Who Chooses Open-Source Software?

Mark A. Lemley† *& Ziv Shafir*††

Economists and legal scholars have debated the reasons people adopt open-source software, and accordingly whether and to what extent the open-source model can scale, replacing proprietary rights as a primary means of production. In this Article, we use the release by a biotechnology company of similar software under both proprietary and open-source licenses to investigate who uses open-source software and why. We find that academic users are somewhat more likely to adopt open-source software than private firms. We find only modest differences in the willingness of open-source users to modify or improve existing programs. And we find that users of open-source software often make business decisions that seem indifferent to the norms of open-source distribution. Our findings cast some doubt on the penetration of the open-source ethos beyond traditional software markets.

I. BACKGROUND

The traditional theory of intellectual property (IP) is well understood. Creators of intellectual content would have insufficient incentive to create if their works could be cheaply and quickly imitated. Thus, the law grants legal control over new creations in order to prevent, delay, or raise the cost of imitation and therefore encourage investment in creation.[1]

The rise of open-source software poses an important challenge to the classic account of the production of intellectual public goods. Instead of using IP rights to optimize monetary benefit, open-source production relies on IP rights to keep software, and any improvements or additions to it, free and widely accessible. Open-source software is provided to others for free; providers profit, if at all, not by selling the software or improvements to it but by providing consulting or other services.

Scholars vigorously debate whether open-source software represents a fundamental new means of collaborative production potentially extendable to other forms of human endeavor[2] or an altruistic fringe

---

† William H. Neukom Professor of Law, Stanford Law School; Partner, Durie Tangri LLP.

†† Medical Device Fellow, US Food and Drug Administration, 2008–2009; JD Candidate 2012, Stanford Law School.

1 See Mark A. Lemley, *The Economics of Improvement in Intellectual Property Law*, 75 Tex L Rev 989, 993–1000 (1997).

2 See Christopher M. Kelty, *Two Bits: The Cultural Significance of Free Software* 66–79 (Duke 2008); Yochai Benkler, *The Wealth of Networks* 219–33 (Yale 2006).

to the dominant market-based model of production.[3] For economists in particular, this debate is intimately bound up with questions about the motivations of those who participate in open-source production. If the classic theory of IP holds—if people are rational economic actors who will create only if the expected rewards exceed the costs—then open-source production is likely to be limited to the creation of relatively low-cost or small-scale products, primarily by those who do it in their spare time out of altruism or intellectual curiosity or who are otherwise subsidized (perhaps by a government or university) to create software without being paid for it.[4] By contrast, if people are collectively motivated to create by nonfinancial incentives, or if there is a sustainable market for the provision of services ancillary to open-source products,[5] the open-source model could conceivably displace proprietary software and even extend to products other than software, such as DNA databases.[6]

The economic literature has made substantial strides in trying to explain and characterize the free and open software movement. Explanations focus on social norms and ethics that reward contributors in nonmonetary ways (for example by enhancing reputation in a community),[7] on the unique characteristics of software as a network good and the corresponding possibility of monetizing services or ancillary products,[8] or on expected benefits from private learning or from reciprocal

---

3 See Andrew George, *Avoiding Tragedy in the Wiki-Commons*, 12 Va J L & Tech 1, ¶¶ 8–9 (Fall 2007); Ronald J. Mann, *Commercializing Open Source Software: Do Property Rights Still Matter?*, 20 Harv J L & Tech 1, 24–25 (2006); Andrea Bonaccorsi and Cristina Rossi, *Altruistic Individuals, Selfish Firms? The Structure of Motivation in Open Source Software*, 9 First Monday 5 (Jan 2004), online at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1113/1033 (visited Oct 18, 2010) (showing that altruism is a motivating factor for individual programmers); David McGowan, *Legal Implications of Open-Source Software*, 2001 U Ill L Rev 241, 260–62.

4 See Jan Eilhard, *Open Source Incorporated* *52–55 (unpublished manuscript, Feb 2009), online at http://ssrn.com/abstract=1360604 (visited Oct 18, 2010).

5 See Michele Boldrin and David K. Levine, *Market Structure and Property Rights in Open Source Industries*, 30 Wash U J L & Pol 325, 330–33 (2009); James Bessen, *Open Source Software: Free Provision of Complex Public Goods*, in Jürgen Bitzer and Philipp J.H. Schröeder, eds, *The Economics of Open Source Software Development* 57, 79–80 (Elsevier 2006).

6 See Andrew W. Torrance, *Open Source Human Evolution*, 30 Wash U J L & Pol 93, 125–28 (2009); Stephen M. Maurer, Arti Rai, and Andrej Sali, *Finding Cures for Tropical Diseases: Is Open Source an Answer?*, 6 Minn J L, Sci, & Tech 169, 171 (2004).

7 See Josh Lerner and Jean Tirole, *Some Simple Economics of Open Source*, 52 J Indust Econ 197, 212–15 (2002); Josh Lerner and Jean Tirole, *The Open Source Movement: Key Research Questions*, 45 Eur Econ Rev 819, 822–23 (2001); Eric S. Raymond, *The Cathedral and the Bazaar* 64–66 (O'Reilly 1999).

