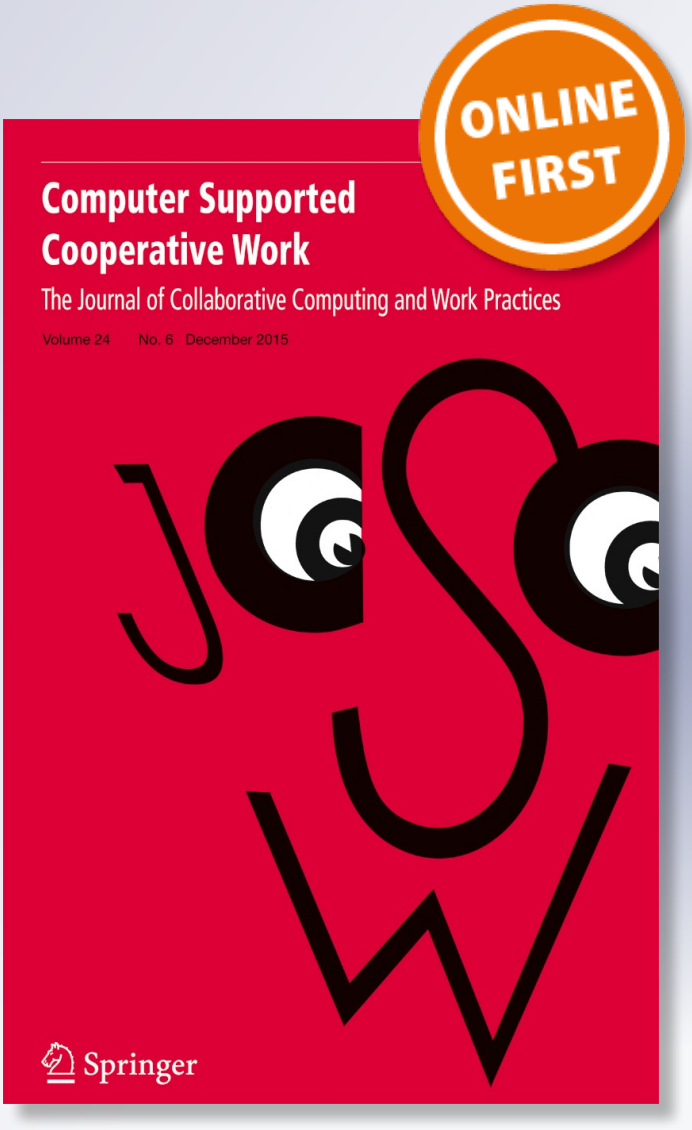
Genomic Big Data and Privacy: Challenges and Opportunities for Precision Medicine

Julie Frizzo-Barker, Peter A. Chow-White, Anita Charters & Dung Ha

Computer Supported Cooperative Work (CSCW)
The Journal of Collaborative Computing and Work Practices

ISSN 0925-9724

Comput Supported Coop Work
DOI 10.1007/s10606-016-9248-7



Springer

Computer Supported Cooperative Work (CSCW)
DOI 10.1007/s10606-016-9248-7

Genomic Big Data and Privacy: Challenges and Opportunities for Precision Medicine

Julie Frizzo-Barker, Peter A. Chow-White, Anita Charters & Dung Ha
School of Communication, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6BC , Canada (Phone: 778-782-9454; E-mail: jfrizzob@sfu.ca; E-mail: petercw@sfu.ca)

Abstract. Genome science is rapidly shifting from research labs and biobanks to the clinical setting. The resulting genomic big data, or large-scale networked genetic material, is a disruptive technology. On one hand, clinical genomics advances life-saving innovation through precision medicine. On the other, the digital databases they are built upon raise new concerns for informational risk to personal privacy. While a traditional biomedical approach focuses on risks and benefits to the human body, our socio-technical analysis sheds lights on the emerging terrain of the human body as digital code. In this paper, we analyze emerging issues related to clinical genomics based on a 3-year collaborative clinical research project to develop a genomic test for Acute Myeloid Leukemia (AML) cancer in British Columbia (BC), the first of its kind in Canada. We found the most pressing issues for genomic researchers and clinicians were challenges around informed consent, return of results and return of incidental findings. In light of technological advances and the emerging context of networked privacy, we outline several recommendations for best practices in diffusing clinical genomics to the healthcare system.

Keywords: Big data, Databases, DNA, Genomics, Healthcare, Precision medicine, Privacy, Informational risk, Information technology

1. Introduction

Genome science is rapidly shifting from research labs and biobanks to the clinical setting. Over the past decade, information technologies such as whole-genome sequencing (WGS) and digital genome databases have transformed the complex process of recording an individual's entire genetic code from a decade-long, billion-dollar, global endeavor, to a week-long, $1000 service (Burn-Murdoch 2012). These advances have sparked debate over the benefits and drawbacks of a future healthcare system based on precision medicine, or the practice of using genetic information and other biological features to tailor healthcare to individual patients (Bush and Moore 2012). For example, a woman with incurable colorectal cancer in British Columbia (BC), Canada, recently received treatment through the BC Cancer Agency's Personalized Onco-Genomics Program (POG). After genomic sequencing, they administered a drug normally used to treat high blood pressure, which reduced her advanced condition to undetectable in just 5 weeks (CTV 2015). Even her oncologist, Dr. Howard Lim, was surprised with the results: "To be quite honest I did not anticipate the kind of response that she would get, but this is what POG is

Published online: 04 March 2016

about—trying to find surprises." This is a prime example of how genomic big data, or large-scale, networked analysis of patient genetic material, is fueling life-saving medical innovation. At the same time, clinical genomics also produces an abundance of individual DNA information to be managed (Caulfield et al. 2010; Wright et al. 2011; PCSB 2012, 2013). As patients' DNA are translated into petabytes of digital data, our shifting socio-technical landscape is also characterized by informational risks to privacy (Frizzo-Barker and Chow-White 2014). While a traditional biomedical approach focuses on risks and benefits to the human body, our analysis sheds lights on the emerging terrain of the human body as digital code.

Advanced information technologies form the bedrock of data-sensitive eScience projects such as genomics. They enable scientists to experiment with reusing experimental information to overcome barriers to knowledge exchange, and connect different databases to facilitate data sharing and publication (Zhao et. al. 2011). Yet genomic big data presents new challenges to both individuals and families. For instance, genetic discrimination may prevent an individual from accessing insurance coverage or employment opportunities. Beyond this risk, a human genome is not only a unique identifier of an individual but a network identifier of familial relations and hereditary diseases—information that may be more easily identified, accessed and replicated via digital databases (Allyse et. al. 2012; Lunshof et. al. 2008). Therefore within our discussion of best practices for clinical genomics, we must now consider the right to know, and the right *not* to know as elements of privacy. Critical research on the social implications of genomic big data is both timely and important. Genomic datafication, governance of digital health information, and networked privacy must be addressed in the earliest stages in order for clinical genomics to be adopted into healthcare systems with socially and ethically sustainable practices.

