The Ethics of Large-Scale Genomic Research

Benjamin E. Berkman, Zachary E. Shapiro, Lisa Eckstein and Elizabeth R. Pike

1 Introduction

Over the past few years, there has been a dramatic evolution of our technological ability to gather and share information. This has enabled the collection, distribution, and analysis of vast amounts of data, in ways never before possible. While there has not been a distinct watershed moment, this type of increasingly large data collection has come to be known as "big data," and is defined by the National Science Foundation as involving "large, diverse, complex, longitudinal, and/or distributed data sets generated from instruments, sensors, Internet transactions, email, video, click streams, and/or all other digital sources available today and in the future." (National Science Foundation 2010).

While big data has applications in many fields, its potential in the biomedical research settings is among the most exciting. Through the creation of big data health repositories, researchers are able to gather information from a multitude of clinical and research sources, greatly expanding the breadth of their data, while allowing them to more widely share information with other researchers (Bollier and Firestone 2010). This enables researchers to study conditions in an entirely new way, ideally allowing advances to be made more quickly. Crucially, big data facilitates more

B.E. Berkman
Department of Bioethics, Clinical Center, and Bioethics Core, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA

Z.E. Shapiro
Harvard Law School, Cambridge, Massachusetts, USA

L. Eckstein
The Faculty of Law, University of Tasmania, Tasmania, Australia

E.R. Pike (✉)
Presidential Commission for the Study of Bioethical Issues, Washington, DC, USA
e-mail: elizabeth.pike@bioethics.gov

J. Collmann and S.A. Matei (eds.), *Ethical Reasoning in Big Data*,
Computational Social Sciences, DOI 10.1007/978-3-319-28422-4_5

efficient analysis of data by allowing complex work to be spread out across multiple investigative sites. Additionally, by allowing data to be aggregated across many different investigational sites, big data allows researchers to solve challenging or rare health problems that had previously proven difficult to investigate.

This potential for big data to advance our understanding of human disease has been particularly heralded in the field of genomics. Recent technological advances have accelerated the massive data generation capabilities of genomic research. Next-generation sequencing techniques now use semiconductors and nanotechnology that increase the speed with which genomes are sequenced, resulting in a dramatic reduction in the time needed to sequence a given genome. This has allowed researchers to undertake larger scale genomic research, with significantly more participants, further spurring the generation of massive amounts of data. The advance of technology has also triggered a significant reduction in cost, allowing large-scale genomic research to be increasingly feasible, even for smaller research sites. This trajectory is likely to continue, as researchers predict that more advanced DNA sequencing technologies will be able not only to generate terabase-scale sequence data in seconds, but they will be able to sequence genomes for little or no cost (Schadt 2012). Along with more advanced methods of sequencing genomes, there have been improvements in the methods for collecting, storing, and sharing the data, particularly using computer-based databases, which have facilitated the rise of big data in genomics. We will use the term Large Scale Genomic Repositories (LSGRs) to refer to these research resources. The rise of genetic research has triggered the creation of many LSGRs, some of which contain the genomic information of more than a million research participants.

While LSGRs have genuine potential, they also have raised a number of ethical concerns. Most prominently, commentators have raised questions about the privacy implications of LSGRs, given that all genomic data is theoretically re-identifiable. Privacy can be further threatened by the possibility of aggregation of data sets, which can give rise to unexpected, and potentially sensitive, information. But beyond privacy concerns, LSGRs also raise questions about participant autonomy, public trust in research, and justice. In this chapter, we explore these ethical challenges, with the goal of elucidating which ones require closer scrutiny and perhaps policy action.

2 The Promise of LSGRs

While all scientific research produces data, genomic analysis is somewhat unique in that it inherently produces vast quantities of data. Every human genome contains roughly 20,000–25,000 genes, each comprised of over three million base pairs, so that even the most routine genomic sequencing or mapping will generate enormous amounts of data (International Human Genome Sequencing Consortium 2004). Since most studies include many different individuals, each with their own unique genomes, sequencing genomes of groups or populations produces huge quantities of data for researchers to analyze.

LSGRs are not merely a useful tool in organizing and compiling genetic research. Genomic research is a natural fit for big data, due to the complex nature of gene-based therapies and investigations, which necessitate the study and comparison of many individual genomes. For common diseases, it has become clear that a range of genomic variants can play a part in determining a given individual's risk for certain diseases or particular health outcomes. In order to find those variants, each of which might only make a small contribution to a given health risk, researchers must study a large number of both healthy and affected individuals, in order to identify the relevant genomic differences.

