The Role of Bioinformatics in Facilitating Translational Science and Medicine

Conor M.W. Douglas
University of British Columbia, Vancouver

Abstract: Significant challenges exist around the translation of the enormous amounts of data generated from large-scale gene and genome sequencing that has been facilitated by the Human Genome Project into tangible medical. Widespread acceptance exists within the biomedical research community about the role that bioinformatics will play in that translational process. While the goal of moving research from "the bench" into socially beneficially applications "at the patient's bedside" has long driven science and technology policy, the picture is now more likely to resemble interactions between very powerful computers, and lab benches. Given the importance of bioinformatics, work presented here reports on a case study of a large Canadian scientific network that has developed a bioinformatics tool designed to facilitate investigations into gene-gene and gene-protein interactions and pathogenomics pathways. By focusing on this kind of bioinformatics system that facilitates a project's own internal biomedical research and simultaneously serves as a free and open resource for a wider group of academic nonpeers, we advocated for a broadening of what translational science and medicine can and should entail. Furthermore, by highlighting the importance of movements between developers and a host of prospective users (and back again) we show how translational bioinformatics systems can be more effectively advanced.

Key words: bioinformatics; translational science; translational medicine; user-configuration; infovis; systems biology.

Corresponding author: Conor M.W. Douglas, 2405 Wesbrook Mall, Faculty of Pharmaceutical Sciences, University of British Columbia, Vancouver, Canada – Email: conor.douglas@ubc.ca

1. Introduction

Biomedical research and development (R&D) is undergoing major transformations as it attempts to achieve translational goals of moving research into the clinic, and deliver on earlier health-related promises issued alongside the Human Genome Project (HGP). One of the components of that transformation has been the development of bioinformatics systems and tools necessary to make sense of enormous amounts of data generated from large-scale gene and genome sequencing that has been facilitated by the HGP and the subsequent (next generation) sequencing activities. There is widespread acceptance within biomedicine that the development of medical interventions derived from that data will only be possible with such bioinformatics systems and tools (Zerhouni 2005; Yang *et al.* 2008; Ostrowski and Wyrwicz 2009; Szalma *et al.* 2010), which has even culminated in an emergent subfield in-and-of itself: translational bioinformatics (Butte 2008; Altman 2012). While biomedical R&D may have once been understood as processes that involve movements between the lab bench and the clinical bed, the picture is now more likely to resemble complex interactions between very powerful computers, lab benches, and maybe some place down the road a clinical bed. That said, bioinformatics systems and tools on their own are not sufficient to facilitate developments in biomedicine. In the interest of understanding the role that bioinformatics systems play in the process of translation, social science research has been conducted on a database and suite of analytical tools called InnateDB (Lynn *et al.* 2008; Breuer *et al.* 2012), which has been developed for the systems-level analysis of the innate immune system as a part of the Pathogenomics of Innate Immunity project (PI2)[1]. This bioinformatics case study was a component of a broader social science endeavor located within the PI2 project that asked which cultural and socio-technical factors constrained and/or enabled the translation of pathogenomics research into medical applications. The argument forwarded here is that bioinformatics systems and tools must be designed with keeping the larger (biological) research community in mind so that biomedical advances can be made more broadly. On top of the need for this particular design mindset, it is argued here that particular design processes and features can also facilitate the development of bioinformatics systems and tools that can be of tangible and far-reaching use in translational science and medicine.

Work in Science and Technology Studies (STS) and beyond has explored the history of bioinformatics (Suárez-Díaz 2010) as well as definitional issues important to understanding these novel systems (Leonelli 2010). Still other work has outlined some of the socio-culture aspects af-

[1] Pathogenomics of Innate Immunity (PI2) project website, *About the project,* http://www.pathogenomics.ca/, accessed 17 December 2009.

fecting the usability of bioinformatics systems (Douglas *et al.* 2011), and the corporeal implications for data that bioinformatics facilitates (Mackenzie 2003). Despite this recent interest, the production processes of bioinformatics systems – such as it is taken up here- has received relatively little attention from a social science perspective.

To position our case study some of the literature and current models of translational science and medicine are first overviewed, along with the acknowledgment of the importance of users in the translation of successful innovation. After detailing the methods through which we have collected and analysed our social science data on this bioinformatics system, we will then outline the functions of InnateDB and the Pathogenomics of Innate Immunity project (PI2) project in more detail. In the body of the text we use the classical sociological concept of *verstehen* to describe the particular mindset that bioinformaticians within the PI2 project adopted when designing a system for translational biomedical work. Further, we will show how specific design processes such as limited release strategy and a particular peer-review system facilitated the development of this translational bioinformatics tool. We will also describe the particular information visualization design features that were integrated into InnateDB so that it would be of use to researchers beyond those with computational backgrounds.

It is our position that resources and systems that are being designed both for internal project-specific use and as platforms that the broader biomedical community of academic non-peers can use for biotechnological development might be conceptualized as form of translational science (TS) that is distinct from other forms of commercial and/or clinical TS. While the iterative movements between bedside and bench (and back again) can be shown in cases of clinical translation, which are mirrored by bench-to-bedside (and back again) movements in the technology transfer and cases of commercial translation, this case of the development of bioinformatics tools suggests that TS needs to be more broadly understood. By including activities that involve movements between developers of research and analysis resources and a host of prospective users (and back again) we not only account for diverse forms of TS, but in doing so we also contribute to the larger goal of translating the masses of genomic data into usable information for health improvements.

