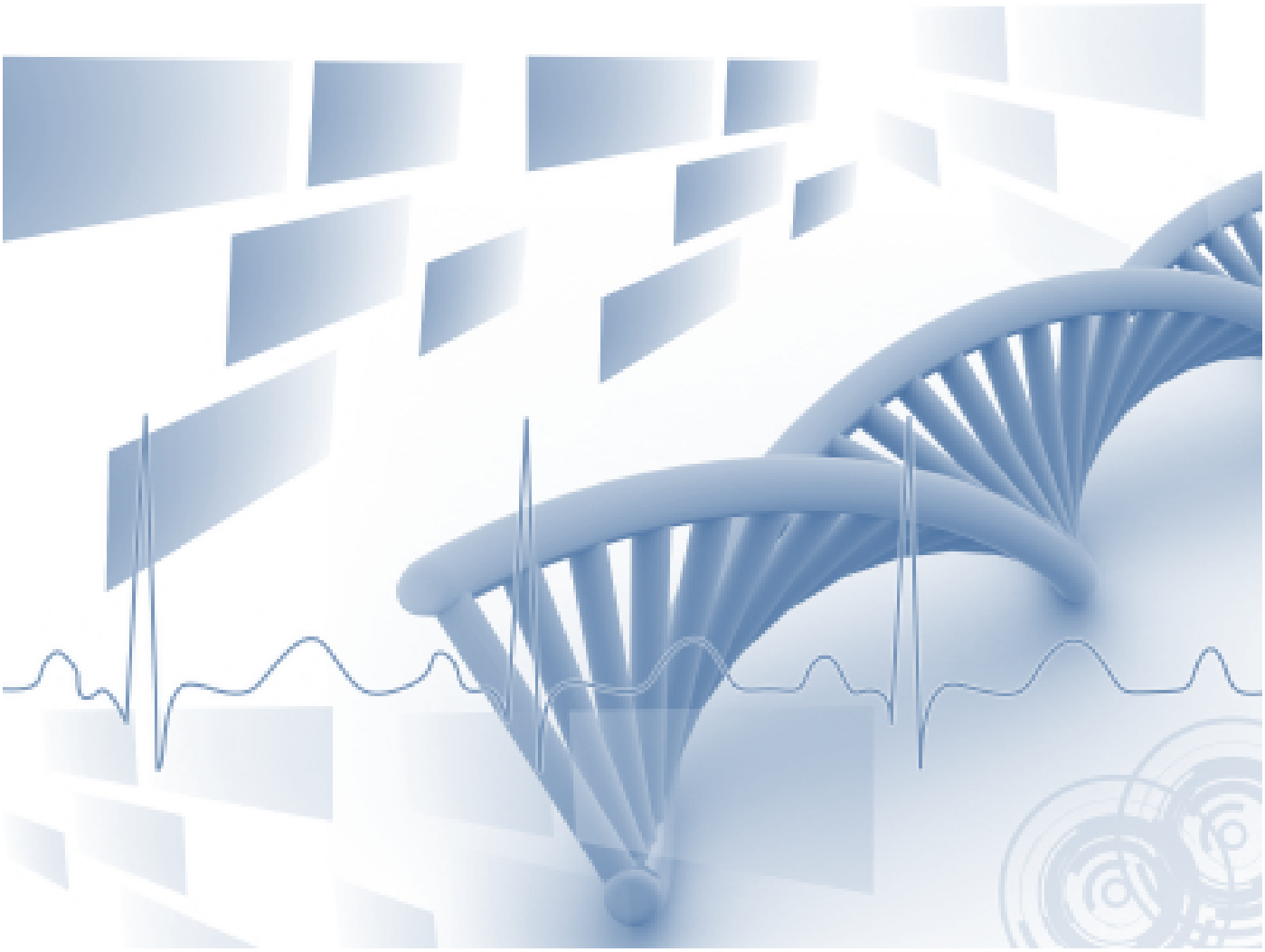
# Using Science, Technology, and Society Studies Research to Move Genomics Discoveries from Bench to Bedside: Identification of Data Integration and Socio-technical Issues Arising in Personalized Medicine & Translational Bioinformatics



**Ellen Balka**, **Ph.D**
Senior Scholar, Michael Smith Foundation for Health Research; Professor, School of Communication, &Director, Assessment of Technology in Context-Design Lab
Simon Fraser University

**Stephan Struve**, **BA**
Research Assistant, Assessment of Technology in Context Design Lab
School of Communication
Simon Fraser University

**Maryam Ali Ficociello, Ph.D.**
Post-doctoral Fellow
Assessment of Technology in Context Design Lab,School of Communication
Simon Fraser University

*ABSTRACT: In this report we apply concepts from science, technology and society studies (STS) as a means of identifying and addressing issues and challenges which arise with the movement of genomics research from discovery environments (the bench) to the bedside (implementation as part of personalized medicine). Starting with concepts from STS, we explored socio-technical issues related to data integration and other challenges in this highly interdisciplinary field. Data were collected using social science methods such as ethnographic studies of two pre-clinical genomic labs and in-depth interviews with stakeholders and researchers in the genomics community. Analysis of data was supported by the use of qualitative data analysis software, which allowed our team to use systematic methods to develop insights about the challenges that arise in moving discoveries from the pre-clinical environment to the bedside.*

# Contents

## Figures



## Tables



**We gratefully acknowledge the financial support from Genome BC through their Strategic Opportunities Program (SOF), to carry out this work.**

## Executive Summary

In this report we apply concepts from science, technology and society studies (STS) as a means of identifying and addressing issues and challenges which arise with the movement of genomics research from discovery environments (the bench) to the bedside (implementation as part of personalized medicine). Starting with concepts from STS, we explored socio-technical issues related to data integration and other challenges in the highly interdisciplinary field of genomics research.

Data were collected using social science methods such as ethnographic studies of two pre-clinical genomic labs and in-depth interviews with stakeholders and researchers in the genomics community in British Columbia. Analysis of data was supported by the use of qualitative data analysis software, which allowed our team to use systematic methods to develop insights about the challenges that arise in moving discoveries from the pre-clinical environment to the bedside.

Recommendations address four areas: improving the capacity of pre-clinical labs; improving the capacity of trainees; reducing barriers through changes in financial support, and reducing barriers through support for cultural changes and cross-stakeholder collaborations. Recommendations are addressed in greater depth in the final section of this report.

***Recommendations***

*Improving Capacity of Labs*

**Recommendation 1**: Support labs in developing organizational memory strategies for written documentation of lab practices, as well as more robust documentation and contextual information about data.

*Improving Capacity of Trainees*

**Recommendation 2:** Develop targeted educational strategies to enhance ability to work across disciplines.

**Recommendation 3**: Develop case examples for teaching that encourage critical thinking about data quality, affordances and constraints of tools, etc. which can be used to encourage awareness of the relationship between tool use and findings.

**Recommendation 4**: Design case based learning resources which highlight issues related to standardization (e.g., the lack of data and tool standards, where standards exist, limitation of standardization, etc.) which can be integrated into varied courses concerned with genomics.

**Recommendation 5**: Develop something akin to a library research guide to assist trainees in identifying resources that are particularly good for addressing certain kinds of issues (description of forums and other resources related to problem solving while undertaking lab-based work).

*Reduce Barriers Through Changes in Financial Support*

**Recommendation 5:** Build funding for core lab technicians into funding programs.

**Recommendation 6:** Increase funding available to validate findings, and move from academic accuracy to clinical accuracy.

*Reduce Barriers Through Support for Cultural Changes and Cross-Stakeholder Collaboration*

**Recommendation 7:** Host a workshop to be attended by senior members of BC's genomics research community, to discuss issues of data ownership, intellectual property and the role these play in willingness to share data across labs.

**Recommendation 8:** Host a cross-sectoral workshop to identify constraints to data sharing and linkages related to genomics, and the development of strategies for addressing public concerns while reducing barriers to data sharing and linkages for research purposes.

**Recommendation 8a:** Support the development of consent language and consider the development of unified and centralized general consent forms to allow researchers to conduct, for instance, secondary analysis of pre-existing samples without re-consenting.

**Recommendation 8b:** Support development of guidelines and/ or standards for de-identification of data as a means of providing data stewards with guidance about how to share data and remain compliant with regulations.

**Recommendation 9:** Provide financial support for deliberative dialogues and other forms of public engagement to address issues of privacy, discrimination, data sharing and secondary use of data and informed consent, as related to genomic data

**Recommendation 10:** Provide funding for a cohort of bio-ethicists to gain exposure to genomics research through becoming an embedded member of genomics research teams, in order to gain more practical experience with the issues and challenges genomics scientists face.

## 1. Introduction

Advances in genomics research and translational bioinformatics have set the stage for the development of new forms of targeted and preventative healthcare. Personalized Medicine (PM), or Personalized Health (PH), is on the cusp of transforming Canada's healthcare system (e.g. Bottinger, 2007; Cascorbi, 2010; Evans, Meslin, Marteau, and Caulfield 2011; Lesko, 2007). Defined as "the application of genomic and molecular data to better target the delivery of health care, facilitate the discovery and clinical testing of new products, and help determine a person's predisposition to a particular disease or condition" (Abrahams, Ginsburg, and Sliver, 2005, p. 396), PM envisions that the knowledge of a person's genomic profile has the potential to guide preventive and acute health care delivery, providing information about the most effective and safest course of treatment on an individual basis (Abraham et al., 2005, also see Figure 1).This has major implications for cost reduction and improved health outcomes in the health care system (Fackler and McGuire, 2009).


*Figure 1: The Economist*

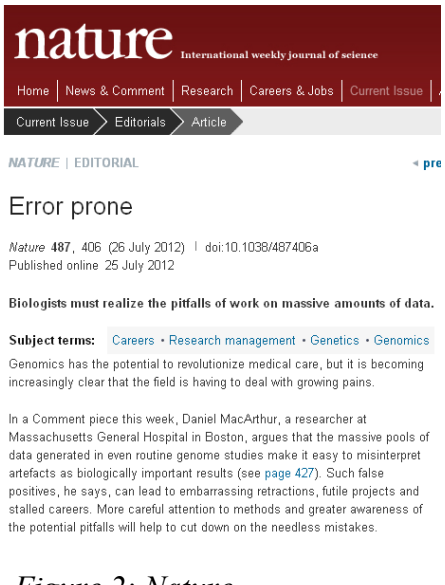



*Figure 2: Nature*

However, research in the area of personalized medicine faces many challenges, which have a socio-technical dimension. These challenges include data representation, linkage of heterogeneous data sets (e.g. Louie et al, 2007), and data integration (Louie et al., 2007; Payne et al., 2009) – just to name a few. Despite the promise of new pipelines and platforms, adoption of genomic discoveries has been lower than hoped, owing in part to socio-technical barriers (see figure 2). As efforts to link genetic data to electronic medical records are reported in the popular press (MacArthur, 2011, see Figure 1), scientists have commented on the small number of cases that demonstrate the efficacy of personalized medicine in a clinical setting (Laurence, 2009) Laurence (2009) states that "there are very few specific examples of its utility in the clinic, and among those gene-based tests that are used, many issues are far from settled" (p. 269). Others have noted that "transforming the great promise of these new technologies into clinical laboratory tests that can help patients directly has happened more slowly than anticipated" (IOM, 2012, p.1). Khoury (2009) echoes this somber

opinion and summarizes that the field of personalized medicine ‘‘is evolving, but it’s still in its infancy. (p.5)”

***The Contributions of this Project***

Many issues and questions related to ethical, environmental, economic, legal and social issues related to genomics ($GE^3LS$) remain. Reflecting “the high degree of hope placed in the promise of omics-enabled technologies and medical care” (IOM, 2012, p.2), this project sought to identify issues and challenges arising with the movement of genomics research from laboratory settings to the practice environment, by exploring the challenges genomics scientists faced in realizing the goals of personalized medicine through application of a science, technology and society lens to emergent issues and challenges in genomics. Using theoretical insights and research methods from science, technology and society studies, this project brings issues and challenges to light in a new way, and offers new approaches to resolution of challenges arising with the movement of genomics research from bench to bedside.

Areas we explored included:

- The use and development of software tools and databases in genomic research;
- The establishment of consistent data standards and data sharing practices;
- Ethical issues related to participant consent in genomics research;
- Ethical issues arising out of the publication and use of human genomic data;
- Issues arising in relation to research-related regulation and funding;
- The use of qualitative research design to explore socio-technical issues in genomics.

Within the field of science, technology and society studies, our research drew in particular on what is known in the Unites States as cyberinfrastructure studies, and in Europe as e-science studies.

In the context of genomics and personalized medicine, few studies to date have examined socio-technical issues arising from the practices of data integration in its most raw form: at the bench (i.e. in health science laboratories). Here, we report findings from research focused on understanding the production of data in genomics and bioinformatics labs, and issues which arise amongst a diverse group of stakeholders, as they attempt to build on such pre-clinical work to move genomic sciences into implementation in varied healthcare settings. We suggest that failing to address some of the issues we have identified here in processes involved in pre-clinical genomic research work at both the micro and macro level will result in the often-promised benefits of personalized medicine remaining elusive.

It was not the goal of this project to discuss or critique pre-clinical genomic science itself, but rather our goal was to identify socio-technical practices and issues surrounding them which, if addressed, may help move the application of genomic discoveries into a health care context more quickly. In our final section of the report, we include recommendations which emerged from our work with genomic scientists, stakeholders, and policy-makers in BC and elsewhere in Canada.

Below, we describe the project and provide an overview of our study approach. This is followed by a description of the research methods and approach to analysis taken in studying two pre-clinical genomics laboratories, and conducting in-depth interviews with varied stakeholders in British Columbia’s genomics research community (section 2). We then present our findings (section 3) followed by a discussion (section 4) and recommendations for action and further research (section 5).

## ***Overview of Project***

The goal of this project is to apply concepts from science, technology and society studies (STS) to issues and challenges arising in the movement of genomics research from the bench (labs) to the bedside (use), in order to yield new insights which may assist British Columbia's genomics research community in achieving their goals.

This project builds on BC's strength in the area of science, technology and society studies, and the strengths of BC's genomics scientists by using genomics research and the business sector as empirical sites for applied research. This project addressed conceptual issues in general and issues related to data in particular which genomics researchers in BC and elsewhere are facing as they work collaboratively across several domains and disciplines in an effort to realize goals in the area of personal medicine and translational bioinformatics. As translation of genomics research from bench to bedside is arguably constrained by the policy environment, we have included a discussion of governance issues and policy as well.

Our research and findings outlined here will help to highlight challenges genomics scientists face in working with multiple data sets, which in turn will contribute to the development of novel approaches to data integration which incorporate insights about social constraints, as well as technical challenges arising with data integration. These insights will help BC maintain a strategic position in realizing benefits from the data intensive genome sciences. Work outlined here takes an original approach to issues arising in a pre-clinical genomics context by focusing on insights from science, technology and society studies, particularly those from e-science and cyberinfrastructural studies, to develop insights about the challenges inherent to integration of data (e.g., from health registries, genomics data and data about social determinants of health) from multiple sources and collaborative work environments.

This project used innovative approaches to apply theory and methods from science, technology and society studies to data integration problems in biomedical genomics. It is of strategic importance to BC's life sciences and genomics sectors, because fully realizing benefits of genomics requires resolution of data integration issues. This report outlines findings from two in-depth case studies of two bioinformatics labs and 12 in-depth interviews with key stakeholders representing diverse areas essential to personalized medicine and translational bioinformatics.

## ***Setting the Stage***

Our current work- which is concerned with data integration practices of pre-clinical genomic scientists[1] can be understood in relation to three related literatures: translational bioinformatics and personalized medicine, information infrastructures, and socio-technical approaches to the study of technology. Each is outlined briefly below.

### *Translational Bioinformatics and Personalized Medicine*

Personalized medicine (which "uses an armamentarium of molecular (i.e., genetic) data, non-genetic data, demographic information, and clinical observations to define the best treatment and health outcome for patients" (Lesko, 2007, p. 812)) and translational bioinformatics are information intensive, multidisciplinary team based areas of study whose success depends on the integration of different kinds of data emanating from multiple databases, software tools and

[1] It is noteworthy to mention here that our two labs are 'research labs', which often experiment with new and emerging tools that are not known to their practice. However, in diagnostic labs which work closely with bench-side medicine, the challenges mentioned above are not common place as researchers there are very well versed in the tools, data handling and analysis; which they use regularly.

instruments. The integration of data from multiple sources has been identified as a deterrent to the success of personalized medicine and translational bioinformatics (Louie et al., 2007; Payne et al., 2009), and Payne et al. have suggested that despite the promise of data integration platforms, adoption has been low, owing in part to socio-technical barriers and ownership and security issues. Payne et al. have suggested that resolution of these issues will require attention to socio-technical issues, and an improved understanding of human factors that need to be overcome in data integration. They suggest that overcoming these challenges will also require community-based consensus. Research outlined here—which is anchored in science, technology and society studies (STS)—is aimed at identifying socio-technical issues related to data integration in BC's biomedical genomics research community, addressing those issues, and developing processes within the BC Genomics research community which can support community-based consensus and new insights about how to support the movement of pre-clinical research in genomics from bench to bedside.

Personalized medicine (PM)—defined as "the application of genomic and molecular data to better target the delivery of health care, facilitate the discovery and clinical testing of new products, and help determine a person's predisposition to a particular disease or condition"(Personalized Medicine Coalition, 2005, cited in Cascorbi, 2010, p. 749) has been the subject of numerous headlines and editorials (e.g., Laurence, 2009; Evans et.al., 2011; Cascorbi, 2010; Lesko, 2007; Bottinger, 2007), which suggest that public expectations of what personalized medicine can deliver in the short term differ markedly from reality (Lesko; Evans et.al.). Lesko (2007) suggested that PM existed more in conceptual terms than in reality—a sentiment echoed somewhat by Fackler & McGuire (2009, p. 1) who suggested that "different categories of stakeholders focus on different aspects of personalized genomic medicine and operationalize it in diverse ways." Fackler & McGuire identified three elements of PM (molecular medicine, pharmacogenomics and health information technology, or HIT), which they suggest, if integrated, have the potential to improve health and reduce costs of care, but which also present many challenges.