The vast quantity of data generated by such an analysis would once have overwhelmed even the most well-funded research labs. However, the use of LSGRs has enabled widespread data sharing, allowing analysis efforts to be spread across any number of investigational sites. This reduces analytic bottlenecks, while permitting more timely data analysis than any one investigative team would be able to accomplish on their own.

Aggregation also facilitates the study of rare diseases, where it is often difficult to find and recruit sufficient numbers of subjects with the relevant condition. LSGRs facilitate the collection of data from a geographically broad range of research sites, allowing advances in understanding that would be impossible to produce from studying small groups of individuals. By allowing data aggregation and pooling of data from many investigational sites, genetic underpinnings of various conditions can be identified, allowing researchers to begin the search for targeted therapies to combat some of the most devastating, and rare, genetic based conditions. LSGRs provide adequate statistical power to address questions that were previously infeasible due to logistical and funding limitations. Aggregation of disparate data sets also can allow researchers to make novel connections, or reveal trends not readily apparent in any one data set.

Given the potential of LSGRs to advance our understanding of disease, it is easy to understand why scholars predict that the use of LSGRs will only accelerate in the coming years. Indeed, there are already signs that LSGRs will become an increasingly common feature of the research landscape. In particular, a recent NIH genomic data-sharing policy requires that any researcher who receives funding for the production of genomic data must deposit their sequence data in a central repository (Genomic Data Sharing 2014). Policies like this are the first step in creating more widespread and informative LSGRs, and indicate that LSGRs may become a common feature of any significant genomic research.

Beyond the 2014 NIH genomic data-sharing policy, there are several examples of well-funded, emerging LSGRs that have already contributed significantly to our understanding of genomics and human disease. One example is the Million Veteran Program (MVP), started by the Department of Veterans Affairs in May of 2011. The MVP contains genomic, and some clinical information, from veterans who receive their care from VA *and* who volunteer to participate in the program. The initial benefits of this database are already being realized, with the VA using this information to identify patterns of illness following deployment.

Additionally, President Obama's "Precision Medicine Initiative" includes as a centerpiece a national repository containing health records and genomic sequence data from more than one million volunteers. The hope is that such a database will allow researchers to study the mechanisms by which peoples' genes, environment, and lifestyle affect their health, in ways not possible without the pooling of large amounts of data. By combining genomic information into population studies, hidden genomic influences may be identified. Beyond potentially revealing the causes of various conditions, this could elucidate opportunities for targeted therapies, allowing the development of cures with maximum efficacy.

3 Privacy and Re-identification

Despite LSGRs' promise for scientific advancement, their increasing ubiquity raises considerable privacy challenges (Lane et al. 2014). Most genomic samples and data are included in LSGRs premised on a promise of anonymity. A major concern is that this promise might be undermined by the possibility of re-identification (Rothstein 2010). While technically very difficult, re-identification can occur when researchers apply bio-informatic techniques that cross-reference existing, identified data sets with the genomic information contained in the LSGRs. These concerns are far from theoretical. Indeed, several groups of researchers have demonstrated that re-identification is possible, even with the limited information contained in de-identified LSGRs. In a seminal study led by Gymrek and colleagues, researchers were able to discover the identity of some individuals whose genomes had been sequenced as part of a genomics project. The research team wrote an algorithm that was able to infer an individual's array of genetic markers, called a haplotype, from the nucleotide sequence of his Y chromosome. The team then searched genealogical databases for the names of men with corresponding Y-chromosome haplotypes, and, after cross-referencing the last names with publicly available records, correctly identified several individuals (Gymrek et al. 2013).

Another study utilized public databases, which make genome-scale RNA abundance profiles (which reveal the amount of RNAs in different cells) available to anybody with the internet. Researchers were able to generate DNA barcodes from these data, which could be screened against DNA databases kept by government agencies (to identify DNA samples associated with unsolved crimes for example). It is possible that comparing these data sets could reveal the identity of a research participant. In 2012, Schadt and colleagues utilized RNA abundance measurements to infer a DNA-based barcode that was specific enough to re-identify individuals whose data was part of a collection of hundreds of millions of individual genotypic profiles obtained in a completely different research context (Schadt et al. 2012). Researchers have also reported that a personal large-scale SNP genotypic profile is sufficient to resolve whether an individual participated in a specific genome-wide association study, even if the study reports only summary statistics such as allelic frequencies (Homer et al. 2008).

