outputs. These are historically difficult measures to obtain in basic research since the level of experimentation is often not observable. In general, the studies would attempt to predict the expected amount of research in genetic engineering given the amount of experimentation:

$$E[publications] = amtExperimentation * Prob(success|experimentation)$$

It would also be possible to expand the analysis to study the translation of experimentation to publications by different sub-groups in the researcher population:

$$E[publications]_i = \sum_i amtExperimentation_i * Prob(success|experimentation)_i$$

where i could be the type of researcher (e.g., initially mammalian or initially bacterial) or a network measure (e.g., number of degrees removed from an initial inventor).

In the longer term, the data will also support studies that determine how experimental tools spread since it will be possible to track when an author would have access to the tool. Such studies would use network analysis to determine the relative importance of geography and co-location, personal relationships as through co-author networks, and third-party institutions that act as distributors of the tools. Although there are limitations to the current paper, the foundation described here and the additional data from Addgene provides the potential for an exciting set of new research on how innovations are used in scientific communities and their influence on future directions of science.

7 References

Aghion, P. and J. Tirole. 1994. "On the Management of Innovation." *Quarterly Journal of Economics* 109: 1185-1207.

Allen, T. J. 1984. *Managing the Flow of Technology: Technology Transfer and the Dissemination of Technological Information within R&D Organization.* MIT Press.

Arora, A., A. Fosfuri, and A. Gambardella. 2001. "Markets for Technology and their Implications for Corporate Strategy." *Industrial and Corporate Change* 102(2): 419-451.

Azoulay, P., W. Ding, and T. Stuart. 2009. "The Effect of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output." *Journal of Industrial Economics* 57(4): 637-676.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, P. Horvath. 2007. "CRISPR provides acquired resistance against viruses in prokaryotes." *Science* 315(5819): 1709–1712.

Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, and U. Bonas. 2009. "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors." *Science* 326(5959): 1509-1512

Bresnahan, T. F., and M. Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65(1): 83-108.

Brooks, H. 1994. "The Relationship Between Science and Technology." *Research Policy* 23(5): 477-486.

Budish, E., B. N. Roin, and H. Williams. 2015. "Do Firms Underinvest in Long-Term Research? Evidence from Cancer Clinical Trials." *American Economic Review* 105 (7): 2044-85

Bush, V. 1945. *Science – The Endless Frontier: A Report to the President on a Program for Postwar Scientific Research.* U.S. Government Printing Office.

Centers for Disease Control and Prevention. 2015. "Recognizing the Biosafety Levels" available at: http://www.cdc.gov/training/QuickLearns/biosafety/

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. 2013. "Multiplex genome engineering using CRISPR/Cas systems." *Science* 339(6121): 819-823.

David, P. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80(2): 355-36.

Doudna, J. 2015. "Genomic Engineering and the Future of Medicine." *JAMA.* 313(8):791-792.

Fan, M., J. Tsai, B. Chen, K. Fan, and J. LaBaer. 2005. "A central repository for published plasmids." *Science* 307 (5717): 1877.

Fehder, D., F. Murray, and S. Stern. 2014. " Intellectual property rights and the evolution of scientific journals as knowledge platforms." *International Journal of Industrial Organization* 36: 83–94.

Furman, J., and S. Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production." *American Economic Review* 101(5): 1933-1963.

Gantz, V., N. Jasinskiene, O. Tatarenkova, A. Fazekas, V. Macias, E. Bier, and A. James. 2015. "Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi." *PNAS* 112(49): E6736–E6743.

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

Gans, J. S., D. H. Hsu, and S. Stern. 2008. "The Impact of Uncertain Intellectual Property Rights on the Market for Ideas." *Management Science* 54(5): 982-997.

Gans, J. S., and S. Stern. 2003. "The Product Market and the Market for Ideas: Commercialization Strategies for Technology Entrepreneurs." *Research Policy* 32: 333-350.

Griliches, Z. 1958. "Research Costs and Social Returns: Hybrid Corn and Related Innovations." *Journal of Political Economy* 66(5): 419-431.

Griliches, Z. 1957. "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica* 25(4):501-522.

Henderson, R. M., and K. B. Clark. 1990. "Architectural Innovation: The Reconfiguration of Existing Product Technologies and the Failure of Established Firms." *Administrative Science Quarterly* 35(1): 9-30.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." *Science* 337(6096): 816-21.

Kahl, L. and D. Endy. 2013. "A Survey of Enabling Technologies in Synthetic Biology." *Journal of Biological Engineering* 7:13.

Liang, P., Y. Xu, X. Zhang, C. Ding, R. Huang, Z. Zhang, ... and J. Huang. 2015. "CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes." *Protein and Cell* 6(5): 363–372.

Ledford, H. 2015. "CRISPR, the disruptor." *Nature* (522): 20–24.

Lerner, J., and U. Malmendier. 2010. "Contractibility and the Design of Research Agreements." *American Economic Review* 100(1): 214-246.

Lerner, J. and R. Merges. 1998. "The Control of Technology Alliances: An Empirical Analysis of the Biotechnology Industry." *Journal of Industrial Economics* 46: 125-56.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. "RNA-guided human genome engineering via Cas9." *Science* 339(6121): 823-6.

Method. 2012. "Method of the Year 2011." *Nature Methods* 9:1.

Mokyr, J. 1990. *The Lever of Riches: Technological Creativity and Economic Progress.* Oxford University Press.

Mokyr, J. 2002. *The Gifts of Athena.* Princeton University Press.

Moscou, M. and A. J. Bogdanove. 2009. "A Simple Cipher Governs DNA Recognition by TAL Effectors." *Science* 326(5959):1501.

Moser, P. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World Fairs." *American Economic Review* 95(4): 1214-1236.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy* 8(1): 212-252.

Murray, F. and S. Stern 2007. "Do formal intellectual property rights hinder the free flow of scientific knowledge? An empirical test of the anti-commons hypothesis." *Journal of Economic Behavior and Organization* 63: 648-687.

Nelson, R. and N. Rosenberg. 1993. *Technical Innovation and National Systems: A Comparative Analysis.* Oxford University Press.

Pennisi, E. 2013. "The CRISPR Craze." *Science* 341 (6148): 833-836.

Platt, R.J., S. Chen, Y. Zhou, M. J. Yim, L. Swiech, H. R. Kempton, J. E. Dahlman, O. Parnas, T. M. Eisenhaure, M. Jovanovic, D. B. Graham, S. Jhunjhunwala, M. Heidenreich, R. J. Xavier, R. Langer, D. G. Anderson, N. Hacohen, A. Regev, G. Feng, P. A. Sharp, and F. Zhang. 2014. "CRISPR-Cas9 knockin mice for genome editing and cancer modeling." *Cell.* 159(2):440-55.

Rosenberg, N. 1990. "Why Do Firms Do Basic Research (with their own money)?" *Research Policy* 19(2): 165-174.

Rosenberg, N., and M. Trajtenberg. 2004. "A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States." *Journal of Economic History* 64(1): 61-99.

Regalado, A. 2014. "Who Owns the Biggest Biotech Discovery of the Century?" *MIT Technology Review,* (December 4).

Sampat, B., and H. L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109(1): 203-36.

Schumpeter, J. 1942. "The Process of Creative Destruction" in *Capitalism, Socialism, and Democracy.* Harper & Row.

Scopus. 2014. "Scopus Content Coverage Guide" available at: https://www.elsevier.com/solutions/scopus/content

Scopus. 2015. "Scopus Author Identifier" available at: http://help.elsevier.com/app/answers/detail/a_id/2845/p/8150/incidents.c$portal_account_name/22659

Siggelkow, N. 2001. "Change in the presence of fit: The rise, the fall, and the renaissance of Liz Claiborne." *Academy of Management Journal* 44: 838-857.

Specter, M. 2015. "The Gene Hackers: A powerful new technology enables us to manipulate our DNA more easily than ever before." *The New Yorker* (16 November).

Stokes, D. 1997. *Pasteur's Quadrant: Basic Science and Technological Innovation.* Brookings Institution Press.

Teece, D. 1986. "Profiting from Technological Innovation: Implications for Integration, Collaboration, Licensing and Public Policy." *Research Policy* 15(6): 285-305.

Teodoridis, F. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64(8): 3469-3970.

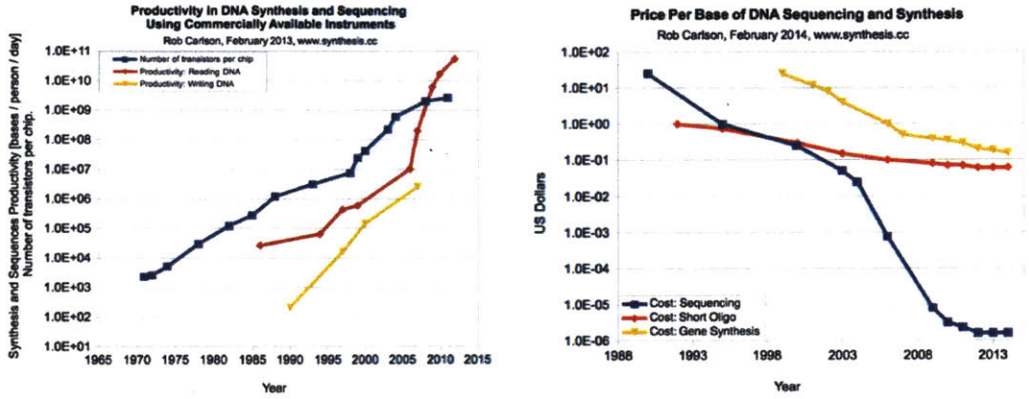
von Hippel, E. 1994. "Sticky Information and the Locus of Problem Solving: Implications for Innovation." *Management Science* 40(4): 429-439.

Wang, Y., X. Cheng, Q. Shan, Y. Zhang, J. Liu, C. Gao, and J. L. Qiu. 2014. "Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew." *Nature Biotechnology* 32(9): 947–951.

Williams, H. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy* 121(1): 1-27.

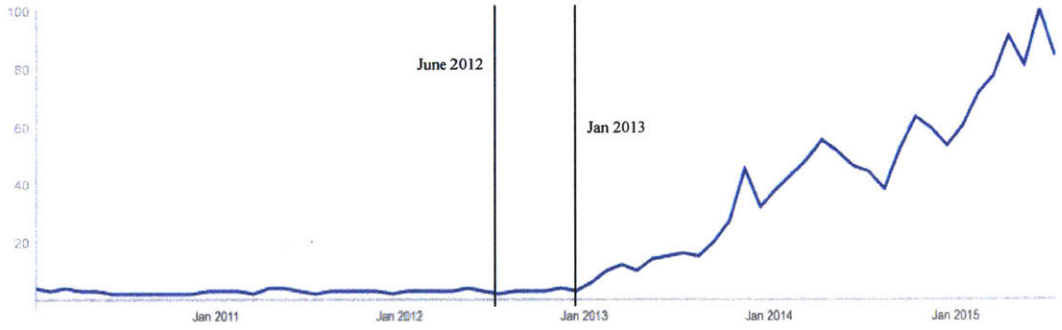
8 Figures and Tables

Figures 1a and 1b. Increasing Productivity and Decreasing Price of DNA Sequencing and Synthesis



Source: Rob Carlson "Time for New DNA Synthesis and Sequencing Cost Curves (corrected pub date)" February 12, 2014, available at: http://www.synthesis.cc/2014/02/time-for-new-cost-curves-2014.html.

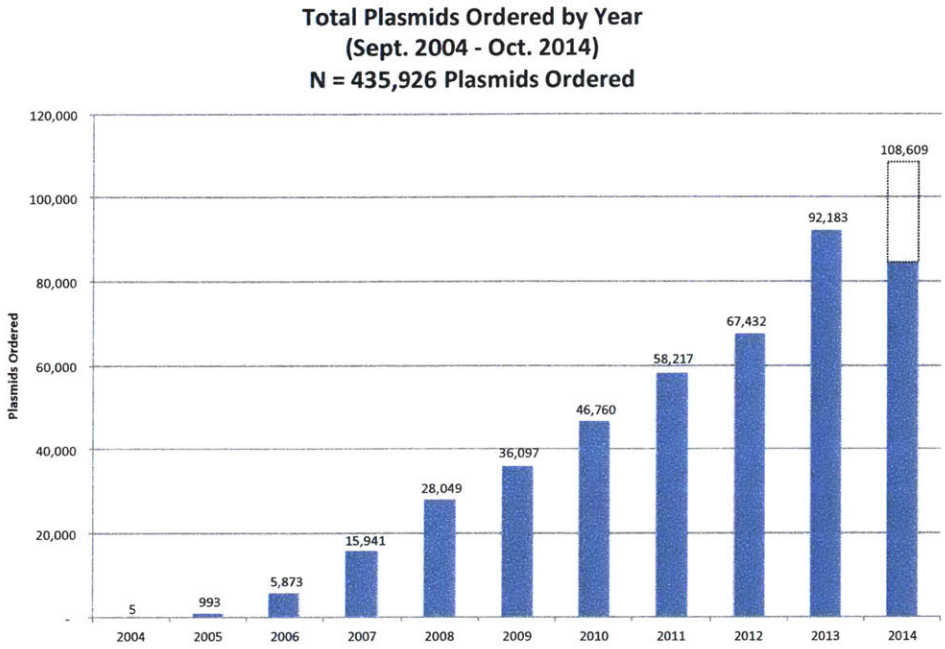
Figure 2. Relative Searches for the Term "CRISPR" Normalized from Most Popular Month (June 2010 – June 2015)



Source: Google Trends (www.google.com/trends).

Notes: Graph shows the relative number of searches on Google for the term "CRISPR" between June 2010 and June 2015. The number of searches per month is indexed against the month with the highest number of searches over the time period and does not represent the total number of searches (May 2015 = 100)

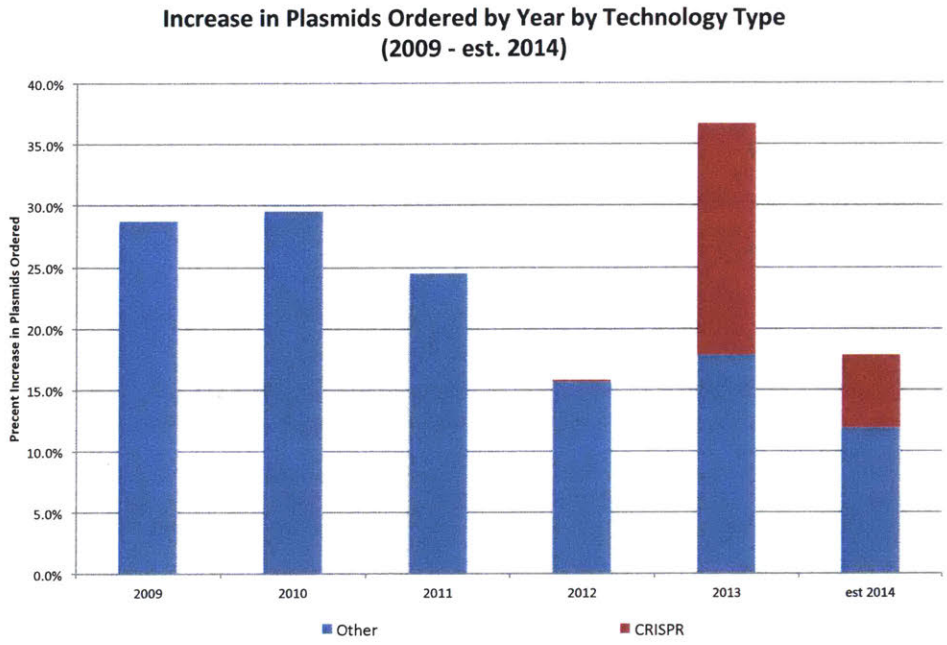
Figure 3.



Source: Addgene Plasmid Database

Notes: This figure shows the increase in the total number of plasmids ordered from Addgene between its inception in September 2004 and October 10, 2014. Each bar represents the number of plasmids ordered by all researchers by year so that a unique plasmid can be ordered many times by different researchers. The dashed bar in 2014 represents estimated orders for the remainder of October, November, and December based on the average monthly orders from January 2014 through October 10, 2014. See the text for more details and variable descriptions.

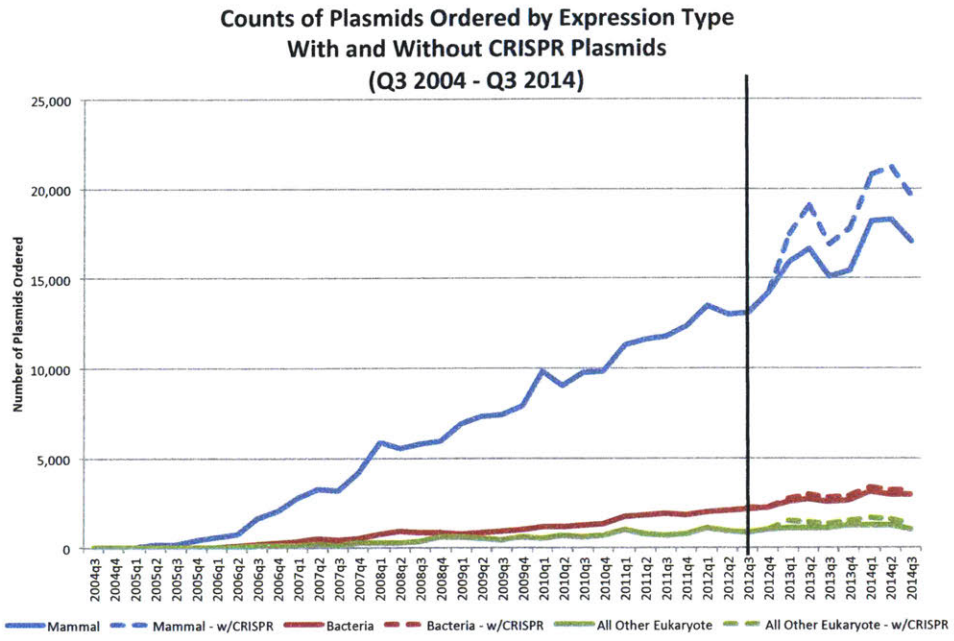
Figure 4.



Source: Addgene Plasmid Database

Notes: This figure shows several descriptive statistics regarding ordering patterns for CRISPR and non-CRISPR plasmids at Addgene. The height of each bar shows the percent increase in the total number of plasmids ordered from Addgene by year from 2009 through 2014 (estimated). The blue bars are the percent increase in orders for plasmids that do not embody the CRISPR/Cas9 system. The red bars are the percent increase in orders for plasmids that contain the CRISPR/Cas9 system. The proportion of CRISPR to non-CRISPR plasmid orders for all of 2014 was estimated to be the same as from January 2014 through October 10, 2014. See the text for more details and variable descriptions.

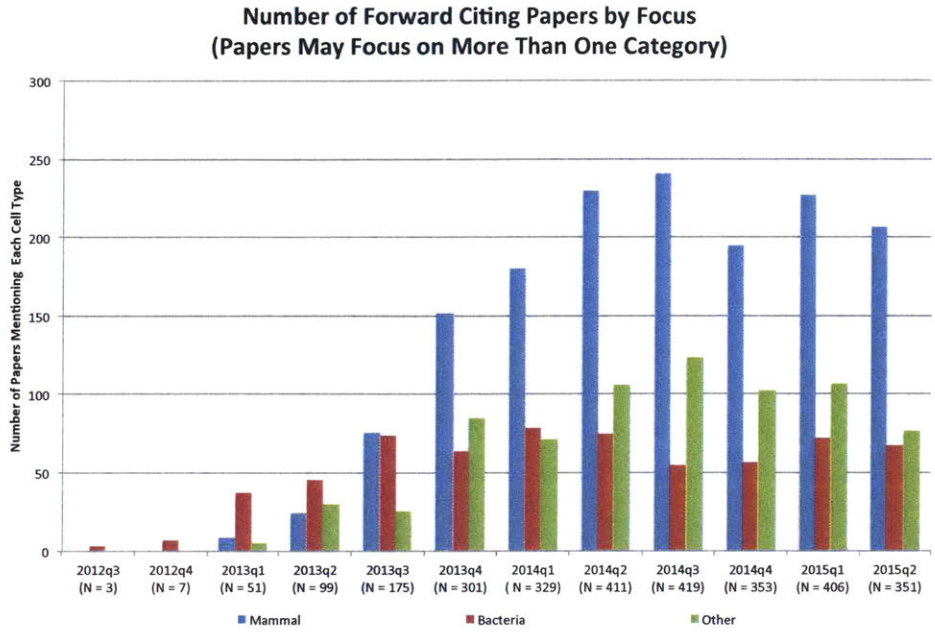
Figure 5.



Source: Addgene Plasmid Database

Notes: This figure plots the effect of CRISPR on the plasmid ordering trends at Addgene by the category of organism for which the plasmids were designed (called Expression Type). Unique plasmids may be ordered more than once by different researchers, categories are mutually exclusive, and plasmids with no category are removed. Aggregate orders by category are calculated by quarter from Q3 2004 through Q3 2014. The solid blue line is the total number of plasmids ordered in each quarter designed for use with mammalian organisms that do not contain the CRISPR/Cas9 system. The dashed blue line adds plasmids ordered that are both designed for use with mammalian organisms and contain the CRISPR/Cas9 system. The red solid and dashed lines contain the same information but for plasmids designed for use with bacterial organisms. The green solid and dashed lines again contain the same information but for plasmids designed for use with other eukaryotic organisms. The vertical line marks the time before and after CRISPR plasmids became available in Q3 2012. See the text for more details and variable descriptions.

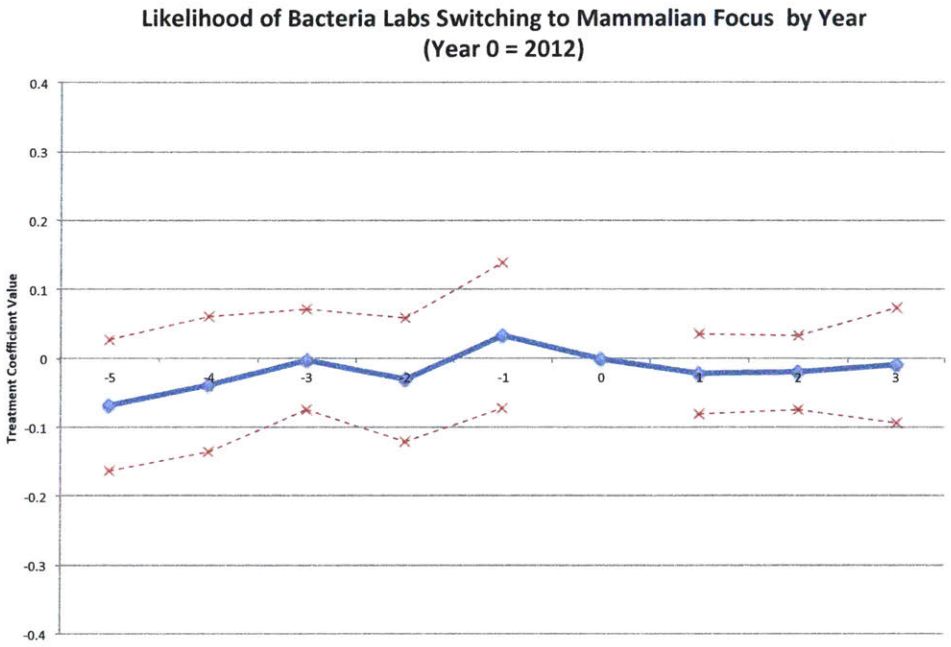
Figure 6.



Source: Web of Science Core Collection

Notes: This figure plots the trends in focus of the articles that cited at least one of the three original CRISPR articles by Doudna, Zhang, or Church. Data is aggregated by quarter from Q3 2012 through Q2 2015. The blue bars are the number of forward citing articles that had at least one keyword in the abstract or Web of Science keywords that indicated a mammalian focus The red bars are the number of forward citing articles that had at least one keyword in the abstract or Web of Science keywords that indicated a bacterial focus. The green bars are the number of forward citing articles that had at least one keyword in the abstract or Web of Science keywords that indicated an other eukaryotic focus. The categories are not mutually exclusive and a paper may be counted more than once. See the text and Appendix A for more details and variable descriptions.

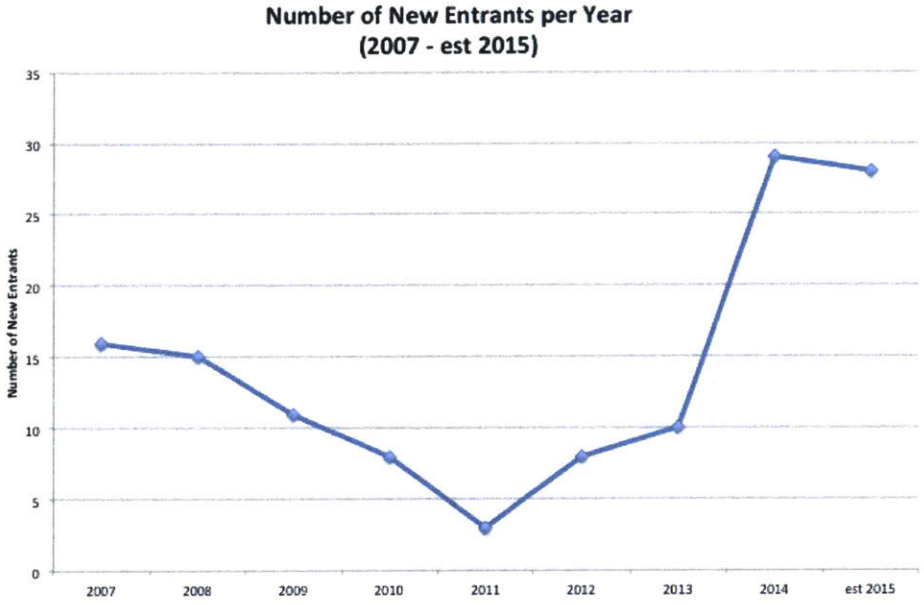
Figure 7.



Source: Scopus Publication History Database

Notes: This figure plots the impact of CRISPR/Cas9 on the likelihood of researchers switching their research focus to mammalian organisms given they began in bacterial organism research. The control group is researchers beginning in mammalian organisms and their likelihood of switching to bacterial organism research. Outcome variable: binary for if an author switched the focus of his or her research in year t from the initial research focus. An author's field of initial focus is determined by which category had an average proportional focus over all the author's publications from 2000-2006 greater than 60 percent. A paper's proportional focus for bacterial or mammalian categories is calculated as the number of keywords in Appendix A appearing in the paper abstract or Scopus keywords for that category over all keywords in Appendix A appearing in the paper abstract or Scopus keywords. An author was considered to switch focus in year t if 60 percent of the proportional focus in that year's papers was in a different category from that author's initial category. On the x-axis is the number of years from a relative year 0, here 2012 when CRISPR/Cas9 was introduced. This specification is based on an author-year level and the coefficients are estimates from OLS model with year and author fixed effects. The sample includes all years 2007-2015 and all authors that are initially classified as bacterial or mammalian that published both before and after 2012. Standard errors are block bootstrapped, clustered at the author level. Full regression results are in Table 7. See the text and Appendix A for more details and variable descriptions.

Figure 8.



Source: Scopus Publication History Database

Notes: This figure plots the number of authors appearing for the first time in the Scopus Publication History Database in years 2007-est. 2015 (estimated from June 2015). The trend shows a large number of new authors appearing for the first time after the introduction of CRISPR/Cas9 in 2012. See the text and Appendix A for more details and variable descriptions.

Table 1. Summary Statistics for Addgene Plasmid Order Database

Variable	N	Mean	Std. Dev.	Min	Max
Addgene Plasmid Order Database (Plasmid-Quarter, 2004Q3 - 2014Q3)					
Plasmid_Order_Total	276,792	1.4747	4.7412	0	488
Plasmid_Age (quarters)	276,792	10.3765	8.0103	0	40
Mammal_Expression	276,792	0.6592	0.4740	0	1
Bacterial_Expression	276,792	0.2375	0.4255	0	1
All_Other_Eukaryotes_Expression	276,792	0.1433	0.3504	0	1
CRISPR	276,792	0.0031	0.0554	0	1
*Mammal_Expression*CRISPR*	276,792	0.0018	0.0428	0	1
*Bacterial_Expression*CRISPR*	276,792	0.0006	0.0235	0	1
*All_Other_Eukaryotes_Expression*CRISPR*	276,792	0.0007	0.0271	0	1
Requestor_First_Quarter	276,792	0.1537	0.3606	0	1

Source: Addgene Plasmid Database

Notes: This data is aggregated to the plasmid-quarter level for Quarter 3 2004 (the first quarter of Addgene's existence) and Quarter 3 2014 (the most recent data available). It is a balanced panel so that quarters where a plasmid has zero orders still appears. Plasmid_Order_Total is the number of orders a unique plasmid received each quarter. Plasmid_Age is the age of the plasmid in quarters from its date of availability on the Addgene website. The Expression variables are binary indicators for whether the plasmid is for use in mammalian, bacterial, or other eukaryotic research as indicated by Addgene and are not time dependent. The categories are not mutually exclusive. The CRISPR variable is a binary indicator for whether the plasmid contains the CRISPR/Cas9 system. These are not time dependent, but are only available after June 2012. The Expression*CRISPR variables are interactions to indicate whether a plasmid contains CRISPR and is for use in mammalian, bacterial, or other eukaryotic research. Requestor_First_Quarter is a binary variable for whether a new requestor first ordered the unique plasmid in the quarter. See the text for more details and variable descriptions.

Table 2. Summary Statistics for Publication Histories

Variable	N	Mean	Std. Dev.	Min	Max
All Authors (Author-Year Panel)					
Total Papers	8,442	7.0142	8.6003	1	180
Percent Change in Papers	7,779	0.2839	1.1359	-0.9565217	28.33333
Year	8,442	2011.0890	2.5681	2007	2015
Age	8,442	17.9147	9.7901	0	56
PostCRISPR	8,442	0.3492	0.4767	0	1
Mammalian Lab	8,442	0.5936	0.4912	0	1
Other Eukaryote Lab	8,442	0.1902	0.3925	0	1
Bacterial Lab	8,442	0.1391	0.3460	0	1
*Mammalian Lab*PostCRISPR*	8,442	0.2113	0.4083	0	1
*Other Eukaryote Lab*PostCRISPR*	8,442	0.0755	0.2641	0	1
*Bacterial Lab*PostCRISPR*	8,442	0.0586	0.2350	0	1
%Mammalian Focus	8,175	0.6154	0.3464	0	1
%Other Eukaryote Focus	8,175	0.2566	0.2850	0	1
%Bacterial Focus	8,175	0.1280	0.2517	0	1
*%Mammalian Focus*PostCRISPR*	8,175	0.2096	0.3547	0	1
*%Other Eukaryote Focus*PostCRISPR*	8,175	0.0911	0.2133	0	1
*%Bacterial Focus*PostCRISPR*	8,175	0.0470	0.1659	0	1

Source: Scopus Publication History Database

Notes: This data is aggregated to the author-year level for 2007-June 2015. Authors included in this database are the set of last listed authors that wrote at least one paper citing to one of the original three CRISPR articles by Doudna, Zhang, or Church. These authors' publication histories were then extracted from Scopus and aggregated by author-year for analysis. Total Papers is the number of papers an author published each year. Percent Change in Papers is the percent increase or decrease in papers for the author from the previous year (no changes are calculated from years where there are no publications). Age is the number of years since the author's first publication listed in Scopus (this can be before 2007). PostCRISPR is a binary variable for the years 2012 and later. The %Focus variables are the average proportional focus on mammalian, bacterial, or other eukaryotic research for an author's papers each year. A paper's proportional focus by category is calculated as the number of keywords in Appendix A appearing in the paper abstract or Scopus keywords for that category over all keywords in Appendix A appearing in the paper abstract or Scopus keywords. The Lab variables are binary indicators for the author's initial field of focus. An author's field of initial focus is determined by which category had an average proportional focus over all the author's publications from 2000-2006 greater than 60 percent. The %Focus*PostCRISPR and Lab*PostCRISPR variables are interaction variables for the different measures of author focus in years 2012 and after. See the text and Appendix A for more details and variable descriptions.

Table 3. Decomposition of Addgene Plasmid Ordering Trends

Model		OLS, Aggregated by Year		
Dependent Variable		Plasmid_Order_Total		
	Coefficient	Standard Error	95% CI Low	95% CI High
Mammal Trend After CRISPR	**338.487*** **	63.153	213.381	463.594
Mammal Trend	**405.524*** **	12.795	380.177	430.870
Mammal	**-1734.849*** **	254.216	-2,238.449	-1,231.250
All Other Eukaryote Trend After CRISPR	-48.817	63.153	-173.923	76.290
All Other Eukaryote Trend	**-37.682** **	12.795	-63.028	-12.335
All Other Eukaryote	205.209	254.216	-298.391	708.808
Bacteria Trend After CRISPR	**96.512***	44.656	8.049	184.976
Bacteria Trend	**71.964*** **	9.047	54.041	89.886
Bacteria (regression constant)	**-379.629***	179.758	-735.727	-23.530
Observations	123			
R squared	0.991			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Addgene Plasmid Database

Notes: The data in the table shows the decomposition of the trends in Figure 5. It provides the statistical significance of the additional effect of CRISPR on the mammalian plasmid ordering trends at Addgene over the other categories (bacterial and other eukaryotic). The analysis controls for the ordering trends of each category as well as the non-CRISPR trend by quarter. The trends after the introduction of CRISPR/Cas9 are the variables of interest and represent the effect of CRISPR/Cas9 on plasmid order totals for mammalian research as compared to the other categories. Unique plasmids may be ordered more than once by different researchers, categories are mutually exclusive, and plasmids with no category are removed. Aggregate orders by category are calculated by quarter from Q3 2004 through Q3 2014. See the text for more details and variable descriptions.

Table 4. Effect of Mammalian CRISPR Plasmids on Total Orders

Model	(1) OLS	(2) Zero-Inflated Poisson
Dependent Variable	Plasmid_Order_Total	Plasmid_Order_Total
*Mammal*CRISPR*	**19.300*** **	**1.555***
		0.442
	(6.055)	(0.250)
Mammal	**0.971*** **	**1.804*** **
		0.590
	(0.043)	(0.044)
CRISPR	**10.781*** **	**9.477*** **
		2.249
	(2.257)	(0.183)
All_Other_Eukaryotes	**-0.233*** **	**0.888** **
		-0.119
	(0.035)	(0.057)
*All_Other_Eukaryotes*CRISPR*	**0.836**	**1.388**
		0.328
	(4.118)	(0.305)
Requestor_First_Year	**1.774*** **	**1.186*** **
		0.171
	(0.065)	(0.016)
constant	**0.078**	**0.626*** **
		-0.469
	(0.188)	(0.098)
Fixed Effects	Quarter, Age	Quarter, Age
Standard Errors	Robust, Clustered	Robust, Clustered
Observations	276,792	276,792
Plasmids	19,147	19,147
Log pseudolikelihood	-	-527,024.7
(pseudo) R squared	0.119	-

Robust standard errors in parentheses, clustered by plasmid in all specifications. OLS specification reports coefficients in bold. Poisson specification reports IRRs in bold.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Source: Addgene Plasmid Database

Notes: This table demonstrates the additional effect being a CRISPR and a mammalian plasmid has on total plasmid orders. The data is aggregated to the plasmid-quarter level for Quarter 3 2004 (the first quarter of Addgene's existence) and Quarter 3 2014 (the most recent data available). It is a balanced panel so that quarters where a plasmid has zero orders still appears. The dependent variable in all specifications is Plasmid_Order_Total, the number of orders a unique plasmid received each quarter. The Expression variables are binary indicators for whether the plasmid is for use in mammalian or other eukaryotic research as indicated by Addgene and are not time dependent (bacterial is the omitted category). The CRISPR variable is a binary indicator for whether the plasmid contains the CRISPR/Cas9 system. These are not time dependent, but are only available after June 2012. The Expression*CRISPR variables are interactions to indicate whether a plasmid contains CRISPR and is for use in mammalian or other eukaryotic research. Requestor_First_Quarter is a binary variable for whether a new requestor first ordered the unique plasmid in the quarter. Model(1) is OLS with FEs for Quarter and Plasmid Age to account for time and plasmid cohort effects (FEs for unique plasmid could not be included since the variables of interest do not vary over time). Model(2) is a Zero-inflated Poisson specification to account for the large number of zero orders in the dataset. See the text for more details and variable descriptions.

Table 5. Number of Orders per Unique CRISPR Plasmid by Sub-field

	Mammal CRISPR	Bacteria CRISPR	Other Eukaryote CRISPR	No Expression CRISPR	Total CRISPR
Total Orders	16,103	1,596	2,497	7,543	27,739
Number of Plasmids	130	23	40	74	267
Orders per Plasmid	***124***	***69***	***62***	***102***	***104***

Source: Addgene Plasmid Database

Notes: The table demonstrates that the additional orders for plasmids that are for Mammalian CRISPR research are not due solely to the fact that there are more unique Mammalian CRISPR plasmids. Total orders is the number of times unique plasmids in each CRISPR-Expression category were ordered from Quarter 3 2012 through Quarter 3 2014. The Number of Plasmids are the number of unique plasmids in each CRISPR-Expression category available for order. Orders per Plasmid are the average number of orders each unique plasmid received by CRISPR-Expression category. See the text for more details and variable descriptions.

