Toward effective software solutions for big biology

Prins, Pjotr; de Ligt, Joep; Tarasov, Artem; Jansen, Ritsert C; Cuppen, Edwin; Bourne, Philip E

Published in:
Nature Biotechnology

DOI:
10.1038/nbt.3240

after the original drug's approval, or they are noted but not adequately appreciated before disclosure would be required—too late to be patented[5] (**Fig. 1**). Furthermore, disclosure crosses international borders: European disclosures will block US patents, and vice versa.

Without patents and with little or no data exclusivity available for originator companies, companies will have little incentive to invest in validating new uses and bringing them into widespread use[5]. Similar challenges may arise even for completely new drugs, such as innovative biologics, where the broad increase to the 'prior art' and 'common general knowledge' created by clinical trial disclosure may render many related new drugs unpatentably obvious.

The new regulations take some account of this. US proposals involve masking the identity of tested drugs; the much-debated[11] new EMA policy and EU regulation protects commercially confidential information. But secondary effects of unrecognized importance are unlikely to be held confidential and would probably still be disclosed under these proposals.

More flexible regulatory data and market exclusivities, enforced by the FDA or EMA, could provide incentives to replace unavailable patents. Even so, firms could still use that clinical trial disclosure for market approvals in countries without equivalent regimes for regulatory exclusivity, potentially leading to contention in trade agreements. Additional clashes may arise with new trade secrecy legislation in the United States and European Union. In particular, the European Commission (EC) recently published a draft directive attempting to harmonize and enhance trade secret protection in Europe[12]; the European Federation of Pharmaceutical Industries and Associations welcomed it and stressed the importance of protecting the "proprietary know-how" of drug development, including in clinical trials[13]. The overlap between these proposals remains wholly unresolved.

Why have these concerns not been sufficiently addressed despite substantial debate on clinical trial disclosure in general? Simply put, few parties have both expertise and incentives in regulatory and intellectual property issues. Regulatory agencies and patent adjudicators each lack the other's expertise and mandate, and large innovator drug companies have few incentives to help smaller companies patent new uses. But to reap the full benefits of clinical trial disclosure, policymakers must consider the overlap between disclosure mandates and intellectual property law. Effective pharmaceutical innovation requires reasonable incentives engaging both private and public actors. We should not block new cures by pursuing a laudable initiative without full consideration.

ACKNOWLEDGMENTS

The authors wish to thank I. Glenn Cohen and the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School for support and advice.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

W Nicholson Price II & Timo Minssen

W. Nicholson Price II is at the University of New Hampshire School of Law, Concord, New Hampshire, USA, and Timo Minssen is at the Centre for Information and Innovation Law, University of Copenhagen, Denmark.
e-mail: nicholson.price@law.unh.edu

1. European Medicines Agency. Publication of clinical reports. http://www.ema.europa.eu/ema/index.jsp?curl=pages/news_and_events/news/2014/10/news_detail_002181.jsp&mid=WC0b01ac058004d5c1 (2 October 2014).
2. The Council of the European Union. Council adopts new rules on clinical trials. http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/lsa/142181.pdf (14 April 2014).
3. US Food and Drug Administration. *Fed. Reg.* **78**, 33421–33423 (2013).
4. Mello, M.M. *et al. N. Engl. J. Med.* **369**, 1651–1658 (2013).
5. Roin, B.N. *Social Science Research Network* http://dx.doi.org/10.2139/ssrn.2337821 (October 10, 2013)
6. Doshi, P. *Br. Med. J.* **347**, f6754 (2013).
7. DiMasi, J.A., Reichert, J.M., Feldman, L. & Malins, A. *Clin. Pharmacol. Ther.* **94**, 329–335 (2013).
8. *MEHL/Biophile Int'l Corp. v. Milgraum*, 192 F.3d 1362 (1999).
9. T128/82 (Pyrrolidin-Derivate) Case T 128/82 - Hoffman-La Roche/Pyrrolidine-derivatives O.J. EPO 1984, 164 (Jan. 12, 1984)
10. G2/08 (Dosage regime/Abbott Respiratory) Case G 2/08 - Dosage regime/Abbott Respiratory 2010 O.J. EPO 456 (Feb. 19, 2010)
11. Torjesen, I. *Br. Med. J.* **348**, g3432 (2014).
12. EU Commission, COM/2013/0813 final - 2013/0402 (COD), Proposal for a Directive of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52013PC0813 (Nov. 28, 2013)
13. European Federation of Pharmaceutical Industries and Associations. EFPIA welcomes the Commission's Proposal on the protection of undisclosed know-how and business information ("Trade Secrets"). http://www.efpia.eu/mediaroom/129/44/EFPIA-welcomes-the-Commission-39-s-Proposal-on-the-protection-of-undisclosed-know-how-and-business-information-quot-Trade-Secrets-quot (28 November 2013).