2. Translational Science/Medicine and the Role of Users in Innovation

There has long been policy pressure to translate investments in a variety of research into socially beneficial applications (Bush 1945), and more recent demands for medical genetics research activities to deliver health benefits is no exception. The novel journal *Translational Medicine*

– published by American Association for the Advancement of Science (AAAS), who produces *Science* among other journals – outlines the need for a specific sub-field to facilitate this process:

> A profound transition is required for the science of translational medicine. Despite 50 years of advances in our fundamental understanding of human biology and the emergence of powerful new technologies, the rapid transformation of this knowledge into effective health measures continues to elude biomedical scientists. This paradox illustrates the daunting complexity of the challenges faced by translational researchers as they apply the basic discoveries and experimental approaches of modern science to the alleviation of human disease. Studies in humans often highlight deep gaps in our fundamental understanding of biology, but the linkages back to basic research to fill these gaps have not been as effective as they could be. Clearly, creative experimental approaches, novel technologies and new ways of conducting scientific explorations at the interface of established and emerging disciplines are now required to an unprecedented degree if real progress is to be made. Nothing short of a true reinvention of the science of translational medicine is likely to suffice.
>
> (*Science Translational Medicine* Mission Statement)[2]

Alongside academic journals, models of translational medicine have also been developed to try and steer translational work. For instance, common models describe the movement of biomedical research into diagnosis or treatment (i.e. phase 1 translation, or T1), which then moves to subsequent development into evidence-based protocols (T2) (Kerner 2006, 73), and their deployment into clinical practice (T3) (Westfall, Mold and Fagnan 2007), and ultimately the verification and evaluation for 'real world' impacts on health (T4) (Khoury *et al.* 2007). Specific areas of research (e.g. autoimmunity) have adapted their work and concurrent challenges to such models for translational medical research (Blumberg *et al.* 2012).

Work in the area of technology transfer and cooperative research centers (CRCs) suggests that advances in medicine need not be restricted to the kind clinical translation described above. A considerable amount of scholarship exists in the area of management sciences and science policy that have sought to facilitate the flow of knowledge and technology between universities and industry (Bozeman 2000). In this way we can come to think of commercial translation in medicine when health technologies or research on medicinal products are transferred to private companies or spun-off into their own market venture.

While it may be the case that these areas of scholarship have some traction with forms of clinical and commercial translation, they are argu-

[2] http://stm.sciencemag.org/site/about/mission.xhtml, accessed March 20, 2014.

ably less well equipped to handle dynamics related to the production of open access bioinformatics research and analysis infrastructures that are being discussed here. What the case presented here shows is that a particular design mindset and specific design processes and design features stand to play a significant role in the production of bioinformatics systems that are critical for the translation of gene and protein data into actionable medical information. As the body of the text shows, what these design mindset, processes, and features share is their attention to –if not direction integration of- system users in the development and production process. To be sure our work is not the first to acknowledge that a reliance on users is beneficial for the innovation processes with considerable attention being given to "user-driven research" (De More *et al.* 2010), customer-active innovation (von Hippel 1978), or "user-producer interactions" (Laursen 2011). Our case marks a slight departure from this perspective and instead suggests a bi-directional flow of innovation between users and creators of technology. Attention to such dynamics has been made in innovation studies (von Hippel 2005), science and technology studies (Oudshoorn and Pinch 2003), and e-commerce and computer programming (Klein and Totz 2004); however, it has yet to be applied in the area of translational science and medicine as is the case here.

3. Methods

The examination of InnateDB was a part of a broader social science project that sought to understand the social, political, economic, cultural, and technological factors that constrain and enable translational biomedical science. As such our team was an integrated component of the PI2 network from 2006 through 2009, and we conducted three translation cases studies within the PI2 network related to the clinical translation in the university hospital (Lander and Atkinson-Grosjean 2011), commercial translation associated with a pharmaceutical spin-off company, and the bioinformatics case presented here. While these three cases do not form an exhaustive list of translational pathways, we selected them because of their respective success in the translational process, their heterogeneity, and because of their connections with the PI2 network.

Given our integration within the PI2 network we knew that the bioinformatics database (InnateDB) and suite of analytical visualizations tools (Cerebral) would play a central role in the development and success of the PI2 project. Not only was clear that InnateDB and Cerebral were critical to the PI2 project, but as it is described in more detail below, these systems and tools were also being developed as a platform technology for those within and outside of the PI2 project to build knowledge, facilitate future discoveries, and assist in the early development of future medical prophylactics and/or therapeutics. Given our research goal of describing

the constraining and enabling factors in translational science, and in light of our recognition of the role that the bioinformatics system and tools were playing in the translational process of PI2 and beyond, we chose to conduct in-depth social science research on the production, maintenance, and use of InnateDB and its associated tools. As a result, from July 2007 to December 2007 we conducted ethnographic participant research and qualitative semi-structured interviews with one part of the bioinformatics collaboration responsible for the design and construction of InnateDB, and in November and December of 2008 we conducted a series of follow-up interviews across the two institutions involved in InnateDB (total n=25). Our interviews included the heads of the bioinformatics lab, the leaders of the PI2 network, the bioinformaticians designing the front-end and logic of the system, the computer scientist writing the programming code, and the curators who were manually inputting and managing the data submitted to the system. Given the relatively small number of researchers involved with InnateDB we choose to interview practically everybody who was significantly involved in the design, production, and maintenance of the system. Our integration within the PI2 network meant we were able to contact and arrange interviews directly with participants who were ready and willing to contribute to the social science component of the project.

Interviews were audio recorded, and transcribed by members of the research team and private transcriptionists. Interviews were then analyzed using a grounded theory approach to guide our exploration of the material (Charmaz 2006). This approach does not assume a theoretical position *a priori* to analysis, but instead allows a theory to grow out of the data in a developmental movement from code to concept to category to theory. In our case this was accomplished by the team constructing a coding matrix containing terms that highlighted important aspects related to the social, political, economic, cultural, and technological factors that constrained and enabled translational biomedical science. Codes were then attached to segments of interviews using qualitative software ATLAS.ti. Some codes were applied across the three cases (e.g. 'role of teaching and learning', 'impact of disciplinary background', or 'patents and intellectual property'), and some specific to the bioinformatics case (e.g. 'limited release strategy', 'manual database curation', or 'problems with database maintenance'). To improve the reliability in applying the coding matrix between team members, several interviews were coded by multiple members. Variations in coding application were discussed and consistent definitions agreed upon. A lead researcher for each case study then coded all remaining transcripts. The software was then used to produce reports on specific codes, which is similar to the 'concept' development phase within the grounded theory method. These reports were then examined for the most salient factors involved in the diverse forms of translation within the PI2 network, which were worked into concepts. It is here that we identified the importance of users, and consequently developed categories that

described the difference facets through which users were included in the translational process. These categories consisted of the importance of the end-user in the design process, the integration of users in developmental processes, and the creation of a system that includes features to enhance the user experience and enlarge the user-community. These categories have formed the core sections in the body of the text presented here. Within qualitative methodologies interview excerpts have been used to illustrate the above mentioned categories. In doing so the code reports that were used to develop our concepts were re-examined, and the most clear and succinct interview responses have been used as quotations in the body of the text to illustrate the specific category.