Although the term PM "commonly refers only to gene-based health care" (Laurence, 2009, p. 269), Ginsberg and Willard (2009, p. 281) suggest that molecular data would be best combined with "various molecular and clinical tools to refine the risk of developing disease as well as screening, diagnosis, prognosis, and therapeutic selection."Many proponents of PM suggest that important gains will come from consideration of genetic data alongside other forms of data (e.g., Lesko, 2007). As efforts to wed genetic data to electronic medical records are reported in the popular press (MacArthur, 2011), scientists are quick to point out that few examples of the utility of PM exist in clinical settings (Laurence), and caution that "if we fail to evaluate the considerable promise of genomics through a realistic lens, exaggerated expectations will undermine its legitimacy" (Evans et al., 2011). Research in the area of PM is data intensive and the creation of knowledge from data requires researchers to integrate large and diverse data sets, which presents numerous challenges in areas such as data representation, and the linking of heterogeneous data sets (data integration) (Louie et al. 2007).

The need to create knowledge from data, particularly through integration and analysis of heterogeneous data has also been addressed within the context of translational research in general, and translational bioinformatics in particular. The term translational research refers to the need to move research from 'bench-to-bedside' by improving the interface between basic science and clinical medicine (Woolf, 2008), while translational bioinformatics refers to "the

development of storage, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health." (AMIA, n.d.). Translational bioinformatics (sometimes referred to as translational biomedical informatics) sits between and (ideally) interoperated with health information technologies (HIT) and electronic medical records, clinical research informatics systems, statistical analysis and data mining. It includes—and draws on—the areas of genomic sciences, biomedical sciences, and health informatics. Payne et.al. (2009, p. 131) suggested that although the need to collect, manage, integrate, analyze and disseminate large-scale heterogeneous (biomedical) data sets is a common theme in translational bioinformatics research, "well-established and broadly adopted theoretical and practical frameworks and models intended to address such needs are conspicuously absent." Louie et.al. (2007) argue that many of the challenges associated with the data intensiveness of PM (and, we suggest, translational informatics) "can be classified as data integration problems, and technologies exist" that can address these challenges. The notion that data integration problems are primarily technical problems is challenged by researchers versed in science, technology and society studies (STS). For example Leonelli (2011) argues that the development and effective use of such technologies is strongly dependent upon researchers' understanding of their fields and the data being mined. Balka (2005) and Schuurman & Balka (2009) have shown how local data collection / production practices present challenges when integrating data from multiple sources, and Bowker and Star (1999) have highlighted the importance of classification systems in knowledge generation.

Payne et al. (2009) outline a number of information management challenges in translational research related to data integration, including socio-technical challenges, and argue that a framework should be developed to enable categorization and conceptual integration of major information sources, which can assist in a) identifying major sources of information and how they relate to one another; b) providing researchers with an ability to understand how their activities contribute to broader knowledge translation goals and evidence generation that span multiple domains; c) supporting development of "cross cutting technology and socio-technical approaches that are specifically targeted at achieving high-level, translational integration spanning what are often distinct data, knowledge and/or evidence silos" (Payne et al., p. 134). Such a framework should address at least three information types—data pertaining to individual and/or population phenotypes (typically generated via public health, clinical care and clinical research); individual and/or population biomarkers (e.g., genomic, proteomic and metabolic expression profiles); domain knowledge (verified biomedical knowledge in specific domains, including public and private databases and ontologies and terminologies that formalize descriptions with specific domains), and biological models and technologies. Payne et al. have suggested that despite the promise of data integration platforms and metadata solutions, data ownership, security and socio-technical barriers exist which will require an improved understanding of human factors and community based consensus to overcome.

*Cyberinfrastructure Studies*

Cyberinfrastructures (CIs) have been developed to support large-scale data sharing and integration initiatives that promise support for scientific communities as well as interdisciplinary activities. Definitions of CI (referred to as e-science in Europe) abound. They vary in the extent to which infrastructural issues are conceptualized as technical problems to be solved or as socio-technical processes. Our use of the term CI here is consistent with STS perspectives of

technology, in that it recognizes the inter-relatedness of the social and technical in the production of CIs, and also recognizes the contingent nature of CIs. Hence here we define CI as "the set of organizational practices, technical infrastructure and social norms that collectively provide for the smooth operation of scientific work at a distance" (Edwards et al., 2007, p. 6). Edwards et.al. (2007) suggest that CIs will fail if any of these aspects are ignored. They suggest that converging histories, path dependencies, serendipity, innovation and 'bricolage' (tinkering) are all required for the eventual growth of CIs, and that the use of language that locates CIs as merely machines that must be built or technical systems that must be designed downplays the importance of social, institutional, cultural and other non-technical problems that arise in the development (and use) of CIs. They turn away from the language of design and engineering and attempt to reframe their discussion of CIs in "a more organic lexicon" (p. 7). Hence they speak of the "the patient art of 'growing' infrastructures" (p. 38), which they suggest will depend less on a Herculean figure or master engineer than it will on a series of modest, pragmatic and "strategically informed interventions undertaken on the basis of imperfect knowledge" (p. 39).

Ultimately, CI studies are concerned with identifying characteristics of large-scale computer infrastructures and issues which arise as scientists working across multiple scientific communities come together to generate knowledge. Databases play a central role in CIs. Schuurman (2008) has coined the term database ethnography to describe the process of investigating and making decisions made about data in databases visible, which is necessary to insure that data are used with integrity, and 'facts' are not stripped from their context.

In order for us to reach personalized medicine that truly makes use of cyberinfrastructures, there is a need to translate much of the research that is carried out in the pre-clinical stage, as well as the data related to discovery and diagnosis so that it is accessible by healthcare professionals who will ultimately implement and use the bench-side research findings. Hence, in carrying out this study we sought to understand two sets of processes—and the issues arising for bioinformatics researchers in relation to each. First, we sought to understand processes surrounding work in bioinformatics labs, and the issues arising in the everyday work of bio-informaticians in lab settings, and particularly those related to the work of producing and using data (the micro-level). Second, we sought to understand issues arising in relation to the broader landscape of bioinformatics research in British Columbia (the macro level). We employed ethnographic participant observation in two bioinformatics labs—one wet lab and one dry lab—to gain an understanding of everyday work practices of bioinformatics researchers in lab settings. Additionally, we conducted interviews with varied stakeholders within the broader bioinformatics research community in British Columbia in order to gain insights about issues and challenges genomic and bioinformatics researchers face in relation to processes which transcend individual labs or research groups.

Our data collection and analysis methods are outlined in Section 2 below, and our findings are summarized in Section 3.

*Science, Technology and Society Studies and Socio-technical Approaches*

Research outlined here draws on the theoretical insights from cognate areas of study within science, technology and society studies (STS) and particularly research concerned with sociology of infrastructures (cyberinfrastructures). Although technology has eluded definition amongst STS scholars, in the context of this work it is nonetheless helpful to define it. Bush's (1981) definition of technology captures many elements of technology which have been important to

contemporary STS scholars, and which lay a foundation for the socio-technical approach to the study of technology which has informed our approach:

> *"technology is a form of human cultural activity that applies the principles of science and mechanics to the solution of problems. It includes the resources, tools, processes, personnel, and systems developed to perform tasks and create immediate particular, and personal and/or competitive advantages in a given ecological, economic and social context" (Bush, 1981, p. 1).*

Central to this definition of technology—and much contemporary STS research—is a focus on the social nature of technological development. New technologies don't simply pop out of the sky, but rather result from the coordination of vast networks of people, machines, processes, materials, what Latour (2005) calls actor networks. The technologies we live with seldom represent the only technological solutions possible; but rather, represent the outcome of a complex series of negotiations which occurred within a web of both human and non-human actors (e.g., the installed base of a computer system). Technology development and use are both social processes, whose outcomes depend upon social and technical factors.

Contributions of social scientists to the study of technology have been notable. For example, Suchman (1987) argued that plans differ from situated actions with respect to human interactions with technology. This seemingly obvious point had long been overlooked by technologists, who typically developed and introduced technologies based on formal descriptions of work (plans) rather than observations of actual work (situated actions), which Suchman's research demonstrated reflected a responsiveness to complexities of work situations that were often addressed through tacit knowledge. Suchman's notion that plans differ from situated actions stimulated interest in ethnographic approaches to the study of technology use in context, which is now undertaken by scholars working in several technology design disciplines (e.g., computer supported cooperative work, usability studies, and human-computer interaction). Work undertaken by our team has been rooted in this tradition. We utilized ethnographic methods to develop an understanding of issues arising for genomics scientists, as they worked with data.

While not all STS scholars are concerned with policy issues, the role of policy in the management of science and technology has been a significant concern within STS communities. Historically STS scholars have looked at a range of topics related to policy, including the relationship of policy to the management of technology (and particularly whether or not policy contributes to just and equitable benefits of technological development), and the role of policy in supporting and/or constraining the development of technologies.

Although policy was not an explicit focus of our work at the onset and a thorough treatment of policy issues was beyond the scope of this project, nonetheless policy issues emerged in our fieldwork as well as background work we undertook in support of the project. As policy issues have played and will continue to play an important role in data collection, handling and integration related to genomics, we have addressed them in somewhat greater depth than other issues in presentation of findings as well as our discussion.

## 2. Research Design and Methodology

Research outlined here follows a multi- method qualitative research design. We conducted two in depth ethnographic case studies (one in a wet lab, the other in a dry lab) which were

supplemented by in-situ interviews in those lab settings, in order to understand the issues and challenges genomics researchers faced in relation to data (producing data, working with data, etc.), in lab settings. Additionally, we conducted in depth interviews with key stakeholders in the British Columbia genomics community, in an effort to gain insights about problems they identified in aspects of their work, related to data and data handling.

## ***Ethnographic Case Studies***

In this section we present the portion of our project that is related to the ethnographic case studies. We discuss the objectives that drove this ethnography, our data collection and analysis methods.

### *Objectives of Case Studies*

The objectives of the ethnographic case studies component of this project were twofold:

1. to contribute to the development of a framework to enable categorization of sources of information required to realize the goals of personalized medicine and translational bioinformatics; and
2. to identify socio-technical challenges inherent to data integration processes, develop strategies for overcoming these challenges, and communicate insights to BC's genomic sciences communities.

Our focus during data collection was on improving our understanding of pre-clinical genomic scientists' work (i.e. in a pre-clinical lab setting). With this aim, we were able to gain insights that can aid us in developing strategies and tools to be used by genomic scientists to improve data quality and data management. In addition, through observing the work of pre-clinical scientists working with gene sequences, we also anticipated identifying a broad range of other issues of interest to natural and social scientists would emerge, which warrant additional attention by scholarly and policy audiences.

## ***Ethnographic Data Collection Methods***

Data collection methods included ethnographic participant observation in two labs, which we refer to here as the Alpha and Zeta Labs. The case studies were undertaken to gain in depth insights about issues arising in relation to genomic data production, use and integration. The focus of observations was on the identification of scientists' work with data, including identification of tacit knowledge; their awareness of affordances and constraints of data sets and tools with which they work to manipulate, view and analyze data, and challenges—both social and technical—they face in their work with data. These in depth case studies were undertaken in order to determine what kinds of data genomics scientists work with, from what instruments/ sources, what software they use to manipulate/ store and view their data, what kinds of data they would like to integrate in the future, and what concerns they have about data integration.

The ethnography comprised of two, in-depth case studies of genomics labs. Each case study employed several means of data collection, including semi-structured interviews with researchers, in situ observations of researchers at work (regular lab meetings, shadowing of key researchers, and interviews with lab members), analysis of documents pertaining to regular meetings (lab presentations and lab meeting minutes), and data from two widely-used discussion forums about bioinformatics, in order to gain insights about genomic scientists' everyday work practices.

Data collection was carried out between September 2011 and July 2012. The majority of interviews were recorded and transcribed. One interviewee opted out from having her interview recorded, and so detailed hand-written notes were taken instead for her interview. Hand-written notes were also taken throughout observations and also transcribed. Data collected through ethnographic observations and interviews related to pre-clinical laboratory activities was captured at the following events:

| Source | Alpha Lab | Zeta Lab |
|---|---|---|
| General Lab Meeting Observations | 35<br>Approx. 1 hour` per meeting | 23<br>Approx. 2 hour per meeting |
| Shadowing | 5<br>Approx. 2 hour per observation | 6<br>Approx. 2.5 hours per meeting |
| Interviews | 6<br>Approx. 0.75 hour per interview | 6<br>Approx. 0.75 hour per meeting |

*Table 1: Ethnographic Research Data Set*

Observations were carried out during regularly scheduled lab meetings (which typically lasted one hour each week in the Alpha lab, and two hours in the Zeta lab). Additionally, bi-weekly meetings were also held for the Alpha lab for an hour. Observations were undertaken during job shadowing, during which time the researcher observed the following types of activities in the wet lab: worm-based experiment preparation and execution, and dry lab: analysis of bioinformatics data and pre-experimental research. Research participants were a combination of graduate students, postdoctoral fellows, lab technicians, and bioinformaticians (web and database developers). All research participants gave their informed consent to take part in this research.

In order to protect the anonymity of those who participated in this research, we have declined to include more specific demographic information about lab members who graciously allowed us to observe, shadow and interview them as the labs are relatively small and it would potentially be easy to identify lab members.

For purposes of data collection, we identified the unit of analysis as data handling processes surrounding two specific technologies that are commonly used in both lab settings: Perl programming language and Structured Query Language (SQL) in the dry lab. Our focus during observations was on how these technologies were used to handle and analyze data, how data were used, produced, and integrated within or across different technological platforms, and identifying the sociotechnical challenges that arose with these practices.

*Overview of Case Study Laboratories*

Here we refer to the two labs which were the focus of our ethnographic case studies as the 'Alpha' and 'Zeta' labs.

*Description of Alpha and Zeta Labs*

Although the Alpha and Zeta Labs are situated in the same department and institution, each lab is slightly different in terms of research orientation and focus. Alpha Lab consists of an interdisciplinary wet lab and dry lab (computer lab) environment used to investigate pathogenomics questions. Computational analyses of genomics are combined with individual proteins and other lab data to facilitate experimenting with new hypotheses and testing them in model host systems. At the time of this research, the lab was comprised of 12 members. The staff members were primarily responsible for the ongoing maintenance of either the wet or dry lab. There was a wet lab technician, as well as web/database developers involved in managing the four databases that were created and continue to be supported by the lab. The remaining eight lab members were mostly scientists-in-training in the form of graduate students and postdoctoral fellows. The majority of research that was observed in this lab was carried out in the dry computer lab, as only one postdoctoral fellow occasionally carries out work in the wet lab. Within the Alpha Lab, our observations were comprised of two settings: general lab meeting observations and lab work observations. The ethnographer (MAF) attended 35 lab meetings, and conducted five dry lab work observations. Each lab meeting lasted an hour on average, while lab work observations lasted an average of two hours. The ethnographer's role as participant observer included volunteering to help with testing the usability and functionality of two of the lab's databases.

In comparison, the Zeta Lab's objective is to develop bioinformatics programs and tools for understanding genomic architecture and expression. There are two parts to this lab, a dry lab and a wet lab both of which are regularly used. In the wet lab, Caenorhabditis elagens (c.elagens) worms were used to carry out experiments. Dry labs are where computers (supported by powerful servers and grid technology) were used to run genomic computations. At the time of this research, the lab consisted of 17 members, including volunteers, master's students, PhD students, postdoctoral fellows and lab technicians. A total of 23 general lab meeting observations, five wet lab work observations, and one dry lab work observation was carried out. Each Lab meeting observations lasted an average of 2.5 hours.

In both labs, during the lab meetings, members would report about recent research progress to the lab director, who mentored and supervised students and staff, led discussions, and was quite involved in helping with on-going analyses of data, fine- tuning of scientists' methods, and carrying-out quality-control checks of the scientist's research standards. The lab meetings allowed each member to share their best practices in relation to tool use. Constructive feedback was constantly given to researchers along with advice about possible future research avenues they might pursue. In contrast to this setting, the wet and dry lab work settings were quite different. Here, the scientists mostly worked in solitude on their individual projects. Collaboration with other scientists was commonplace, but mostly in the form of trouble shooting in the dry lab, or carrying out a part of an experiment in the wet lab.

*Technology in Use: Perl and SQL*

Through our ethnographic work, we came to understand that most of the work of integrating data from multiple sources in our two lab field sites was accomplished with the use of two

technologies: Perl programming language and SQL database quarrying language. Both of these technologies are situated in the dry labs. The Perl programming language is a high-level, general purpose, dynamic programming language that has become widely used by biologists in bioinformatics research. It was originally developed in 1987 by Larry Wall as a general purpose scripting language to make report processing easier (Sheppard 2000). Perl facilitates easy manipulation of text files, with its ability to processes and detect patterns in data. The release of Perl version 5, which supports object-oriented programming, made it much easier to develop reusable modules of biology across research centers (Tisdall 2000). Perl was commonly used in the dry lab for string processing of biological data such as gene or protein sequences. The core work of dry lab scientists was in the manipulation of textual data sets from multiple databases and resources. Structured Quarry Language (SQL), on the other hand was vital for running quarries across databases. Both programming languages allowed scientists to access, extract, divide, or insert data sets from one or a combination of databases and convert them to other forms of output. Many scientists in bioinformatics did not necessarily receive formal training in the use of Perl or SQL, yet these two languages (amongst others) had become such important tools in bioinformatics in recent years that scientists routinely learned how to use them either on their own or from peers.

