The Bioinformatics Knowledge Commons: The Relationship between Governance Frameworks and Intellectual Property Rights in the Development of Open Source Bioinformatics Software

James Scheibner[1]

Abstract:

Whilst the majority of bioinformatics software, or software for the statistical analysis of genomic sequence data, is openly licensed, there is increasing interest in the commercialisation of bioinformatics software, and by extension the potential for patent protection. However, the impact of this commercialisation on other important features of scientific software development, such as sustainability and reproducibility, remains heavily debated. This article describes a mixed methods approach for examining bioinformatics software development using the knowledge commons framework, a modified version of Ostrom's Institutional Analysis and Design (IAD) framework. This framework involved both examining bioinformatics patents filed in the US, EU, Australia and New Zealand, as well as interviewing a broad range of molecular biologists and bioinformatics developers from these jurisdictions. Whilst there has been an incremental rise in the number of bioinformatics patents, interviewees broadly expressed a disinclination towards patenting due to the transaction costs associated with patent filing. Nevertheless, the results of this study reveal a tension between scientific software developers (and ecosystem stewards), and mainstream scientific researchers over both publication and commercialisation rights. The interviewees interviewed for this article reported the greatest success in open source development when intellectual property rights and organisational policies provided sufficient flexibility for open source licensing and collaboration, not only between different departments but also between different disciplines. In line with these observations, this article posits a set of preliminary observations on when open source licensing is appropriate for bioinformatics software development.

[1] PhD (Doctor of Laws) candidate, Centre for Law and Genetics, University of Tasmania. The author is extremely grateful to his supervisors Professor Dianne Nicol, Dr Jane Nielsen and Associate Professor Michael Charleston (research supervisor) for their contributions to this paper. In addition, the author wishes to thank Dr Linda Kahl, Dr Janet Hope, Dr Andrezj Killian, Professor Drew Endy and Professor Tania Bubela for their contributions.

1. Introduction:

This article examines the relationship between intellectual property laws and the governance of scientific software production in the field of bioinformatics and computational biology. In this context, bioinformatics refers to software for the statistical analysis and alignment of genomic and protein data (Hogeweg, 2011). This article approaches open source bioinformatics software as a common pool resource, or a shared resource that is available to all users on terms that encourage efficiency, equitable use and sustainability (Ostrom, 1990). In particular, the availability of content versioning software (CVS) such as Github has made it increasingly easier for scientists and scientific software developers to share source code (Crusoe & Brown, 2016, p. 27). This article uses a modified version of Ostrom's Institutional Analysis and Design (IAD) theoretical framework (henceforth referred to as the 'knowledge commons' framework) to analyse the role of intellectual property (IP) rights in the production of scientific software (Frischmann, Madison, & Strandburg, 2014; Poteete, Janssen, & Ostrom, 2010). This framework has been previously applied to both open source software and biomedical genomic data sharing consortiums (Contreras, 2011; C. Schweik & English, 2013). However, this article represents the first study that attempts to ascertain the factors behind successful open source bioinformatics development and determine whether these factors are affected by IP rights. In particular, this article posits that there may be additional factors in the development of a commons of scientific research software beyond those under consideration for the development of open source software or of a commons of research data.

Bioinformatics software development is a useful case study in understanding the evolution of a knowledge commons because of the long-standing support for open access to data in publicly funded genomic science (V. Stodden, Guo, & Ma, 2013). For example, data sharing arrangements such as the Bermuda Agreement mandated rapid release of human genome sequence data to ensure that this information entered the public domain and could not be commercialised (Stevens, 2015, p. 492). However, this article is largely concerned with the open source release of bioinformatics software rather than data and in particular whether IP rights may play a positive, negative or neutral role in this open licensing. Although prima facie open source licences operate through copyright law, which is a completely separate doctrine and focusses on different aspects of software to patent law, the process of acquiring or enforcing a patent may discourage open source licensing, which in turn may undermine the 'commons' of bioinformatics software.

To this end, this article relies on a 'multiple methods' approach to assess the impact of IP rights on the production of open source bioinformatics software (Nielsen, 2010, pp. 932–5; Poteete et al., 2010, pp. 15–20). Firstly, this article uses doctrinal analysis of patent and copyright law to analyse the boundaries of IP laws (Frischmann et al., 2014). Secondly, this article integrates this doctrinal analysis with an empirical analysis of bioinformatics patents using a search heuristic based on the International Patent Classification (IPC) marks for bioinformatics related patents held by research institutes and commercial research companies. This heuristic was used to ascertain the extent to which these institutions rely on patent protection and as a sampling strategy to identify interviewees (Park, 2012; Rasmussen, 2005; Vishnubhakat & Rai, 2015). Thirdly, this article includes an analysis of interviews with both academic and commercial bioinformatics developers and computer scientists. These interviews were used to explore the relationship between open source project success (and failure) and IP rights (Six, Van Zimmeren, Popa, & Frison, 2015). These interviews, in concert with bioinformatics patent data, were used to build a grounded theory on the relationship between open source licensing and bioinformatics patent acquisition.

Part 2 first explores Ostrom's IAD framework and the application of this framework to open source software as a knowledge commons. Part 2 then examines the relationship between this knowledge commons framework and IP rights, as well as the impact of IP rights on scientific software development. Part 3 articulates the methods described above for analysing the relationship between IP rights and scientific software development. Part 4 concludes by offering a comparison of the different private ordering strategies that can be used for the dissemination of scientific research software. Part 4 notes the difference between bioinformatics communities centred around academic institutes and commercial bioinformatics operators. Whilst the latter attempt to commercialise software itself, for the former the transactional costs for software are too great and accordingly there is an inclination towards open publication. Approaching the development of open source bioinformatics software from a governance perspective can help software developers understand what IP rights attach to co-produced software and how those rights influence the use of that software.

2. Theoretical Perspectives on the Knowledge Commons Framework and Scientific Software Development

In analysing scientific research software through a knowledge commons framework, this article links three separate perspectives on the development of open source scientific software. These are the development of open source software as a knowledge commons resource, the role of IP rights in building the knowledge commons and the development of scientific software ecosystems.

2.1 The IAD framework, the knowledge commons framework and open source software

Schweik and English have already conducted significant research into the factors that influence the success and failure of open source projects, noting that open source software shares many of the same characteristics as socioecological or natural common pool resources (C. Schweik & English, 2013; C. M. Schweik & English, 2007, 2012; C. M. Schweik, English, Kitsing, & Haire, 2008). These are sociological resources which can be jointly managed by a group who collectively use that resource (Poteete et al., 2010, pp. 42–3). Because socioecological resources can be overused, a socioecological common pool resource may experience a 'tragedy of the commons' when the resource is overused (Hardin, 1968). Ostrom formulated the Institutional Analysis and Design (IAD) framework to explain the rules that uses to ensure the sustainability of that resource (otherwise known as rules in use) and prevent a tragedy of the commons (Ostrom, 1990). Ostrom's IAD framework also emphasises the roles that different stakeholders play in the community governance of a common pool resource. It is this emphasis on community roles that allows open source software communities to be examined as a common pool resource through the lens of the IAD framework compared with other competing theories (such as regulatory capitalism or social capital theory) (Adler & Kwon, 2002; Aoki, 2008, pp. 2295–2302; Braithwaite, Nicol, & Hope, 2008). Schweik and English argue that an open source software project reaches a tragedy of the commons when development on the project stalls or the software is no longer used. This is because an open source software project, as an information resource, is not rivalrous and cannot be overused, but can fail when that project is abandoned or does not result in the production of useful software. Schweik and English's research identifies the sociological and economic factors that can lead to this situation, such as project size, characteristics and organisation (C. M. Schweik & English, 2012, p. 33). These success factors are tied to another important aspect of the common pool resource literature, which considers what IP rights should vest in a knowledge commons to guide control of that commons. Accordingly, the next section of this chapter will consider the potential for IP rights to encourage or discourage open source development in bioinformatics.

2.2 The knowledge commons framework and IP rights

In examining natural common pool resources, Ostrom (and later Frischmann and colleagues) note the importance of formal property laws in providing a framework for common pool resources (B. M. Frischmann, 2013, p. 392; Poteete et al., 2010, p. 31). Both Benkler and Frischmann have modified elements of Ostrom's framework to consider the role of IP rights in

protecting intangible and informational goods (Benkler, 2006). The knowledge commons framework has previously been used to examine the advantages and limitations of IP laws (specifically patents and copyright) in promoting or preventing the existence of a knowledge commons (Frischmann et al., 2014; Ghosh, 2007; Strandburg, 2008; Vertinsky, 2013). The literature in this area focuses on potential blocking effects of patents and copyright in scientific research and innovation (Barnett, 2014; Sag, 2014). Both patents and copyright may play a significant role in the commercialisation of bioinformatics software.