Table 6. The Effect of CRISPR/Cas9 on Sub-Field Productivity

Model	(1) OLS	(2) OLS	(3) OLS, FE	(4) OLS, FE
Dependent Variable	Percent Change in Papers	Percent Change in Papers	Percent Change in Papers	Percent Change in Papers
*Mammalian Lab*PostCRISPR*	0.097	0.090	-0.005	
	(0.072)	(0.071)	(0.055)	
*Other Eukaryote Lab*PostCRISPR*	0.298*	0.285*	0.075	
	(0.117)	(0.116)	(0.088)	
*%Mammalian Focus*PostCRISPR*				0.050
				(0.117)
*%Other Eukaryote Focus*PostCRISPR*				0.199
				(0.155)
Mammalian Lab	-0.002	-0.016		
	(0.035)	(0.036)		
Other Eukaryote Lab	0.077	0.010		
	(0.050)	(0.053)		
%Mammalian Focus				0.000
				(0.137)
%Other Eukaryote Focus				-0.099
				(0.135)
constant	0.351***	0.311***	0.737***	0.745***
	(0.043)	(0.094)	(0.126)	(0.158)
Fixed Effects	Year	Year, Age	Year, Age, Author	Year, Age, Author
Controls	No	No		No
Standard Errors	Robust, Clustered	Robust, Clustered	Robust, Clustered	Robust, Clustered
Observations	6,897	6,897	6,897	6,736
R squared	0.005	0.013	0.003	0.003

Robust standard errors in parentheses, clustered by author in all specifications.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Scopus Publication History Database

Notes: The table shows that there is no relative increase in productivity for mammalian researchers after CRISPR. This data is aggregated to the author-year level for 2007-2014. Authors included in this database are the set of last listed authors that wrote at least one paper citing to one of the original three CRISPR articles by Doudna, Zhang, or Church. These authors' publication histories were then extracted from Scopus and aggregated by author-year for analysis. The dependent variable in all specifications is Percent Change in Papers, the percent increase or decrease in papers for the author from the previous year (no changes are calculated from years where there are no publications). Age is the number of years since the author's first publication listed in Scopus (this can be before 2007). PostCRISPR is a binary variable for the years 2012 and later. The %Focus variables are the average proportional focus on mammalian, bacterial, or other eukaryotic research for an author's papers each year. A paper's proportional focus by category is calculated as the number of keywords in Appendix A appearing in the paper abstract or Scopus keywords for that category over all keywords in Appendix A appearing in the paper abstract or Scopus keywords. The Lab variables are binary indicators for the author's initial field of focus. An author's field of initial focus is determined by which category had an average proportional focus over all the author's publications from 2000-2006 greater than 60 percent. The %Focus*PostCRISPR and Lab*PostCRISPR variables are interaction variables for the different measures of author focus in years 2012 and after. Fixed effects for Year are included in all specifications. Fixed Effects for Age of Author's Career are added in Models(2)-(4) and FEs for Author are added in Models(3)-(4). See the text and Appendix A for more details and variable descriptions.

Table 7. Treatment Effect of CRISPR on the Likelihood of Switching to an Alternative Focus

Model	(1) OLS	(2) OLS, FE	(3) OLS, FE	(4) OLS, FE
Dependent Variable	Shift in Paper Focus	Shift in Paper Focus	Shift in Paper Focus	Shift in Paper Focus
*Bacteria*PostCRISPR*	0.010	-0.001	-0.001	
	(0.031)	(0.024)	(0.023)	
Bacteria	0.115***			
	(0.019)			
PostCRISPR	0.006	0.006		
	(0.006)	(0.005)		
*Bacteria*2007*				-0.069
				(0.049)
*Bacteria*2008*				-0.038
				(0.050)
*Bacteria*2009*				-0.002
				(0.037)
*Bacteria*2010*				-0.031
				(0.046)
*Bacteria*2011*				0.033
				(0.054)
*Bacteria*2013*				-0.023
				(0.030)
*Bacteria*2014*				-0.020
				(0.028)
*Bacteria*2015*				-0.010
				(0.043)
constant	0.0200***	0.0348***	0.0363***	0.0364***
	(0.003)	(0.005)	(0.009)	(0.008)
Fixed Effects		Author	Year, Author	Year, Author
Standard Errors	Block Bootstrap	Block Bootstrap	Block Bootstrap	Block Bootstrap
Observations	4,962	4,962	4,962	4,962
R Squared	0.043	0.000	0.004	0.009

Block Bootstrap standard errors in parentheses, clustered by author in all specifications.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Scopus Publication History Database

Notes: This table shows the lack of impact of CRISPR/Cas9 on the likelihood of researchers switching their research focus to mammalian organisms given they began in bacterial organism research through differences-in-differences specifications. The control group is researchers beginning in mammalian organisms and their subsequent likelihood of switching to bacterial organism research. Dependent variable: binary for if an author switched the focus of his or her research in year t from the initial research focus. An author's field of initial focus is determined by which category had an average proportional focus over all the author's publications from 2000-2006 greater than 60 percent. A paper's proportional focus for bacterial or mammalian categories is calculated as the number of keywords in Appendix A appearing in the paper abstract or Scopus keywords for that category over all keywords in Appendix A appearing in the paper abstract or Scopus keywords. An author was considered to switch focus in year t if 60 percent of the proportional focus in that year's papers was in a different category from that author's initial category. The sample includes all years 2007-2015 and all authors that are initially classified as bacterial or mammalian that published both before and after 2012. Standard errors are block bootstrapped, clustered at the author level in all specifications. FEs for Author are in Models(2)-(4) and FEs for Year are added in Models(3)-(4). Model(4) provides the coefficients plotted in Figure 7. See the text and Appendix A for more details and variable descriptions.

Table 8. Change in New Entrant Focus after CRISPR/Cas9

Model	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS
Dependent Variable	%Mammalian Focus	%Bacterial Focus	%Other Eukaryote Focus	Dummy Mammalian Focus	Dummy Bacterial Focus	Dummy Other Eukaryote Focus
PostCRISPR	0.296*	0.136*	-0.431**	0.321*	0.152**	-0.447*
	(0.119)	(0.062)	(0.137)	(0.160)	(0.054)	(0.175)
constant	0.315**	0.063	0.622***	0.222	0.000*	0.556**
	(0.104)	(0.035)	(0.131)	(0.141)	(0.000)	(0.169)
Standard Errors	Robust	Robust	Robust	Robust	Robust	Robust
Observations	55	55	55	55	55	55
R squared	0.081	0.025	0.230	0.057	0.029	0.184

Robust standard errors in parentheses. Data is restricted to new entering authors each year from 2011 forward.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Source: Scopus Publication History Database

Notes: The table shows that new authors first appearing in the publication history database after the introduction of CRISPR/Cas9 may have more of a mammalian research focus. Authors included in this database are the set of last listed authors that wrote at least one paper citing to one of the original three CRISPR articles by Doudna, Zhang, or Church for only their first year in the database. The sample is new entering authors from 2011-2015. The dependent variable in Models(1)-(3) is %Focus for each category, the average proportional focus on mammalian, bacterial, or other eukaryotic research for an author's papers in their first year. A paper's proportional focus by category is calculated as the number of keywords in Appendix A appearing in the paper abstract or Scopus keywords for that category over all keywords in Appendix A appearing in the paper abstract or Scopus keywords. The dependent variable in Models(4)-(6) is Dummy Focus for each category, a binary indication for if %Focus is greater than 60 percent. PostCRISPR is a binary variable for the years 2012 and later. All standard errors are robust in all specifications. See the text and Appendix A for more details and variable descriptions.

9 Appendix A

Author Focus Keywords

Keywords for Scopus	Bacteria	Mammal	Other
archaea	1		
bacillus	1		
bacter	1		
clostridium	1		
coccus	1		
coli	1		
haemophilus	1		
listeria	1		
prokaryot	1		
pseudomonas	1		
ralstonia	1		
streptomyces	1		
xanthomonas	1		
aav		1	
adeno		1	
animal		1	
bovine		1	
cattle		1	
chimpanzee		1	
human		1	
macaque		1	
mammal		1	
mice		1	
monkey		1	
mouse		1	
pigs		1	
porcine		1	
primate		1	
rats		1	
rodent		1	
anopheles			1
arabidopsis			1
candida			1
cerevisiae			1
ctenopharyngodon			1
cucumber			1
drosophila			1
elegans			1
eukaryot			1
fish			1
flea			1
flounder			1
frog			1
fungal			1
fungi			1
insect			1
lepidoptera			1
mosquito			1
newt			1
nicotiana			1
orchid			1
paralichthys			1
plant			1
pristionchus			1
saccharomyces			1
urchin			1
silkworm			1
tobacco			1
tomato			1
vanilla			1
wheat			1
worm			1
xenopus			1
yeast			1
zebrafish			1

Chapter 2

Who tries (and who succeeds) in staying at the forefront of Science: Evidence from the DNA-editing technology, CRISPR

(with Neil Thompson)

1 Introduction

When promising new scientific tools appear, a researcher faces difficult considerations around adoption: should she invest time and resources into experimenting with the new tool? If she does experiment, is she likely to be able to convert that experimentation into a useful output? Since the probability of someone adopting is the probability that she experiments multiplied by the probability that she can successfully convert that experimentation into a new idea,[1] both of these aspects will matter for her decision. Thus, innovation scholars, policy makers, and firms interested in the rate of scientific technology adoption must consider both. If a technology is adopted quickly, is it driven by an abundance of experimentation or an ease in converting experimentation into a useful outcome? Or conversely, if adoption is slow, is it because of a lack of experimentation or difficulties in conversion? The answers to these questions matter because they reflect differences in how a technology diffuses and what policies or interventions would be conducive to promoting adoption.

The pattern of adoption for new innovations has been well documented. For example, Rogers (1962) provides a thorough explanation of the general S-shaped pattern of innovation adoption and provides sets of factors that describe different categories of adopters. Although Rogers' work was focused on adoption in the general population, many of the factors apply directly to academic researchers considering whether to try to become Early Adopters of a new scientific tool. Most other research on adoption curves relies only on successful adoption to characterize diffusion, as the attempts to adopt (i.e. experimentation) are historically difficult to observe. But in measuring only successful adoption, the role of experimentation (whether or not successful) is lost. This despite the well-known importance of experimentation to the innovative process and adoption (e.g., March 1991, Lazear 2005, Azoulay et al. 2011, Manso 2011, von Hippel et al. 2011, von Hippel et al. 2012).

The inability to distinguish between those who experiment and those who adopt can create ambiguities that make policy formulation and product strategy more difficult, and thus make interventions less effective. For example, consider the classic technology adoption scenario: the

[1] Mathematically: $p(adoption) = p(experiment) * p(success|experiment)$

usage of new crop varieties. A failure to adopt might be because farmers were unaware of a new varietal, because they did not see the benefit of adopting, or because they tried to adopt but were unable to make the new plants prosper. Without disaggregating adoption into its components, a policy-maker or manager would not be able to distinguish where the failure to adopt occurred and thus would not know whether to invest in awareness programs, cost-sharing plans, or agricultural outreach programs.

Historically, a technology adoption setting that allows for the measurement of both experimentation and conversion, particularly an important one, has been difficult to find. But, the adoption of the CRISPR[2] gene editing system provides such a setting. A discovery of similar importance to the seminal discovery of PCR (Polymerase Chain Reaction), CRISPR allows for the first time the cheap and easy cutting-and-pasting of DNA from one organism into another. For scientists working on mammalian biology the effect of CRISPR has been particularly transformative. For society, this promises not only better academic science, but more efficacious medical treatments.

We are able to track attempts to experiment with CRISPR by academic scientists because it is instantiated as a biological part that was distributed to them almost exclusively by a single non-profit repository, Addgene. We can also track journal publications that used CRISPR, which we use to identify which academics were able to convert their experimental efforts into research output.[3]

We find that different factors characterize scientists who experiment with CRISPR from those who are able to successfully convert that experimentation into a CRISPR paper. Holding all else constant, academics with prior influential papers across multiple research areas are more likely to both experiment and convert. Having more research funding is associated with more experimentation, but interestingly not with higher conversion rates. Experience has the opposite effect, where scientists with more experience experiment less but are better at converting when they do.

Commensurate with many adoption studies, we find that early adoption rates are highest among those with the most to gain. In the case of CRISPR, these are researchers working with mammalian cells, where CRISPR represents the largest improvements over the previous techniques. This increased adoption, however, is entirely accounted for by higher experimentation by those scientists as compared to those working in bacteria. If anything, these mammalian scientists are less able to convert their experimentation into adoption, although this effect is mitigated if the experimenting scientist is in close proximity to the original discovery (for the case of mammalian cell CRISPR, in Cambridge, MA).

---

[2] Short for Clustered Regularly Interspaced Short Palindromic Repeats.

[3] We do not attempt to look at industrial scientists using CRISPR because both the experimentation and conversion pieces are substantially harder to track. Addgene generally does not distribute to for-profit organizations.

2 Technology Adoption

Technology adoption is a key factor for economic growth. The Solow (1956) model shows that an economy's growth does not occur solely due to the accumulation of capital or labor. Instead, a residual, denoted "A," overcomes diminishing returns to labor and capital and allows for growth. This "A" is a factor that enhances the quality of physical or human capital and is usually thought of as the advancement of ideas or technology. Others such as Romer (1990) and Aghion and Howitt (1992) enhanced this model by explicitly introducing idea creation and showing that it is endogenous to growth. Economies rely on the introduction and adoption of new ideas to grow. In such models, either explicitly or implicitly, the technological advancements from new ideas must be adopted for the economy to benefit. A number of studies show the importance of technology adoption for economic growth and technological progress empirically. For example, adoption effects on growth has been shown in semiconductors (Bresnahan and Trajtenberg 1995), in hybrid corn (Griliches 1957, 1958) and for the steam engine (Rosenberg and Trajtenberg 2004). New technologies can also affect the direction of technological progress (e.g. Hanlon 2015).

In scientific research specifically, adoption of new technologies can have profound effects on scientific progress. Access and adoption of new tools has been shown to lead to more paper citations (Furman and Stern 2011), more follow-on research (Murray et al. 2016), and increased research in the field that received a new tool (Teodoridis 2018, Thompson et al. 2017a). New tools may also encourage new directions of research inquiry. Murray et al. (2016) found that restrictions on access to the Oncomouse led to a very diverse set of research paths that seem to encourage more horizontal exploration. Teodoridis (2018) found that new technologies bring into the field scientists who either specialize or are generalists. The generalists connect specialists to new research opportunities opened by the new technology and create novel research avenues. Thompson et al. (2017a) found that the adoption of new tools for writing DNA sequences generated more impactful work and led to more follow-on research.

The importance of technology adoption on innovation can also be seen in the actions of scientists who are denied it. For example, Murray (2010) documents the contentious history of the Oncomouse, where researchers were denied the ability to adopt an animal model for cancer unless they agreed to onerous legal terms. Scientists were furious and influential organizations like MIT and Johns Hopkins refused to agree to the terms. Ultimately, the backlash was sufficiently strong that DuPont, the IP owner, reversed its stance.

The literature on the adoption of tools discussed above focuses mostly on productivity outputs, but does not cover the types of people likely to experiment with and be early adopters of new tools. That is its own separate literature on the adoption curve. In 1962, Everett Rogers published Diffusion of Innovations and first categorized adopters, starting with the Innovators and Early Adopters (the first people to adopt a new technology) to the Laggards (the last people to

adopt). Rogers noted that Early Adopters in an innovation's lifecycle can be characterized broadly by Socioeconomics (e.g., people with higher social status and more wealth), Personality (e.g., people that are more rational and have a favorable attitude towards change), and Communication Behavior (e.g., people seen as opinion leaders and those who are highly interconnected with others from diverse locations). Although Rogers was writing about Early Adopters in general, many of the factors he highlights apply directly to technology adoption by scientists.

By necessity, to successfully adopt a new tool or idea a scientist must try it (experiment with it). The economic literature in experimentation often focuses on the benefits of experimentation more generally, but a few authors focus specifically on the types of individuals who are likely to engage in experimentation. Not surprisingly, certain types of people are more likely to be attracted to experimentation than others. For example, Azoulay et al. (2011) compare scientists appointed to be HHMI investigators to those who are similar but did not become HHMI investigators to study the incentives that encourage exploratory rather than exploitative behavior from scientists. The authors measure explorative behavior through the novelty of keywords, the breadth of the scientists' research before and after the appointment to HHMI, and the number of times a scientist has a top cited paper by publication cohort. They find that HHMI investigators fare better in these measures after their appointment.

The characteristics of individuals that experiment is also explored in another branch of the literature, the work by von Hippel and co-authors on user innovators (e.g. von Hippel et al. 2011 and von Hippel et al. 2012). User innovators are people who develop or modify products for their own use on their own time. Such users experiment with designs to suit their needs and sometimes their product designs will later be used by large manufacturers. Surveys conducted by von Hippel and co-authors find that user innovators are willing to spend personal resources on their products, are highly educated, usually have a technical degree relevant to the product domain, and are often male.

Finally, a third type of literature focuses on the characteristics of a different type of experimenter – the entrepreneur. Lazear (2005) provides some evidence that students that go on to start businesses are "jacks of all trade," having a wider variety of roles at work prior to starting their own company or studied more than one subject in school. The implication is that people with more balanced backgrounds are more willing to take on the risk of starting their own business.

These literatures on experimentation touch on many of the same traits Rogers identified for Early Adopters, suggesting that individuals who experiment and adopt early in the technology's lifecycle have similar characteristics. Given that one is a necessary precursor to the other, this isn't surprising. Based on the findings from these literatures, we focus our analyses of individual experimentation and conversion on the following characteristics of the scientists:

- Location
- Availability of Resources
- Ability to Apply the New Technological Knowledge
- Opinion Leadership
- Social Participation
- Novelty
- Breadth of Research

We discuss appropriate measures for these factors and their mapping to our data in Section 4.

3 Gene editing with CRISPR

For most industries, it is difficult to observe experimentation and successful adoption separately. Rarely shown is the testing process where technologies are either incorporated into the science or discarded. We study an instance where this experimentation is visible, the introduction of the new gene-editing tool CRISPR into genetic engineering. This is a particularly fertile setting because CRISPR is an important research breakthrough, because its use has been growing rapidly, and because the primary means of academic experimentation starts with a request for the CRISPR biological parts from a central biological resource center, Addgene. As such, it is a setting where adoption is important and interest in experimentation is widespread and can be observed. To understand the factors that influence the experimentation with CRISPR and its conversion into successful ideas, it is helpful to understand its biological function.

3.1 DNA and gene editing

DNA is the parts list and blueprint for life; it determines the proteins that make up the human body and the regulatory processes that manages them. In the early 1970s, Herbert Boyer and Stanley Cohen first demonstrated that DNA could be transferred between organisms by cutting the DNA of one organism and inserting a part from another. This launched the field known today as genetic engineering.

Follow-on research demonstrated the benefits of editing DNA. These included the creation of model organisms that help researchers understand and fight human diseases, and the adding or deleting of functionality for existing organisms, for example by making crops resistant to pesticides. Editing DNA before CRISPR was technically challenging. For bacteria or other prokaryotes (that have cells without nucleuses) there are some standard and effective alternative editing methods. For higher order species, like mammals, even very-recent alternatives[4] are challenging as these

[4] Called Zinc Finger Nucleases (ZFN) and Transcription Activator-Like Effector Nucleases (TALENs) (Moscou and Bogdanove, 2009; Boch, et al., 2009)

techniques were more difficult and time-consuming to work with. It is a testament to the importance of gene editing in mammals and the prior paucity of good tools that one of these was nevertheless chosen as "Method of the Year" in 2011 by Nature Methods (Method, 2012). But, in a true shock to the field, CRISPR appeared only months later. This new technology, CRISPR, is a universal cutting technology that works across organisms and is a much more flexible option for editing than any other, especially for organisms like mammals.

3.2 CRISPR

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were discovered in bacteria as early as 1987, but their exact purpose and function were not well understood until 2007 when it was discovered they are an adaptive part of the bacterial immune system (Barrangou et al, 2007). Bacteria use CRISPR to recognize the DNA sequences of viruses and then use the Cas9 protein to cut up the virus' DNA, rendering it harmless to the bacteria. Over five years, researchers were able to repurpose this natural system of matching and cutting virus DNA.

In June 2012, Professors Jennifer Doudna and Emmanuelle Charpentier at the University of California, Berkeley first introduced the CRISPR system for genome editing (Jinek, et al., 2012). Doudna and Charpendier proved in test tubes that the CRISPR system could be used to find and edit any sequence (not just viral) through the use of guide sequences, which direct the cutting enzyme to the right place on the DNA. Doudna and Charpentier's work did the necessary first steps to allow the CRISPR system to be used to edit the DNA of organisms like bacteria. In January 2013, MIT Professor Feng Zhang and his collaborators showed that the CRISPR system could be used to also edit mammalian cells, including human cell lines (Cong et al., 2013). This work and the related work of George Church and his colleagues at Harvard Medical School (Mali et al. 2013) demonstrated the flexibility and ease of use of the new CRISPR tool. This was particularly welcomed in the mammalian research community,[5] including those working on human cells, because of the enormous improvement over previous methods.

The power of the CRISPR system to find and cut the correct target DNA, and its contrast with the previous methods, can be understood by an analogy to cutting-and-pasting text in documents. For example, say that in a Word document we are looking to replace the word "absolutely" with "completely" (the analogy being the change of one section of DNA to another). With the previous editing technologies, it was only possible to search for a short string, such as "abs." This will find "absolutely," but also "absolute," and "abstract," with the consequence that our new document might not be sensible (corresponding to a non-functional DNA sequence). With

[5] This actually applied to all eukaryotes (organisms with nucleuses in their cells), but we refer to these in the text as mammalian cells to keep the technical jargon at a minimum and because this was probably the most important use case in the biomedical community.

CRISPR, it is possible to search for the string "absolutely" and have only that string identified. In addition, a researcher can use CRISPR to make multiple cuts at the same time in the same cell, something that was not possible with the previous editing techniques.

The enthusiasm from the genetic engineering community to the release of the CRISPR tools was almost immediate because of its accuracy, flexibility, and relative ease of use (Pennisi, 2013; Regalado, 2014). Between June 2012 and June 2016, almost 4,000 CRISPR-related articles were published worldwide, an average of ~100 articles a month since Doudna and Charpendier's first.

Patent applications mentioning CRISPR and funding for venture backed firms licensed to use CRISPR technology have also soared (Ledford 2015). In 2014, over 150 patent applications involving CRISPR were published. Funding also flowed to CRISPR commercialization efforts. The top biotech firms founded on CRISPR technology, Caribou Biosciences (Berkeley, CA), Editas Medicine (Cambridge, MA), CRISPR Therapeutics (Basel, Switzerland), and Intellia Therapeutics (Cambridge, MA) collectively raised initial funding of more than $150 million. The last three all had IPOs in 2016, each currently valued at over $500 million.

CRISPR's effect on biological research has been profound as geneticist John Schimenti at Cornell University noted in a recent article, "I've seen two huge developments since I've been in science: CRISPR and PCR... CRISPR is impacting the life sciences in so many ways" (Ledford 2015). One of the original inventors, Jennifer Doudna, stated in a February 2015 JAMA editorial, "This discovery has triggered a veritable revolution as laboratories worldwide have begun to introduce or correct mutations in cells and organisms with a level of ease and efficiency not previously possible." (Doudna, 2015). For example, CRISPR has already been used to modify crops to be blight resistant (Wang et al. 2014) and to create "malaria-proof" mosquitoes that are genetically unable to be affected by or transmit malaria (Gantz et al. 2015). The introduction of CRISPR is likely to be especially useful in medical applications since it may ultimately allow us to actively correct genetic errors.[6] In the interim it is allowing researchers to easily build mouse and human cell disease models with specific mutations that are useful for testing drugs. For example, Feng Zhang's lab has created a "Cas9 mouse" (Platt et al. 2014) that can be modified to model lung cancer. Before CRISPR, creating such a mouse model took huge teams of people and a decade of time, but this model was designed by one person in four months (Specter 2015).

Even researchers who believed that CRISPR would be beneficial for their research might delay adopting because the tool was still developing. Rather than being an early adopter today, such researchers could wait until CRISPR development matured, potentially making adoption easier and perhaps yielding a more-powerful tool.[7] Further, some research areas (e.g., neurodegenerative diseases like Huntington's) have only recently been able to use CRISPR effectively. Because of the

[6] This is already being done in human embryos that are not implanted (Ma et al. 2017).

[7] For example, almost immediately research began on new and better cutting enzymes that could replace Cas9.

large benefits of CRISPR, but also its continued development and shortcomings, researchers have reasons both for and against becoming an Early Adopter.

3.3 Access to CRISPR

Traditionally, obtaining biological materials in order to experiment with them has been challenging for technical, social, and legal reasons. As with most genetic tools, the DNA representation of the CRISPR tool is usually encoded in a plasmid, a circular DNA construct that preserves the tool and facilitates getting the new DNA into host cells. Labs create plasmids in the course of an experiment and keep them for use in future studies. Labs sometimes share their plasmids with other academics who want to replicate or build on their research, but they will sometimes refrain from sharing to deter competition (Thompson et al., 2017b), or impose legal terms that are unattractive or delaying (Walsh et al. 2003 and 2007). Even when requests do not face these hurdles, they can be delayed by logistics when a lab with a popular plasmid (like CRISPR) is inundated with more requests than they can handle.

To circumvent the difficulties associated with lab-to-lab material transfers, many research communities set up biological repositories, such as the American Type Culture Collection, that centralize the distribution process (Furman and Stern 2011). The focal repository for CRISPR is Addgene, a non-profit organization where researchers deposit their plasmids and which maintains and distributes them to researchers on request. Addgene is the key distributor of CRISPR plasmids for academic scientists, especially in the years we study.

3.4 Addgene

In 2004 Melina Fan, Kenneth Fan, and Benjie Chen founded Addgene as a not-for-profit biological resource center for scientists to easily share plasmids for use in biological research (Fan et al., 2005). Addgene not only stores plasmids donated by academic researchers all over the world, but also validates the materials and facilitates their distribution to other academics in 85 different countries.

When the CRISPR tool was first introduced in 2012 and 2013, Doudna, Charpendier, and Zhang created plasmids for CRISPR and donated them to Addgene at the time the original papers were published. As Zhang mentioned in a talk at MIT in 2015, he gave the plasmids to Addgene for distribution because his lab wouldn't have time to do science if they responded to all the requests from other researchers. As new CRISPR plasmids were developed, they also were donated to Addgene. Scientists still actively refer requests for plasmids to Addgene. From 2004 to Q1 2016, the growth rate of Addgene orders has been increasing over time, as shown in Figure 1. Even during

this fast overall growth, CRISPR plasmid orders outpaced others, rising from 0.1% of all Addgene orders in 2012 to over 20% by Q1 2016.

[Figure 1 here]

Addgene has a price of $65 per plasmid which has not changed since it began in 2004, meaning that researchers have consistently had easy access to experimentation with the genetic parts that Addgene distributes. This low cost and quality control process encourages labs to order directly from Addgene rather than make their own or attempt to get it from the original lab. This alleviates the burden on the individual research labs and eases the access to the tools.

4 Data and Measures

This paper's outcomes of interest are experimentation with and conversion of CRISPR. We define experimentation in this context to be when a researcher becomes sufficiently interested in the technology to order a CRISPR plasmid, so that they can use it for experiments in their lab. For example, this might occur because the researcher is working on an existing project that includes gene editing, and decides to try CRISPR rather than older tools. It might also occur if a lab sees the potential to launch a new project that is only possible because of CRISPR.

Conversion in this context occurs when a researcher that has been experimenting with CRISPR is able to successfully use it to produce a new research result, as measured by the appearance of an academic publication that uses CRISPR. We ascertain that the paper uses CRISPR by looking for telltale keywords in the abstract or paper meta-data (described in more detail in Section 4.5).

This section describes how we gather measures on experimentation and conversion, how we merge them, and how we build our sample. We also discuss the secondary data sources used to construct the researcher-level covariates described in Section 2.

4.1 CRISPR Experimentation Data

To get an indicator of scientists experimenting with CRISPR, we work with Addgene to analyze their internal data on plasmid orders.[8] Addgene has a set of unique identifiers for each order, including the name and location of the scientist, the date of orders, the plasmids ordered, the

[8] The CRISPR Experimentation Dataset was used to make geographic and publication associations but the resulting data has been coded so no specific requestors or authors will be revealed in this study.

cell type each plasmid is used with, and whether the plasmid is a CRISPR plasmid. This becomes our CRISPR Experimentation Dataset.

Our 'at-risk' set for those ordering CRISPR are all other authors who publish in the same genetic engineering journals. To construct this set of authors, we use Web of Science (WOS). WOS (owned by Clarivate Analytics), is a citation database that provides bibliographic information on academic publications including data on authors, author location, journal of publication, and date of publication. In particular, we chose all authors who publish papers in genetic engineering journals from Q1 2000 – Q2 2016. We include journals in this set if they fall into any of five broad WOS categories: Genetics & Heredity, Cell Biology, Biotechnology & Applied Microbiology, Biochemistry & Molecular Biology, and Biochemical Research Methods. We also add the specific interdisciplinary journals: Nature, Science, PLOS One, and PNAS to capture the very-high prestige publications in these fields. We choose these journals because, collectively, they host the vast majority of CRISPR papers published after June 2012. We use WOS to collect information on the authors, including their names, geographic locations, organizational affiliations, and unique WOS IDs.

4.2 CRISPR Conversion Data

We determine that a scientist has successfully converted when we observe that author publishing a paper in an academic journal that uses CRISPR. To measure this, we again use the Web of Science data, but focus on the publications themselves. Our initial WOS subset includes over 1.3 million papers and almost 4.4 million authors. For each paper we collect data on the title, abstract, journal, keywords, date of publication, grants acknowledged, number of citations received, and the WOS Research Category.

To determine whether a publication involved the CRISPR tool, we searched for the terms "CRISPR" and "Cas9" in the abstract or keywords, restricting ourselves to publications after June 2012 (to exclude those who could only be talking about the bacterial immune system and not the gene-editing tool).

4.3 Disambiguating and Matching Authors

In order to construct our full dataset, we must both consolidate researchers within the CRISPR Experimentation Dataset and then link them to publication authors from the Web of Science data. This section describes the algorithm that we use to match authors in each of these stages. A full discussion of the matching process is included in Appendix C, but we describe the basic steps here.

4.3.1 Consolidating the CRISPR Experimentation Dataset

We consolidate researchers in the CRISPR Experimentation Dataset because information entered with slight differences can show up as different people. For example, Jane A. Scientist, Dr. Jane A. Scientist, and Scientist Jane all at "Research U" would appear as three people. But, given the distinct last name and identical location, these indicate the same person. For the consolidation algorithm, we use three pieces of information: full name, location, and (unconsolidated) ID number. We use these in an iterative process that tests whether a name is a match for any of the other 67,466 names in the CRISPR Experimentation Dataset.

In the first stage of the algorithm we accept as a match any two individuals with the same ID number.

In the second stage, we accept as a match any two individuals with identical names and the same location, but with different ID numbers.

In the third stage, we calculate twelve different measures of name closeness, each of which could indicate a higher likelihood that the two names are for the same person and use an ensemble approach from machine learning to determine whether the two names are a likely match.

The twelve measures that we employ are text based. For example, we calculate the 'string distance'[9] between each individual's full name and those other 67,466 names. If these exceeds a certain threshold, it is deemed to have passed that indicator. Since first names are often missing or replaced by only a first initial, our measures also include comparisons of only last names, but requires a tighter threshold than whole name matching and which varies based on the frequency of the last name.[10] We deem two names to be a match if 75% of our name measures are passed and their locations are the same.

Having established a high quality set of primary matches from the algorithm above, we search for secondary matches via a graph closure algorithm. The intuition here is a simple one that will be familiar to economists from transitivity: if we find a match between A and B, and one from B to C, then we should infer a match between A and C, even if a one-to-one comparison between A and C would be inconclusive.

We test the effectiveness of our algorithm by hand coding the matches of 50 randomly selected names against all the 67,466 other names (without knowing the algorithm's prediction for them). We find that 100% of the algorithm matches are true matches. We also test whether the algorithm is conservative, not matching some true matches. We find that, of 86 'true' matches (i.e.

[9] See, for example, van der Loo (2014) for a discussion of these measures.

[10] For the distance measures, all name strings are scrubbed of punctuation, accents, spaces, numbers, and uppercase letters. We also remove titles like Dr. and Professor from the strings.

hand-coded ones), the algorithm successfully captures 83. That is, our algorithm is 96.5% effective at identifying the hand-coded 'true' matches.

Our algorithm reduces the number of distinct individuals from 67,466 to 54,133.

4.3.2 Matching Researchers between the CRISPR Experimentation Dataset and Web of Science

Matching the 54,133 consolidated researchers in the CRISPR Experimentation Dataset to the 4.4 million authors in our WOS subset is done using the same consolidation algorithm described above, but with a pre-processing step for computational tractability and with a manual matching of institution names. In our pre-processing step, we limit pairwise comparisons to instances with the same last name and first initial, which greatly reduces the number of distance calculations that we need to perform. This restriction excludes 1.8% of the CRISPR Experimentation Dataset names because they have no first name information (in any of their orders). Once this matching is complete, we again do graph closure to obtain our final author data set.

Through these two sets of name matching we are able to associate authors with both their publications (before and after CRISPR's discovery) and their CRISPR orders.

4.4 Sample Construction

Because of CRISPR's recent introduction, we focus our analysis on Early Adopters, which we define as those who tried experimenting with CRISPR between its invention in June 2012 and the end of 2014.[11] We choose this end date because it gives us sufficient time afterwards to observe researchers turning their experimentation into a journal article (i.e., conversion). If we had instead chosen a later end date, there would be substantial truncation bias from later experimenters who would not have enough time to get their work published. For robustness, we also consider ending our sample earlier, at the end of 2013, and find very little difference in our results apart from those that arise from having a smaller dataset (e.g., less statistical power).

To look at heterogeneous adoption behavior by researchers we construct a series of covariates. To avoid endogeneity concerns, we use only pre-June 2012 factors to construct these covariates. However, this introduces the additional requirement that authors publish prior to June 2012 for us to be able to gather this data. Our sample consists of authors who have at least one publication pre-2012 (for covariate construction) and one publication post-2012 (for outcomes). This

[11] Because of this definition, authors who first order CRISPR in 2015 or early 2016 are included in the control group. Another possible choice would have been to exclude them altogether, but this would have required conditioning on an ex-post variable, which could have introduced selection effects. In any case, the inclusion or exclusion of these authors makes no difference to our results.

excludes new Ph.D. students and focuses on the behavior of incumbent academic scientists. We also restrict our data to US authors, where we are able to obtain richer covariate information.

After incorporating our data requirements, our final sample contains 164,993 authors, including 2,982 (1.8%) individuals who order CRISPR from 2012-2014. Of these 2,982 that experiment with CRISPR, 337 (11.3%) manage to publish a paper in CRISPR by Q2 2016. This puts in stark contrast the differences between those that try to be Early Adopters (experiment) and those that are successfully able to do so (conversion) by Q2 2016. Indeed, there are 8.85 experimenters for each successful Early Adopter. The experimentation and conversion curves for these 2,982 authors by quarter are shown in Figure 2.

[Figure 2 here]

Figure 3 and Figure 4 show the geographic spread of U.S. academic scientists experimenting and converting CRISPR. Figure 3 is split into three groups of states, with each color representing one-third of the experimenters with CRISPR. So, for example, California, Massachusetts, Maryland, and New York collectively house one-third of all CRISPR experimenters. Figure 4 parallels Figure 3, but shows conversion with CRISPR, rather than experimentation. Not surprisingly, the geographic concentrations are similar.

[Figure 3 and Figure 4 here]

There are a few limitations to the dataset as constructed. First, the matching algorithm is designed to be conservative, so there may be names that should be included in a group, but are not. This could affect the ultimate count of how many people ordered CRISPR. Second, we are only concerned with direct orders of CRISPR in the CRISPR Experimentation Dataset for purposes of this analysis and make no assumptions as to whether co-authors receive the plasmid too.

4.5 Measures

We construct the following measures for the outcomes and covariates of interest, with outcome variables measured only on data after June 2012 and covariates only on data before June 2012. We discuss the details of the calculation of each variable in Appendix B, but provide a summary here:

Category	Measure	Sources
Outcomes	Experimentation (Ordered CRISPR in 2012-2014)	CRISPR Orders
Outcomes	Conversion (Publication | Experimentation)	WOS abstracts and keywords where "CRISPR" and/or "Cas9" appears
Outcomes	Adoption (Published a CRISPR Paper by Q2 2016)	WOS abstracts and keywords where "CRISPR" and/or "Cas9" appears
Location	Central Location: Author is in Cambridge, MA or Berkeley, CA in 2011 or 2012	WOS author locations
Availability of Resources	US Rank	US News & World Report, 2012
Availability of Resources	University R&D Expenditures per PI	NSF HERD Survey, 2011
Availability of Resources	Average Number of Granting Agencies per Paper	WOS grant information
Ability to Apply New Technical Knowledge	Research Focus - %	Proportion of papers where the abstract contains terms indicating a particular organism type
Ability to Apply New Technical Knowledge	Dominant Research Focus – (Indicator Variable)	As for Research Focus, but indicating that 60%+ of the research is focused on that organism type
Ability to Apply New Technical Knowledge	Total Number of Papers (2009-2012)	WOS publications
Ability to Apply New Technical Knowledge	Experience in Years	Years since the earliest WOS publication in our dataset (note: right censored at 12)
Opinion Leadership	Average 5 Year Impact Factor (2009-2012)	Weighted average of Impact Factor across publications based on InCites Journal Citation Reports, 2011
Social Participation	Average Number of Co-Authors (2009-2012)	WOS authors and publications
Novelty	Number of Highly Cited Articles by Publication Cohort (2009-2012)	WOS publication date and forward citations
Breadth of Research Field	Number of Published Subject Categories	WOS Research Categories

Table 1 presents the pairwise correlations of the covariates of this analysis. It shows that there is significant correlation between some variables, for example a 35% correlation between the number of papers that an author published in the 3 years prior to CRISPR and the number of papers they have in the top 1% of the citation distribution. There are also significant negative correlations, for example a -45% correlation between authors that have a mammalian focus and those that have a bacterial focus.[12]

[Table 1 here]

4.6 Summary Statistics

Table 2 shows the summary statistics for the full experimentation 'at-risk' set, which is the set we use to understand the drivers of experimentation. Table 3 shows the summary statistics for those that have experimented (i.e., ordered CRISPR) and are 'at-risk' of conversion.