*Data Analysis*

The main objective of qualitative data analysis is "the transformation of data into findings and inferences" (Azeem & Salfi, p. 264). It involves "reducing the volume of raw information, sifting trivia from significance, identifying significant patterns, and constructing a framework for communicating the essence of what the data reveal" (Patton, 2002, p. 432).

Field notes were taken during observations, and later transcribed. After carrying out observations in these settings, more formal interviews were conducted (12 in total, ranging in length from 30 minutes to 60 minutes) with members of both labs. The interviews were carried out in order to answer more specific questions related to the challenges the researchers face during their work, and to illicit their views on their experience in terms of training and multidisciplinary work. All but one interview (see above) were recorded, and subsequently transcribed.

All data were systematically coded following Glaser and Strauss' (1967) qualitative data coding method. A computer based qualitative data analysis program (Nvivo 9) was used to facilitate this process. Whilst there are many approaches to coding, we have followed an approach in which some coding reflects our project's key questions (e.g., data integration). Coding categories and themes initially emerged from the data; and were subsequently verified by more than one researcher. Our coding scheme, which included topics and sub-topics (parent, child and grandchild nodes) appears in Appendix A. Our focus during coding was on identifying common themes, which occur across observations and interviews.

Another major component of our analysis was to draw on science and technology studies (STS) literature and particularly e-science/cyberinfrastructure studies (this literature is addressed in more detail the project introduction section of this report). STS is an interdisciplinary field of study that makes the production of science and technology subjects of study. It explores and seeks to understand how science and technology shape culture, values, and institutions, and how such factors shape science and technology. There have been a number of influential studies that followed the tradition of 'lab studies' and focused for example on how we understand tacit knowledge, and epistemology.

Data analysis was undertaken on an iterative, ongoing basis (Strauss and Corbin 1990, Miles and Huberman 1994). As we began our iterative coding processes (a method involving going back to already coded data when new coding categories emerged), six high level (or parent) categories (referred to as nodes in the context of NVivo) emerged. These were: analyses, challenges, cycles of credit, practicing science, using bioinformatics tools, and cases of interest. We present all our node definitions in Appendix A. Additionally, each 'parent' nodes was further categorized into sub-nodes. For example, the node Challenges has the following sub-nodes:

- Cost and funding
- Time
- Errors and technical problems
- Ethics
- Inherent knowledge biases
- Insufficient computing power
- Learning new tools
- Politics of data ownership

Furthermore, some sub-nodes also had sub-sub-nodes. Each code was defined and described in a memo to allow for consistent coding. Figure 3 (next page) shows the final codes for the Alpha Lab. Furthermore, transcripts from interviews with members of Alphaand Zeta labs and observations were systematically reviewed and segments of the transcripts were coded with the node or nodes that reflected that portion of our data. This process of transcribing observations and interviews into text and subsequently coding text makes it possible to perform searches, or queries, and retrieve (for example) all instances from our data set which have been coded as the same node. This in turn supports a systematic and thorough approach to our data analysis.

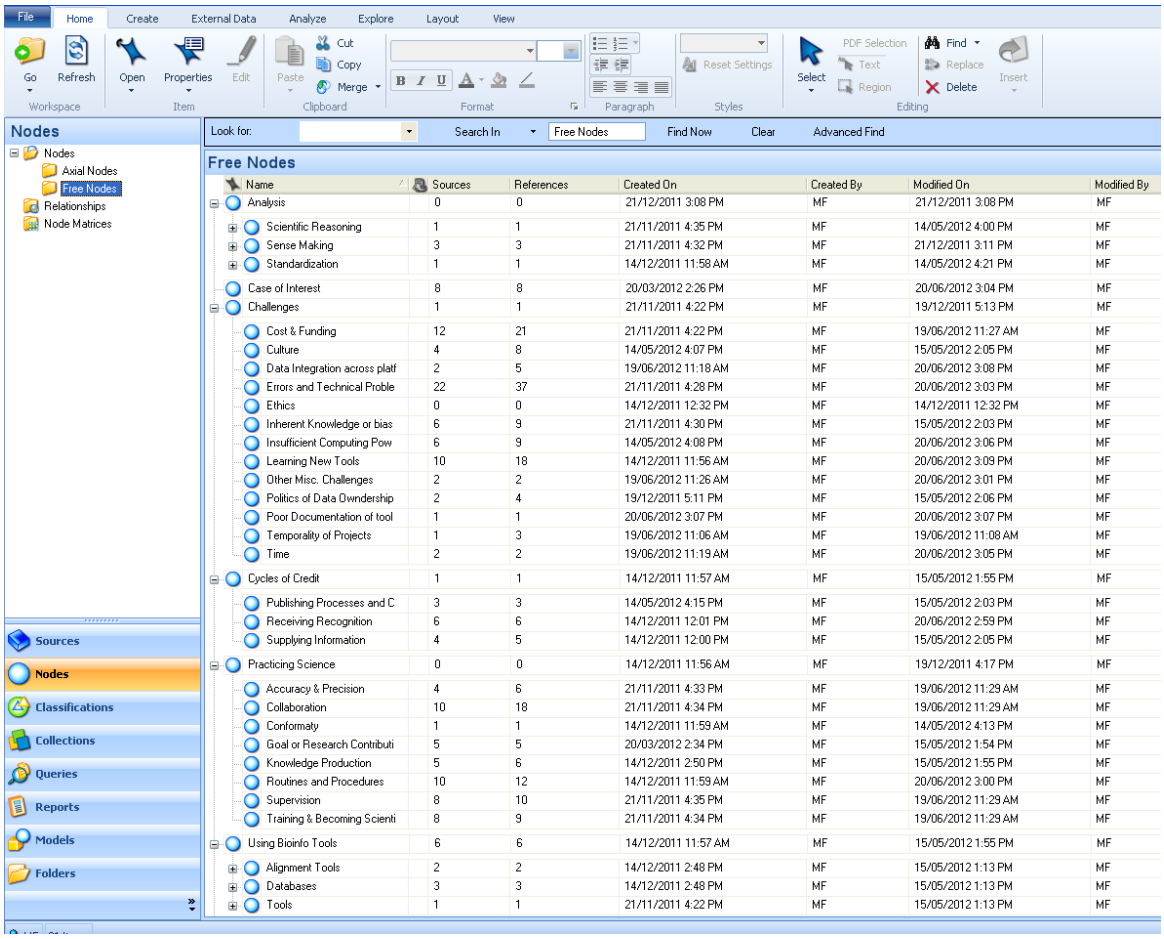



*Figure3: Snapshot of NVivo9 Project Nodes for the Alpha Lab*

NVivo 9, which we used for our data coding and analysis captures a range of information which supports a systematic approach to data coding and analysis. For example, in Figure 4, the column 'sources' refers to how many different sources of data (e.g., interviews, notes from each single observations, etc.) have had a particular code applied to them. The column 'references' refers to the total number of times a particular code has been applied across all sources of data. The program also allows users, for example, to capture similar information for a single source (e.g., how many times each code has been applied within a single source of data such as an interview or observation). The program also allows users to view on which date a node was created (which is important in terms of maintaining a systematic approach during iterative coding), captures which team member last applied that node to data, when the last application of coding for a particular node occurred, and which team member carried out that work.

During our analysis, we used memos – a feature of the NVivo software--as a way of recording our ongoing reflective notes and interpretations of data, and what might be learned from it, as it emerged, while we looked over different codes and transcriptions. The practice of keeping memos allowed us to keep track of our ideas as they developed throughout this research, and supported a systematic approach to coding data.

## ***In-Depth Interviews***

Our research plan included conducting interviews with a wide array of genomics stakeholders in BC. In light of our orientation towards science, technology and society studies, we were interested in the production of science and technology as subjects of study. While genomics research uses genomic technologies and bioninformatics as tools, science, technology and society scholars make the researchers and the tools they use the empirical site of our study, while exploring how science and technology shape culture, values, and institutions, and how such factors shape science and technology in return. A number of studies that followed the tradition of 'lab studies' have highlighted, for instance, how we understand tacit knowledge and how technologies are used in different settings.

We undertook in-depth interviews in order to gain new insights into challenges which arose as data from pre-clinical studies moved from the pre-clinical environments in which it was generated, into healthcare and business settings. This work served as a means through which pre-clinical studies were supplemented, and the breadth of stakeholders whose perspectives we learned about were expanded. Our preparation for this portion of the work included reviewing literature, and attending bioinformatics lectures over the course of a semester to understand the basics of the science and elements of this collaborative research area. Here we were introduced to a multitude of platforms, pipelines and databases, which also highlighted the very dynamic nature of the genomic research environment.

The genomic landscape is vast and includes various research domains and locations. In order to navigate the myriad network of actors and linkages, we mapped the landscape of institutions, centres, entities, programs and even stakeholders that were engaged in genomic research locally, provincially and federally.(Appendix B).We mapped the network of actors we identified using visual thinking software (Inspiration). We subsequently colour coded institutions and entities which appeared in the graphic as follows: federal and provincial entities (grey), academic institutions (green) and commercial groups (red). As we expected, genomic research is a Canada-wide endeavor in which BC's actors are closely interconnected with larger institutions, such as ministries or federal funding bodies. This visualization helped us to understand the interconnectedness of BC's genomics environment, and served as a foundation for identification of stakeholders for subsequent in-depth interviews.

### *Identification and Selection of Interview Participants*

What started as an attempt to understand and visualize the landscape of actors in the genomic sphere also helped us identify groups of stakeholders from which we sought interview subjects. We used a strategy of mapping stakeholders as a means to both identify potential interview participants, and as a means of better understanding the interconnectedness of the BC genomics community. From this work, we identified the following groups of stakeholders who we subsequently sought representation from in our in-depth interviews: 1) genomics researchers working in health-related genomic areas in British Columbia, in both the public and private sectors; 2) bioinformaticians, 3) professionals involved in cancer research or clinical health care delivery, 4) other stakeholders (including health sector administrators, etc.).

As anticipated, the size of the community is small enough that it was possible to conduct interviews with representatives from all research groups actively conducting biomedical genomics research in BC. Interviews followed a semi-structured format, to elicit information about areas of interest to the research team (e.g., research area and disciplines, tools/ software/

databases/ data used; recent and anticipated changes; challenges; future plans including data integration; views about data ownership and re-use, etc.), and allowed respondents to raise issues of interest to them. Interview recordings were transcribed, analyzed with qualitative data analysis software, following an approach to qualitative data analysis as outlined above.

*Identification of Genomic Scientists and Clinicians*

We compiled a comprehensive list of researchers working in pre-clinical and health-related research communities using university research directories, Genome BC directories, and by conducting searches in bibliographic databases (such as Medline). A review of the Genome BC web site (including past and currently funded projects), search of the Michael Smith Foundation for Health Research and Canadian Institutes for Health Research researcher directories, BC university research directories, the Canadian Life Sciences Databases, as well as BC wide directories identifying R&D efforts carried out in commercial businesses (Industry Canada registry, SFU Library's Biotechnology Industry Resources guide) was undertaken to discover local genomic researchers and related research entities. Additionally, staff in the two pre-clinical labs where our ethnographic work was carried out suggested potential interview subjects as well.

*Table 2: Overview of Potential Interview Participants*

| Researchers Identified | Contacted by e-mail | Respondents | Interviewed |
|---|---|---|---|
| **170** | 35 | 15 | 12 |

We identified and listed a total of 170 researchers / scientists as potential interview participants. From this master list (and with the aid of our map of stakeholders), we purposely selected scientists that had multiple affiliations (and hence could provide very complex insights). This process allowed us to strategically seek out members of the community to help us better understand issues of concern to the genomics health research sector, with a focus on researchers working in a pre-clinical setting, a health delivery context or working towards the commercialization of genomic discoveries. Thus, we gathered data across a wide spectrum of experiences.

After this selection process, we contacted a total of 35 researchers. We received 15 responses[2], 12 of which were available for interviews in our proposed time frame. Potential interview participants were initially contacted by telephone. If they indicated interest we subsequently emailed them a one-page information sheet that summarized our project goals and the interview process. We followed up with another telephone call and/or e-mail to schedule the interview. Table 2 below summarizes information about interview participants.

*Conducting the Interviews*

Semi-structured interviews were conducted (see Appendix C for the interview guide). This semi-structured format helped elicit information about areas of interest to us (e.g., research area and disciplines, tools/ software/ databases/ data used; recent and anticipated changes; challenges; future plans including data integration; views about data ownership and re-use, etc.), and allowed respondents to raise issues of interest to them.

[2] We attribute the high number of non-responses to the high absence during the summer season which was our field work time frame.

We also used the interviews as a way to obtain conceptual clarification about various aspects of genomics research and clarify questions we had about the tools involved. Our questions sought to explore the meaning and the conceptual dimensions of different challenges, as well as their significance within the research community. Similar to Kvale's (2007) definition, we hoped to take the scientists and researchers on "a joint endeavour to uncover the essential nature of a phenomenon," such as the socio-technical challenges in genomic research (Kvale, 2007, p.7).

| Participant ID | Gender | Current Role(s) / Job Title(s) | Areas of Focus |
|---|---|---|---|
| **SOF1001** | Male | Professor, Chair | Pharmaceutical Science, Drug Research and Development |
| **SOF1002** | Male | Professor, CEO | Medical Genetics, Diagnostic Biomarkers, Drug Development, Proteomics |
| **SOF1003** | Female | Scientist, Associate Professor | Medical Genetics and Disease Pathways |
| **SOF1004** | Male | Co-Director, Senior Scientist, Professor, Chair, | Medical Genetics, Developmental Genetics, Epigenetics, Genomics, |
| **SOF1005** | Female | Executive Director, | health policy research, hospital management, pharmaceutical market research |
| **SOF1006** | Male | Associate Director, Senior Scientist, Associate Professor | Bioinformatics, cancer genomics, pipeline development and evaluation |
| **SOF1007** | Male | Co-Director, Senior Scientist, Professor | Medical Genetics, Bioinformatics, comparative genome analysis |
| **SOF1008** | Male | Senior Clinician Scientist, Professor, Physician | Developmental Neurosciences, Child Health |
| **SOF1009** | Male | Senior Scientist, Associate Professor | Medical Genetics, Bioinformatics, High throughput data analysis |
| **SOF1010** | Male | Professor, Chief Informatics Officer | Genomic data mining, health informatics, development of biomarker panels |
| **SOF1011** | Male | Director, Co-Director, Professor | Pathology and Laboratory Medicine, Disease pathways, biomarker discovery, bioinformatics |
| **SOF1012** | Male | Director, Co-Director, Professor | Genetic Pathology, Cancer biomarker detection and development |

*Table 3: Attributes of Interview Participants*

*Data Analysis*

*Transcription Process*

Variation exists in how interview data are transcribed. We sought out and followed norms for transcription and analysis of bio-medical interview data as outlined by Halcomb and Davidson

(2006). Verbatim transcription has become a common strategy to deal with qualitative research in a health-care context and is widely considered to be integral to the analysis and interpretation of verbal data (Halcomb & Davidson, 2006). In line with our mixed-methods research design, using verbatim transcription allowed us to carry over the fine nuances of each interview. In our transcriptions, we transferred all sounds, word and pauses uttered during the interview into our transcripts. The verbatim transcription allowed us to trace verbal and non-verbal behavior throughout the interview, a strategy thought to bring analysts closer to their data.

*Data Analysis Software*

We imported the de-identified verbatim transcripts in NVivo 9, which is widely regarded as the most sophisticated qualitative data analysis software (Gibbs, 2002). This approach allowed us to "search for an accurate and transparent picture of the data whilst also providing an audit of the data analysis process as a whole—something which has often been missing in accounts of qualitative research" (Welsh, 2002, p.1). This made it possible for us to interact with our interview data and invite researcher's comments and reflections on coding and data analysis. This was particularly important because different members of the research team assumed responsibility for coding and preliminary analysis of specific portions of the data they had been engaged in collecting, and this annotation process served as a means of communication between various team members.

*Coding Process*

Throughout the entire coding process we followed an interpretive approach in line with a science, technology and society framework. As suggested by Glaser and Strauss' (1967) and parallel to the coding of our case studies, we systematically coded all qualitative data in conjunction with our project's key questions (e.g., data integration). Because we wanted to connect both sites of field work, we carried over the nodes from the ethnographic case studies to keep our coding framework consistent. Hence, we started out with the six high-level categories (which are referred to as nodes in the context of NVivo): 1) *analysis, 2) challenges, 3) cycles of credit, 4) practicing science, 5) using bioinformatics tools, and cases of interest* which had emerged during prior analysis of our case study data.

During the coding process categories and themes surfaced from the data and were subsequently verified by more than one researcher[3]. This step included open coding as well as substantive coding, in which we started to conceptualize the emerging themes and challenges. Similar to the approach taken in the ethnographic case studies (described above), we followed Bradley et.al. (2007) in such a way that we utilized inductive reasoning and the constant comparison method while employing predetermined code types (e.g., type of participant, type of materials, etc.). Emerging themes were categorized into sub-nodes of related pre-existing categories. For instance, the major high-level category (and therefore node in NVivo) 'Challenges' contained some of the following sub-nodes at the end of the coding process: Not surprisingly, there was some overlap with themes which emerged from data collected during the ethnographic case studies.

- Commercialization
- Cost and Funding

[3] We attached our coding scheme (including parent, child and grandchild nodes) and node definitions in Appendix A.

- Interdisciplinarity
- Time
- Ethics
- Politics of Data Ownership

Because the interview process was dynamic and participant-driven, we had to add many sub-nodes to account for the newly raised issues, which emerged as themes. In addition to adding sub-nodes, we also created sub-sub-nodes for further distinction and nuances raised by interviewees. Altogether we had a total of 107 nodes, sub-nodes and sub-sub-nodes in our coding scheme for the in-depth interviews. Appendix D serves as an example and shows the node structure in NVivo, as it evolved with the addition of the codes (and nodes, within NVivo) required for analysis of in-depth interview data.