The final step in the grounded theory approach is to use identified categories, and the associated quotations, as the basis for a theory of the phenomenon in question. In our case that theoretical supposition is that if bioinformatics systems are going be of use to those beyond the development team for the translation data into useable health information, then they need to be constructed with a particular a mindset (i.e. *verstehen*) that take users into account, and they need to integrate users in the design process (i.e. through the peer review and a limited release strategy), and design the system with tools that facilitate systems-level analysis for those without a computational background.

4. InnateDB Case Study and the PI2 Project

The PI2 project/network was funded largely by Genome Canada to improve the systems-level understanding of the innate immune system. The human immune system has two general components: the adaptive immune system that response defensively against microbial infection and is stimulated by medical interventions like vaccination, and the innate immune system which acts as the first line defense against all foreign pathogens. According to the project's webpage, innate immunity can be understood as a:

> …part of our natural biological makeup – [and because of it] we are able to withstand a daily onslaught of tens of thousands of potentially pathogenic microbes in air, food and water, and in our interactions with other people and animals. But our innate immunity can sometimes get over-stimulated, leading to inflammation of tissue and even sepsis – a deadly infection of blood or tissue. Understanding the balance between infection resolution and inflammation is the goal of the new Pathogenomics of Innate Immunity Genome Canada Competition III project.
>
> (PI2 2006) [3]

[3] Pathogenomics of Innate Immunity (PI2) project website (2006) *About the project,* http://www.pathogenomics.ca/, accessed 17 December 2009.

If the goal of the PI2 project is to understand "the balance between infection resolution and inflammation", then InnateDB's role in that project was to create a roadmap of the immune system. The metaphor of the roadmap is apt for non-scientific writers and audiences to deploy when trying to make sense of InnateDB, and was also a guiding metaphor for members of the InnateDB development team. It is worthwhile for one of bioinformaticians to explain themselves how this metaphor of the roadmap can facilitate an understanding of InnateDB:

> We just kind of want to make a roadmap to the immune system that, you know, when people… If you just look at a list - say you go to an atlas and you look at the index. Oh, very exciting; it's just a list of places. You can't really picture that. But then when you open things, when you open your atlas to a map page, you say, "Oh, this city is connected to this city by this road. Oh, these cities are in the same country. Oh, these cities are in a different country". And it's just like that. In the past, people have been analyzing their array data by just looking at a list. And they've never really put that list into biological context. So we are giving them a map. And we are giving them a map that's laid out well... But once you lay things out in their proper context - this goes here, this goes there, this goes there, this is in this part of the cell, this is in this part cell- then it makes it so much clearer, and people can start to follow relationships and trace pathways [and think]: "Oh, this receptor up here is being activated. And all these genes down here are getting turned on. Maybe that receptor is linked to this set of genes somehow".

While the broader PI2 project had numerous goals, one of the distinct objectives was to identify the key molecules involved in infectious disease response, which might ultimately give rise to new prophylactics or treatments. According to one of the InnateDB Project Leaders, "if you can target those key central molecules, perhaps you can predict therapeutic effects on the outcome of disease or the outcome of information". As a result when genes are identified to have an association to disease response InnateDB can be used to model the pathways and networks of those genes and proteins across different datasets. If the concurrent systems-level analysis does identify mechanisms within the pathway, then lab biologists will conduct wet experiments for the confirmation or dismissal of the mechanisms within the identified pathway.

Importantly, InnateDB also boasts supplementary interactions that innate immunity genes participate in, and because it has been created as resource to include all human and mouse pathways of interactions at systems biology level its relevance is not limited to innate immunity. Further, unlike bioinformatics resources that contain large amounts of annotated data, InnateDB comes equipped with a suite of tools through which re-

searchers can conduct analysis directly in the InnateDB website. What is more InnateDB is an open source and open access database and analysis environment. While there is a tradition (or even convention) in computer science (and perhaps even in bioinformatics) to design databases to be open access and open sources, the bioinformatics Primary Investigator describes the importance of open source and open access characteristics in some length:

> Yeah, so for open source it's important that you realize that open source doesn't mean "free", you know, so it just means that when you make software you can actually… the way you make software you write a program, and you can release that program to somebody and then they can run that program. Or you can package it up into an executable - a wxe file - and so that it's actually just in this binary code that you can actually see what the original program was, and you can release that…The open source model is where you just keep that package open so…you still have the ability to see that code, see that program, see how it works, know exactly how it works so you can either modify it for your own uses, or…redistribute it as some other version, or you just might want to see how it works to understand why it's doing… and so there's definitely been a growing movement of people that really want that, because they're frustrated with the sort of closed black box kind of software and for example in our pathogenomics project we found it very useful because there was definitely some microarrays software that was black box like that…

While the open source characteristic of InnateDB refers to the process of keeping lids lifted on black boxes so that users can see the computational processes that have gone into making the database function the way that it does, the open access refers to a similar characteristic of transparency. Within the open access model all users are provided free right of entry into the database, and the data contained within the database is free to access, download, and use for one's own research purposes.

Part of the data contained within InnateDB is itself an amalgamation of three or four different types of data that come from four or five different categories of open access data. This data that is present within the database is mostly "gene, proteins, and interactions and signalling responses involved in the mammalian innate immune response"[4]. These different kinds of data are collected from gene lists, external interaction databases, and external pathway databases, which are all integrated within InnateDB, and all open access. There are also links to external databases containing immunology-relevant data, but it is not clear if this information is integrated within InnateDB. As new data is compiled in these external

[4] InnateDB website *Home page*, www.innatedb.ca/index.jsp, accessed 1 March 2014.

databases it is regularly uploaded into InnateDB by website administrators. Keeping abreast of novel pathway and interaction data can be achieved in part by those with computer science backgrounds as they amalgamate existing databases into InnateDB; however, one of the distinct characteristics of this tool for translational biomedicine is that it is also manually curated - a point which we will return to as a key design feature of the system.