Compared to mainstream biotechnology research (where the imperative to seek patent protection is high for both commercial return and translational reasons) there may be different considerations in the development of scientific research software (Rai, Allison, & Sampat, 2009, p. 1520). As Howison notes, there is an ongoing conflict between the production of software as part of experimental design and the production of software as part of an ongoing engineering process (Howison & Herbsleb, 2011, p. 513). In addition, as Stodden notes, the incentive to patent can be a powerful countervailing consideration against publication if prior publication would have an impact on the patent application (Reich & Stodden, 2012). In other words, although an academic software developer may receive recognition through publication citation, there may be competing pressure between seeking further publication or acquiring a patent. Although Love and colleagues's survey of academic computer scientist patent holders suggest that many see patents as a substitute for academic citation, patents may also represent an opportunity for further commercialisation. This incentive may be particularly true for academics who wish to perform consulting work as part of their research (Love, 2013). However, the acquisition of a patent may also act as a disincentive to reuse the software, either through potentials or actual threats of patent infringement. Further, as discussed in the next section of this article, whether this disincentive is present may depend on the jurisdiction in which the researcher operates.

The potential for the existence of patent thickets raises questions as to the utility of software patents in academic research, particularly on foundational or 'upstream' algorithms (McLeod, 2016, p. 10). Stodden has conducted a comprehensive survey of academic researchers in machine learning to determine their perspective on the relationship between software source code sharing and IP laws. Both software source code and data were relevant in this study, as machine learning is a subset of computer science that involves developing algorithms that respond to statistical trends in data (Goth, 2015, p. 17). In Stodden's study, both code and data were considered given the role of machine learning and artificial intelligence algorithms to identify particular statistical trends in data. Stodden's study directly relates to this study, as bioinformatics represents an application of many commonly

used machine learning algorithms (Libbrecht & Noble, 2015). Stodden's theoretical underpinnings evolved from Merton's study of the norms of science (V. Stodden, 2010, p. 5). On the issue of IP rights, Merton's norm of communalism proposed that, in following 'the ethos of science', scientific proprietary rights only extend as far as naming and publication rights, and all other proprietary interests should be forgone.

However, as Stodden identifies in a later piece, the development of any form of scientific research software also involves aspects of software engineering. As a discipline, software engineering involves resolving solving real world problems using computational mechanisms and therefore involves an element of applied research (Snir, 2011, p. 39). Accordingly, software engineering can be best classified as use-inspired basic scientific research in Pasteur's quadrant, a leitmotiv that conceptualises scientific research depending on the intended outcome of that research (Snir, 2011, p. 40). Specifically, Pasteur's quadrant can be used to determine whether the benefits from a particular piece of research are related to the advancement of knowledge, industrial benefit or both (Gans, Murray, & Stern, 2017, p. 821).

	Considerations of Use	
Quest for Fundamental Understanding	No	Yes
No	No	Pure applied research (Edison)
Yes	Pure basic research (Bohr)	Use-inspired basic research (Pasteur)

In Pasteur's quadrant bioinformatics (as a subset of information and communication technology and software engineering) can be considered use applied basic research, as it encapsulates technology which supports further research in genomics and molecular biology (Cook-Deegan & Dedeurwaerdere, 2006, p. 315; Harvey & McMeekin, 2009, p. 484; McLeod, 2016, p. 3). In other words, although bioinformatics software was originally developed with the goal of applying informational approaches to genomics, bioinformatics software may also have a translational impact (such as in biomedicine or agriculture) (Harvey & McMeekin, 2009, p. 484). However, the blurring of the boundaries between basic and applied research in bioinformatics development represents a key challenge in determining the appropriate regime and scope of IP protection for bioinformatics tools

(Harvey & McMeekin, 2009, p. 487). These engineering aspects of software require sustaining software infrastructure long after the period of initial construction (V. Stodden, 2010, p. 3).

2.2.1 Jurisdictional Differences in Patent Rights and Enforcement

The propensity of developers to patent can also be broken down along jurisdictional boundaries. In the 1980s the US Congress enacted two laws, the *Stevenson-Wydler Act* and the *Bayh-Dole Act*, which permitted licensing arrangements between US universities and private sector researchers. As a result, US universities and research institutes could freely commercialise technology resulting from their research. On a technology neutral basis, economic studies demonstrate that patents are associated with higher academic citations for the university filing for the patent (Etzkowitz & Leydesdorff, 1997; Leydesdorff, 2000). A further factor under US patent law is the availability of triple damages for wilful infringement, or infringement where the infringer knew that they were infringing on a patent. The potential for punitive damages may discourage other inventors examining the boundaries of the patent system to determine whether they are engaging in infringement or not (Lei, Juneja, & Wright, 2009), thereby undermining the disclosure function of the patent system (Ouellette, 2011). Given the uncertainty about disclosure requirements for software, the presence of wilful infringement damages may act as a particular disincentive for software developers to examine the patent literature (Ouellette, 2017, p. 42).

This doctrine can be contrasted with some (but not all) EU patent systems, where generally speaking, the average award of damages is significantly smaller than is the case in the US and punitive damages are not available for wilful infringement (Chao, 2014, p. 80; Cotter, 2016, p. 270). Likewise, Australia does not have a doctrine of wilful infringement, although as with the UK and Canada patent damages are conditional upon the defendant having constructive knowledge (Cotter, 2016, p. 270; Weatherall & Jensen, 2005, p. 252). Therefore, although awareness of the presence of the patent is likely to be considered sufficient to amount to patent infringement, additional damages will not be available where there is evidence the patent infringer has deliberately sought information about the patent (Cotter, 2016, p. 270). The interview themes analysed in this article explicitly canvassed whether US bioinformaticians and computer scientists reported being less likely to examine the patent literature than EU, Australian and New Zealand bioinformaticians. A negative response in interviews to examining patents by US bioinformaticians compared to non-US bioinformaticians could indicate whether patent damages for wilful infringement would discourage US bioinformaticians and academics from releasing software or examining the patent literature for fear of patent infringement.

2.2.2 Jurisdictional Differences in Copyright Protection and Policy

The US copyright system provides flexibility for fair use of copyrighted materials, which includes the use of software and software components (*Oracle America Inc v Google Inc* (May 5 2016) Case No. CV 10-03561 WHA). However, the extent of this fair use protection largely depends on how the court evaluates the accused infringement according to four separate criteria. These factors include the amount taken, the nature of the work, the amount used or the market effect of use (17 U.S.C section 107). In contrast, the EU and Australia have implemented explicit exemptions to copyright infringement, both for research-related purposes and the development of interoperable software (both of which are pertinent to bioinformatics development) (Council Directive 2001/29, 2001 O.J. (L 167), *Copyright Act 1968* (Cth), sections 40, 47AB-H). A further divergence exists with respect to the treatment of compilations of data and database rights. In respect of compilations of data, both the US and Australia require a sufficient degree of originality for copyright to vest in either a collection of data or a database (*Feist Publications Inc v Rural Telephone Service Co* 499 U.S. 340 (1991); *IceTV Pty Limited v Nine Network Australia Pty Limited* [2009] HCA 14). On the other hand, the EU has implemented a dedicated sui generis regime for the protection of compilations of data (Directive of the European Parliament and of the Council of 11 March 1996 of the Legal Protection of Databases, Directive 96/9/EC, 1996 OJ (L77) 20, 21 (EC)).

From the EU perspective, *sui generis* protection for databases is highly controversial, with considerable uncertainty as to whether the rights extend to the compilation of data or data itself. The European Union First Evaluation revealed equivocal support for the *sui generis* regime and industry consultation recommended clarification of whether the regime extended to data that has been created and placed in a database or the collection of data (Evaluation of 96/9/EC; *British Horseracing Board Ltd v William Hill Organisation* Case 203/03). In other words, a database that only draws data from a sole source is unlikely to satisfy the threshold for protection, whereas a database composed of collected data is more likely to be protected (Carroll, 2015). Although these database rights have no impact on the protection of algorithms per se, bioinformatics is a data driven discipline and so therefore depends on access to multiple datasets to identify relationships between different genotypes and phenotypes (Dawyndt, Dedeurwaerdere, & Swings, 2006, p. 250). In particular, Reichman, Dedeurwaerdere and Uhlir note that adopting the distinction between sole data sources and collection of data, the vast majority of scientific databases would automatically qualify for protection under the sui generis regime. As a result, the use of bioinformatics tools to collate different data sets could amount to a prima facie breach of database rights that would not necessarily be protected by the

limited research exemption in the Database Directive (Reichman, Dedeurwaerdere, & Uhlir, 2016, pp. 336–342).