Memos were used as described above in the section on data coding for ethnographic case studies, to ensure the reliability in the coding process. As was the case for the case study data, every time we created and defined a new node we systematically reviewed and re-coded segments of the each transcript.

This thorough and tiered coding process allowed us to systematically identify themes across interviews and helped us find themes across data sets (e.g., interviews only or case studies only) as well as themes which were common to both datasets (interviews and case studies of pre-clinical labs).

## 3. Findings

***Ethnographic Studies***

Two important levels of challenges emerged during this research: micro level challenges and macro level challenges.

Micro level challenges include issues that arise on a day-to-day basis while researchers carry out their work, and have a direct impact on their ability to continue or complete their work. Micro level challenges included seven main challenges: working with interdisciplinary teams, questioning research quality and validity, learning on the fly, challenges of technical compatibility, searching for support resources, top-down collaborations, and dead-end projects.

Macro level challenges are those issues that have an indirect influence on the work of scientists, and include: funding scarcity and biases, limitations of interdisciplinary collaboration/networking, lack of coherent standards, difficulty in developing and training personnel, and limited access to appropriate computing power.

Micro and macro level challenges are addressed in more detail below.

*Micro-Level Challenges*

We first present issues that arise on a day-to-day basis during the scientists work, and which hinder their ability to carry on their routine work activities. We refer to these as ‘micro’ since they directly come to bear on the progression of work. Each of the seven micro-level challenges is outlined below.

*a. Working Within Interdisciplinary Teams*

Scientists often mentioned the importance as well as the challenges of working within interdisciplinary teams. It had become the norm for the two case study laboratories to build teams within the lab as well as external to their labs with researchers that are from very different disciplines from their own. There were three categories of discipline-combinations we observed: some researchers were purely trained in biological sciences (usually molecular biology and biochemistry); some were solely computer scientists, and a third group were the converted group: those who came from one discipline (e.g., computer science) and were now training in the other discipline (e.g., biology). The marriage of the two disciplines and creation of this third emerging discipline of bioinformatics has rendered it common place for scientists to ask one another about their training background.

During our fieldwork, it was noted repeatedly that communicating across the disciplinary boundaries was often challenging. One biologist noted that when he tried to explain to the computer scientists what he was trying to do with his analyses, it was just hard to explain it in simple terms.

Because of this difference in training backgrounds and skills, and an ongoing need for insights from both disciplines (biology and computer science) to accomplish their everyday research, scientists often paired up with other researchers in the lab that had strengths that would complement their weaknesses. For example, skill-tradeoffs were frequently practiced. These consisted of (for example) scientists who were good at setting up worm plates performing this task for a colleague in return for a Perl script written by the scientists who were good at Perl.

*b. Questioning Research Quality and Validity*

Another issue that the researchers found quite challenging revolved around knowing whether or not they were asking 'the right questions' in their research, meaning questions that would be of interest to the academic community.

They were also concerned about the quality of their research. This concern emerged both in relation to the way they were executing experiments or running analyses of data, and also their overall analyses and sense-making of the subsequent results. They often looked for approval and confirmation from both colleagues as well as laboratory directors.

There was a constant consideration of biological and computational affordances and subsequent explanations of results.

*c. Leaning on the Fly*

Researchers referred to learning new skills 'on the fly' as they were carrying out their research. As it became evident that they did not know a particular component (sometimes a new programming language), they were trained to just go out and learn it on their own. They did find this however a frustrating as well as time consuming practice.

*d. Challenges of Technical Compatibility (e.g., across platforms or tools)*

Researchers often had to question their tool's biases as well as the different affordances each tool had. Each tool that was used had a different level of accuracy, carried out the analyses in a particular way, or presented the data in a certain way. All these affordances had consequences for how the results of using a particular tool could be interpreted.

One example is the use of different alignment tools when analyzing read sequences. There is no gold standard, and the researchers tended to use SSHAHA2, NovoAlign, Sametool or BWA to perform sequence alignment. Each one of these different tools presented slightly different results, with matches for Indels (insertions and deletions) for example being higher or lower depending on the tool used. This is an important difference as it has implications for how the data from use of each tool could subsequently be interpreted, as some of these tools result in slightly skewed data which could inappropriately influence their interpretation.

It is also noteworthy that research participants considered this item only fourth in order of importance, while one of the lab directors felt it should have been ranked the most problematic. This is perhaps an indication of the extent to which it has moved to the background for trainees as they have gotten used to coping with this issue by enlisting different workarounds rather than discussing the issue more or complaining about it. It is perhaps indicative of a culture of 'moving on' in whichever way this research community can to carry on their work in a timely fashion. At the same time, the importance given to this issue by one lab director signals its scientific significance.

i. Technical Errors

At the same time that scientists had to be aware of the affordances and constraints of each of the tools they used to perform various aspects of their analyses, they also had to maintain an awareness of technical problems or errors. These can appear as a result of poorly written code for example. Such problems and corresponding trouble shooting happen often, but often remain un-documented. They are however communicated verbally during general lab meetings. Both lab directors encouraged briefing the lab about both experiments that do work and those that do not work to avoid duplication of effort.

ii. Workarounds

Workarounds, defined as improvised methods for overcoming a problem or limitation in a program or a system, were commonplace due to the abundance of technical obstacles. Some workarounds were so common that the real method of completing the job was no longer referred to. For example, because the analysis of large sets of data requires so much computing power that is not available to many researchers, the workaround to 'sneaker it' (i.e. using your sneakers to run from one PC to another) to segment the analysis across many machines and hence speed up the processing time was common place for anyone who had data of a certain size. It became commonplace to use the term 'sneaker it' to refer to practicing this workaround in these labs.

While there were multiple workarounds related to creating a code, for example that does what the researcher wants it to do without going through the 'proper' way, there were also many workarounds that were practiced to deal with the scientist's limited access to sufficient computing power, as well as to eliminate time wasting.

*e. Searching for 'Support' Resources*

Scientists looked for support resources in the form of specialized electronic community forums, or other online discussion forums that they could turn to when 'things don't work' or for extra help in understanding specific issues related to their research. This was more common place for the bioinformatics component than it was for the biological or wet-lab component in these two bioinformatics lab case studies. For example, during a few observations we saw how one

scientist was seeking information on different web forums to understand the different affordances and constraints of the sequence alignment tools that were available for use.

*f. Top-Down Driven Collaborations*

Most collaboration is top-down, owing to the reality that lab directors sought and received funds that matched their research agendas. Funding bodies expect that students match their research programs with that of lab directors, which often results in students following the research program of the lab they belong to, rather than following their own research interests. Some researchers found this challenging, and desired more autonomy over choice of research program.

*g. Dead-End projects*

Many research projects were halted months or sometimes years into their existence, after it became clear that a particular avenue of inquiry or experiment was unlikely to yield promising results. Some scientists considered this poor use of project management tools. Scientists are very culturally bound by the traditional biological methods of doing things, and junior researchers at times felt that many of the projects lost momentum due to bad planning or execution of the project.

*h. Data-Related Challenges*

As a result of the work undertaken, we have learned that genomics researchers and bioinformaticians collect, use and re-produced data from various sources, including:

i. Multiple Sources

This includes sources such as previous or current work or experiments carried out by the scientists, a colleague, or other published materials;

ii. Multiple Mediums

Data was carried and transferred in multiple material carriers, including lab notebooks, email messages, presentations, published works, and excel sheets, word documents, to name a few;

iii. Multiple Modes

Some data was transferred in written formats, while others were transferred or handed-over verbally;

iv. Multiple Processes

Data underwent multiple levels of processing, including transforming it from raw data and numbers to more contextual data, carrying out computational algorithms to sum, average, or carry out other arithmetic functions.

The above four data-related processes create new challenges for the data and how it is being interpreted. For example, we wondered if these data handling methods need to be made visible for later stages of integration, and if there might be ways to represent the tacit/implicit knowledge for users later on. Additionally, we were left with questions about the mediating effects that different tools play and have on the data. These questions became important as this research unfolded.

In addition to the above micro-level challenges we observed which directly affect the day-to-day work of these researchers, macro-level challenges, which occurred at a distance from labs, yet still had a profound ‘indirect affect’ on the scientists’ work. These are discussed in detail below.

*Macro-Level Challenges*

Macro-level challenges are issues which occur at a distance from a specific lab or workplace—usually at a provincial or national level- but which have an indirect influence on what is done in a specific lab setting, or influences how the work is carried out. Macro-level issues are those that in a sense set the broad parameters surrounding the labs operation, and, as such, they indirectly come to bear on everyday activities in a lab. Macro-level issues comprise a sort of operating environment in which labs operate, and set down a system of opportunities and constraints, which lab directors must work within to carry out their genomics work. While the macro-level challenges are at a distance from the scientist's everyday work, they have an indirect influence on scientists work as they have many implications for them. We observed five different macro-level challenges, which we outline below.

*a. Funding Scarcity and Biases*

Most research projects are funded for a given period of time, for example, 6-12 months. Once project funding runs out; it is hard to dedicate more resources to the continuity of that project, even if there are clear benefits of extending that research. Additionally, funding bodies tend to favor funding 'new' initiatives, rather than funding the 'maintenance' or further development of existing projects. Scientists acknowledge this funding bias as a major limitation to their research.

For example, one of the lab directors once commented on how important it is that researchers realize that this is the way funding is allocated, for new and 'low-hanging fruit'- type projects, and that the only way to get money for 'maintenance research related to tool/website further development or to carry-on older research projects is to use money that comes in for new projects to hire staff and have those same staff spend some of their allocated time on the maintenance work. This is a good example of how a macro issue (funding requirements) influences everyday work (e.g., projects not maintained, which leads to other issues such as failure to document things perhaps).

*b. Limitations of Interdisciplinary Collaboration and Networking*

Scientific research is becoming more interdisciplinary with the provision of technology that makes it easier to collaborate across geographical boundaries. While this has potential for promising results, it raises a number of concerns related to differing practices and norms between different scientific cultures. One emerging challenge is the growing centrality of sharing and disseminating data, which makes standards of data formatting even more important. Closely related to this is the need to sometimes discard old data due to storage limitations, which is a large cultural change for most scientists and has been faced mostly with resistance.

*c. Lack of Coherent Standards (e.g., data integration across multiple platforms)*

Genomics and Bioinformatics researchers often need to import, export, or otherwise integrate data from multiple sources (different databases) to carry out their work. However, the multiplicity of formats that much of this data is organized in, as well as the embedded biases each data set carries (related to the individual tools used for sequencing) slows down as well as limits the progress of work. To date, there is no standard platform for most data output nor is there a standard for data curation in general, or any way of assessing levels or quality of data curation in particular. This poses many challenges when such data are subsequently used by other scientists who are removed from the local context where the original data were produced. The common practice has been to commence in an oral discussion (within an individual lab)

about the specific biases or limitations a dataset might have. But this oral handoff of the dataset between lab members is largely contextual and customized to the needs of the researcher who are using a given dataset. Scientists are aware of the need for more systematic recording of quality 'meta data' (data about data), which could make any biases that do exist in a dataset known to potential users.

Also, there is a need for established policies and legislation. Researchers believe that there is a need for more regulation in terms of establishing technical standards for data formats and research protocols related to quality of data annotation, methods of sharing and dissemination of data, workflow management, and technical compatibility of individual systems. There were also concerns of the absence of established legislation enforcing standards and the availability of standards for big projects or initiatives, rather than for smaller projects, which makes coherent work even more difficult to achieve. One solution that is sought after is an aim to make standardization a culture- by training scientists at early stages to think about and invest in standardization.

*d. Difficulty in Developing and Training Personnel (i.e., expensive, and not usually funded)*
There is a constant need to train and learn new computer science techniques in order to keep up with the rapid pace of development in genomics and bioinformatics research. Most scientists are either trained as biologists, or biochemists. Limitations in funding make it difficult to invest in training personnel to become more interdisciplinary, which is the current trend in the genomics and bioinformatics domain.

*e. Limited Access to Appropriate Computing Power*
One of the major constraints that scientists face is that the move towards more intensive 'dry lab' research (using computers and computer programs to run algorithms and test large sets of data against different hypotheses) requires access to extensive computing power, or grid technology, which is expensive. Apart from accessing university-based super computers, researchers constantly apply for grants to allow access to more computing power to run their large computations.

## ***In-Depth Interview Findings***

Four major sets of challenges arose during the in-depth interviews. Challenges associated with interdisciplinarity, financial constraints, inconsistent intellectual property (IP) policies and inconsistent solutions for addressing issues related to consent and ethical clearance procedures. Additional themes that arose included the politics of data ownership, culture, the challenges of making sense of data, information management, standards, learning new tools, data validation and commercialization. Each of these is discussed briefly below, and illustrated with excerpts from the interviews.

*Interdisciplinarity*

Genomic research is considered the functional marriage of biology and computer science (Chow-White& Garcia-Sancho, 2012). As both disciplines converge in a new field of socio-technological practices, various actors come together and have to bridge the gap between theory-driven academic research and the health care setting. The degree of interdisciplinarity is unprecedented as these actors come from very different fields of scientific discovery (such as bioinformatics or medical genetics) and collaborate with practitioners in clinical settings.

All of our interviewees acknowledged that working together with researchers from other domains (such as bioinformatics, genetics, etc.) presents a challenge. Genomics research requires a high degree of collaboration and all interviewees recognized that it takes a certain skill set as well as time to bridge the gap between scientific domains and to coordinate genomic discoveries successfully. Yet at the same time, interviewees spoke of the need for people with a high-level overview and a certain macro-level approach. One interviewee noted that *"we don't train people and not just in Canada but worldwide to think systematically" (SOF1002).* The same researcher also pointed out that the challenge of interdisciplinary stretches from beginning to end, starting with the grant application process for an applied clinical problem.

> *"Principle investigators are going to write grants, which are not hypothesis driven grants. I've got difficulties, to do drug discovery, which very few investigators are truly in an academic setting equipped to do that. Couple people I know, but very few"(SOF1002).*

Even if projects get funded, the challenges of bringing diverse actors together remain. In particular, some interviewees (with many years of experience serving as principle investigators) describe how they have to close that gap between their respective domain and other researchers and to "speak the language of biology" for instance.

> *"So one of the challenges is getting that um line of communication open, everyone we're all such busy people […] so that those are the kind of issues that impact me the most, and impact my colleagues the most" (SOF1003).*

Because of the convergence of disciplines, committing time to such collaborations is not enough; it also requires the ability to understand a foreign field and research. The same principle investigator (PI) expands on her previous statements and explains:

> *"The other challenge is how do you talk to someone so far out of your field? So [that particular scientist] is unique in having that skill, that she/he can speak really logically and just get the concepts and understand what I'm saying, given my non-mathematical way, I speak about, at the biological idea and she/he can translate that into a mathematical idea" (SOF1003).*

Naturally, PIs spend considerable amounts of time and effort on establishing, as well as maintaining successful collaborations. At some point, this 'networking' even becomes the core of their work.*"Most of my work is involved in setting up collaborations with investigators who are having problems analyzing their data"(SOF1009).* This participant acknowledged that researchers need his help to carry out the data analysis. It appears that his skill set – required to navigate this interdisciplinary field – is in high demand.

Indeed, it appears that it is crucial for the interviewees to work with people who are trained and qualified to work in this highly interdisciplinary area.

> *"But most people don't have the language or the understanding to be able to do [work across disciplines], and so how do you foster that? For me it seems to be more of an innate skill and I'm lucky to hit on people who can do that, like she's so good at that. But ah it's a challenge right?"(SOF 1003).*

The same participant explains how successful collaborations are so important and yet seem to be so rare:

> *"[It is] just the lack of manpower. We just don't have enough bioinformaticians who can talk to biologists because really, since everything has so much, so much of this has to be, um, ah, you need a lot of creativity and imagination in how you solve these challenges right, since this is not all out of the box, and if you can't really talk to each other at a really high level you can't share new ideas" (SOF1003).*

This particular participant is concerned about a shortage of well-equipped collaborators and why they are so important.

> *"Health researchers are crying out for more people who can do computational work, and maybe they won't be first author on a paper, last author, but it's so important and you need someone who's really high level, you can't just put a student or even a post doc in, right? They need a lot of oversight and, so people say "well you can solve lack of manpower by creating training programs, but that's never, or do collaborative training programs but I just find you're not going to get anywhere unless you have a really smart PI behind it" (SOF1003).*

Beyond the ability of understanding a foreign area of scientific inquiry, the diversity of "trades" within the genomic community presents many other challenges, such as conflicting time schedules, research practices or teaching duties. As PIs have to collaborate across disciplines, weekly schedules get stretched very thin and finding time to work together on such collaborative projects becomes very small.

> *"So it is pretty tough, first you have to find a collaborator, so where do you find them, there's um, a lot of people in Canada for good computational scientists, how many of them are interested in biology, or know it and they're in a different funding stream, in a different department, with different academic needs and criteria, they've got heavy teaching loads, they've very little time for research, yeah. They're not rewarded necessarily or appreciated for any work they do with health research, um, they're not rewarded financially, they might be middle of the papers, and that might not be recognized by anybody, despite all the work they put into it, so the career structures are quite different the way grants are written are very different so it's very hard for them to – they might need to develop whole new methods that are just amazing and innovative to solve your problem, and so that's something-- innovative methods development needs to be funded, yet it's very difficult to get that funded" (SOF1003).*

The collaboration between scientists and clinicians can be difficult to manage. While clinicians' time is tied up in health care delivery, collaborators often struggle to work within the time constraints of such interdisciplinary teams. *"In every single collaboration I've been on it's [the] physician['s] time [that presents a barrier]"(SOF1009).*With short time to invest, clinicians can have different expectations of the projects and the exchange of de-identified data takes up too much time.