In this paper, we analyze emerging socio-technical issues related to clinical genomics within the Canadian context. The authors of this article have recently completed a 3-year collaborative clinical research project to develop a genomic test for Acute Myeloid Leukemia (AML) cancer. It is potentially the first molecular test in Canada and a key moment in the translation of genomics from the research lab to the hospital bedside. Over a 3 year period, we conducted documentary and policy analysis, as well as interviews with active genome researchers, clinicians and decision-makers in the Vancouver, BC, area to examine opportunities, challenges, and risks to different stakeholders in clinical genomics. We also worked side by side with genome scientists, bioinformaticians, clinicians, and health economists. In our research and collaborations, we identified key issues regarding privacy and management of sensitive genomic information related to informed consent, return of results and incidental findings at the point of care. We found the adoption of genome technologies presented privacy challenges for both clinicians and patients. In the policy environment, we found that although many regulations and guidelines exist, the state of best practices is uncertain. Our findings reveal a number of socio-technical problems with the shaping of genomic technologies in the clinical setting.

Our recommendations are intended to inform researchers, clinicians, and policy and decision-makers in government and the health care system.

One of the central concerns of Computer-Supported Cooperative Work (CSCW) is to close the social-technical gap between social problems and technical systems (Ackerman 2000). As genomics diffuses from the bench to the bedside, bridging this gap will be of chief importance in order to avoid "the last mile problem," when new technologies with enormous investment backing fail to integrate into society in a widespread, useful way (Blumenthal-Barby et al. 2015). Scientists and bioinformaticians can now conduct large scale genomic data analysis, however there is a lack of consensus on the social implications of predictive analytics, or guidelines on how to apply results clinically or procedurally.

In many ways, genomic research represents a quintessential CSCW project. The mapping of the first human genome, spanning from 1990 to 2003, set the precedent for genomic research as a collaborative field of scientific research involving a global group of scientists, organizations, and funding (Collins et al. 2003). This endeavor involved participants from disparate geographical locations and disciplinary fields, using technological infrastructures to support communication and painstakingly large-scale data analysis, in order to produce results that could not have been otherwise realized. The significance of this research model reflected the ethos of the human genome project itself: sharing such important data via Internet databases made it widely accessible for the greater good of society.

Since the late 1990s, the rise of the Internet and related information technologies amplified the networked potential for genomic research to reach a wider sphere of medical innovation. CSCW scholars have developed key concepts that help to inform our analysis of genomic research, including "context-based metadata" (Schuurman and Balka 2008, p. 83), "cyberinfrastructures" (Bietz et al. 2010, p. 245), and "metagenomics," which highlights the social aspect of databases as "contested sites around which particular research questions are supported or disenfranchised through the inclusion or exclusion of necessary metadata" (Jirotka et al. 2013, p. 688).

In addition, scholars of communication, science and technology studies (STS) and critical race studies have traced the coevolution of biomedicine and information technologies for over 30 years. Scholars of communication and STS have analyzed the socio-cultural impacts of mediated information, information technologies, and the network society (McLuhan 1964; Williams 1975; Castells 2000; Bowker and Star 2000; Boczkowski and Lievrouw 2008). And over the past decade, scholars of biotechnology, race and politics have explored key tensions at the intersection of this rapidly evolving field (Thacker 2004, 2005; Condit et al. 2004; Condit 2007). Chow-White and Garcia-Sancho (2012) conceptualized DNA databases as "spaces of convergence for computing and biology" that have evolved in both form and function over the past 50 years, setting the stage for today's genomic research in which "the biological and computational are currently indivisible" (p. 128). The convergence between biology and computing transformed genomics from wet lab

science into a big data project. The emergence of clinical genomics is another transformative moment.

Kitchin (2014) highlights the need for empirical studies of data assemblages to provide "holistic accounts of how they are constituted and operate in practice" (p.190). Our paper contributes to this gap in the literature through a socio-technical, ethnographic study of front-line stakeholders bridging genomic research and medicine. We begin by exploring some of the benefits, challenges and risks that arise in the space of convergence where personal genomes become digital big data. Next we outline the concept of networked privacy and its implications for researchers, healthcare providers, policy makers, and the general public. We then describe the methodology for our empirical study of genomic researchers, clinicians and decision-makers. Finally, we report our findings and recommendations for best practices in diffusing clinical genomics to the healthcare system. While our study is based in the Canadian context, it has implications for researchers, clinicians and policy makers globally.

2. Genomic big data

Both technically and culturally, genomics is big data. A single human genome is a complex formation made up of six billion bases of information. The file size of a single genome can range from about 700 MB of raw data to 200GB of annotated variant and metadata. Genomic sequencing is no longer a process built on test tubes and pipettes, but on information technologies and databases. The world's largest genomics research institute, China-based BGI, sequences the equivalent of 2000 human genomes per day, limited only by the fact that "the ability to determine DNA sequences is starting to outrun the ability of researchers to store, transmit and especially to analyze the data" (Pollack 2011).

The most popular definition of big data focuses on these technical aspects: the immense volume, variety, and velocity of available data (IBM 2012). Kitchin (2014) expands on the 3 Vs, identifying other features of big data: exhaustive data sets (n = all), fine grained resolution, relational in nature, flexibility and scalability. Others highlight that big data is all about predictions, connections, and relationships amongst vast data sets (Mayer-Schönberger and Cukier 2013). These approaches to defining the phenomenon are helpful in identifying its material capacities. However, greater volumes, variety and velocity of data are not necessarily revolutionary (Strasser 2012). Rather, we argue that beyond bigger, better, networked data, today's big data is defined by its novel applications.

Big data represents a diffusion of data-driven approaches to decision-making into new industries and enterprises and an expansion of uses in data-friendly environments. For example of the former, actors in professional sports use new analytical approaches and collect new forms of data to understand player valuation and team dynamics. The development of genomics for clinical practice is an example of the latter. Big data also represents a cultural shift where actors gather around databases

and form new types of collaborations. Chow-White and Garcia-Sancho (2012) show how early big data ventures, such as genomics, brought together different groups in the academy, such as molecular biologists and computer scientists, which had not traditionally worked together. This type of collaboration became a new scientific field, with new methodological approaches that influenced the development of both disciplines.

The mapping of the first human genome, spanning from 1990 to 2003, set several important precedents. First, it represented a global, collaborative field of big data research involving a global group of scientists, organizations, and funding (Collins et. al. 2003). Second, it marked a victory for the open-access approach of the multidisciplinary Human Genome Project (HGP), which trumped the simultaneous attempts of Celera, a private company, to map and patent the first human genome (Marris 2005). The significance of this collaborative, open access research model reflected the ethos of the HGP itself: sharing such important data via Internet databases made it widely accessible for the greater good of society. In the middle of the HGP, American and British project scientists met in neutral Bermuda to discuss the progress of the project. One of the main issues they raised was how to ensure genomic data would be a public good rather than privately owned. They developed a set of principles requiring all DNA sequence data to be copyright-free and released to open access networks within 24 hours of generation, in stark contrast to traditional scientific practices of releasing experimental data only after publication (Contreras 2010). The resulting "Bermuda Principles" policy initiative shaped contemporary open access scientific practices and the very concept of viewing information as a global knowledge resource.

