Who Chooses Open Source Software?¹

Mark A. Lemley² & Ziv Shafir³

Abstract

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 study, 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.

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

¹ © 2009 Mark A. Lemley & Ziv Shafir.

² William H. Neukom Professor, Stanford Law School; partner, Durie Tangri LLP.

³ Medical Device Fellow, U.S. Food and Drug Administration; Student, Stanford Law School, J.D. expected 2012.

We thank the National Human Genome Research Initiative for a generous grant supporting this work, and Jeremy Gollub, Earl Hubbell, Katie Kong, Martha Manion, and Wei Zhou for assistance in designing the study, contacting the participants, and collecting the data, and Vinita Kailasanath for statistical assistance.

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. (Lemley, 1997).

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 either 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 (Benkler, 2005), or an altruistic fringe to the dominant market-based model of production (Epstein?; Mann, 2006). For economists in particular, this debate is intimately bound up with questions about the motivation 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 – open source production is likely to be limited to the creation of relatively low-cost products, primarily by those who do it in their spare time out of altruism or who are otherwise subsidized (perhaps by a government or university) to create software without being paid for it. By contrast, if people are collectively motivated to create by non-financial incentives, or if there is a sustainable market for the provision of services ancillary to open source products, the open source model could

conceivably displace proprietary software and even extend to products other than software, such as DNA databases. (McGowan, 2001; Maurer & Rai, 2006).

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 (Lerner & Tirole, 2001, 2002); on the unique characteristics of software as a network good and the corresponding possibility of monetizing services or ancillary products (Benkler 2002; Mann, 2006; Maurer & Scotchmer, 2006; McGowan, 2001); or on expected benefits from private learning or from reciprocal contribution by others (Lakhani & von Hippel, 2003).

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

2. *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 (www.affymetrix.com), provides software that enables purchasers to use the chips and analyze the results. 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 gene chip to analyze and understand the results of the chip probes. They have in turn implemented those algorithms in a variety of software programs.

Importantly, Affymetrix makes available both open source and proprietary versions of the same basic algorithms and software.⁴ Called “dual licensing” or “versioning,” this dual release strategy is increasingly common in the open source software environment. (Comino & Manenti, 2007; Gomulkiewicz, 2004). 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.⁵

To test those usage patterns, we conducted two studies. First, we reviewed publications from Affymetrix’s “Scientific Publications page” that appeared from the period

⁴ 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 the GNU Public License, or GPL, which requires the user to release the source code for any modifications to others without charge. 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 to be linked from this page.”

⁵ 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 aren’t necessarily independent variables. For example, MAS 5.0 does not run the RMA algorithm, and Genespring doesn’t run the MAS algorithm. As a result, some users might have chosen their software based on the algorithms rather than the open source/proprietary issues. Second, since we study Affymetrix microarrays, our study sample is naturally biased towards users that have a greater propensity to choose the GCOS/MAS 5.0/Affymetrix 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.

2007 to 2008.⁶ 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 to cite the software they relied on to generate those results.⁷ 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 disproportionately academic and non-profit users. And indeed we find in the next section that most publishers are academics. To ensure that private firm users are represented, we oversampled commercial users in the publication search.

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 Gene Chip. We sent out surveys to a wide range of individuals who have used the Affymetrix software, and received 437 complete and usable responses.⁸ Of the 437 remaining respondents, 305 (or 69.8%) were academics, 52 (or 11.9%) were non-profit entities, and 80 (or 18.3%) were at private firms. The

⁶ For a full list of these sources, see Appendix B.

⁷ This is confirmed by survey data: 95.4% of academic users, 90.9% of other non-profit users, and 80% 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%, cited the software by name in the paper. See Appendix C, Question 14.

⁸ This represents a response rate below 50%, but given that the survey was sent by email and without a follow up, that rate is not abnormally low.

results indicate a predominance of academic users, but not surprisingly include more private firms than did the publication search.

3. Results

A. Publication Data

For each published article, we studied the nature of the authors (academic, non-profit, commercial, or mixed), if commercial the industry (biotechnology or pharmaceuticals) in which the authors work, 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 where the data was first collected, as well as articles that did not specify the software they used. The summary results are presented in Table 1.