Reichman and colleagues note that it was for this reason that the US scientific community vigorously opposed the introduction of further database rights in the US beyond copyright protection. However, Reichman and colleagues are also sceptical of the fair use exemption contained in US copyright law and the extent of its operation with respect to scientific research. This perspective is supported by an earlier work by Sag, who notes that although the question of whether the alleged infringing use caused market harm remained a predominant consideration in early US case law on fair use, the question of whether the use was non-transformative has become an increasingly important within electronic research (Sag, 2009, pp. 1607–8, 1611–12). Non-transformative fair use is inextricably linked to both bioinformatics techniques and open source licensing in digital data science. Firstly, bioinformatics involves sequencing massive sets of genomic data, where the contents may be drawn from multiple sources (Preeyanon, Pyrkosz, & Brown, 2014, pp. 186–7). Secondly, applying bioinformatics software to these files in turn raises the question of whether this sequencing amounts to a transformative use of the original data sources. This situation may also depend on the licence attached to the data; although some primary databases may be open access or openly licensed, the licence on others remains uncertain (Contreras & Cuticchia, 2013, pp. 8–9).

To this end, the lack of precedent on the relationship between fair use and large scale digital technologies may act as an enormous roadblock on the use of bioinformatics software to sequence openly available data, even where this software is openly licensed (Reichman et al., 2016, pp. 333–4; Sag, 2009). Furthermore, the question of whether a work is transformative to determine fair use is closely tied to the question of whether a copyrighted work is a derivative work of an original work and accordingly whether the original creator's copyright extends to that derivative work. This is an important consideration with respect to determining the extent of the operation of open source licences with respect to software source code. Accordingly, the interview questions were themed to determine whether EU interviewees reported greater difficulty with the use of shared data than their US or Australian counterparts.

2.2.3 Copyright and Patent Protection: Irrelevant in the Governance of the Bioinformatics Commons

An alternative theory proposed by Ghosh is that rather than having either a positive or negative impact on the governance of a knowledge commons, patents and copyright (and by extension the jurisdictional differences for each regime) have no impact at all on knowledge commons formation

(Ghosh, 2007, p. 220). Ghosh's theory is based on Coase's contractual theory that formal legal property rights are irrelevant where the transactional costs of private bargaining are sufficiently low (Coase, 1960). When applied to a knowledge commons, Ghosh argues that because the resources in a knowledge commons can be shared easily, patents and copyright neither to prevent nor facilitate the formation of rules for governing that commons. Instead, the utility of copyrights and patents varies by context, and the parties that govern that resource may rely on contractual and informal norms where it is less expensive for the parties to do so (Ghosh, 2007, p. 221; Kapczynski, 2011). Contreras has already considered the implications of Ghosh's statements regarding the irrelevance of patents and copyright for genomic data sharing (Contreras, 2010, p. 1670). Contreras concluded that although early genomic science agreements mandated the rapid release of data unfettered by patent rights, the continuing use of this policy would largely depend on the context in which that data was used (Contreras, 2011, p. 111). In addition, Dedeurwaerdere notes that the managers of microbiological common pool resources rely on a dual licensing scheme for commercial and non-commercial use. This dual licensing scheme is equivalent to similar licences used in software development to provide an opportunity for the commercialisation of open source software (Dedeurwaerdere, 2006, p. 352). Accordingly, the interview questions articulated in this article were designed to interrogate whether patents and copyright lack relevance in the formation and maintenance of the bioinformatics commons. In addition, the interview questions were also designed to determine whether parties were able to negotiate around jurisdictional differences in patent and copyright policy.

2.3 The knowledge commons framework and scientific software development

The concept of multiple user groups with competing interests in the use and development of scientific software is reflected in the studies that examine the 'scientific software ecosystem' (Donoho, 2010). On the one hand, reproducibility (in the form of software that will produce the same output when run again or that can be installed and used on the same hardware) is an important consideration, given the heavy reliance of modern scientific research on computational methods (Stodden, 2009). On the other, maintaining sustainability (in the sense that software should be subject to ongoing maintenance so that the broader user base can continue to use it after the research it was developed for is concluded) is also an important consideration (Crusoe & Brown, 2016). These competing considerations create a conflict between those users who wish to continue to develop scientific software and those users who are interested in the results that software produces (Howison & Herbsleb, 2011, p. 513). To this end, Howison and colleagues identify four different types of stakeholders in the scientific software ecosystem, each of whom may have competing interests in the development of scientific research software (Howison, Deelman, McLennan, da Silva, & Herbsleb, 2015). Firstly, *scientist end users*

may be using their software to undertake scientific research within a scientific domain. The domain science within which a scientist end user uses bioinformatics software would be molecular biology or genomics. Secondly, *dedicated scientific software developers* may be involved in the production of scientific software components without considering the development of software further. Thirdly, *scientific computing administrators* may be responsible for administering high performance computing to make software available to a wider audience of scientific programmers. Finally, *senior scientists* or 'ecosystem stewards' may be broadly concerned with the overall functioning of the scientific software ecosystem (Howison et al., 2015, pp. 455–8).

This institutional heterogeneity means that the knowledge commons framework is an ideal mechanism to analyse these roles. Each of these stakeholders may have competing interests in terms of research objectives and organisational placement. Outside of information systems research modelling, these competing interests are echoed in the institutional tension described by Lewis & Bartlett between biologists and bioinformaticians, where the former is dependent on the latter for data (Lewis & Bartlett, 2013, p. 245). Each of these stakeholders may also have competing considerations on the role of IP rights. Within the field of bioinformatics, this tension may manifest itself as a contest between biologists (domain scientists and ecosystem stewards) and computer scientists (software developers and scientific software maintainers) over the role that IP rights play in balancing openness and commodification in the scientific software ecosystem (Calvert, 2008, pp. 383–4). However, before empirically examining the role that IP rights play in bioinformatics software development, it is first necessary to define the grounded theory model that can be used to illustrate these relationships.

2.4 Understanding a Knowledge Commons through Grounded Theory

Scientists use theories to explain observations made using evidence gathered over time (Ostrom, Gardner, & Walker, 1994, pp. 23–24). Because open source production is dependent on the institutions and people who control development, developing a theory of open source bioinformatics development requires a focus on the people and institutions involved in the development of software (Benkler, 2006, p. 177; C. M. Schweik & English, 2012, pp. 44–5). This approach to theory building dovetails the knowledge commons framework to explain how the rules that underpin the commons emerge. There are significant differences between open source development across different fields of software engineering, and pure quantitative approaches alone (such as case studies) with a focus on defined hypothesis testing can yield limited information about group dynamics in an open source project (Howison, 2009, p. 5; von Krogh, Spaeth, & Lakhani, 2003, pp. 1218–20). In addition, as Crowston and colleagues note, a key weakness of qualitative case studies on open source software is

a tendency to focus on large scale, successful open source projects (Crowston, Wei, Howison, & Wiggins, 2012) whilst ignoring less successful or failed projects. To avoid this oversight, this article uses a grounded theory approach to gather information about the development of open source bioinformatics software (Howison, 2009, p. 5; von Krogh, Spaeth, & Lakhani, 2003, pp. 1218–20). Grounded theory refers to a specific qualitative approach where theory emerges from the data through the data collection and analysis process (Strauss & Corbin, 1994, p. 273). This grounded theory approach can be contrasted with quantitative studies into the organisation of open source communities, which focus on using archived data to analyse communal interactions or surveys (Crowston et al., 2012; Howison, 2009, pp. 7–8). The methodology used to articulate this grounded theory is described in further detail in the methods section below.

3. Methods

3.1 Sampling for Interviews

Qualitative interviewing was used to build a grounded theory about the use of IP rights in bioinformatics research (Ostrom et al., 1994, p. 23). Appendix A includes the interview questions that were used. Interviews were conducted with 23 bioinformaticians and computer scientists using a snowball sampling method (Fusco, 2011, p. 14; Goodman, 1961; Love, 2013, p. 299; Rai et al., 2009, p. 1528). Approximately 18% of potential interviewees contacted responded and agreed to an interview. These snowball samples were cross referenced using the BioDirectories website maintained by Dr Geoffrey Routh (http://www.grouthbio.com/), which provides a list of both scientific software providers and genomics research institutes. The resulting list was also crossed referenced with the patent search described below to identify bioinformaticians and computer scientists who were listed as patent applicants.

The purpose of this approach was to isolate the bioinformaticians and computer scientists who had competing interests between publishing open source bioinformatics software and seeking patent protection (V. Stodden, 2010, pp. 20–21). Unlike Stodden, who exclusively interviewed computer scientists and machine learning experts from the United States, this study considered interviewees from the United States, Europe, Australia and New Zealand. These jurisdictions were selected to determine whether any of the jurisdictional differences described in Part 2 would affect how the interviewees interacted with intellectual property laws. This patent search heuristic is described in further detail below.

3.2 Patent Searching Algorithm

The CAMBIA Foundation's Patent Lens was used to find bioinformatics patents that had been filed for by academic developers (Jefferson, 2006, p. 28, https://www.lens.org/lens/). The results from this search were cross referenced with results from the patent databases provided by the World Intellectual Property Office (WIPO), the United States Trademark and Patent Office (USPTO), the European Patent Office (EPO), IP Australia and IP New Zealand to confirm that they were accurate, as well as to gather additional details about the patent applicant and patent classification.