Kitchin (2014) points to some of the new uses and insights facilitated by big data in his definition of *data assemblages*: "amalgams of systems of thought, forms of knowledge, finance, political economies, governmentalities and legalities, materialities and infrastructures, practices, organizations and institutions, subjectivities and communities, places, and marketplaces" (p. xvi). This highlights the fact that big data is not a single technology. Rather, it is a cluster of different information technologies and techniques for finding patterns in large data sets, with social and cultural ramifications. In light of this definition, we identify a weak spot in big data approaches: they are useful in terms of *what* to analyze, but not *why* or *how* we ought to go about it. The *how* acknowledges the challenges of data analysis, and the *why* points to a larger range of social and ethical issues, before we focus on the *what*, smallest and most crucial aspect of the data itself. As Bowker notes, "raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care" (2005, p. 184). In communication terms, data are representations. They are "cultural objects that stand in for stimuli and mediate relations" (Chow-White and Green 2013 p. 6). Genomic data holds new social significance, not only because it represents an individual's unique DNA, but also because it can take on a tangible, material life of its own as it enters the digital database. This is especially true for genomic data,

which may be accessed, replicated, or analyzed in unforeseen ways with unknown implications, especially with the rise of data mining techniques.

Data mining refers to the process of attempting to discover patterns and meanings from large data sets, with the Internet being the most obvious example of such a database (Manovich 2001). Data mining involves gathering "disparate types of information from users and consumers—sometimes with the users' knowledge, sometimes without—and turn[ing] this information into analytical data points for measurement, sorting, and classification to achieve different organizational and institutional goals" (Chow-White and Green 2013, p. 556). Medical researchers and clinicians are eager to employ such information technologies for the purposes of predictive analytics in genomics to aid in medical discovery, streamline policies and programs, and evaluate critical data (Gordon and Pai 2012). For example, the International HapMap Project traces genes associated with human disease, the 1000 Genomes Project aims to identify common genetic variants, and the $20 million International Cancer Genome Consortium gathers comprehensive data on genomic, transcriptnomic and epigenomic changes in 50 different tumor types (Gulland 2010). Data is continually uploaded to these open access databases, which are accessible to anyone with the ability to make sense of the information (HapMap 2015; ICGC 2015; 1000 Genomes 2015). In comparison, direct-to-consumer genetic testing services such as 23andMe involve Google-backed databases closed to public access, but have sparked their own controversy such as an FDA lawsuit over accuracy of test results (Rukovets 2014).

Data can be seen as the life-blood of the scientific realm, and a perceived advantage of eScience is precisely this "commons of information" or "data that can be easily accessed, reused and shared in collaborations across both geographical and disciplinary boundaries" (Jirotka et al. 2013, p. 671). Yet however altruistic in design, open access genomic databases raise several thorny questions. For example, where does responsibility for the stewardship and governance of data ultimately lie? As we bargain for medical innovation, how is personal privacy challenged? Although genomic data is intended for research purposes, what happens if corporations access it for unintended purposes? Scholars of CSCW and STS have shed light on the dynamic nature of networks by studying infrastructure, which "represent complex sets of relationships embedded in and constrained by other systems, making it impossible to predict perfectly in advance what the infrastructure will be or how it will be used" (Bietz et al. 2010 p. 246). Infrastructures evolve through use: "because it means different things locally, it is never changed from above. Changes take time and negotiation, and adjustment with other aspects of the systems are involved. Nobody is really in charge of infrastructure" (Star 1999, p. 382). These perspectives reinforce the importance of studying the social, legal and communication-based elements necessary for genomics to effectively and ethically diffuse into the healthcare setting. With this in mind, we turn to look at networked privacy in clinical genomics.

3. Networked privacy and informational risk

The general public has become more comfortable sharing highly personal information online as the Internet has become increasingly enmeshed in everyday life over the past decade. In the early 2000s we disguised our identities behind pseudonyms and avatars and felt insecure about putting our credit card information online. Now we give out our credits cards regularly and post our names and photos across all sorts of social media platforms. Latour describes one of the effects of this social shift, which further exposes our subjectivities to the realm of empirical inquiry: "it is as if the inner workings of private worlds have been pried open because their inputs and outputs have become thoroughly traceable" (2007, p. 2). More and more of our personal data is accessible and can be downloaded, shared, analyzed and sorted. One significant consequence of the pervasive nature of information and communication technologies (ICTs) in contemporary society is that they have become part of an invisible infrastructure of daily life that only becomes noticeable when it breaks down (Boczkowski and Lievrouw 2008). Precisely because they are so embedded, convenient and reliable, ICTs make our personal data more prone to invasions of privacy in unforeseen ways.

Issues of privacy in clinical genomics may seem like a highly specialized area affecting a small percentage of the population. However, privacy of health information in the context of digital networks increasingly affects the vast majority of us who use the Internet, whether or not we are aware of it. For example, a recent study of over 80,000 health-related websites revealed that nine out of ten visits result in personal health information being leaked to third parties, including online advertisers and data brokers (Libert 2014). Why is this a problem? First, one's health interests and name may be linked and publicly identifiable. Second, many online databases and algorithmic tools are designed to sort web users into groups such as "target" and "waste," with serious ramifications in terms of insurability, employability and access to public or retail services.

Bowker and Star (2000) have exposed the informational risks involved when classification systems and algorithms are used as social tools for surveillance. In the network society (Castells 2000), personal information becomes much more significant. At times, having our personal preferences tracked can be experienced as a benefit, such as when Amazon recommends a book. However, the downside to tracking and sorting individuals by their data trails, is that certain people *pass* through various borders and barriers in society with greater ease than others. Frequent flier cards and customer loyalty cards are the visible tip of a personal information iceberg (Bennett et al. 2014).

Information privacy, typically based on "fair information practices," pertains to "communication control, that is, how far data subjects have a say over how their personal data are collected, processed, and used," (Lyon 2005, p. 19). Both Lyon and Fuchs (2013) argue that current legislation on privacy rights disproportionately favor financial privacy of the elite, while exploiting personal privacy of the masses. These advantages can go unnoticed by the average citizen because in Canada and the

United States discourses of privacy have historically been defined by individuals' rights, characterized by secrecy, anonymity, and solitude in personal matters (Gavison 1980; Solove 2008). In contrast, countries such as Iceland have a well-established communitarian approach to privacy and biotechnological innovation. In 1998, the Icelandic government partnered with deCode Genetics to map the genome of the nation as part of a broad computerized medical database (Pálsson and Rabinow 2001). But, as open access genomic databases continue to develop globally, we are forced to redefine traditional notions of privacy created in earlier techno-social contexts, and consequently, address new informational risks. The reality is that in today's context of the network society, personal data is "public by default, private by effort" (Boyd 2010). Rainie (2015) further clarifies that privacy is not a simple binary—context matters, personal agency matters, and various trade-offs are increasingly becoming part of the bargain of big data innovation.

The future of data is characterized by "networked privacy," and it is therefore important to shift toward "a model that focuses on usage and interpretation. Who has the ability—and the right—to interpret what data they think they see?" (Boyd 2012, p. 349). These new sensibilities will require a solid understanding of the Internet, databases and the social shaping of technology (MacKenzie and Wajcman 1999). Nissenbaum (2010) argues that social and contextual-relative information norms can help guide our thinking about privacy in a networked society. Early regulation on human research, based on the Tuskegee experiments of the 1970s, focused mainly on protecting participants against physical and psychological harm, and thus informational risk is only mentioned tangentially (Hudson 2011). Informational risk refers to the economic and social consequences involved in making one's private data public, the impact of incidental findings that may also affect one's family members, and the long-term participatory risks for research participants (Allyse et al. 2012).