Toward effective software solutions for big biology

To the Editor:

Leading scientists tell us that the problem of large data and data integration, referred to as 'big data', is acute and hurting research. Recently, Snijder *et al.*[1] suggested a culture change in which scientists would aim to share high-dimensional data among laboratories. It is important to realize that sharing data is only part of the solution. The elephant in the room is bioinformatics and bioinformatics software development in particular—which, despite being crucially important, mostly fails to address the requirements of 'big data'.

Whereas Internet companies such as Google, Facebook and Skype have built infrastructure and developed innovative software solutions to cope with vast amounts of data, the bioscience community seems to be struggling to realize big data software projects. This has led to problems in sharing, annotation, computation and reproducibility of data[2–4].

Before we can devise software solutions for big data, there are more basic pressing concerns with bioinformatics software development that need to be resolved. Biologists are not formally trained for software engineering, so much of the bioinformatics software available today has been developed by PhD biologists in relative isolation on the back of funded experimental research programs. This model of software development tied to wet-lab research can work well but has resulted in a culture of 'one-offs'. The aim of most research projects is to obtain results in the shortest possible time, and this is often achieved by writing prototype software rather than developing well-engineered and scalable solutions. Even when funding is obtained to develop software, there are usually no long-term resources allocated to software maintenance, which results in problems with bug fixing, continuity and reproducibility.

Instead of working alone to develop software, researchers can join or start collaborative free and open-source software (FOSS) projects, thereby improving their coding skills through the scrutiny of their peers. True FOSS projects have licenses that allow continuation of projects that

were abandoned by the original developers, thereby enabling modular development. We published a bioinformatics manifesto as a practical guide for FOSS-style development (https://github.com/pjotrp/bioinformatics/blob/master/README.md) that aims to provide process and architecture guidelines for early-career bioinformaticians and their supervisors. Bioinformatics already has vibrant collaborative FOSS projects, such as Galaxy, Cytoscape, BioPerl and Biopython, but these projects are often worked on part-time owing to lack of or inadequate funding and will not service the requirements of big biology without major additional investment. For example, after initial funding from the US National Institutes of Health (NIH) and the National Science Foundation (NSF), the Galaxy project is now seeking new funding to continue its work, and no funds at all have been granted by scientific agencies to work on Biopython. The amount of dedicated funding for bioinformatics software development remains small. For example, the NIH has a budget of $30 billion, of which an estimated 2–4% is allocated to computation and bioinformatics grants. We estimate that only a small fraction of this funding is used for big data software development. By comparison, the nonprofit Mozilla Foundation turns over $300 million annually for software development and FOSS promotion, and Google invests an estimated $6.7 billion annually in R&D. Private donors could, in principle, establish a foundation to support software development for integrative web-based services on large computer clusters. If investments in sharing data resources for biomedical research, such as the NIH Big Data to Knowledge (BD2K) initiative, with an annual budget of $24 million, and the European Bioinformatics Institute's smaller BioSamples project, were matched by serious investments in software development, maintenance and reproducibility, these projects would render better returns.

One way to solve the challenge is to wait for companies, such as 23andMe, that have made multimillion-dollar deals with pharma to realize large-scale investments and create big data solutions. However, such solutions would need to be purchased and, owing to their proprietary nature, would be difficult to adapt or benchmark. Another solution would be for biology funding agencies to establish initiatives for centralized software development. A different solution, and the one that we favor, is to use FOSS as a distributed development effort and develop collaborative software projects, such as those developed by the Linux, Mozilla and Apache foundations, which include private sector participation. For example, the goal of the Linux Foundation (which includes members such as IBM and Intel) is to fund Linux development.