[Table 2 and Table 3 here]

On average, the authors at risk for experimenting are an accomplished group, averaging more than one paper published per year and listing an average of 1.7 granting agencies per paper. 4.1% of the authors are located in Berkeley, CA or Cambridge, MA at the time that CRISPR was discovered in these places. 63.3% of the authors work primarily in mammals. The papers they write are impactful, although some authors have far more impact than others, as would be expected. Most authors do not have a paper in the top 1% of citations in 2009-2012 (0.11 papers on average), but some are part of large influential labs and have as many as 25 papers in the top 1% of citations during those years.

Authors who experimented and are at risk of conversion are a little more likely to be in Berkeley or Cambridge, are more likely to be mammalian focused, and are even more accomplished. In this much smaller set of authors, 6.3% are located in Berkeley or Cambridge and 70.5% are mammalian focused. Their papers published in the three years before CRISPR have an average 5-year impact factor of 10.9, but with a standard deviation of 6.8. They are also more likely to have a paper in the top 1% of citations in 2009-2012, with an average of 0.55 papers.

[12] This is by construction since authors are categorized exclusively by their dominant focus (mammal, bacteria, other eukaryote) or by a no dominant focus category.

5 Methodology

5.1 Identification Strategy

This paper aims to understand the characteristics of researchers that do early experimentation and conversion with new technologies, particularly CRISPR. An ideal experiment to test this would have a large exogenous technology shock that is available to all scientists where experimentation and conversion rates would be visible. The effect of the technology shock could then be estimated by looking at the heterogeneous treatment effects across the variation present in the underlying population before the shock.

For many technologies, such an identification strategy would be challenging because adoption happens slowly. This long adoption period would create identification concerns, because 'pre-treatment' variables could substantially pre-date real decision-making. For example, the institutional affiliation for a scientist may not be predictive of technology adoption if 'pre-treatment' is 10 years earlier, and many scientists have switched affiliations in the interim. This would bias estimates towards zero because of measurement error.

Measurement for many technologies can also be difficult for the pragmatic reason that tracking outcomes, particularly experimentation with the technology, can be hard to monitor. This can introduce selection issues that also introduce bias in estimates.

The invention of CRISPR provides an apt technological setting to study early experimentation and conversion because:

i. The discovery of CRISPR was a surprise.

ii. CRISPR provided a substantial (indeed likely Nobel-worthy) improvement on the previous technology.

iii. The development of CRISPR as a gene-editing technology happened very rapidly once the potential of the technology was understood.

iv. There was near-simultaneous discovery of CRISPR's functionality for editing genes like those of mammals.[13] The Zhang and Church labs published on CRISPR within 6 months of the Doudna lab. This competition between discoverers suggests that there was little room for any of them to endogenously affect the timing of the discovery.

v. Overwhelmingly the transfer of the CRISPR materials to other scientists happened via the Addgene platform, which allows us to track experimentation.

[13] For technical reasons, the biggest impact is on most eukaryotic organisms, e.g. mammals, fish, plants, insects, but was less so on prokaryotes.

vi. The key output of successful adoption for many of these scientists are academic publications, which we can observe and in which they are obliged to recognize the usage of CRISPR, which they often do by citing the Addgene part directly.

To summarize, this paper treats the discovery of CRISPR as a shock that exogenously pushes out the frontier of knowledge by making gene editing substantially easier. Because the size of this innovation is so large, and the development of the tool so rapid, the period of experimentation and conversion begins very quickly. As such, 'pre-treatment' variables are quite recent in time, and thus variation in these is likely to be informative of scientist characteristics. We can track both experimentation with the technology (via the CRISPR orders)) and conversion of CRISPR by its usage in an academic paper.

5.2 Specifications

To study the components of adoption, we split it into experimentation (attempted adoption) and conversion (success at adopting given experimentation):

$$p(adoption) = p(experiment) * p(success|experiment) \qquad (1)$$

We analyze experimentation and conversion separately to understand the factors that contribute to each.

5.2.1 Univariate regressions

We begin with univariate analyses of the effect of each covariate. Econometrically this is problematic because the correlations shown in Table 1 could lead to substantial risk of omitted variable bias. Thus, we do not present them as valid estimates of causal effects, but as the types of data that are often presented to policy makers or managers. For example, "technology adoption is X times more likely if someone is in close proximity to the original discovery than if they are not".

We estimate the univariate effects with a linear probability model that analyzes each of the co-variates in Section 4:

$$Experimentation_{i,post2012} = \alpha_0 + \beta Factor_{i,pre2012} + \varepsilon_i \qquad (2)$$

Here Experimentation = 1 if an author i ordered CRISPR after June 2012 and 0 otherwise. Factor represents one of the measures listed in Table 1 that is hypothesized to influence adoption. To guard against reverse causation, all these factors are calculated prior to June 2012, when experimentation can begin.

For each model described in equation (2), we also run a similar model for conversion (3) where the independent variables are identical, but the dependent variable is $Publication|Experimentation_i$, whether author i published a paper on CRISPR after ordering it:

$$Publication|Experimentation_{i,post2012} = \alpha_0 + \gamma Factor_{i,pre2012} + \varepsilon_i \qquad (3)$$

Together, equations (2) and (3) allow us to separately estimate both stages in the successful adoption of CRISPR.[14]

We calibrate the coefficients that result from our univariate analyses by comparing them against the baseline rates of (a) scientists in genetic engineering experimenting with CRISPR, and (b) the average success rate in converting the experiments into research outcomes (successful adoption). In all cases, we present these as ratios of the form $\frac{Effect\ Size_{target\ group}}{Effect\ Size_{overall}}$, that is as multipliers of the baseline probabilities. Because the effect of experimentation and success | experimentation on adoption are multiplicative, as shown in equation (1), such normalization also clarifies the effect of each factor on overall adoption. For example, if the effect size for a factor on experimentation is 2x the baseline, it implies a 2x change on overall adoption.

5.2.2 Multivariate Analyses

While our analyses in 5.2.1 provide a useful overview of the heterogeneous treatment effects, it fails to disambiguate among correlated factors, for example, because authors with top cited papers are more likely to be at highly ranked schools. To address this, we control for further sets of variables, also using a linear probability model. With this adjustment, our estimator for Experimentation becomes:

$$Experimentation_{i,post2012} = \qquad (4)$$

$$\beta_0 + \beta_{Loc} Location_{i,pre2012} + \beta_{Res} Resources_{i,pre2012} + \beta_{TK} TechKnowledge_{i,pre2012} + \beta_{OL} OpinionLeader_{i,pre2012} + \beta_{Soc} Social_{i,pre2012} + \beta_N Novelty_{i,pre2012} + \beta_{Breadth} Breadth_{i,pre2012} + \varepsilon_i$$

- Location is the measure for Location.
- Resources is the set of measures for Availability of Resources.
- TechKnowledge is the set of measures for Ability to Apply New Technological Knowledge.
- OpinionLeader is the measure for Opinion Leadership.
- Social is the measure for Social Participation.

[14] Notice, we are not doing a Heckman selection model here. Doing so would allow us (if we had the right selection model) to estimate the conversion estimates for the general population of potential technology adopters (i.e., all the at-risk group for experimentation). This is not the group whose effects interest us. We are interested in the conditional estimates – how are these variables affecting the selected group – and hence we do not try to adjust for the selection into experimentation.

- Novelty is the measure for Novelty.
- Breadth is the measure for Breadth of Research Focus.

We similarly model the effect on publication, conditional on experimentation, yielding our linear probability estimator:

$$\boldsymbol{Publication|Experimentation}_{i,post2012} = \tag{5}$$

$$\begin{aligned}\beta_0 + \beta_{Loc} Location_{i,pre2012} + \beta_{Res} Resources_{i,pre2012} + \beta_{TK} TechKnowledge_{i,pre2012} \\ + \beta_{OL} OpinionLeader_{i,pre2012} + \beta_{Soc} Social_{i,pre2012} \\ + \beta_N Novelty_{i,pre2012} + \beta_{Breadth} Breadth_{i,pre2012} + \varepsilon_i\end{aligned}$$

where the factor descriptions are the same as in equation (4).

In some cases, we also consider versions of these specifications with only a subset of the covariates to illustrate the iterative effect of conditioning.

5.3 Challenges to our Identification Strategy

The largest threat to our identification would come from the transfer of materials without our being able to observe them. As noted earlier, the traditional model of material transfer was bilateral, with one scientist directly sending materials to another. In the case of CRISPR, statements to the contrary by Feng Zhang and the scale of the distribution by Addgene make it unlikely that substantial quantities of materials were transferred from the Zhang Lab to scientists directly. Nevertheless, once materials have been transferred to a lab by Addgene, there could well be sharing within that lab across projects. This could indicate that more scientists than we realize are experimenting with CRISPR, and thus our estimates for the baseline effect would be overestimates because the size of the at-risk set in the denominator would be too small.[15] To test for this, we examine instances where publications about CRISPR do not correspond to orders in the CRISPR Experimentation Dataset (either to the focal author or a co-author). If there is significant experimentation outside our data, we would also expect there to see significant numbers of academic publications that result from it. In practice we find that while there are papers that publish on CRISPR without any order of the source materials, these are overwhelmingly papers that are about CRISPR, rather than papers that use it. For example, review articles summarizing the impact of CRISPR, rather than those using it as a gene-editing tool. Since these are articles are precisely the ones that do not require the materials, they do not provide evidence of bias and thus we conclude that unobservable material transfer is not substantially affecting our results.

[15] The effect on the estimates of contributing factors would be ambiguous.

5.4 Generalizability and Estimate Interpretation

This paper is a case study of the adoption of a particularly important technology that is transforming biology. Because of the importance of CRISPR, this question is interesting in its own right. But how generalizable are these lessons to other contexts? We hypothesize that many of these lessons will be more broadly applicable, even to technology improvements that are less transformative. This hypothesis rests on the observations by researchers of similar dynamics across areas with different gains from technology adoption. For example, in Griliches 1957, he documents that while some areas (e.g. Iowa) adoption hybrid corn very rapidly, others with less to gain (e.g. Alabama) adopt more slowly. Nevertheless, he argues that such differences can be understood as variants on a larger underlying process. We would argue that CRISPR's high value to scientists makes it more like the value of hybrid corn in Iowa, generating rapid adoption, and that less-transformative discoveries would be more like hybrid corn in Alabama, generating slower adoption. Thus, we would argue that while the pace of CRISPR adoption is likely to be much rapider than other technologies, many of the underlying dynamics are likely to be shared.

Because CRISPR was a surprise discovery, many (but not all) of our coefficients will have potential causal interpretations as estimates of how scientists adopt large technological discoveries like CRISPR. Of particular interest among these causal estimates will be the heterogeneous effects – some groups will experiment or convert more than others. It will be tempting to interpret these causally, for example claiming that if we observe that researchers that have more of some characteristic (e.g., grants or top papers) are more likely to experiment with a technology, then a policy initiative to increase the prevalence of that characteristic among scientists will increase experimentation. Because we do not have random assignment (or an equivalent) of researchers for such characteristics, such claims would be overreaching. We caution against such direct causal interpretations, instead viewing our findings as informative of potential interventions that would themselves need to be tested.

6 Results

Overall, we find that of the 164,993 US authors in our data who publish in genetic engineering, 1.81% experiment with CRISPR (ordered CRISPR in 2012-2014) with an average success rate of 11.30% (published a CRISPR paper after ordering). Only 0.20% of the authors in our at-risk set are successful Early Adopters of CRISPR.[16]

[16] By the definition of Rogers (1962) this would actually imply that we were looking at "Innovators" rather than "Early Adopters". But we believe that this narrow definition of the term "Innovators" would be confusing to the reader because of its generally accepted broader meaning, and thus we continue to use the term Early Adopter.

Our ability to interpret our results as causal is nuanced and worth discussion. Many of our results should be interpreted causally insomuch as they reflect the heterogeneous responses of scientists with differing characteristics to a new technology adoption shock. Notable exceptions to this would include location and social network effects, since it is not as-good-as-random that particular researchers in Berkeley, CA and Cambridge, MA were the ones to discover CRISPR.

There are also areas, particularly around policy interventions, where we can only make correlational claims because the variation we observe is not exogenous. For example, since our variation in grants per researcher is not random, we can only report the associations between aspects of CRISPR adoption and grant funding. We cannot make a causal claim that a change in grant funding would change adoption behaviour.

6.1 The importance of separating experimentation and successful adoption

This paper contends that the separation between experimentation and conversion success is important for policy makers and managers, since factors may influence each component differentially. The importance of such analysis is illustrated in Table 4 which considers the effect that a scientist's organism focus has on their adoption of CRISPR. The intuition is that there will be more adoption of CRISPR by scientists studying mammals because (i) the CRISPR tool is a bigger improvement upon previous editing techniques in this area, and (ii) the evolutionary proximity of other mammals to humans makes them excellent models for the development of impactful (and lucrative) disease treatment in humans. Model 1 (Adoption) provides evidence that this is indeed the case, with mammalian scientists adopting 0.13 percentage points*** more often than bacteria scientists (recall: this is against an average across all authors of 0.20%). However, Model 2 (Experimentation) and Model 3 (Conversion) reveal that the result in Model 1 comes exclusively from experimentation, not conversion. Mammalian scientists are more likely to experiment with CRISPR, but if anything they are less likely to be successful in their conversion to publication.

[Table 4 here]

There are many potential reasons why mammalian scientists might adopt with these patterns. Perhaps they are overall less familiar with gene-editing since it has been so much more difficult historically? We return to this question later, but regardless of the reason, this distinction is important for policy-makers and managers, since it implies that different types of interventions might be valuable to promote adoption. For example, it might be more valuable to have a workshop on how to use CRISPR, rather than funding general awareness-raising.

As this illustration shows, there are meaningful differences in how different factors affect experimentation and successful adoption given experimentation. That nuance is lost if analysis only focuses on overall adoption.

6.2 Univariate Results

We begin by considering univariate analyses to assess overall statistical associations and, more importantly, to indicate the kinds of data that policy makers and managers are likely to be presented with (e.g., "the center for technology A should be located here because technology A was discovered here and those scientists are more likely to adopt it").

Many of our variables of interest are correlated. For example, successful scientists are likely to publish more papers, receive more grant funding, and be published in journals with higher impact factors. This means that analyzing them individually may to produce biased estimates. Indeed, the divergence between the univariate results and the multivariate ones will be a theme we return to, because it has important policy implications.

Table 5 shows the differences in means for each factor, both within the group at risk of experimenting and those at risk of converting. Many of these variables affect both experimentation and conversion in the same direction, for example scientists who publish in higher impact factor journals are both more likely to experiment and to convert. Of more interest, are the variables where the effects move in opposite directions, including the amount of experience an author has in years or having a mammalian research focus. More experienced authors are less likely to experiment with CRISPR, but if they do, they are more likely to convert. Conversely, mammalian focused authors are more likely to experiment, but are less likely to convert. There are also many more variables positively associated with experimentation than conversion. For example, getting grants prior to June 2012 is significant at the 0.01% level for experimentation, but is not at all significant for conversion.

[Table 5 here]

We summarize the relative importance of each factor to experimentation and conversion by plotting all the coefficients from the univariate regressions, as described in Section 5, on the same graph (Figure 5). To make these comparable, we present them as incremental multiples of the baseline effect (e.g., a 50% estimate indicates that the factor's marginal effect is equivalent to increasing the base-rate 50%).

[Figure 5 here]

The scatter plot suggests there is a positive correlation between the factors that promote experimentation and conversion. It also suggests that the most impactful variables are being located centrally (near one of the discoverers of CRISPR), the number of papers an author has that are in the top 1% of papers receiving citations, and the type of research focus (e.g., mammalian). Thirdly, most factors seem to be associated with larger effects on experimentation than conversion.

The initial finding of the importance of being in a CRISPR founding city location is particularly interesting as it is consistent with many observations about the clustering of scientists and the importance of implicit knowledge in technology adoption. However, our later analysis will show that there is more to this story than the univariate analysis would suggest.

6.3 Multivariate Results

6.3.1 Importance of Being Centrally Located

Our univariate regressions indicate that a scientist located in Cambridge, MA or Berkeley, CA is much more likely to experiment with CRISPR and to then go on to publish a paper. This is consistent with tacit information theories that suggest new adoption is easier if there is access to someone who uses the tool and can show the new adopter how the it works. In the case of CRISPR, the people who had a strong working knowledge of the CRISPR system early on were the original inventors and those closest to them in Cambridge and Berkeley. But these locations are also areas where high quality scientists reside and where there is abundant research funding. Thus, estimating the effect of a scientists being co-located with a CRISPR inventor requires controlling for these other correlates.

Table 6 shows how the estimates for the effect of being centrally located on experimentation changes when controls are added. Model (1) replicates the univariate analysis, showing that being centrally located is important for experimentation. Adding in the average impact factor of the journals for an author (Model 2) reduces the effect of being centrally located. The central location effect is eliminated completely if we instead include the number of papers that are in the top 1% of citations (Model 3) or include both the impact factor and top citations (Model 4). These suggest that being centrally located does not promote experimentation, but rather that there are just many high quality scientists in Cambridge or Berkeley. Thus, our data provides no evidence that meeting around the water cooler promotes experimentation with CRISPR.

[Table 6 here]

If being at a central location does not improve the likelihood that a scientist experiments, might it still have an impact on conversion (the ability to turn that experimentation into a

publication on CRISPR)? Table 7 repeats the analysis from Table 6, but on conversion rather than experimentation. It shows that the large effect of central location is diminished once you condition on publication quality measures, but that nevertheless a significant impact remains. Experimentation may occur regardless of location for excellent scientists, but it still seems helpful to be near the original inventors when trying to successfully use CRISPR.

[Table 7 here]

6.3.2 Specific Tacit Information

As Table 6 and Table 7 show, the effects of being co-located with discoverers is much smaller than a naïve univariate analysis would suggest, and indeed are only at all impactful on conversion. But might there be subtler effects underlying this? In particular, if we separate out the work being done on bacteria and mammals, do those researchers benefit from being in either Berkeley and Cambridge? (Recall that evidence for CRISPR use in bacteria was first provided in Berkeley and CRISPR use in mammalian cells was first demonstrated in Cambridge). Since it is much more difficult to work with mammalian cells, we might reasonably hypothesize that tacit information would be most useful to mammalian researchers in Cambridge. Table 8 and Table 9 provide support for just that.

As in the previous findings, the effect of central location on experimentation disappears once we control for researcher quality. Table 8, Models 1 and 2 are restricted to only authors with a mammalian research focus. For those authors, being in Cambridge is only significant for experimentation when we do not control for quality. The same holds for authors with a bacteria focus in Models 3 and 4. Interestingly, being in Berkeley does not seem to be significant for CRISPR experimentation, regardless of the research focus considered. Model 5 combines authors from research focuses, but controls explicitly for author quality, each location, research focus, and interactions for being a mammalian author in Cambridge and a bacterial author in Berkeley. Again, we find that evidence that controls for researcher quality make any location effects small and statistically insignificant. Thus, being an excellent researcher is more important than location for the likelihood of experimenting with CRISPR even if you are a mammalian researcher in Cambridge.

[Table 8 here]

In contrast to the lack of location effects for experimentation, we find that, for conversion, being a mammalian researcher in Cambridge has a large effect. In Table 9, Models 1 and 2 show that for mammalian researchers, being in Cambridge remains important for conversion even after controlling for quality. For bacteria focused authors, being in Cambridge makes them less likely to convert, although being in Berkeley may still be helpful as Models 3 and 4 suggest (the effect on

Berkeley is not significant but is large after controlling for author quality). Model 5 shows that the effect of the interaction term for mammal authors in Cambridge is large and significant. So it may be the case that having tacit information is important for converting a new tool into a successful new idea, but there must also be a relevant match in the type of research conducted to reap the largest benefits. For example, because CRISPR is highly in demand for mammalian use but more difficult to implement, mammalian researchers in close proximity to the original inventors of mammalian CRISPR are more likely to benefit from such tacit information than a bacterial researcher in Cambridge. The same effect may be true for bacterial researchers in Berkeley, but it is difficult to provide clear evidence as there are fewer who try CRISPR in the first place. Likely this is because the biological properties of bacteria make the use of CRISPR less critical.

[Table 9 here]

6.3.3 Additional Importance of Resources, Experience, and Research Breadth

Thus far, our analyses have highlighted the role of location, journal impact factor, highly-cited papers, and research focus. But the literature suggests a number of other factors that could be important: available resources, author's years of experience, and research breadth. Table 10 explores these factors for their effects on experimentation, and Table 11 explores them for conversion.

For experimentation, Table 10 Model 1 is the base regression showing the importance of author quality and research focus that overtake the value in being centrally located. The Models 2-8 separately add a new variable to the base model to observe each effect of US university rank, years of experience, the number of Web of Science subjects the author publishes in (a measure of breadth), average R&D expenditures per PI for the researcher's university, the average number of granting agencies an author receives money from, the total number of papers, and the average number of co-authors a focal author has. Model 9 combines all factors into one specification.

We find that, after controlling for the base specification variables, there is little or no effect from the average number of co-authors that a researcher has on their papers or from the host university rank, or the average number of granting agencies listed per paper. We find positive and statistically significant effects for publishing in more subjects, the average R&D expenditures per PI, and the number of papers published by the academic in the previous years. We find a negative effect on experimentation if a researcher has more years of experience.

[Table 10 here]

Table 11 is designed similarly to Table 10 except that the dependent variable is conversion. Here again, the base model coefficients are stable across models. As can be seen in the full specification in Model 9, many fewer of the factors have statistically significant associations with conversion, which is itself interesting. It suggests that policy makers or managers may have fewer potential levers for influencing this component of adoption.

Effects that do have a positive and statistically significant effect on conversion after controlling for the base specification include university rank, the number of subjects that the author publishes in and the number of papers they published in the previous 3 years. There is also a negative and statistically significant effect from having more co-authors per paper on average.

[Table 11 here]

Table 12 compares the multivariate models for adoption (Model 1), experimentation (Model 2), and conversion (Model 3) that include all variables. One of the most striking facts from Table 12 is that different sets of variables describe experimentation and conversion. Not only are fewer factors significant for conversion, but some switch in the direction of effect. For example, R&D expenditures per PI by university is correlated with more CRISPR experimentation by scientists, but has either no correlation or a negative one with producing a new paper (at least in this earliest stage). Additionally, authors with more experience tend to experiment less, but might have no or a slightly positive effect on getting a publication from that experimentation. And, as in our first example, the research focus matters for experimentation and conversion. Mammalian authors experiment more, but likely have a harder time converting, possibly due to the relative difficulty of using CRISPR.

Table 12 also shows how only looking at adoption may not reflect all the levers policy makers or managers have at their disposal or may not provide a clear picture of which types of early adopters will be influenced. For example, raising the amount of available resources to an author by $1 million, holding all else constant, is only associated with an increase in the likelihood of adoption by 0.15%, mostly through the likelihood of experimentation. Thus, affecting adoption through this method might also require outreach targeting the people naturally drawn to experimenting to help them convert.

Figure 6 uses the coefficients from Table 12 to highlight the relative importance of each for adoption, experimentation, and conversion. By column, the highest significant variable is the darkest green, the lowest is the darkest red, and insignificant or small significant variables are white.

[Table 12 and Figure 6 here]

Figure 7 plots the adjusted coefficients for the full multivariate models as a comparison to the original univariate plot (Figure 5). Here it is the case that experimentation is the dimension with more variation and that central location is now only important for conversion. Top citations continues to be important for both experimentation and conversion.

[Figure 7 here]

6.4 Standardized Multivariate Results

The previous analyses presented results in their original units, which helps with intuition since it is easy to visualize writing one more paper or adding another million dollars in R&D funding. However, in the original units it is difficult to compare the effect of one more paper on experimentation to the effect of an extra million dollars in research funding. To make the effect sizes of the factors more directly comparable, we convert each variable into Z-scores, so that the coefficients in the regressions represent the number of standard deviations from each mean. Using the standardized values, the direction and significance levels of the regressions do not change, but the relative effect of each factor does. This can be easily seen in the standardized scatterplots.

Figure 8 shows the standardized coefficients from the multivariate regressions for experimentation and conversion. The total number of papers an author writes in the previous three years before CRISPR is now one of the most important factors for both experimentation and conversion, but some variables, like the research focus of authors is relatively less important as compared to the natural unit multivariate scatterplot in Figure 7.

[Figure 8 here]

We can observe the relative effect of the standardized coefficients in Table 13, which is comparable to Table 12. Converting to standardized coefficients does not change the overarching result that scientists who experiment and those who convert can be characterized by different sets of factors. Table 13 shows explicitly the factors that are relatively more important for experimentation (Model 2) and conversion (Model 3) with standardized coefficients. We also note that if we only had information on adoption, our conclusions as to which factors matter most would be different from either experimentation or conversion (Model 1). To highlight the relative importance of the standardized factors, Figure 9 provides the corresponding heat map to the one presented in Figure 6.

[Table 13 and Figure 9 here]

6.5 Robustness

One concern stemming from our preferred specifications is that our results are being influenced by the fact that scientists ordering in 2014 would not have had enough time to publish in our dataset. Publications in biology occur on a more rapid schedule than other fields like the social sciences, but experiments do take time. To address this concern, we re-ran all of our models defining experimentation as scientists who order CRISPR in 2012 or 2013 only. The full set of regressions are on file with the authors, but in general this restriction does not impact our overall results. The direction and magnitudes remain similar, although we do lose some significance due to our smaller sample size, and hence larger standard errors, in conversion models.[17] Appendix Table A1 provides comparisons for the full experimentation and conversion models. Here, Models 1 and 2 compare our preferred experimentation specification for the full multivariate model to one that defines the dependent variable as CRISPR orders from 2012 and 2013 only. Direction and magnitudes are very similar. Likewise, Models 3 and 4 compare the full multivariate models for conversion where Model 3 is our preferred specification and Model 4 only includes scientists experimenting with CRISPR in 2012 and 2013.

Many of our authors do not have data for every variable we use in the specifications with all controls. To maintain statistical power, we try to keep as many authors in each regression as possible. However, if we restrict our models to only include authors that have data for all variables, the reported results are of the same direction and magnitude. Only the significance level changes due to the reduction in power.

7 Discussion

The slow adoption of socially beneficial technologies occurs frequently enough that the promotion of technological adoption by policy-makers is common. Marketing campaigns are mounted to raise awareness, outreach programs are created to help potential adopters work through technical challenges, and incentives are introduced throughout the system. For example, one of the earliest US programs was the Agricultural Conservation Program (ACP) of the 1930s that reimbursed farmers for adopting technologies that mitigated soil erosion from the Dust Bowl. This was coupled with a federal effort to develop new soil conservation techniques and to spread awareness to local farmers (History of NRCS 2017). More recently, the California Solar Initiative (CSI) was designed to incentivize the adoption of solar energy by businesses and residents in California by providing rebates for installing and using solar power. The program ran from 2007 through 2016 with a total budget of $2.17 billion and a goal of installing 1,940 MW of new solar

[17] The most noticeable change is in the bacteria focus models in Table 9 where being from Berkeley becomes significant and negative. Because that model has very few observations, we suspect this is just a spurious correlation.

capacity (California Solar Initiative 2017). The U.S. Department of Energy also launched the SunShot Initiative in 2011, that works with local public and private partners to encourage solar adoption and to lower the cost of solar energy. SunShot funds projects within communities that focus on outreach, training programs, and R&D (SunShot Solar Projects Map 2017). Since 2011, SunShot has funded almost 300 projects across the U.S. (SunShot Initiative 2017). Likewise, firms roll out specific strategies to encourage customers to test out and continue to use their products. For example, IBM originally relied on a large salesforce to not only sell mainframe computers to skeptical customers but to provide personalized service and maintenance to keep customers using their computers (e.g., Bresnahan et al. 2012).

Such programs affect adoption on multiple levels, but it is not always clear that the intervention chosen is the one that will incentivize the necessary people. As our paper has demonstrated, overall adoption by individuals can be influenced through two important channels: (1) awareness and willingness to experiment with the new technology and (2) ability to convert the experimentation into successful use of the technology. The results of our study provide evidence that the types of individuals successful at each stage of early adoption are generally different. Because of this, policy-makers and managers should assess where in the process adoption is stalling – in experimentation or conversion as targeting the wrong people could lead to wasted funds and under-adoption. For example, in our CRISPR context, targeting generally successful individuals who are primed to adopt and have the demonstrated skill to convert may be an inefficient use of resources since they are correlated with successful adoption.

As noted in Section 5.4, our ability to make causal claims for interventions is limited because our exogenous variation comes from CRISPR itself, not from variation in the underlying characteristics of the scientists. Nevertheless, we can observe trends in the heterogeneous effects that correlate with experimentation and conversion. These could be valuable to policy makers or managers for generating hypotheses about which interventions to try and how much variation in response to expect. For example, we observe much more variation in experimentation than in conversion rates, suggesting that this dimension of adoption may be more easily influenced by external factors or, potentially, by policy makers.

8 Conclusion

This paper has argued that it is important for innovation scholars, policy makers, and firms to separate their analysis of early stage technology adoption into two components: a willingness to experiment with the technology, and the likelihood of converting that experimentation into a successful research output. This distinction is important because it informs how a technology diffuses and what policies or interventions might best guide its adoption. To highlight the importance of decomposing the adoption curve, we construct a unique dataset for the breakthrough gene-editing technology, CRISPR, that allows us to observe both experimentation and conversion. Historically it has been difficult to observe experimentation and conversion separately, but because the CRISPR tool was originally distributed by one main source, Addgene, the authors experimenting and publishing in CRISPR can be reliably tracked.

Our findings confirm the importance of splitting adoption into these two components. Different factors are associated with scientists who experiment with CRISPR in the first three years and those who can go on to publish a successful CRISPR paper. Holding all else constant, academics with many influential papers in multiple research areas prior to the introduction of CRISPR are more likely to both experiment and convert. However, having more resources available is associated with more experimentation but not more conversion. In the reverse direction, experience is correlated with less experimentation but those who experiment may be more likely to publish a paper.

How useful CRISPR is to the research focus of the scientist is also important. The more useful a tool is to a researcher, the more incentive she has to experiment. This hypothesis is borne out in the data and from anecdotal evidence supplied by CRISPR scientists. Mammalian researchers are more likely to experiment with CRISPR than bacteria researchers, but may ultimately have a lower probability of converting. Although this paper does not explore fully the underlying mechanisms of the main model, we find that it could be the case that mammalian researchers have a harder time converting on average due to a lack of specific tacit information in close proximity. For example, if a mammalian author is in Cambridge, MA (where CRISPR for mammalian use was first discovered) she has a much better chance of converting her experimentation into a paper.

Finally, we find that there is much more variation triggered in experimentation than in conversion rates, suggesting that this dimension of adoption may be more easily influenced by external factors or, potentially, by policy makers. Thus, any technology adoption policies should try to identify where bottlenecks are occurring and tailor programs to that part of the adoption process.

9 References

Aghion, P. and P. Howitt. 1992. "A model of growth through creative destruction." *Econometrica* 60(2): 323-351.

Azoulay, P., J. G. Zivin, and G. Manso. 2011. "Incentives and creativity: Evidence from the academic life sciences." *RAND Journal of Economics* 42(3): 527–554.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, and P. Horvath. 2007. "CRISPR provides acquired resistance against viruses in prokaryotes." *Science* 315(5819): 1709–1712.

Bresnahan, T. F., S. Greenstein, and R. M. Henderson. 2012. "Schumpeterian competition and diseconomies of scope: Illustration from the histories of Microsoft and IBM" in *The Rate & Direction of Inventive Activity Revisited*, J. Lerner and S. Stern S eds. University of Chicago Press.

Bresnahan, T. F., and M. Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65(1): 83-108.

Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, and U. Bonas. 2009. "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors." *Science* 326(5959): 1509-1512.

California Solar Initiative. 2017. CSI (October 19), http://www.gosolarcalifornia.ca.gov/about/csi.php.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. 2013. "Multiplex genome engineering using CRISPR/Cas systems." *Science* 339(6121): 819-823.

Doudna, J. 2015. "Genomic engineering and the future of medicine." *JAMA* 313(8):791-792.

Fan, M., J. Tsai, B. Chen, K. Fan, and J. LaBaer. 2005. "A central repository for published plasmids." *Science* 307(5717): 1877.

Furman, J., and S. Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production." *American Economic Review* 101(5): 1933-1963.

Gantz, V., N. Jasinskiene, O. Tatarenkova, A. Fazekas, V. Macias, E. Bier, and A. James. 2015. "Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi." *PNAS* 112(49): E6736–E6743.

Griliches, Z. 1958. "Research Costs and Social Returns: Hybrid Corn and Related Innovations." *Journal of Political Economy* 66(5): 419-431.

Griliches, Z. 1957. "Hybrid Corn: An Exploration in the Economics of Technological Change." *Econometrica* 25(4):501-522.

Hanlon, W. W. 2015. "Necessity is the mother of invention: Input supplies and directed technical change." *Econometrica* 83(1): 61-100.

Heidenreich, M. and F. Zhang. 2016. "Applications of CRISPR–Cas systems in neuroscience." *Nature Reviews Bioscience* 17: 36-44.

History of NRCS. 2017. NRCS (October 19, 2017) available at: https://www.nrcs.usda.gov/wps/portal/nrcs/detail/national/about/history/?cid=nrcs143_021392.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." *Science* 337(6096): 816-21.

Jones, B. F. 2009. "The burden of knowledge and the 'Death of the Renaissance Man': Is innovation getting harder?" *Review of Economic Studies* 76(1): 283-317.

Lazear, E. 2005. "Entrepreneurship." *Journal of Labor Economics* 23(4): 649-680.

Ledford, H. 2015. "CRISPR: The disruptor." *Nature* 522(7554): 20-24.

Ma, H., N. Marti-Gutierrez, S. W. Park, J. Wu, Y. Lee, K. Suzuki, A. Koski, D. Ji, T. Hayama, R. Ahmed, H. Darby, C. Dyken, Y. Li, E. Kang, A. R. Park, D. Kim, S. T. Kim, J. Gong, Y. Gu, X. Xu, D. Battaglia, S. Krieg, D. Lee, D. Wu, D. Wolf, S. Heitner, J. C. Belmonte, P. Amato, J. S. Kim, S. Kaul, and S. Mitalipov. 2017. "Correction of a pathogenic gene mutation in human embryos." *Nature* 548(7668): 413-419.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. "RNA-guided human genome engineering via Cas9." *Science* 339(6121): 823-6.

Manso, G. 2011. "Motivating innovation." *Journal of Finance* 66(5): 1823-1860.

March, J. G. 1991. "Exploration and exploitation in organizational learning." *Organization Science* 2(1): 71-87.

Method. 2012. "Method of the year 2011." *Nature Methods* 9:1.

Moscou, M. and A. J. Bogdanove. 2009. "A simple cipher governs DNA recognition by TAL effectors." *Science* 326(5959): 1501.

Murray, F. 2010. "The Oncomouse that roared: Hybrid exchange strategies as a source of distinction at the boundary of overlapping institutions." *American Journal of Sociology* 116(2): 341-388.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy* 8(1): 212-252.

NSF. 2012a. HERD Survey FY 2011, Data Tables (October 19, 2017) available at: https://nsf.gov/statistics/nsf13325/content.cfm?pub_id=4240&id=2

NSF. 2012b. HERD Survey FY 2011, Technical Notes (October 19, 2017) available at: https://nsf.gov/statistics/nsf13325/content.cfm?pub_id=4240&id=3

Pennisi, E. 2013. "The CRISPR craze." *Science* 341(6148): 833-836.

Platt, R.J., S. Chen, Y. Zhou, M. J. Yim, L. Swiech, H. R. Kempton, J. E. Dahlman, O. Parnas, T. M. Eisenhaure, M. Jovanovic, D. B. Graham, S. Jhunjhunwala, M. Heidenreich, R. J. Xavier, R. Langer, D. G. Anderson, N. Hacohen, A. Regev, G. Feng, P. A. Sharp, and F. Zhang. 2014. "CRISPR-Cas9 knockin mice for genome editing and cancer modeling." *Cell.* 159(2):440-55.

Regalado, A. 2014. "Who owns the biggest biotech discovery of the century?" *MIT Technology Review* (December 4).

Rogers, E. 1995[1962] *Diffusion of Innovations.* 4 ed. The Free Press: Simon & Schuster.

Romer, P. M. 1990. "Endogenous technological change." *Journal of Political Economy* 98(5): S71-S102.

Rosenberg, N. and M. Trajtenberg. 2004. "A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States." *Journal of Economic History* 64(1): 61-99.

Solow, R. M. 1956. "A contribution to the theory of economic growth." *Quarterly Journal of Economics* 70(1): 65–94.

Specter, M. 2015. "The gene hackers: A powerful new technology enables us to manipulate our DNA more easily than ever before." *The New Yorker* (16 November).

SunShot Initiative. 2017. SSI (October 19) available at: https://energy.gov/eere/sunshot/about-sunshot-initiative.

SunShot Solar Projects Map. 2017. SSI (October 19) available at: https://energy.gov/eere/sunshot/sunshot-solar-projects-map.

Teodoridis, F. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64(8): 3469-3970.

Thompson, N. C., P. Pfingstag, A. Kunjapur, and J. Henkel. 2017. "Lightening the burden of knowledge: How tools help expand the frontiers of science." Working paper.

Thompson, N. C., A. Ziedonis, and D. Mowery. 2017. "University licensing and the flow of scientific knowledge." Working paper.

US News & World Report. 2012. Available at: https://www.washingtonpost.com/apps/g/page/local/us-news-college-ranking-trends-2014/1292/.

van der Loo, M. P. J. 2014. "The stringdist package for approximate string matching." *R Journal* 6(1): 111-122. (July 31, 2017) available at: https://cran.r-project.org/web/packages/stringdist/stringdist.pdf.

von Hippel, E., J. De Jong, and S. Flowers. 2012. "Comparing business and household sector innovation in consumer products: Findings from a representative study in the United Kingdom." *Management Science* 58(9): 1669-1681.

von Hippel, E., S. Ogawa, J. and De Jong. 2011. "The Age of the consumer innovator." *MIT Sloan Management Review* 53(1): 27-35.