8 See Stephen M. Maurer and Suzanne Scotchmer, *Open Source Software: The New Intellectual Property Paradigm*, in Terrance Hendershott, ed, *Economics and Information Systems* 285, 290–92 (Elsevier 2006); Mann, 20 Harv J L & Tech at 33–34 (cited in note 3); Robert P. Merges, *A New Dynamism in the Public Domain*, 71 U Chi L Rev 183, 192–93 (2004); Yochai

contribution by others.[9] It has also emphasized that roughly half of all open-source contributors are paid for their contribution, usually by corporate sponsors,[10] though that raises the question of why the corporations are willing to pay that money.

Critical to advancing this debate is understanding who chooses open-source software and why.

II. OUR STUDY

In this study, we take advantage of a natural experiment in the provision of software in the bioinformatics industry to test motivations and usage patterns for open-source and proprietary software. Affymetrix, a company that sells microarrays (branded as "GeneChips") for use in DNA-based laboratory tests, provides software that enables purchasers to use the chips and analyze the results.[11] Specifically, Affymetrix and other companies have developed several algorithms that summarize the chip probe-set results and normalize the resulting data, allowing users of the Affymetrix GeneChip to analyze and understand the results of the chip probes. They have in turn implemented these algorithms in a variety of software programs.

Importantly, Affymetrix makes available both open-source and proprietary versions of the same basic algorithms and software.[12] Called "dual licensing" or "versioning," this dual release strategy is

---

Benkler, *Coase's Penguin, or, Linux and* The Nature of the Firm, 112 Yale L J 369, 411–12 (2002); McGowan, 2001 U Ill L Rev at 250–53 (cited in note 3).

9 See Karim R. Lakhani and Eric von Hippel, *How Open Source Software Works: "Free" User-to-User Assistance*, 32 Rsrch Pol 923, 936–37 (2003) (finding that programmers post answers to open-source questions because they have benefitted from the answers in the past or expect to in the future).

10 See Eilhard, *Open Source Incorporated* at *15, 26–27 (cited in note 4).

11 See Affymetrix, *GeneChip-Compatible Software Providers*, online at http://www.affymetrix.com/estore/partners_programs/genechip_compatible/genechip_compatible.affx (visited Oct 19, 2010).

12 A full list of both open-source and proprietary algorithms is provided in Appendix A. The open-source algorithms are released under a variety of different licenses. Bioconductor is released under version 2 of the GNU Public License (GPL 2), which requires the user to release the source code for any modifications to others without charge. GNU, *General Public License, Version 2*, online at http://www.gnu.org/licenses/gpl-2.0.html (visited Oct 19, 2010). dChip, by contrast, is released under a looser standard that seems to approach freeware; the source code is made available, and the license says that "if you build a Unix or Linux version, we will appreciate it if you can share the codes or binaries or be linked from this page." *dChip: Introduction and Installation*, online at http://biosun1.harvard.edu/complab/dchip/install.htm (visited Oct 19, 2010). On the bewildering array of open-source and related licenses, see Robert W. Gomulkiewicz, *Open Source License Proliferation: Helpful Diversity or Hopeless Confusion?*, 30 Wash U J L & Pol 261, 264–77 (2009). See also Robert W. Gomulkiewicz, *Conditions and Covenants in License Contracts: Tales from a Test of the Artistic License*, 17 Tex Intel Prop L J 335, 337–39 (2009).

increasingly common in the open-source software environment.[13] Affymetrix provides parallel software in two different versions because it believes that some users will want the low cost of open-source software, while some private firms will want to improve on the software or add their own proprietary extensions without being subject to the obligation to release those extensions under an open-source license. This provides a natural test of customer selection of open-source or proprietary software.[14]

To test those usage patterns, we conducted two studies. First, we reviewed publications from Affymetrix's "Scientific Publications" page for the period from 2007 to 2008.[15] In all, we found 178 publications that make reference to algorithms of software for interpreting data from the Affymetrix GeneChip. We collected data on the entity status of each author, and on the algorithm and the software those authors used. When users write papers disclosing results made using the GeneChip, they almost always cite the software they relied on to generate those results.[16] As a result, we can learn about usage patterns by observing publication patterns.

Those usage patterns will be biased, however, because they represent not all users of Affymetrix GeneChips but only those who publish their results. That group is likely to include disproportionate numbers of academic and nonprofit users. And indeed we find in the next Part that most publishers are academics. To ensure that private firm users are represented, we oversampled commercial users in the publication search. Publication also cannot account for intensity of use, though there should be some relationship between the two: those who use the chip to make a publishable discovery are likely to be

---

13 See Stefano Comino and Fabio M. Manenti, *Dual Licensing in Open Source Software Markets* *3–4 (unpublished manuscript, Jan 2010), online at http://ssrn.com/abstract=985529 (visited Oct 19, 2010); Robert W. Gomulkiewicz, *Entrepreneurial Open Source Software Hackers: MySQL and Its Dual Licensing*, 9 Computer L Rev & Tech J 203, 208–11 (2004).