With re-identification existing as an increasingly real possibility, attention has shifted to the challenges associated with offers of anonymity in genetic research. Re-identification concerns are heightened further by the aggregation of ever-greater amounts of information on the internet. This aggregation problem creates a novel threat to privacy, as cross-referencing this information with LSGRs can give rise to unexpected and potentially sensitive inferences and information. Furthermore, recent research has raised the possibility that scientists could use genetic markers from DNA in order to create a fairly accurate picture of an individual's face, highlighting that we are only beginning to realize some of the privacy implications raised by access to genetic information (Claes et al. 2014).

The above discussion highlights the potential for re-identification of genetic research participants. However, a more nuanced understanding of the *risks* that such re-identification poses to participants warrants closer scrutiny. A helpful way of assessing such risks is separating out participants' welfare and non-welfare interests (Tomlinson 2009). Welfare risks are best thought of as individual direct harms that represent a real personal risk to the individual. In contrast, non-welfare risks do not present a risk of immediate personal harm, but rather represent abstract harms to an individual's wishes, desires, or preferences. A non-welfare risk can be said to occur when an individual loses control over their personal information (Tomlinson 2009). We address these different kinds of harms separately in the next sections.

4 Welfare Interests

Genomic big data research may expose subjects to psychological, social, and economic harms, particularly if the research reveals sensitive information about re-identified individuals, or racial/ethnic/geographic groups with which they identify. Psychological harms include undesired changes in thought processes and emotion (e.g., episodes of depression, confusion, feelings of stress, guilt, and loss of self-esteem). Social and economic harms might include embarrassment within a participant's business or social group, loss of employment, or criminal prosecution caused, for example, by invasions of privacy and breaches of confidentiality. Additionally, some social and behavioral research may yield information about individuals that could "label" or "stigmatize" the subjects, either as individuals or through association with a specific group. While these harms are often cited as reasons to worry about genomic research, evidence of these harms is thus far quite low.

4.1 Psychological Harms

Arguments about psychological harms assume that research participants will be given distressing information about their genetic health risks, which will cause undue negative emotions. There is a robust psychological literature, however, that suggests

that people are more emotionally adaptable than they think, and that we are terrible at affective forecasting, or predicting our future emotional reactions to negative events. While we often assume that learning about genetic risk for serious diseases will be devastating, in reality, the data suggest that the negative psychological effects of learning such information are generally transient and mild. This has been attributed to two psychological concepts: immune neglect and the focal illusion. Immune neglect refers to "the failure to anticipate how easily and quickly we make sense of and adapt to negative events." (Peters et al. 2014). The related focal illusion bias "is the tendency to focus on the affective consequences of a single, focal future event, while ignoring the emotional impact of non-focal events on well-being." (Peters et al. 2014).

The minimal psychological impact of negative genetic information has been demonstrated in a range of contexts (Heshka et al. 2008). For example, the REVEAL studies (Risk Evaluation and Education for Alzheimer's disease) were the first randomized controlled trials designed to evaluate the impact of susceptibility testing using the Alzheimer's Disease ("AD") susceptibility gene *APOE-ε4*. These comprised a series of four multi-site, randomized clinical trials examining psychosocial and behavioral responses to genetic risk assessment for AD using *APOE* disclosure (Roberts et al. 2011). The studies found little negative emotional impact (Green et al. 2009). Another systematic review similarly found no increased distress within the year after testing, and actually demonstrated a decrease in stress for many participants post-test (Broadstock et al. 2000). Similarly, a review of the literature on responses to genetic testing of cancer susceptibility found that there was very little evidence of adverse psychological effects observed among people who learn that they have a genetic predisposition to certain cancers (Meiser 2005). Similar data exists for testing range of other conditions, including Huntington's disease, breast cancer, and colon cancer, among others.

While we do not mean to minimize the possibility of psychological harms resulting from disclosure of genetic risk information, the existing literature should force us to consider whether our society is "systematically overestimate[ing] the durability and intensity of the affective impact of events on well-being." (Peters et al. 2014). Our argument is merely that policy makers and the scientific community should be cautious about using the psychological concerns of receiving genetic test results to justify regulations that will have a profound impact on the scientific enterprise.