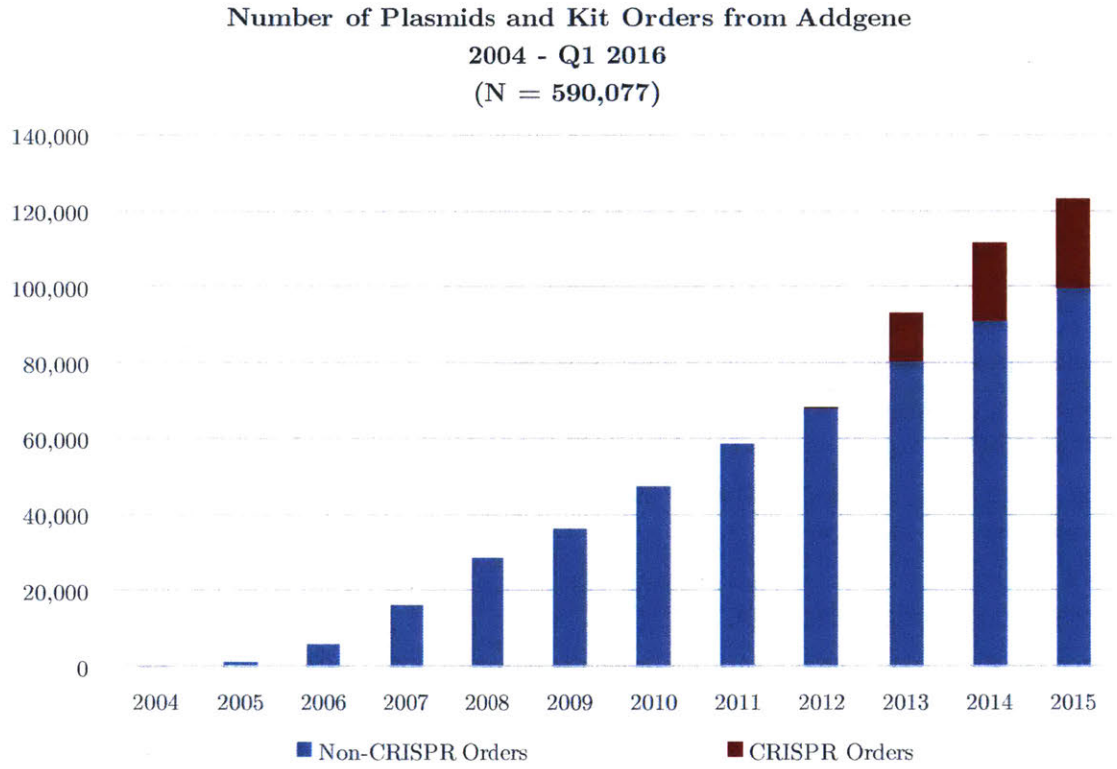
Walsh, J. P., A. Arora, and W. M. Cohen. 2003. "Effects of research tool patents and licensing on biomedical innovation" in *Patents in the Knowledge-Based Economy*, S.A. Merrill and W.M. Cohen, eds. The National Academies Press.

Walsh, J. P., W. M. Cohen, and C. Cho. 2007. "Where excludability matters: Material versus intellectual property in academic biomedical research." *Research Policy* 36(8):1184–1203.

Wang, Y., X. Cheng, Q. Shan, Y. Zhang, J. Liu, C. Gao, and J. L. Qiu. 2014. "Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew." *Nature Biotechnology* 32(9): 947–951.

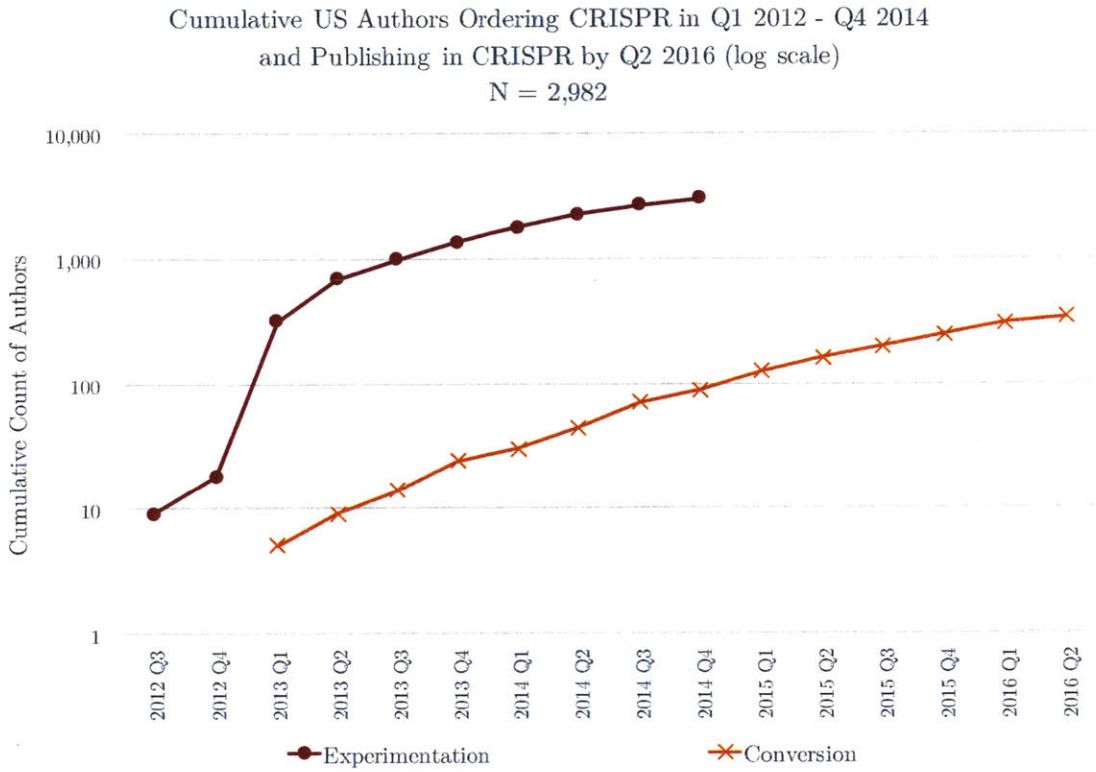
10 Figures and Tables

Figure 1. Addgene Plasmid and Kit Orders by Year



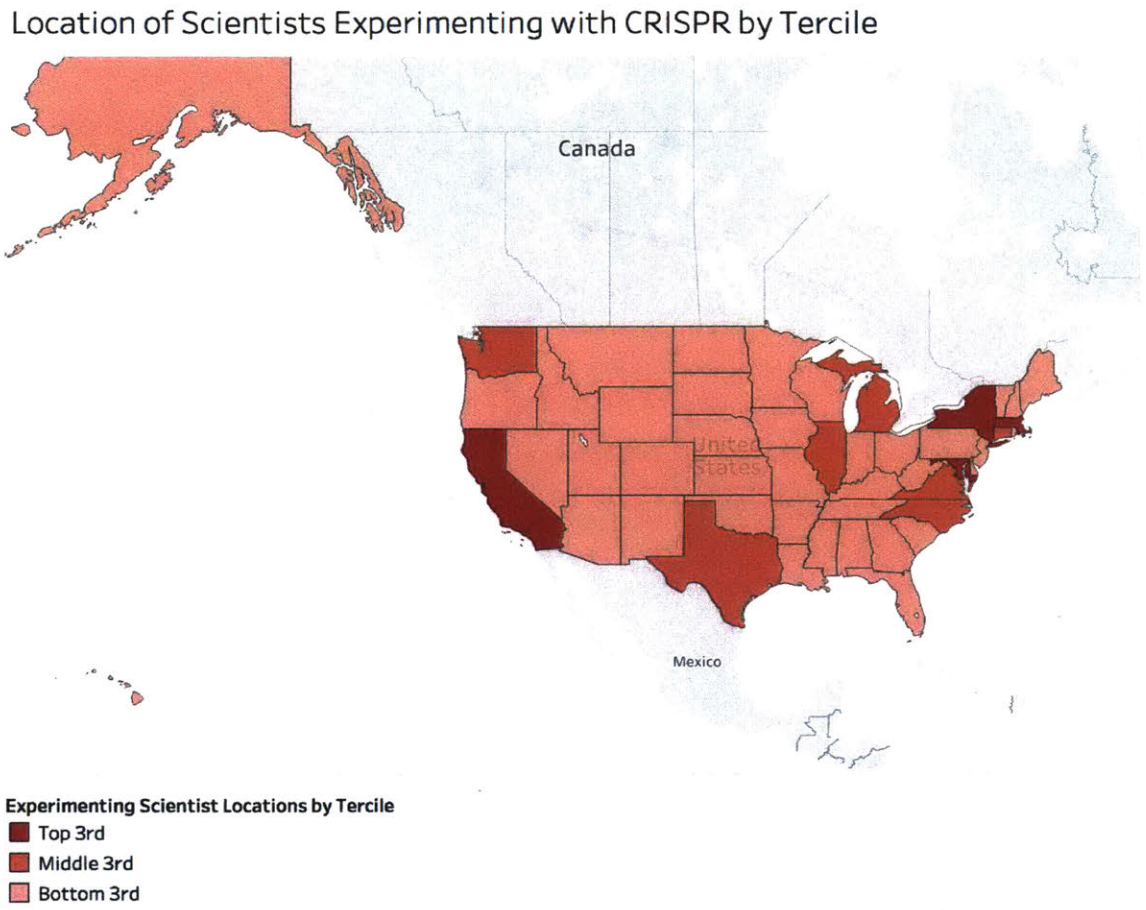
Notes. This graph shows the number of individual plasmids and sets of plasmids (kits) that Addgene sold per year from its start in 2004 through 2015. The blue bars are orders for non-CRISPR plasmids and the red bars are orders for CRISPR plasmids.

Figure 2. Experimentation and Conversion in CRISPR over Time



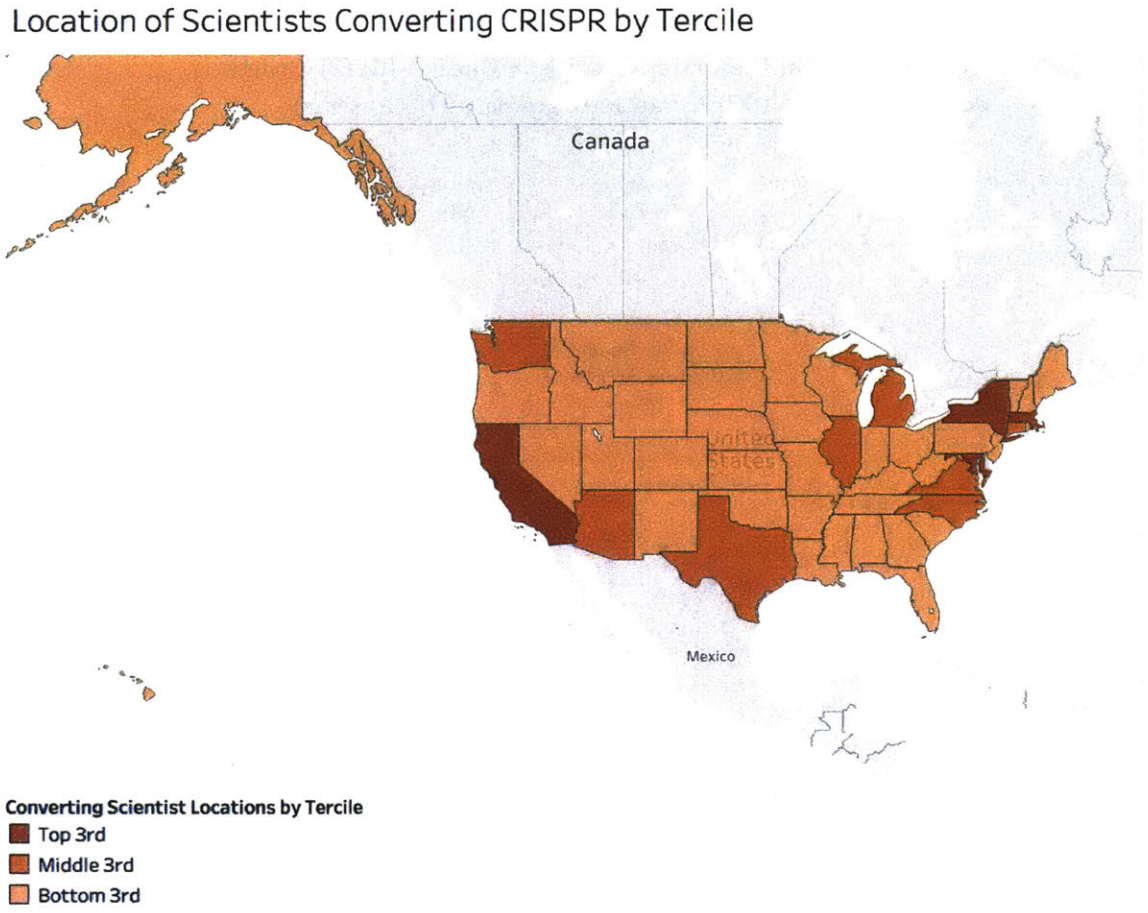
Notes. This graph is built from data in the CRISPR Experimentation Dataset and in Web of Science. The top red line represents the cumulative number of researchers who experimented with CRISPR (ordered CRISPR in the CRISPR Experimentation Dataset) from Q3 2012 – Q4 2014. The bottom orange line represents the cumulative number of researchers who converted (experimented and then went on to publish a paper in CRISPR) from Q3 2012 – Q2 2016.

Figure 3. Locations of Scientists Experimenting with CRISPR



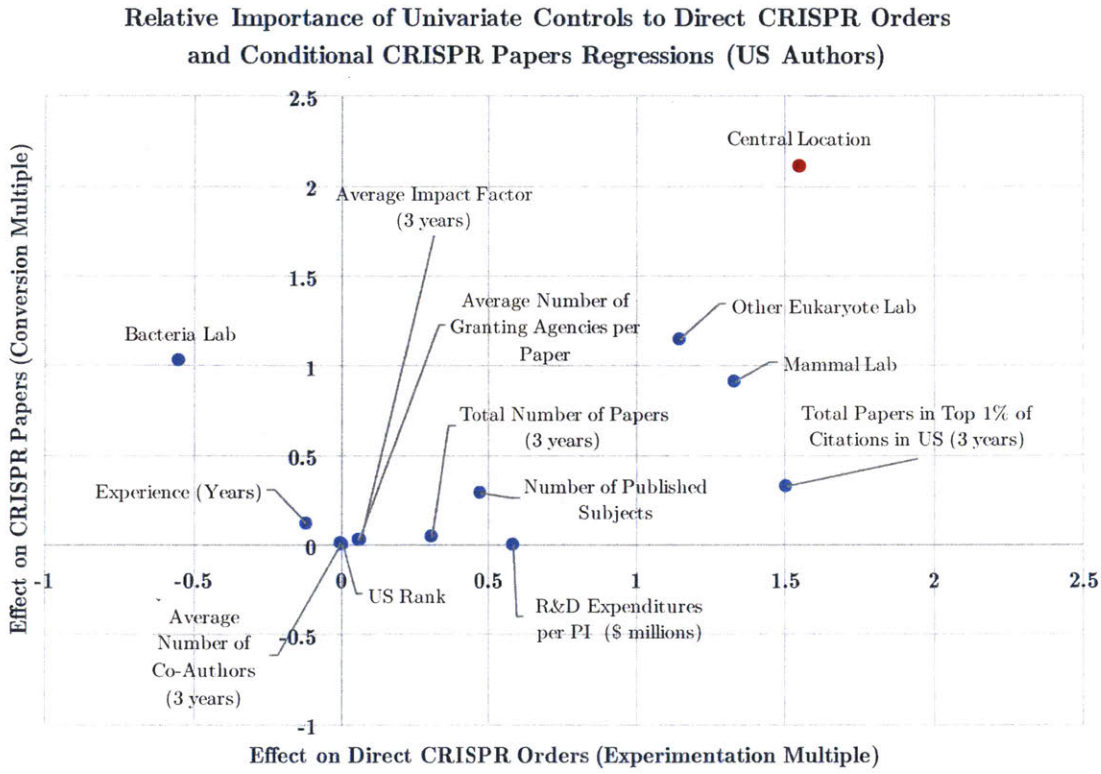
Notes. This figure uses the CRISPR Experimentation Dataset to map the states of CRISPR orderers. The coloring partitions these states by the cumulative number of orderers represented by those states.

Figure 4. Locations of Scientists Converting CRISPR



Notes. This figure uses the Web of Science Dataset to map the states of researchers converting CRISPR (those who experimented and then went on to publish a paper in CRISPR). The coloring partitions these states by the cumulative number of converters represented by those states.

Figure 5. Scatter Plot Illustrating the Most Important Univariate Variables for Experimentation and Conversion



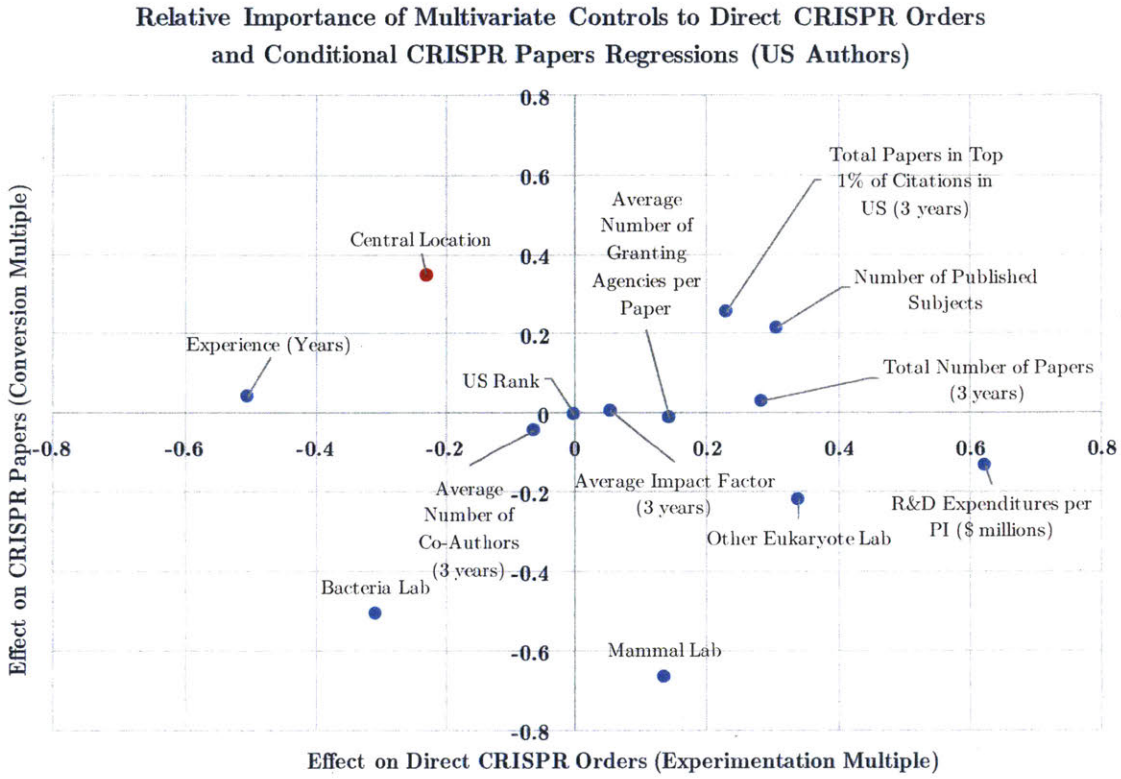
Notes. This graph represents the relative importance of each researcher characteristic, considered on its own, to experimentation and conversion. Each point is built by two underlying linear probability models where the characteristic (e.g., Experience) is the independent variable and the dependent variables are Experimentation (ordered CRISPR in the CRISPR Experimentation Dataset) and Conversion (experimented and then went on to publish a paper in CRISPR) respectively. To make the coefficients comparable, they are presented as incremental multiples of the baseline effect (e.g., a 50% estimate indicates that the factor's marginal effect is equivalent to increasing the base-rate 50%). Each characteristic's effect on Experimentation is measured on the horizontal access and its effect on Conversion is measured on the vertical axis.

Figure 6. Heatmap of Significant Coefficients for Adoption, Experimentation, and Conversion

	Adoption (Published in CRISPR)	Experimentation (Ordered CRISPR)	Conversion (Published | Order)
Central Location	0.0011	-0.0041	0.0396
Average Impact Factor (3 years)	0.0001	0.0010	0.0007
Total Papers in Top 1% of Citations in US (3 years)	0.0043	0.0041	0.0286
Mammal Lab	-0.0012	0.0024	-0.0753
Other Eukaryote Lab	0.0006	0.0061	-0.0250
Bacteria Lab	-0.0018	-0.0056	-0.0573
US Rank	0.0000	0.0000	-0.0002
Experience (Years)	-0.0012	-0.0091	0.0047
Number of Published Subjects	0.0004	0.0055	0.0242
R&D Expenditures per PI (Millions)	0.0015	0.0112	-0.0151
Average Number of Granting Agencies per Paper	0.0000	0.0026	-0.0011
Total Number of Papers (3 years)	0.0013	0.0051	0.0032
Average Number of Co-Authors (3 years)	-0.0002	-0.0011	-0.0048

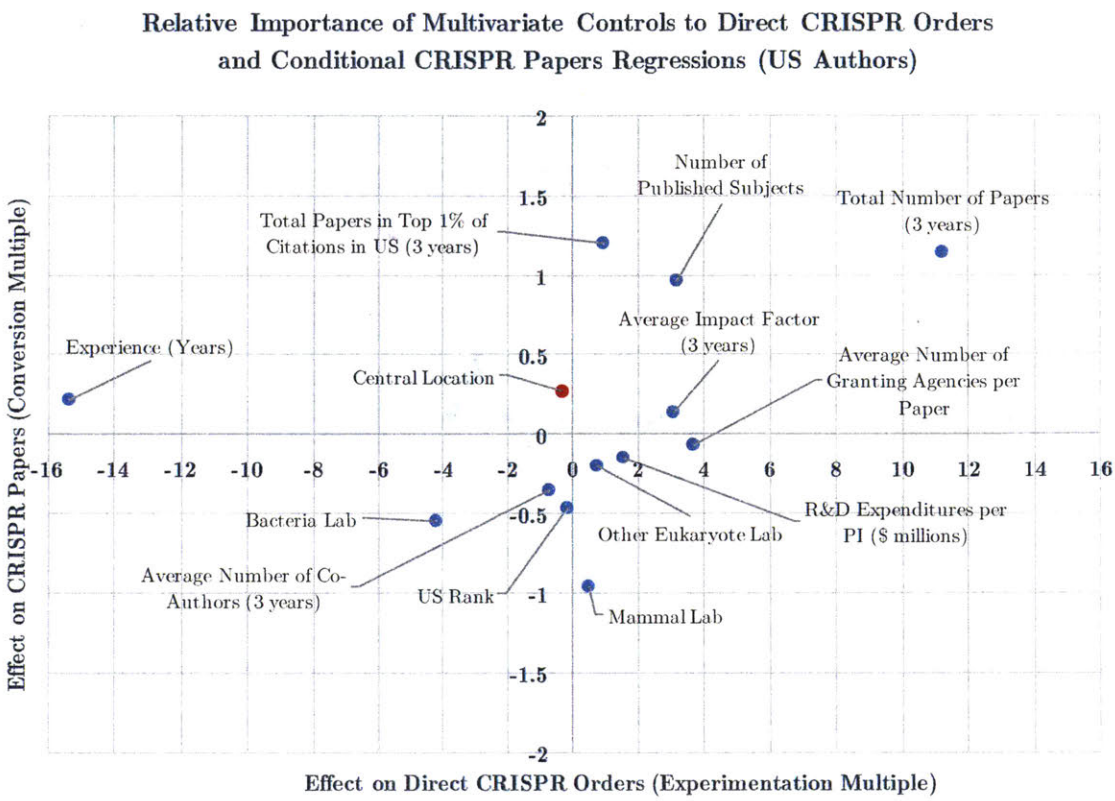
Notes. This graph highlights the differences in significant researcher characteristics for the multivariate linear probability models on Adoption (published a paper in CRISPR), Experimentation (ordered CRISPR in the CRISPR Experimentation Dataset) and Conversion (experimented and then went on to publish a paper in CRISPR) in Table 12. By column, the darkest green is the largest positive significant researcher characteristic and the darkest red is the largest negative significant researcher characteristic. Other colored characteristics are significant at the 10% level. White characteristics are insignificant.

Figure 7. Scatter Plot Illustrating the Most Important Multivariate Variables for Experimentation and Conversion



Notes. This graph represents the relative importance of each researcher characteristic, considered together, to experimentation and conversion. Each point represents the coefficient of the characteristic in two underlying multivariate linear probability models where the dependent variables are Experimentation (ordered CRISPR in the CRISPR Experimentation Dataset) and Conversion (experimented and then went on to publish a paper in CRISPR) respectively (both in Table 12). To make the coefficients comparable, they are presented as incremental multiples of the baseline effect (e.g., a 50% estimate indicates that the factor's marginal effect is equivalent to increasing the base-rate 50%). Each characteristic's effect on Experimentation is measured on the horizontal access and its effect on Conversion is measured on the vertical axis.

Figure 8. Scatter Plot Illustrating the Most Important Multivariate Variables for Experimentation and Conversion with Standard Units



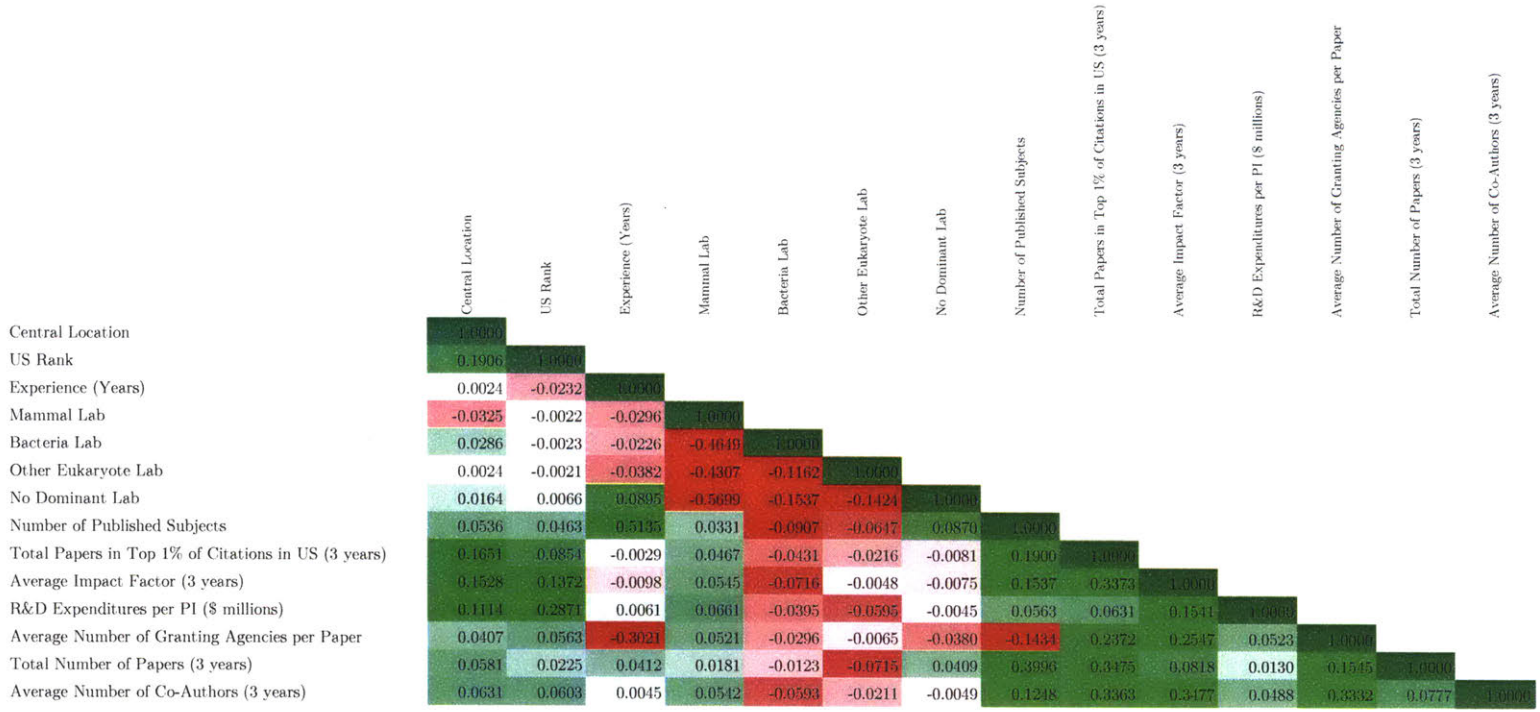
Notes. This graph represents the relative importance of each researcher characteristic, considered together and standardized, to experimentation and conversion. Each point represents the coefficient of the characteristic after being converted to Z-scores, in two underlying multivariate linear probability models where the dependent variables are Experimentation (ordered CRISPR in the CRISPR Experimentation Dataset) and Conversion (experimented and then went on to publish a paper in CRISPR) respectively (both in Table 13). To make the coefficients comparable, they are presented as incremental multiples of the baseline effect (e.g., a 50% estimate indicates that the factor's marginal effect is equivalent to increasing the base-rate 50%). Each characteristic's effect on Experimentation is measured on the horizontal access and its effect on Conversion is measured on the vertical axis.

Figure 9. Heatmap of Significant Coefficients for Adoption, Experimentation, and Conversion (Standard Units)

	Adoption (Published in CRISPR)	Experimentation (Ordered CRISPR)	Conversion (Published | Order)
Central Location	0.0046	-0.0061	0.0304
Average Impact Factor (3 years)	0.0168	0.0556	0.0156
Total Papers in Top 1% of Citations in US (3 years)	0.0526	0.0173	0.1354
Mammal Lab	-0.0125	0.0088	-0.1084
Other Eukaryote Lab	0.0042	0.0136	-0.0229
Bacteria Lab	-0.0127	-0.0131	-0.0406
US Rank	-0.0094	-0.0033	-0.0535
Experience (Years)	-0.1099	-0.2773	0.0246
Number of Published Subjects	0.0131	0.0573	0.1095
R&D Expenditures per PI (Millions)	0.0112	0.0276	-0.0174
Average Number of Granting Agencies per Paper	-0.0036	0.0663	-0.0076
Total Number of Papers (3 years)	0.1502	0.2030	0.1287
Average Number of Co-Authors (3 years)	-0.0479	-0.0762	-0.0619

Notes. This graph highlights the differences in significant researcher characteristics for the multivariate linear probability models on Adoption (published a paper in CRISPR), Experimentation (ordered CRISPR in the CRISPR Experimentation Dataset) and Conversion (experimented and then went on to publish a paper in CRISPR) in Table 13. All characteristics in these models were converted to Z-scores. By column, the darkest green is the largest positive significant researcher characteristic and the darkest red is the largest negative significant researcher characteristic. Other colored characteristics are significant at the 10% level. White characteristics are insignificant.

Table 1. Pairwise Correlations Between Control Factors



	Central Location	US Rank	Experience (Years)	Mammal Lab	Bacteria Lab	Other Eukaryote Lab	No Dominant Lab	Number of Published Subjects	Total Papers in Top 1% of Citations in US (3 years)	Average Impact Factor (3 years)	R&D Expenditures per PI ($ millions)	Average Number of Granting Agencies per Paper	Total Number of Papers (3 years)	Average Number of Co-Authors (3 years)
Central Location	1.0000													
US Rank	0.1906	1.0000												
Experience (Years)	0.0024	-0.0232	1.0000											
Mammal Lab	-0.0325	-0.0022	-0.0296	1.0000										
Bacteria Lab	0.0286	-0.0023	-0.0226	-0.4649	1.0000									
Other Eukaryote Lab	0.0024	-0.0021	-0.0382	-0.4307	-0.1162	1.0000								
No Dominant Lab	0.0164	0.0066	0.0895	-0.5699	-0.1537	-0.1424	1.0000							
Number of Published Subjects	0.0536	0.0463	0.5135	0.0331	-0.0907	-0.0617	0.0870	1.0000						
Total Papers in Top 1% of Citations in US (3 years)	0.1651	0.0854	-0.0029	0.0467	-0.0431	-0.0216	-0.0081	0.1900	1.0000					
Average Impact Factor (3 years)	0.1528	0.1372	-0.0098	0.0545	-0.0716	-0.0048	-0.0075	0.1537	0.3373	1.0000				
R&D Expenditures per PI ($ millions)	0.1114	0.2871	0.0061	0.0661	-0.0395	-0.0595	-0.0045	0.0563	0.0631	0.1541	1.0000			
Average Number of Granting Agencies per Paper	0.0407	0.0563	-0.3021	0.0521	-0.0296	-0.0065	-0.0380	-0.1434	0.2372	0.2547	0.0523	1.0000		
Total Number of Papers (3 years)	0.0581	0.0225	0.0412	0.0181	-0.0123	-0.0715	0.0409	0.3996	0.3475	0.0818	0.0130	0.1545	1.0000	
Average Number of Co-Authors (3 years)	0.0631	0.0603	0.0045	0.0542	-0.0593	-0.0211	-0.0049	0.1248	0.3363	0.3477	0.0488	0.3332	0.0777	1.0000

Table 2. Summary Statistics for Authors in the At-risk Set

Variable	N	Mean	Std. Dev.	Min	Max
Outcomes					
Experimentation (Ordered CRISPR)	164,993	0.018	0.133	0.000	1.000
Conversion (Published | Order)	164,993	0.002	0.049	0.000	1.000
Location					
Central Location	164,993	0.041	0.198	0.000	1.000
Availability of Resources					
US Rank	137,566	-80.363	75.157	-200.00	-1.000
Average Number of Granting Agencies per Paper	164,993	1.723	3.389	0.000	148.000
R&D Expenditures per PI ($ millions)	125,929	0.586	0.328	0.010	2.782
Ability to Apply Technological Knowledge					
Mammal Lab	127,022	0.633	0.482	0.000	1.000
Bacteria Lab	127,022	0.111	0.315	0.000	1.000
Other Eukaryote Lab	127,022	0.097	0.296	0.000	1.000
No Dominant Lab	127,022	0.159	0.365	0.000	1.000
Total Number of Papers (3 years)	103,004	4.237	5.310	0.000	141.000
Experience (Years)	164,993	5.673	4.038	0.000	12.000
Opinion Leadership					
Average Impact Factor (3 years)	100,245	7.416	7.543	0.000	44.026
Social Participation					
Average Number of Co-Authors (3 years)	103,004	4.505	9.199	0.000	283.000
Novelty					
Total Papers in Top 1% of Citations in US (3 years)	103,004	0.114	0.557	0.000	25.000
Breadth of Research Field					
Number of Published Subjects	164,993	2.581	1.379	1.000	6.000

Table 3. Summary Statistics for Authors that Ordered CRISPR (Experimenters)

Variable	N	Mean	Std. Dev.	Min	Max
Outcomes					
Conversion (Published | Order)	2,982	0.113	0.317	0.000	1.000
Location					
Central Location	2,982	0.063	0.243	0.000	1.000
Availability of Resources					
US Rank	2,365	-69.328	75.533	-200.000	-1.000
Average Number of Granting Agencies per Paper	2,982	2.407	2.153	0.000	36.667
R&D Expenditures per PI ($ millions)	2,273	0.649	0.366	0.031	1.959
Ability to Apply Technological Knowledge					
Mammal Lab	2,742	0.705	0.456	0.000	1.000
Bacteria Lab	2,742	0.053	0.225	0.000	1.000
Other Eukaryote Lab	2,742	0.093	0.290	0.000	1.000
No Dominant Lab	2,742	0.148	0.356	0.000	1.000
Total Number of Papers (3 years)	1,997	12.259	12.882	0.000	141.000
Experience (Years)	2,982	3.745	1.675	0.000	12.000
Opinion Leadership					
Average Impact Factor (3 years)	1,986	10.883	6.800	0.000	38.159
Social Participation					
Average Number of Co-Authors (3 years)	1,997	4.431	4.059	0.000	53.286
Novelty					
Total Papers in Top 1% of Citations in US (3 years)	1,997	0.548	1.500	0.000	25.000
Breadth of Research Field					
Number of Published Subjects	2,982	3.487	1.432	1.000	6.000

Table 4. An Example of Obscured Information When Only Observing Successful Adoption

Author's Lab Type (Bacteria, Mammal, Other Eukaryote) for Adoption, Experimentation, and Conversion

	(1) Adoption (Published in CRISPR)	(2) Experimentation (Ordered CRISPR)	(3) Conversion (Published | Order)
Mammal Lab	0.0013***	0.0137***	-0.0130
	(0.0003)	(0.0010)	(0.0275)
Other Eukaryote Lab	0.0015**	0.0103***	0.0130
	(0.0005)	(0.0015)	(0.0339)
No Dominant Lab Category	0.0022***	0.0099***	0.0531
	(0.0005)	(0.0013)	(0.0324)
Constant	0.0012***	0.0103***	0.1164***
	(0.0003)	(0.0008)	(0.0266)
Observations	127,022	127,022	2,742
Adjusted R2	0.0001	0.0008	0.0043

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is Bacteria Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5. Tests of Differences of Means (Unequal Variances) for Authors Who Experiment and Do Not and for Authors Who Convert and Do Not