> *"And because we're working so closely with the clinicians and because the only thing we're working on is de-identified data, we have to give the clinicians a matching sheet that says, you know we de-identified the data, here's how you match our…we de-identified with your identity that you can match up. But the clinicians have no time at all for research" (SOF1009).*

> *Naturally, the handover of de-identified genomic data between collaborators is complicated and time is short. "So it seems like no matter what simple thing you're trying to do there's never a one size fits all answer to any kind of data collection or analysis, it seems like, and if you go to your bioinformatics collaborator and it's never any easy answer right? […]we're all such busy people" (SOF1003).*

Large scale collaborations seem to attract more funding but come with a set of challenges on their own. PIs acknowledge that genomic research collaborations grow in size and in diversity, which can actually hinder the discovery progress. One interviewee made the argument that the quantity of funding should be increased, as well as the overall number of funded projects. It is very interesting to note that competing group members might hinder the process.

> *"I think you have to have a critical mass to the size of a research group, so that you get those [discovery] synergies, but if the group size gets too big, and too bureaucratic, what happens is actually quite counterproductive. I've observed being in some of these big labs, myself, but as the funding is more and more established, and there's more money, and the groups get bigger, the competition is actually not just with other groups. This is within the group itself, which is very counterproductive as well" (SOF1002).*

Investing in many smaller groups would potentially solve that problem. The same interviewee offers a solution:

> *"I think it's better to fund more investigators with less money and then have natural collaborations evolve that are dynamic amongst those investigators, as is by necessity, than to try to force groups together in unnatural, unholy alliances, for the sake of getting large amounts of money which I don't think that they spend very effectively" (SOF1002).*

In the midst of competition for research grants and publications, different job interests can collide. Hypothesis driven research and the need to establish oneself can put different time pressures on the collaborators.

> *"That's a problem and who does the work, and even the time it takes, so we're out of sync, biologists, I have the data, I need to identify the things to follow up in the lab right away, so I need the analysis to be done instantly [chuckles] and if they need to develop a whole new algorithm and put someone on it and then go and teach a heavy teaching load, minimum, it could take way longer than that. And I can't wait a year or two, for my data to be analyzed, before I can start working on it, because I have to publish much sooner than that"(SOF1003).*

Similarly, bioinformaticians are very well aware of the complex interplay of various disciplines. They have to negotiate with PIs about the increasingly complex analysis which does not always yield the expected results.

> *"At some point, you have to say we've run the analysis through, and yes we could get to even more esoteric levels, uh but they're probably not going to give you anything that's, that's useful. And so you have to have some place to stop. And so that becomes a problem, because the clinicians don't accept that [as an answer]" (SOF1009).*

*Cost & Funding*

Scientific research is dependent on appropriate funding and almost all interviewees mentioned that funding can be challenging, especially when trying to work in a fast-paced research field which requires a balance between basic research and its application.

*Intellectual Property*

One way funding can present a challenge is in the form of a trade-off. As PIs are applying for funding, they have to either rally for financial support from various institutions or consider trading the potential commercial application of a genomic discovery for funding. When it comes to commercializing genomic discoveries, PIs may have to decide, for instance, if they want to trade intellectual property (IP) of a genomic discovery for appropriate funding from other institutions.

> *"The starting point is anything I invent is my own. Now I can enlist [a university] to help me develop it, in which case we come up with a joint ownership agreement. But it doesn't have to happen. If I get the money I'll pay [and] going to do it all by myself and I'm going to go off and commercialize it [on my own]" (SOF1001).*

> *"I mean there's all these options [to assign IP] and you have to come up with some plan. Generally it's a bit of a negotiation. [The funders] come and say, well what we would like is for you to give us all the IP and then we'll develop it. We say, well no. And uh so then we come with some arrangement" (SOF1001).*

Although the negotiation of IP for commercial purposes was mentioned in interviews, it was generally not perceived as a problem, though one stakeholder did point out that often the negotiation of IP rights occurred alongside project work and was not concluded until nearly the end of a project. While commercialization of IP rights did not appear to act as a constraint to the movement of genomics science from the bench to the bedside, several stakeholders identified challenges associated with securing labour to help them make sense of their genomic data.

*Pay scales and Tool Maintenance*

Data analysis becomes increasingly complex and so is sense making. PIs recognize that there is a shortage of money to pay people who can make sense of their data, and who are capable of moving genomic research forward.

> *"It's not shortage of people. It's the money they pay the people [laughs]. I haven't had any funding from CIHR for about seven years now, even though I have applied many times" (SOF1002).*

Some interviewees mentioned that is it hard to attract funding for software tool maintenance, when it is paramount for commercialization and clinical research to work with previously benchmarked tools and applications. In this sense, continuously working with the same pipelines or software essentially establishes a status-quo for quality control and benchmarks. However, it is easier to gain access to funds for re-creating pipelines or platforms then it is to maintain existing platforms and pipelines. This, in turn, leads to challenges associated with benchmarking findings and replicating results.

> *"[Scientists are] actually doing molecular genetics and, and trying to come up with targets and then, and then develop uh you know show that they're worth*

> *something. And [scientists] doing just what I was saying that they, they sort of did what they could do based on their own capabilities using computer analysis and compounds and came up with some data that looks interesting. And ah now they're trying to get a bigger grant to do it better, but they say, well what can we do? I said, well you know you have to validate what you've done. You know you may have this compound that does something but who's to say it doesn't, just because it fits your theory—and it doesn't even fit that well. I mean the, the results are kind of marginal" (SOF1001).*

The longevity of funding is a huge factor as well. As maintenance of tools and data curation take time, researchers hope for longer lasting projects to improve already existing platforms or advance benchmarked pipelines:

> *"How you can get leading edge research and predict work that if you're lucky to get funded six months to a year down the road, that you're going to be doing for a plan for up to five years. How can that ever be, how does research work like that? It's just not possible" (SOF1002).*

Interviewees who have experience in the field of commercialization of genomic technologies recognized that financing genomic research is increasingly more competitive. Well-connected, highly experienced interviewees seem to stress the challenge of funding even more. Sometimes, their connection to the commercial sector can create barriers for PIs.

> *"If you go [for funding] as an academic that has an industrial connection, it's counted against you. I mean, when you have resources that are scarce and ultimately it doesn't matter what the granting agency wants to do in its mandate. If you have a panel of peers that are from academia, and they're voting and they see how rough it is out there, they're going to have a little more sympathy for the investigator whose career depends on them getting funded, than someone who has 'well he has an industrial connection he'll be okay'" (SOF1002).*

On the other hand, when commercial investors help fund genomic discoveries, revenue pressures limit the scope of research and can leave the PIs with a dilemma.

> *"The problem we face is that for most companies in the bio-tech industry, you get a lot of investors early on if you're fortunate, and a tremendous amount of capital is raised, and then it's basically deployed to try to usually find a drug, and develop a drug. There's not much investor support for diagnostics. There's certainly not much investor support for sort of like a basic research company. They want to make sure there's a practical outcome that will generate revenue" (SOF1002).*

> *"Our obstacle is basically running a business that um, can meet its you know monthly payroll right? So what it requires ultimately is that I have to put more money into the company personally, enough to keep the company running. And then, I have to be very philosophical about it. I have goals as an academic, and I have goals as basically someone who's got an obligation to my shareholders, they're going to get a return. So. What I try to do is strike a balance between those, and I do believe that if I continue in this direction, ultimately it might take another five to ten years we'll have something of real value here, that will be an*

> *asset to another company one day, if it's purchased or, maybe there's other opportunities for it to be fragmented into pieces that are useful for different organizations" (SOF1002).*

Funding directly influences the properties and materials of genomic research. Financial strains can present a barrier for discovery processes, especially when PIs chose to work with 'cheaper' compounds or biomarkers that 'fit the bill'.

> *"A lot of sort of genomic discoveries are languishing and not going where they could go is because of sort of naïveté. The people involved, sort of see it as their little domain and they work on it in a way in their little place with a few people and solve the issues of getting the molecules by sort of cheap and dirty ways. They basically go and buy them somewhere. It's probably really a poor substitute but it's what they can get" (SOF1001).*

However, even attracting sufficient funding is not always enough. When it comes to moving genomic discoveries from bench to bedside, sense making is crucial and has to be funded as well. While costs for genomic technologies have decreased, the state of technology may have outpaced our knowledge.

> *"The technology is increasingly getting cheaper, but I think you would be well aware the data analysis part of it costs are not coming down, um, there's just too much unknowns. In terms of what those biomarkers mean individually, and also in combination. Our state of knowledge is just so far behind in this respect" (SOF1002).*

*Federal or Provincial Agendas*

As the main funding bodies, federal and provincial agendas directly influence the longevity and focus of genomic research. PI's have indicated that they would like to see the provincial government getting more involved.

> *"Now the provincial government doesn't make it easy of course. They tend to be a brick wall of sort of uh ignorance and sometimes uh although it's not as bad as it might seem, but I think there needs to be more worry about bringing provincial governments into this game somehow. So they can see a role there, because they're payers and they are people hopefully worried about outcomes [of genomic discoveries and Personalized Medicine]. I mean it's not just paying. The whole thing is to have good health come out of this" (SOF1001).*

> *"Now certainly the Personalized Medicine initiative is very much trying to bring the provincial government into this business. And the provincial government is […] funding many things" (SOF1001).*

This also entails an overarching provincial or federal strategy for moving genomic technologies forward. One PI is unsatisfied with the quality of Canadian –omics technologies.

> *"We were using microarrays from [a Canadian research training centre] but they're not good enough to give us the quality of what we're looking for. The turnaround, the cost of printing, the quality of the chips, […] and you know there are limitation to the resources themselves, and the funding. […] Now we go to the*

> *States […] where it's cheaper, and the quality of the printing's higher. But you know we're a company, and that's really important to us" (SOF1002).*

One interviewee mentioned for instance how 'niche' genetic disorders that could potentially only affect a small group of Canadians attract less research funding due to the potential lack of payout or return, and cited a mis-match between federal testing requirements and the economics of bringing some drugs to market.

> *"The regulatory environment and the general public's views of life are going to have to change a bit. […] Right now you have to come with any kind of cardiovascular drug and they'll say we need four years of safety in twenty thousand people. You know, go away and come back when you've got that. And of course you know that ends up costing a billion dollars. You can't spend a billion dollars for a fifty million dollar drug. And so there you have scenarios like that where a new directed therapy for pain if it were a relatively small population would be tough to develop" (SOF1001).*

On a macro level, interviewees acknowledged that provincial and federal agendas are not always aligned and that a unified strategy would help researchers to attract funding for more diverse –omic disciplines. This being said, some interviewees suggest that other disciplines and –omic branches (proteomics, metaboleomics) have great promise to advance PM as well but tend to be left out of federal or provincial funding programs.

> *"The other problem that we're seeing is that, in academia, there's a really strong emphasis towards hypothesis driven research, and although these amazing technologies for different types of -omic analyses have emerged, there actually hasn't really been, um, much support at the Canadian Institutes for Health Research level for example, or the NSERC level for systems biology, there's been some token um funding, controlled by a small group of people that funds systems biology research in certain directions" (SOF1002).*

> *"90 percent plus of that money always goes down the genomics route. So if you look at that past history, and you wanna be successful, it's pretty obvious what you should do. Don't do proteomics. Proteomics is also a hell of a lot tougher to do. I think it's going to be more fruitful. It is much more difficult to do" (SOF1002).*

*Privacy, Consent and Ethics*

The legal landscape surrounding genomic research and PM is slowly catching up to socio-technical questions. Nearly all interviewees mention that concerns around the privacy of research participants and patients present a challenge in their research. And while privacy and a patient's right to privacy, ethical use of data and indeed ethical issues surrounding genomics research in general and gaining consent of research subjects are arguably separate issues, for most interview participants, they were inextricably linked.

In terms of clinical application and discoveries, de-identified data presents its own challenges to researchers. Severed links between genomic profiles and clinical charts limit the discovery process and potential for new genomic findings. Maintaining links between patient data and genomic data requires what many interviewees seemed to suggest were insurmountable consent and ethical clearance processes. Interviewees spent a lot of time navigating various ethical

guidelines, and overcoming challenges in securing ethics clearance from institutional review boards for their data and application. In some cases, researchers indicated that they might avoid working with more robust de-identified data sets because the task of obtaining ethics clearance is deemed *"too daunting"*(SOF1011). Some interviewees mention that de-identified data could be better handled in hands of centralized custodians / stewards to ease access for researchers and clinicians. This would also improve issues around consent and information management of genomic data. While interviewees were not unconcerned with ethical issues arising in relation to genomics and genomic data, they spoke more frequently about issues associated with securing ethical clearance to conduct their research, which for most were closely tied to issues of privacy and challenges associated with consenting research subjects/ patients.

*Ethics*

The ethical issues surrounding genomic discoveries are extensive, and were often mentioned by our interview respondents.

> *"There's a lot of you know children whose parents aren't exactly at least one of them who they think it is. Usually their mother is okay. But the father may not be the real father. And, when you start having a genomic information reveals this to be the case, and it's of knowledge to the father, until then, then there's all kinds of ethical issues there that need to be dealt with. And is really a genie in a bottle" (SOF002).*

While conducting interviews for this project, we learned of a court case concerning the re-use of blood spots collected by a local hospital. Consent had been securing for the collection of the data, however, the plaintiff alleged that she would not have consented to the collection of blood samples from her children if it had been disclosed to her that the blood samples would be stored after being used for initial testing. The plaintiffs argued therefore there was no informed consent to the collection of the blood samples, and no consent to their storage. The case also raised issues about whether or not the collected blood samples amount to a legally unauthorized fully functional DNA database for every child (and his or her parents), that may be accessed by as yet unknown persons and/ or agencies, for reasons other than those provided when the samples were collected (see http://www.courts.gov.bc.ca/jdb-txt/SC/11/06/2011BCSC0628.htm and http://www.courts.gov.bc.ca/jdb-txt/CA/12/04/2012BCCA0491co1.htm for additional information). The case (which remains in the courts having been cleared for hearing in the Supreme Court of Appeal) points to public perceptions of genomic data, and the need for discussion and debate about ethical issues associated with genomic data with the public.

Another stakeholder suggested that while there are ethicist**s** in the research community, that genomics research would benefit if ethicists had more practical experience:

> *"And you know and we convened various use case based uh, uh biomedical ethics workshops and in the early days, in the early 2000's one could understand why it was uneven and so on. But the ethi-…in my opinion — and I know them all and love them all — but the ethicists have failed uh to get their act together and they're too interested in being academic about it. And not i-, they're not interested eno-, enough to be practical. And uh and you know 97% of the public in Canada would consent to have any of their information used for research if asked provided it was asked by a reputable researcher and was, was you know secure" (SOF1011).*

This stakeholder viewed ethicists as an impediment, rather than facilitator of work:

> *"Oh. I mean there…I mean let's face it uh ethicists are, are a hu-, are lightyears behind where they need to be. Uh you know they're, they're not much of a solution, they're mostly a problem. Uh and, and the reason is that they, they spend way too much time um deliberating and not enough time acting. And so um like I say to many people um the eth-, the field of, of health ethics or biomedical ethics is so far behind where the actual work is today that, that they'll never catch up. Um so in that sense they're somewhat of an impediment to ah…because you know Michael Smith tried to fund a harmonization process for ethics in this, ethics review in this community, and do you think that the ethics boards could agree on it?" (SOF1011).*

Ethical issues also arose in relation to media coverage of genomic discoveries, where early—and arguably inappropriate reporting of what some viewed as scientifically questionable results was seen by some as unethical.

Consent

The lack of standardized consent guidelines was identified as a constraint by many. Unified and centralized, general consent forms would allow researchers to conduct, for instance, secondary analysis of pre-existing samples without re-consenting. It was suggested that reducing the complexity and variety of consent forms would rapidly advance the discovery process and application of lab based findings. However, many issues and challenges were raised by researchers concerning challenges associated with gaining ethical clearance and consent, particularly in the context of projects in which researchers had a desire to link bio-samples with other forms of data.

> *"Again that's not technically difficult, but it does raise the question of, of consent. Was that blood obtained with the knowledge that the, the HIV and HepC screening would be done but what about all these other studies that one wouldn't know about? So that's where a GE3LS type project would be really helpful in being able to address that and we've talked to Mike Burgess about it and there is interest it, but we need really to get a whole organization…" (SOF1008).*

This respondent described how a series of questions he wanted to pursue led him into an ethics and consent quagmire:

> *"And going forward, you could then see, you could build a case for how does a set of biologic factors uh interacting with the environment predict early brain development, which is the thing I'm interested in. So you can see, sorta see that out of a simple set of questions, and excitement about a biorepository population level data, I began to see some challenges like who owns the blood, who gave consent for it, what are the legal and ethical parameters in which that blood can be used. At first, I was under the understanding that it actually, because it's going to be thrown away, it falls under Tri-council uh um concept of its available for research, and therefore the question of ownership no longer applies, etc. etc. I'm giving you one argument…" (SOF1008).*

Another respondent commented about the blood spot case outlined above. The respondent explained that

> *"the parents [were] saying I didn't consent to this use. And so then you implement consent and consent as you know is heavily restrictive, when we're thinking about we don't know what's going to happen in the future, once you, lay consents on, you've got a set of rules that you've gotta operate by and manage it, etcetera etcetera. And so that's also in the case of um [inaudible 29:45] you know, everything but the kitchen sink, so it's [chuckles] which to me seems really ill advised, I think we really do need to move, we need to be careful about the issues of data, we need to be, you know, really cognizant of the public's response and, ah, the harm to the individual but, I don't think we're getting at it by consent" (SOF1005).*

Bio-repositories posed additional challenges as well:

> *"The technique of, of linking these things is not that significant, but the issue is really what do you, how do, how do you confront issues like consent, legal issues around ownership of the biorepository data. And then how do you deal with a biorepository source that is almost infinite in its capacity to tell you a story about human biology, behaviour?" (SOF1008).*

However, for many scientists, issues of ethics seemed to be synonymous with issues associated with securing informed consent of those they hoped to include as subjects in their research, and to gain access to previously collected data or to link data for scientific purposes.