14 The test is not perfect, because the algorithms and software implementing those algorithms are not identical. First, we note that the algorithms and the software are not necessarily independent variables. For example, MAS 5.0 does not run the RMA algorithm, and GeneSpring does not run the MAS algorithm. As a result, some users might have chosen their software based on the algorithms rather than the open-source or proprietary issues. Second, since we study Affymetrix microarrays, our study sample is naturally biased toward users who have a greater propensity to choose the Affymetrix GCOS or MAS 5.0 software than do people buying Illumina microarrays. Notwithstanding these limitations, the programs are used for the same purpose, written by the same people, and have significant similarities.

15 For a full discussion of how those sources were selected, see Appendix B.

16 This is confirmed by survey data: 95.4 percent of academic users, 90.9 percent of other nonprofit users, and 80 percent of private firms who wrote a paper using the results cited the software in that paper. Overall, 250 of 270 survey respondents who published papers, or 92.6 percent, cited the software by name in the paper. For the original survey question, see Appendix C, question 14.

more intensive users than those who do not, all other things being equal. Still, this is a limitation of the publication data.

To explore the motivations of users and the uses they make of the software, we supplement the literature search with a survey of users of the Affymetrix GeneChip. We sent out surveys to a wide range of individuals who have used the Affymetrix software and received 437 complete and usable responses.[17] Of the 437 respondents, 305 (or 69.8 percent) were academics, 52 (or 11.9 percent) were nonprofit entities, and 80 (or 18.3 percent) were at private firms. The results indicate a predominance of academic users, but not surprisingly include more private firms than did the publication search.

III. RESULTS

A. Publication Data

For each published article, we studied the nature of the authors (academic, nonprofit, commercial, or mixed); if commercial, the industry (biotechnology or pharmaceuticals) in which the authors worked; whether the authors were domestic or foreign; and whether they used commercial or open-source software. We excluded from consideration review articles or meta-analyses in which the authors were not themselves running the experiments in which the data were collected, as well as articles that did not specify the software used. The summary results are presented in Table 1.

---

17 This represents a response rate below 50 percent, but given that the survey was sent by email and without a follow-up, that rate is not abnormally low.

TABLE 1. DESCRIPTIVE STATISTICS FOR PUBLICATIONS

Category	Number
Total papers	178
Academic papers	110
Nonprofit papers (including government)	11
Commercial papers	55
Commercial biotech	14
Commercial pharma	41
Joint academic–commercial papers	8*
Papers by US authors	116
Papers by foreign authors	62**
Open-source algorithms/software	35
Proprietary algorithms/software	148

* When there were joint papers, we included them in either the commercial or academic category, but not both. We made this assessment based on who the lead author was and which side had the greater number of authors. Of the joint papers, six were primarily by academic authors and two were primarily commercial.

** As with joint authors, we treated papers as primarily either US or foreign in authorship.

In Table 2, we explore how different actors vary in their use of open-source and proprietary software.

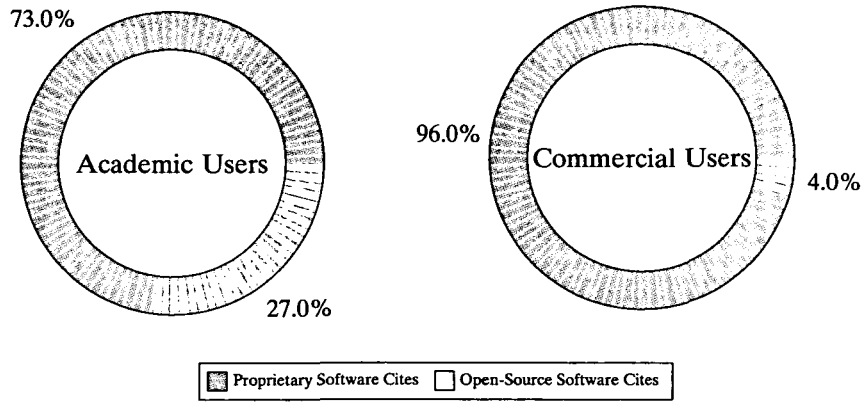
TABLE 2. USAGE OF OPEN-SOURCE SOFTWARE BY AUTHOR CHARACTERISTICS

Author Category	Proprietary Cites*	Open-Source Cites
Academic papers	85 (77%)	31 (28%)
Nonprofit papers	9 (82%)	2 (18%)
Commercial papers	54 (96%)	2 (4%)
Commercial biotech	13 (93%)	1 (7%)
Commercial pharma	41 (100%)	1 (2%)
Joint academic–commercial papers	7 (88%)	2 (25%)
US authors	106 (91%)	24 (22%)
Foreign authors	52 (84%)	13 (21%)

* Some authors cited both open-source and proprietary software, so the percentages may exceed 100 percent.