A challenge in defining a heuristic for searching for bioinformatics patents is the actual definition of bioinformatics, given the ubiquitous nature of bioinformatics software in life sciences research and resultant definitional uncertainty (Wiechers, Perin, & Cook-Deegan, 2013, p. 87). This definitional uncertainty makes it difficult to identify what is truly a 'bioinformatics patent'. Indeed, there is no accepted definition as to what exactly is a 'software patent' (Rai, Allison, & Sampat, 2009, p. 1523). However, significant research has been conducted into developing search heuristics for identifying software patents (Bessen & Hunt, 2007; Hall & MacGarvie, 2010) using either keyword searching patent claims or searching by IPC marks. The World Intellectual Property Office (WIPO) issues IPC marks under the *Strasbourg Agreement 1971* to categorise patents by technological field, both as an aid for international patent filing and prior art searching (International Patent Classification (IPC), World Intellectual Property Organization (WIPO), http://www.wipo.int/classifications/ipc/en/).

Bessen and Hunt argue in favour of keyword searching for generic software patents, noting that changes in the classification scheme make it difficult to conduct longitudinal surveys of patent activity in a single technological field, and that patent attorneys file strategically to influence the examination process (Bessen & Hunt, 2007, p. 166). However, patent attorneys may also use idiosyncratic language for patent claims (Bessen & Hunt, 2007, p. 164; Bessen & Meurer, 2009, pp. 204–5), which can lead to a high number of false positives and false negatives for classification. In contrast, searching by IPC marks allows for comparisons to be drawn between the patent filing strategies of different examiners between technological fields. Vishnubhakat and Rai have demonstrated the effect of examiner background and education with respect to bioinformatics and generic software patents filed at the USPTO filed under different USPC marks with different prior art examination divisions (Vishnubhakat & Rai, 2015, p. 221). Further, Park notes that IPC searching is the most effective strategy for searching in convergent disciplines such as bioinformatics, which has been influenced by both computer science and genomics research (Park, 2012, p. 271; Vishnubhakat & Rai, 2015).

For this reason, bioinformatics patents held by research institutes and commercial software developers by a primary IPC mark of G06F19/10 to G06F19/28, which were introduced with the 2011

edition of IPC and which were used to reindex bioinformatics patents that were filed for prior to the 2011 edition (Hall & MacGarvie, 2010, p. 997). Of the search results returned from the Patent Lens, those patents that did not have a primary mark of G06F19/10 to G06F19/28 were removed. In this context, primary IPC mark refers to the mark that best represents the patented invention (*Guide to the International Patent Classification 2017 edition* at http://www.wipo.int/export/sites/www/classifications/ipc/en/guide/guide_ipc.pdf, section 156). Applying this search technique removed patents referring to public domain bioinformatics algorithms, such as BLAST, and left a set of patents on novel bioinformatics algorithms. The remaining patents identified were then filtered to remove pharmaceutical companies that did not specialise in bioinformatics research. As suggested by Rasmussen, these patent applicants were more likely to specialise in cross licensing with other pharmaceutical firms and therefore less likely to have competing considerations between patent filing and patent applications (Rasmussen, 2010; Reich & Stodden, 2012). In other words, by examining software held by research institutes and software developers, the patented software identified clearly sat in the use inspired basic research segment of Pasteur's quadrant as software that could be both published and patented (Gans et al., 2017, p. 821).

In addition to providing an overview of the companies and research institutes that were filing for bioinformatics patents, as well as the bioinformatics patents that they were filing for, this search heuristic also provided a list of potential interviewees in listed inventors. However, this search strategy could not identify the potential issues with respect to copyright management. Accordingly, this article will also rely on interviews to assess the role of copyright in the development of open source bioinformatics software.

3.3 Qualitative Interviews

3.3.1 Recording and Analysis of Data

Interviews were 45 to 60 minutes long and were conducted over a VoIP client, such as Google Hangouts or Tox. A list of the interview questions used is included in Appendix A below. These interviews were supplemented with information about the technology transfer policies held by each institution or software publisher where available.

The transcripts were anonymised and analysed using Transana (https://www.transana.com/), following the approach used by Schweik and English (English & Schweik, 2007, p. 2). The interview questions were amended following the initial interviews to focus on how bioinformaticians and computational biology software developers focus on promoting sustainable software development.

Educational Background	Number of Interviewees
Computer science/Statistics/Mathematics/Bioinformatics	13
Biology	7
Chemistry	1
Physics	1
Other	1

Table 1: Educational Background

Jurisdiction	Number of Interviewees
Australia/New Zealand	6
United States	12
Europe	5

Table 2: Jurisdiction where interviewee located

4. Results

4.1 Patents Statistics

A search of the patent literature revealed that the number of patents on bioinformatics algorithms and database management software filed by research institutes and software developers was comparatively small relative to university software patent activity. This trend can be demonstrated by considering bioinformatics patent activity in the US as an example. In a survey of the top 20 computer science universities in the United States, Stodden and Reich identified 891 software patents granted between 2000 and 2009. Likewise, Rai, Allison and Sampat identified a ten-fold increase in the number of software related patents filed from 1982 to 2002, due to case law that allowed for more liberal patent filing strategies (Rai et al., 2009, pp. 1525, 1551). The search documented in this article identified 129 bioinformatics algorithm patents granted in the US to universities, research institutes and software developers between 2000 and 2016. These 129 patents were drawn from a sample of 206 patent applications. This grant rate of 62.6% is lower than the 70% average success rate for patent

filing identified by Lemley and Sampat (Lemley & Sampat, 2008, p. 193). There was a modest increase in the number of patents granted over this period, from 4 in 2006 to 28 in 2016.

The relatively low number of patents that have been granted on pure bioinformatics inventions suggests some interesting trends regarding the patenting patterns of bioinformatics software by software developers, universities and research institutes. Love and colleagues note that more than 50 percent of patent filing US computer science and electrical engineering academics regarded patents as providing insufficient motivation to produce more or better research (Love, 2013, pp. 314, 320). Nevertheless, Love and colleagues's study still concludes that academic computer scientists grudgingly accept the presence of software patents. On the other hand, the raw statistics on granted and filed bioinformatics patents suggest that bioinformaticians and computational biology software developers involved in computational biology often bypass the patenting process. These raw statistics suggests that bioinformatics developers rely on other mechanisms to support sustainable software development. These strategies are elaborated in further detail with the interviews described below.

4.2 Interview Results

4.2.1 Disinclination towards Patent Protection and 'Cultural Openness' in Bioinformatics

As mentioned in the methodology, the purpose of the thematic approach for interview questions was to determine whether attitudes towards IP rights and acquisition for bioinformatics algorithms differed depending on the background discipline for the interviewee. A key theme that emerged from the interviews was a disinclination towards patent protection amongst interviewees who listed their educational background as computer science or physics. This divide supports an expectation stated in Part 2 that researchers from a biology background (where commercialisation of research is necessary for technologies to be translated from an academic to a practical setting) would be more amenable to IP protection. Correlating closely to the responses received by Love and colleagues, many interviews argued that because they had received taxpayer funding '[the public] should not have to pay for [their research] again.' In contrast, those who had received training in the biological sciences, or a combination of both, were more willing to acknowledge that, at least in theory, patents should be available for bioinformatics software and methods. On the other hand, those interviewees from a computer science or bioinformatics background noted many of the software developers and statisticians joining early genomics initiatives were from free and open source software backgrounds, and therefore brought their perspectives on software sharing to their new institutions. Statements of this nature conform with Stevens's observations regarding the early evolution of computational biology, where large sequencing centres eschewed patenting in favour of rapid public domain

sequence releases (Stevens, 2015, p. 467). These sharing practices have been stimulated by increasing improvements in network technology, from the initial development of Wide Area Networks (WANs) that enabled distribution of early protein sequence information to Content Revision Systems such as Github that allow for researchers to collaboratively work on software and papers (Crusoe & Brown, 2016, p. 1; Stevens, 2015, pp. 478–9).

4.2.2 Institutional Preference towards Patenting and Bioinformatics Patenting

One of the key factors that Stevens attributes towards this disinclination towards patent protection was due to rapid sequence release strategies by government funded research institutes to prevent commercial or semi-commercial operators from acquiring patent protection on genome sequences (Stevens, 2015, p. 467). However, this perspective was not shared by all interviewees; several interviewees who worked at semi-commercial institutes with high patenting activity and industry collaboration (particularly for synthetically edited genome sequences) mentioned that '[the institute] allowed them to continue releasing open source software without seeking patent protection.' Rather than being a mandated policy that all bioinformatics software had to be commercialised, these interviewees mentioned that their parent institutes 'left the decision [on whether to commercialise or not] to us.' This perspective was supported by the comments of a computational chemistry researcher, who was at pains to distinguish between software patents and other research (such as chemical compounds) that were capable of being subject to patent protection. Interestingly, these individual perspectives seem to run contrary to trends reported by Rai and colleagues that US universities which acquired large numbers of software patents were also likely to patent aggressively across technological fields (Rai et al., 2009, p. 1526). Although these responses do not nullify the effect that pre-existing open source development norms have on the development of bioinformatics software, they demonstrate that open source approaches can exist where there might otherwise be a strong institutional emphasis on patent protection.