Table 1
Descriptive Statistics for Publications

Category	Number
Total papers	178
Academic papers	110
Non-profit papers (incl. gov't)	11
Commercial papers	55
Commercial biotech	14

Commercial pharma	41
Joint academic/commercial papers	8 ⁹
Papers by US authors	116
Papers by Foreign authors	62 ¹⁰
Open source algorithms/software	35
Proprietary algorithms/software	148

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	Open Source Cites	Proprietary Cites¹¹
Academic papers	31 (28%)	85 (77%)
Non-profit papers	2 (18%)	9 (82%)
Commercial papers	2 (4%)	54 (96%)
Commercial biotech	1 (7%)	13 (93%)
Commercial pharma	1 (2%)	41 (100%)
Joint academic-commercial papers	2 (25%)	7 (88%)
US authors	24 (22%)	106 (91%)

⁹ 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, 6 were primarily academic authors and 2 were primarily commercial.

¹⁰ As with joint authors, we treated papers as primarily either US or foreign in authorship.

¹¹ Some authors cited both open source and proprietary software, so the percentages may exceed 100%.

Foreign authors

13 (21%)

52 (84%)

The difference between academic and commercial users of open source software is dramatic.

The results are depicted in Figures 1:

Figure 1

Software Usage of Academic vs. Commercial Users

Academic Users

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 gene chip software and to charge for those products, something open source licenses would not permit them to do. We investigate that possibility in the next section. 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 Gene Chips. 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 if so to whom, and whether they published their results. [The complete list of questions is attached is Appendix C]. 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.¹²

¹² 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

We begin by replicating the analysis from the publication survey, measuring academic, non-profit, 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 ¹³
Academic respondents	305 (68.1%)
Non-profit 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%)

Table 4

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 that 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, and note that below.

¹³ Five respondents did not fit into one of the listed categories.

Usage of Open Source Software by Author Characteristics

Author Category	Open Source Uses ¹⁴	Proprietary Uses ¹⁵
Academic users	174 (71%)	245 (68.6%)
Non-profit users	25 (10.2%)	43 (12.0%)
Commercial users	43 (17.6%)	62 (17.4%)

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.¹⁶

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, MAS5.0..._)	67.4%	225
It was	35.0%	117

¹⁴ N = 245; 6 did not answer the question. Of the 245, 3 (1.2%) were independent or reported no affiliation.

¹⁵ N = 357; 12 did not answer the question. Of the 357, 7 (2.0%) were independent or reported no affiliation.

¹⁶ The statistical tests are reported in Appendix D.

cheaper		
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
<i>answered question</i>		334
<i>skipped question</i>		35

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	74.6%	176

RMA, GC-RMA, MAS5.0..._)		
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
<i>answered question</i>		236
<i>skipped question</i>		15

There are some differences here that line up with expectations – users of open source software put more emphasis on cost (47.5% vs. 35%) and on ability to modify the software (42.4% vs. 29.6%), for instance. But the differences are overall surprisingly modest. Indeed, the cost and

ability to modify answers are the only ones that are statistically significant at a 99% confidence level.¹⁷

We do 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. 42.3% of open source users provided bug reports, compared with 37.8% among proprietary software users; 23.5% of open source users improved the software, compared with 19.1% of proprietary users. Fifty percent of the open source users who modified the software released that software to others, compared with 45.6% of those who modified proprietary software.

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 gene chip software at different times. This seems the most logical way to explain the result that 35% of users of proprietary software stated “lower cost” as a reason for choosing the software they did, for example.¹⁸ As a result, we also report a restricted sample limited to those who only used open source or only used proprietary software.

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 are statistically significant.

¹⁷ With only 90% 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.

¹⁸ 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.

Table 5

Usage of Solely Open Source or Proprietary Software by Author Characteristics

Author Category	Solely Open Source Uses ¹⁹	Solely Proprietary Uses ²⁰
Academic users	33 (73.3%)	114 (64.8%)
Non-profit users	4 (8.9%)	22 (12.5%)
Commercial users	8 (17.8%)	34 (19.3%)

The differences users offer for why they chose 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:

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, MAS5.0..._)	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
<i>answered question</i>		37
<i>skipped question</i>		10

¹⁹ N = 47; 2 did not answer the question.

²⁰ N = 185; 9 did not answer the question. Of the 185, 6 (3.2%) were independent or reported no affiliation.