As STS scholars have shown, when a new technology moves from a small group of expert users into another context, in this case a population-wide health care system with a broad set of stakeholders including patients and clinicians, new challenges arise (Hackett 2008). For instance, the gold standard for privacy in health sciences has traditionally been the de-identification of personal data and, with genomic data, aggregation. However, advanced big data techniques of advanced algorithms and database linkage challenge the security of this method of anonymization. In a key study, Homer et al. (2008) used high-density single nucleotide polymorphism (SNP) genotyping microarrays from a complex DNA mixture to correctly re-identify individual identities, and to demonstrate privacy limits through simulation. The implications of their findings show that de-anonymization of genetics data from a database is possible by using simple allele frequencies or genotype counts. A similar study was conducted by Israeli scientists which demonstrated that DNA evidence can be fabricated for manipulating a crime scene or medical records (Pollack 2009). Schadt et al. (2012) also released a study proving that it is possible to confirm the identity of a study participant from public genetic data, if one already knows the

person's genetic makeup. In short, genome-wide association studies cannot completely conceal identities of patients or subject participants.

The studies outlined above also demonstrate the fact that in the context of genomics, it is nearly impossible to gain patient consent in the traditional sense, since the potential uses of one's data cannot be identified at the time of collection. For instance, privacy concerns were recently raised over the UK's 100,000 Genome Project which is the first phase of the projected 50 million Genome Project, a national genomic database of all National Health Services patients in England and Wales (Hockings and Coyne 2015). Like many open access databases, the project claims to make anonymous clinical and genomic data available to academics, researchers, and industry members, yet a Freedom of Information request to the Department of Health clarified that in fact pseudonymized data would be available to third parties including commercial entities.

Data breaches present another informational risk to the privacy of digital genomic data. Since October 2009, health and business organizations have reported 1142 large-scale data breaches to the U.S. Department of Health and Human Services, affecting at least 500 people, and of those only seven cases have resulted in fines (Wei et al. 2015). The US National Institute of General Medical Sciences (NIGMS) reacted to a breach of their database by removing the data from public view, but computational biologist Eric Schadt says such measures do not constitute a real solution to the privacy issues at hand; instead, "we should be up front with participants that we can't protect their privacy completely, and we should ensure that the most appropriate legislation is in place to protect participants from being exploited in any way" (Check Hayden 2013). This is an increasing concern, especially as black markets for medical data emerge (Pasquale 2015a).

4. Spaces of convergence

Genomics is a disruptive technology in the clinical work and healthcare setting. It requires scientists and doctors to work together in a new space of convergence, not just alongside one another but to co-produce results. This culture of collaboration poses a challenge: while these groups share an epistemological background, they differ in organizational culture and goals. For instance, the failure of a scientific experiment may represent progress in the discovery process. Whereas the failure of an oncologist's prescribed treatment can result in the death of a patient. Genomics is a new type of medical literacy that doctors and health workers are grappling to understand and integrate into their work practices. A recent survey of 329 BC physicians showed that despite believing genomic knowledge is of great importance, 67.8 % assessed their own knowledge of genomic technologies as poor (Friedman 2013). Genome scientists need to collaborate with doctors to understand what information is useful and actionable. In this sense, the communication between the groups is a critical technology.

Since we included both researchers (non-physician) and clinical researchers (dual role as researcher and physician) in our sample population, we captured the blurring of this boundary as patient samples crossed between these jurisdictions. Our data illustrated that this divide is very significant for collaboration. For example, a clinical researcher may collaborate with a genomic scientist (researcher) thereby allowing the researcher to analyze the patient samples collected for a clinical study. One genomic researcher who works in a team with clinicians explained,

> The separation between the clinic and the research, the church and the state, is an uncomfortable space in which we have to operate. There are, I think, very blurry lines that start to interfere with our recognition of what is what. And not everyone agrees on what is the definition. So I have a debate with a surgeon colleague all the time, [in which] I say, "You know, what we are doing for research is not clinically relevant," and I get a lot of feedback on that. (1027, researcher).

This clinician identifies the gap between the clinical setting and the research setting as an "uncomfortable" space, which illustrates the uncertainty in the space of convergence between different institutional cultures. As Pool (1983) reminds us, convergence is always pushing different organizational actors together while producing a dynamic tension towards change. This uncertainty often arises when determining what type of variation is valid and relevant for clinical practice. In clinical genomics, for instance, the two camps work through an iterative feedback process where colleagues co-construct the facts and meanings of research innovations. Their interaction is not simply instrumental. They become inherently interdependent.

5. Methodology

The goal of the AML project was to create a genome-based test for an acute form of cancer, to understand the social and economic issues for deploying the test in a public health system, and to develop strategies for addressing professional and social risks. The funder, a non-profit publically supported research organization, mandated this type of collaborative approach. We worked closely with a diverse project group to identify and understand issues that connect the different sub-teams (clinical, genomic, economic, communication). Our own team focused on understanding the potential social and professional benefits, challenges, and risks for key stakeholders in the diffusion and uptake of the new technology. We use the word "potential" because genomic tests had not been used in the healthcare system in Canada at the time. However, scholars and policy makers have been discussing issues such as privacy for some time. We focused on the clinicians who would be the front line users of the technology, healthcare decision makers at the local and provincial levels who would enable or constrain its adoption, and the provincial privacy commissioner who would be a key regulator of data and information management. We found a need

to understand the role of the Internet, digital databases, and the social shaping of information technologies in the operationalization of scientific discoveries in public healthcare. The test is now in use and is the first of its kind in Canada.

We identified 98 active clinical genome researchers in the province of British Columbia. We excluded those who work solely on animals, contacted the 67 remaining via email to request an interview, and ultimately interviewed 36 people. In addition to this, we interviewed seven policy officials' for a total of 43 interviewees. Over half of the interviewees (57 %) use whole genome sequencing in their projects and all used genomics in their research and/or practice. We conducted semi-structured interviews based on a series of questions we developed from the literature and our project experiences with genome and clinical experts and tested during the initial pilot interviews. We conducted the majority of the interviews face-to-face and four interviews over the phone when a face-to-face situation was not possible. Shortly after completing the interview, the researcher reviewed her notes and made annotations for issues and items that could be addressed in subsequent interviews or analysis. After transcribing the interviews verbatim, the interviewer checked the transcripts against the recordings for accuracy. Finally, we assigned the interviewees numbers in order to anonymize their identities when reporting their quotes.

We conducted a qualitative analysis of the interviews for expected and emergent themes as well as generated descriptive statistics. Quantitative data analysis involved turning words into numbers to understand general trends and inform qualitative analysis. The process involved two individuals coding independently to ensure a high level of intercoder reliability, which tested at over 90 %. A team of four researchers developed content categories using an iterative process of reviewing data and literature, which was subsequently tested and validated. When disagreements could not be resolved, the coders consulted the principle investigator to develop a consensus. The researcher then used a qualitative analysis software program, *NVivo*, to code the interviews. The resultant themes were then presented to our research team and project leader for further review and discussion. The most pressing topics for analysis, as determined by the literature and the interviews, included informed consent, return of research results, and incidental findings.

6. Findings

6.1. Managing privacy in informed consent

One of the most pressing challenges we found for clinical genomics revolves around the process of gathering informed consent. As outlined above, the Internet plays a significant role in clinical genomics as a benefit to the development of personalized medicine, a risk to individual privacy, and a challenge to the management of patient information in the healthcare system. Despite its central role in genomic medicine, we found that most stakeholder discussions fail to address the role of the Internet and information technologies with patients. In particular, we found a complete lack of

direct references to the Internet within information for patients, including informed consent documents and education manuals.