4.2.3 Combining Patenting and Open Source Approaches to Bioinformatics

Not all interviewees saw an open source approach to development as incompatible with patent protection. An emergent theme, particularly amongst interviewees who wrote both proprietary and open source bioinformatics software, was the notion of using patents as a private ordering strategy to guide software development towards certain research goals. One interviewee expressed frustration over the idea that software was to be 'used and then never touched again [once a particular grant had been closed]' and wanted to develop his own software for his own research goals. Accordingly, this interviewee developed his own spinoff company and was planning on acquiring one or two patents

for methods underlying algorithms which he deemed to have sufficient novelty, and releasing the remaining software under an open source licence. Responses of this nature cut to the heart of the institutional tension within bioinformatics development, where bioinformaticians, as producers of technical software, are subordinate and dependent on biologists for biological data and experiments (Lewis & Bartlett, 2013, pp. 243, 249). For at least one interviewee, their frustration at their dependence on molecular biologists for experimental designs and data to develop software manifested itself in pursuing an IP and commercialisation strategy revolving around patent protection to provide funding for a research agenda. This interviewee planned to use the funds from commercial software development to fund their own original research rather than rely on other biologists for experimental design or data. This response was indicative of the potential for tension between different stakeholders.

The other listed interviewees who held patents offered more conventional rationalisations for seeking patent protection. One interviewee who worked for an institute relying on dual licensing explained that although they had sought patents to protect the underlying methods, the unmodified software was still available for academic use:

> *We treat open source software as almost a trial version of what we can offer. If a researcher comes to us with a specific problem that they want to solve with our software, we fork it and develop a bespoke version that they pay for. With the bespoke version, we then attach headers to the bespoke forked version so that if we see this version online, we know that [infringement has occurred]* (2017).

Despite this 'digital watermarking' approach to dual licensing, this interviewee still expressed scepticism about the use of patent protection for bioinformatics software, noting that the process of acquiring patents had been 'time consuming' and questioning whether it had its intended benefit. Another patent holder who is also heavily involved in open source bioinformatics software development described the acquisition of the patent as follows:

> *We have chosen to patent and commercialise an aspect of basic research that we did… we realised in the course of doing something that was a basic research project was something that might have commercial value. [Although the release of the bioinformatics invention is a long way down the line] it has something of value that is worth being protected. Most bioinformatics software isn't like that… someone has a better way of doing [the task you are attempting to commercialise] in a few months time.*

This delineation between bioinformatics software based on value raises questions about the sustainability of bioinformatics development practices. For other interviewees, the value of patented software received very little attention. One interviewee, who had transitioned from a commercial research institute to a publicly funded research institute, argued that in their old place of employment, patents were little more than a 'key performance indicator' and did not contribute any value to encourage the development of software. There was also significant concern amongst scientific software developers about whether the existence of patents could act as a disincentive for software reuse. As one interviewee noted:

> *Getting the software out there and lowering the barriers to entry is the main thing… [Particular subfields of bioinformatics] are arguably tiny markets, and the community worldwide for each specific area of research. If I have a thousand users, I consider that to be a success. So I think the issue of open source versus closed source… I think open source is a better way to get people to use the software… To acquire patents would keep people from using that software.*

This interviewee mentioned a particular university software patent that, due to institutional technology transfer requirements, was unavailable for use by the rest of the scientific community. This interviewee also mentioned that whilst the technology transfer staff at his institute were prepared to file for software patents, the developer would have to initiate this process and in his case '[I would prefer] to get the software out there'. This statement is reflective of both Rai et al and Love et al's observations regarding the disinclination of academic software developers to seek patent protection (Love, 2013; Rai et al., 2009). The disinclination against patent protection is a distinct question from the commercialisation of open source software (and data) which will be discussed in sections 4.2.5 and 4.2.6 of this article. However, this article will first turn to address the question of whether bioinformatics researchers and software developers read patents as a source of technical information, and what impact this has on either supporting or prohibiting open source development.

4.2.4 Bioinformatics Patents and Examining the Patent Literature for Disclosure

All the interviewees were asked whether they were aware of the presence of patents when they developed software, whether they searched for patents or whether they regarded patents as a useful source of technical information. This directly relates to the impact that both wilful infringement requirements and the lack of source code disclosure in patents have on researcher behaviour. This builds off the research of Larrimore, who conducted a survey of academic scientific software developers and found that 30% reported identifying useful information in patent applications (Ouellette, 2017, p. 422). This was in spite of the existence of the wilful infringement doctrine in US

patent law that provides plaintiffs with the right to seek additional damages where the alleged infringer is aware of the presence of the patent. It is therefore somewhat surprising that only three of the researchers interviewed for this study (two of whom had previously worked in joint academic commercial enterprises) when asked admitted that they examined patents in their field. The first noted that they were 'aware of [significant] patents in the field' but that they left the matter of determining the question of patent infringement to in house counsel. The second interviewee (whose experience with seeking a bioinformatics patent is described below) mentioned that they and the rest of their research team used the patent filing process to assess what other researchers had been investigating in the same field. The third interviewee (who described the rationale behind the patenting process being to free themselves from being bound by the research interests of biologists) reported examining external patent activity to determine the activity of other researchers.

However, none of the other interviewees who described their work as reliant on open source development admitted to searching for patent information. Some argued strongly for releasing software under a restrictive open source licence (such as the GPL version 3) due to the requirement under a restrictive licence to relicense derivative source code as well as express patent licences prohibiting patent royalty collection. Nevertheless, a common theme amongst interviewees was the notion that 'this is basic science, so there shouldn't be any patents in this field' as a means of seeking protection against potential patent infringement action. Another interviewee gave a more nuanced explanation, that '[having acted as an expert witness in a patent infringement trial] my perspective is that you need a lot of funding to run a patent infringement trial and I don't know that this would be justified for [the majority of bioinformatics software]'. This statement correlates with Asay's observations that the risk of patent litigation to community produced open source software is not as high as suggested by open source advocates (Asay, 2014, p. 435). Nevertheless, despite the relatively low risk of potential patent litigation, some interviewees (particularly software developers) expressed concern about the presence of patents and their potential impact on community engagement. Interestingly, none of the interviewees had considered the role that standards (and related patents) might play in hindering or helping open innovation; to quote one of the patent holding interviewees mentioned above:

> *I don't think we've established a standard… as long as other people want to use those, we've got a right to do that. But that's not to say that everyone else has to follow our lead. The subject is not quite ready for standards yet. The main reason for acquiring the patent [and licensing it to a company] was because we deemed the techniques disclosed valuable.*

In part, this answer reflects another aspect of patent protection; that is, the fast pace at which bioinformatics development moves and the difficulty in developing bioinformatics standards when also seeking patent protection (van Zimmeren, Rutz, & Minssen, 2016, p. 1479). However, the development of standards can also create issues for long term assignment of copyright, as discussed in the section below.

4.2.5 Copyright Protection for Bioinformatics Software and Sustaining Community Engagement

Some of the interviewees were significantly more concerned about the impact of copyright regimes and restrictive licensing on open source bioinformatics development. A consistent theme here was the time delay involved in negotiating the cross licensing of bioinformatics software source code between the technology transfer offices at different institutions. Because of the ease with which a developer can create a new source code repository using a content revision system many interviewees (particularly developers) expressed frustration at the time taken to negotiate source code sharing agreements through their technology transfer offices.

Another challenge in source code licensing involves ensuring that the perspectives of the original developer of the software are respected. One interviewee who had transitioned into the business development department within their institution mentioned that initially it had been a challenge to ensure that bioinformaticians and software developers who openly licensed their software did so in a way that accurately reflected how they intended their software to be used. This interviewee mentioned that over time, teams associated with different bioinformatics institutes had developed informal strategies on how they would licence source code. These strategies depended on whether the software was designed to be useable downstream (such as if the software was a library file) or whether the software was part of a community initiative to create a uniform software package to be abstracted to a variety of research situations. In the former case, a permissive licence such as a BSD or MIT licence would be more appropriate due to the absence of restrictions on patenting of downstream research in these licences. In contrast, GPL licence would be better suited to the latter case due to the requirement for contributors to re-licence their modifications of the source code under the same GPL. Consistent with the knowledge commons framework, although the official licensing strategy provided an overarching mechanism for the sharing of software, informal norms were used to enforce software sharing and use. Some of these norms spring from the scientific norms described by Stodden that have evolved with respect to citation networks and attribution; as such, the attribution requirements in BSD or MIT licences (which require enforcement) carry additional weight.