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, MAS5.0..._)	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
<i>answered question</i>		154
<i>skipped question</i>		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% vs. 21.4%) and freedom to modify their software (54.1% vs. 16.9%); proprietary software users were more likely to point to an existing institutional license (32.5% vs. 2.7% for open source) and the convenience of the software (37.7% vs. 18.9% for open source users).²¹

²¹ The "cheaper," "freedom to modify," and "license" answers are statistically significant at the 99% confidence level; the "convenience" answer is significant at the 95% confidence level.

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

Table 6
Bug Reporting and Software Improvement Differences between
Open Source vs. Proprietary Users

Type of Usage	Open Source Users	Proprietary Users ²²
Report bugs or send fixes	35.6%	32.8%
Improved the software	22.2%	16.4%
Released modified software to others	60.0%	48.3%
*		

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

²² Differences between open source and proprietary users are not statistically significant in any of the categories.

This is a curious result. Open source software licenses generally require that those who modify or contribute new software make their 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, rather than simply distributed on request.²³ But even so, we find that most open source users who released improved software did so only to their own customers (67.9% in the unrestricted sample; 83.3% 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% for open source users (100% in the restricted sample) and 72.4% for proprietary software users (61.5% 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. 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 for software with their own mark. Again, however, we note that these differences are modest and not statistically significant.

²³ The GPL v. 2.0 provides that “You must make sure that they, too, receive or can get the source code.” This could be satisfied by a distribution on request.

4. *Conclusions*

What stand out in this data are not the differences between open source and proprietary software 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. (Black Duck Software, www.blackducksoftware.com; Gomulkiewicz, 2004). Mixing is relevant only 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 those improvements to others.²⁴ 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.

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

²⁴ Other studies have found even higher rates of hybrid software use. (The 451 Group 2008, p. 11)

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.²⁵

But if our results do generalize, it has 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 – draw 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 doesn’t appear to affect the behavior of software users.

That in turn inclines us to a policy of legal neutrality with regard to open source and proprietary software. (Hahn 2002, ch. 5) Open source software isn’t “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.”²⁶ The same, it seems, can be said of technology licenses.

²⁵ Others have surveyed computer programmers in more traditional software industries (Haruvy et al., 2003; Hertel et al., 2003)

²⁶ **William Gibson, *Burning Chrome* (1986).**

References

- Belenzon, Sharon & Mark Schankerman, *Motivation and Sorting in Open Source Software Innovation*, working paper March 3, 2009, <http://www.ssrn.com>
- Benkler, Yochai, *Coase's Penguin, or Linux and the Nature of the Firm*, 112 **Yale L.J.** 369 (2002)
- Benkler, Yochai, **The Wealth of Networks** (2006)
- Bessen, James, *Open Source Software: Free Provision of Complex Public Goods*, <http://www.researchoninnovation.org/opensrc.pdf> (July 2002)
- Comino, Stefano & Fabio M. Manenti, *Dual Licensing in Open Source Software Markets*, <http://ssrn.com/abstract=985529> (working paper May 2007)
- David, Paul A. & Joseph S. Shapiro, *Community-Based Production of Open Source Software: What do we know about the developers who participate?*, <http://ssrn.com/abstract=1286273> (Sept. 23, 2008)
- Free Software Foundation, GNU Public License v. 2.0
- Gomulkiewicz, Robert W., *Entrepreneurial Open Source Software Hackers: MySQL and Its Dual Licensing*, 9 **Comp. L. Rev. & Tech. J.** 203 (2004)
- Gomulkiewicz, Robert W., *Conditions and Covenants in License Contracts: Tales From a Test of the Artistic License*, 17 **Tex. Intell. Prop. L.J.** 335 (2009)
- Gomulkiewicz, Robert W., *Open Source License Proliferation: Helpful Diversity or Hopeless Confusion?*, 30 **Wash. U. J.L. & Pol'y** 261 (2009)
- Hahn, Robert W., *Government Policy Toward Open Source Software* (AEI-Brookings Joint Center for Regulatory Studies 2002)
- Haruvy, Ernan, Fang Wu, & Sujoy Chakravarty, *Incentives for Developers' Contributions and Product Performance Metrics in Open Source Development: An Empirical Investigation*, working paper (2003)
- Hertel, G., M. Krishnan & Sandra Slaughter, *Motivation in Open Source Projects: An Internet-based Survey of Contributors to the Linux Kernel*, 32 **Res. Pol'y** 1159 (2003)
- Lakhani, Karim & Eric von Hippel, *How Open Source Software Works: 'Free' User-to-User Assistance*, 32 **Res. Pol'y** 923 (2003)
- Lemley, Mark A., *The Economics of Improvement in Intellectual Property Law*, 75 **Tex. L. Rev.** 989 (1997).
- Lerner, Josh & Jean Tirole, *The Open Source Movement: Key Research Questions*, 45:4-6 **Eur. Econ. Rev.** 819 (2001)