During the time period we were conducting interviews, a conference held in Vancouver about data was attended by several health researchers. Many subsequently cited poll results indicating the public's willingness to consent to have information about them used for research purposes, and suggested there was a significant disconnect between public perceptions about data and privacy and the currently regulatory environment and ethical approval processes.

> *"And uh and you know 97% of the public in Canada would consent to have any of their information used for research if asked provided it was asked by a reputable researcher and was, was you know secure" (SOF1011).*

Another interview participant reflected on a conference presentation they had heard abroad, which suggested that the more knowledgeable members of the public became about research use of data, the more willing they were to share their data:

> *"Yeah it was very exciting, and ah there's some interesting talks that I went to around consent incidentally and how…ah as the participant got more knowledgeable about the consent and understanding the various issues, they became more permissive in the uses of the data, and that was kind of the monitoring some of this movement towards what um, this deliberate democracies Mike Burgess was talking about and what not, is that as you move towards this model where people have really healthy debate and really hash through the issues, are thinking through it, that it's um, that you it's almost counterintuitive, that you're getting more permissiveness, in that, these people have really carefully thought through their decisions. And I think the public would appreciate that knowledge that there is neutral third party that is going to kind of be tasked with that, because this is too complicated for me to want to wade into or figure out, etcetera, so" (SOF1005).*

Consent is sometimes seen as a constraint to science.

> *"Um the ah Tri-council's kind of ah move or suggestion of a move towards um duty to re-contact uh is massively disturbing on several levels. Um it's an ill-defined um threat at this point, um which could um make us uh you know face a very difficult problem where we have IRBs which say we would never re-contact you and consents where patients have signed something that the undersignee will not be re-contacted and then a national body telling us that re-contacting people…" (SOF1012).*

One interview participant suggested that because to data few genomic projects had attempted to make use of linked data, that issues related to ethics and privacy in this area were still in their infancy, and that as scientists increasingly sought to work with linked data, that issues related to privacy and ethics would increase.

> *"Another random thought around genomic data, is that, I'm not sure to what extent privacy has really trickled into the work or discussion or debate. Um, because in some ways that it's not ah they've been and I'm making some [inaudible 33:04] beliefs so do correct me if I'm wrong but there are more often standalone projects and analysis within the data set. and with any of that, there's not once you're thinking of moving to the linked data environment, ah, one of my background concerns that I haven't had verified at all, is that this whole new swap of requirements on the privacy front, maybe a bit of a barrier or a challenge, so" (SOF1005).*

The need to gain consent for the use of data—particularly data collected previously (either in the normal course of care or for a previous study)—was generally seen as something which could improve the quality of findings and hasten the movement of genomic results from bench to bedside, but also as something slowing down the progress of science.

> *"And then used for, it's found to have alternate uses, so new worm blood spots would fall under that blood test labs data, um, x-ray data, you know that sort of thing. Um but then you, then the other side, ah that is very challenging is as you pointed out the um, where researchers collect data for a research project. And then, that's where again it sort of triggers this need for consent and then the, the um need for consent to be adequately vague to allow for these incredibly valuable future uses, I think is something that the research community really needs to um, come together to work out with privacy experts etcetera" (SOF1005).*

Data linkages also presented practical challenges for genomic scientists in terms of consent:

> *"They sign off, um, okay, so the I… there's two streams here. One is thinking of biological specimens in the context of a one specific project where that PI collected the data themselves. So then that is their consent form for that defined research purpose. And in that case, we have a practice of whenever you're going to field and doing primary data collection that the data stewards, in advance when you're going to field, review that consent form to ensure that it meets their needs. So it's not a requirement but it's a good practice on the part of researchers so that it meets those anticipated data needs. We can kind of get that signed off, and then they go to field, the data comes back, and so then the consent is not*

> *viewed as an issue in some cases that pre-review hasn't happened and the data stewards have said we don't feel this consent allows us to link the data [inaudible 7:00] responsible to the data that's been collected. And so there's a bit of ah out of synch there on the consent forms. Um, but so your question was five different data stewards and, let's say five different data stewards plus primary collected ah genomics data" (SOF1005).*

Operationalizing consent has practical implications:

> *"So again, if you're wanting to link it to those five other data sources, that Popdata facilitates, um, access to, we ah, those data stewards would look at the consent and see if it's still valid or that it's not still valid I should say that the consent, that the language in there around the ah, the uses of the data, and the planned linkages of data, ah, cover the proposed requests" (SOF1005).*

One researcher acknowledged the time required to improve ethical clearance processes and suggested that investment in centralized data stewardship services could lead to improvements in availability of genomic data in the future:

> *"I think there is some, would be some value in setting up a form of ah data stewardship committee or you know information privacy and stewardship committee or something like that, that has representatives from you know, definitely the public, researchers, ah, I would say privacy experts, ethicists who'd want, and that can be delegated responsibility for reviewing applications where there is this fuzzy stewardship component. 'Cause as a PI, on a project, you may not want to take that on for your remaining years but it would be nice for us to set up the structures and processes that when PIs are launching on a major data collection initiative, that we can say 'can you add this to your consent and that will open up this huge door for future research purposes and you don't have to worry about anything else' [laughs]" (SOF1005).*

Other issues related to informed consent related to challenges associated with scaling consent, under a variety of circumstances, which included use of data collected for a pilot study in a larger study; participating in national or international collaborations (where several at times competing consent processes exist), and related issues.

> *"…the critical issues are. The real question is…so doing the feasibility or pilot studies were relatively easy because we just go to the administrators of all those data biorepositories and they were able to anonymously give us…they were able to give us anonymized samples. Fifteen here. Ten here. Whatever. That was fine. But if we wanted to go to the next level, which is, you know, four hundred, ten thousand, fifteen thousand case cohort type studies, then we, then we had to go to this bigger question about who owns the re-…the sample, under what conditions was consent given, and uh, and, and if these are samples that are going to be thrown away, what kind of ethical framework do they fit into? So there's questions of legal and ethical barriers that, that need to be sorted out. And that's sort of where the conversations ended. And you know if we could, if we could ahhh develop um, you know and it became… Yep. If we could harness enough people in the province who are interested in biorepository samples, we could address the question of how the samples are collected, how they're stored, consent issues, and*

> *uh and, and research uh… and developing an infrastructure for research capacity. If we can do that we're..." (SOF1008).*

While some stakeholders suggested that some issues associated with gaining informed consent could be accommodated with the use of language on consent forms that would be more permissive of uses of data that had not been anticipated when initial consent forms were developed, others suggested that this strategy would not be adequate.

> *"like if you think of a consent, am I going to sign out a consent that lists every single use?" (SOF1005).*

> *"So the issue around consent there is that it's typically you know like in this project that your consent is for a specific purpose for a specific research question. And with biological specimen which I sort of lump together with genomics is that there it's quite often for a specific research question. But there are increasingly cases like BC Generations Project, or um, or or even sort of Biolibrary or what not where it's being collected but there's recognition that there are likely to be subsequent purposes that are as yet undefined so when you mix that, reality and the importance of that and the value in that with what our concurrent consent approach and framework is, ah, it creates struggle in that, that that sort of if you think of bringing in say ah the biological specimens that are being used in the BC Generations Project, as an example, that um, that we have to work out with our other data stewards, aspects around that consent, and is it considered legitimate and so I'm feeling like you're not I'm not clear with the message there" (SOF1005).*

Indeed, this is among the issues at the heart of the court case concerning blood spots.

When describing how they might handle being told existing data – arguably administrative data which could be anonymized – could not be used for research, one respondent's comment about how they would likely handle the situation shed light on the complexities of working with data for which consent had been obtained in the distant past.

> *"Or request to re-consent, or something like that. But we get into even murkier areas when you take say the no warm blood spots, and ah go back to that because is that then secondary use? Is it sort of administrative data that has secondary uses which is allowed to be used without consent? Ah, under privacy legislation? Um, or is it something that should've had consent?" (SOF1005).*

Challenges of dealing with consent to use data can also be magnified if several PIs were involved in the initial collection of data, and addressing these issues also has implications for data stewardship. As one respondent explained:

> *"So in one ah case, it's I've mentioned in various fuzzy ways about the consent, and that ties also to stewardship - who is, who does take the decisions on ah some of these specimens, when they're not developed by an organization they might be developed by a series of PIs so getting clarity around the stewardship of the data, and then in general, um… a third piece for us is actually where does the analysis take place? And in typically, um, I would anticipate that it would need to not be on our secure research environment because of capacity requirements, the*

> *computing capacity, um… I mean depending I think there's a whole series um… I'm jumping around with my thoughts here but…"(SOF1005).*

The above comment hints at the challenges we may just be on the brink of facing, related to each research ethics board wanting data to be securely stored in their facility. The practicality of this situation breaks down when several jurisdictions each own a portion of data to be combined in a single analysis. While BC's Popdata facility addresses this challenge for several holders of large datasets, it is equally true that there are numerous data sets which may be the subject of genomic linkage in the future, which do not fall within the BC Popdata mandate.

Another respondent who was involved in a worldwide tumor research initiative described how consent issues in general and the need to insure that patients with a tumor had consented to have their tissue samples included in an international research project had precluded the use of existing tissue samples.

> *"One of the things that people have to consent to is that their data is going to be available on uncontrolled access to scientists worldwide to do biomedical research and so, ah, that is a sort of, that reason alone is a reason why a lot of preexisting samples could not be used and so everybody has to be re- consenting – you to be participating in this project because ah very few sort of older consent forms would say ah, you're free to put my genome on a website and share with the rest of the scientific community" (SOF1006).*

Finally, one interviewee suggested that the scientific bar for research consent was becoming so high that few could meet it, and increasingly scientists were finding ways to gain consent for use of genomic data outside of traditional research settings.

> *"So ah… so the main one is that I'm sort of concerned about right now is this cloud computing one, the other main one which I've already mentioned is sort of the barrier – making the barrier to controlled access data, apparently so high that very few people bother by doing it. And there are a few initiatives in the US for example, like George Church's ah PGP, their Personal Genome Project, we're aware of what he's doing so it's not a Cancer and Genome Project, it's a sort of healthy normal individual genome project, but what he gets is, the people whose genome are consented to, the consent form that he gets people to sign off on, has been having no restriction whatsoever on the data" (SOF1006).*

Managing consent processes—particularly when multiple data sets are involved which researchers would like to link—presents several challenges.

> *"But it does get to a bit of rounding back to my initial point about consent is that once you've got all these five different data stewards, in that example we had, they all start whether you like it or not, kind of paying attention to the other's business. So it's, even though they're really only adjudicating based on whether they're willing to let their data, um, be used, they're also wanting to have assurances that the appropriate legal structures are in place for the other data sources, for other public bodies, let's say Ministry of Health, if there's Education and there's Children and Family generally speaking not such an issue but once you're dealing with aspects that are fuzzy, like the genomics data would be, there would be extra scrutiny there, and sort of, because the release of data for research*

> *purposes is under the discretion of the data stewards, not a requirement, that if there's any level of discomfort about the types of data they're linking to, or the, or whether there's proper authorities in place and their judgment of the data that they're linking to, um, it either would slow it down or make it not receive. So to your question about these, extra ah curricular data set construction that happens in just about any researcher in the world, um,… where things are currently, I think unless we're able to trace back a chain of consent like, if it is a chart reviews, or some of his chart reviews and a perinatal database, we're able to engage those data stewards and say, we've been using this data, are you willing to sign off on, it says per use, that's how we would have to structure it, but we sort of need to find the source of accountability for every piece of data that's getting linked in" (SOF1005).*

To address this type of issue, this respondent suggested setting up a data stewardship committee.

> *"that I think there is some, would be some value in setting up a form of ah data stewardship committee or you know information privacy and stewardship committee or something like that, that has representatives from you know, definitely the public, researchers, ah, I would say privacy experts, ethicists who'd want, and that can be delegated responsibility for reviewing applications where there is this fuzzy stewardship component. 'Cause as a PI, on a project, you may not want to take that on for your remaining years but it would be nice for us to set up the structures and processes that when PIs are launching on a major data collection initiative, that we can say " can you add this to your consent and that will open up this huge door for future research purposes and you don't have to worry about anything else [laughs]" (SOF1005).*

The potential benefits of educating the public about research use of data was also identified as a means of increasing the permissiveness of data use.

> *"there's some interesting talks that I went to around consent incidentally and how…ah as the participant got more knowledgeable about the consent and understanding the various issues, they became more permissive in the uses of the data, and that was kind of the monitoring some of this movement towards what um, this deliberate democracies Mike Burgess was talking about and what not, is that as you move towards this model where people have really healthy debate and really hash through the issues, are thinking through it, that it's um, that you it's almost counterintuitive, that you're getting more permissiveness, in that, these people have really carefully thought through their decisions. And I think the public would appreciate that knowledge that there is neutral third party that is going to kind of be tasked with that, because this is too complicated for me to want to wade into or figure out, etcetera, so. Another random thought around genomic data, is that, I'm not sure to what extent privacy has really trickled into the work or discussion or debate. Um, because in some ways that it's not ah they've been and I'm making some [inaudible 33:04] beliefs so do correct me if I'm wrong but there are more often standalone projects and analysis within the data set and with any of that, there's not once you're thinking of moving to the linked data environment, ah, one of my background concerns that I haven't had*

> *verified at all, is that this whole new swap of requirements on the privacy front, maybe a bit of a barrier or a challenge, so" (SOF1005).*

This respondent also highlighted some of the challenges researchers and data stewards face when working with data from multiple sources, and suggested that both groups would benefit from more guidance.

> *"And I think that's an example like when I, not so directly but genomics data is an area that we very much know that it's coming, or it's here but we need to do some work to make it ah more readily, seamlessly integratable with population data BC, and there's many fronts that that covers, but one of which is, this idea of having this idea of package of information or a structure for, that genomics research community is able to consider or to have, to add to their consents or to identify that this is how you these are some of the downstream requirements you would have regarding stewardship, and these are ways you might approach, identifying a steward, so. This does get like, and guide me away if this is getting too focused on one area but, in cases where you have the multiple funding sources which I think most are, that stewardship is really, um, surprisingly left to assumptions about who can use that data, that often if the government is a co-funder, they believe they own the data, and because they say it's been done under contract, and [laughs] so" (SOF1005).*

While there is a data stewards working group in BC that has reached agreement on several issues such as a common application form for data access, a research data access framework and common approaches to what is required on consent forms, much work remains to be done.

Finally, one respondent suggested that existing consent procedures simply went too far, in that the research community has been focused on consent processes, while the general public is more concerned about whether or not appropriate processes are in place to safeguard data and the use of data.

> *"Um… I think that the public is supportive of uses of data but um… you use the word consent and I think that's going a bit too far, I think that they're trusting that there is some process in place that has appropriate checks and balances, so it's not anybody to use the data for any purposes, it's for you know identified individuals to use it for defined purposes" (SOF1005).*

*Politics of Data Ownership*

Not all genomic data are created or shared equally. Ethical concerns and a myriad of formats make genomic data ownership very complex. Hence, each institution and even each PI have their own intellectual property policies and managing large-scale collaborations can be difficult; various teams may claim ownership of data or handle data in formats which are not compatible across labs. Only a few interviewees mentioned that they had no problem getting the data they needed. However, since the interviewees were senior researchers, we do not know about any potential problems related to handling the data on a smaller scale. In the case of individual data stewards, sharing practices are dictated by trusted individuals, whose legacy of stewardship can be inconsistent. Interviewees identified fears and worries on the side of custodians and provincial authorities about what researchers might do with the data. On a larger scale, access to large cohort data brings up various questions about ownership of samples, ethical guidelines and

regulation. On the scale of individual labs, the tough interplay between publishing results and potentially protecting genomic signatures threatens to slow the advance of clinical applications.

*Culture*

The road to PM will likely require a combination of different types of research and practice. Similar to the concept of interdisciplinarity, different practices and agendas can slow down the discovery and application of genomic research. Pharmaceutical companies for instance will seek to develop mass-suitable test and treatments, whereas hypothesis-driven, academic genetic research might look at outliers, or so called 'black sheep'. Interviewees have to maneuver between contradictory expectations, which can be time consuming and requires people with a certain skill set and training. Moreover, academic culture in itself presents challenges: junior researchers who are looking to publish and raise their profile might work continuously with new tools, whereas interviewees focused on commercialization of genomic tools may prefer to work with established tools. Needless to say that each focus means that the labs (including the graduate students) have to accommodate these potentially conflicting needs. Holding on to data for publication purposes or other uses can delay the application of tools. Related to this point, the question of authorship and project lead creates a certain barrier. It appears that bioinformaticians are a coveted group of people who are infrequently given a central authorship or project lead status, which can potentially constrain careers.

*Sense Making*

Despite the advancement of sequencing tools, making sense of genomic discoveries is still challenging. Determining the clinical relevance of genetic discoveries is very complex and time consuming. Because of the interdiscplinarity of the field and the complexity of the data, determining the clinical value of genetic research takes a very specific skill set. For instance, clinician researchers or geneticists rely on the bioinformatician to carry out the analysis and interpret the data. One consequence of this is that the separation of tasks somewhat obfuscates methodology as a 'black box' for geneticists (SOF 1009) but they trust the methods and the results. Interviewees mentioned that large data sets with various puzzle pieces (phenotype and genotype) would provide the best foundation for clinical discoveries, but are, however, very tough to obtain. This access to large robust data sets presents a challenge when analyzing the meaning of genomic discoveries.