	Mean Don't Experiment	Mean Experiment	Diff. Experiment	P-value	Mean Don't Convert	Mean Convert	Diff. Convert	P-value
Central Location	0.04	0.06	0.02 ***	0.00	0.05	0.13	0.08 ***	0.00
US Rank	-80.56	-69.33	11.23 ***	0.00	-68.69	-74.35	-5.66	0.27
Experience (Years)	5.71	3.75	-1.96 ***	0.00	3.70	4.08	0.38 ***	0.00
Mammal Lab	0.63	0.71	0.08 ***	0.00	0.72	0.63	-0.09 ***	0.00
Bacteria Lab	0.11	0.05	-0.06 ***	0.00	0.05	0.05	0.00	1.00
Other Eukaryote Lab	0.10	0.09	-0.01	0.45	0.09	0.10	0.01	0.51
No Dominant Lab	0.16	0.15	-0.01	0.13	0.14	0.22	0.08 ***	0.00
Number of Published Subjects	2.56	3.49	0.93 ***	0.00	3.41	4.07	0.66 ***	0.00
Total Papers in Top 1% of Citations in US (3 years)	0.11	0.55	0.44 ***	0.00	0.45	1.19	0.74 ***	0.00
Average Impact Factor (3 years)	7.35	10.88	3.53 ***	0.00	10.70	12.10	1.40 ***	0.00
R&D Expenditures per PI ($ millions)	0.59	0.65	0.06 ***	0.00	0.65	0.65	0.00	0.65
Average Number of Granting Agencies per Paper	1.71	2.41	0.70 ***	0.00	2.39	2.53	0.14	0.32
Total Number of Papers (3 years)	4.08	12.26	8.18 ***	0.00	11.31	18.63	7.32 ***	0.00
Average Number of Co-Authors (3 years)	4.51	4.43	-0.08	0.43	4.41	4.58	0.17	0.47

$^{+}p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 6. The Effects of Location and Author Quality on Experimentation

Author Location and Quality Effects on Experimentation

	(1) Experimentation (Ordered CRISPR)	(2) Experimentation (Ordered CRISPR)	(3) Experimentation (Ordered CRISPR)	(4) Experimentation (Ordered CRISPR)
Central Location	0.0103***	0.0075**	0.0021	-0.0002
	(0.0020)	(0.0028)	(0.0026)	(0.0027)
Average Impact Factor (3 years)		0.0012***		0.0006***
		(0.0001)		(0.0001)
Total Papers in Top 1% of Citations in US (3 years)			0.0270***	0.0244***
			(0.0020)	(0.0021)
Constant	0.0177***	0.0108***	0.0162***	0.0124***
	(0.0003)	(0.0005)	(0.0004)	(0.0005)
Observations	164,993	100,245	103,004	100,245
Adjusted R2	0.0002	0.0044	0.0120	0.0130

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 7. The Effects of Location and Author Quality on Conversion

Author Location and Quality Effects on Conversion

	(1) Conversion (Published | Order)	(2) Conversion (Published | Order)	(3) Conversion (Published | Order)	(4) Conversion (Published | Order)
Central Location	0.1349***	0.1164**	0.0745*	0.0744*
	(0.0317)	(0.0360)	(0.0348)	(0.0347)
Average Impact Factor (3 years)		0.0026*		0.0000
		(0.0011)		(0.0011)
Total Papers in Top 1% of Citations in US (3 years)			0.0335***	0.0334***
			(0.0065)	(0.0069)
Constant	0.1045***	0.0921***	0.1054***	0.1051***
	(0.0058)	(0.0137)	(0.0078)	(0.0136)
Observations	2,982	1,986	1,997	1,986
Adjusted R2	0.0104	0.0117	0.0296	0.0291

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8. The Importance of Being a Mammalian Researcher in Cambridge for Experimentation

Specific Tacit Information Effects on Experimentation

	(1) Mammal Focus Experimentation (Ordered CRISPR)	(2) Mammal Focus Experimentation (Ordered CRISPR)	(3) Bacteria Focus Experimentation (Ordered CRISPR)	(4) Bacteria Focus Experimentation (Ordered CRISPR)	(5) All Authors Experimentation (Ordered CRISPR)
Cambridge	0.0193***	-0.0038	0.0214*	0.0064	0.0074
	(0.0043)	(0.0051)	(0.0085)	(0.0082)	(0.0060)
Berkeley	-0.0001	-0.0016	0.0116	0.0010	-0.0005
	(0.0056)	(0.0074)	(0.0072)	(0.0082)	(0.0051)
Average Impact Factor (3 years)		0.0008***		0.0006**	0.0008***
		(0.0001)		(0.0002)	(0.0001)
Total Papers in Top 1% of Citations in US (3 years)		0.0253***		0.0252**	0.0238***
		(0.0025)		(0.0097)	(0.0021)
Mammal Lab					0.0046***
					(0.0013)
Bacteria Lab					-0.0075***
					(0.0015)
Other Eukaryote Lab					0.0015
					(0.0019)
Cambridge*Mammal Lab					-0.0098
					(0.0078)
Berkeley*Bacteria Lab					0.0013
					(0.0097)
Constant	0.0235***	0.0147***	0.0093***	0.0040***	0.0104***
	(0.0005)	(0.0007)	(0.0008)	(0.0011)	(0.0011)
Observations	80,386	55,486	14,155	9,433	88,588
Adjusted R2	0.0004	0.0163	0.0015	0.0087	0.0145

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab in Model 5

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9. The Importance of Being a Mammalian Researcher in Cambridge for Conversion

Specific Tacit Information Effects on Conversion

	(1) Mammal Focus Conversion (Published | Order)	(2) Mammal Focus Conversion (Published | Order)	(3) Bacteria Focus Conversion (Published | Order)	(4) Bacteria Focus Conversion (Published | Order)	(5) All Authors Conversion (Published | Order)
Cambridge	0.2531***	0.1566**	0.0490	-0.1435*	-0.1043*
	(0.0490)	(0.0541)	(0.1049)	(0.0688)	(0.0511)
Berkeley	0.0205	0.0135	0.1174	0.2273	0.0203
	(0.0744)	(0.0801)	(0.1428)	(0.1913)	(0.0648)
Average Impact Factor (3 years)		-0.0007		-0.0079*	-0.0004
		(0.0012)		(0.0036)	(0.0012)
Total Papers in Top 1% of Citations in US (3 years)		0.0314***		0.0655	0.0318***
		(0.0074)		(0.0543)	(0.0070)
Mammal Lab					-0.0790**
					(0.0240)
Bacteria Lab					-0.0607
					(0.0409)
Other Eukaryote Lab					-0.0178
					(0.0362)
Cambridge*Mammal Lab					0.2576***
					(0.0738)
Berkeley*Bacteria Lab					0.2389
					(0.2152)
Constant	0.0907***	0.0979***	0.1048***	0.1787**	0.1731***
	(0.0067)	(0.0154)	(0.0278)	(0.0608)	(0.0252)
Observations	1,934	1,374	146	87	1,923
Adjusted R2	0.0316	0.0481	-0.0048	0.0349	0.0385

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014.; Dropped is No Dominant Lab in Model 5

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10. The Effects of Resources, Experience, and Breadth on Experimentation

Resource, Experience, and Breadth Effects on Experimentation

	(1) Experimentation (Ordered CRISPR)	(2) Experimentation (Ordered CRISPR)	(3) Experimentation (Ordered CRISPR)	(4) Experimentation (Ordered CRISPR)
Central Location	0.0010	-0.0005	-0.0005	-0.0003
	(0.0030)	(0.0032)	(0.0030)	(0.0030)
Average Impact Factor (3 years)	0.0008***	0.0008***	0.0008***	0.0007***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Total Papers in Top 1% of Citations in US (3 years)	0.0237***	0.0193***	0.0237***	0.0214***
	(0.0021)	(0.0022)	(0.0021)	(0.0021)
Mammal Lab	0.0044***	0.0026+	0.0008	0.0053***
	(0.0013)	(0.0014)	(0.0012)	(0.0013)
Other Eukaryote Lab	0.0014	0.0003	-0.0035+	0.0040*
	(0.0019)	(0.0020)	(0.0019)	(0.0019)
Bacteria Lab	-0.0075***	-0.0073***	-0.0106***	-0.0042**
	(0.0015)	(0.0016)	(0.0015)	(0.0015)
US Rank		0.0000***		
		(0.0000)		
Experience (Years)			-0.0088***	
			(0.0002)	
Number of Published Subjects				0.0066***
				(0.0004)
Constant	0.0105***	0.0139***	0.0870***	-0.0113***
	(0.0011)	(0.0014)	(0.0023)	(0.0017)
Observations	88,588	73,391	88,588	88,588
Adjusted R2	0.0145	0.0108	0.0478	0.0179

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.00$

Resource, Experience, and Breadth Effects on Experimentation (cont.)

	(5) Experimentation (Ordered CRISPR)	(6) Experimentation (Ordered CRISPR)	(7) Experimentation (Ordered CRISPR)	(8) Experimentation (Ordered CRISPR)	(9) Experimentation (Ordered CRISPR)
Central Location	-0.0005	0.0005	0.0004	0.0002	-0.0041
	(0.0033)	(0.0030)	(0.0029)	(0.0030)	(0.0031)
Average Impact Factor (3 years)	0.0008***	0.0004***	0.0010***	0.0011***	0.0010***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Total Papers in Top 1% of Citations in US (3 years)	0.0211***	0.0197***	0.0055**	0.0274***	0.0041+
	(0.0023)	(0.0021)	(0.0021)	(0.0023)	(0.0023)
Mammal Lab	0.0035*	0.0034**	0.0071***	0.0046***	0.0024+
	(0.0015)	(0.0013)	(0.0013)	(0.0013)	(0.0014)
Other Eukaryote Lab	0.0011	0.0018	0.0103***	0.0010	0.0061**
	(0.0022)	(0.0019)	(0.0019)	(0.0019)	(0.0021)
Bacteria Lab	-0.0068***	-0.0073***	-0.0046**	-0.0082***	-0.0056**
	(0.0018)	(0.0015)	(0.0015)	(0.0015)	(0.0017)
US Rank					-0.0000
					(0.0000)
Experience (Years)					-0.0091***
					(0.0003)
Number of Published Subjects					0.0055***
					(0.0005)
R&D Expenditures per PI (Millions)	0.0091***				0.0112***
	(0.0020)				(0.0021)
Average Number of Granting Agencies per Paper		0.0106***			0.0026***
		(0.0007)			(0.0007)
Total Number of Papers (3 years)			0.0054***		0.0051***
			(0.0002)		(0.0002)
Average Number of Co-Authors (3 years)				-0.0010***	-0.0011***
				(0.0001)	(0.0001)
Constant	0.0063***	0.0019	-0.0163***	0.0124***	0.0408***
	(0.0017)	(0.0012)	(0.0014)	(0.0011)	(0.0029)
Observations	66,975	88,588	88,588	88,588	64,709
Adjusted R2	0.0130	0.0227	0.0513	0.0175	0.0894

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11. The Effects of Resources, Experience, and Breadth on Conversion

Resource, Experience, and Breadth Effects on Conversion

	(1) Conversion (Published | Order)	(2) Conversion (Published | Order)	(3) Conversion (Published | Order)	(4) Conversion (Published | Order)
Central Location	0.0698*	0.0647+	0.0685+	0.0623+
	(0.0355)	(0.0373)	(0.0356)	(0.0351)
Average Impact Factor (3 years)	-0.0003	-0.0001	-0.0002	-0.0007
	(0.0012)	(0.0013)	(0.0012)	(0.0012)
Total Papers in Top 1% of Citations in US (3 years)	0.0353***	0.0393***	0.0348***	0.0288***
	(0.0070)	(0.0078)	(0.0070)	(0.0070)
Mammal Lab	-0.0649**	-0.0843**	-0.0650**	-0.0635**
	(0.0236)	(0.0265)	(0.0236)	(0.0234)
Other Eukaryote Lab	-0.0150	-0.0504	-0.0142	-0.0054
	(0.0361)	(0.0383)	(0.0361)	(0.0359)
Bacteria Lab	-0.0504	-0.0638	-0.0509	-0.0447
	(0.0416)	(0.0447)	(0.0415)	(0.0418)
US Rank		-0.0003*		
		(0.0001)		
Experience (Years)			0.0053	
			(0.0066)	
Number of Published Subjects				0.0351***
				(0.0067)
Constant	0.1596***	0.1581***	0.1352***	0.0256
	(0.0249)	(0.0299)	(0.0382)	(0.0338)
Observations	1,923	1,523	1,923	1,923
Adjusted R2	0.0329	0.0351	0.0328	0.0480

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Resource, Experience, and Breadth Effects on Conversion (cont.)

	(5) Conversion (Published | Order)	(6) Conversion (Published | Order)	(7) Conversion (Published | Order)	(8) Conversion (Published | Order)	(9) Conversion (Published | Order)
Central Location	0.0520	0.0702*	0.0562	0.0686+	0.0396
	(0.0361)	(0.0355)	(0.0351)	(0.0354)	(0.0375)
Average Impact Factor (3 years)	-0.0002	0.0000	0.0007	0.0005	0.0007
	(0.0013)	(0.0012)	(0.0012)	(0.0012)	(0.0014)
Total Papers in Top 1% of Citations in US (3 years)	0.0424***	0.0361***	0.0215**	0.0401***	0.0286***
	(0.0076)	(0.0069)	(0.0076)	(0.0069)	(0.0082)
Mammal Lab	-0.0868**	-0.0643**	-0.0573*	-0.0612**	-0.0753**
	(0.0274)	(0.0236)	(0.0232)	(0.0236)	(0.0277)
Other Eukaryote Lab	-0.0447	-0.0146	0.0068	-0.0161	-0.0250
	(0.0404)	(0.0361)	(0.0359)	(0.0360)	(0.0405)
Bacteria Lab	-0.0664	-0.0516	-0.0358	-0.0538	-0.0573
	(0.0457)	(0.0417)	(0.0411)	(0.0418)	(0.0465)
US Rank					-0.0002+
					(0.0001)
Experience (Years)					0.0047
					(0.0085)
Number of Published Subjects					0.0242*
					(0.0094)
R&D Expenditures per PI (Millions)	-0.0126				-0.0151
	(0.0218)				(0.0220)
Average Number of Granting Agencies per Paper		-0.0053			-0.0011
		(0.0043)			(0.0049)
Total Number of Papers (3 years)			0.0038***		0.0032**
			(0.0008)		(0.0010)
Average Number of Co-Authors (3 years)				-0.0054***	-0.0048*
				(0.0016)	(0.0019)
Constant	0.1836***	0.1676***	0.1017***	0.1712***	0.0261
	(0.0310)	(0.0259)	(0.0260)	(0.0251)	(0.0559)
Observations	1,475	1,923	1,923	1,923	1,357
Adjusted R2	0.0383	0.0331	0.0495	0.0358	0.0658

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 12. Comparison of the Experimentation and Conversion Models, All Factors

Comparison of Experimentation and Conversion Models

	(1) Adoption (Published in CRISPR)	(2) Experimentation (Ordered CRISPR)	(3) Conversion (Published | Order)
Central Location	0.0011	-0.0041	0.0396
	(0.0014)	(0.0031)	(0.0375)
Average Impact Factor (3 years)	0.0001**	0.0010***	0.0007
	(0.0000)	(0.0001)	(0.0014)
Total Papers in Top 1% of Citations in US (3 years)	0.0043*	0.0041+	0.0286***
	(0.0018)	(0.0023)	(0.0082)
Mammal Lab	-0.0012+	0.0024+	-0.0753**
	(0.0006)	(0.0014)	(0.0277)
Other Eukaryote Lab	0.0006	0.0061**	-0.0250
	(0.0009)	(0.0021)	(0.0405)
Bacteria Lab	-0.0018*	-0.0056**	-0.0573
	(0.0007)	(0.0017)	(0.0465)
US Rank	-0.0000+	-0.0000	-0.0002+
	(0.0000)	(0.0000)	(0.0001)
Experience (Years)	-0.0012***	-0.0091***	0.0047
	(0.0001)	(0.0003)	(0.0085)
Number of Published Subjects	0.0004	0.0055***	0.0242*
	(0.0003)	(0.0005)	(0.0094)
R&D Expenditures per PI (Millions)	0.0015*	0.0112***	-0.0151
	(0.0006)	(0.0021)	(0.0220)
Average Number of Granting Agencies per Paper	-0.0000	0.0026***	-0.0011
	(0.0002)	(0.0007)	(0.0049)
Total Number of Papers (3 years)	0.0013***	0.0051***	0.0032**
	(0.0002)	(0.0002)	(0.0010)
Average Number of Co-Authors (3 years)	-0.0002***	-0.0011***	-0.0048*
	(0.0000)	(0.0001)	(0.0019)
Constant	0.0050***	0.0408***	0.0261
	(0.0011)	(0.0029)	(0.0559)
Observations	64,709	64,709	1,357
Adjusted R2	0.0293	0.0894	0.0658

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 13. Comparison of the Experimentation and Conversion Models, All Factors (Standard Units)

Comparison of Experimentation and Conversion Models

	(1) Adoption (Published in CRISPR)	(2) Experimentation (Ordered CRISPR)	(3) Conversion (Published | Order)
Central Location (std)	0.0046	-0.0061	0.0304
	(0.0059)	(0.0046)	(0.0288)
Average Impact Factor (3 years) (std)	0.0168**	0.0556***	0.0156
	(0.0057)	(0.0048)	(0.0304)
Total Papers in Top 1% of Citations in US (3 years) (std)	0.0526*	0.0173+	0.1354***
	(0.0223)	(0.0097)	(0.0390)
Mammal Lab (std)	-0.0125+	0.0088+	-0.1084**
	(0.0067)	(0.0051)	(0.0399)
Other Eukaryote Lab (std)	0.0042	0.0136**	-0.0229
	(0.0058)	(0.0047)	(0.0371)
Bacteria Lab (std)	-0.0127*	-0.0131**	-0.0406
	(0.0050)	(0.0041)	(0.0330)
US Rank (std)	-0.0094+	-0.0033	-0.0535+
	(0.0052)	(0.0047)	(0.0312)
Experience (Years) (std)	-0.1099***	-0.2773***	0.0246
	(0.0094)	(0.0081)	(0.0449)
Number of Published Subjects (std)	0.0131	0.0573***	0.1095*
	(0.0086)	(0.0052)	(0.0426)
R&D Expenditures per PI (Millions) (std)	0.0112*	0.0276***	-0.0174
	(0.0047)	(0.0051)	(0.0254)
Average Number of Granting Agencies per Paper (std)	-0.0036	0.0663***	-0.0076
	(0.0183)	(0.0172)	(0.0331)
Total Number of Papers (3 years) (std)	0.1502***	0.2030***	0.1287**
	(0.0220)	(0.0096)	(0.0415)
Average Number of Co-Authors (3 years) (std)	-0.0479***	-0.0762***	-0.0619*
	(0.0082)	(0.0058)	(0.0240)
Constant	0.0694***	0.1679***	0.0096
	(0.0087)	(0.0079)	(0.0350)
Observations	64,709	64,709	1,357
Adjusted R2	0.0293	0.0894	0.0658

Robust standard errors in parentheses

Orders can occur from Q3 2012 - Q4 2014; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

11 Appendix A: Robustness

Table A1. Comparing CRISPR Orders from 2012-2014 to Orders from 2012-2013

Robustness Checks: Comparison of Experimentation Models

	(1) Experimentation (Ordered CRISPR) 2012-2014	(2) Experimentation (Ordered CRISPR) 2012-2013	(3) Conversion (Published | Order) 2012-2014	(4) Conversion (Published | Order) 2012-2013
Central Location	-0.0041	-0.0014	0.0396	0.0868
	(0.0031)	(0.0024)	(0.0375)	(0.0589)
Average Impact Factor (3 years)	0.0010***	0.0005***	0.0007	0.0036
	(0.0001)	(0.0001)	(0.0014)	(0.0026)
Total Papers in Top 1% of Citations in US (3 years)	0.0041+	0.0049*	0.0286***	0.0200
	(0.0023)	(0.0019)	(0.0082)	(0.0138)
Mammal Lab	0.0024+	0.0005	-0.0753**	-0.0940*
	(0.0014)	(0.0010)	(0.0277)	(0.0444)
Other Eukaryote Lab	0.0061**	0.0047**	-0.0250	-0.0428
	(0.0021)	(0.0016)	(0.0405)	(0.0623)
Bacteria Lab	-0.0056**	-0.0024+	-0.0573	-0.1514*
	(0.0017)	(0.0012)	(0.0465)	(0.0729)
US Rank	-0.0000	-0.0000	-0.0002+	-0.0005*
	(0.0000)	(0.0000)	(0.0001)	(0.0002)
Experience (Years)	-0.0091***	-0.0044***	0.0047	0.0069
	(0.0003)	(0.0002)	(0.0085)	(0.0133)
Number of Published Subjects	0.0055***	0.0031***	0.0242*	0.0412**
	(0.0005)	(0.0004)	(0.0094)	(0.0159)
R&D Expenditures per PI (Millions)	0.0112***	0.0085***	-0.0151	-0.0456
	(0.0021)	(0.0015)	(0.0220)	(0.0367)
Average Number of Granting Agencies per Paper	0.0026***	0.0007+	-0.0011	-0.0273*
	(0.0007)	(0.0004)	(0.0049)	(0.0121)
Total Number of Papers (3 years)	0.0051***	0.0029***	0.0032**	0.0030*
	(0.0002)	(0.0002)	(0.0010)	(0.0013)
Average Number of Co-Authors (3 years)	-0.0011***	-0.0006***	-0.0048*	-0.0061+
	(0.0001)	(0.0001)	(0.0019)	(0.0034)
Constant	0.0408***	0.0148***	0.0261	0.0491
	(0.0029)	(0.0019)	(0.0559)	(0.0950)
Observations	64,709	64,709	1,357	648
Adjusted R2	0.0894	0.0544	0.0658	0.0741

Robust standard errors in parentheses

Models 1 and 3: Orders from Q3 2012 - Q4 2014, Models 2 and 4: Orders only from Q3 2012 - Q4 2013; Dropped is No Dominant Lab

+ $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

12 Appendix B: Measures of Scientist Characteristics

Although we identify seven broad factors for experimentation and adoption in the literature, we need to adapt these measures to fit with our sample of academic scientists.

12.1 Experimentation (Ordered CRISPR)

We use orders from the CRISPR Experimentation Data as an indicator of scientists experimenting with the tool. If a unique identifier set is associated with a CRISPR order[28] in 2012, 2013, or 2014, then we consider that author to be experimenting with CRISPR. It would also be possible for an author to receive a CRISPR plasmid indirectly through a co-author, or someone else who transfers it to them, but we see little evidence of this.

12.2 Adoption (Published in CRISPR)

We consider an academic author to have successfully 'adopted' if they publish a paper using CRISPR any time through Q2 2016. We identify a CRISPR paper as any in our WOS subset that mention "CRISPR" or "CRISPR/Cas9" in the WoS Keywords or in the abstract.

This definition may pick up papers that reference CRISPR but do not use the tool in an experiment. To address this, we limit to cases where the author experimented with CRISPR first and then published a CRISPR paper. Manual checking confirms that this restriction accomplishes the goal we intend it to.

It is also possible that this definition misses papers where authors use CRISPR as a tool, but do not mention it in the abstract or keywords. This is often a concern once a tool becomes established, but in this very early segment of the CRISPR lifecycle, almost everyone wanted to mention that they were using CRISPR. It was often a joke that including "CRISPR" in the abstract of a paper submission would almost guarantee a publication in the earliest years.

Due to the early stage of CRISPR development (only five years since the first publication), our definition of conversion will not capture papers where there are delays in publication. We mitigate this by restricting orders to no later than 2014, but this is nevertheless a key part of our definition of being an Early Adopter of CRISPR.

[28] This characterization of the plasmid as being CRISPR is identified in the plasmid description, as a distinct tag.

12.3 Conversion (Published in CRISPR having ordered it)

Conversion is defined by experimentation and adoption. We deem an author has converted their experimentation if we see them adopt after having experimented with CRISPR, as described above.

12.4 Location

Our dataset of authors is restricted to any author who has a US affiliation in WOS at the time CRISPR's introduction (in either 2011 or 2012). Since being located near the original discovery of a tool is expected to increase the likelihood of adoption, we go further to define an author as being centrally located to the discovery of the CRISPR tool if they have a correspondence address in WOS in Berkeley, CA or Cambridge, MA in 2011 or 2012.

12.5 Availability of Resources

The literature also suggests that authors with more resources are more likely to experiment and adopt new tools possibly because they have less to lose from failure. Due to the large number of authors in our dataset, it was not possible to get information on the amount of money available to each individual. However, we developed proxy measures based on individual's university affiliations and grants associated with their publication histories.

The first proxy we use is the rank of the US university that the author is associated with in WOS. Universities with higher ranks usually have more funds available through grants and endowments. Although there are many other factors that tie into university rankings. If an author is associated with more than one US university in our dataset, we take the best rank (the rank closest to 1). Ranks come from the 2012 US News & World Report on the best colleges in the US (US News & World Report 2012). US university names between WOS and US News & World Report were hand matched. In some cases WOS reported school names at a higher level than US News & World Report, for example for some papers WOS would report "University of California System" and US News & World Report would report "UC – Berkeley" and "UCLA" separately. In those cases, only exact matches on the campuses were used. Any US universities remaining without a rank after the hand match were given a rank of 200, lower than the lowest reported rank in US News & World Report, 2012. Robustness tests show that this assumption does not influence the results we observe.

In order to have the intuitive direction of the coefficients in our tables, i.e. a positive estimate indicates that better ranks are more likely to explore / convert, we multiply ranks by negative one, so that a higher rank is indeed better.

We create a university-level variable for funding per principal investigator (PI) using the National Science Foundation Higher Education Research and Development (HERD) Survey from FY 2011. The FY 2011 HERD survey covers 954 US universities that granted a bachelor's degree or higher and had at least $150,000 in R&D in FY 2011 (NSF 2012b). It reports expenditures on research and development by university with headcounts. For our calculations, we use total R&D expenditures by university for FY 2011 (Table 13) normalized by the total number of PIs (Table 42) (NSF 2012a). Again, university names were hand matched between WOS and HERD. If authors were affiliated with more than one US university, the average expenditures per PI were used for that individual.

Even within a university, scientists have different levels of available resources. We use grant information available for each paper in the WOS to construct an individual-level resource variable. Most papers in our dataset include a list of the number of grants and granting agencies that funded the research (although dollar amounts are not included). We hypothesize that authors who are able to receive money from more granting agencies on average have more available resources. To control for authors with more years of experience, we calculate this variable as the average number of granting agencies per paper for the 3.5 years prior to CRISPR (January 2009 – June 2012). To be included, authors must have their first paper in our dataset prior to 2009 (i.e., we do not include new entrants to the dataset in 2009-2012).

12.6 Ability to Apply New Technical Knowledge

The ability to apply the new technical knowledge is important for experimentation and conversion because the more experience an individual has with a field and the more useful the tool for that field, the more likely that person is to experiment in that field with the new tool. In the CRISPR setting, scientists tend to work in three broad sub-communities based on biological cell type: bacteria, mammals, and other eukaryotes (including yeast, plants, and insects). From conversations with scientists who use CRISPR and the staff at Addgene, CRISPR can be used in all cell types, but it is most useful for mammalian cells since there are very few alternatives available for DNA editing there. It is easiest to use in bacteria since their biology is more amenable to DNA editing.

To determine the primary focus of each lab we worked with an expert in synthetic biology and genetic engineering, Aditya Kunjapur, a PhD student at MIT to create a dictionary of terms that would most likely indicate whether a paper focused on bacteria, mammals, or other eukaryotes. The terms are in the following table:

Words from Abstracts	Bacteria	Mammal	Other Eukaryotes
bacillus	1		
bacter	1		
bacteria	1		
coccus	1		
coli	1		
prokaryot	1		
streptomyces	1		
xanthomonas	1		
AAV		1	
adeno		1	
animal		1	
human		1	
macaque		1	
mammal		1	
mice		1	
mouse		1	
rodent		1	
arabidopsis			1
Drosphilia			1
elegans			1
eukaryot			1
fish			1
frog			1
fungi			1
insect			1
plant			1
Xenopus			1
zebrafish			1

To construct the percent focus of each author, we created dummy variables for each term and assigned a 1 if the term was in a paper abstract and a 0 otherwise. This was done for all papers from 2000 – June 2012. Then by author, we summed the number of terms associated with bacteria, mammal, and other eukaryotes and divided each by the total number of terms in all papers for that author. This results in the proportion of terms that focus on bacteria, mammal, or other eukaryote an author uses prior to the introduction of CRISPR.

The continuous percent focus variables can be difficult to interpret, so we use binary variables that put each author into four mutually exclusive categories based on their publication history prior to June 2012. An author is considered having a dominant mammalian focus if her

mammalian percent focus is greater than 60%, she has a dominant bacteria focus if her bacteria percent focus is greater than 60%, or she has an other eukaryote focus if her other eukaryote percent focus is greater than 60%. If an author is not in one of these categories because their prior research covers a mixture of mammal, bacteria, or other eukaryotic organisms, they are placed in a No Dominant Lab category.

Another useful measure for author experience is the total number of papers an author has been able to publish, indicating success in shepherding a new idea to fruition in the past. Again, to control for authors with more experience, we calculate this variable as the total number of papers in WOS for the 3.5 years prior to CRISPR (January 2009 – June 2012). To be included, authors must have their first paper in our dataset prior to 2009 (i.e., we do not include new entrants to the dataset in 2009-2012).

Our final measure of experience is the total number of years an author has in our dataset. This is calculated as 2012 (the year CRISPR was introduced) minus the first year an author's paper appears in our WOS dataset. So if an author's first paper is observed in 2000, they are considered having 12 years in our dataset. Because our dataset begins in 2000, there is some early truncation, but the majority of first papers come after 2005.

12.7 Opinion Leadership

Individuals who are early adopters of a technology often have influential opinions that persuade others to adopt in the future. In the academic sciences, one way to measure influence of an author is to observe the types of journals in which he or she publishes. Journals with high impact factors are highly cited and positively influence future research. To obtain the journal impact factor for each paper we match the journal title in WOS to the journal title in the InCites Journal Citation Reports published by Thompson Reuters (now Clarivate Analytics) for 2012. The report provides the impact factor for 2012 and the average impact factor for the previous 5 years.

Although impact factors remain relatively consistent year to year, we choose to use the average 5-year impact. By author, we take the average 5-year impact over her papers to measure influence. Again, to control for authors with more experience, we use the number of papers in WOS for the 3.5 years prior to CRISPR (January 2009 – June 2012). To be included, authors must have their first paper in our dataset prior to 2009 (i.e., we do not include new entrants to the dataset in 2009-2012).

12.8 Social Participation

Early adopters tend to be well connected in their field and socially. For academic scientists, one way to measure the relative number of connections they have in their field is to calculate the

average number of co-authors a focal scientist has per paper. To calculate this we count the number of co-authors per paper and then take the average by focal author. This is the simple average of co-authors, so repeat co-authors of a focal author will be counted every time they appear. We control for authors with more experience by using the number of papers in WOS for the 3.5 years prior to CRISPR (January 2009 – June 2012). To be included, authors must have their first paper in our dataset prior to 2009 (i.e., we do not include new entrants to the dataset in 2009-2012).

12.9 Novelty

Azoulay et al. (2011), showed that one way to measure creativity and novelty is to count the number of top citations an author has adjusted for the publication cohort year. WOS provides the number of citations a paper has received as of Q2 2016, so we calculate the number of papers in the top 1% of citations by author. To be included, authors must have their first paper in our dataset prior to 2009 (i.e., we do not include new entrants to the dataset in 2009-2012). We first rank the papers in our WOS dataset for these authors by the number of citations they have received by Q2 2016 and by year of publication. We adjust by year since papers published in 2009 have more years to accumulate citations than papers published in 2012. If the paper ranks in the top 1% of citations for the included papers each year, it receives a 1 and a zero otherwise. We then sum the papers in the top 1% by author.

12.10 Breadth of Research Field

Previous research has also suggested that experimenters and entrepreneurs are able to connect with a broader number of fields rather than focusing on one unique specialty. To measure breadth in the CRISPR setting, we count the number of WOS categories an author's work appears in from 2000 – June 2012. The categories are restricted to those six used to create our subset: Genetics & Heredity, Cell Biology, Biotechnology & Applied Microbiology, Biochemistry & Molecular Biology, Biochemical Research Methods, and Multidisciplinary Studies for the journals Nature, Science, PLOS One, and PNAS.

13 Appendix C: Name Disambiguation and Publication Matching

13.1 Overview

For every plasmid order in the CRISPR Experimentation Dataset we have unique identifiers, including the full name of the principal investigator (PI), the organization they were affiliated with, the country, and the state if in the US. The initial PI database has 67,466 PI-organization pairs since PIs can be associated with more than one organization. The system initially treats every variation in a PI name as a separate researcher, so it was necessary to first consolidate the list of PIs before matching to publications in Web of Science (WOS). For example, Jane A. Scientist, Dr. Jane A. Scientist, and Scientist Jane all at "Research U" would appear as three people. But, given the distinct last name and identical location, these indicate the same person.
The principal steps in our consolidation algorithm are:

(1) Clean the name strings for easier comparison,
(2) Calculate 'string distance' measures between a focal name and the other 67,466 names and define potential matches using reasonable thresholds that change based on frequency of last name,
(3) Run a set of tests that determine whether the two names are a likely match, and
(4) Use transitive connections to complete the formation of groups.

We create a validation set of 50 random records and find that the algorithm has a 96.5% accuracy rate in that set. We provide additional details on the algorithm steps below.

13.1.1 Name String Cleaning and New Variable Creation

To create the consolidation algorithm, we begin by cleaning the PI names including removing punctuation, eliminating numerical digits and non-ASCII characters, and removing any trailing or leading titles including variations on Dr. and Professor. We convert the name strings to lowercase to avoid differences due to capitalization choices. Before removing all spaces, we separate out the last string after the last space and assume that is the PI's last name.

We make some manual changes to the cleaned full and last names based on the number of spaces appearing in the original (pre-cleaned) PI name, assuming there should be at least one space in a name and no more than three. We made the following manual changes on top of the cleaned full and last names:

- In all cases we removed additional suffixes and titles that were not found by our initial cleaning of the data due to lack of a space.
- Changed the order of names if entered as "last,first."

- Did a web search for common names and locations to determine which may be different people.
- Removed extraneous notes in name fields.
- Used the first PI if a list of PIs was given.

The remainder of the algorithm uses the full and last names that incorporate both the coded and manual cleaning for the best results. Based on the last name column, we calculated the frequency of the last name in the dataset for use in the matching algorithm.

13.1.2 Determining Potential Matches

Because the names in the CRISPR Experimentation Dataset are sometimes missing, misspelled, or otherwise idiosyncratic (even for the same author), the algorithm uses fuzzy matching to determine whether any two names are potential matches. We first create a list of the 67,466 names and set a loop to compare one focal name to every other name in the list, so name 1 is compared to names 1-67,466, then name 2 is compared to names 1-67,466, and so on until name 67,466 is compared to names 1-67,466. We use a number of different tests of 'string distance' in the comparisons so that closeness is not determined by only one measure. Thresholds for each of the tests are determined based on how common the last name is in the full dataset.

For each last name, the loop first determines how common it is (below the median, in the top 25% of names, or between the median and top 25% of names). Then, each pair of full names is compared using cosine distance (based on letters), Levenshtein distance, and longest common substring distance (van der Loo 2014). For each distance measure an exact match between two names is zero so we set broad and narrow thresholds greater than zero for being "close." The narrow thresholds require the names to be near exact. The broad thresholds allow for more flexibility in differences between the names. For example, the Levenshtein distance counts the number of deletions, insertions, and substitutions needed to turn string A into string B. For the least common names in our database, the narrow threshold allows three letters to be different and the broad allows six letters to be different. This process is repeated for last names only. This results in 12 tests per pair of names with the thresholds determined by how frequently the last name appears in the dataset.

Since each of these tests has some power, but is not decisive, we evaluate whether two names are a potential match using an ensemble method from machine learning: if 75% of the tests return positively, we consider the pair to be a potential match. The thresholds used for each test by last name frequency are in the table below.

String Tested:	Narrow Threshold			Broad Threshold		
	Cosine Distance Threshold	Levenshtein Distance Threshold	Longest Common Substring Distance Threshold	Cosine Distance Threshold	Levenshtein Distance Threshold	Longest Common Substring Distance Threshold
	Last Name in Top 25% of Last Names					
Full Name	0.01	1	1	0.05	2	2
Last Name	0.02	2	2	0.07	3	3
	Last Name Between Top 20% and 50% of Last Names					
Full Name	0.05	3	3	0.1	6	6
Last Name	0.05	3	3	0.1	6	6
	Last Name in Bottom 50% of Last Names					
Full Name	0.05	3	3	0.1	6	6
Last Name	0.05	3	3	0.1	6	6

13.1.3 Tests for Paired Name Matches

Once the algorithm determines that a pair of names is a potential match, it will consider the pair a true match if at least one the following conditions are met. For names in the top 25% of the distribution and between the top 25% and 50%, a pair of names are a match if (1) the PI IDs are the same, (2) if the full names are an exact match and the locations are the same, or (3) the full names are a close match and the locations are the same. For names in the lower half of the distribution a pair is a match if any of the same criteria hold or (4) the last names are an exact match and the locations are the same.

13.1.4 Including All Paired Names in One Group

Because the algorithm tests each name separately against all others, it is possible that name A will be considered a match for name B and name B is a match for name C, but that names A and C will not be a match. For example, consider the following records:

Row	ID	Full Name	Last Name	Location ID
1	1	janescientist	scientist	1
2	1	janescientist	scientist	2
3	2	scientistjane	jane	1

Here, the algorithm would consider rows 1 and 2 a match because they have the same ID. Rows 1 and 3 would also be a match because the names are close and the location is the same. Rows 2 and 3 are not a match however, because although the names are close, the ID and location are different.

These are clearly the same person, though and should be considered one group. To solve this problem, we find the transitive closure[18] of each group using the iGraph package in R. This results in a matrix that has identical rows for each group of names. For each group of identical rows, we assign the group the minimum PI ID number to maintain consistency with the original CRISPR Experimentation Dataset. This process results in 54,133 consolidated IDs.