Alongside the massive amounts of curated and standardized interaction data that is accessible through InnateDB there is also a multitude of search mechanism available to mine the data. A researcher can make use of the search functions included in the suite of tools provided by InnateDB to investigate genes and proteins of interest, or view statistics for manually-curated molecular interactions that are relevant to innate immunity and submitted weekly by curators. Further searches can be conducted for "experimentally-verified molecular interactions by gene/protein name, interaction type, cell type, etc." as well as searches for 147,240+ interactions & 4,400+ pathways[5].

Not only can a researcher mine the gene, protein, and interaction data that is provided through InnateDB, but because it has been concurrently constructed as a suite of tools researchers can also upload their own data and conduct particular kinds of analysis immediately on the InnateDB website. Gene expression data can be uploaded by anyone, and then through the use of a piece of software called Cerebral researchers are "able to interactively visualize interaction networks with expression data overlaid; carry out Pathway, Gene Ontology and Transcription Factor Binding Site over-representation analysis, construct orthologous interaction networks in other species and much more"[6]. Not only are these tools provided as an integral part of InnateDB, but video tutorials also exist on the website so to help users familiarize themselves with how these tools can be most effectively used. In light of the central role that is played by Cerebral in the InnateDB analysis environment, it will receive more attention later when we more directly describe the particular design features of the system that can facilitate the translation of data into valuable biomedical information.

5. Designing Bioinformatics Systems for Translational Science and Medicine

[5] InnateDB website *Home page,* www.innatedb.ca/index.jsp, accessed 1 March 2014.

[6] InnateDB website *Home page*, www.innatedb.ca/index.jsp, accessed 1 March 2014.

5.1. Particular Design Mindset

In order for InnateDB to be usable tool for translational science to both those inside the PI2 network and to those outside of it who may or may not have a computational background, the researchers building the system had to adopt a particular design mindset. Classical German sociologist Max Weber used the word "*verstehen*" to describe the process through which the social researcher would develop an interpretive understanding of meaning and human activity (Ritzer 2007). By approaching a human's actions from their point of view Weber hoped to gain an appreciation of the way in which they constructed and gave meaning to their own world. In doing so the social actor is not seen as the mere object of investigation, but rather as a subject. Here we can adapt this concept to the social study of science and technology to explore the case of InnateDB and the particular mindset the bioinformatics system designers deployed in making a tool that would be broadly usable for translational activities. The system architects -whom were largely computer programmers coming from a computer science background- needed to develop a level of interpretive understanding (or *verstehen*) of diverse prospective users of InnateDB, so that it could be appropriately configured to their needs. While the system architects and designers had a general understanding of what members of the PI2 team would be using the system for, they had to develop a more nuanced understanding of systems biology so that InnateDB would be equipped for the kinds of translational work that such researchers would be undertaking. One of the developers described this learning process in these words:

> I didn't anticipate that people would be uploading entire GeneChips of data. I thought it would probably be 100 or 200 queries at a time. And so I sort of…the way it was designed, it was sort of a design in the manner to handle these many pieces of data. But when you get into, sometimes…like 25,000 or 30,000 genes being uploaded is quite a load on the server. And it sort of brought it to a crawl at first until we said: "Okay. I've got to step back and rewrite this". So, it was a few extra months, but it definitely paid off.

InnateDB also boasts a team of curators that manually keep gene, pathway, and interaction data current. Their training in biologically-relevant disciplines means that they can sift through individual pieces of published data – as opposed to already curated data that is found in the other interaction and pathway databases. They are then able to make decisions with regards to the accuracy and relevance of that data to InnateDB, and submit it to the system. Without this curation the database becomes a rather static entity, and its practical value concurrently decreases to the PI2 project team as well as those interested in innate immunity and

systems biology more broadly. Curating originally began as an examination of data concerning single genes, and if the quality of that data could be confirmed then it would be uploaded into InnateDB for subsequent use. However as InnateDB grew, curating extended towards the examination of specific pathways with curators themselves playing an increasing role in deciphering the balance between infection resolution and inflammation. While these curators are not the analytical bioinformaticians who conduct the systems-level analyses that identify mechanisms within the pathway, nor are they the lab biologist that produce experimental confirmation or refutation of the mechanisms within the pathway, they are key players in the pathway identification process. Having a level of interpretive understanding of how systems-level bioinformaticians go about assembling these pathways greatly facilitates the work of the curators by sensitizing them to the kind of data that they should be on the look-out for. In turn, systems-level bioinformaticians increasingly grow to trust the data within InnateDB when they know its character, quality, and standards, which then facilitates their analytical work. One of the curators explains this dynamic when asked about the potential for training to increase her analytical role in the project:

> But from a bioinformatics point of view, to understand how it kind of is related to this database, like that's the whole point right, is to analyze data basically. So from me, I think it would be more interesting to kind of learn the aspects of that [analysis], but our job description is to look for particular protein-protein, or protein-gene interactions. So you don't necessarily need [added understanding of the analytical processes], it's just kind of an added thing that might actually increase the analysis, or maybe things that you kind of pick-up on that other people may need later on. Because I think [the project leader] also kind of looks at it with the perspective of: "How he would analyze his data", but when it comes to curating, I ask for certain things that maybe weren't on the website, but might help us later on to do the pathway curation. But over all, it's supposed to help out data analysis.