*Information Management*

Connected to the issue of data ownership and analysis is information management. Sharing data across collaborators can be difficult due to inconsistent field names or versioning of tools. Moreover, some custodians and stewards grow weary of regulating access and modifying data sets for every single study. Instead, one interviewee suggested that the general principle should be allowed to minimize workload for REBs (SOF1012). Data quality management and tests for robustness become more important as most labs are trying to run *"like businesses"* (SOF1011). Because data formats seem to differ a lot, most interviewees prefer to work with raw data, such as original sequence reads.

*Standards*

Genomic technologies are evolving fast and the lack of standards as well as guidelines can present barriers. Documenting the tools, all the steps of analysis or data modification are not common, which makes is almost impossible to reproduce results or integrate foreign data

(SOF1003). The lack of data standards and meta-tagging conventions place a burden on the researcher to establish and maintain their own data standards. However, for standards to be effective and consistent, larger groups of stewards and journals have to establish and enforce them. For instance, clinical-grade research standards differ from academic research standards and these differences should be documented. This also includes the need to unify what de-identified data should look like and the need to develop common ways to anonymize data.

*Learning New Tools*

The continuous development of new tools is challenging researchers who must keep up with the technology. New platforms and pipelines carry the promise of delivering better results, saving money, as well as saving time, but they also come with a set of challenges. Sometimes samples get tested on new platforms without delivering the expected results (SOF1012). Pioneering genomic discoveries with new platforms presents a different set of challenges than commercializing genomic discoveries. PIs constantly ask themselves if there is *"anything really new out there"* that they should be using (SOF1011). When it comes to learning and applying new tools, PIs still have to walk the tightrope between using 'trusted and benchmarked tools' or 'new, cutting edge technology.' Using new platforms or pipelines on pre-existing data sets can deliver very inconsistent results which have to be accurately documented.

*Validation*

Some interviewees feel that there is not enough funding or time spent on the validation of tools and data. Especially in the area of commercialization and moving genomic discoveries forward into the practice setting, or 'bed side,' academic research often ends after publication. One reason is the lack of funding and another is the need to publish more, faster.

> *"It's easy you know you go and buy it. Sure yeah it's a kind of crappy compound, but um but at least it sort of, sort of fits my bill. And they get some results out, which often are totally spurious because the compound really isn't properly designed. And there's actually, the literature's sort of rife with all sorts of crap because people have done these things" (SOF 1001).*

> *"…and came up with some data that looks interesting. And ah now they're trying to get a bigger grant to do it better, but they say, well what can we do? I said, well you know you have to validate what you've done. You know you may have this compound that does something but who's to say it doesn't… 'just because…it fits your theory—and it doesn't even fit that well. I mean the, the results are kind of marginal. But they're sort of in the right direction" (SOF1001).*

Moving genomic discoveries into clinical trials is a crucial and increasingly expensive endeavor. Researchers have to consider various barriers, such as size and costs of clinical trials.

> *"But then they say, well wait a minute, you're only tested 90 000 people how do we know it's safe? You know. This drug has tested a million people. We found that I don't know point one percent you know their hair fell out. If it's a thousand people, point one percent is only one person so maybe statistically you won't see it. So ah you're going to have to test that for five years to ensure it's safe [inaudible]. Yet I have to test it broadly, which costs a fortune, to serve the small number again. Yeah. And so that's really going to be a big issue in um, in the near future, is we develop some more personalized medicine" (SOF1001).*

Validation of drug targets and compounds is still expensive. When it comes to a return of investment, this problem of expensive validation creates a problem for new drugs and clinical trials for a potentially small market.

> *"Right now you have to come with any kind of cardiovascular drug and they'll say we need four years of safety [and] twenty thousand people.[...] And of course you know that ends up costing a billion dollars. You can't spend a billion dollars for a fifty million dollar drug" (SOF1001).*

*Commercialization*

Engagement with the commercial sector for leveraging funds in the event a promising discovery with commercial potential was one strategy many were aware of to move their science from labs to the applied domain.

> *"Interviewer: If it was something like Alzheimer's it'd be a different deal?*
>
> *Interviewee: Then there might be worthwhile to pursue it. And again that's because you're trying to…you want to have the potential to leverage and bring in the investment that comes from the, the commercial side…" (SOF1007).*

Commercialization and working with industry was seen as a means to raise funds required to move genomic discoveries into market:

> *"They…lots…so well, we just, you know, I think we came to realize that a lot of those uh commercial partners, uh they have money. They have a sales force. I mean it's…I mean it's not just about FDA. Even if you have an approved test, it doesn't mean that doctors will use it, right?" (SOF1010).*

Yet stakeholders also expressed concern (see above) that standards for academic publication and those for commercialization differed, at times significantly.

## 4. Discussion

The focus of our research was the identification of socio-technical issues which may be constraining the movement of genomic research from lab settings to practical application. We were particularly interested in issues related to data integration. Although at the onset of our project one of our objectives was to contribute to the development of a framework to enable categorization of sources of information required to realize the goals of personalized medicine and translational bioinformatics, part way through our observations in the two pre-clinical genomics labs we realized this would be of limited utility. First, the field is changing so rapidly that by the time such a tool might be developed, it would be out of date. Second, it became clear that genomics researchers were aware of an array of tools—for example computer discussion groups – which they could access to gain clarification of varied technical issues they faced in their work, in the event other lab members were unable to assist them.

It also became clear that knowing more about the data was but one part of a larger set of inter-connected challenges, which here we have identified as micro-level challenges. Lab staff worked in interdisciplinary teams, and, particularly early in their training, often struggled to gain a big-picture or overview. Even more senior stakeholders and researchers at times struggled with interdisciplinarity at times.

Perhaps a consequence of working in such interdisciplinary environments (and that they are trainees) leaves especially junior scientists prone to questioning their work, a phenomenon which was amplified perhaps by biological and computational affordances of the materials they worked with. At the same time, they worked with a frequently changing set of tools, often required the acquisition of new skills to use. Tools too carried affordances and constraints and had implications for subsequent findings.

Lab cultures are generally verbal cultures. While individuals record their own findings in lab notebooks, lab notebooks travel with the individual, and not the lab. Consequently, issues such as technical errors that had been found, workarounds which had been developed, where to look for help with particular issues or tools, what projects had been pursued but led nowhere, etc. were verbally communicated. While some of these issues might be particular to a lab, others—such as projects which had been pursued but failed—are arguably of interest to the broader genomics community.

In an environment where concerns have been raised about the quality of some compounds (for example) and the potential of data linkage and integration have been heralded, the less formal means of communicating about data (which seemed to be related to where data came from, medium used for their transfer, mode of transfer and transformation processes) are likely to slow progress of genomic work. While clearly the reliance on an oral culture to communicate about data is partly a cultural issue, it also partly reflects macro issues such as funding constraints. The hardest type of position to obtain funding for in an academic setting is a lab manager, and it is arguably lab managers who can contribute to standardization of communication practices in lab settings (e.g., documenting information about technical errors centrally rather than in individual lab notebooks, logging affordances and constraints of various tools, etc.

Cultural issues also influence the movement of genomics science from bench to bedside. Emphasis on publications in academic settings and competition for scarce funds may adversely influence sharing of data amongst genomic scientists. Stakeholders on the commercial end of things suggested that standards differed in academic and commercially oriented genomics, and often what was publishable in academic settings lacked validation making it suitable for commercial consideration. However, academic work was also significantly influenced by macro-level issues.

Macro level issues—those at a distance from the day to day work of researchers, came to bear on each of these issues. Both the scarcity of funds and the biases built into funding calls (e.g., certain kinds of positions may not be funded with grant resources; a focus on tool development rather than maintaining tools, etc.) influence the day to day work in labs, how that work is organized, etc. Funding guidelines influence team size and often team composition as well, all of which influence team productivity and success. Among the most significant policy issues for genomics researchers revolve around ethics, privacy and informed consent.

Sometimes analysts refer to complex trans-sector policy landscapes as being the site of ‘wicked problems.’ These are the kinds of problems that have so many contributing influences and emerging issues and contexts that it is almost impossible to sort out the best approach to untangling them. While this idea is more commonly applied while seeking to trace the complexities in other domains, it can also be applied to the policy landscape in personalized medicine (PM) and genomics. Looking at any one of the four ethical issues that are most commonly discussed in relation to genomics research—privacy, confidentiality, discrimination

and informed consent—leads to the other three (Jamieson, 2001) as well as toward questions about private sector agendas and balancing individual rights with what is the best for society as a whole (Knoppers 2010).

***Canadian Governance Instruments***

The global diversity of governance is reflected in both national and provincial jurisdictions in Canada. The major regulatory instruments in Canada are:

- The Tri-Council Policy statement: Ethical Conduct for Research Involving Humans (TCPS) (which includes the "tri-agencies" -- Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council (NSERC) and the Social Sciences and Humanities Research Council of Canada (SSHRC) (Hadskis, 2011).
- Research ethics boards (REBs) appointed by research institutions as part of TCPS requirements (Hadskis, 2011).
- Clinical Trail Regulations under the Food and Drugs Act and Good Clinical Practice (CGP): Consolidated Guidelines which apply to all clinical drug trials in Canada, irrespective of how the research is being funded" (Hadskis, 2011).
- Quebec and Newfoundland and Labrador Instruments described as "enacted legislation which directly deals with the conduct of human research" (Hadskis, 2011).
- Medical codes of ethics; legislation regarding informed consent; privacy legislation including the Personal and Electronic Documents Act (PIPEDA) and provincial acts (Hadskis, 2011; Levin and Nicholson, 2005).

All of these instruments address, to some extent or another, a variety of key issues emerging in relation to genetic research. Broadly, these issues can be described as privacy, confidentiality, discrimination and informed consent but this short list only skims the surface of the ethical issues arising from genetic research. While elsewhere several other issues not addressed in depth here have been addressed, such as conflict of interest (Hadskis, 2011), lack of resources for reviews of ongoing research (Hadskis, 2011; Peterson-Iyer, 2008), reassertion of discrimination against people with disabilities (Jamieson, 2001) and the creation of new marginalized groups such as the "not-yet-ill"(Jamieson, 2001, 35),issues raised consistently in in-depth stakeholder interviews highlighted the challenges the genomics community faces in relation to "secondary use" of information already gathered and "the confusion and frustration biomedical research stakeholders experience when attempting to navigate the current regulatory regime" (Hadskis, 2011, p. 499). Described by commentators as a "patchwork" and a "unwieldy hodgepodge" (Anderson et.al., p. 36; Kosseim, 2003, p. 115), Canada's collection of instruments for governing research inspire one scholar to state that this country's "regulatory framework for biomedical research falls short of offering a comprehensive research oversight system" (Hadskis, 2011, p. 450). Indeed, there is much work to be done coordinating and clarifying policy issues related to protection of personal information, informed consent and use of secondary data and as these issues pertain to genomics research.

## 5. Recommendations

Recommendations below span the full range of issues identified as challenges through our data collection and analyses. As such, they target change in both everyday work practices of labs through to provincial and federal initiatives which come to bear on genomics research.

***Improving Capacity of Labs***

**Principle:** Support improvement of lab management practices in order to realize practice efficiencies and improve the environment for data sharing amongst genomic scientists.

**Recommendation 1**: Support labs in developing organizational memory strategies for written documentation of lab practices, as well as more robust documentation and contextual information about data, so that information about technical errors and issues, as well as data context (e.g., data origin, prior transformations, tools used in data production and analysis, etc.) that can travel with data sets that might be re-used.

This can be supported through encouragement of

a) use of an e-documentation tool, which is properly indexed and which can capture meaningful information, within labs, about issues such as
b) development of non-verbal (e.g., written to be shared by the lab) means of organizational memory within labs (e.g., documentation tools for making tacit knowledge more visible (e.g., trouble-shooting list of 'what to do if…')
c) development of a publically accessible (and anonymous) log-file of 'failed' experiments, interventions, or tools.

Strategies to encourage sharing of these ideas might include contests (similar to ImagineNation Challenges run by Canada Health Infoway—see http://imaginenationchallenge.ca/about-imaginenation/) where organizations are encouraged to share best practices in exchange for an incentive such as a small cash reward or, when larger amounts of work are involved, the opportunity (for example) to travel to a conference of interest to target audiences.

***Improving Capacity of Trainees***

**Principle:** Address knowledge gaps through targeted programming delivered through existing training mechanisms such as the NSERC – Collaborative Research and Training Experience (CREATE) training program and existing bioinformatics programs.

**Recommendation 2:** Develop targeted educational strategies to enhance ability to work across disciplines, and deliver them through existing training mechanisms such as NSERC – Collaborative Research and Training Experience (CREATE) bioinformatics training program.

While ideally more work should be undertaken to determine both specific needs of trainees concerning interdisciplinary thinking and how to support those needs, past work suggests that exploring differences in how each discipline frames problems, speaks about research problems (e.g., vocabulary used, etc.) and sharing of information about issues and challenges particular to each disciplinary perspective would all be useful. Problem based learning in a classroom setting may also be a fruitful avenue for addressing issues of cross-disciplinarity.

**Recommendation 3**: Develop case examples for teaching that encourage critical thinking about data quality, affordances and constraints of tools, etc. which can be used to encourage awareness of the relationship between tool use and findings.

Ideally case based learning examples would be developed such that they could be easily incorporated into varied courses, which would allow trainees to gain exposure to issues related to tool affordances, data quality and sense making throughout their curriculum.

**Recommendation 4**: Design case based learning resources which highlight issues related to standardization (e.g., the lack of data and tool standards, where standards exist, limitation of standardization, etc.) which can be integrated into varied courses concerned with genomics.

Thinking about standards goes hand in hand with thinking about data quality. While there are extensive activities going on at national and international levels related to standards and standardization within genomics research, engagement with issues related to standardization within BC's genomics community appear to be somewhat limited. Hence, the above recommendation is designed to encourage engagement with issues and questions related to standardization of data and processes among trainees, who will hopefully become more active in addressing these issues nationally and internationally as their careers progress.

**Recommendation 5**: Develop something akin to a library research guide to assist trainees in identifying resources that are particularly good for addressing certain kinds of issues (description of forums and other resources related to problem solving while undertaking lab-based work.

***Reduce Barriers Through Changes in Financial Support***

**Principle:** Develop financial mechanisms to support research in areas where gaps have been identified.

**Recommendation 5:** Build funding for core lab technicians into funding programs, to support both organizational memory activities outlined above, as well as maintenance of existing datasets and tools.

Support for this recommendation should also positively impact validation of findings, as validation can often be achieved more readily when older datasets and tools known to be robust can be used.

**Recommendation 6:** Increase funding available to validate findings, and move from academic accuracy to clinical accuracy. This might be conceptualized as supporting research beyond initial academic publication, or as proof of principle funding.

***Reduce Barriers Through Support for Cultural Changes and Cross-Stakeholder Collaboration***

**Principle:** Sharing of data across labs, research groups and institutions requires resolution of issues concerning data ownership and data quality, the development of data stewardship policies, and would benefit from support of harmonization of consent processes and forms.

**Recommendation 7:** Host a workshop to be attended by senior members of BC's genomics research community, to discuss issues of data ownership, intellectual property and the role these play in willingness (or lack of willingness) to share data across labs. Have as a goal of this workshop development of a set of principles for data sharing policies, to be further developed by one or more working groups (e.g., if data quality and or data standards arise as issues during the workshop, these might need to be addressed following the workshop by working groups dedicated to each of these topics).

**Recommendation 8:** Host across-sectoral workshop with senior representatives from research ethics boards, the privacy commissioner, the Ministry of Health, Health Canada, senior staff with

operational responsibilities at PopData BC and senior genomic scientists who have engaged (or attempted to engage) in cross-institutional / multi-institutional data sharing or data linkage projects. The focus of the workshop should be identification of constraints to data sharing and linkages related to genomics, and the development of strategies for addressing public concerns while reducing barriers to data sharing and linkages for research purposes.

**Recommendation 8a:** Support the development of consent language which is robust enough to protect individual rights to privacy but also allows secondary use of data by academic researchers. Consider the development of unified and centralized general consent forms to allow researchers to conduct, for instance, secondary analysis of pre-existing samples without re-consenting.

**Recommendation 8b:** Support development of guidelines and/ or standards for de-identification of data as a means of providing data stewards with guidance about how to share data and remain compliant with regulations. This should also be undertaken with support from chairs of research ethics boards, who will be responsible for accepting processes developed.

**Recommendation 9:** Provide financial support for deliberative dialogues and other forms of public engagement to address issues of privacy, discrimination, data sharing and secondary use of data and informed consent, in order to raise awareness amongst members of the public about issues related to genomic data. Undertake planning with research ethics board chairs, the privacy commissioner, senior members of the genomics research community that have experienced difficulty gaining approval for data linkage and/or secondary use of data. Include members of the public who have been visible in relation to these issues. Consider hosting this in consort with an event such as the annual data effect conference, in order to enhance visibility.

**Recommendation 10:**Provide funding for a cohort of bio-ethicists to gain exposure to genomics research through becoming an embedded member of genomics research teams, in order to gain more practical experience with the issues and challenges genomics scientists face.

## References

Abrahams, E., Ginsburg, G. S., & Silver, M. (2005).The Personalized Medicine Coalition. *American Journal of PharmacoGenomics*, *5*(6), 345-355.

American Medical Informatics Association.(2007). AMIA Strategic Plan. 2006.

Anderson, J.A., McDonald, M., Preto, N., Pullman, D., Sampson, H. (2011). Research ethics in 2020: Strengths, weaknesses, opportunities and threats. *Health Law Review. 19(3)* (Summer): 36.

Atkins, D. E., Droegemeier, K. K., & Feldman, S. I. (2003). Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. *Revolutionizing Science and Engineering Through Cyberinfrastructure*.