Lerner, Josh & Jean Tirole, *Some Simple Economics of Open Source*, 52 **J. Indus. Econ.** 197 (2002)

Lerner, Josh & Jean Tirole, *The Scope of Open Source Licensing*, 21 **J. L., Econ. & Org.** 20 (2009)

Mann, Ronald J., *Commercializing Open Source Software: Do Property Rights Still Matter?*, 20 **Harv. J. L. & Tech.** 1 (2006).

Maurer, Stephen M., Arti Rai & Andrej Sali, *Finding Cures for Tropical Diseases: Is Open Source an Answer?*, 6 Minn. J.L. Sci. & Tech. 169, 169 (2004)

Maurer, Stephen & Suzanne Scotchmer, *Open Source Software: The New Intellectual Property Paradigm*, NBER Working Paper 12148 (2006)

McGowan, David, *Legal Implications of Open Source Software*, 2001 **U. Ill. L. Rev.** 241

Merges, Robert P., *A New Dynamism in the Public Domain*, 71 **U. Chi. L. Rev.** 183 (2004)

Raymond, Eric, **The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary** (O'Reilly Press 2001)

The 451 Group, *Open Source Is Not a Business Model* (monograph October 2008)

Vetter, Greg R., *Commercial Free and Open Source Software: Knowledge Production, Hybrid Appropriability, and Patents*, 77 **Fordham L. Rev.** 2087 (2009)

Vetter, Greg R., *The Collaborative Integrity of Open Source Software*, 2004 **Utah L. Rev.** 563

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:

1. Year of publication
2. Article name
3. Journal Name
4. Number of authors cited
5. Number of different departments and institutions cited
6. Primary location of the authors
7. Whether the publication was from an academic source, an NGO, a commercial company, or a combination
8. If publication is from a company, then what industry that company is in (either pharma or biotech)
9. Whether the software cited was proprietary or open source
10. The name of the company releasing the software
11. Name of the particular software

Academic Publications

Academic publications were randomly chosen in the following fashion: 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 1-9 was produced, which would correspond to a publication in the group of nine, 1 denoting the publication on top, and 9

denoting the publication on bottom. There are 3622 total publications in the year of 2007, on a total of 363 different pages displaying groups of nine publications for each page. At first an article from each page was chosen, but by the page 50, the selection switched to choosing articles only from odd-numbered pages. A screenshot of the webpage is shown below:

SCIENTIFIC PUBLICATIONS

[New Search](#)

Full List

12483 publications found in the database

2008 (460)	2007 (3622)	2006 (3158)	2005 (1805)	2004 (1316)
2003 (1065)	2002 (589)	2001 (239)	2000 (109)	1999 (52)
1998 (31)	1997 (9)	1996 (11)	1995 (2)	1994 (6)
1993 (5)	1992 (2)	1991 (2)		

2007 PUBLICATIONS

- Identification of CD36 as a new surface marker of marginal zone B cells by transcriptomic analysis
Zhang, P. *et al.* *Molecular Immunology* 44(4), 332-7, 2007 [PubMed](#)
- Identification of IRF-8 and IRF-1 target genes in activated macrophages
Dror, N. *et al.* *Molecular Immunology* 44(4), 338-46, 2007 [PubMed](#)
- Prediction of chromosomal aneuploidy from gene expression data
Hertzberg, L. *et al.* *Genes Chromosomes & Cancer* 46(1), 75-86, 2007 [PubMed](#)
- Gene expression in 16q is associated with survival and differs between Sorlie breast cancer subtypes
Wennmalm, K. *et al.* *Genes Chromosomes & Cancer* 46(1), 87-97, 2007 [PubMed](#)
- Overlay analysis of the oligonucleotide array gene expression profiles and copy number abnormalities as determined by array comparative genomic hybridization in medulloblastomas
Lo, K. C. *et al.* *Genes Chromosomes & Cancer* 46(1), 53-66, 2007 [PubMed](#)
- Differential expression of drug resistance-related genes between sensitive and resistant blasts in acute myeloid leukemia
Eisele, L. *et al.* *Acta Haematologica* 117(1), 8-15, 2007 [PubMed](#)
- Qualitative assessment of gene expression in affymetrix genechip arrays
Nagarajan, R. *et al.* *Physica A-Statistical Mechanics and Its Applications* 373(3), 486-96, 2007 [PubMed](#)
- Large-scale inference and graph-theoretical analysis of gene-regulatory networks in B. Subtilis
Christensen, C. *et al.* *Physica A-Statistical Mechanics and Its Applications* 373(3), 796-810, 2007 [PubMed](#)
- Use of microarrays to find novel regulators of periodontal ligament fibroblast differentiation
Lallier, T. E. *et al.* *Cell and Tissue Research* 327(1), 93-109, 2007 [PubMed](#)