By demonstrating a level of end-user *verstehen*, manual curators also affect the development of the data that goes into InnateDB, which influences the analytical applications that the data is used for. Seeing the relationship between curators and system-level analysts in this way conforms to the colloquialism of 'garbage-in-garbage-out' which is well worn within database and bioinformatics cultures. The PI2 analysis emergent from InnateDB will only be as good as the interaction and pathway data that is boxed-up inside of it, so it is clear why high quality curated data is central to the project. Part of the process of obtaining high quality curated data is to equip curators with a bigger picture of what the data would be used for in bioinformatics terms. Therefore it seems useful -if not necessary- for each member of the project team to have an appreciation of what oth-

er team members are doing and what their job entails so that they can do their own job better. For analysts to do their job well they need to know that they have good interaction or pathway data to conduct their analysis with, and for curators to do their job well it is good to know the larger analytical picture (i.e. the purpose of the databases is about, what it is meant for and what it is meant to accomplish) so that they can input the right kind of data with the right kind of annotations. This suggests that understanding the roles and goals of other project team members is highly relevant to the success of multidisciplinary research, and ultimately to the achievement of translational goals. While we have seen here how a particular design mindset that takes into account the prospective user of the technology is critical in the construction of a useful translational tool, the following section explores how particular design processes are similarly important in achieving this goal.

5.2. Particular Design Processes

InnateDB was first released for public use in May of 2008. However, before it could effectively "go live" a number of design processes were undertaken that included a limited release strategy and a rigorous peer review, which helped it to become a useful tool to the PI2 network and beyond. In his work *Democratizing Innovation*, Eric von Hippel shows that "much of the information needed by product and service designers is 'sticky'" (von Hippel 2005, 67). Different users have diverse needs and capabilities that require inscription into a system so that its' utility can be maximized. As a result, unsticking those needs and capabilities and getting them to the designers is of paramount importance in the development of useful technologies such as bioinformatics systems. One of the ways through which this was accomplished with InnateDB was by releasing drafts of the system prior to its public release to select colleagues in the innate immunity community and to the PI2 project team. The role of the limited release strategy should not be underestimated, as prospective users of a technology are proving to play an increasingly central role in up-stream innovation processes. With a working version of the database in place, and with some data now loaded in, PI2 team members from other components of the project were invited to access the system and experiment with its uses while the database was still in its developmental stages. Incorporating project team members outside of the database development team at this stage was important for a number of reasons. First, the development of any system is bound to have bugs, and identifying problems with the operations of InnateDB would be crucial before it was to be released to the public. More importantly PI2 project team members were brought into the development process so that their needs could be readily identified and configured into the design of the technology. While the developers of InnateDB would certainly consider themselves bioin-

formaticians their familiarity with biological sciences varied. As a result experimental biologists were consulted to provide feedback on the system. One of the InnateDB's designers explained the content and function of that feedback:

> [Biologists provided feedback on] all sorts of levels to, you know, to broadly… kinds of things you want to do, you know, feedback to the extent of what are the biological questions that they want to be able to use the system to use, down to pretty nitty-gritty questions of, you know, in our visualization system, you know: "Do you want to see broad spectrums of colours or do you just want to keep it pretty simple? Yes/No, kind of colours? [up-down] kind of thing?" Yeah, so from quite a broad spectrum of very nitty-gritty stuff to big picture types of big questions they want answered.

Two bioinformaticians who conducted system-level research and who were familiar with the challenges of databases and tools were also a part of the process through which InnateDB would be improved upon. Both of these bioinformaticians were members of the PI2 project, and involved with the construction of InnateDB, but were not the developers responsible for the schema, submission system, or search mechanisms. As one of these key figures point out:

> The other thing that I think was kind of critical is that, although I've worked in bioinformatics for about 10 years now, my background and interest is on the biological sciences side of things. I think having someone with that background making the key decisions on the direction of the thing was very beneficial to ensuring that it was relevant to a biologist, and a lot of these things are developed by people with computer science backgrounds who, you know, can come up with great algorithms or whatever but don't have the same insight into how a biologist wants to see things.

As the above excerpt demonstrates, these two bioinformaticians were of a particular ilk which made them crucial to the development of InnateDB. Not only did these two actors have more familiarity with the computer science end of databases, which allowed them to engage with the developers on a deep level that the average experimental biologist was unable to do, but more importantly they were prime examples of the kinds systems-level end-users of InnateDB. Problem areas of InnateDB were identified through the early deployment of the suite of tools for high-level analysis and future improvements were also prospected.

Another important aspect of the design process that has facilitated the construction of a translational bioinformatics system was the peer-review process, and subsequent publication of both the article that describes InnateDB (i.e. Lynn et al. 2008) and the actual database system itself. While

it may seem obvious that a journal would access the functionality of a system like InnateDB in the peer-review process, this strategy seems to represent a departure from traditional articles in bioinformatics. One of the members of the PI2 network described it this way:

> I mean bioinformatics when it started, and bioinformatics really only started to become big ten to fifteen years ago. And for the longest time, it was such a specialized field that people were doing it for sort of discovery sake, and not really making tools that were ideally suited to an end-user. Even when I was doing my PhD [2001-2005], I'd say half of the papers I read that reported sort of a relevant method to what I was doing were just algorithm papers. There was no software; there was no website to go along with it. It was just telling you the method, "We did this, and he's our paper with some math showing how we did it, but you can't actually do this unless you create this entire system and do this entire training dataset". So that was the prevailing mindset in bioinformatics, and I think that was probably, I don't know maybe it was sort of a cultural thing. The scientists that first got there, they were these specialized scientists. They didn't really care; they were just doing this for discovery sake. But then the people that have gotten into bioinformatics more recently, people of sort of my generation, or a couple of years older and sometimes younger realized the importance of the user community. Because we had to do our Master's and our PhD's seeing these methods that looked really interesting, and not being able to use them anywhere. So I think to our generation of bioinformaticians the notion of open source is a big thing. Making your work available to people. And people realize that your tool can be open source, and available to the world, but if it's not designed well people aren't going to use it.