Atkinson, P., Glasner, P. and Lock, M. (2009) Handbook of genetics and society: Mapping the new genomic era, Routledge, London.

Avgerou, C. (2002) New socio-technical perspectives of IS innovation in organizations, in C. Avgerou and R.L. LaRovere (Eds.): *ICT innovation: Economic and Organizational Perspectives,* Edward Elgar, Cheltenham.

Azeem, M., Salfi, N. A., &Dogar, A. H. (2012).Usage of NVivo software for qualitative data analysis. *Academic Research International*, *2*(1), 262-266.

Balka, E. (2005, September). The production of health indicators as computer supported cooperative work: reflections on the multiple roles of electronic health records. In *Reconfiguring Healthcare: Issues in Computer Supported Cooperative Work in Healthcare Environments. Workshop Organized by Ellen Balka & Ina Wagner European Computer Supported Cooperative Work Conference* (pp. 67-75).

Bostrom, R.P., and Heinen, J. S. (1977) MIS problems and failures: A socio-technical perspective, Part 1: The causes, *MIS Quarterly,* 1, 3, 17-32.

Bottinger, E. P. (2007). Foundations, promises and uncertainties of personalized medicine. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine*, *74*(1), 15-21.

Bowker, G. C., & Star, S. L. (2000). Sorting things out: Classification and its consequences.MIT press.

Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative data analysis for health services research: developing taxonomy, themes, and theory. *Health services research*, *42*(4), 1758-1772.

Bush, C. G. (1982). Taking Hold of Technology: A Topic Guide for the 80's.*American Association of University Women, Washington*.

Cascorbi, I. (2010). The promises of personalized medicine. *European journal of clinical pharmacology*, *66*(8), 749-754.

Chow-White, P. A., &García-Sancho, M. (2012).Bidirectional shaping and spaces of convergence interactions between biology and computing from the first DNA sequencers to global genome databases.*Science, Technology & Human Values*,*37*(1), 124-164.

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, *25*(5), 528-540.

Edwards, P. N., Jackson, S. J., Bowker, G. C., &Knobel, C. P. (2007). Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures". *Understanding Infrastructure: Dynamics, Tensions, and Designs*.

Evans, J. P., Meslin, E. M., Marteau, T. M., & Caulfield, T. (2011).Deflating the genomic bubble.*Science*,*331*(6019), 861-862.

Fackler, J. L., & McGuire, A. L. (2009).Paving the way to personalized genomic medicine: steps to successful implementation.*Current pharmacogenomics and personalized medicine*,*7*(2), 125.

Gaskell, G., and Bauer, M. W. (2006) Genomics and society: legal, ethical, and social dimensions, Earthscan, London.

Ginsburg, G. S., & Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications.*Translational research*,*154*(6), 277-287.

Glaser, B., & Strauss, A. (1967). The discovery grounded theory: strategies for qualitative inquiry.*Aldin, Chicago*.

Hadskis, M. (2011)."The Regulation of Human Biomedical Research in Canada" in J.Downie, T. Caulfield, and C. Flood, (Eds.).*Canadian Health Law and Policy*, 4th ed. (Markham: Lexis/Nexis, 2011) 437-500.

Halcomb, E. J., & Davidson, P. M. (2006). Is verbatim transcription of interview data always necessary?.*Applied Nursing Research*,*19*(1), 38-42.

Hughes, T. P. (1983) Networks of power: Electrification in Western Society 1880-1930, John Hopkins University Press.

Jamieson, C. (2001). Genetic testing for late onset diseases: Current research practices and analysis of policy development. *Health Policy Working Paper Series*, Health Canada: Ottawa.

Khoury, M. J., McBride, C. M., Schully, S. D., Ioannidis, J. P., Feero, W. G., Janssens, A. C. J.,& Xu, J., (2009). The scientific foundation for personal genomics: recommendations from a National Institutes of Health–Centers for Disease Control and Prevention multidisciplinary workshop.*Genetics in Medicine*,*11*(8), 559-567.

Kling, R. and Scacchi, L. (1982) The web of computing: Computer technology as social organization, *Advances inComputers,* 21, 1-90.

Knoppers, B.M. (2010). Consent to 'personal'genomics and privacy. *European Molecul EMBO reports*, *11 (6),* 416 – 419.

Kosseim, P. (2003). The Landscape of Rules of Governing Access to Personal Information for Health Research: A View from Afar. *Health Law Journal*, 113.

Kvale, S. (2007). *oing interviews*. Sage.

Latour, B. (2005). Reassembling the social-an introduction to actor-network-theory. Reassembling the Social-An Introduction to Actor-Network-Theory, by Bruno Latour, pp. 316. Oxford University Press, Sep 2005

Laurence, J. (2009). Getting personal: the promises and pitfalls of personalized medicine. *Translational Research*, *154*(6), 269-271.

Leonelli, S. (2010). Machine science: The human side. *Science(Washington)*,*330*(6002), 317-317.

Lesko, L. J. (2007). Personalized medicine: elusive dream or imminent reality?.*Clinical Pharmacology & Therapeutics*, *81*(6), 807-816.

Levin, A., and Nicholson, M. J. (2005).Privacy Law in the United States, the EU and Canada: The Allure of the Middle Ground. *University of Ottawa Law and Technology Journal*. 2(2), 357-395.

Lin, A., & Cornford, T. (2000). *Sociotechnical perspectives on emergence phenomena* (pp. 51-60).Springer London.

Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A., &Tarczy-Hornoch, P. (2007).Data integration and genomic medicine. *Journal of biomedical informatics*, *40*(1), 5-16.

MacArthur, D. (2011, February 18). When "Cautious" Means "Useless." Retrieved November 27, 2015, from http://www.wired.com/2011/02/when-cautious-means-useless/

Micheel, C. M., Nass, S. J., & Omenn, G. S. (Eds.). (2012). Evolution of translational omics: lessons learned and the path forward. National Academies Press.

Miles, M. B. and Huberman, A. M. (1994). Qualitative data analysis, Thousand Oaks, Sage, CA, USA.

Mumford, E. (1983) Designing human systems, Manchester Business School, Manchester, UK.

Orlikowski, W. (1992) The duality of technology: Rethinking the concept of technology in organizations, *Organization Science,* 3, 3, 398-427.

Orlikowski, W., and Scott, S.W. (2008) The entangling of technology and work in organizations, in *Working Paper#168, London School of Economics and Political Science*.

Patton, M. Q. (2002). Qualitative research & evaluation methods (3rd ed.). Thousand Oaks, CA:Sage.

Payne, P. R., Embi, P. J., & Sen, C. K. (2009). Translational informatics: enabling high-throughput research paradigms. *Physiological genomics*, *39*(3), 131-140.

Peterson-Iyer, K. (2008). Pharmacogenomics, ethcis and public policy.*Kennedy Institute of Ethics*, 18(1), 35 - 56

Schuurman, N. (2008). Database ethnographies using social science methodologies to enhance data analysis and interpretation. *Geography Compass*, *2*(5), 1529-1548.

Schuurman, N., & Balka, E. (2009).alt. metadata. health: Ontological Context for Data Use and Integration. *Computer Supported Cooperative Work (CSCW)*,*18*(1), 83-108.

Sheppard, D. (2000). Beginner's Introduction to perl, Retrieved 2012-02-17 from http://www.perl.com/pub/2000/10/begperl1.html

Star, S. L., and Ruhleder, K. (1996). Steps towards an ecology of infrastructure, *Information Systems Research*, 7, 1, 111-134.

Strauss, A. L, and Corbin, J. (1990) Basics of qualitative research, Sage Publications, Inc; 3rd edition.

Suchman, L. A. (1987). Plans and situated actions: the problem of human-machine communication. Cambridge university press.

Tilson, D., Lyytinen, K., &Sørensen, C. (2010). Research commentary-digital infrastructures: the missing IS research agenda. *Information systems research*,*21*(4), 748-759.

Tisdall, J. (2000) Mastering BioPerl for bioinformatics, Retrieved 2012-02-17 from http://oreilly.com/catalog/mperlbio/chapter/ch09.pdf

Trist, E.L. and Bamford, K.W. (1951) Social and psychological consequences of the Longwall method of coal-getting, *Human Relations*, 4, 3-28.

Walker, B.H, Holling, C.S., Carpenter, S.R., and Kinzig, A. (2004).Resilience, adaptability and transformability in social–ecological systems, *Ecology and Society,* 9, 2, 5.

Welsh, E. (2002, May). Dealing with data: Using NVivo in the qualitative data analysis process. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 3, No. 2).

Woolf, S. H. (2008). The meaning of translational research and why it matters.*Jama*, *299*(2), 211-213.

## Appendix A: Table of Node Structure and Node Definitions (Observation and Interview Data)

| Parent Nodes (Node) | Child Nodes (sub node) | Grand Child Node (sub-sub node) | Node Definitions |
|---|---|---|---|
| Analysis | 1. Scientific Reasoning | 1.1 No prior evidence | |
| | 2. Sense Making | 2.1 Good Interpretations | Data is coded to this node when the researcher is able to anchor his/her research into a sound justification, or when the PI and colleagues show sings of approval for the quality of interpretation carried out. |
| | | 2.2 Justification | |
| | | 2.3 Problem solving | |
| | 3. Standardization | 3.1 Following practical standards | When researchers justify the method/tool they used as a 'practical' one that they chose to use either to save time or money, or because there is experience with it in the lab, or even resources. |
| | | 3.2 Following scientific standards | This is when researchers emphasize the use of a tool/method due to it being the better one as a scientific standard rather than it being a practical tool/method. |
| | | 3.3 Lack of standards | |
| Case of Interest | | | These are stories/observational notes that I thought work as a perfect example that could be used later on when writing the project report/publishing. |
| Challenges | Cloud Computing | | References to cloud computing and cloud solutions. |
| | Cost, time and funding | - | Anything related to cost of doing research (time or money), and funding constraints or biases (favoring the low-hanging fruit). |
| | Culture | - | |

| | | | |
|---|---|---|---|
| | Data integration across platforms | - | Any challenges related to integrating more than one data set from different platforms. Mostly these challenges relate to difficulties in findings common keys/columns/identifiers because of different formatting methods and no unique standard followed by the majority of websites/dbs. |
| | Errors and technical problems | - | Any challenges faced when using tools, such as a tool not working accurately. Or technical errors such as some databases being offline sometimes. |
| | Ethics | - | Any mention of ethical problems faced when doing research. This could be related to getting ethical clearance, or real ethical issues that scientists might be debating about. |
| | Failed wet lab experiments | - | Any mention of failed attempts to run analyses in the wet lab. |
| | Failed dry lab experiments | - | Any mention of failed attempts to run analyses in the dry lab. |
| | Inherent knowledge or biases | | Mentions of implicit assumptions in databases or tools, such as facts related to why a tool was conceived, and how it 'should' be used, or its limitations due to what it was originally designed for. |
| | Insufficient computing power | | |
| | Learning new tools | - | |
| | Other misc. challenges | - | |
| | Politics of data ownership | - | |
| | Poor documentation of tools | - | |
| | Temporality of | - | |

| | | | |
|---|---|---|---|
| | projects | | |
| Cycles of credit | Time | - | |
| | Ownership of data or research | - | Nodes related to when scientists explicitly give credit to another scientist for basing research on their past research, or acknowledge their current input or help; or mentioning that they have extended their help to others. I'm not coding at this parent node. |
| | Publication motives | - | |
| | Receiving or giving credits | - | |
| Practicing science | Receiving or Giving Information | - | |
| | Accuracy and precision | - | Scientist's work that resembles working toward quality, or striving to be exact and accurate. |
| | Collaboration | - | Working in teams or in pairs to co-produce knowledge. Learning from each other, or basing research on others past work. The practice of working collaboratively for the greater good. |
| | Conformity | - | When scientists conform to standards, or agree to supervisor's perspective, or other proven research. Also includes conforming to the common taken-for-granted and known scientific 'methods' , or those proven or known as accurate, and research methods generally. |
| | Goals or research contribution | - | Mention of research goal, or main contribution to research. |
| | Knowledge production | - | |
| | Routines and procedures | - | |
| | Supervision | - | |

| | | | |
|---|---|---|---|
| Using bioinformatics tools | Training and becoming a scientist | - | |
| | Alignment tools | - | Any mention of Alignment tool use, which are defined by wiki as: "In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences". |
| | Databases | - | Databases used and/or mentioned in lab meetings or during observations and/or interviews. |

## Appendix B: Map of Genomic Landscape

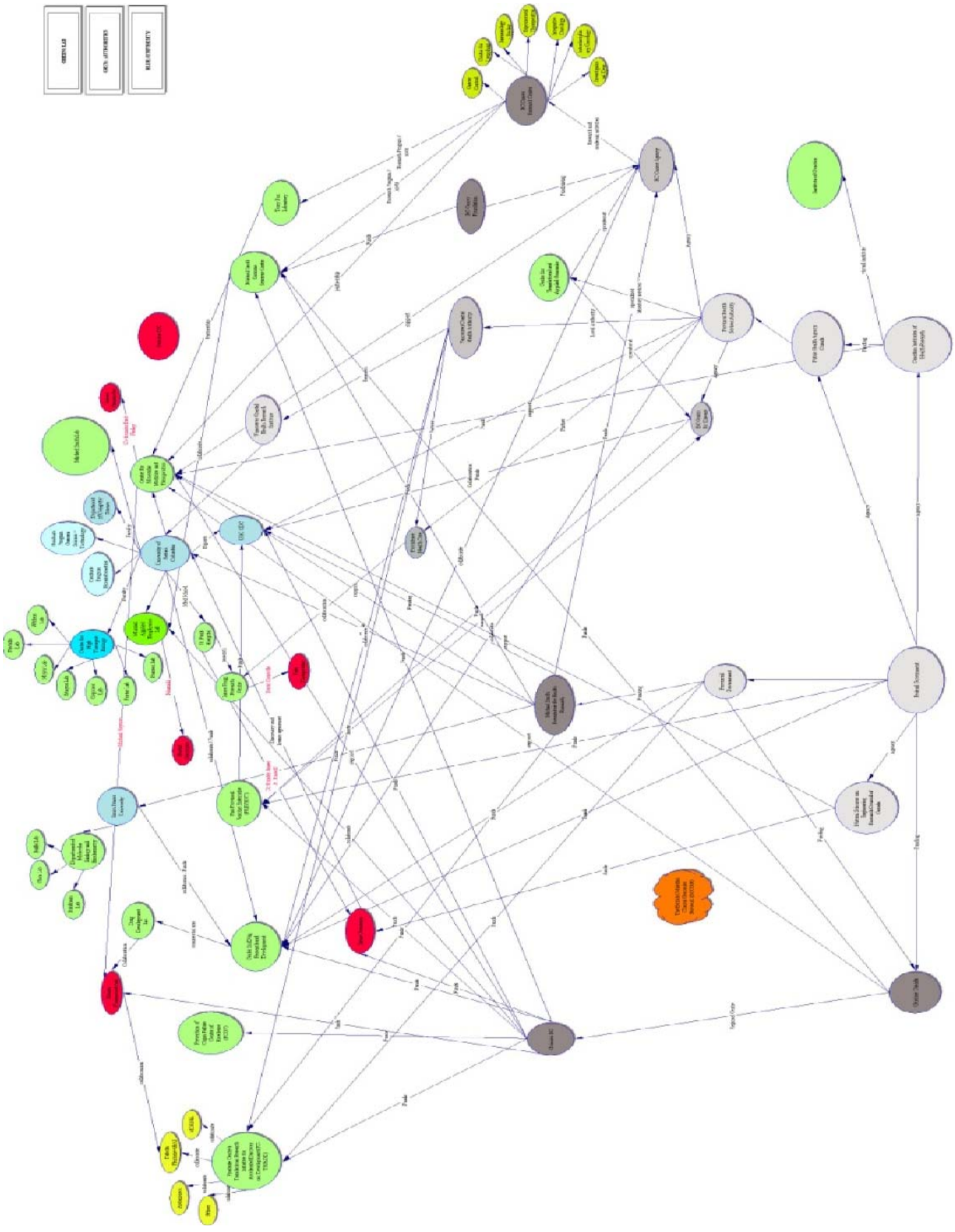


Figure5: Map of institutions and stakeholders.

## Appendix C: Interview Guide: In-depth Semi-structured Interviews

PART 1: Get respondents involved (general questions, i.e. “facts” or definition)

Can you tell me about your job in general, and about working with data in particular? (Can you describe to us a usual day for you?)

PART 2: Cover potential ‘controversial’ issues

PART 2.1 General (might be ethics, policies, financial, data ownership, standardization...)

What obstacles do you face in your line of work?

In terms of working with data, what are the biggest obstacles you face in your job? Can you give me some examples? Have others tried to resolve these issues or have you?

Have you done anything in terms of trying to respond to these sorts of challenges? If so, what?

PART 2.2. Socio-technical challenges

From your perspective, what are the key issues related to data use in general and data integration in particular, in genomics?

On a day to day basis, do you face any data integration issues?

(e.g. using one set of data output across different tools)

Do you have to reformat data often?

What is the most pressing issue in Genomics/Bioinformatics today? (in relation to translation)?

In your view, what are the key challenges that you would like to have addressed so that genomics research in general and bioinformatics research can improve?

Our interest in these issues partly relates to the long term goals of personalized medicine. We’re interested in hearing from you about how you understand the term personalized medicine, whether you see your work as part of that, etc. What does it mean to you? Do you think of your work as part of it?

What is translational bioinformatics? Do you feel part of it?

From your perspective, what are the biggest issues arising in genomics research you are involved with now that related to data which are constraining movement of genomic discoveries into use?

Where do you see genomics and personalized medicine in 5 / 10 years?

CLOSING QUESTION:

1. Is there anything else you would like to add?

# Appendix D: Screenshot of Node Structures Including In-depth Interview Data