13.1.5 Algorithm Validation

To test the accuracy of the algorithm, we randomly selected 50 rows and hand-coded whether each name was a match to the 67,466 names. We then ran the algorithm on the same 50 rows and compared the resulting matches. We find that 100% of the algorithm matches are true matches. We test whether the algorithm is conservative, perhaps not matching some true matches. We find that of 86 'true' matches (i.e. hand-coded ones), the algorithm successfully captured 83 of these. Thus, even with our conservative approach, we correctly identify 96.5% of the true matches (83/86) in our test set.

13.2 CRISPR Experimentation Dataset to WOS Matching

We use a nearly identical process to the consolidation algorithm described above to then match the 54,133 CRISPR Experimentation Dataset names to authors in Web of Science, but with a little pre-processing. In particular, because there are 4.4 million authors in the WOS data, it would be computationally intractable to computer our approach for a 54K by 4.4 million matrix. To address this, we preprocess to exclude WOS names that would never be a match for CRISPR Experimentation Dataset name. We assign each CRISPR Experimentation Dataset name and each WOS name a "group name" that consisted of the first initial and last name. We then created a list of group names that were common to the CRISPR Experimentation Dataset and WOS. For each of those common group names, we created separate CSV files for the CRISPR Experimentation Dataset and WOS. The matching algorithm then loops through each group name, selects the appropriate CSV files to be compared, and runs the same consolidation algorithm on each name within the group files. The matches are recorded in a separate file and duplicate records are removed at the end. The final author IDs, by construction, are still the minimum PI ID.

The only additional modification is that CRISPR Experimentation Dataset and WOS locations are not identical. The CRISPR Experimentation Dataset and WOS locations were hand matched and assigned a location ID so that the algorithm could be used in the same way as above.

[18] In Graph Theory this is also called graph closure. It simply means that if A is matched to B, and B to C, then A should be matched to C, even if they wouldn't independently be matched.

Chapter 3

Running with (CRISPR) Scissors: Tool Adoption and Team Assembly

1 Introduction

Some of history's greatest technological advances can be attributed to research tools. Research tools include physical inputs into the process of discovery and can be inventions of a method of invention with large economic impacts across a range of domains (e.g., Griliches 1957; Walsh, Arora, and Cohen 2003; Cockburn, Henderson, and Stern 2017). For example magnification provided by microscopes allowed Antony Van Leeuwenhoek to first observe bacteria, critical to today's understanding of biology and medicine (Wills 2018). The introduction of Polymerase Chain Reaction (PCR) by Kary Mullis allowed scientists to make copies of DNA which improved the speed of diagnostic tests and revolutionized the way scientists manipulated genetic material (Rabinow 2011). The statistical software package STATA improved researcher productivity in many domains, including economics, finance, epidemiology, political science, and sociology (Pinzon 2015).

Technological progress requires domain knowledge as well as tools (e.g., Rosenberg 1982, 1994, 2009; Nelson 1981, 2003; David 1990; Bresnahan and Trajtenberg 1995; Rosenberg and Trajtenberg 2004). For example, advances in microscopes and molecular pathway knowledge lead to medical breakthroughs. Over time, the combination of new tool and domain knowledge leads to an ever-larger knowledge base from which innovation emerges (Cohen and Levinthal 1989; Weitzman 1998; Fleming 2001; Mokyr 2002; Wuchty et al. 2007; Schilling and Green 2011).

Research tools require adopters to both have access and the ability to apply the tool in a domain (Teece 1986; Scotchmer 1991; Weitzman 1996; Fleming 2001). Tools often embed sufficient know-how so that primarily access to the tool lowers research costs to innovators in the domain and lowers entry barriers to external innovators (e.g., Furman and Stern 2011; Williams 2013; Murray et al. 2016). But not all new tools embed their necessary know-how. In order to effectively apply such tools, early adopters must acquire complementary tool-specific know-how. This additional input to knowledge production is distinct from human capital in the domain and physical capital in the tool. Since the complementary know-how is crucial for the adoption of a new tool that can advance technological progress this paper explores the question: *How do early teams acquire the tool-specific know-how necessary to innovate with a new research tool?*

The amount of complementary specialized tool know-how is not binary, but rather varies along a continuum. Ex ante, newly introduced tools can require more complementary know-how in

some domains than others. Over a set of these domains on the continuum, the complementary know-how can come from teams internal to a domain or external specialists in the tool not associated with the domain. Over the range of domains where there is a choice of where to acquire complementary know-how, internal domain teams face a form of the "make or buy" problem over whether to learn the complementary know-how or collaborate with external specialists (e.g., Coase 1937; Williamson 1975, 1985). For example, consider a research team's decision to use advanced Artificial Intelligence (AI). If the team does not have employees trained in the technology, it must either provide incentives for employees to learn AI themselves or it must find and pay for external specialists, either through training or hiring.

When tools are new, external tool specialists are scarce, giving them a separate choice of which domain teams to join, introducing a two-sided market matching problem (e.g., Gale and Shapley 1962; Roth 1984). External tool specialists can either choose to join teams in easier domains where they can apply the tool quickly and broadly or they can choose to join teams in difficult domains where the problems are complex and solutions are potentially more influential. The decision of where and how to work with scarce external tool specialists to acquire complementary tool know-how is one firms, mangers, and individual innovators face repeatedly as they innovate.

Given the costs and benefits of collaboration and a scarce supply of external tool specialists, it is not immediately clear which domains are likely to attract external tool specialists into effective collaborations more often. However, it is reasonable to hypothesize that team assembly to acquire complementary tool know-how varies by the difficulty of using the tool in the domain and that the ex ante complexity of the domain determines where external tool specialists match with domain teams most often.

One concern with empirically studying knowledge recombination using research tools is that external tool specialists can contribute to the innovation process by providing access and by sharing complementary information about how to use the new technology (Polanyi 1962; Zucker and Darby 2001; Murray 2002). However, access does not necessarily lead to the ability to use a tool (e.g., Cohen and Levinthal 1990; Ahuja & Katila 2001; Zahra and George 2002; Thompson and Zyontz 2017). External specialists' contributions to the application of a new technology apart from access are difficult to identify empirically because access is commonly conflated with the ability to use the tool or tool and domain knowledge develop concurrently (e.g., Furman and Stern 2011; Murray et al. 2016; Teodoridis 2018).

To identify whether and where external tool specialists provide specialized complementary tool know-how to create early innovations, an ideal setting would separately identify external tool specialists. The setting would also allow access to be tracked separately from the ability to use the tool. Finally, the ex ante difficulty of applying the tool should vary across domains, and the tool

should enter different domains randomly. Such an environment does not occur naturally, but the recent introduction of the DNA-editing tool, CRISPR, provides a novel and approximate setting.

CRISPR is a naturally occurring immune response in bacteria that proved to be a powerful editing tool for DNA modification in almost any organism. The CRISPR tool was first introduced between June 2012 and January 2013 when researchers from Berkeley, MIT, and Harvard demonstrated that CRISPR could be used to edit DNA in both bacteria and mammals. CRISPR is a substantial improvement over existing DNA editing tools especially for researchers working in mammalian cells (Zyontz 2016). For mammalian researchers, it represented a long-term reduction in research costs and an unanticipated increase in new opportunities across a wide range of domains.

CRISPR has incredible promise in human disease domains caused by genetic mutations, such as infections, viruses, and inherited genetic diseases. However, not every such disease received access to CRISPR at the same time, as much as researchers wanted to adopt the tool. Although CRISPR works similarly once delivered into a cell affected by a disease, certain cells are biologically more difficult to edit than others. This imposes natural delays on when CRISPR can be applied to a particular disease, since researchers must first overcome this delivery problem. The unanticipated timing of the CRISPR tool introduction to different human disease domains mitigates some of the endogeneity inherent in adoption and helps to identify the knowledge bases of the innovators responsible for the articles that use CRISPR in a disease.

Specifically, the CRISPR tool was not originally developed for any specific domain application because it was a shock to gene editing. It took almost another year for a human disease application to appear. This established a set of separate CRISPR tool specialists not associated with a disease but whose know-how would be useful in any disease domain. In the earliest years (2012 – 2016), the primary adopters of CRISPR were academic scientists so it is possible to use historical publications from PubMed to separately identify external tool specialists from domain specialists in CRISPR.

Because biological materials are often transferred between labs, science settings can conflate the different access and complementary know-how contributions of an external tool specialist. Material Transfer Agreements (MTAs) can delay access or require co-authorships for the receipt of the materials (Walsh, Cohen, and Cho 2007; Strandberg 2010). The CRISPR setting circumvents these concerns by having a biological resource center that breaks the link between contracting for access to the tool and acquiring complementary know-how from external specialists. From 2012-2016, Addgene was the primary distributor of CRISPR to academic researchers and as a third-party, eliminated the need for scientists to sign MTAs directly with other academics. Thus any external tool specialists observed entering a new domain in this setting primarily represent value added know-how rather than the price of access.

The results show that the share of external tool specialists on new CRISPR papers in a disease is significantly larger in more difficult disease domains, suggesting that the match between external tool specialists and domain teams occurs more often in domains that focus on solving influential problems rather than the breadth of implementation. The share of external tool specialists also increases for subsequent CRISPR papers in difficult disease domains, so the effect does not attenuate immediately. The paper is the first to introduce CRISPR as a setting to empirically study how teams form to effectively overcome tool adoption barriers in know-how. It also uniquely shows that effective team composition is driven by the specific nature of the problem and the nature of tools available for innovation, not just features of management, organizational structure, or industry. As more tools emerge that require the acquisition of complementary tool know-how, like AI, research teams looking to be early adopters of such tools will have to weigh the complexity and importance of their possible solutions in considering how best to attract and collaborate with external tool specialists.

Section 2 discusses the relevant literature on knowledge inputs to innovation production and the choices faced by internal domain teams and external tool specialists. Section 3 provides details on the CRISPR setting. Section 4 outlines the identification and empirical specifications used for the analysis. Sections 5 and 6 describes the measures constructed and results. Section 7 provides a discussion of the results and concludes.

2 Tools as Inputs to Innovation Production

The recent emergence of AI can be used to innovate in a range of applications including natural language processing, image recognition, enhanced data security, and smart products like cars (Marr 2016). Using AI tools as inputs to innovation in these areas requires more than just accessing an off-the-shelf product. AI also requires adopters to acquire specialized knowledge and skills including coding, training models, building computing infrastructure, and scaling for firm-wide implementation. The amount of complementary knowledge needed varies by application area. From a firm's perspective, it can either have their employees learn the complementary AI know-how internally or can bring in external AI specialists. External specialist know-how can be useful across AI applications, but when external specialists are scarce they also get to choose the application areas in which to collaborate. For example, the initial team at Google working on the self-driving car consisted of Google-X employees with engineering and AI experience. The team could have eventually learned the complex complementary AI know-how internally, but instead Google-X collaborated with a specialist previously running the Stanford Artificial Intelligence Laboratory at Stanford University, Sebastian Thrun (Dallon 2017). However, Dr. Thrun's also had the viable choice to join the team or work on different applications since his skillset was rare.

The AI anecdote illustrates how recombining external specialized know-how with internal domain knowledge can help research teams move from access to ability to use a new tool. It also suggests that complexities in a domain may influence the match between internal domain teams and external know-how. External tool specialists may be attracted to teams doing more complex and influential work when their skillsets are in demand and scarce.

2.1 Research Tool Adoption and Innovation

Research tools are types of inputs to innovation distinct from human capital in a domain. Tools are integral to technological progress in an application domain because they often embed their own know-how, allowing users to apply the tool without understanding why it works (Mokyr 2002). Access to these tools helps to lower the costs of research to innovators in the domain and invites those outside of the domain to make new contributions (e.g., Furman and Stern 2011; Williams 2013; Murray et al. 2016; Teodoridis 2018; Furman and Teodoridis 2018).

Much of the traditional adoption literature focuses on the diffusion of products throughout their lifecycles and their role in economic growth or social returns. Related literature discusses the types of individuals who adopt these products, but neither focuses directly on the role of research tools in creating innovations. Some of the earliest work on product adoption looked at the social rate of return to hybrid corn research (Griliches 1957, 1958) and showed that although there is a high return to investment, it takes time for products to diffuse due to both availability and acceptance. Locations most in need of the new product will likely adopt it sooner, but even within an area diffusion occurs in an "S-shaped" pattern as some individuals wait to adopt. Complementary work on adopter types showed that the earliest product adopters at the beginning of the S-curve are influential in their fields, have resources to adopt, and have the ability to incorporate the new products (Rogers 1962).

The literature on general purpose technologies (GPTs) also focuses on the role of technological progress to economic growth. GPTs can be technologies such as semiconductors (e.g., Bresnahan and Trajtenberg 1995), steam engines (Rosenberg and Trajtenberg 2004), or information and communication technologies (e.g., David 1990). GPTs generally are considered enabling technologies that encourage economic growth in a large range of downstream sectors through a positive feedback loop between a GPT producer and downstream markets as each makes complementary improvements to the GPT (Bresnahan and Trajtenberg 1995). Different sectors may delay adopting the GPT depending on when it is most valuable.

Innovation as an outcome for research tool adoption is more common in empirical work that focuses on access. Access to research tools has been shown to lead to an increase in the rate and changes to the direction of innovation in a number of settings (e.g., Moser 2005; Azoulay et al 2009; Furman and Stern 2011; Sampat and Williams 2019; Murray et al. 2016; Teodoridis 2018). However,

these papers tend to focus on access to tools that reduce the cost of research quickly with little tool-specific know-how needed. Further, it is assumed that the new and broader work is due to innovators using the tool they can now access. That mechanism is not assured though since the tool is rarely tied directly to the new papers or products.

Furman and Teodoridis (2018) have one example of tool know-how interacting with different domains, but even in the case of Kinect, the tool is assumed to reduce the cost of research at the time of access with little variation. Therefore it is not possible to use the difficulty of applying the tool to understand how tool specific know-how is incorporated. Further the authors, by necessity, assume that the internal and external researchers they study are using the Kinect tool. Nagle and Teodoridis (2017) take a more direct look at the role of researchers who use Kinect and show that it is generalists who tend to bring the new tool into teams. Once again, because the costs of Kinect do not vary by domain, they cannot address how their outcomes might change as the tool is more difficult to use and the problems become more complex or influential.

This paper adds to the literature on research tool adoption and innovation by introducing a way to empirically observe innovations directly due to a tool that not only requires complementary know-how to use but the amount necessary varies by application domain. It also uniquely shows both the nature of the problem and the nature of tools available for innovation affect successful team composition, not just features of management, organizational structure, or industry.

2.2 Research Tools and Complementary Know-How

Many tools that have been historically important for technological advancement are those that eventually embed the necessary know-how in the physical product including hammers, scissors, microscopes, steam engines, automobiles, or telephones. For these, the decision to adopt is mostly rooted in access to the tool.[1] However, some tools when first introduced do not embed necessary know-how and require the user to learn complementary know-how even after obtaining access. For example, tools like early wind tunnels, the first computers, or early software packages like STATA all embedded some of the underlying knowledge of physics, computing algorithms, or statistics. However, users required additional knowledge of the tool in order to apply it to different problems. For example, the earliest versions of STATA could run multiple regression analysis directly, but time-series analysis and other more advanced statistical models still needed to be programmed by the user (Pinzon 2015).

The above might suggest that tools should be classified dichotomously – those that need no additional know-how and immediately lower learning costs versus those that need a host of complementary know-how to bring learning costs down. However, this is an oversimplification.

[1] Although, at introduction, many tools embed less of their own know-how than they do after they become more routinized.

Instead, it is possible to think of tools appearing on a continuum based on the amount of complementary know-how needed to employ the tool in a domain at a particular time. Some tools like hammers are introduced with all necessary embedded know-how so their positions on the continuum generally do not change over time. Other tools embed a greater amount of knowledge over time. Continued use of these tools bring about improvements in performance and modifications that embedded more knowledge in the tool itself. For example, as users created code for more advanced statistical models, later versions of STATA included those updates so that non-statisticians could apply the models just as easily. Thus later versions of the tool can appear in different locations on the continuum than the original. Finally, the same tool can appear in different locations on the continuum based on the application domain at a given point in time.

If tools lie on a continuum of least to most necessary complementary know-how, then there is a range of tools that require the user to learn complementary know-how as an additional input to innovation. One option for adoption of tools within this range is to delay until the tool is improved and embeds enough needed know-how, which is an aspect of adoption discussed in the adoption literature (e.g., Griliches 1957, 1958). However, our understanding of how early users adopt tools when this complementary know-how is still needed is incomplete. Understanding how early teams form to effectively adopt tools that require complementary know-how and use them in innovations provides a better idea of the choices that shape early adoption and the innovative paths that are formed from these early decisions by successful teams.

For teams already in an application domain that want to be early adopters, they can choose to learn the necessary know-how internally or they can collaborate with external tool specialists, those that acquired the tool know-how independent of a domain. However, external tool specialists are scarce in this early stage and can choose in which domains to work. Effective collaborations will only occur in domains where there are sufficient incentives for both sides.

2.3 Domain Team Choices

Over a range of domains where a newly introduced tool requires the user to learn complementary know-how, teams that want to be early adopters face a form of the "make or buy" problem (e.g., Coase 1937; Williamson 1971, 1975, 1985; Grossman and Hart 1986). Teams can invest time learning the tool know-how internally (through the team leader or another team member). Alternatively, they can choose to acquire the know-how from external tool specialists who developed human capital in the tool independent of the domain (e.g., Cassiman and Veugelers 2006 and Grigoriou and Rothaermel 2017). However, in order to successfully use external knowledge, an organization must not only have access to the new knowledge but must also have the resources and ability to incorporate it (Cohen and Levinthal 1990).

The choice to internally learn the complementary know-how or acquire it from external tool specialists involves weighing the costs and benefits of each option. For example, learning internally has the benefit of providing the internal domain team control over its work, making it more self-sufficient for future innovations. The drawback is that learning a new tool can take time, possibly causing the team to give up a first mover advantage (Jones 2009). Collaborating instead with external specialists can reduce the time it takes to apply the tool since the specialists bring the complementary know-how with them (e.g., Arora and Gambardella 1994; Wuchty et al. 2007; Uzzi et al. 2013). However, collaborations have inherent frictions that need to be overcome before the tool can be applied so there is a risk that the collaboration could fail (Cummings and Kiesler 2007; Bikard et al. 2015). Internal domain teams will seek external tool specialists if the difference between the costs and benefits of collaborating is greater than the difference between the costs and benefits of learning internally.

2.4 External Tool Specialists Choices

Although innovation often emerges from a recombination of previous ideas (e.g., Scotchmer 1991; Fleming 2001; Kaplan and Vakili 2015), researchers need both access to a tool and the ability to incorporate specialized tool know-how (Teece 1986; Scotchmer 1991; Weitzman 1996; Fleming 2001) to successfully innovate. However, access does not necessarily lead to the ability to use a tool (e.g., Ahuja & Katila 2001; Zahra and George 2002; Thompson and Zyontz 2017). External tool specialists can contribute to the innovation process by sharing information about how to use the new technology (Polanyi 1962; Zucker and Darby 2001; Murray 2002).

When a research tool is introduced, the stock of external tool specialists is initially small. If there is a rush to use the new tool, scarce external tool specialists can choose the teams and domains where they wish to share their complementary tool know-how. If external tool specialists choose to collaborate with a team, it may be because collaborations are increasingly common and they have been shown to result in higher productivity and higher quality ideas (Adams et al. 2005; Stephan 2012; Gans and Murray 2014). By collaborating with teams in easier domains where the tool needs less complementary know-how, external tool specialists can be more productive and broadly apply the tool to more domains quickly. On the other hand, by collaborating with teams in difficult domains where the tool needs more complementary know-how, external tool specialists can work on more challenging and complex problems where the possible solutions are highly influential. The external tool specialists' choice creates a two-sided matching market for the scarce human capital in the complementary tool know-how. Internal domain teams and external tool specialists will only collaborate in domains where there is a match on both sides.

Given the costs and benefits of collaboration to internal domain teams and external tool specialists, it is not ex ante obvious which kinds of domains are likely to attract external tool

specialists into collaborations more often. However, it should be the case that the ex ante difficulty of an application domain will determine where the successful matches occur. To test how early teams acquire specialized complementary tool know-how to produce innovations, the recent introduction of the DNA-editing tool, CRISPR, provides a novel setting where the knowledge bases of the earliest innovators can be separately identified using natural variation in CRISPR's entry into different domains. The next section describes CRISPR and useful factors of the setting.

3 Gene Editing with CRISPR

CRISPR provides a unique setting for exploring how innovative teams acquire specialized complementary tool know-how. First, CRISPR was an unexpected shock to gene editing researchers and the DNA-editing tool's use has exploded since its introduction. Second, access to the tool is widely available through the biological resource center Addgene, alleviating the need for material transfer agreements between labs. Third, CRISPR can be applied to many different disease domains but was not developed for a particular application. The difficulty of applying CRISPR to different diseases varies based on natural properties of the cells to be edited.[2]

3.1 Gene Editing Before CRISPR

DNA editing has led to many advances including the creation of model organisms and the modification of existing organisms since its introduction in the early 1970s. These advances allowed researchers to better understand human disease and to create useful products like pesticide resistant crops. For bacteria or other prokaryotes that have cells without nuclei, relatively easy editing techniques existed prior to CRISPR. However, for higher-order species, like mammals, even recent editing alternatives like Zinc Finger Nucleases (ZFN) and Transcription Activator-Like Effector Nucleases (TALENs) (Moscou and Bogdanove 2009; Boch, et al. 2009) are difficult and time-consuming to use. Despite this, TALENs was chosen as "Method of the Year" in 2011 (Method 2012) because of the advancements it represented. Only a few months later, CRISPR came as a surprise to researchers working in gene editing. The new CRISPR tool acts like a pair of universal DNA scissors that work across organisms and is a much more flexible option for DNA editing than any other tool, especially for complex organisms like mammals.

[2] Further details on CRISPR and its introduction to gene editing beyond those provided in this section can be found in Zyontz (2016) and Thompson and Zyontz (2017).

3.2 CRISPR

The exact purpose and function of CRISPRs, short for Clustered Regularly Interspaced Short Palindromic Repeats, were not well understood until 2007 when it was discovered that the unique DNA sequences are an adaptive part of the bacterial immune system (Barrangou et al. 2007). Bacteria use CRISPR sequences to recognize viral DNA and then use a related enzyme (often the Cas9 protein) to cut up invading viral DNA and destroy the virus.

In June 2012, Professors Jennifer Doudna and Emmanuelle Charpentier at the University of California, Berkeley first introduced a modifiable CRISPR system for DNA editing (Jinek et al. 2012). Doudna and Charpendier proved in test tubes that the CRISPR system could find and edit any DNA sequence (not just viral) using programmed guide sequences to direct the cutting enzyme to the right place in the DNA. Doudna and Charpentier's work provided proof of concept that the CRISPR system could edit organisms like bacteria. In January 2013, MIT Professor Feng Zhang and his collaborators showed that the CRISPR system could also edit mammalian cells, including human cell lines (Cong et al. 2013). This work and the related work of George Church and his colleagues at Harvard Medical School (Mali et al. 2013) demonstrated the flexibility and ease of use of the new CRISPR tool. This was particularly welcomed in the mammalian research community, including those working on human cells, because of the enormous improvement in terms of accuracy and difficulty over previous methods.

To understand how CRISPR works, consider the find-and-replace function in a word processing program. To replace the word "absolutely" with "certainly," the user only need to put in the two strings and the program will find all instances of the string "absolutely," cut it out of the document, and replace it with the second string. CRISPR works the same way: program a DNA sequence to find the same string in the DNA of an organism, use an enzyme to cut out that string, and then use a second programmed string as a replacement in the organism's DNA. One of the main benefits to CRISPR is that it can look for the equivalent of the full word "absolutely." Previous editing technologies could only search for a short string, like the equivalent of "abs." In the find-and-replace analogy, this short string would find "absolutely" but also "abstract," making the edited document (and edited DNA) unreadable.

The enthusiasm from researchers conducting gene editing to the release of the CRISPR tools was almost immediate because of its accuracy, flexibility, and relative ease of use (Pennisi 2013; Regalado 2014). Since Doudna and Charpendier's first paper in June 2012 through December 2016, over 4,500 CRISPR-related articles were published according to the medical publication database PubMed. Patent applications mentioning CRISPR and funding for venture backed firms licensed to use CRISPR technology have also soared (Ledford 2015). By December 2016, over 3,000 patent applications published worldwide mentioned CRISPR. Funding also flowed to CRISPR commercialization efforts. The original biotech firms founded on CRISPR technology, Caribou

Biosciences (Berkeley, CA), Editas Medicine (EDIT; Cambridge, MA), CRISPR Therapeutics (CRSP; Basel, Switzerland), and Intellia Therapeutics (NTLA; Cambridge, MA) collectively raised initial funding of more than $150 million. The last three all had IPOs in 2016, each currently with market capitalizations of over $1 billion.

CRISPR's effect on biological research has been profound, as geneticist John Schimenti at Cornell University noted: "I've seen two huge developments since I've been in science: CRISPR and PCR... CRISPR is impacting the life sciences in so many ways" (Ledford 2015). One of the original inventors, Jennifer Doudna, stated in a February 2015 JAMA editorial, "This discovery has triggered a veritable revolution as laboratories worldwide have begun to introduce or correct mutations in cells and organisms with a level of ease and efficiency not previously possible." (Doudna 2015). CRISPR has already been used to create blight resistant crops (Wang et al. 2014) and "malaria-proof" mosquitoes that are genetically unable to transmit malaria (Gantz et al. 2015). The introduction of CRISPR is likely to be especially useful in medical applications since it may ultimately allow for the correction of genetic errors.[3] Currently, it is allowing researchers to build mouse and human cell disease models more easily with specific mutations that are useful for testing drugs. For example, Feng Zhang's lab has created a "Cas9 mouse" (Platt et al. 2014) that can be modified to model lung cancer. Before CRISPR, creating such a mouse model took large teams of people and a decade to complete, but this model was designed by one person with CRISPR in four months (Specter 2015).

3.3 Access to CRISPR with Addgene

Laboratories often use material transfer agreements to transfer tools from one institution to another, often delaying adoption of the tool (e.g., Mowery and Ziedonis 2007; Walsh, Cohen, and Cho 2007; Strandberg 2010). To circumvent the difficulties associated with lab-to-lab material transfers, biological resource centers, such as the American Type Culture Collection, are created to centralize the distribution process (Furman and Stern 2011). The dominant central repository for CRISPR, from the first day, is Addgene. In 2004 Melina Fan, Kenneth Fan, and Benjie Chen founded Addgene as a non-profit biological resource center for scientists to easily share tools[4] for use in biological research (Fan et al., 2005). Addgene not only stores biological tools donated by academic researchers all over the world, but also validates the materials and facilitates their distribution to other academic institutions in more than 85 different countries and counting.

When the CRISPR tool was first introduced in 2012 and 2013, Doudna, Charpendier, and Zhang donated their versions to Addgene at the time the original papers were published. As Zhang

[3] This has already been done in non-viable human embryos (Ma et al. 2017) and may have been done in viable embryos resulting in gene edited babies as recently as November 2018 (Cyranoski 2018).

[4] The tools donated to Addgene, including CRISPR, are usually distributed as plasmids. A plasmid is a form of circular DNA that is commonly used to replicate or expand upon gene editing experiments. See https://www.addgene.org/ for available tools.

said in a talk at MIT in 2015, he gave CRISPR to Addgene for distribution because his lab "wouldn't have time to do science" if they responded to all the requests from other researchers. Indeed, to date, Addgene has sent more than 42,000 CRISPR tools from Zhang's lab to over 2,000 institutions (Zhang 2018). As new CRISPR tools have been developed, they have also been donated to Addgene. CRISPR orders quickly rose from 0.1% of all Addgene orders in 2012 to about 18% in 2015 (Figure 1). To date, CRISPR is one Addgene's most popular tools, making up over 20% of total orders.

[Figure 1 here]

Addgene has a price of $65 per plasmid which has remained constant since 2004. This stable low cost and consistent quality control process encourages labs to order directly from Addgene rather than make their own or attempt to get it from the original lab. Addgene alleviates the burden on the individual research labs and separates access to the tool from the original inventor.

3.4 Variation in Availability of CRISPR by Disease Domain

When the CRISPR tool was first introduced in 2012, it came as such a surprise that it was not designed with a specific application in mind. However, some of the greatest strides forward have come from mammalian gene editing particularly in human diseases (Zyontz 2016). The co-founders of CRISPR argue that the tool could be useful for the study and eventual treatment (or even cure) of most human diseases caused by to genetic mutations (Doudna and Sternberg 2017; Whitaker 2018), making CRISPR a valuable tool for research in all of these domains. Despite CRISPR's generality, natural barriers in gene editing prevent CRISPR from being available for all disease domains at the same time. Availability variation is due in part to the type of cell primarily targeted by the disease, which affects how CRISPR must be delivered to the nucleus of the cell (e.g., Regalado 2016, Stockton 2017, Kaiser 2016, Wang et al. 2016, LaFountaine et al. 2015). One of the co-founders of CRISPR, Jennifer Doudna, notes in her book, "That's not to say that it'll be easy to get CRISPR inside the cells themselves. This delivery problem is one of the greatest challenges" (Doudna and Sternberg 2017). The more complicated the disease and cell type, the more difficult it is to deliver and use CRISPR, which provides natural delays that vary by disease as researchers overcome the delivery barrier.

Some of the easiest diseases to edit with CRISPR involve cells that quickly self-replicate and can be edited ex vivo.[5] For example, blood cells can be easily removed from an organism, edited with CRISPR in a dish, and then the modified cells are put back in the organism (Regalado 2016). Because blood cells replicate easily, the new ones will replicate with the edit and eventually overtake

[5] In *ex vivo* (exterior) gene editing, target cells are first modified outside a living organism. The edited cells are then returned to the organism as an effective treatment for the disease.

the old damaged cells. T-cells associated with immune deficiencies can also be successfully edited ex vivo. Recently CRISPR was used to edit infected T-cells and eliminate HIV in mice (Stockton 2017).

More complicated diseases to edit with CRISPR involve cells that can self-replicate but may not be prime targets for ex vivo editing. For example, using CRISPR to study diseases in muscle tissue is more difficult than blood cells (Regalado 2016). Studies are underway to treat Duchenne muscular dystrophy (DMD) (LaFountaine et al. 2015), but edits must be made to all damaged cells, which cannot be done effectively ex vivo. Treating DMD requires a delivery mechanism that targets the affected cells, is large enough to deliver the CRISPR system in the organism, and does not make the individual sicker.

Some of the most difficult diseases to study and treat with CRISPR are those involving cells that do not replicate and cannot be edited ex vivo, such as diseases in the brain or nervous system (Regalado 2016, Kaiser 2016, LaFountaine et al. 2015). Brain cells and nerve cells are generally difficult to manipulate in a lab setting and because their interconnections matter, studies generally are conducted in vivo.[6] Again, this results in a delivery problem, where the engineered CRISPR tools are too large for standard delivery mechanisms. Work is being conducted to shrink the size of the cutting enzymes and to find alternative delivery mechanisms (Wang 2016), but the additional limitations have delayed research in these areas. For example, CRISPR has only appeared in Huntington's Disease publications since 2017.

There are already a number of reported possible therapeutic applications of CRISPR (LaFountaine et al. 2015) including cystic fibrosis (2013), HIV-1 (2013), sickle cell anemia (2014), Hepatitis B (2014), Duchenne muscular dystrophy (2014), HPV (2014), and a range of cancers. This suggests that applications where CRISPR is easier to use generally gain access to CRISPR sooner. Although no clinical trials with CRISPR were approved in the U.S. until 2018.

4 Methodology

4.1 Using the CRISPR Setting

As discussed in Section 2, external tool specialists can add value to the innovation process by sharing information about how to use the new technology across a range of domains. The match between internal domain teams and external tool specialists to create new innovations with the tool could occur more frequently in easier domains, where less complementary tool know-how is needed, in order to apply the tool more broadly or in more difficult domains, where more complementary tool know-how is needed, where the problems are more complex but influential. However,

[6] In *in vivo* (interior) gene editing, target cells are modified while still inside the living organism.

empirically identifying where external tool experts contributed most often to the application of a new tool, beyond just access, is a complex task as tool and domain knowledge often develop concurrently or access is conflated with the ability to use the tool (e.g., Furman and Stern 2011; Murray et al. 2016; Teodoridis 2018).

In order to identify where external tool specialists appear more often to create early innovations with a tool, an ideal setting would need to have several main features. First, it must be possible to separately track external tool specialists and domain specialists within teams that generate early innovations. Second, the sharing of know-how by external tool specialists must be separable from any access to the tool they may provide in order to identify their value-added contributions. Finally, when the new tool is introduced, the difficulty of using the tool must vary across domains and the tool should randomly enter different domains to mitigate selection based on the value of the tool to the domain. Such an environment does not occur naturally, but the recent introduction of CRISPR provides a unique and approximate setting for this ideal.

As discussed in Section 3, CRISPR was an unexpected, powerful tool that had the potential to lower the costs of research and create new possibilities in gene editing. Because of its unexpected nature, gene editing scientists did not anticipate its arrival or the future advances the tool would eventually allow. CRISPR can be used in a wide range of organisms including bacteria, yeasts, plants, insects, and mammals. However, previous research has shown that scientists who conduct gene editing on organisms from different branches on the biological tree of life (e.g., bacteria versus mammal) have different uses for the tool and value CRISPR differently (Zyontz 2016). This argues for focusing on researchers most at risk for using CRISPR, namely scientists conducting research on mammalian cells. During the earliest period of CRISPR adoption (2012-2016), these scientists are mostly academic scientists publishing in academic journals.

To ensure that different domains do not adopt CRISPR only when it is most valuable, it is necessary to find areas of mammalian research where scientists want to use CRISPR immediately, but cannot for some biological reason. Fortunately, research in almost every DNA-altering human disease would benefit from CRISPR due to its improvements over the previous tools. CRISPR co-inventor Feng Zhang supported this claim in a recent interview saying, "There are about 6,000 or more diseases that are caused by faulty genes. The hope is that we will be able to address most if not all of them" (Whitaker 2018). However, not every disease received CRISPR at the same time. The timing of CRISPR's introduction to each disease was not anticipated due to natural delays caused by the biology of the cells affected by the disease, as highlighted in Section 3.[7]

By restricting attention to human diseases caused by mutated genes, including infections, viruses, and inheritable diseases, the setting provides a way to separately identify external tool

[7] Certain diseases attract more attention and funding which could mitigate delays in receiving CRISPR. However, the exact timing of CRISPR's arrival in these diseases still could not be anticipated *ex ante*.

specialists and domain specialists. CRISPR was introduced first as a tool in June 2012 with no specific human disease application. Since the first human disease application occurred in 2013, there was time for scientists to gain CRISPR-specific tool knowledge outside of any particular disease domain. Applying this initial delay and the unanticipated timing of CRISPR appearing in different human diseases, it is possible to identify which authors were specialists in CRISPR prior to publishing in the disease. Figure 2 provides an example of the natural delays in CRISPR entry for selected diseases from first introduction of the tool.

[Figure 2 here]

Finally, science settings are not always ideal for identifying the contributions of external specialists beyond providing access to the tool. Biological materials are often transferred from one lab to another using Material Transfer Agreements (MTAs) that can delay access or have been known to require co-authorships for the receipt of the materials (Walsh, Cohen, and Cho 2007; Strandberg 2010). This can confound the value added by external tool specialists in innovation. The CRISPR setting circumvents these concerns by having a biological resource center that breaks the link between contracting for access and needing to work with external specialists to use the tool. Addgene is the primary third-party distributor of CRISPR to academic researchers that eliminates the need for scientists to sign MTAs directly with other academics. Thus any external specialists observed entering a new domain in this setting primarily represent value added knowledge rather than the price of access to CRISPR.

4.2 Empirical Specifications

To explore how early teams acquire complementary know-how to use new tools for innovation, the empirical specifications in this paper test the relationship between the share of external CRISPR tool specialists authoring CRISPR papers in a set of disease domains and the difficulty of editing the cell targeted by the disease. The focal population consists of all successfully published CRISPR articles in a set of human diseases and their authors. The level of analysis is at the disease-quarter from Q1 2013, the first quarter a CRISPR disease application appeared to Q4 2016. The vast majority of the disease quarters only contain one CRISPR-disease paper, so the share of external CRISPR specialists can be interpreted as the relative participation of external CRISPR specialists on the team for one paper (or innovation). The specifications also consider the unanticipated timing of CRISPR's entry by disease.

The main specification tests the overall impact of target cell editing difficulty on the share of external CRISPR specialists who are authors on CRISPR publications in a disease. The model controls for the quarter of publication and the amount of time since the first CRISPR paper in the disease to account for factors specific to the quarter or the trend of additional papers. The model

also controls for the total number of publications in a disease quarter as a proxy for the attention and funding a disease may receive.

$$\textit{Share of External CRISPR Specialists}_{it} = \beta_0 + \beta_1 \textit{Edit Difficulty}_i + \beta_2 \textit{Total Disease Pubs}_{it} + \delta_t + \delta_{age} + \varepsilon_{it}$$

Share of External CRISPR Specialists$_{it}$	=	The number of CRISPR paper authors that are external CRISPR specialists divided by the total number of CRISPR paper authors by disease domain (i) and quarter (t).
Edit Difficulty$_i$	=	1 if the target cells of the disease (i) cannot be edited ex vivo (*No Ex Vivo*) or do not self-replicate (*No Cell Replication*); 0 otherwise. Two separate variables.
Total Disease Pubs$_{it}$	=	The total number of papers in disease domain (i) and quarter (t).
δ_t	=	Fixed effects for the quarter (t) of publication.
δ_{age}	=	Fixed effects for the difference between the focal quarter of publication (t) and the quarter of the first CRISPR publication in a disease domain (i).
ε_{it}	=	Error term.