4.2 Discrimination

Genetic discrimination ("GD") commonly refers to "the differential treatment of asymptomatic individuals or their relatives on the basis of their real or assumed genetic characteristics." (Otlowski et al. 2012). Differential treatment can occur within interpersonal and institutional domains, but institutional domains have been the focus of regulatory efforts. Objective evidence of GD has been difficult to establish and, until recently, its prevalence and depth has been largely undocumented.

Some studies have even presented positive evidence suggesting skepticism about GD's scope. A U.S. study on insurance outcomes published in 2009 surveyed 47 unaffected individuals with a genetic predisposition to breast cancer, concluding, "[r]esults suggest fear of GD is prevalent, yet data do not support evidence that GD exists." (McKinnon et al. 2009). Two adverse events were reported to have occurred when individuals changed health insurance. The study found no reports of job discrimination due to genetic status or family history of cancer. Furthermore, we are not aware of any instances in which GD has arisen from genetic research projects. In the closest available report, Kathy Hudson and others reported a case study in 1995 in which a research geneticist determined—outside the context of a research project—that a four-year old boy carried a genetic alteration that causes long QT-syndrome. His father subsequently was unable to obtain insurance coverage for his son because of this mutation (Hudson et al. 1995).

In the U.S., early experience with the Genetic Information Non-Discrimination Act (GINA) similarly suggests that perhaps there is less cause for concern than previously thought. Enacted in 2008, GINA was passed as a way to combat fears that genetic discrimination was a barrier to adoption of clinical genetic testing (Prince and Berkman 2012). The law works both prospectively (prohibiting employers and health insurance companies from receiving genetic information) and retrospectively (punishing bad actors who have illegally used genetic information as the basis for employment or actuarial decisions). While a watershed achievement, there have been remarkably few cases brought under the law (Genetic Information Non-Discrimination Act Charges 2014). Since 2010, there has been an annual average of just 48 cases reaching merit resolution and damages have not been substantial, averaging less than $1 million in total annual awards. While there have been more documented instances of discrimination in the life insurance and long-term care insurance areas, a systematic review of existing data led researchers to conclude that no policy intervention is currently justified, concluding that "with the notable exception of studies on Huntington's disease, none of the studies reviewed here (or their combination) brings irrefutable evidence of a systemic problem of GD that would yield a highly negative societal impact." (Joly et al. 2013).

As with the discussion of psychological harms, we do not mean to minimize the problem of genetic discrimination. It is certainly possible that genetic discrimination could eventually become a serious problem. While policy-makers should be cautious about imposing burdens on the research enterprise when there is little evidence of a current widespread problem, there is reason to guard against dismissing GD too quickly. As genetic information becomes more available and as our knowledge of the links between phenotype and genotype improves, insurance companies and others may take the opportunity to incorporate the information into decision-making. In one study, researchers at Georgetown University asked underwriters from insurance companies to underwrite hypothetical applicants who had received a genetic test result indicating increased risk of a future health condition. In seven of 92 total decisions, underwriters said they would deny coverage, place a surcharge on premiums, or limit covered benefits based on an applicant's genetic information. Adverse determinations were dispersed among the surveyed underwriters, across the hypothetical examples and despite relevant state-level proscriptions on genetic discrimination (Politz et al. 2007).

5 Non-welfare Interests

5.1 Trust

Even though there might not be current evidence of extensive individual welfare harms, one still must be concerned about the threat of harm to the non-welfare interests of participants. This can result from the lack of control over their samples and data, as well as the harm of broken promises, as participants participate in research with the expectation that they will not be personally identified by the data, and that their data will not be publically linked to them. Even if this does not result in tangible economic or mental harm to the individual, participant's non-welfare interests can be harmed by the release of this information. For these reasons, maintaining trust in the research enterprise and in the process of developing LSGRs is fundamental to the ongoing success of LSGRs and the research enterprise. And yet, the way that LSGRs are currently being created falls short of best practices for establishing and fostering trust.

Although some of the samples stored in LSGRs are collected from people who have provided consent for the genetic material to be used in a wide array of research projects, in other cases, the samples stored in LSGRs were collected without the source's knowledge or consent. For example, researchers often rely on samples collected as medical waste—blood or bodily tissue obtained in the clinic in excess of what was strictly necessary for testing or diagnosis. Current U.S. laws and guidelines allow the excess medical waste to be collected and stored in LSGRs without the source's knowledge or consent. Collection and storage of medical waste is generally governed by the Health Insurance Portability and Accountability Act's Privacy Rule, which places only limited restrictions on the ability to collect and store medical waste without consent. When samples are de-identified, the Privacy Rule places no restrictions on their use or disclosure. Even samples stored with specific identifiers can be used or disclosed under the Privacy Rule if the information is released as part of a "Limited Data Set" or if an Institutional Review Board has waived the requirement that individuals provide informed consent.