In light of transformations within bioinformatics to publish functional tools rather than a theoretical algorithms and methods, InnateDB had to be up and running before the review process could get under way. In this respect the publication process that the InnateDB paper had to navigate – before it could be deployed to grow its user community – had to pass a kind of usability test in the form of peer review. While peer reviewers may not have embodied the traditional notion of 'user' that is conjured up in one's mind when technologies are discussed, and nor does the publication process meet conventional understandings of technological 'use', both would prove to be an essential hurdle that the team had to overcome in their attempt to manifest InnateDB's translational potential:

> Writing any paper takes a while to complete, a big paper like this. The reviewer comments were probably the most positive comments I've got on anything I've ever been involved in before. They were hugely positive comments and the suggestions that they wanted to do were very relevant and things that we would have wanted to do and that we just, we did it for them. For example, we

> used to allow users to upload just four gene expression datasets at one time and they wanted to increase the capacity to be able to do more datasets in one go, so we increased it so that you can now do up to 10 different conditions at any one time. We had some other limits to do with computational power in terms of the number of interactions you could return in any one search, and they felt that if we could find a way around that it would be better not to have any limits. Our original thinking was that, you know, we had pretty generous limits - like we're talking you could return up to 10,000 data points kind of - and if you really wanted to do any more than that you were probably at a fairly advanced bioinformatics level and you can just download the entire database including all the data and then analyse it, but they would have preferred that the limits be removed. So we came up with a computational approach to mean that we could do that. And so now we don't have limits in our searches now, you can return all the data in any search.

What this section shows is that there is a clear link between the publication process, the role of users in the design and development stage of bioinformatics tools, and potentialities of translation.

5.3. Particular Design Features

For InnateDB to be a useful tool for making sense of vast amounts of sequence data it also had to include a suite of tools to aid researchers conducting analyses into problems of systems biology. One of the analytical tools that can be found within InnateDB is called Cerebral. This tool was created as a Postdoc project by researcher within the PI2 network to facilitate the research into innate immunity by the team, but also to act as a tool for the wider biological community in general. As she explains herself:

> Sure, well I was always sort of peripheral to the InnateDB project, I was brought on to work on Cerebral, which is a spin-off, you know it's a component of InnateDB, but in and of itself it's, its own project. And so I really, you know when I was doing the Cerebral work; I tried to develop it for the larger community.
>
> [Researcher]: Which larger community, sorry?
>
> Biology in general, anybody interested in visualizing networks in a pathway like fashion. So you know, it's all basically, its creation was inspired by InnateDB, and sort of went along with InnateDB, but I always kept my eye towards a larger audience when developing it. So, I was always sort of on the periphery. I'd be included in some of the InnateDB meetings, and things, just to provide guidance as one of the ultimate users of the database.

Cerebral – or CEll REgion-Based Rendering And Layout – is a tool that allows analysts to visualize biological information in traditional sig-

naling pathway/system diagrams. It is not a stand-alone tool, but rather a plug-in for one of the most widely used bioinformatics tools called Cytoscape, which: "is an open source bioinformatics software platform for visualizing molecular interaction networks and biological pathways and integrating these networks with annotations, gene expression profiles and other state data" (Cytoscape 2012). As a plug-in, Cerebral brings many features useful for pathway and interaction analysis that Cytoscape lacks, and seeks to supplement -rather than supplant- the existing visualization tool: "Cerebral is a plug-in that enhances Cytoscape's functionality by using extra annotation provided by the user to both automatically generate a more pathway-like representation of a network and to provide an environment for the visualization, comparison, and clustering of expression data from multiple conditions" (Barsky *et al.* 2007; Cerebral 2012). While Cerebral could have been developed as a standalone tool, Cytoscape has created a certain degree of technological lock-in within the bioinformatics user community that has been facilitated by its open source and open access character. Releasing a tool outside of Cytoscape software platform would undoubtedly reduce the numbers of users accessing Cerebral thereby diminishing its capacity as a piece of translational science. One of the developers of Cerebral explained it this way:

> "By piggybacking on a big endeavour like that, there's two main advantages to the plug-in developer both of which are entirely selfish. 1) Is its way less work…if you're looking at it [Cytoscape] from an infovis perspective, you're kind of like: "Oh, why did they decide to do this?". And the rendering engine is goofy, and all that stuff. So initially [our collaborators] looked at Cytoscape, and they're like, "Oh gees this is a piece of crap, can we please just build our own version". And I was pretty adamant that, "No we gotta do it in Cytoscape", I mean there's so many functions beyond the visualization that we would have to code into one of these bits of software that would take years, and years, and years to do something that even did a tenth of what Cytoscape does. So it's saving you a pile of work by piggybacking on something, and 2) it's also giving a huge user community too.

One of the ways that Cerebral enhances Cytoscape's functionality is by integrating ideas and lessons from the emergent interdisciplinary fields of information visualization and visual analytics. Information visualization – or infovis – is "the use of computer supported, interactive, visual representations of abstract data to amplify cognition" (Card *et al.* 1999). Data can take both numerical and non-numerical form, such as genes and proteins. Visualization can aide users from various disciplines to address a variety systems-level problems in biology because "visual representations and interaction techniques take advantage of the human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once. Information visualization

focused on the creation of approaches for conveying abstract information in intuitive ways" (Thomas and Cook 2005). Visual analytics (VA) on the other hand is an outgrowth of infovis, which "combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets" (Keim *et al.* 2008). Whereas infovis is concerned with principles, ideas, and assumptions concerning how users see and use information, VA is more about the development of tools resultant from these visualizations to facilitate analytical reasoning.

In the case of InnateDB infovis principles were used in the development stages of Cerebral and the tool can boast of both infovis and VA characteristics in its most recent incarnation. When asked what role infovis would play in developing bioinformatics systems that are useful tools in the translation process and resolving biological problems, one of the developers of Cerebral responded this way:

> Infovis is going to be huge, huge, huge. And Cerebral and a few other sort of similar type tools are really the first ones to bring visualization to bioinformatics. I think Cerebral was probably the first one to bring principles from information visualization to bioinformatics. You know tools like Cytoscape were obviously around for a while that would create a visual representation of data so you could interact with it easily. But they didn't really do any research into infovis principles and ideas when they built Cytoscape. But when we built Cerebral, we had our two infovis collaborators so they brought in all these things that we sort of never heard of before and never considered in biology that just made Cerebral that much better. Because all this research into a how a user looks at screen or where do they look, what colours do they respond to, what shapes do they respond to all of that went into Cerebral, and it really was the first instance of that happening. But, I think visual analytics are going to be huge…So if you can make things as simple and as universal as possible, then you're well on your way forward to satisfying as many people and getting a huge user community as you possibly can. So I think as bioinformatics professionals recognize this, they're going to be making their tools more usable by adopting visual methods.