In this model, the coefficient of interest is β_1, which is the relationship between disease cell editing difficulty and the share of external CRISPR specialists in teams that publish CRISPR papers in a disease. A negative coefficient would support the idea that internal domain teams and external tool specialists match more often in easier domains that can help increase productivity and encourage broader use of the tool. A positive coefficient would support the idea that internal domain teams and external tool specialists match more often in difficult domains to effectively address more complex but influential problems.

The second specification tests whether the share of external CRISPR specialists increases or decreases for subsequent innovations after the first in difficult diseases. It is similar to the first specification above, but adds an interaction term and disease (i) fixed effects.

$$\textit{Share of External CRISPR Specialists}_{it} = \gamma_0 + \sum_{age} \gamma_{age} \textit{Edit Difficulty}_i * \textit{Qtr from First Pub}_{age} + \gamma_2 \textit{Total Disease Pubs}_{it} + \delta_i + \delta_t + \delta_{age} + \varepsilon_{it}$$

Qtr from First Pub is the difference between the focal quarter of the publication (t) in a disease (i) and the first CRISPR publication quarter in the disease. The coefficients of interest in this model are γ_{age} which are the changes in the share of external CRISPR specialists authoring subsequent CRISPR papers in a disease for difficult to edit target cells. Positive coefficients indicate that a higher share of external CRISPR specialists participate in additional innovations in domains where CRISPR is more difficult to use, even after controlling for time effects, age effects, disease effects, and the attractiveness of the disease.

All specifications are run initially as OLS models since the outcomes have a continuous response between 0 and 1. However, because the outcome is fractional and has some weight on 0 and 1 values, simple linear models lead to predictions outside the possible range. To mitigate this concern, all specifications are also run as Generalized Linear Models (GLM) with binomial family and logit link. The latter specification is as outlined by Papke and Wooldridge (1996) who showed that quasi-maximum likelihood estimation (QMLE) for pooled fractional response models results in robust estimators. The direction and significance of the results are similar regardless of model used, although the coefficients have different interpretations.

5 Data and Measures

5.1 Database Construction

Because CRISPR can be traced to a handful of initial papers and because of the explosion of interest in the tool, it is possible to find the entire population of academic scientists using CRISPR in human diseases, and not just a sample. The database starts with all articles and authors in the U.S. National Library of Medicine's (NLM) PubMed database from 2007-2016, providing approximately five years before CRISPR and five years after. PubMed includes the NLM's MEDLINE journal citation database and contains 28 million citations for biomedical literature including the fields of biomedicine and health, making it a definitive source for original research in human diseases.

In order to identify the relevant disease domains used in this study the keywords assigned by the NLM to each paper were used. These keywords are the Medical Subject Headings (MeSH Terms), a controlled vocabulary that consistently classifies each document in PubMed. First, a list of terms was collected from Category C (Diseases) in the 2017 MeSH Tree. Only keywords describing human diseases caused by DNA mutations were used to define the domains at risk of using CRISPR, including terms for cancers, infections (such as HIV), and inheritable monogenic

diseases.[8] Next, all papers in PubMed from 2007-2016 containing the MeSH Terms for the at-risk diseases were identified. Papers were restricted to original scientific articles and do not include documents like reviews, news, or other non-experimental articles. From there, the database was further restricted to papers (and associated authors) that contained both a disease MeSH term and a CRISPR MeSH term.[9] The final database contains the CRISPR papers in 228 disease domains published between Q1 2013 – Q4 2016.[10] Because of the different CRISPR entry dates, some disease domains appear earlier than others for N = 442 disease-quarters in the database.

Within the 228 disease domains, there are 611 papers containing both disease and CRISPR MeSH Terms. For each author on the 611 joint papers, his or her publication history in a disease domain and CRISPR was constructed using the following procedure.[11] First for each disease domain, a sub-database at the author-paper level was constructed containing every person in PubMed that authored a paper in the focal disease or CRISPR (as defined by the MeSH Terms) from 2007 - 2016.

Second, within that sub-database, author names across papers were matched using full first names, last names, and middle initials. In order to mitigate false matches across papers, authors with last names in the top 5% of all author names had to have exact matches for the last name, first name, and middle initial in order to be considered the same person. Authors with last names in the lower 50% of all author names (very uncommon names) only had to have their last name and first initial match to be considered the same person. All others required an exact match of the full first name and last name to be considered the same person.[12]

Third, scientists were classified as CRISPR-disease paper authors if they authored one of the 611 papers that contained both disease and CRISPR MeSH Terms. Records from all other authors were dropped, leaving only the publication histories of the CRISPR-disease paper authors.

Fourth, using MeSH Terms, these remaining papers were classified as CRISPR Only (if the paper only had CRISPR MeSH Terms), Disease Only (if the paper only had Disease MeSH Terms), or CRISPR-disease (if the paper had CRISPR and Disease MeSH Terms). A CRISPR-disease author was classified as an external CRISPR specialist if he or she published a CRISPR Only paper first before any CRISPR-disease papers or Disease Only papers.[13]

Finally, authors with no identifiable publication history in CRISPR or the disease, usually graduate students or non-key contributors, were dropped from the sub-database. All non CRISPR-

[8] Monogenic diseases, although rare, affect a wide range of cell types and are identified by a well-known inherited single mutation in a gene, for example sickle cell anemia. These are prime targets for CRISPR since there is only one gene to edit and both the mutation and the correct sequence are known.

[9] Because CRISPR was unexpected, the MeSH Terms for the tool did not appear immediately. They were added to papers in 2013, however. A check for CRISPR in the abstracts did not reveal any earlier disease domains or papers.

[10] The list of disease domains and associated MeSH terms are on file with the author.

[11] Papers are generally exclusive to one disease category.

[12] The code for the grouping algorithm is on file with the author.

[13] The majority of authors in a sub-database only have one CRISPR paper in a disease domain.

disease papers were dropped as well to focus only on the published innovations with CRISPR in each disease domain.

This process was repeated for each disease domain for a total of 228 sub-databases. These were joined into one large author-paper-disease level database with 3,019 authors, 611 joint papers, and 228 disease domains. The final database used for the analyses collapses the data to the disease-quarter level by calculating the number of total authors, external CRISPR specialists, and papers present in each quarter for each disease domain from Q1 2013 – Q4 2016 in 442 observations. The majority of disease-quarters only contain one paper, so the final database is similar to one constructed at the paper level. More than half the disease domains only have one CRISPR-disease paper over the entire time period.

5.2 Measures

The main dependent variable is a measure of external tool specialist know-how that is used to generate innovations with the tool in a domain. Share of External CRISPR Specialists is calculated as the number of external CRISPR specialist authors on CRISPR-disease papers divided by the total number of authors on CRISPR-disease papers. External CRISPR specialists are those that published in CRISPR first before publishing in the disease or a CRISPR-disease paper. This is measured quarterly by disease domain but due to the small number of CRISPR-disease papers, almost all disease-quarters only contain one paper. Therefore, this measure is the relative participation of external CRISPR specialists on teams that create new papers (or innovations) with CRISPR in a disease.

The main independent *Edit Difficulty* measures are binary and are different indicators of how difficult affected cells are to edit in each disease domain. The difficulty of cell editing can be measured by biological factors of the cells primarily targeted by each disease. Two key factors are (1) whether the cell can be edited *ex vivo* (edited outside a living organism and placed back in) and (2) whether the cell can self-replicate. If the cell a disease targets cannot be edited ex vivo or if the cell does not self-replicate, then it will be far more difficult to use CRISPR in that disease (see Section 3). The variable, *No Ex Vivo* is equal to 1 if the target cells cannot be edited ex vivo and 0 otherwise. The variable *No Cell Replication* is equal to 1 if the target cells do not replicate easily on their own and 0 otherwise.

To code these two variables, for each disease domain the primary target cell category was determined using information from a number of sources including the Online Mendelian Inheritance in Man database (OMIM), the Genetic and Rare Disease Information Center (GARD), and a set of gene editing review articles (LaFountaine et al. 2015; Barrangou and Doudna 2015; Cox et al. 2015; Kelton et al. 2016; Riordan et al. 2016; Scott and DeFrancesco 2016; Wang et al. 2016; Xiong et al. 2016; Bachtarzi 2017; Pandey et al. 2017; Singh et al. 2017; Song et al. 2017; Bakhrebah et al.

2018).[14] Then for each target cell category *No Ex Vivo* (National Academies of Science 2018; Cox et al. 2015) and *No Cell Replication* (Weizmann Institute of Science 2018) were coded. For example, diseases that target blood or T-cells have *No Ex Vivo* and *No Cell Replication* both equal to 0 since they are easiest to edit. Diseases that target muscle tissues have *No Ex Vivo* equal to 1 and No Cell Replication equal to 0 since effective editing is done within the organism. Diseases that target neurons have *No Ex Vivo* and *No Cell Replication* both equal to 1 since they are hardest to edit.

The independent variable *Quarters from First Pub* is the difference in quarters from the publication of the first CRISPR paper in a disease domain to the publication quarter of the focal CRISPR paper in the disease. The publications quarters for CRISPR-disease papers range from Q1 2013 – Q4 2016 since there are no human disease applications for CRISPR prior to 2013.

Finally, the independent variable *Total Disease Pubs* is a proxy for the attractiveness of the disease domain in terms of attention and possible funding. It is calculated as the total number of academic articles by disease domain (not including reviews, news, and other similar documents) published from Q1 2013 through Q4 2016. Articles are considered part of a disease domain if they contain MeSH Terms for those diseases.

5.3 Summary Statistics

The summary statistics for the CRISPR-disease papers and authors in the 228 disease domains by quarter is in Table 1. Table 1 presents the statistics for all diseases and further breaks down the results by Easy Diseases and Difficult Diseases as defined by whether the target cell can be edited ex vivo or not. On average by quarter, about 25% of the authors are external CRISPR specialists reinforcing the idea that domain-specific knowledge is important to the production of innovation in the domain but indicating that external CRISPR know-how plays a meaningful role across domains. The share of external CRISPR specialists is higher for teams writing CRISPR papers in difficult diseases at 33% versus 22% for easier diseases.

The average team size on all CRISPR-disease papers is just under 7 people. Yet the make-up of the team depends on the edit difficulty of the disease. For example, CRISPR papers in more difficult diseases have more CRISPR tool specialists but also have fewer people overall, making the individual contributions of external CRISPR specialists larger.

Further, a minority of diseases by quarter are those where CRISPR is more difficult to use. For example, 31% of diseases in a quarter have target cells that cannot be edited ex vivo and 17% have target cells that do not self-replicate. On average, CRISPR-disease papers are published in Q4 2015, supporting the fact that a number of diseases did not receive CRISPR until 2016, especially

[14] The associated cell category for each disease domain and the coding for *No Ex Vivo* and *No Cell Replication* are on file with the author.

difficult diseases. Also, additional papers generally occur more than six months after the first CRISPR-disease paper.

[Table 1 here]

Table 2 provides the raw counts of authors and papers by the difficulty of cell editing (No Ex Vivo) and year of CRISPR-disease paper publication. The results suggest that there was a delay in attracting authors and publications in the more difficult diseases as compared to their easier counterparts. However, the share of external CRISPR specialists on teams publishing in difficult diseases quickly outpaced the share in easier diseases.

[Table 2 here]

This shift in more difficult diseases for authors and publications is due in part to the delayed entry of CRISPR into difficult to edit diseases. Figure 3a plots the number of diseases by the year CRISPR entered and by whether the target cell can be edited ex vivo. Figure 3b plots the number of diseases by the year CRISPR entered and by whether the target cell can self-replicate. For diseases where the target cells are more difficult to edit, CRISPR took additional time to enter as compared to the easier to edit diseases. Although, diseases that cannot be edited ex vivo overtook easier to edit diseases in 2016.

[Figures 3a and 3b here]

6 Results

6.1 Disease Edit Difficulty and Share of External CRISPR Specialists

As a way to explore where the match occurs between internal domain teams and external tool specialists for the earliest innovations, it is possible to test the relationship between disease edit difficulty and share of external tool specialists. If domain teams and external tool specialists match most often to solve simpler problems where the tool can be applied more quickly to a larger range of domains, then the share of external CRISPR specialists on CRISPR paper teams should decrease for hard to edit diseases on average. If internal domain teams and external tool specialists match most often instead to solve complex and influential problems, the share of external CRISPR specialists on CRISPR paper teams should increase for hard to edit diseases on average.

First, to establish that external tool specialists are used across domains, Figure 4 plots the distribution of the dependent variable, Share of External CRISPR Specialists, for publishing teams

by disease edit difficulty (here where the affected cell can be edited ex vivo). External tool specialists are part of teams in domains that are both easier and difficult to edit, but the distribution is shifted towards a higher share of external CRISPR specialists for more difficult to edit domains.

[Figure 4 here]

Second, it is important to note that there are an increasing number of difficult to edit diseases with CRISPR availability and more available external authors with CRISPR know-how over time. Due to these facts, it should be expected that the share of external CRISPR specialists should increase over time for all diseases as the stock of CRISPR specialists increases each quarter. Figure 5a shows the increase in the number of authors over time for both easier to edit and difficult to edit diseases. Figure 5b shows similar trends for the number of external tool specialists on the research teams over time.

[Figures 5a and 5b here]

Figure 1A shows the relationship between the timing of publications and the share of external CRISPR specialists. Between Q4 2014 and Q4 2016 the share of external CRISPR specialists increases each year with some cyclicality over quarters.

Using the above facts, Table 3, Models 1-3 show fixed effect OLS regressions at the disease-quarter level to look at the effect of disease edit difficulty on the share of external CRISPR specialists contributing to new CRISPR-disease papers. From Table 3, Model 1, diseases with target cells that cannot be edited ex vivo have an increased share of external CRISPR specialists. The direction is different for diseases with target cells that cannot self-replicate, although the estimate is much lower and is not significant (Table 3, Model 2). Taken together the different types of edit difficulty do play different roles in affecting the percentage of external CRISPR authors (Table 3, Model 3). Here, the effect of being a No Ex Vivo disease is enhanced by controlling for being a No Cell Replication disease.

The first three models in Table 3 run OLS specifications and control for the stock of external CRISPR specialists through the quarter of publication, the number of quarters the publication is from the first CRISPR paper in a disease, and the attractiveness of the disease domain. Table 3, Models 4 – 6 have the same covariates and dependent variables but use a GLM with binomial family and logit link to account for the fractional outcome as discussed in Section 4. Although the coefficients have a different interpretation, as expected, their direction and significance are largely the same. The marginal effects of these coefficients are very similar to those of the standard linear models.

[Table 3 here]

Overall, the results support the idea that the successful match between internal domain teams and external tool specialists to create new innovations occurs in the difficult diseases where the problems are complex but solutions are highly influential. These are the diseases that had no prior viable gene therapy alternatives.

6.2 Share of External CRISPR Specialists in Subsequent Innovations

If the increase in the share of external CRISPR specialists was a simple transfer of tacit information, it might be expected that the effect would attenuate for CRISPR-disease papers after the first. Once the know-how is in the domain, it should not be necessary to have external CRISPR specialists on the team. However, contrary to this expectation, the share of external CRISPR specialists increases for subsequent innovations in more difficult diseases.

Figure 6 shows the differences in means for the share of external CRISPR specialists on CRISPR-disease papers for diseases that are more difficult to edit and those that are easier (by ex vivo editing), over specific periods of time. For example, the first pair of bars illustrates the difference in the average percentage of external CRISPR specialists between diseases where target cells can be edited ex vivo and not for the very first paper. CRISPR papers in difficult diseases (no ex vivo editing) have a higher share of external CRISPR specialists when only considering the first CRISPR paper published in a disease. The next pairs show the difference in means for additional CRISPR papers in a disease after the first one (i.e., excluding the first) in the first year, in the next year after, and in the remaining years. More difficult diseases have higher average shares of external CRISPR specialists, increasing over each time period. Table A1 provides the underlying data for Figure 6.

[Figure 6 here]

Because Figure 6 highlights the results from simple T-tests, Table 4a considers the interaction between subsequent innovations after the first and the edit difficulty of the disease while controlling for disease attractiveness as well as quarter, age, and disease fixed effects. The results in Table 4a echo the trends found in Figure 6. Table 4a, Model 1 includes an interaction term between the time distance of an additional innovation in a disease and disease edit difficulty (using the No Ex Vivo measure). The results suggest that for diseases that are more difficult to edit and as additional innovation is more time-distant from the original joint paper, the share of authors with external CRISPR specialization increases.

Table 4a, Model 2 breaks these interactions down further by the length of time the innovation appeared after the first. As more time passes after the first paper, the share of external CRISPR specialists continues to increase during this very early stage. Whether this pattern will continue as the tool matures is the subject of a future study as more time is allowed to pass. Table 4a, Models 3-4 show results in the same direction and significance levels for the analogous GLMs. Figure 7 plots the coefficients in Table 4a, Model 2 to illustrate the increasing trend in the share of external CRISPR specialists for subsequent innovations.

[Table 4a and Figure 7 here]

Because some diseases only have one CRISPR paper, it might be a concern that these diseases are different and influencing the results in Table 4a. In order to account for that, Table 4b performs an identical analysis but only for the 93 diseases that have more than one CRISPR paper in the dataset. The results are very similar, so including the entire dataset is not influencing the core findings for the OLS or GLM specifications.

[Table 4b here]

6.3 Future Research

The results in Sections 6.1 and 6.2 are the first steps towards empirically understanding how the earliest teams acquire complementary tool know-how across domains with different levels of ex ante difficulty. However, the analysis presented suggests several other areas of inquiry that can help to further this understanding. For example, are there characteristics of external CRISPR specialists that attract some to difficult problems and others to easier ones? As more data becomes available, it will be possible to exploit heterogeneity in adopters that does not currently exist in the data to test whether the tenure status of external CRISPR specialists plays a role in their attraction to more complex and influential problems. It could be the case that tenured external CRISPR specialists have more incentives to aim for the Nobel Prize while non-tenured external CRISPR specialists just need to publish. This could cause tenured external CRISPR specialists to seek out and match in more difficult domains.

Other analyses are aimed at assessing the benefits to building versus acquiring the complementary tool know-how. Specifically, the analyses address the questions do teams that use external CRISPR specialists get to publication faster? Does this change given domain difficulty? For this, team members on CRISPR-disease papers are being matched to their CRISPR order histories at Addgene to determine the earliest date the team received CRISPR and the length of time between the first order and the eventual publication.

7 Discussion and Conclusion

When a new tool is introduced that requires some investment in tool-specific know-how, how to combine that know-how with domain-specific knowledge is a decision firms, mangers, and individual innovators face as they innovate. In general, innovators in a domain can either learn to use the tool themselves or can acquire the complementary tool know-how from an external source. For example, firms face this choice when deciding how to use AI and individual academics may make similar decisions when deciding how to use STATA or Python in their work. However, when external tool specialists are scarce, they also have a choice over which teams to join. They could collaborate with teams in easy domains to expand the use of the tool as far as possible or in difficult domains where the problems are more complex but the solutions are highly influential.

Previous research has shown that access to research tools increases innovation but often access is conflated with the ability to use the tool, limiting investigations into how innovators use tools to generate innovations, especially if they are trying to be first. One mechanism this paper highlights is the role of external tool specialists in generating early innovation by providing complementary know-how needed to use the tool.

Using the introduction of the new breakthrough DNA-editing tool, CRISPR, and applying the unanticipated timing of CRISPR entry in different human disease domains, it is possible to separately identify the knowledge bases of the academic scientists responsible for the new articles that use CRISPR in a disease. Because the analysis focuses on the first days of CRISPR (2012-2016), these papers represent the earliest innovations in each domain with the tool. It is also possible to create measures for the difficulty of applying CRISPR in the different disease domains. Human diseases primarily target certain cell categories and each category is easier or harder to edit based on biological factors specific to the cells. For example, CRISPR is more difficult to apply in target cells that cannot be edited ex vivo or cannot self-replicate.

The variation in the difficulty of applying CRISPR provides a novel lever to presents a direct mechanism for how new tools are adopted and incorporated into early innovations. First, a higher share of external CRISPR tool specialists participate in early innovations with the tool in difficult disease domains. This suggests the match between internal domain teams and external tool specialists occurs more often in domains with complex and influential problems. To understand this result, consider the different experiences of CRISPR adoption in HIV and muscular dystrophy. The human immunodeficiency virus (HIV) had one of the earliest introductions of CRISPR in part due to the targeted T cells being easy to edit and the large amount of previous research conducted on gene editing alternatives. The first study using CRISPR to make new advances in HIV was published in 2013 by Yoshio Koyanagi, the PI of a Viral Pathogenesis lab at Kyoto University in Japan (Ebina et al. 2013). He and his three co-authors, all part of the lab, had previous experience in HIV, but none specifically in CRISPR. Although external CRISPR know-how would be useful

in this case, the internal team ordered CRISPR from Addgene and modified it for their application but did not collaborate on the paper with an external CRISPR tool specialist outside of the domain. In contrast, muscular dystrophy targets muscle cells that are more challenging to edit and success would represent an enormous advance in medicine. For the first paper that successfully used CRISPR inside a mouse to treat Duchenne muscular dystrophy, Charles Gersbach, a leading muscular dystrophy researcher at Duke University and his lab were having problems delivering CRISPR to the nucleus of the target cells. To overcome this problem, the team incorporated the knowledge of Feng Zhang, a CRISPR co-founder, to create a new delivery solution. The resulting paper lists both professors as contributing authors (Nelson et al. 2016; Duke Today Staff 2015). The example suggests that authors in internal teams working on difficult and influential diseases are more likely to look for and attract external CRISPR specialists to collaborate.

Second, there is evidence that the higher share of external CRISPR specialists persists for subsequent innovations in more difficult disease domains. This result is not immediately intuitive. If research in the domain using CRISPR became easier after the tool was first introduced, then complementary know-how about the tool should be less valuable and external tool specialists less necessary. However, in this setting, the ultimate goal is to use CRISPR to create commercial therapies and drugs for human use. In order to do this, additional research in each disease will try to use the tool in increasingly complex organisms. Even within mammals, as the organism gets closer to humans, editing becomes more difficult and the solutions more notable, attracting a higher share of external CRISPR specialists. As an example, one of the first CRISPR experiments in muscular dystrophy was to deliver CRISPR inside a living mouse (Nelson et al. 2016). The next step was to deliver CRISPR inside living dogs (Amoasil et al. 2018). Note that the author teams for each organism are different. This pattern of research is repeated in other diseases with more difficult to edit target cells.

Although the current study was conducted in the context of CRISPR, the overarching findings have implications for firms and individuals thinking about when and how to adopt new tools for innovation. Variations in the ex ante domain difficulty and solution novelty is not unique to CRISPR and research teams looking to be early adopters in tools like AI will have to weigh the complexity and influence of their goals when considering how best to attract and collaborate with external tool specialists. In order to be first to innovate with a new tool, external specialists are not always necessary to acquire complementary tool know-how. However, external tool specialists may be more likely to find successful matches with internal domain teams that focus on more complex problems with highly influential solutions.

The findings contribute to the literature on innovative teams and team structure by uniquely showing that not just features of management, organizational structure, or industry are important for effective team design. Team composition is also driven by the specific nature of the problem to be solved and the nature of tools available for innovation. This paper is one of the

first in a series that uses CRISPR to answer key questions in innovation, management, and economics. For example, future papers can build on the results established here to study the effect of breakthrough technologies on academic entrepreneurship, the impact of policies regarding genetically modified organisms on agricultural product development, the effect of unresolved intellectual property disputes on scientific innovation, and how to incentivize ethical innovation without stifling important technological advances.

8 References

Adams, J. D., G. C. Black, J. R. Clemmons, and P. E. Stephan. 2005. "Scientific Teams and Institutional Collaborations: Evidence from US Universities, 1981–1999." *Research Policy* 34(3): 259–85.

Ahuja, G. and R. Katila. 2001. "Technological acquisitions and the innovation performance of acquiring firms: A longitudinal study." *Strategic Management Journal* 22(3): 197-220.

Amoasi, L., J. Hildyard, H. Li, E. Sanchez-Ortiz, A. Mireault, D. Caballero, R. Harron, T. Stathopoulou, C. Massey, J. Shelton, R. Bassel-Duby, R. Piercy, and E. Olson. 2018. "Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy." *Science* 362(6410):86-91.

Arora, A., and A. Gambardella. 1994. "The changing technology of technological change: general and abstract knowledge and the division of innovative labour." *Research Policy* 23(5): 523-532.

Azoulay, P., W. Ding, and T. Stuart. 2009. "The Effect of Academic Patenting on the Rate, Quality, and Direction of (Public) Research Output." *Journal of Industrial Economics* 57(4): 637-676.

Bachtarzi, H. 2017. "Ex vivo and in vivo genome editing: a regulatory scientific framework from early development to clinical implementation." *Regenerative Medicine* 12(8): 1015-1030.

Bakhrebah, M.A., M.S. Nassar, M.S. Alsuabeyl, W.A. Zaher, and S.A. Meo. 2018. "CRISPR technology: new paradigm to target the infectious disease pathogens." *European Review for Medical and Pharmacological Sciences* 22: 3448-3452.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, P. Horvath. 2007. "CRISPR provides acquired resistance against viruses in prokaryotes." *Science* 315(5819): 1709–1712.

Barrangou, R. and J. Doudna. 2016. "Applications of CRISPR technologies in research and beyond." *Nature Biotechnology* 34(9): 933-941.

Bikard, M., F. Murray, and J. Gans. 2015. "Exploring Tradeoffs in the Organization of Scientific Work: Collaboration and Scientific Reward." *Management Science* 61(7): 1473 -1495.

Boch, J., H. Scholze, S. Schornack, A. Landgraf, S. Hahn, S. Kay, T. Lahaye, A. Nickstadt, and U. Bonas. 2009. "Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors." *Science* 326(5959): 1509-1512.

Bresnahan, T. F., and M. Trajtenberg. 1995. "General Purpose Technologies: Engines of Growth?" *Journal of Econometrics* 65(1): 83-108.

Cassiman, B. and R. Veugelers. 2006. "In Search of Complementarity in Innovation Strategy: Internal R&D and External Knowledge Acquisition." *Management Science* 52: 68-82.

Coase, R.A. 1937. "The nature of the firm." *Economica* 16(4): 386–405.

Cockburn, I.A., R. Henderson, and S. Stern. 2017. "The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis" in *The Economics of Artificial Intelligence: An Agenda*, A. K. Agrawal, J. Gans, and A. Goldfarb, eds. Chicago University Press. Forthcoming.

Cohen, W. M. and D.A. Levinthal. 1989. "Innovation and learning: the two faces of R & D." *The Economic Journal* 99(397): 569-596.

Cohen, W. M. and D.A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation." *Administrative Science Quarterly* 128-152.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, and F. Zhang. 2013. "Multiplex genome engineering using CRISPR/Cas systems." *Science* 339(6121): 819-823.

Cox, D., R. Platt, and F. Zhang. 2015. "Therapeutic Genome Editing: Prospects and Challenges" *Nature Medicine* 21(2): 121–131.

Cummings, J. N. and S. Kiesler. 2007. "Coordination costs and project outcomes in multiuniversity collaborations." *Research Policy* 36(10): 1620-163.

Cyranoski, D. 2018. "CRISPR-baby scientist fails to satisfy critics." *Nature* 564: 13-14.

Dallon, A. 2017. "Everything you need to know about Waymo's self-driving car project." *Digital Trends* (April 26) available at: https://waymo.com/journey/.

David, P. 1990. "The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox." *American Economic Review* 80(2): 355-36.

Doudna, J. 2015. "Genomic engineering and the future of medicine." *JAMA* 313(8):791-792.

Doudna, J. and S. Sternberg. 2017. *A Crack in Creation: Gene Editing and the Unthinkable Power to Control Evolution.* Houghton Mifflin Harcourt.

Duke Today Staff. 2015. "CRISPR Treats Genetic Disorder in Adult Mammal." *Duke Today* (December 31) available at: https://today.duke.edu/2015/12/crisprmousedmd.

Ebina, H., N. Misawa, Y. Kanemura, and Y. Koyanagi. 2013. "Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus." *Scientific Reports* 3: 2510.

Fan, M., J. Tsai, B. Chen, K. Fan, and J. LaBaer. 2005. "A central repository for published plasmids." *Science* 307 (5717): 1877.

Fleming, L. 2001. "Recombinant Uncertainty in Technological Search." *Management Science* 47: 117-132.

Furman, J., and S. Stern. 2011. "Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production." *American Economic Review* 101(5): 1933-1963.

Furman, J. and F. Teodoridis. 2018. "The Cost of Research Tools and the Direction of Innovation: Evidence from Computer Science and Electrical Engineering." Working paper.

Gale, D. and L. Shapley. 1962. "College Admissions and the Stability of Marriage," *American Mathematical Monthly* 69: 9-15.

Gans, J. and F. Murray. 2014. "Markets for Scientific Attribution." NBER Working Paper 20677.

Gantz, V., N. Jasinskiene, O. Tatarenkova, A. Fazekas, V. Macias, E. Bier, and A. James. 2015. "Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi." *PNAS* 112(49): E6736–E6743.

GARD. "Genetic and Rare Diseases Information Center" available at: https://rarediseases.info.nih.gov/.

Grigoriou, K. and F. Rothaermel. 2017. "Organizing for Knowledge Generation: Internal Knowledge Networks and the Contingent Effect of External Knowledge Sourcing." *Strategic Management Journal* 38: 395-414.

Griliches, Z. 1957. "Hybrid corn: An exploration in the economics of technological change." *Econometrica* 25(4):501-522.

Griliches, Z. 1958. "Research Costs and Social Returns: Hybrid Corn and Related Innovations." *Journal of Political Economy* 66(5): 419-431.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. "A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity." *Science* 337(6096): 816-21.

Jones, B. F. 2009. "The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder?" *Review of Economic Studies* 76(1): 283-31.

Kaiser, J. 2016. "The Gene Editor CRISPR won't fully fix sick people anytime soon. Here's why." *Science* (May 3).

Kaplan, S. and K. Vakili. 2015. "The Double-edged Sword of Recombination in Breakthrough Innovation." *Strategic Management Journal* 36: 1435-1457.

Kelton, W., T. Pesch, S. Matile, and S. Reddy. 2016. "Surveying the Delivery Methods of CRISPR/Cas9 for ex vivo Mammalian Cell Engineering." *CHIMIA* 70(6): 439-442.

LaFountaine, J., K. Fathe, and H. Smyth. 2015. "Delivery and therapeutic applications of gene editing technologies ZFNs, TALENs, and CRISPR/Cas9." *International Journal of Pharmaceutics* 494: 180–194.

Ledford, H. 2015. "CRISPR: The disruptor." *Nature* 522(7554): 20-24.

Ma, H., N. Marti-Gutierrez, S. W. Park, J. Wu, Y. Lee, K. Suzuki, A. Koski, D. Ji, T. Hayama, R. Ahmed, H. Darby, C. Dyken, Y. Li, E. Kang, A. R. Park, D. Kim, S. T. Kim, J. Gong, Y. Gu, X. Xu, D. Battaglia, S. Krieg, D. Lee, D. Wu, D. Wolf, S. Heitner, J. C. Belmonte, P. Amato, J. S. Kim, S. Kaul, and S. Mitalipov. 2017. "Correction of a pathogenic gene mutation in human embryos." *Nature* 548(7668): 413-419.

Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. "RNA-guided human genome engineering via Cas9." *Science* 339(6121): 823-6.

Marr, B. 2016. "The Top 10 AI And Machine Learning Use Cases Everyone Should Know About." *Forbes* (September 30).

Method. 2012. "Method of the year 2011." *Nature Methods.* 9: 1.

MeSH Terms. 2017. PubMed Database, MeSH Tree, available at: ftp://nlmpubs.nlm.nih.gov/online/mesh/2017/ .

Mokyr, J. 2002. *Gifts of Athena: Historical Origins of the Knowledge Economy.* Princeton University Press.

Moscou, M. and A. J. Bogdanove. 2009. "A Simple Cipher Governs DNA Recognition by TAL Effectors." *Science* 326(5959):1501.

Moser, P. 2005. "How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World Fairs." *American Economic Review* 95(4): 1214-1236.

Mowery, D.C. and A.A. Ziedonis. 2007. "Academic patents and materials transfer agreements: substitutes or complements?" *Journal of Technology Transfer* 32(3): 157–172.

Murray, F. 2002. "Innovation as co-evolution of scientific and technological networks: exploring tissue engineering." *Research Policy* 31(8-9): 1389-1403.

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern. 2016. "Of mice and academics: Examining the effect of openness on innovation." *American Economic Journal: Economic Policy* 8(1): 212-252.

Nagle, F. and F. Teodoridis. 2017. "Jack of All Trades and Master of Knowledge: The Role of Generalists in Novel Knowledge Integration." USC Dornsife Institute for New Economics Thinking, Working Paper No. 17-23.

National Academies of Science. 2018. *Human Genome Editing: Science, Ethics, and Governance*, available at: https://www.nap.edu/read/24623/chapter/6#84.

Nelson, C., C. Hakim, D. Ousterout, P. Thakore, E. Moreb, R. Castellanos Rivera, S. Madhavan, X. Pan, F. Ran, W. Yan, A. Asokan, F. Zhang, D. Duan, and C. Gersbach. 2016. "In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy." *Science* 351(6271):403-7.

Nelson, R.R. 1981. "Research on productivity growth and productivity differences: Dead ends and new departures." *Journal of Economic Literature* 19(3): 1029–1064.

Nelson, R.R. 2003. "On the uneven evolution of human know-how." *Research Policy* 32(6): 909–922.

OMIM. Online Mendelian Inheritance in Man Database, available at: https://www.omim.org/.

Pandey, V., A. Tripathi, R. Bhushan, A. Ali, and P. Dubey. 2017. "Application of CRISPR/Cas9 Genome Editing in Genetic Disorders: A Systematic Review Up to Date." *Journal of Genetic Syndromes and Gene Therapy* 8(2): 321-331.

Papke, L.E., and J.M. Wooldridge. 1996. "Econometric methods for fractional response variables with an application to 401(k) plan participation rates." *Journal of Applied Econometrics* 11: 619–632.

Pennisi, E. 2013. "The CRISPR craze." *Science* 341(6148): 833-836.

Pinzon, E. 2015. *Thirty Years with Stata: A Retrospective*. Stata Press.

Platt, R.J., S. Chen, Y. Zhou, M. J. Yim, L. Swiech, H. R. Kempton, J. E. Dahlman, O. Parnas, T. M. Eisenhaure, M. Jovanovic, D. B. Graham, S. Jhunjhunwala, M. Heidenreich, R. J. Xavier, R. Langer, D. G. Anderson, N. Hacohen, A. Regev, G. Feng, P. A. Sharp, and F. Zhang. 2014. "CRISPR-Cas9 knockin mice for genome editing and cancer modeling." *Cell.* 159(2):440-55.

Polanyi, M. 1962. "Tacit knowing: Its bearing on some problems of philosophy." *Reviews of Modern Physics* 34(4): 601.

PubMed. 2017-2018, available at: https://www.ncbi.nlm.nih.gov/pubmed.

Rabinow, P. 2011. *Making PCR: A Story of Biotechnology.* University of Chicago Press.

Regalado, A. 2014. "Who owns the biggest biotech discovery of the century?" *MIT Technology Review* (December 4).

Regalado, A. 2016. "Can CRISPR Save Ben Dupree?" *MIT Technology Review*, 119(8): 81-87.

Riordan, S., D. Heruth, L.Q. Zhang, and S.Q. Ye. 2015. "Application of CRISPR/Cas9 for biomedical discoveries." *Cell and Bioscience* 5:33-44.

Rogers, E. 1995[1962] *Diffusion of Innovations.* 4 ed. The Free Press: Simon & Schuster.

Rosenberg, N. 1982. *Inside the Black Box: Technology and Economics.* Cambridge University Press.

Rosenberg, N. 1994. *Exploring the Black Box: Technology, Economics, and History.* Cambridge University Press.

Rosenberg, N. 2009. "Some critical episodes in the progress of medical innovation: An Anglo-American perspective." *Research Policy* 3(2): 234–242.

Rosenberg, N. and M. Trajtenberg. 2004. "A General-Purpose Technology at Work: The Corliss Steam Engine in the Late-Nineteenth-Century United States." *Journal of Economic History* 64(1): 61-99.

Roth, A. E. 1984. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy* 92:991-1016.

Sampat, B., and H. L. Williams. 2019. "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome." *American Economic Review* 109 (1): 203-36.

Schilling, M. A. and E. Green. 2011. "Recombinant search and breakthrough idea generation: An analysis of high impact papers in the social sciences." *Research Policy* 40(10): 1321-1331.

Scotchmer, S. 1991. "Standing on the Shoulders of Giants: Cumulative Research and the Patent Law." *Journal of Economic Perspectives* 5(1): 29-41.

Scott, C. T. and L. DeFrancesco. 2016. "Gene therapy's out-of-body experience." *Nature Biotechnology* 34(6): 600-607.

Singh, V., N. Gohil, R.R. Garcia, D. Braddick, and C. K. Fofie. 2018. "Recent advances in CRISPR-Cas9 genome editing technology for biological and biomedical investigations." *Journal of Cellular Biochemistry* 119: 81-94.

Song, M. 2017. "The CRISPR/Cas9 System: Their Delivery, In Vivo and Ex Vivo Applications and Clinical Development by Startups." *Biotechnology Progress* 33(4): 1035-1045.

Specter, M. 2015. "The gene hackers: A powerful new technology enables us to manipulate our DNA more easily than ever before." *The New Yorker* (16 November).

Stephan, P. 2012. *How Economics Shapes Science.* Harvard University Press.

Stockton, N. 2017. "CRISPR kills HIV and eats Zia 'like Pac-man'. It's Next Target? Cancer." *WIRED.*

Strandberg, K. 2010. "Norms and the Sharing of Research Materials and Tacit Knowledge" in *Working Within the Boundaries of Intellectual Property: Innovation Policy For The Knowledge Society.* R. C. Dreyfuss, D. L. Zimmerman, and H. First eds. Oxford University Press.

Teece, D. J. 1986. "Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy." *Research Policy* 15(6): 285-305.