4.2.6 Funding Strategies for Bioinformatics Software and Sustaining Community Engagement

Another strategy used to enforce open source licensing for software is through funding sources; two interviewees mentioned that they modified employment contracts for laboratory staff and grant requests to mandate the use of a specific open source licence. Returning to Howison's split of stakeholders into scientific software developers, scientific researchers, scientific computing managers and ecosystem stewards, this funding mechanism created an additional layer of enforcement for the open source licence. In addition, this funding mechanism created an additional layer of financial security for scientific software developers, whose contributions may or may not be included as part of a domain science publication (Howison et al., 2015, p. 459). Where available, this funding model could represent a solution to the conflicting goals that different stakeholders in the scientific software ecosystem have.

However, there are still improvements which can be made to assist connections between technology transfer offices and bioinformaticians. Agreements for software licensing and collaborative open source development are often predicated on the assumption that source code is protected under copyright law as opposed to patent law. The interviewee mentioned in the previous paragraph discussed how it was conceivable that one institution might hold the copyright associated with the software source code for the project, but another institute might hold the patent associated with the innovative concept embodied in the algorithms underlying the source code. This split in ownership creates enormous uncertainty as to determining the scope of open source licensing. Beyond defensive publication, very few interviewees (including interviewees who might be described as senior scientific staff) had considered any form of standards licensing for patented bioinformatics methods.

Closely related to these non-legal norms is the question of copyright assignment to particular universities or research institutions. The presence of legal rights attached to software is meaningless without the power to enforce those rights. Some interviewees discussed official copyright assignment strategies at their institutions to ensure that external contributors at other institutions could share source code. However, these contributions are largely executed through what one interviewee described as 'pairwise' negotiation between different institutions. In other words, rather than a formal agreement executed for ongoing sharing between two institutions (such as through a technology transfer office), the developers themselves would negotiate an agreement for data sharing. Without these agreements (or with agreements which lack adequate support for open source licensing), interviewees occasionally described working outside the scope of their technology transfer department or directly disobeying policy on source code sharing to work with other institutes. These

problems were commonly reported for joint projects between academic institutions and commercial developers.

4.2.7 Bioinformatics Software and Copyright Protection for Databases

An additional level of contention amongst interviewees concerned copyright and other IP protection for data sequenced with bioinformatics software. Surprisingly, none of the European interviewees mentioned *sui generis* database rights as having a significant impact on research. However, all European interviewees were opposed to the concept due to the lack of a fair use exemption and a lack of information on derivative works under the *sui generis* regime. In addition, the majority of interviewees appeared to be more concerned about copyright protection for bioinformatics sequence data. As one interviewee noted:

> *[The whole field of bioinformatics] concerns itself with data transformation… without being derogatory, there's a lot of hacking of different files together [to view] different associations.*

At least one interviewee argued that the easiest way to resolve this conflict was through the development of common bioinformatics file formats that could aid with identifying statistical frequencies, such as genotype and phenotype distribution within a population. Such languages do not need to be necessarily specific to bioinformatics; for example, the Common Workflow Language (CWL) was designed as a standard for encouraging homogeneity in workflow language standards (Cohen-Boulakia et al., 2017; Karczewski et al., 2016, p. 157). One interviewee who advocates using CWL in bioinformatics and computational biology noted that different storage formats were associated with different custom licences for what should otherwise be publicly available data:

> *If you look at the data licences, the majority are custom and therefore need legal interaction for combination. After a year, we've failed to openly licence the publicly licensed data set. Licence information is sometimes great but most of the time it is undetermined. 95% of [data licences] are custom which require pairwise negotiation.*

Given the enormous public investment in biomedical research, this interviewee argued that the lack of standardised formats for data sharing amounted to a serious impediment on reproducibility, and perhaps more of an impediment than the existence of overlapping patents on bioinformatics methods.

In response to this concern, scientific software developers have had to adapt different strategies to process the data that they receive to assist with the development of new software. An interesting connection which confirmed Howison's classification of different stakeholders in the production of scientific research software was the relationship between scientific software developers and biologists

who collect genomic data (Howison et al., 2015, p. 459; Lewis & Bartlett, 2013, p. 254; Penders, Horstman, & Vos, 2008, p. 748). 6 of the interviewees with a computer science background mentioned that they received access to data collections in exchange for early versions of sequencing software for optimisation and testing purposes. However, this reciprocal style arrangement still requires negotiation between the data repository and the software developer. The interviewee above also noted that the complications with this model are compounded by the difficulty in distinguishing between academic and commercial uses of data, which can further undermine sustainability:

> *Ideally we would have a situation where source providers could share their data for commercial and academic uses so that it could be seen as more valuable by the commercial entities. [Whilst mandating public domain licensing such as Creative Commons 0 licence] would make things easier for everyone, because it puts everyone on an even footing, it completely destroys any prospect for sustainability… it is less sustainable and not self correcting.*

These problems can in part be attributed to the inflexibility of the overarching statutory regimes for copyright and patent protection. However, this interviewee (along with others) expressed concern that adopting a blanket public domain model may have implications for the sustainability of scientific software production because software developers may not receive appropriate recognition for the use of their work in research. This concern is more broadly reflective of the significant divide between what is nominally academic, and what is commercial software development. As discussed previously, one of the key justifications for the release of open source software more broadly is recognition or signalling that may provide an open source developer with further work in the future (C. M. Schweik & English, 2012, pp. 51–52). This signalling effect in turn provides an incentive for developers to continue to develop open source software, thereby sustaining the open source project and preventing an anticommons. If these efforts are not recognised, the case for scientific software developers to receive ongoing research funding may be undermined, adversely affecting the sustainability of bioinformatics software development. The question of sustainability versus reproducibility is a matter that will be considered in greater depth in the conclusion of this article.

4.2.8 Managing Bioinformatics Development Commercially

The thematic analysis of these interviews suggested ambivalence towards patents across other technological fields as opposed to scepticism towards bioinformatics software patents specifically. Although this sample was largely composed of academics involved in bioinformatics research, a general sentiment echoed by the group was the potential for patenting activity to discourage use of bioinformatics software. Similar sentiment was expressed with respect to commercialisation of

bioinformatics software, although some interviewees were more than willing to use commercial or proprietary software if that software is superior to an open source equivalent. For example, the PAUP software package used in phylogenetics research is released under a dual commercial and academic licence without patent protection (Howison & Herbsleb, 2011, p. 521). However, one interviewee noted that the specific example of PAUP was the exception rather than the rule due to its highly user friendly graphical user interface. A clear thematic trend from the interviews amongst computer scientists and bioinformaticians was that most bioinformatics algorithms are simply not sufficiently commercially valuable to be released under a proprietary licence, let alone subject to the arduous patent application process. Instead, these computer scientists and bioinformaticians measured success through other measures, including raw usage rates for software or user feedback on the quality of the software. These success requirements will be discussed in further detail below.

4.2.8 Managing Bioinformatics Development as a Knowledge Commons

Underlying the conflict is the ongoing debate as to whether computational reproducibility (insofar as creating software that can be reinstalled and run to produce the same results) is best served through the open licensing of software. For other scientific software, measurements such as the proportional size of the user base for a software package or collaborative development between different institutions may represent more conclusive measures of success (such as ongoing contributions to a project) than commercialisation. It was also apparent from the interviews that there are competing considerations in the development of a robust commons of bioinformatics research software. On the one hand, the interview data suggested that biologists and scientific advisors have the broad objective of designing software to solve specific biological inquiries as opposed to using bioinformatics software to represent computational problems. Applying these solutions may in turn entail commodifying or commercialising either the software itself or its application in a clinical or translational setting.

However, commodifying bioinformatics software (and to a lesser extent data) largely depends on how that software can be best commodified. Competing against commodification are the desires of software developers and scientific software maintainers, who may have their own publication requirements or may see commercialisation or patent protection as a threat to the sustainability of the bioinformatics software ecosystem. Pushing against this are the wishes and policies of both ecosystem stewards and scientific computing administrators, who establish organisational policies on sharing source code and data. Outside these organisational policies, a thematic analysis of the interviews revealed that different organisational stakeholders developed their own norms for the exchange and use of data as an aid to the experimental design of bioinformatics software. The presence of these

pairwise arrangements would in part confirm Ghosh's theory that parties can develop their own rules for the governance of common pool resources when transaction costs are low enough (Ghosh, 2007, p. 223). These rules in use were most likely to lead to the creation of successful open source projects where the parent institute had a flexible organisational policy on collaboration. On the other hand, interviewees at institutes which pursued an aggressive or technologically neutral patent acquisition approach or which had inflexible technology transfer policies were more likely to report an inability to collaborate. Therefore, in an interdisciplinary field such as bioinformatics, the acquisition of IP rights itself does not act as a bar on open sharing. Instead, it is the consequences of inflexible or technology neutral approaches to technology transfer that have the greatest impact on open source licensing.