The difference between academic and commercial users of open-source software is dramatic. The results are depicted in Figure 1.

FIGURE 1. SOFTWARE USAGE OF ACADEMIC VERSUS COMMERCIAL USERS



The publication data suggest that academic users are far more likely to use open-source software than commercial entities; only a very small percentage of commercial users made use of open-source software, and half of those did so in the course of academic–commercial collaborations.

One possible explanation for this result is that commercial users want to release products using or incorporating the GeneChip software and charge for those products, something open-source licenses would restrict.[18] We investigate that possibility in the next Part. Relatedly, commercial users may be worried about the risk of legal entanglement that comes with open-source software even if they are not now planning to release a product that incorporates that software. Another possibility is that commercial users expect proprietary software to be superior, with additional features or better usability, and are more willing than academic users to pay for that additional functionality.

B. Survey Data

With the assistance of Affymetrix, we surveyed users of Affymetrix GeneChips. We asked survey respondents a variety of questions, including which algorithms and which software they used, whether they were aware of other software, why they chose the version they did, whether they provided bug reports or fixes, whether they modified or improved the software, whether they released those modifications and

---

[18] Exactly what form this restriction takes is a complicated question. A few open-source licenses, such as the BSD license, do not prevent licensees from charging for products. The GPL and related licenses, strictly speaking, allow the licensee to charge for products made using the license, but they also require that the code be made available for free, which makes the collection of revenue more difficult.

if so to whom, and whether they published their results.[19] We can sort these results on the answer to any previous question; in much of what follows, we distinguish between those who used open-source software and algorithms and those who used proprietary software and algorithms.[20]

We begin by replicating the analysis from the publication survey, measuring academic, nonprofit, and commercial users and users of open-source or proprietary software. The results appear in Tables 3 and 4.

TABLE 3. DESCRIPTIVE STATISTICS FOR SURVEY DATA

Category	Number
Total respondents	442*
Academic respondents	305 (68.1%)
Nonprofit respondents	52 (11.7%)
Commercial respondents	80 (17.9%)
Respondents using open-source algorithms/software	251 (57.4%)
Respondents using proprietary algorithms/software	369 (84.4%)

* Five respondents did not fit into one of the listed categories.

TABLE 4. USAGE OF OPEN-SOURCE SOFTWARE BY AUTHOR CHARACTERISTICS

Author Category	Proprietary Uses*	Open-Source Uses**
Academic users	245 (68.6%)	174 (71%)
Nonprofit users	43 (12.0%)	25 (10.2%)
Commercial users	62 (17.4%)	43 (17.6%)

* N = 357; 12 did not answer the question. Of the 357, 7 (2.0 percent) were independent or reported no affiliation.

** N = 245; 6 did not answer the question. Of the 245, 3 (1.2 percent) were independent or reported no affiliation.

---

19 The complete list of questions is attached as Appendix C.

20 We included Bioconductor and dChip in the open-source category and Stratagene ArrayAssist, Stratagene ArrayAssist Lite, Rosetta Resolver, GeneChip Operating System (GCOS), GeneSpring, Affymetrix Power Tools, Affymetrix Expression Console, Microarray Suite (MAS), and Spotfire in the proprietary category. There was also a survey option of "Other," in which the vast majority of people listed other less well-known proprietary software—these answers were included in the "proprietary" category.

We also had many users who used both open-source and proprietary software and algorithms. As a general matter, we have included joint users in both the open-source and proprietary software categories, but at various points we have separated them into their own group. We note that in Table 5.

Curiously, the data here show no similar distinction between academic and commercial users. Indeed, commercial users made up the same percentage of users of both open-source and proprietary software. And we were unable to reject the hypothesis that each type of user was equally likely to use open-source or proprietary software.[21]

We do find modest differences in the stated reasons why people employed particular programs. Those who used proprietary software gave the following answers:

Why did you choose the version you did? [check all that apply]

Answer Options	Response Frequency	Response Count
It was compatible with/used the desired processing algorithm (such as RMA, GC-RMA, MAS 5.0)	67.4%	225
It was cheaper	35.0%	117
It gave me greater freedom to modify the software	29.6%	99
It was more reliable	30.8%	103
It was more convenient	37.7%	126
It had better support	22.5%	75
My institution already had a license to it	30.2%	101
Other	9.6%	32
If "Other," please explain in your own words why you chose the version you did:		41
Answered question		334
Skipped question		35

21 The statistical tests are available from the authors upon request.

By contrast, those who used open-source software gave the following answers:

Why did you choose the version you did? [check all that apply]

Answer Options	Response Frequency	Response Count
It was compatible with/used the desired processing algorithm (such as RMA, GC-RMA, MAS 5.0)	74.6%	176
It was cheaper	47.5%	112
It gave me greater freedom to modify the software	42.4%	100
It was more reliable	31.8%	75
It was more convenient	34.7%	82
It had better support	21.6%	51
My institution already had a license to it	23.7%	56
Other	7.2%	17
If "Other," please explain in your own words why you chose the version you did:		24
Answered question		236
Skipped question		15