Teodoridis, F. 2018. "Understanding Team Knowledge Production: The Interrelated Roles of Technology and Expertise." *Management Science* 64(8): 3469-3970.

Thompson, N. and S. Zyontz. 2017. "Who tries (and who succeeds) in staying at the forefront of science: Evidence from the DNA-editing technology, CRISPR." Working paper.

Uzzi, B., S. Mukherjee, M. Stringer, and B. Jones. 2013. "Atypical combinations and scientific impact." *Science* 342(6157): 468-472.

Walsh, J., A. Arora, and W. Cohen. 2003. "Effects of Research Tool Patents and Licensing on Biomedical Innovation" in *Patents in the knowledge-based economy.* W.M. Cohen and S. A. Merrill eds. National Academies Press.

Walsh, J.P., W.M. Cohen, and C. Cho. 2007. "Where excludability matters: Material versus intellectual property in academic biomedical research." *Research Policy* 36(8):1184–1203.

Wang, L., F. Li, L. Dang, C. Liang, C. Wang, B. He, J. Liu, D. Li, X. Wu, X. Xu, A. Lu, and G. Zhang. 2016. "In Vivo Delivery Systems for Therapeutic Genome Editing." *International Journal of Molecular Science* 17:626-645.

Wang, Y., X. Cheng, Q. Shan, Y. Zhang, J. Liu, C. Gao, and J. L. Qiu. 2014. "Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew." *Nature Biotechnology* 32(9): 947–951.

Weitzman, M. 1996. "Hybridizing Growth Theory." *American Economic Review* 86: 207-212.

Weitzman, M. 1998. "Recombinant Growth." *Quarterly Journal of Economics* 113(2):331-360.

Weizmann Institute of Science. 2018. "Bionumbers" available at: http://bionumbers.hms.harvard.edu/search.aspx AND http://book.bionumbers.org/how-quickly-do-different-cells-in-the-body-replace-themselves/.

Whitaker, B. 2018. "CRISPR: The Gene Editing Tool Revolutionizing Biomedical Research." *60 Minutes*, available at: https://www.cbsnews.com/news/crispr-the-gene-editing-tool-revolutionizing-biomedical-research/.

Williams, H. 2013. "Intellectual Property Rights and Innovation: Evidence from the Human Genome." *Journal of Political Economy* 121(1): 1-27.

Williamson, O. 1975. *Markets and Hierarchies.* Free Press.

Williamson, O. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting.* Free Press.

Wills, M. 2018. "The Evolution of the Microscope." *JSTOR Daily* (March 27).

Wuchty, S., B. Jones, and B. Uzzi. 2007. "The Increasing Dominance of Teams in Production of Knowledge." *Science* 316(5827):1036-9.

Xiong, X., M. Chen, W.A. Lim, D. Zhao, and L.S. Qi. 2016. "CRISPR/Cas9 for Human Genome Engineering and Disease Research." *Annual Review of Genomics and Human Genetics* 17: 131-54.

Zahra, S.A. and G. George. 2002. "Absorptive capacity: A review, reconceptualization, and extension." *Academy of Management Review* 27(2): 185-203.

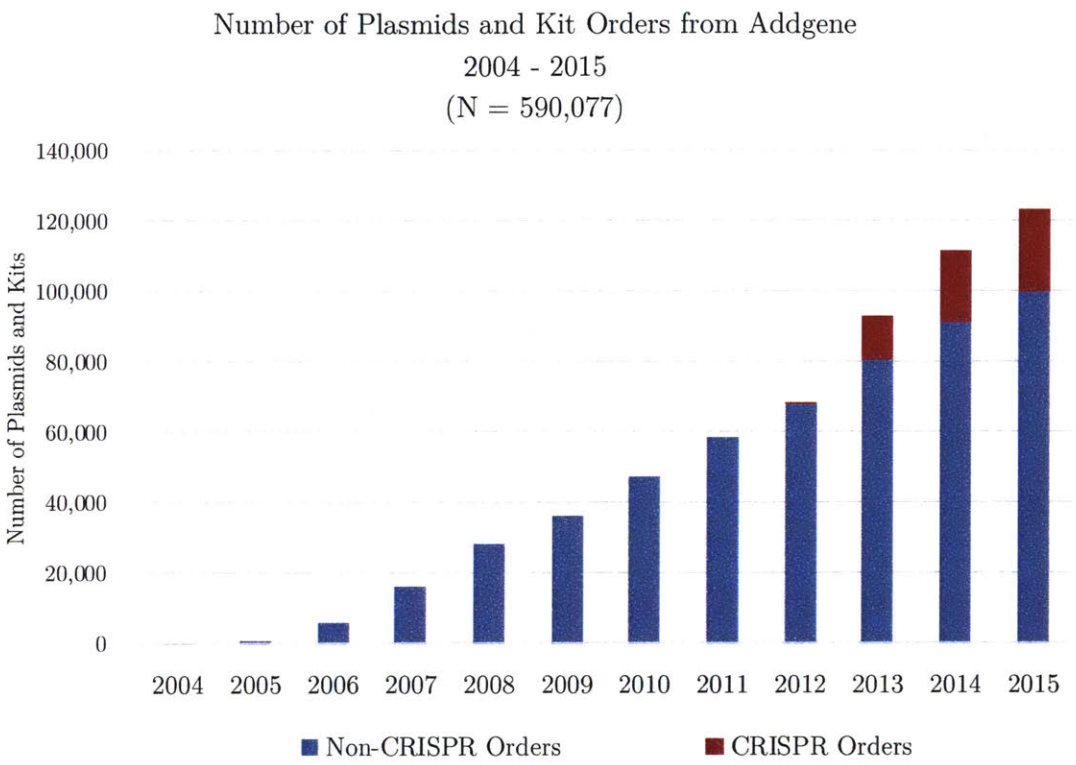
Zhang, S. 2018. "The Little-Known Nonprofit Behind the CRISPR Boom." *The Atlantic* (June 13).

Zucker, L. G. and M.R. Darby. 2001. "Capturing technological opportunity via Japan's star scientists: Evidence from Japanese firms' biotech patents and products." *The Journal of Technology Transfer* 26(1-2): 37-58.

Zyontz, S. 2016. "Technological breakthroughs, entry, and the direct of scientific progress: Evidence from CRISPR/Cas9." Working paper.

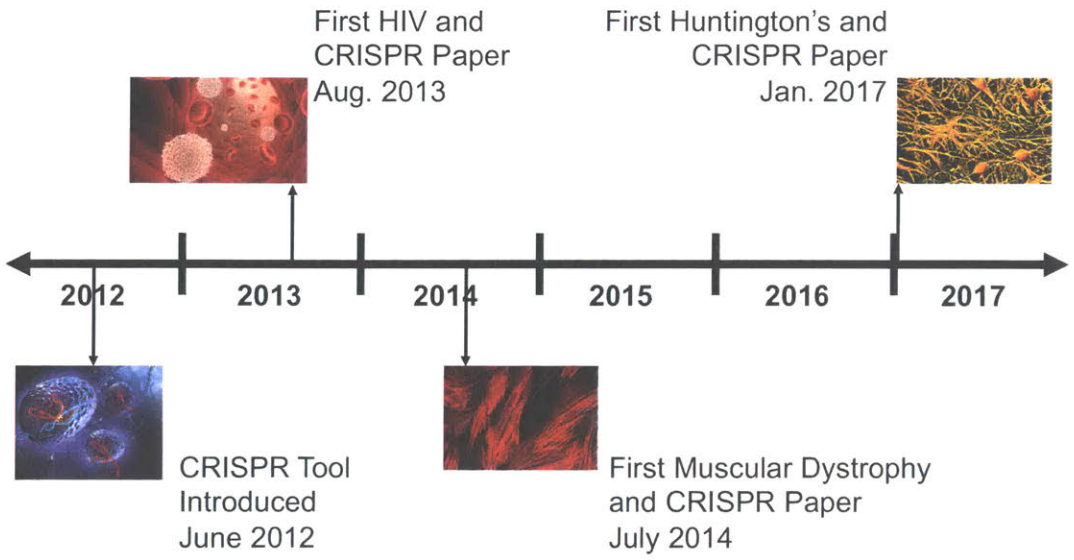
9 Figures and Tables

Figure 1. Addgene Plasmid and Kit Orders by Year

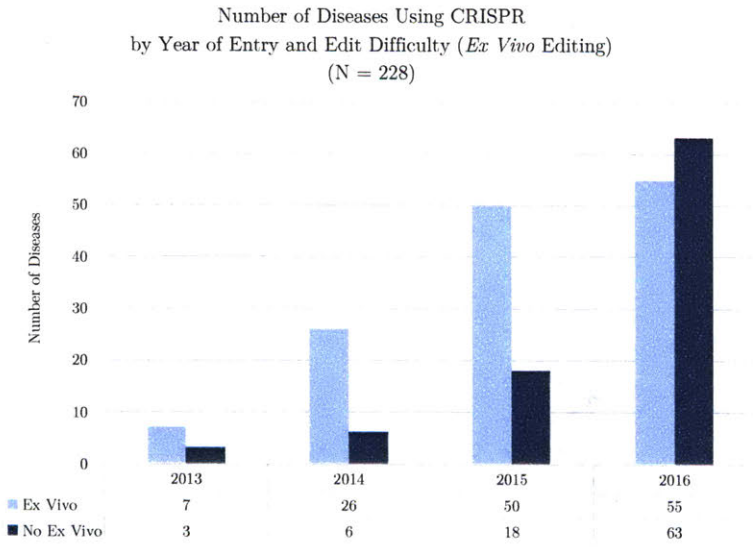


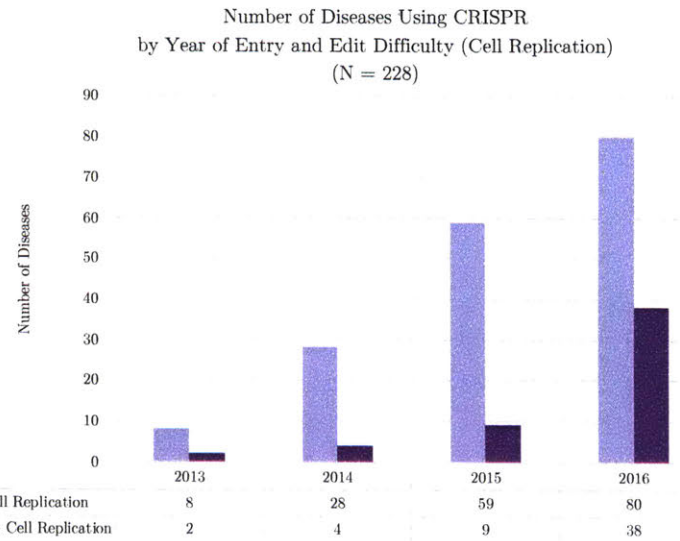
Notes. This graph shows the number of individual plasmids and sets of plasmids (kits) that Addgene sold per year from its start in 2004 through 2015. The blue bars are orders for non-CRISPR plasmids and the red bars are orders for CRISPR plasmids. Source: Addgene internal records, 2004 – 2015.

Figure 2. Example Timeline of CRISPR Introduction and Applications



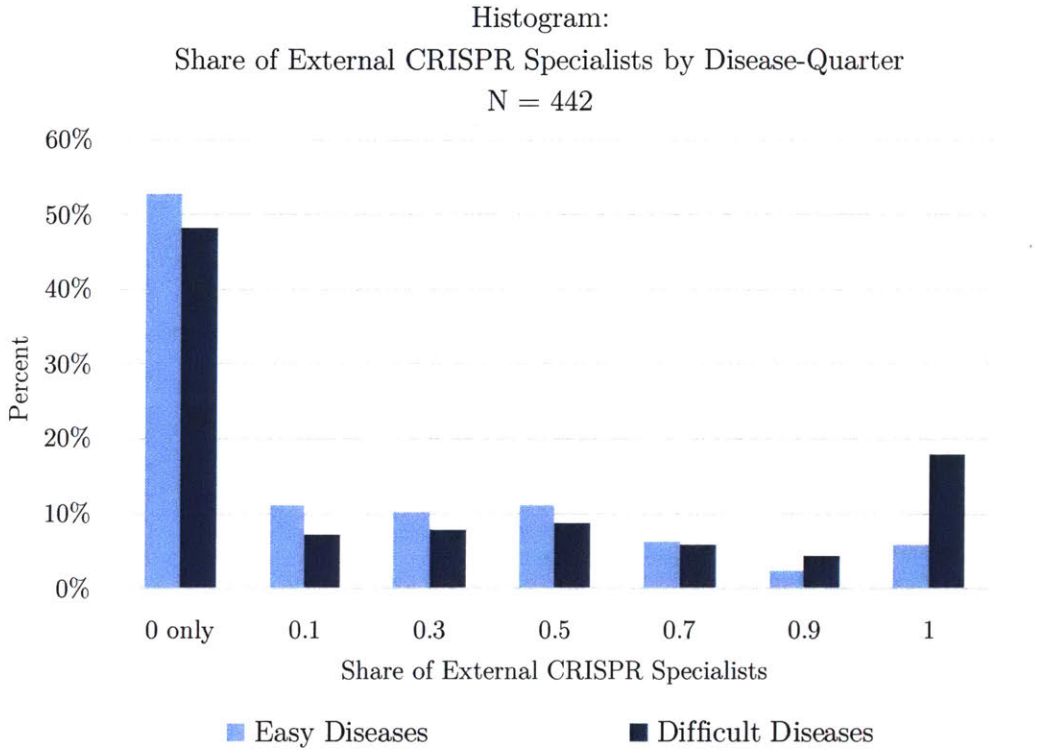
Figures 3a and b. CRISPR Entry by Year and Disease Cell Edit Difficulty





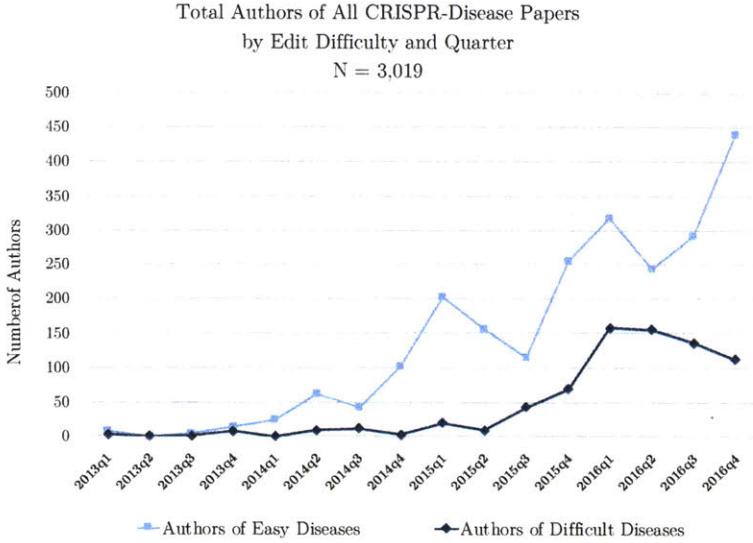
Notes. Figure 3a shows the number of diseases that first receive CRISPR each year by whether the target cell can be edited ex vivo or not. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. The light blue bars are the number of diseases that receive CRISPR if their affected cells can be edited ex vivo. The dark blue bars are the number of diseases that receive CRISPR if their affected cells cannot be edited ex vivo. Figure 3b shows the number of diseases that first receive CRISPR each year by whether the cell can self-replicate or not. If the cell a disease targets does not self-replicate, then it will be harder and costlier (in resources) to edit. The light purple bars are the number of diseases that receive CRISPR if their target cells do self-replicate. The dark purple bars are the number of diseases that receive CRISPR if their target cells cannot self-replicate. For both measures, CRISPR entry is delayed in the difficult to edit diseases. Source: PubMed publications and MeSH Terms, 2013 – 2016.

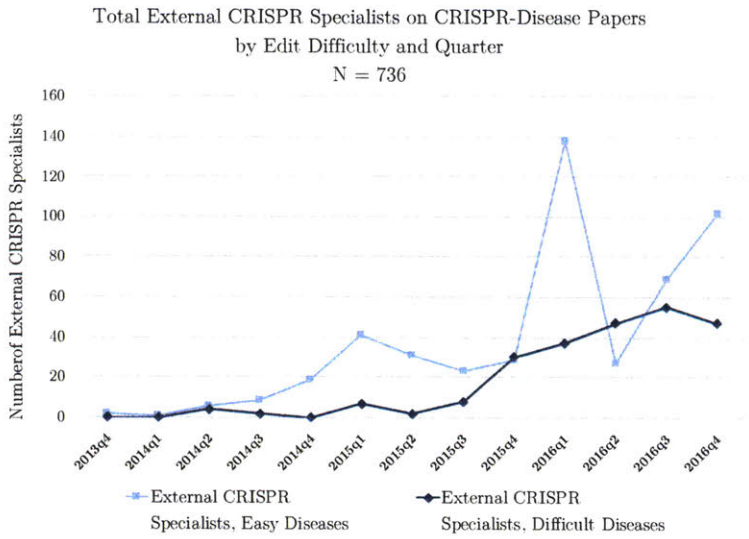
Figure 4. Distribution of the Share of External CRISPR Specialists by Disease-Quarter and Disease Cell Edit Difficulty (Ex Vivo Editing)



Notes. This figure shows the distributions of the share of external CRISPR specialists on teams that publish CRISPR papers in easy and difficult diseases by disease and quarter. The light blue bars represent the distribution pattern for easy diseases and the dark blue bars represent the distribution pattern for difficult diseases. The distribution of the dark blue bars shifts to the right suggesting that teams publishing in difficult diseases have a higher share of external CRISPR specialists. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

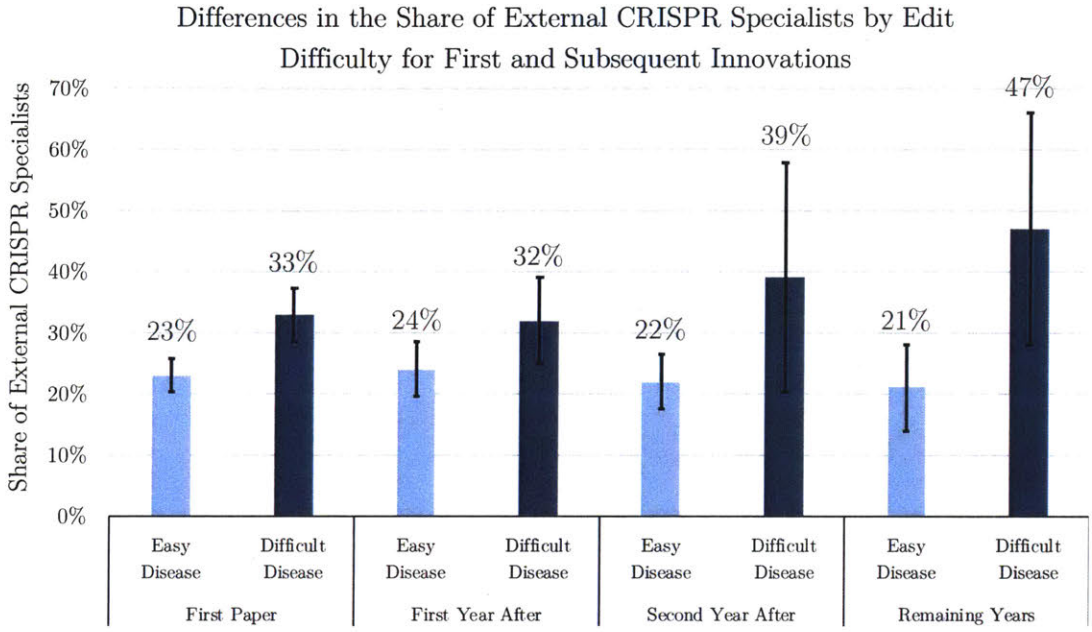
Figures 5a and b. Number of Authors on CRISPR-Disease Papers by Quarter and Disease Affected Cell Edit Difficulty (Ex Vivo Editing)





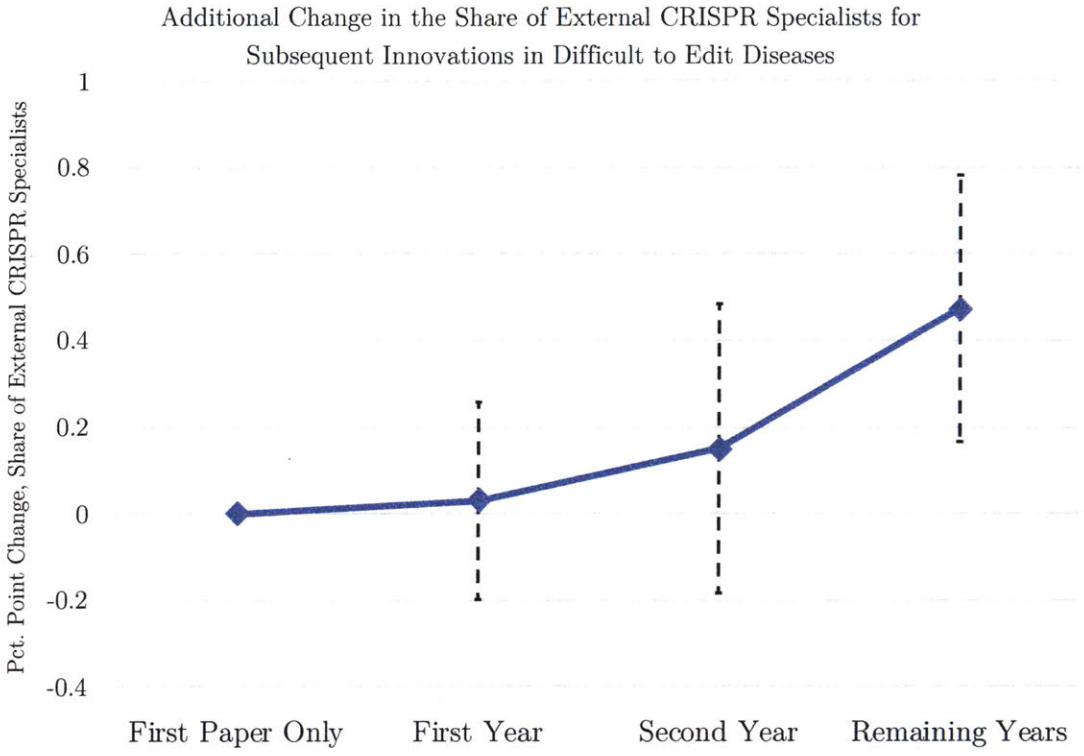
Notes. Figure 4a shows the increase in the total number of authors on CRISPR-disease papers by disease difficulty and quarter. Figure 5b shows the increase in only external CRISPR specialist authors on CRISPR-disease papers by disease difficulty and quarter. In both figures, the light blue line represents the trend for easy diseases and the dark blue line represents the trend for difficult diseases. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Figure 6. Mean Differences in the Share of External CRISPR Specialists by Disease Cell Edit Difficulty (Ex Vivo Editing) for First and Subsequent Innovations



Notes. This figure shows the difference in means between the share of external CRISPR specialists for CRISPR papers in easy diseases versus difficult diseases based on whether the paper was the first or subsequent CRISPR paper in the disease. The light blue bars represent the means for easy diseases and the dark blue bars represent the means for difficult diseases. The share of external CRISPR specialists is higher on average for teams publishing CRISPR papers in difficult diseases for both the first and subsequent papers. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Figure 7. Additional Change in the Share of External CRISPR Specialists for Subsequent Innovations in Difficult to Edit Diseases (Ex Vivo Editing)



Notes. This figure shows the change in the share of external CRISPR specialists on teams publishing subsequent CRISPR-disease papers in difficult diseases an corresponds to the coefficients estimated in Table 4a, Model 2. As later subsequent CRISPR papers are published in difficult diseases as compared to easy diseases, the share of external CRISPR specialists authors increases. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Table 1. Summary Statistics

Variable	Description	All Diseases					Easy Diseases (No *Ex Vivo* = 0)					Difficult Diseases (No *Ex Vivo* = 1)				
		N	Mean	Std. Dev.	Min	Max	N	Mean	Std. Dev.	Min	Max	N	Mean	Std. Dev.	Min	Max
Share of External CRISPR Specialists	Share of authors with a CRISPR publication before a Disease or CRISPR-Disease publication	442	0.251	0.342	0	1	303	0.215	0.307	0	1	139	0.328	0.398	0	1
Total External CRISPR Specialists	Total authors with a CRISPR publication before a Disease or CRISPR-Disease publication	442	1.665	2.920	0	22	303	1.640	3.062	0	22	139	1.719	2.593	0	12
Total Authors	Total authors on a CRISPR-Disease publication	442	6.830	7.186	1	60	303	7.541	7.900	1	60	139	5.281	4.993	1	41
Edit Difficulty (No Ex Vivo)	= 1 if disease affects cells that cannot be edited *ex vivo*: = 0 otherwise	442	0.314	0.465	0	1	303	0.000	0.000	0	0	139	1.000	0.000	1	1
Edit Difficulty (No Cell Rep)	= 1 if disease affects cells that do not self replicate; = 0 otherwise	442	0.172	0.378	0	1	303	0.000	0.000	0	0	139	0.547	0.500	0	1
Quarter	Quarter of focal CRISPR-Disease paper	442	2015q4	3.187	2013q1	2016q4	303	2015q4	3.277	2013q1	2016q4	139	2016q1	2.809	2013q1	2016q4
Quarters from First Pub	Difference in quarters from the focal to the first CRISPR-Disease paper in a Disease	442	2.394	3.345	0	15	303	2.779	3.439	0	15	139	1.554	2.971	0	14
Max Quarters from First Pub	Difference in quarters from the most recent to the first CRISPR-Disease paper in a Disease	442	4.541	4.272	0	15	303	5.304	4.231	0	15	139	2.878	3.885	0	14
Total Disease Pubs	Number of Disease papers published	442	506.380	662.740	2	2967	303	573.574	705.242	6	2967	139	359.907	532.585	2	2346
Total CRISPR-Disease Papers	Number of CRISPR-Disease papers published	442	1.382	0.981	1	8	303	1.426	1.023	1	7	139	1.288	0.878	1	8

Notes. Summary statistics by Disease-Quarter for CRISPR publications and authors in 228 Disease categories. CRISPR-Disease papers are the underlying population of the dataset and are defined as articles containing MeSH Terms for both the disease domain and CRISPR. Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Table 2. Counts of Authors and Papers by
Disease Cell Edit Difficulty (*Ex Vivo* Editing) and Year

Total Number of CRISPR-Disease Authors by Edit Difficulty (No *Ex Vivo*)

	2013	2014	2015	2016	Total
Easy Diseases	27	232	730	1296	2285
Difficult Diseases	10	22	141	561	734
Total	37	254	871	1857	3019

Total Number and % of External CRISPR Specialists by Edit Difficulty (No *Ex Vivo*)

	2013	2014	2015	2016	Total
Easy Diseases	2	35	124	336	497
Difficult Diseases	0	6	47	186	239
Total	2	41	171	522	736

	2013	2014	2015	2016	Total
Easy Diseases	7.4%	15.1%	17.0%	25.9%	21.8%
Difficult Diseases	0.0%	27.3%	33.3%	33.2%	32.6%

Total Number of CRISPR-Disease Papers by Edit Difficulty (No *Ex Vivo*)

	2013	2014	2015	2016	Total
Easy Diseases	7	49	136	240	432
Difficult Diseases	4	7	30	138	179
Total	11	56	166	378	611

Notes. A CRISPR-Disease paper is defined as an article containing MeSH Terms for both the disease domain and CRISPR. CRISPR-Disease authors are the authors on these publications. External CRISPR Specialists are authors of CRISPR-Disease papers that published in CRISPR first. Year is when a CRISPR-Disease paper was published. Easy Diseases are those where the affected cell can be effectively edited ex vivo. Difficult Diseases are those where ex vivo editing is not available. Source: PubMed publications and MeSH Terms, 2013 – 2016.

Table 3. Disease Cell Edit Difficulty and Share of External CRISPR Specialists

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	GLM, Logit	GLM, Logit	GLM, Logit
DV =	Share Ext. CRISPR Spec.	Share Ext. CRISPR Spec.	Share Ext. CRISPR Spec.	Share Ext. CRISPR Spec.	Share Ext. CRISPR Spec.	Share Ext. CRISPR Spec.
Edit Difficulty	0.0754*		0.1538***	0.3891*		0.7943***
(No *Ex Vivo*)	(0.0400)		(0.0483)	(0.2015)		(0.2198)
Edit Difficulty		-0.0291	-0.1459**		-0.1662	-0.7483**
(No Cell Rep)		(0.0506)	(0.0640)		(0.2532)	(0.3010)
Tot Disease	-0.0001***	-0.0001***	-0.0001***	-0.0007***	-0.0007***	-0.0007***
Pubs	(0.0000)	(0.0000)	(0.0000)	(0.0002)	(0.0002)	(0.0002)
Constant	0.0613	0.0931	0.0416	-15.2530***	-15.0626***	-15.4207***
	(0.0715)	(0.0624)	(0.0910)	(0.6577)	(0.7006)	(0.6997)
FE	Quarter, Age	Quarter, Age	Quarter, Age	Quarter, Age	Quarter, Age	Quarter, Age
Observations	442	442	442	442	442	442
Diseases	228	228	228	228	228	228
R^2	0.0853	0.0775	0.0991			
Log Likelihood	-132.9693	-134.8536	-129.6273	-204.7815	-205.7884	-202.8224

Standard errors in parentheses

Based on authors of all CRISPR-Disease papers, errors clustered by disease

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes. This table shows the difference in the share of external CRISPR specialists on CRISPR-disease papers by the difficulty of disease target cell editing using both OLS and GLM with Logit Link models. In diseases with cells that cannot be edited ex vivo, the share of external CRISPR specialist authors increases controlling for quarter, age, and disease attractiveness. The effect is stronger when also controlling for the target cell's ability to self-replicate. The difficulty of cell editing by disease can be measured by biological factors of the cells each disease primarily targets. Two key factors are (1) whether the cell can be edited ex vivo and (2) whether the cell can self-replicate. If the cell a disease targets cannot be edited ex vivo or if the cell does not self-replicate, then it will be harder and costlier (in resources) to edit. A CRISPR-disease paper is defined as an article containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR-disease paper has an external CRISPR background if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Table 4a. Share of External CRISPR Specialists in Subsequent Innovations by Disease Cell Edit Difficulty (Ex Vivo Editing) (Complete Dataset)

DV =	(1) OLS, FE Share Ext. CRISPR Spec.	(2) OLS, FE Share Ext. CRISPR Spec.	(3) GLM, Logit Share Ext. CRISPR Spec.	(4) GLM, Logit Share Ext. CRISPR Spec.
Edit Difficulty* Qtr. from First Pub	0.0442***		0.3331***	
	(0.0151)		(0.1292)	
Edit Difficulty* First Year (no first paper)		0.0294		0.0898
		(0.1152)		(0.8116)
Edit Difficulty* Second Year		0.1508		1.0440
		(0.1708)		(0.9814)
Edit Difficulty* Remaining Years		0.4737***		4.1928***
		(0.1562)		(1.3324)
Tot Disease Pubs	-0.0000	0.0000	-0.0009	-0.0003
	(0.0001)	(0.0001)	(0.0008)	(0.0009)
Constant	0.2345	0.1490	-18.4380***	-17.9954***
	(0.1780)	(0.1546)	(1.0464)	(1.0471)
FE	Quarter, Age, Disease	Quarter, Age, Disease	Age, Disease	Age, Disease
Observations	442	442	442	442
Diseases	228	228	228	228
R^2	0.1015	0.0966		
Log Likelihood	104.1781	102.9659	-115.3090	-115.2413

Standard errors in parentheses;

Based on authors of all CRISPR-Disease papers, errors clustered by disease

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes. This table shows the difference in the share of external CRISPR specialists for subsequent CRISPR-disease papers by disease difficulty using both OLS and GLM with Logit Link models. As later subsequent CRISPR papers are published in difficult diseases as compared to easy diseases, the share of external CRISPR specialists authors increases controlling for disease, quarter, age, and disease attractiveness. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

Table 4b. Share of External CRISPR Specialists in Subsequent Innovations by Disease Cell Edit Difficulty (Ex Vivo Editing) (Diseases w/Many Papers)

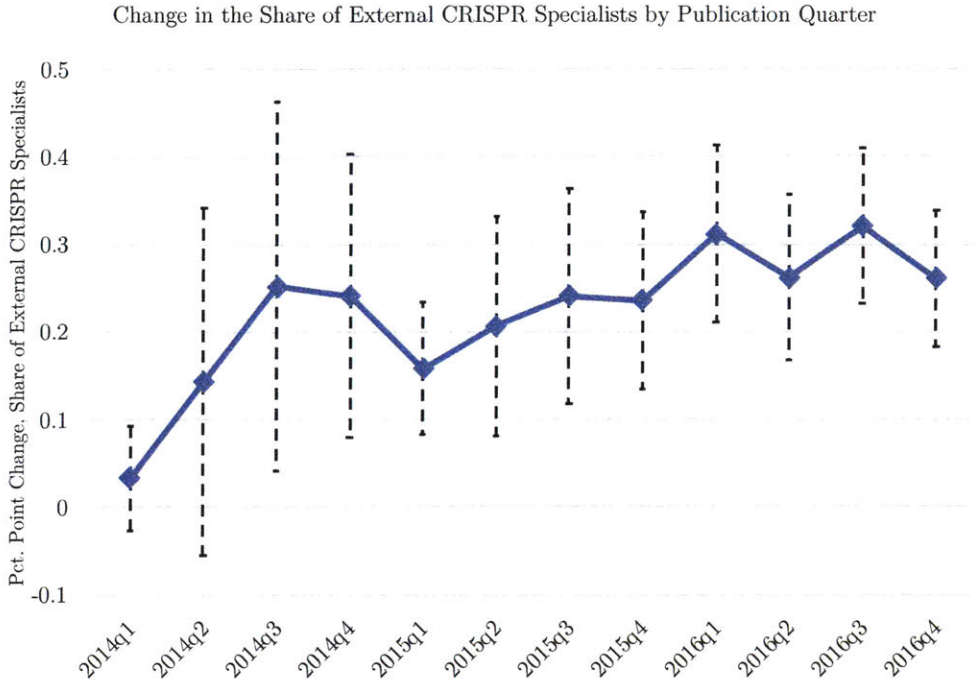
DV =	(1) OLS, FE Share Ext. CRISPR Spec.	(2) OLS, FE Share Ext. CRISPR Spec.	(3) GLM, Logit Share Ext. CRISPR Spec.	(4) GLM, Logit Share Ext. CRISPR Spec.
Edit Difficulty* Qtr from First Pub	0.0442***		0.3858**	
	(0.0153)		(0.1763)	
Edit Difficulty* First Year (no first paper)		0.0294		0.3494
		(0.1169)		(0.8458)
Edit Difficulty* Second Year		0.1508		1.3501
		(0.1732)		(1.1542)
Edit Difficulty* Remaining Years		0.4737***		4.0363***
		(0.1585)		(1.4842)
Tot Disease Pubs	-0.0000	0.0000	-0.0001	-0.0000
	(0.0001)	(0.0001)	(0.0012)	(0.0012)
Constant	0.1680	0.0895	-16.3330***	-17.4666***
	(0.1893)	(0.1658)	(2.8533)	(2.9530)
FE	Quarter, Age, Disease	Quarter, Age, Disease	Quarter, Age, Disease	Quarter, Age, Disease
Observations	307	307	307	307
Diseases	93	93	93	93
R^2	0.1015	0.0966		
Log Likelihood	16.4140	15.5721	-94.0049	-94.4182

Standard errors in parentheses, errors clustered by disease;
Based on authors of all CRISPR-Disease papers for diseases with more than one paper;
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes. This table shows the difference in the share of external CRISPR specialists for subsequent CRISPR-disease papers by disease difficulty using both OLS and GLM with Logit Link models but run only for diseases with multiple papers. As later subsequent CRISPR papers are published in difficult diseases as compared to easy diseases, the share of external CRISPR specialists authors increases controlling for disease, quarter, age, and disease attractiveness. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.

10 Appendix

Figure A1. Quarter of Publication and Share of External CRISPR Specialists



Notes. This figure shows the change in the share of external CRISPR specialists by quarter of publication. Each point is the estimated coefficient and standard errors from an OLS model that regresses each quarter of publication from Q1 2014 – Q4 2016 on the share of external CRISPR specialists publishing CRISPR-disease papers. A CRISPR-disease paper is defined as an article containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR-disease paper has an external CRISPR background if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2014 – Q4 2016.

Table A1. Mean Differences in the Share of External CRISPR Specialists by Disease Cell Edit Difficulty (*Ex Vivo* Editing) for First and Subsequent Innovations

Difference in Means by Time Period and the Availability of *Ex Vivo* Editing

Share of External CRISPR Specialists	N (*Ex Vivo*)	N (No *Ex Vivo*)	*Ex Vivo*	No *Ex Vivo*	Diff	P-val
First Paper Only	138	90	0.23	0.33	0.10	0.05
In First Year - No First Paper	50	23	0.24	0.32	0.08	0.36
In Second Year - No First Paper	36	5	0.22	0.39	0.17	0.43
Remaining Years - No First Paper	13	6	0.21	0.47	0.26	0.24

Notes. This table shows the difference in means between the share of external CRISPR specialists for CRISPR papers in easy diseases versus difficult diseases based on whether the paper was the first or subsequent CRISPR paper in the disease and is the raw data for Figure 6. The share of external CRISPR specialists is higher on average for teams publishing CRISPR papers in difficult diseases for both the first and subsequent papers. The difficulty of the disease is measured by whether the targeted cell can be edited ex vivo. If the cell a disease targets cannot be edited ex vivo, then it will be harder and costlier (in resources) to edit. CRISPR-Disease papers are defined as articles containing MeSH Terms for both the disease domain and CRISPR. An author of a CRISPR paper in a disease is considered an external CRISPR specialist if he or she published in CRISPR first (and not in the disease). Source: PubMed publications and MeSH Terms, Q1 2013 – Q4 2016.