5 Conclusion and Preliminary Policy Considerations

The interviews described in this article would suggest that the prevailing attitude towards bioinformatics (at least in academic or industry-academic collaborations) is towards open source or public domain licensing, with commercialisation of bioinformatics algorithms being a rarity. The responses from these interviews would suggest that for the most part, patents do not have a significant impact on the bioinformatics software commons (a statement supported by the survey of patent activity in bioinformatics), and that copyright (through open source licensing) plays a predominant role. However, as for the genomic data sharing commons, the role that IP rights play in open source success in bioinformatics is dependent on both the technology under consideration and the institutional rules where the software is being developed. Ultimately, interviewees reported that successful open source bioinformatics projects were most likely to emerge where there was broad institutional support and policies to support open source licensing, as well as support for scientific software developers to receive recognition for their work. Sometimes, these policies emerged from pre-existing cultural norms from open source software communities or a recognition of the difficulty of commercialising bioinformatics software by other stakeholders. However, in the absence of flexible policies, the transaction costs associated with sharing of source code or data quickly became burdensome for parties. Accordingly, at an institutional level ecosystem stewards and staff scientists should restructure their technology transfer policies to explicitly support open licensing of source code. Allowing scientific developers to work from the 'ground up' and develop their own pairwise arrangements can in turn support the development of a successful bioinformatics commons.

Accordingly, the development of these commons rules in turn requires a more nuanced approach between both different stakeholders and different forms of technology. Applying a uniform public

domain licensing model to all bioinformatics software (irrespective of user base size or intended software use) is unlikely to lead to maximum utilisation of software. On the other hand, software development and bioinformatics development is too far removed from other forms of research and development due to the low transactional cost of development to apply a uniform commercialisation approach that is also applied to other forms of technological research. This article demonstrates that the solution is to find a balance between the two extremes that recognises the publication rights of all stakeholders involved in the scientific software ecosystem, as well as the inherently interdisciplinary nature of bioinformatics development. Although standardised file formats and workflows can encourage standardised licensing of data (which in turn can encourage reuse of data), licensing of bioinformatics software must be explicitly linked to the objective behind the development of that software. Developing an appropriate licensing strategy depends on both co-operation between bioinformaticians and technology transfer offices to develop more effective reach through agreements for collaboration that can cope with differences in IP laws between jurisdictions.

Bibliography

Adler, P. S., & Kwon, S.-W. (2002). Social Capital: Prospects for a New Concept. *The Academy of Management Review*, *27*(1), 17–40. https://doi.org/10.2307/4134367

Aoki, K. (2008). Free Seeds, Not Free Beer: Participatory Plant Breeding, Open Source Seeds, and Acknowledging User Innovation in Agriculture. *Fordham Law Review*, *77*, 2275–2310.

Asay, C. D. (2014). Enabling Patentless Innovation. *Maryland Law Review*, *74*, 431.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale University Press.

Bessen, J., & Hunt, R. M. (2007). An Empirical Look at Software Patents. *Journal of Economics & Management Strategy*, *16*(1), 157–189. https://doi.org/10.1111/j.1530-9134.2007.00136.x

Bessen, J., & Meurer, M. J. (2009). *Patent Failure: How Judges, Bureaucrats, and Lawyers Put Innovators at Risk*. Princeton: Princeton University Press.

Braithwaite, J., Nicol, D., & Hope, J. (2008). Regulatory captialism, business models and the knowledge economy. In J. Braithwaite (Ed.), *Regulatory Capitalism: How it Works, Ideas for Making it Work Better*. Edward Elgar.

Calvert, J. (2008). The Commodification of Emergence: Systems Biology, Synthetic Biology and Intellectual Property. *BioSocieties*, *3*(4), 383–398. https://doi.org/10.1017/S1745855208006303

Carroll, M. W. (2015). Sharing Research Data and Intellectual Property Law: A Primer. *PLoS Biol*, *13*(8), e1002235. https://doi.org/10.1371/journal.pbio.1002235

Coase, R. H. (1960). The Problem of Social Cost. *The Journal of Law and Economics*, *3*(4), 1. https://doi.org/10.1086/674872

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., … Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2017.01.012

Contreras, J. L. (2010). Data Sharing, Latency Variables, and Science Commons. *Berkeley Technology Law Journal*, *25*(4), 1601.
Contreras, J. L. (2011). Bermuda's Legacy: Policy, Patents, and the Design of the Genome Commons. *Minnesota Journal of Law, Science & Technology*, *12*, 61.
Cook-Deegan, R., & Dedeurwaerdere, T. (2006). The science commons in life science research: structure, function, and value of access to genetic diversity. *International Social Science Journal*, *58*(188), 299–317. https://doi.org/10.1111/j.1468-2451.2006.00620.x
Crowston, K., Wei, K., Howison, J., & Wiggins, A. (2012). Free/Libre Open-Source Software Development: What We Know and What We Do Not Know. *ACM Computing Surveys*, *44*(2), 7–7:35. https://doi.org/10.1145/2089125.2089127
Crusoe, M., & Brown, C. (2016). Walking the Talk: Adopting and Adapting Sustainable Scientific Software Development processes in a Small Biology Lab. *Journal of Open Research Software*, *4*(1). https://doi.org/10.5334/jors.35
Dawyndt, P., Dedeurwaerdere, T., & Swings, J. (2006). Contributions of bioinformatics and intellectual property rights in sharing biological information. *International Social Science Journal*, *58*(188), 249–258. https://doi.org/10.1111/j.1468-2451.2006.00617.x
Dedeurwaerdere, T. (2006). The institutional economics of sharing biological information. *International Social Science Journal*, *58*(188), 351–368. https://doi.org/10.1111/j.1468-2451.2006.00623.x
Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, *11*(3), 385–388. https://doi.org/10.1093/biostatistics/kxq028
English, R., & Schweik, C. M. (2007). Identifying Success and Tragedy of FLOSS Commons: A Preliminary Classification of Sourceforge.net Projects. In *First International Workshop on Emerging Trends in FLOSS Research and Development, 2007. FLOSS '07* (pp. 11–11). https://doi.org/10.1109/FLOSS.2007.9
Frischmann, B., Madison, M., & Strandburg, K. (2014). Governing Knowledge Commons – Introduction & Chapter 1. In B. M. Frischman, M. Madison, & K. Strandburg (Eds.), *Governing Knowledge Commons*. Retrieved from http://lsr.nellco.org/nyu_plltwp/477
Gans, J. S., Murray, F. E., & Stern, S. (2017). Contracting over the disclosure of scientific knowledge: Intellectual property and academic publication. *Research Policy*, *46*(4), 820–835. https://doi.org/10.1016/j.respol.2017.02.005
Ghosh, S. (2007). How to Build a Commons: Is Intellectual Property Constrictive, Facilitating, or Irrelevant? In *Understanding Knowledge as a Commons: From Theory to Practice* (pp. 209–245). Cambridge, MA: MIT Press.
Goth, G. (2015). Bringing Big Data to the Big Tent. *Communications of the ACM*, *58*(7), 17–19. https://doi.org/10.1145/2771299
Hall, B. H., & MacGarvie, M. (2010). The private value of software patents. *Research Policy*, *39*(7), 994–1009. https://doi.org/10.1016/j.respol.2010.04.007
Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243–1248. https://doi.org/10.1126/science.162.3859.1243
Harvey, M., & McMeekin, A. (2009). Public or private economies of knowledge: The economics of diffusion and appropriation of bioinformatics tools. *International Journal of the Commons*, *4*(1), 481. https://doi.org/10.18352/ijc.144
Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLoS Comput Biol*, *7*(3), e1002021. https://doi.org/10.1371/journal.pcbi.1002021
Howison, J. (2009). *Alone Together: A Socio-technical Theory of Motivation, Coordination and Collaboration Technologies in Organizing for Free and Open Source Software Development*. Syracuse University, Syracuse, NY, USA.