There are some differences here that line up with expectations—users of open-source software put more emphasis on cost (47.5 versus 35 percent) and on ability to modify the software (42.4 versus 29.6 percent), for instance. But the overall differences are surprisingly modest. Indeed, differences between respondents' answers to the cost and ability to modify questions are the only ones that are statistically significant at a 99 percent confidence level.[22]

We also find modest differences in the use people make of the different types of software. For example, users of open-source software were marginally more likely to send fixes or bug reports back to the software developers. Among open-source users, 42.3 percent provided bug reports, compared with 37.8 percent of proprietary software users; 23.5 percent of open-source users improved the software, compared with 19.1 percent of proprietary users. Fifty percent of the open-source users who modified the software released that software to others, compared with 45.6 percent of those who modified proprietary software.

22 With only 90 percent confidence, we can say that users preferred open-source software because it was compatible with what they already used, and proprietary software if their firm already had a license to it.

One reason the data appear so noisy may be that these measures include a significant number of users who employed both open-source and proprietary GeneChip software at different times. This is the most logical way to explain the result that 35 percent of users of proprietary software stated "lower cost" as a reason for choosing the software they did, for example.[23] As a result, we also report a restricted sample limited to those who used only open-source or only proprietary software.[24]

The results of the restricted sample are more consistent with our expectations. We report the descriptive statistics in Table 5. We emphasize, however, that the restricted sample is smaller and that none of these differences is statistically significant.

TABLE 5. USAGE OF SOLELY OPEN-SOURCE OR PROPRIETARY SOFTWARE BY AUTHOR CHARACTERISTICS

Author Category	Solely Proprietary Uses*	Solely Open-Source Uses**
Academic users	114 (64.8%)	33 (73.3%)
Nonprofit users	22 (12.5%)	4 (8.9%)
Commercial users	34 (19.3%)	8 (17.8%)

* N = 185; 9 did not answer the question. Of the 185, 6 (3.2 percent) were independent or reported no affiliation.

** N = 47; 2 did not answer the question.

The differences between the reasons users offer for choosing solely open-source or proprietary software, by contrast, are more pronounced in the restricted sample. Respondents who used only open-source software gave the following answers:

---

23 It is also possible that some users considered proprietary software cheap because their institution already owned a license to it, but that should at most equalize the cost vis-à-vis open-source software.

24 This restricted sample will of necessity exclude those who switched from proprietary to open-source software (or, more problematically, from open-source to proprietary software) during the time of the study. See Greg R. Vetter, *Commercial Free and Open Source Software: Knowledge Production, Hybrid Appropriability, and Patents*, 77 Fordham L Rev 2087, 2129–31 (2009) (discussing examples of companies that switch modes).

Why did you choose the version you did? [check all that apply]

Answer Options	Response Frequency	Response Count
It was compatible with/used the desired processing algorithm (such as RMA, GC-RMA, MAS 5.0)	64.9%	24
It was cheaper	64.9%	24
It gave me greater freedom to modify the software	54.1%	20
It was more reliable	27.0%	10
It was more convenient	18.9%	7
It had better support	16.2%	6
My institution already had a license to it	2.7%	1
Other	2.7%	1
If "Other," please explain in your own words why you chose the version you did:		2
Answered question		37
Skipped question		10

By contrast, those who used only proprietary software gave the following answers:

Why did you choose the version you did? [check all that apply]

Answer Options	Response Frequency	Response Count
It was compatible with/used the desired processing algorithm (such as RMA, GC-RMA, MAS 5.0)	53.9%	83
It was cheaper	21.4%	33
It gave me greater freedom to modify the software	16.9%	26
It was more reliable	28.6%	44
It was more convenient	37.7%	58
It had better support	23.4%	36
My institution already had a license to it	32.5%	50
Other	16.2%	25
If "Other," please explain in your own words why you chose the version you did:		30
Answered question		154
Skipped question		31

Both sets in the restricted sample emphasized compatibility as a strong factor in software choice, suggesting a fair degree of path dependence driving the results. Open-source software users were more likely to emphasize the lower cost (64.9 versus 21.4 percent) and freedom to modify their software (54.1 versus 16.9 percent); proprietary software users were more likely to point to an existing institutional license (32.5 versus 2.7 percent) and the convenience of the software (37.7 versus 18.9 percent).[25]

We also find modest differences in the use people make of the different types of software in our restricted sample. We report those differences in Table 6.

TABLE 6. BUG REPORTING AND SOFTWARE IMPROVEMENT DIFFERENCES BETWEEN OPEN-SOURCE AND PROPRIETARY USERS

Type of Usage	Proprietary Users*	Open-Source Users
Reported bugs or sent fixes	32.8%	35.6%
Improved the software	16.4%	22.2%
Released modified software to others	48.3%	60.0%

* Differences between open-source and proprietary users are not statistically significant in any of the categories.