< back | **page: 1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177

203 Academic publications were analyzed. Of these, 121 proved useable for analysis. The rest were left out of analysis if 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
- were not using Affymetrix "Expression Analysis Arrays"

Commercial Publications

Publications from commercial companies were the second type of publication analyzed. For commercial publications, a list of all the scientific publications that cite Affymetrix for 2007 and 2008 was obtained from Martha Marion, head librarian of Affymetrix. This list denotes which publication is from a commercial entity, thus all publications by commercial entities were obtained, ensuring that there was not a selection bias in choosing the commercial publications. 110 publications were analyzed, of these, 55 proved usable for analysis. Only publications that were composed of authors solely from commercial companies were analyzed; no academic/commercial combinations were analyzed in this particular round of analysis. Exclusion criteria of 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.

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 (MAS5.0 or MAS4.0)		
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		
Answer Options	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	<u>dChip</u> <u>Bioconductor</u> <u>Stratagene ArrayAssist</u> <u>Stratagene ArrayAssist Lite</u> <u>Rosetta Resolver</u> <u>GeneChip Operating System (GCOS)</u> <u>Genespring</u> <u>Affymetrix Power Tools</u> <u>Affymetrix Expression Console</u> <u>Microarray Suite</u> <u>Spotfire</u> <u>Other</u>												
	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, MAS5.0...)		
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 non-profit 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		

add

Appendix D

Results of Statistical Tests

Author Category	Open Source Cites		Proprietary Cites[1]		
Academic papers	34	##	96	74%	
Non-profit papers	3	##	9	75%	
Commercial papers	3	5%	53	95%	
Commercial biotech	1	7%	13	93%	
Commercial pharma	2	5%	42	95%	
Joint academic-commercial papers	2	##	6	75%	
	45		219		264
US authors	24	##	96	62%	
Foreign authors	13	##	60	38%	
	37		156		193

[1] Some authors cited both open source and proprietary software, so the percentages may exceed 100%.

Testing hypothesis that there is no difference between the proportion of respondents in proprietary and open source categories for each category

Can we reject the null hypothesis that the proportion is the same in both sets?

	Academic papers	Non-profit papers	Commercial papers	Commercial biotech	Commercial pharma	Joint academic-commercial papers	US authors	Foreign authors
Are means significantly different at the 90% confidence level?	Yes	No	Yes	No	Yes	No	No	No
Are means significantly different at the 95% confidence level?	Yes	No	Yes	No	Yes	No	No	No
Are means significantly different at the 99% confidence level?	Yes	No	Yes	No	Yes*	No	No	No
frequency of open source users	34	3	3	1	2	2	24	13
frequency of proprietary users	96	9	53	13	42	6	96	60
proportion of open source users	0.76	0.07	0.07	0.02	0.04	0.04	0.65	0.35
proportion of proprietary users	0.44	0.04	0.24	0.06	0.19	0.03	0.62	0.38
p hat	0.49	0.05	0.21	0.05	0.17	0.03	0.62	0.38
q hat	0.51	0.95	0.79	0.95	0.83	0.97	0.38	0.62
standard error	0.08	0.03	0.07	0.04	0.06	0.03	0.09	0.08
z value	3.88	0.75	-2.62	-1.01	-2.42	0.61	0.38	-0.42
p-value (AREA BTWN +/- Z)	0.9999	0.5467	0.9912	0.6875	0.9845	0.4581	0.296	0.326