Through their integration of Cerebral into the construction of InnateDB the project team was not simply coupling a suite of analytical tools with a database; rather, they were creating a research resource that would be as widely usable as possible. Further, it is important to note that this was not required by their funders, and they were under no obligation to create their own project tool this way; rather, it was an initiative they took of their own volition. Not only would their choices of particular design features allow them to tap into a larger research community associated with Cytoscape, but by designing a visual analytics tool like Cerebral with infovis principles the PI2 team were purposefully creating a research re-

source that would extend beyond their own project and into the wider biological research community without a hardcore computational background.

6. Conclusion

We have argued here that a number of features of InnateDB have functioned to make it a research and development resources both for internal use of the PI2 team and as platforms for the broader biomedical community to engage in translational work. Specifically we have shown how the development team took on a particular design mindset throughout the construction process in which they constantly envisaged who their diverse users might be, and how they might use the system. By deploying a level of interpretive understanding – or *verstehen* – of their users the InnateDB team was able to construct a tool more suitable to diverse user needs. Furthermore, through an appreciation of how systems biologists would use InnateDB the architects of database were able to make important alterations to the amount of gene data that could be uploaded by users, and the curators were able to improve the data that they were inputting into the database so as to minimize the 'garbage-in-garbage-out' phenomenon. Both of these changes stand to have an impact on the ability of systems biologists to move their work along in the translational process. We have also shown how particular design processes related to the limited release strategy and peer review worked to not only debug the system, but to construct a tool that was more useful for the kinds of system-level analyses needed to advance translation in innate immunity and beyond. Finally, work here has made clear how design features related to information visualization and visual analytics make InnateDB a resource and tool increasingly usable to those who may not have a computational background. By designing a system that is more usable, the potential users of the system expand, and then so too does the potential to make sense of data contained within the database.

By creating resources and tools for the broader biological community to use, activities like the construction of InnateDB could be considered a particular form of translational science, or what we have referred to elsewhere as 'civic translational science' (CTS) (Atkinson-Grosjean and Douglas 2010; Lander and Atkinson-Grosjean 2011). The motivation behind the labelling of CTS practices is not to construct a hard and fast definition that will be true for a specific set of activities, but rather to call attention to a broader set translational dynamics that exist beyond the clinic or the market. Iterative movements between bedside and bench (and back again) can characterize clinical TS (see Lander and Atkinson-Grosjean 2011 for e.g. within PI2 network), which are mirrored by bench to bedside (and back again) movements in the commercial TS and tech-

nology transfer. However, the development of an open source and open access resource like InnateDB that facilitates the translation of massive amounts of gene and protein data into usable health information for clinical and commercial developments is not well suited to such clinical or commercial representations of TS. What InnateDB shows is the importance of movements between developers and a host of prospective users (and back again) in the production of research and analysis tools. In the case presented here those users were the wider scientific polis or academic non-peers who would use InnateDB, but the concept needn't be applied strictly to such users. For instance, in other cases users might also include factions of the public, as "civic science" has elsewhere been "used interchangeably with participatory, citizen, stakeholder and democratic science, which are all catch words that signify various attempts to increase public participation in the production and use of scientific knowledge" (Bäckstrand 2003). Rather than exacerbating ambiguities that already exists around the notion of "civic science" or "civic scientist" (Clark and Illman 2001), our intention here has been to broaden what counts as translation science and medicine to include the construction of bioinformatics systems, and to show how such systems can be more beneficially constructed to fulfil translational tasks.

Acknowledgments

Funding for this research was provided by Genome Canada through the Pathogenomics of Innate Immunity (PI2) project. This work was made possible by the interview participants within the PI2 project, for which much thanks is owed. At the time the research was conducted, the author was affiliated with the Centre for Applied Ethics at the University of British Columbia, Canada and would like to thanks his project team partners there for their support and contributions to this work: Janet Atkinson-Grosjean, Bryn Lander, Cory Fairley, and Lilly Farris. Thanks to the two anonymous reviewers who provided helpful feedback and to the Journal's Special Issue Guest Editors Federico Neresini and Assunta Viteritti for facilitating this process. The author has no conflicts of interests relating to this work.

References

Altman, R.B. (2012)*Translational Bioinformatics: Linking the Molecular World to the Clinical World*, in "Clinical Pharmacology & Therapeutics", 91 (6), pp. 994-1000.

Atkinson-Grosjean, J., Douglas, C. (2010) *The Third Mission and the Laboratory: How Translational Science Engages and Serves the Community*, in H. Schuetze, H. and P. Inman (eds.), *Community engagement and service mission of universities*, Leicester, NIACE, pp. 309-23.

Bäckstrand, K. (2003) *Civic Science for Sustainability: Reframing the Role of Experts, Policy-Makers and Citizens in Environmental Governance*, in "Global Environmental Politics", 3 (4), pp. 24-41.

Barsky, A., Gardy J.L., Hancock, R.E.W., Munzner, T. (2007) *Cerebral: A Cytoscape Plugin for Layout of and Interaction with Biological Networks Using Subcellular Localization Annotation*, in "Bioinformatics", 23 (8), pp. 1040-1042.

Blumberg, R.S., Dittel, B., Hafler D., von Herrath, M., Nestle, F.O. (2012) *Unraveling the Autoimmune Translational Research Process Layer by Layer*, in "Nature Medicine", 18 (1), pp. 35-41.

Bozeman, B. (2000) *Technology Transfer and Public Policy: A Review Of Research and Theory*, in "Research Policy", 29 (4/5), pp. 627–655.

Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C.A., Sribnaia, R.L., Winsor, G.L. , Hancock, R.E. W., Brinkman, F.S.L., Lynn, D.J. (2012) *InnateDB: Systems Biology of Innate Immunity and Beyond - Recent Updates and Continuing Curation*, in "Nucleic Acids Research", 41 (D1), D1228-D1233.