Users of open-source software were not appreciably more likely to send fixes or bug reports back to the software developers—35.6 percent of open-source users provided bug reports, compared with 32.8 percent among proprietary software users; 22.2 percent of open-source users improved the software, compared with 16.4 percent of proprietary users. Sixty percent of the open-source users who modified the software released that software to others, compared with 48.3 percent of those who modified proprietary software. As with the unrestricted sample, the differences are surprisingly modest, and none is statistically significant.

This is a curious result. Under the general terms of open-source software licenses, those who modify or contribute new software that they sell or distribute must make their new or modified code available to others. There might be an argument that open-source rules do not require that the source code be publicly distributed at all but simply

---

25 The "cheaper," "freedom to modify," and "license" answers are statistically significant at the 99 percent confidence level; the "convenience" answer is significant at the 95 percent confidence level.

provided on request.[26] But even so, we find that most open-source users who released improved software did so only to their own customers (67.9 percent in the unrestricted sample; 83.3 percent in the restricted sample), despite the presumed obligation to make the code available to anyone who wanted it. Those who did release the software improvements usually did so under their own brand names—78.6 percent for open-source users (100 percent in the restricted sample) and 72.4 percent for proprietary software users (61.5 percent in the restricted sample).

Interestingly, the rationale that drove Affymetrix to release parallel programs in both open-source and proprietary formats—the belief that customers would want proprietary software because they hoped to improve on it and sell the improvement—does not appear to be borne out in the data. Some users of both open-source and proprietary software modified the software and released their modifications to the public as products. Users of open-source software were slightly more likely to modify the program than users of proprietary software. They were marginally more likely to release that software to the public, and marginally more likely to brand that improved software with their own mark. Again, however, we note that these differences are modest and not statistically significant.

IV. IMPLICATIONS

What stand out in these results are not the differences between open-source and proprietary software users but the similarities. Users of the probe-set summarization and normalization software seem largely indifferent to whether the software they use is open-source or proprietary. They often use multiple programs and algorithms, some of which are open source and some of which are proprietary. That sort of mixing is a big worry for open-source lawyers; indeed, there are companies designed to audit software code to make sure there is no inappropriate mixing.[27] Mixing is relevant primarily for those who change or improve code, not for passive users. But we find that roughly a quarter of users, including many users of both open-source and proprietary software, do in fact improve the software and release their

---

[26] The GPL 2 provides that "[y]ou must make sure that [recipients], too, receive or can get the source code." GNU, *General Public License, Version 2* (cited in note 12). This could be satisfied by a distribution on request. And the BSD license does not impose that constraint.

[27] See, for example, Black Duck Software, *Multi-source Development with Open Source Software*, online at http://www.blackducksoftware.com/multisourcedevelopment (visited Oct 20, 2010); Gomulkiewicz, 9 Computer L Rev & Tech J at 207, 210–11 (cited in note 13).

improvements to others.[28] Curiously, even these improvers seem largely indifferent to the open-source–proprietary line, to the extent that many of them appear to be ignoring the fundamental constraint of open-source software—that you release your improvements to everyone. In fact, it is not clear that they are even aware of the limits on behavior that open-source licenses impose.

This is a small study in a single industry, one in which the users are not primarily computer programmers. Further research may reveal whether our conclusion is merely an artifact of our study universe, whether similar behavior exists in more traditional open-source software contexts, or whether the result is driven by the fact that users outside the central open-source community have not fully internalized the norms of open-source software. More study is required.[29]

But if our results are generalizable, they have a broader implication: to bring to bear the "law in action" literature to the open-source–proprietary divide. While the law—and the intent of the open-source movement—draws a sharp distinction between open-source and proprietary software, placing them effectively in different worlds, users of the software in bioinformatics appear to observe no such sharp distinction. They appear to employ a mix of open-source and proprietary software tools chosen for a variety of reasons, not merely or even primarily for their openness or appropriability. And they seem to use those software tools as they will—and not as the niceties of open-source contracts would suggest. Law may matter to makers of open-source software, but it does not appear to affect the behavior of software users.

---

28 Other studies have found even higher rates of hybrid software use. See The 451 Group, *Executive Overview: Open Source Is Not a Business Model: How Vendors Generate Revenue from Open Source Software* *1 (Oct 2008), online at http://www.the451group.com/reports/executive_summary.php?id=694 (visited Oct 31, 2010) (finding that 50 percent of open-source software vendors use a hybrid model); Eilhard, *Open Source Incorporated* at *26–27 (cited in note 4) (finding that nearly half of open-source programmers are paid for their work).

29 Others have surveyed computer programmers in more traditional software industries. See, for example, Sujoy Chakravarty, Ernan Haruvy, and Fang Wu, *The Link between Incentives and Product Performance in Open Source Development: An Empirical Investigation*, 9 Global Bus & Econ Rev 151, 159–60 (2007); Guido Hertel, Sven Niedner, and Stefanie Herrmann, *Motivation of Software Developers in Open Source Projects: An Internet-Based Survey of Contributors to the Linux Kernel*, 32 Rsrch Pol 1159, 1166 (2003); Sharon Belenzon and Mark Schankerman, *Motivation and Sorting in Open Source Software Innovation* *21–24 (unpublished manuscript, Oct 2008), online at http://ssrn.com/abstract=1401776 (visited Oct 20, 2010). And one study measured open-source adoption in hospitals. See Gilberto Munoz-Cornejo, Carolyn B. Seaman, and A. Güneş Koru, *An Empirical Investigation into the Adoption of Open Source Software in Hospitals*, 3 Intl J Healthcare Info Sys & Informatics 16, 26 (July–Sept 2008).