The traditional bioethical model for privacy of patients positions anonymization and informed consent as two sides of the same coin. However, with clinical genomics it is nearly impossible to outline and attain genuinely informed consent since health data in open access databases may be used in numerous unforeseen ways. As outlined above, these infrastructures represent a space of convergence – each time data is uploaded, something new is created, and no person or process is in charge of when or how this information may be accessed or used. As one researcher explained: "We usually have a very broad [consent] where we say we will use the material for genomic research related to lung cancer, for example, because it's hard to predict what information will come out of it in the future" (1014, clinical researcher). Another challenge we found is that the majority of interviewees felt that the forms are too long and not written in accessible language that most research participants understand. Yet participants, especially those with diseases they want researched and treated, are eager to participate in potentially life-saving research, and genomics represents an exciting new avenue: "Patients like technology so they like the idea that they're getting something new and cutting-edge" (1004, researcher).

Patients with incurable diseases (by conventional treatment) may enter clinical trials offering alternative or experimental treatments. Genomic profiling may be part of these studies. One clinical researcher expressed how motivated patients could be in these situations,

> Because patients, as soon as they understand that we don't know which drugs will work for them, they are highly motivated to help us figure out which drugs might work for them. And the problem of course is this information may or may not help us. But this is the only way to figure that out. So our patients are highly motivated. And so are clinicians because we all learn a tremendous amount. (1024, clinical researcher)

Interviewees also expressed the over-hyping of genomics in the media and how patients may misunderstand what the information can do for them. One researcher working with clinicians remarks, "Oh, nobody understands what it means. They get the idea that these fancy machines read the DNA of cancer and these guys have a fancy way to find the mistake," (1027, researcher).

The very idea of "biotechnology" can be confusing to the general public as it "often encompasses existing, emerging and imaginary scientific techniques" (Gerlach and Hamilton 2005, p. 80). We can see the lack of knowledge of informational risk in one study of patients with diabetes, in which participants were more concerned with the privacy of their physical tissue samples entering biobanks than with their digital health data entering an online registry, despite similarities in their purposes and long-term uses (Gibson et al. 2008). The physical samples were seen as more tangible, bearing more serious privacy implications, where the digital data was

seen as more anonymous. One of our interviewees reported a similar example, contrasting a patient's contradictory perception of risk in sharing her health information online:

> I think right now we still have a lot of old fashion concepts of privacy and we're trying to apply those same rules now and it may not work. There's a case now—say somebody has TB, a young person, and she didn't want anyone to know she had TB. So she wanted all the nurses to treat her [privately]. "Don't treat me at school, don't show up, don't talk to me, you can hand me the drug and I'll sneak it in and nobody will see it." But then at same point she went on Facebook and told everybody, "I have TB and am taking all these gross drugs." I was like, people don't really understand what privacy means. We really don't understand what should be private. And it's changing so quickly that I think we're caught in two different worlds. (1011, researcher).

This case exemplifies the problem of digital dualism (Jurgenson 2012), treating the online and offline worlds as discrete realms, when in fact information technologies are so pervasively enmeshed in contemporary society that this notion no longer holds water. The online and offline spheres are complexly intertwined in continuous, bi-directional feedback loops, and most web users remain unaware of the potential ramifications. The hidden complexity of Internet algorithms can anesthetize the public to informational risk. They can also mask exploitation by "extend[ing] an open invitation for quants or traders or managers to bully their way past gatekeepers, like rating agencies, accountants, and regulators" (Pasquale 2015a, p. 137).

Based on our findings, we developed recommendations for best practices on managing privacy in terms of informed consent. Underlying each is the need for increased transparency in the data collection process for patients, researchers and practitioners. First, researchers should develop a proactive consent process that stresses risks and benefits of digital genetic information. Hand in hand with this, medical practitioners require a greater understanding of the digital pathways genomic data may take. We found that even some of the top oncologists on the front lines of genomic testing were not aware of which data may enter open access databases. Second, the role of the Internet and digital information should be clearly emphasized in the process of sharing, managing and storing of genetic data. This may seem like a given nowadays, but among the dozens of consent documents we analyzed including the one used in the AML project, none of them mentioned the Internet at all. Third, consent forms should contain a clear disclaimer that privacy is not absolute. Details of data release and the potential for re-identification of anonymized data should be included. Fourth, consent forms should note that patients cannot always exercise the choice to withdraw their data. While it is generally an option for research participants to withdraw from clinical research, once genomic data is uploaded to a database in a clinical trial, it cannot always be withdrawn. Finally, researchers, organizations and companies ought to be transparent in the use of patients' genetic information,

whether it will be shared with another third party or where and for how long biological and digital samples will be stored for research purposes.

In the AML project, we have argued that the frame of reference in patients' minds when they consider their privacy is closer to Facebook, not biomedical history. This is problematic, because Facebook's privacy controls do not represent informed consent. Rather, they are user agreements to gain access to the site, with nothing preventing Facebook from reversing the original terms of agreement in place when a user joined. Indeed, Facebook has done this on a number of occasions in the past. Open access genomic databases represent a similar challenge to personal privacy since its various future uses remain unknown. Scholars have proposed various solutions to this. On a macro level, Kosseim et. al. suggest a data sharing model for global genomic research based on international business networks that are "flexible, externally endorsed, multilateral arrangements, combined with an objective third-party assurance mechanism" (2014, p. 430). On a meso level, Kaye (2012) advocates for an e-governance system to complement existing legislation, and make better use of available technologies to ensure compliance with ethical and legal requirements. On a micro level, Pasquale (2015b) recommends that "genetic data companies should set aside 10 to 20 % of revenues from data sales to compensate victims of information breaches if they occur." In other words, genomic research organizations and companies should be held responsible for the misuse of genetic information against its participants.

6.2. Return of results and incidental findings

One of the major tensions we identified in the AML project was the right to know versus the right *not* to know in terms of the return of results and incidental findings. Incidental findings are discoveries of genomic conditions that may cause disease, unrelated to the original clinical tests. The emerging consensus from the interview data and the literature is to return results, including incidental findings, if they are material, scientifically valid, and clinically actionable. In particular, we found that specific reasons to return results to include: result indicates a change in treatment, result answers clinical question, or result is asked for by the patient. The literature suggests that whether a result is returned should be based on the ACCE model (CDC 2010), which states results should be evaluated on four criteria: 1) Analytic validity: how accurately and reliably it measures the results; 2) Clinical validity: how consistently and accurately the test can detect or predict outcomes; 3) Clinical utility: how likely the result may significantly improve health or health related decisions made by the participant; and 4) Ethical, legal, and social implications (ELSI) (Lévesque et al. 2011). However, this does not solve the problem of who decides whether the results meet this criteria or how to actually return the results.

We found one of the major obstacles expressed by clinicians and researchers was a lack of consensus of whether or how to return results. This is a significant issue to be resolved as genomics shifts from the bench to the bedside. Researchers are not

obligated to return any results to participants. Whereas doctors work under a different ethical regime guided by the Hippocratic Oath that requires them to prioritize patient care and act on any medical information that has a clinical action. The problem right now is that there is very little genomic information that has clear and agreed-upon clinical action. So many decisions about the return of results are currently made on an *ad hoc* basis which highlights the need for agreed-upon protocols in this new space. In some cases, the "return of results" and "the right to know" was positive and welcomed by patients. Some of our interviewees shared encouraging medical breakthroughs facilitated by genomic databases, especially in cases of rare or undiagnosed conditions:

> It's important because families often feel like there's nobody else in the world like that—that their kid is just immensely weird. And finding a cause often means you can find other affected patients and it's pretty amazing. One of the conditions that one of my graduate students found was the first case [of an extremely rare condition]. She recognized the second case clinically and tested it and showed the same thing. And a third case was recognized and referred to her. Those three were published. She's had about a half dozen other cases referred to her from all over the world because she's the only one who's written anything about this. So for those families, knowing that there are just 2 or 3 others around the world, they have asked to be put in touch with each other. (1026, clinical researcher).