Howison, J., Deelman, E., McLennan, M. J., da Silva, R. F., & Herbsleb, J. D. (2015). Understanding the scientific software ecosystem and its impact: Current and future measures. *Research Evaluation*, *24*(4), 454–470.
Howison, J., & Herbsleb, J. D. (2011). Scientific Software Production: Incentives and Collaboration. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work* (pp. 513–522). New York, NY, USA: ACM. https://doi.org/10.1145/1958824.1958904
Jefferson, R. (2006). Science as Social Enterprise: The CAMBIA BiOS Initiative. *Innovations: Technology, Governance, Globalization*, *1*(4), 13–44. https://doi.org/10.1162/itgg.2006.1.4.13
Kapczynski, A. (2011). The Cost of Price: Why and How to Get beyond Intellectual Property Internalism. *UCLA Law Review*, *59*, 970–1027.
Karczewski, K. J., Tatonetti, N. P., Manrai, A. K., Patel, C. J., Titus, B. C., & Ioannidis, J. P. (2016). Methods to Ensure the Reproducibility of Biomedical Research. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (Vol. 22, p. 117). Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/27896967
Lemley, M. A., & Sampat, B. (2008). Is the Patent Office a Rubber Stamp. *Emory Law Journal*, *58*, 181–206.
Lewis, J., & Bartlett, A. (2013). Inscribing a discipline: tensions in the field of bioinformatics. *New Genetics and Society*, *32*(3), 243–263. https://doi.org/10.1080/14636778.2013.773172
Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, *16*(6), 321–332. https://doi.org/10.1038/nrg3920
Love, B. J. (2013). Do University Patents Pay Off-Evidence from a Survey of University Inventors in Computer Science and Electrical Engineering. *Yale Journal of Law and Technology*, *16*, 285.
McLeod, A. (2016). *Returns on investment: considerations on publicly funded ICT research and impact assessment*. University of Melbourne. Retrieved from http://minerva-access.unimelb.edu.au/handle/11343/124272
Nielsen, L. B. (2010). The Need for Multi-Method Approaches in Empirical Legal Research. In P. Cane & H. M. Kritzer (Eds.), *The Oxford Handbook of Empirical Legal Research*. Retrieved from http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199542475.001.0001/oxfordhb-9780199542475-e-40
Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press.
Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, Games, and Common-pool Resources*. University of Michigan Press.
Ouellette, L. L. (2017). Who reads patents? *Nature Biotechnology*, *35*(5), 421–424. https://doi.org/10.1038/nbt.3864
Park, H.-S. (2012). Preliminary Study of Bioinformatics Patents and Their Classifications Registered in the KIPRIS Database. *Genomics & Informatics*, *10*(4), 271–274. https://doi.org/10.5808/GI.2012.10.4.271
Penders, B., Horstman, K., & Vos, R. (2008). Walking the Line between Lab and Computation: The "Moist" Zone. *BioScience*, *58*(8), 747–755. https://doi.org/10.1641/B580811
Poteete, A. R., Janssen, M., & Ostrom, E. (2010). *Working Together: Collective Action, the Commons, and Multiple Methods in Practice*. Princeton University Press.
Rai, A. K., Allison, J. R., & Sampat, B. N. (2009). University Software Ownership and Litigation: A First Examination. *North Carolina Law Review*, *87*(5), 1519–1570.
Rasmussen, B. (2005). *Commercialisation processes in bioinformatics: analysis of bioinformatics patents*. Citeseer. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.574.801&rep=rep1&type=pdf
Rasmussen, B. (2010). *Innovation and Commercialisation in the Biopharmaceutical Industry: Creating and Capturing Value*. Edward Elgar Publishing.

Reich, I. R., & Stodden, V. C. (2012). Software Patents as a Barrier to Scientific Transparency: An Unexpected Consequence of Bayh-Dole. Presented at the 7th Annual Conference on Empirical Legal Studies, Stanford Law School: Stanford University Press. Retrieved from http://academiccommons.columbia.edu/catalog/ac:155777
Schweik, C., & English, R. (2013). Preliminary steps toward a general theory of internet-based collective-action in digital information commons: Findings from a study of open source software projects. *International Journal of the Commons*, *7*(2). https://doi.org/10.18352/ijc.397
Schweik, C. M., & English, R. (2007). Tragedy of the FOSS commons? Investigating the institutional designs of free/libre and open source software projects. *First Monday*, *12*(2). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/1619
Schweik, C. M., & English, R. C. (2012). *Internet Success: A Study of Open-Source Software Commons*. MIT Press.
Schweik, C. M., English, R. C., Kitsing, M., & Haire, S. (2008). Brooks' Versus Linus' Law: An Empirical Test of Open Source Projects. In *Proceedings of the 2008 International Conference on Digital Government Research* (pp. 423–424). Montreal, Canada: Digital Government Society of North America. Retrieved from http://dl.acm.org/citation.cfm?id=1367832.1367924
Six, B., Van Zimmeren, E., Popa, F., & Frison, C. (2015). Trust and social capital in the design and evolution of institutions for collective action. *International Journal of the Commons*, *9*(1), 151. https://doi.org/10.18352/ijc.435
Snir, M. (2011). Computer and Information Science and Engineering: One Discipline, Many Specialties. *Commun. ACM*, *54*(3), 38–43. https://doi.org/10.1145/1897852.1897867
Stevens, H. (2015). The Politics of Sequence: Data Sharing and the Open Source Software Movement. *Information & Culture: A Journal of History*, *50*(4), 465–503. https://doi.org/10.1353/lac.2015.0022
Stodden, V. C. (2009). Enabling Reproducible Research: Open Licensing for Scientific Innovation. *International Journal of Communications Law and Policy*, *13*, 25.
Strandburg, K. J. (2008). User Innovator Community Norms: At the Boundary between Academic and Industry Research Symposium: When Worlds Collide: Intellectual Property at the Interface between Systems of Knowledge Creation: Panel II: University Research and Commercial Science. *Fordham Law Review*, *77*, 2237–2274.
Strauss, A., & Corbin, J. (1994). Grounded theory methodology. *Handbook of Qualitative Research*, *17*, 273–85.
van Zimmeren, E., Rutz, B., & Minssen, T. (2016). Intellectual property rights, standards and data exchange in systems biology. *Biotechnology Journal*, *11*(12), 1477–1480. https://doi.org/10.1002/biot.201600109
Vertinsky, L. S. (2013). Making Room for Cooperative Innovation. *Florida State University Law Review*, *41*, 1067–1124.
Vishnubhakat, S., & Rai, A. (2015). When Biopharma Meets Software: Bioinformatics at the Patent Office. *Harvard Journal of Law & Technology*, *29*(1), 206.
von Krogh, G., Spaeth, S., & Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: a case study. *Research Policy*, *32*(7), 1217–1241. https://doi.org/10.1016/S0048-7333(03)00050-7
Wiechers, I. R., Perin, N. C., & Cook-Deegan, R. (2013). The emergence of commercial genomics: analysis of the rise of a biotechnology subsector during the Human Genome Project, 1990 to 2004. *Genome Medicine*, *5*(9), 83. https://doi.org/10.1186/gm487
(9), 83.

Appendix A: Interview Questions

1. Please tell us briefly about your background and your role at the institution

2. What kinds of bioinformatics research/development does your institute perform?

3. What are the main outputs of research from your institute?

4. How frequently is open source licensing used for software released from your institution?

5. Does your organisation promote the use of a specific open source licence?

6. Do you or your organisation prefer permissive (MIT, BSD, Mozilla, Apache) or restrictive (GPL) open source licences?

7. Do you use open source licensing for data as well as source code?

8. What do you see as the purpose for open source licensing for bioinformatics software? Is it the same as for bioinformatics data?

9. What measures would you regard as indicative of the success or failure of open source bioinformatics projects? (Use? User satisfaction? Code quality? Information quality?)

10. What factors do you or your organisation use to measure the success of an open source project? (Number of downloads? Project completion? Number of users? Ongoing project activity?)

11. Do these factors vary depending on the project in question (commercial project versus open source project)?

12. Do you believe the choice of open source licence for your software has an impact on these success factors?

13. Do you or your organisation use dual licensing for software or for data?

14. When a previously open source project is dual licensed, do the success or failure factors change?

15. Does your institution have an intellectual property or licensing policy?

16. Does this policy explicitly or implicitly mention the transfer of source code or data under an open source licence?

17. a) Have you ever shared source code following this policy?

18. b) Have you ever shared source code informally or without following this policy?

19. Does your organisation have a licence for technology transfer from other research institutions?

20. Do you allow contributors from outside your organisation to work on your software?

21. Is there a formal policy on outside contributions to software hosted by your organisation?

22. Have you ever been aware of the presence of intellectual property rights (patents, copyright, trade secrets, plant breeders rights) held by your institute when conducting bioinformatics research?

23. Have you ever been aware of the presence of intellectual property rights (patents, copyright, plant breeders rights) held by other institutions or research entities when conducting bioinformatics research?

24. Do you actively search for patents, copyrights and plant breeders rights to determine whether you might be infringing on these rights when you develop bioinformatics software?

25. Have you or your organisation ever received a notification that your work was infringing on other intellectual property rights?

26. Have you ever not released source code under an open source licence due to the presence (or potential presence) of intellectual property rights held by your institute?

27. Overall, what do you think about the utility of patents, copyright and other intellectual property rights in bioinformatics?

28. What impact do you believe that patents, copyrights and other intellectual property rights have on the success or failure of open source projects?

29. What do you think about patenting bioinformatics algorithms or sequencing machines?

30. What do you think about copyright protection for bioinformatics algorithms?

31. Do you think software more broadly should be patentable?

32. What aspects of the copyright and patent systems would you change to make them more amenable for software?

33. Would you support a formalised strategy in your organisation to encourage source code sharing?

34. Do you support friendly, reasonable and non-discriminatory (FRAND) licensing between different research institutes for standards essential patents for bioinformatics software?

35. Do you support the formation of patent pools within bioinformatics?

36. Do you support transnational instruments for the sharing of source code?

36. a) With respect to collaborative agreements like the G4GH, do you prefer a ‘top down’ or a ‘bottom up’ approach to management?

36. b) As part of the development of this commons, would you support royalty free licensing of patented algorithms and/or proprietary data?

37. Do you support corporate spinoffs as an appropriate way of commercialising bioinformatics software?