Bush, V. (1945) *The Endless Frontier*, in "Transactions of the Kansas Academy of Science", 48 (3), pp 231-264.

Butte, A.J. (2008) *Translational Bioinformatics: Coming of Age*, in "Journal of the American Medical Informatics Association", 15 (6), 709 -714.

Card, S.K., Mackinlay J.D., Shneiderman B. (1999) *Readings in Information Visualization: Using Vision to Think,* Morgan Kaufmann.

Cerebral (2012) *About Cerebral. Cerebral-Cell Region-Based Rendering And Layout, Multiple Experiment Comparison*. Available from http://www.pathogenomics.ca/cerebral, accessed 2 November 2013.

Charmaz, K. (2006) *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*, Pine Forge Press.

Clark, F., Illman D.L. (2001) *Dimensions of Civic Science Introductory Essay*, in "Science Communication", 23 (1), pp. 5–27.

Cytoscape (2001-2013) *What is Cytoscape? Cytoscape*. Available from http://www.cytoscape.org/what_is_cytoscape.html, accessed 2 November 2013.

De Moor, K., Berte K., De Marez L., Wout J., Deryckere T., Martens L. (2010) *User-driven Innovation? Challenges of User Involvement in Future Technology Analysis*, in "Science and Public Policy", 37 (1), pp. 51-61.

Douglas, C., Goulding, R., Farris, L., Atkinson-Grosjean, J. (2011) *Socio-Cultural Characteristics of Usability of Bioinformatics Databases and Tools* in "Interdisciplinary Science Reviews", 36 (1), pp. 55-71.

Kerner, J.F. (2006) *Knowledge Translation Versus Knowledge Integration: A "Funder's" Perspective*, in "Journal of Continuing Education in the Health Professions", 26 (1), pp. 72-80.

Keim, D., Andrienko G., Fekete J. D., Carsten, G., Kohlhammer J., Melançon G.

(2008) *Visual Analytics: Definition, Process, and Challenges* in *Information Visualization*, in A. Kerren, J.T. Stasko, J.D. Fekete, C. North (eds.), Berlin, Springer, pp. 154-175.

Khoury, M.J., Gwinn M., Yoon P.W., Dowling N., Moore C.A., Bradley L. (2007) *The Continuum of Translation Research in Genomic Medicine: How Can We Accelerate the Appropriate Integration of Human Genome Discoveries into Health Care and Disease Prevention?*, in "Genetics in Medicine", 9 (10), pp. 665-674.

Klein, S., Totz, C. (2004) *Prosumers as Service Configurators-Vision, Status and Future Requirements*, in B. Preissl, H. Bouwman and C. Steinfield (eds.), *E-life after the Dot Com Bust*, Heidelberg, Physica-Verlag, pp. 119-34..

Lander, B., Atkinson-Grosjean J. (2011) *Translational Science and the Hidden Research System in Universities and Academic Hospitals: A Case Study*, in "Social Science & Medicine", 72 (4), pp. 537-544.

Laursen, K. (2011) *User–producer Interaction as a Driver of Innovation: Costs and Advantages in an Open Innovation Model*, in "Science and Public Policy", 38 (9), pp. 713-723.

Leonelli, S. (2010) *Documenting the Emergence of Bio-Ontologies: Or, Why Researching Bioinformatics Requires HPSSB*, in "History and Philosophy of the Life Sciences", 32 (1), pp. 105-125.

Lynn, D.J., Winsor, G.L. Chan, C., Richard, N., Laird, N., Barsky, A., Gardy, J.L., F.M. Roche, T.H.W. Chan, N. Shah, R. Lo, M. Naseer, J. Que, M. Yau, M. Acab, D. Tulpan, M.D. Whiteside, A. Chikatamarla, B. Mah, T. Munzner, K. Hokamp, R.E.W. Hancock, F.S.L. Brinkman (2008) *InnateDB: Facilitating Systems-Level Analyses of the Mammalian Innate Immune Response*, in "Molecular Systems Biology", http://www.nature.com/msb/journal/v4/n1/synopsis/msb200855.html.

Mackenzie, A. (2003) *Bringing Sequences to Life: How Bioinformatics Corporealizes Sequence Data*, in "New Genetics & Society", 22 (3), pp. 315-332.

Ostrowski, J. Wyrwicz, L.S. (2009) *Integrating Genomics, Proteomics and Bioinformatics in Translational Studies of Molecular Medicine*, in "Expert Review of Molecular Diagnostics", 9 (6), pp. 623-630.

Oudshoorn, N., Pinch, T. (2003) *How Users Matter: The Co-Construction of Users and Technologies*, Cambridge, MA, MIT Press.

Ritzer, G. (2007) *Sociological Theory*, Boston, McGraw-Hill.

Suárez-Díaz, E. (2010) *Making Room for New Faces: Evolution, Genomics and the Growth of Bioinformatics*, in "History and Philosophy of the Life Sciences", 32 (1), pp. 65-89.

Szalma, S., Koka V., Khasanova T., Perakslis, E.D. (2010) *Effective Knowledge Management in Translational Medicine*, in "Journal of Translational Medicine", 8 (1) pp. 8-68.

Thomas, J.J., Cook K.A. (2005) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE Computer Society Press.

von Hippel, E. (1978) *A Customer-Active Paradigm for Industrial Product Idea Generation* in "Research Policy", 7 (3), pp 240–266.

von Hippel, E. (2005) *Democratizing Innovation*, MIT Press.

Westfall, J.M., Mold J., Fagnan, L. (2007) *Practice-based Research "Blue Highways" on the NIH Roadmap*, in "JAMA", 297 (4) pp. 403-406.

Yang, J.Y., Yang, M.Q. Arabnia, H.R., Deng Y. (2008) *Genomics, Molecular Imaging, Bioinformatics, and Bio-Nano-Info Integration Are Synergistic Components of Translational Medicine and Personalized Healthcare Research*, in "BMC Genomics", 9 (Suppl 2) I1.

Zerhouni, E.A. (2005) *Translational and Clinical Science — Time for a New Vision*, in "New England Journal of Medicine", 353 (15), pp. 1621-1623.