Most of the bioinformatics software in use today does not scale for terabytes of data. R software programs typically load all data in RAM and suffer from its memory and runtime inefficiencies, and they are not designed for simultaneous use of multiple CPUs to speed up computations[3]. Where programming languages such as R, Python, Perl and Ruby are great for prototyping and quick analysis, they fail to deliver when it comes to big data processing. Solving the scalability problem will require embracing programming languages that are more efficient and have abstractions for multi-CPU computations[3], even if switching languages proves hard for most bioinformatician programmers.

Attribution for bioinformatics software development is also problematic. In a post titled 'You're not allowed bioinformatics anymore' on his blog *Opiniomics* (https://biomickwatson.wordpress.com/2014/07/21/youre-not-allowed-bioinformatics-anymore/), Mick Watson eloquently explains that bioinformatics is a scientific discipline in its own right and that bioinformaticians need career development. Ironically, in many of the most-cited biology research publications, there is a substantial bioinformatics contribution (usually the analytic method), often delivered as novel software solutions and data. However, it is rare for bioinformaticians to feature either as first or last authors on publications in high-impact journals. Authorship of community software projects can be troublesome as well, because the original authors tend to receive credit for the lifetime of the project, even when later code amendments and added functionality are equally or more important than the initial software. Lack of scientific attribution for software development hurts career development and can force bioinformaticians to opt for careers in traditional biology.

To solve the issue of attribution and related career development, we propose that the software contribution itself counts toward scientific track record. Every versioned software release and accompanying source code can be assigned a digital object identifier (DOI) with clear attribution for all contributors. The relative contribution of authors could be checked by visiting the software version control, such as that delivered by web services such as GitHub. This would make published software accountable, reproducible and citable. DOI citations could count as conventional citations, because they express the impact of a piece of software by its use.

In conclusion, our view is that to tackle the challenge of big biology software development, leading scientists need to acknowledge that software development is an integral part of research and not just an underpinning method. Projects need to promote bioinformatics collaborations and create scientific rewards. Universities need to increase their efforts to promote interdisciplinary research, to ensure that informatics is embedded in the life-sciences curriculum and encourage talented software developers and biologists to get involved in big data by tailoring individual career-development plans. Funding agencies can add institutional focus; emphasize collaborative FOSS approaches; build on existing grassroots initiatives[5]; create split funding streams for software and hardware; support maintenance of projects; encourage collaboration with experts in high-performance computing and software engineering; and fund larger projects dedicated to big biology software solutions.

ACKNOWLEDGMENTS

This document benefited from many reviewers. We especially thank K.W. Broman, T. Casci, V. Guryev, T. Seemann and J. Vilo for constructive comments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Pjotr Prins[1,2,3,7], Joep de Ligt[2,7], Artem Tarasov[4], Ritsert C Jansen[5], Edwin Cuppen[1,2] & Philip E Bourne[6]

[1]Department of Medical Genetics, University Medical Centre, Utrecht, the Netherlands. [2]Hubrecht Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), CancerGenomics.nl, Utrecht, the Netherlands. [3]Department of Nematology, Wageningen University, the Netherlands. [4]St. Petersburg State University, St. Petersburg, Russia. [5]University of Groningen, Groningen Bioinformatics Centre, Groningen, the Netherlands. [6]Office of the Director, The National Institutes of Health, Bethesda, Maryland, USA. [7]These authors contributed equally to this work.
e-mail: j.c.p.prins@umcutrecht.nl

1. Snijder, B., Kandasamy, R.K. & Superti-Furga, G. *Nat. Biotechnol.* **32**, 755–759 (2014).
2. Collins, F.S. & Tabak, L.A. *Nature* **505**, 612–613 (2014).
3. Trelles, O., Prins, P., Snir, M. & Jansen, R.C. *Nat. Rev. Genet.* **12**, 224 (2011).
4. Marx, V. *Nature* **498**, 255–260 (2013).
5. Möller, S. *et al. BMC Bioinformatics* **15**, S7 (2014).