This in turn inclines us to a policy of legal neutrality with regard to open-source and proprietary software.[30] Open-source software is not "better" than proprietary software, nor the reverse. Users want—and should have—the freedom to choose the right software for their particular purposes. But legal scholars need to understand that those users are often less concerned with the niceties of software licenses than with using whatever tool seems best suited to the job at hand. William Gibson famously said that "the street finds its own uses for things."[31] The same, it seems, can be said of technology licenses.

---

30 See generally Robert W. Hahn, ed, *Government Policy toward Open Source Software* (Brookings 2002).

31 William Gibson, *Burning Chrome*, in William Frucht, ed, *Imaginary Numbers: An Anthology of Marvelous Mathematical Stories, Diversions, Poems, and Musings* 195, 212 (Wiley 1999).

APPENDIX A. LIST OF PROGRAMS AND ALGORITHMS STUDIED

Algorithms	Software	
	Proprietary	Open Source
MAS 5.0	Microarray Analysis Suite 5.0 (MAS 5.0)	Bioconductor
RMA/GC-RMA	GeneChip Operating System (GCOS)	dChip
PLIER	GeneSpring	
Rosetta Resolver	Affymetrix Power Tools	
	Affymetrix Expression Console	
	Stratagene ArrayAssist	
	Stratagene ArrayAssist Lite	
	Rosetta Resolver	
	Spotfire	
	Other	

APPENDIX B. METHODOLOGY FOR IDENTIFYING RELEVANT PUBLICATIONS

We analyzed scientific publications to identify the software used to conduct primary probe-set summarization on Affymetrix Expression Analysis GeneChips. Probe-set summarization refers to the step of calculating raw expression data from the scanned image of a microarray. A publication was said to have used a particular software package if it was clear that probe-set summarization was being performed by the software in question. All other data-analysis steps such as normalization were not considered in deciding which software was used.

The following information was captured for each of the publications:

- Year of publication
- Article name
- Journal name
- Number of authors cited
- Number of different departments and institutions cited
- Primary location of the authors
- Whether the publication was from an academic source, an NGO, a commercial company, or a combination
- If the publication was from a company, then which industry that company is in (either pharma or biotech)
- Whether the software cited was proprietary or open source
- The name of the company releasing the software
- The name of the particular software

A. Academic Publications

Academic publications were randomly chosen in the following fashion: As of the time we collected the data in 2009, the complete list of publications that cite Affymetrix are listed on the Affymetrix website, by year, under the "Scientific Publications" page. Publications are displayed in groups of nine for any particular year.

Using the random number generator from www.random.org, a random number from one through nine was produced, which would correspond to a publication in the group of nine, with one denoting the publication on top, and nine denoting the publication on the bottom. There were 3,622 total publications in 2007, on a total of 363 different pages displaying groups of nine publications for each page.[32] At

---

[32] For a complete list, see Affymetrix, *Results*, online at http://www.affymetrix.com/estore/publications/pub_query_result.affx?year=2007 (visited Jan 16, 2011).

first an article from each page was chosen, but by page 50, the selection switched to choosing articles only from odd-numbered pages.

In total, we analyzed 203 academic publications. Of these, 121 proved useable for analysis. The rest were left out of the analysis because they:

- were not accessible;
- did not explicitly specify the software they were using to get probe-set summarization data;
- were review articles or were meta-analyses of others' data, rather than reports of primary use of GeneChips; or
- were not using Affymetrix "Expression Analysis Arrays."

B. Commercial Publications

We also analyzed publications from commercial companies. For commercial publications, we obtained a list of all of the scientific publications that cited Affymetrix in 2007 and 2008 from Martha Manion, head librarian of Affymetrix. This list denotes which publications are from commercial entities. As a result, we collected all publications by commercial entities, ensuring that there was not a selection bias in choosing the commercial publications. In total, we analyzed 110 publications; of these, 55 proved usable for analysis. We analyzed only publications by authors solely from commercial companies; no academic–commercial combinations were analyzed in this particular round of analysis. The exclusion criteria used for academic publications also applied here.

A full list of publications studied is available from the authors upon request.