In other cases, returning incidental findings turned into an unwelcome situation that reflects the "right not to know" position:

> When I first started in medical genetics, a friend of mine who is now a retired medical geneticist, had discovered in a family in Vancouver a particular allele of a gene with devastating consequences. And he learned that the rest of the family were in Michigan. This was 1966. So he flew to Michigan out of his own pocket, to visit the family and explain what he'd found. Because of the genetics tree certain members of the family there were carriers of that gene. They might want to be tested to see what their status was. [Depending on your feelings], it might make no difference if your future kids might have muscular dystrophy, or it might make a big difference. And he got sued. He got sued for revealing information to them that they didn't want to know. (1022, researcher).

Researchers expressed clear reasons for not returning research results (including incidental findings). Research involves investigating genomic questions that have yet to be solved. Often data is aggregated and researchers are looking for trends, not individual results. The consent document clarifies the use of data and returning research results might not be ethically appropriate. A researcher may analyze a sample for a specific mutation in a research study, but returning the mutation status to the sample donor may be inappropriate since the result is not clinically validated or

necessarily linked to a medical diagnosis. One exception seemed to be researchers involved in population health who have access to anonymized genomic data to look for trends in a group of samples. If they find a result that needs to be communicated to the individual, there is a protocol in place to inform the individual's physician. One researcher explains,

> We're looking at the population and you're basically looking for trends and then as I pointed out we have those consents where there's no consent to the patient, but the approval enables us to re-link the patient if necessary. And there's a process to do the re-linking and that process is outside of my hands. That would go up to the medical health officer, but I would clearly have the responsibility to highlight that there is a finding that is suspicious. (1006, clinical researcher).

In contrast, clinical researchers navigate a very different ethical paradigm. As physicians, they have an ethical responsibility to do what is best for the patient. The genomic samples they use are de-identified for research purposes but can be re-linked if necessary. This allows the possibility of returning research results and incidental findings. One clinical researchers explains,

> So all of our information is reversibly de-identified. It's rendered anonymous, in a reversible fashion. So there is always a master key, that allows us to reconnect any data, proportion of the data set, with the individual from which it came because that's necessary in doing any kind of clinically connected work because [what] gathers additional information about people with the passage of time that you may need to relink to the information. (1002, clinical researcher).

Clinical researchers expressed three reasons in favor of returning research results. First, if the result can change treatment for the patient it is imperative to communicate the result. Second, if the result is part of the clinical question then it would naturally be returned. Third, many clinicians indicated patients often ask for results and they are either required to (by funder or consent agreements) or desire to fulfill the request. Some highly knowledgeable patients ask for specific information, especially those with rare diseases who have become advocates for themselves. The detail of result given depends on the patient as described by one clinical researcher: "So there's some patients who want to go through all the different permutations and there are some that just don't have that level of sophistication and they just want to know the basics of what we came up with so we have a meeting about the results" (1024, clinical researcher).

At the start of the AML project in 2011, we found the discussion in the literature and among our interviewees focused on the issues of whether or not to return results and incidental findings: "There's not one consensus on how data should be explained or discussed with a patient or how much or what to do with germline mutations and all that. So it's an ongoing discussion and we have a lot of discussion about it" (1024,

clinical researcher). Just a few years later as we concluded our data collection, the conversation had notably shifted from whether or not to return the results, toward the need for guidelines and processes to help determine which results and findings are scientifically valid, clinically useful, and actionable. One clinical researcher discusses his thoughts on returning results:

> Well again, if there's a pipeline in place then yes we should. It's my opinion that we should. But that has to be done in an ethically sustainable manner. It has to be done…it's not up to me to decide, it's up to the team to decide. There has to be policy put into place for these types of things (1010, clinical researcher).

Based on our findings regarding the dilemma of return of results and incidental findings, we recommended the development of a "green-yellow-red" light decision-making matrix. "Green" class findings includes the results with both clinical utility and scientific validity; and thus, it should be returned. "Yellow" results could have scientific validity but no consensus on clinical utility. Therefore, clinical researchers or medical practitioners should take this "yellow" class result into discrete consideration. "Red" class results have no scientific validity nor clinical utility and should not be returned. In addition, we put forth a recommendation to develop a multi-stakeholder network and a multidisciplinary effort in British Columbia to address the limitations of current healthcare and privacy policies, as well as the potential need for new health information and data guidelines.

7. Conclusion

The expansion of genomic big data presents new challenges for medical practitioners, researchers, policymakers and the public, as it introduces a disruptive new type of personal information in the healthcare system. Both medical innovation and personal privacy hang in the balance—not as opposing forces, but as equally important factors to be addressed in the socio-technical gap of clinical genomics. As technological systems, social requirements and medical advances collide in spaces of convergence, solutions developed today will not be seen as adequate tomorrow. We have outlined some of the informational risks to do with genomic data above. In closing, we also highlight the risks to medical innovation if sustainable practices in genomic data sharing fail to develop. For example, "between 2011 and 2013, a network of Canadian geneticists uncovered the precise molecular causes of 146 conditions, solving 55 % of their undiagnosed cases" (Regalado 2015). But success rates are now plateauing for more obscure cases, which require broader systematic data sharing.

On first look, the informational risk associated with genomic big data is similar to other types of personal information. Privacy is a major concern for handling any sensitive medical data. Yet the information-sharing model of genomics has further-reaching implications than traditional electronic health records. Researchers wanting

to treat genomic data as a public good set up open access databases. This approach facilitated genomic research globally and enabled one of the original goals of the HGP specifically and academic and public researchers more generally of open access genomics. More recently, however, researchers have identified some of the unforeseen risks for individuals because of open access databases due to technological advances. Computers have become faster and smarter due to innovations by programmers in data mining and database linkage. Our research uncovered another information risk issue, which merits future research. Scientists and medical practitioners are working with data that is highly uncertain. In the translation process, scientists and doctors are negotiating the validity of genomic variation. They collaborate to identify genomic variation, understand which ones are meaningful, and which of those are actionable clinically and should be communicated to the patient. This type of risk does not fit typical privacy or surveillance discussions. However, it is of equal importance and may be more critical to the effective adoption of genomic technologies hospitals and clinics.

Unresolved issues in the diffusion of clinical genomics include re-identification of anonymized samples, determining effective privacy protection measures for downstream data, and creating effective processes of informed consent to allow scientists and clinicians to realize the full potential of genomic data while still respecting participants' privacy and autonomy. Other challenges include the dissemination of research results and incidental findings such as what to return, who should return, who is liable, and whether results should be returned to family members. In light of our findings in the AML project, we highlight the need for greater transparency with patients, acknowledging the role of the Internet around informed consent. We also recommend the development of a systematic "green-yellow-red" decision scheme to guide the return of results and incidental findings. These steps will begin to address the socio-technical gap of informational risk around clinical genomics, by acknowledging the way networked privacy challenges the traditional biomedical approach.