APPENDIX C. LIST OF SURVEY QUESTIONS

1. Do you use software algorithms to summarize the probe set and/or normalize raw expression data from Affymetrix GeneChips? These are the steps in GeneChip analysis that turn raw intensity data (.CEL files) into expression signal values (e.g., .CHP files) that can be compared from microarray to microarray for further analysis.		
Answer Options	Response Frequency	Response Count
Yes		
No		

2. Are you the person within your organization who decided which algorithm and software to use to summarize and/or normalize raw expression data?		
Answer Options	Response Frequency	Response Count
Yes		
No		
If "No," please tell us who within your organization made that decision (contact information is most appreciated)		

3. Which of the following algorithms have you used for probe-set summarization or normalization of Affymetrix GeneChip data? [check all that apply]		
Answer Options	Response Frequency	Response Count
Microarray Analysis Suite (MAS 5.0 or MAS 4.0)		
RMA/GC_RMA		
PLIER		
Rosetta Resolver		
None		
I don't know		
Other		
If "Other," please list the algorithms here		

4. Which of the following software programs have you used for probe-set summarization or normalization of Affymetrix GeneChip data? [check all that apply]		
Answer Options	Response Frequency	Response Count
dChip		
Bioconductor		
Stratagene ArrayAssist		
Stratagene ArrayAssist Lite		
Rosetta Resolver		
GeneChip Operating System (GCOS)		
GeneSpring		
Affymetrix Power Tools		
Affymetrix Expression Console		
Microarray Suite		
Spotfire		
I don't know		
None [if you answered "none" to both questions 3 and 4, please skip to the end of the survey]		
Other		
If "Other," please list the software programs here		

5. What do you use the software for? [If different software is used for different steps, please choose the appropriate package used most often in the right-most column]		
Check "Yes" if you use software cited in Question 4 for this function		
	Yes	
Probe-set summarization only		
Normalization of data only		
Both probe-set summarization and normalization		
Background subtraction		
Plotting and analysis		
Further downstream analysis (clustering, fold change analysis, etc.)		
Other		
I don't know		

Other software used for this particular function

Answer Options	Chip	Bioconductor	Stratagene ArrayAssist	Stratagene ArrayAssist Lite	Rosetta Resolver	GeneChip Operating System (GCOS)	GeneSpring	Affymetrix Power Tools	Affymetrix Expression Console	Microarray Suite	Spotfire	Other
Probe-set summarization only												
Normalization of data only												
Both probe-set summarization and normalization												
Background subtraction												
Plotting and analysis												
Further downstream analysis (clustering, fold change analysis, etc.)												
Other												
I don't know												

6. Were you aware of the existence of different software programs that served these purposes?		
Answer Options	Response Frequency	Response Count
Yes		
No		

7. Why did you choose the version you did? [check all that apply]		
Answer Options	Response Frequency	Response Count
It was compatible with/used the desired processing algorithm (such as RMA, GC-RMA, MAS 5.0)		
It was cheaper		
It gave me greater freedom to modify the software		
It was more reliable		
It was more convenient		
It had better support		
My institution already had a license to it		
Other		
If "Other," please explain in your own words why you chose the version you did:		

8. Have you provided reports of bugs or other problems with summarization or normalization software since you began using it?		
Answer Options	Response Frequency	Response Count
Yes		
No		

9. Have you modified or improved summarization or normalization software since you began using it?		
Answer Options	Response Frequency	Response Count
Yes		
No		
If you answered "Yes," please briefly describe how you modified it:		

10. If so, did you release that modified version to others outside your organization?		
Answer Options	Response Frequency	Response Count
Yes		
No		

11. If so, did you release the modified version under your own brand name, or under the original publisher's brand name?		
Answer Options	Response Frequency	Response Count
Our own		
Original publisher		

12. Who did you market, sell, or give the modified version to?		
Answer Options	Response Frequency	Response Count
Our existing clients or partners only		
Anyone who wanted it		

13. Have you published results based on research you did using microarray analysis software?		
Answer Options	Response Frequency	Response Count
Yes		
No		

14. If so, did you cite the software by name?		
Answer Options	Response Frequency	Response Count
Yes		
No		

15. Are you at an academic research institution or a for-profit company?		
Answer Options	Response Frequency	Response Count
Academic institution		
Other nonprofit entity		
Private company		
Independent/no affiliation		

16. Which role or roles best define your usage of Affymetrix GeneChips?		
Answer Options	Response Frequency	Response Count
I am the primary user of the chip, seeing the work through from when the GeneChip is first used in an experiment to when it is analyzed		
I use the GeneChip in my experiments, but outsource the analysis of it to a third party (ex. a core facility)		
I conduct the GeneChip analysis for other people's experiments (ex. worker at a core facility)		
I don't use GeneChips myself; I extract GeneChip data from other sources (such as the Gene Expression Omnibus) and analyze that		
I utilize microarray data and/or microarray analysis programs to develop my own software tools, which I then market to outside parties		
I utilize microarray data and/or microarray analysis programs to develop my own software tools, which I then release under an open source license		
I utilize microarray data and/or microarray analysis programs to develop my own software tools, which I then use only internally		
I have no association with GeneChips or their data		

17. If you would like to be entered into a drawing for an iPod Touch, then please enter your information below. All identifying information entered will be kept anonymous and separate from the survey responses.		
Answer Options	Response Frequency	Response Count
Name		
Email		
Phone Number		