Acknowledgments

Funding for this research was provided by a grant from Genome British Columbia Personalized Medicine Program (Grant #121AML PMP).

References

1000 Genomes. (2015). Retrieved June 14, 2015, from http://www.1000genomes.org/data

Ackerman, Mark S. (2000). The Intellectual Challenge of CSCW: The Gap between Social Requirements and Technical Feasibility. *Human–Computer Interaction*, vol. 15, nos. 2–3, 179–203.

Allyse, Megan, Katrina Karkazis, Sandra Soo Jin Lee, Sara L. Tobin, Henry T. Greely, Mildred K. Cho, and David Magnus (2012). Informational Risk, Institutional Review, and Autonomy in the Proposed Changes to the Common Rule. *IRB: Ethics and Human Research*, vol. 34, no. 3, pp. 17–19.

Bennett, Colin J, Kevin D. Haggerty, David Lyon, and Valerie Steeves (2014). *Transparent Lives: Surveillance in Canada*. Edmonton, AB: Athabasca University Press.

Bietz, Matthew J, Eric P. S. Baumer, and Charlotte P. Lee (2010). Synergizing in Cyberinfrastructure Development. *Computer Supported Cooperative Work (CSCW)*, vol. 19, nos. 3–4, 245–281.

Blumenthal-Barby, Jennifer S., Amy L. McGuire, Robert C. Green, and Peter A. Ubel (2015). How behavioral economics can help to avoid "The last mile problem" in whole genome sequencing. *Genome Medicine*, vol. 7, no. 1, art. 3.

Boczkowski, Pablo, and Leah A. Lievrouw (2008). Bridging STS and communication studies: Scholarship on media and information technologies. In Edward J. Hackett, O. Amsterdamska, M. Lynch, & J. Wajcman (Eds.), *The Handbook of Science and Technology Studies* (pp. 949–977). Cambridge, MA: MIT Press.

Bowker, Geoffrey C. (2005). *Memory practices in the sciences*. Cambridge, MA: MIT Press.

Bowker, Geoffrey C., and Susan Leigh Star (2000). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Boyd, Danah. (2010). Making Sense of Privacy and Publicity. http://www.danah.org/papers/talks/2010/SXSW2010.html, accessed 10 March 2015.

Boyd, Danah. (2012). Networked Privacy. *Surveillance & Society*, vol. 10, nos. 3/4, pp. 348–350.

Burn-Murdoch, John (2012, Oct. 26). Big data: what is it and how can it help? http://www.theguardian.com/news/datablog/2012/oct/26/big-data-what-is-it-examples, accessed 12 March, 2015.

Bush, William S., and Jason H. Moore (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Comput Biol*, vol. 8, no. 12, e1002822.

Castells, Manuel (2000). *The Rise of the Network Society: The Information Age: Economy, Society, and Culture*. Malden, MA: Blackwell.

Caulfield, Timothy A., and Bartha Maria Knoppers (2010). *Consent, privacy & research biobanks (Policy Brief No.1)*. Genome Canada. http://www.genomecanada.ca/medias/pdf/en/GPS-Policy-Directions-Brief.pdf

CDC, Public Health Genomics at. (2010). Genomics|Genetic Testing|ACCE. http://www.cdc.gov/genomics/gtesting/ACCE/index.htm, accessed 14 March, 2015.

Check Hayden, Erika (2013). Privacy loophole found in genetic databases. *Nature News*.

Chow-White, Peter A. and Miguel García-Sancho (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*, vol. 37, no. 1, 124–164. 10.1177/0162243910397969

Chow-White, Peter A. and Sandy Green, Jr. (2013). Data Mining Difference in the Age of Big Data: Communication and the Social Shaping of Genome Technologies from 1998 to 2007. *International Journal of Communication*, 7(0), 28. Retrieved from http://ijoc.org/index.php/ijoc/article/view/1459, accessed 17 June, 2014.

Collins, Francis S., Michael Morgan, and Aristides Patrinos (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science*, vol. 300, no. 5617, pp. 286–290.

Condit, Celeste M. (2007). How Culture and Science Make Race "Genetic": Motives and Strategies for Discrete Categorization of the Continuous and Heterogeneous. *Literature and Medicine*, vol. 26, no. 1, pp. 240–268.

Condit, Celeste Michelle, Roxanne L. Parrott, Tina M. Harris, John Lynch, and Tasha Dubriwny (2004). The Role of "Genetics" in Popular Understandings of Race in the United States. *Public Understanding of Science*, vol. 13, no. 3, pp. 249–272.

Contreras, Jorge L. (2010). *Bermuda's Legacy: Policy, Patents and the Design of the Genome Commons* (SSRN Scholarly Paper No. ID 1667659). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=1667659

CTV (2015). Blood pressure drug shrinks cancer in "miracle" clinical trial. http://bc.ctvnews.ca/blood-pressure-drug-shrinks-cancer-in-miracle-clinical-trial-1.2271586, accessed 13 March, 2015.

Friedman, Jan (2013). The UBC Medical Curriculum and the Genomic Revolution. *University of British Columbia Medical Journal*, vol. 4, no. 2. www.ubcmj.com/pdf/ubcmj_4_2_2013_6-8.pdf

Frizzo-Barker, Julie, and Peter A. Chow-White (2014). From Patients to Petabytes: Genomic Big Data, Privacy, and Informational Risk. *Canadian Journal of Communication*, vol. 39, no. 4.

Fuchs, Christian (2013). *Social Media: A Critical Introduction*. London: SAGE.

Gandy, Oscar H. (1993). *The Panoptic Sort: A Political Economy of Personal Information*. Boulder, CO: HarperCollins Canada.

Gavison, Ruth. (1980). Privacy and the Limits of Law. *The Yale Law Journal*, vol. 89, no. 3, pp. 421–471.

Gerlach, Neil, and Sheryl N. Hamilton (2005). From Mad Scientist to Bad Scientist: Richard Seed as Biogovernmental Event. *Communication Theory*, vol. 15, no. 1, pp. 78–99.

Gibson, Elaine, Kevin Brazil, Michael D. Coughlin, Claudia Emerson, Francois Fournier, Lisa Schwartz, Karen V. Szala-Meneok, Karen M. Weisbaum, and Donald J. Willison (2008). Who's minding the shop? The role of Canadian research ethics boards in the creation and uses of registries and biobanks. *BMC Medical Ethics*, vol. 9, no. 1, art. 17.

Gordon, Dan, and Aditya Pai (2012). BIG data, BIG opportunity. *Canadian Healthcare Manager*, vol. 1, no. 2, pp. 25–27.

Gulland, Anne. (2010). Project to decode genomes in cancer samples promises new treatments. *BMJ*, vol. 340, c2149.

Hackett, Edward J. (2008). *The Handbook of Science and Technology Studies*. Cambridge, MA: MIT Press.

HapMap Project. (2015). http://hapmap.ncbi.nlm.nih.gov/, accessed 14 June, 2015.

Hockings, Edward, and Coyne, Lewis (2015). Privacy and the 100,000 Genome Project. The Guardian. Retrieved from https://www.theguardian.com/science/political-science/2015/mar/10/privacy-and-the-100000-genome-project, accessed 19 Febuary 2015.

Homer, Nils, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig (2008). Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. *PLoS Genetics*, vol. 4, no. 8.

Hudson, Kathy. L. (2011). Genomics, Health Care, and Society. *New England Journal of Medicine*, vol. 365, no. 11, pp. 1033–1041.

IBM (2012). What is big data? http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, accessed 12 March 2015.

ICGC: International Cancer Genomics Consortium Data Portal (2015). https://dcc.icgc.org/, accessed 14 June 2015.

Jirotka, Marina, Charlotte P. Lee, and Gary M. Olson (2013). Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW)*, vol. 22, nos. 4–6, pp. 667–715.

Jurgenson, Nathan (2012). When Atoms Meet Bits: Social Media, the Mobile Web and Augmented Revolution. *Future Internet*, vol. 4, no. 1, pp. 83–91.

Kaye, Jane (2012). The Tension Between Data Sharing and the Protection of Privacy in Genomics Research. *Annual Review of Genomics and Human Genetics*, vol. 13, no. 1, pp. 415–431.

Kitchin, Rob (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (1st edition). Thousand Oaks, CA: SAGE Publications Ltd.

Kosseim, Patricia, Edward S. Dove, Carman Baggaley, Eric M. Meslin, Fred H. Cate, Jane Kaye, Jennifer R- Harris, and Bartha M. Knoppers (2014). Building a data sharing model for global genomic research. *Genome Biology*, vol. 15, no. 8.

Latour, Bruno. (2007). *Beware your imagination leaves digital traces*. Times Higher Literary Supplement. http://www.bruno-latour.fr/node/245

Lévesque, Emmanuelle, Yann Joly, and Jacques Simard (2011). Return of Research Results: General Principles and International Perspectives. *The Journal of Law, Medicine & Ethics*, vol. 39, no. 4, pp. 583–592.

Libert, Tim (2014). *Privacy Implications of Health Information Seeking on the Web* (SSRN Scholarly Paper No. ID 2423006). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2423006

Lunshof, Jeantine E., Ruth Chadwick, Daniel B. Vorhaus, and George M. Church. (2008). From genetic privacy to open consent. *Nature Reviews Genetics*, vol. 9, no. 5, pp. 406–411.

Lyon, David (2005). Surveillance as social sorting: computer codes and mobile bodies. In *Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination* (pp. 13–30). London: Routledge.

MacKenzie, Donald, and Judy Wajcman (Eds.). (1999). *The Social Shaping of Technology* (2nd edition). Buckingham Eng.; Philadelphia: McGraw Hill Education / Open University.

Manovich, Lev (2001). *The Language of New Media*. Cambridge, MA: MIT Press.

Marris, Emma (2005). Free genome databases finally defeat Celera. *Nature*, vol. 435, no. 7038, 6–6.

Mayer-Schönberger, Viktor, and Kenneth Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Eamon Dolan/Houghton Mifflin Harcourt.

McLuhan, Marshall (1964). *Understanding Media: The extensions of man.* New York, NY: McGraw Hill.

Nissenbaum, Helen F. (2010). *Privacy in context: technology, policy, and the integrity of social life*. Stanford, CA: Stanford Law Books.

Pálsson, Gisli and Paul Rabinow (2001). The Icelandic genome debate. *Trends in Biotechnology*, vol. 19, no. 5, pp. 166–171. doi:10.1016/S0167-7799(01)01607-9

Pasquale, Frank (2015a). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Pasquale, Frank. (2015b, March 2). Can a Genetic Test Be Anonymous? *The New York Times*. http://www.nytimes.com/roomfordebate/2015/03/02/23andme-and-the-promise-of-anonymous-genetic-testing-10/insure-people-against-genetic-data-breaches

Pollack, Andrew (2009, August 18). DNA Evidence Can Be Fabricated, Scientists Show. *The New York Times*. http://www.nytimes.com/2009/08/18/science/18dna.html

Pollack, Andrew (2011, November 30). DNA Sequencing Caught in Deluge of Data. *The New York Times*. http://www.nytimes.com/2011/12/01/business/dna-sequencing-caught-in-deluge-of-data.html

Pool, Ithiel de Sola (1983). *Technologies of freedom*. Cambridge, MA: Belknap Press.

Presidential Commission for the Study of Bioethical Issues (PCSB) (2012). *Privacy and Progress in Whole Genome Sequencing*. http://bioethics.gov/sites/default/files/PrivacyProgress508_1.pdf

Presidential Commission for the Study of Bioethical Issues (PCSB) (2013). *Anticipate and Communicate: Ethical Management of Incidental and Secondary Findings in the Clinical, Research and Direct-to-Consumer Contexts.* bioethics.gov/sites/default/files/FINALAnticipateCommunicate_PCSBI_0.pdf

Rainie, Lee (2015). Networked Privacy in the Age of Surveillance, Sousveillance, Coveillance. http://www.pewinternet.org/2015/01/23/networked-privacy-in-the-age-of-surveillance-sousveillance-coveillance/, accessed 2 February, 2015.

Regalado, Antonio (2015 2–18). Networks of Genome Data Will Transform Medicine. http://www.technologyreview.com/featuredstory/535016/internet-of-dna/ , accessed 8 March, 2015.

Rukovets, Olga (2014). FDA to 23andMe: "Stop Marketing Genetic Tests." *Neurology Today*, vol. 14, no. 2, pp. 1,11–14.

Schadt, Eric E., Sangsoon Woo, and Ke Hao (2012). Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics*, vol. 44, no. 5, pp. 603–608.

Schuurman, Nadine, and Ellen Balka (2008). alt.metadata.health: Ontological Context for Data Use and Integration. *Computer Supported Cooperative Work (CSCW)*, vol. 18, no. 1, pp. 83–108.

Solove, Daniel J. (2008). *Understanding privacy.* Cambridge, Mass: Harvard University Press.

Star, Susan Leigh (1999). The Ethnography of Infrastructure. *American Behavioral Scientist*, vol. 43, no. 3, pp. 377–391.

Strasser, Bruno J. (2012). Data-driven sciences: From wonder cabinets to electronic databases. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, vol. 43, no. 1, pp. 85–87.

Thacker, Eugene (2004). *Biomedia* (1st edition). Minneapolis: Univ. of Minnesota Press.

Thacker, Eugene (2005). *The Global Genome: Biotechnology, Politics, and Culture* (1st edition). Cambridge, Mass: The MIT Press.

Wei, Sisi, and Charles Ornstein (2015). Over 1,100 Health Data Breaches, but Few Fines. *ProPublica*, 27 February 2015 https://projects.propublica.org/graphics/healthcare-data-breaches, accessed 7 March, 2015.

Williams, Raymond (1975). *Television: Technology and cultural form*. New York, NY: Schocken Books.

Wright, Caroline, Hilary Burton, Alison Hall, Sowmiya Moorthie, Anna Pokorska-Bocci, Gurdeep Sagoo, Simon Sanderson, and Rosalind Skinner (2011). *Next steps in the sequence: The implications of whole genome sequencing for health in the UK*. PHG Foundation. http://www.phgfoundation.org/reports/10364/.

Zhao, Jun, Oscar Corcho, Paolo Missier, Khalid Belhajjame, David Newmann,David de Roure and Carole A. Goble (2011). eScience. In John Domingue, Dieter Fensel, and James A. Hendler (Eds.), *Handbook of Semantic Web Technologies* (pp. 701–736). Springer Berlin Heidelberg.