Researchers also rely on samples collected through the process of newborn screening—a public health screening process whereby newborns' heels are pricked and blood is collected and tested in the first few days of the child's life. Newborn blood spots are thought to be "an especially rich source of research material: they are stable over time, they constitute an unbiased collection of samples since they represent the entire population, and they can potentially be linked to basic demographic information" (Suter 2014). In many cases, the collection of a newborn's blood occurs without the parents' knowledge or consent (Suter 2014). The samples are then retained, in some cases indefinitely, for a range of subsequent uses (Citizens' Council on Health care 2009). The research use of newborn samples accelerated in 2009 due to an NIH grant that funded the *Newborn Screening Translational Research Network*, a national repository of newborn blood samples for use in research (Scutti 2014).

Although the federal Common Rule governing research with human subjects generally requires that investigators obtain informed consent from research participants, consent often is not required for research involving genetic samples. First, to the extent research samples are de-identified, the research is not considered human subjects research at all such that the Common Rule requirements (including the requirement of informed consent) do not apply. Second, even research using identifiable biospecimens may nevertheless be exempt from the Common Rule requirements of informed consent if data is "recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects." Third, even if the research is not considered exempt, an IRB is permitted to waive the requirements for informed consent in certain circumstances. Research using identifiable biospecimens can often qualify for waiver because the sheer number of people from whom genetic data has been collected renders re-contact and obtaining informed consent impracticable or impossible (Geetter 2011).

Given the lack of legal limitation, it is unsurprising that there are vast numbers of samples that are likely to have been collected without people's knowledge or consent. As of 1999, the RAND Corporation estimated that U.S. research repositories contained 307 million tissue samples. These samples were taken from 178 million individuals, accounting for almost two-thirds of the American population (Eiseman 2000). The RAND report conservatively assumed that the number of samples would grow by 20 million per year, which would mean that more than 600 million samples are being stored today, which does not even fully account for new sources of biological samples (direct-to-consumer genetic testing, criminal databases, etc.) that were just emerging at the end of the 20th century. It does not seem like much of an exaggeration to conclude, therefore, that "virtually everyone has his or her tissue on file" (Dunn 2012).

The potential for loss of trust in LSGRs when people learn that their genetic material has been collected, stored, and used without their knowledge or consent is high, and hugely consequential. This loss of trust has already occurred at the state level. In two states, Texas and Minnesota, parents learned that blood samples from their newborns had been collected without their consent and had been stored and used for a range of purposes including research. They subsequently brought suit. The Texas lawsuit, *Beleno v. Texas Department of Health Services*, ultimately led to the state agreeing to incinerate approximately 5.3 million newborn blood samples (Waldo 2010). The Minnesota lawsuit, *Bearder v. Minnesota*, ended with the state agreeing to "destroy all blood samples in long-term storage … and to pay nearly $1 million in legal costs." (Olson 2014).

These cases go beyond potential legal and financial consequences to highlight the less tangible ramifications of insufficiently informing and accommodating the views of potential participants in large-scale genetic research. Notably, Andrea Beleno, the named plaintiff in the Texas lawsuit, stated that she might have consented to the collection and subsequent use of her newborn's genetic data if she had trust in the enterprise: "If they had asked me … I probably would have consented. The fact that it was a secret program really made me so suspicious of the true motives, there's no way I would consent now" (Roser 2009). Surreptitious

collection and use—collecting and using samples without the knowledge and consent of the source—leads to lack of trust in the enterprise. Without trust in the mission of LSGRs, LSGRs are at risk of the type of lawsuits that resulted in incineration of millions of samples along with a more widespread loss of faith in the medical research establishment more broadly.

5.2 Autonomy

Informed consent is a cornerstone of research ethics. However, LSGRs have forced a reexamination of existing regulations and norms. Traditionally, there was a clear distinction between data that included identifiers (e.g., name, date of birth, social security number, etc.) and data that had been de-identified. Under the Common Rule, secondary research involving de-identified data has not been considered to be human subjects research, and thus has not required IRB review. This regulatory distinction ultimately meant that consent has not been required for much of the genomic data contained in research repositories.

In large part because of concerns about re-identification of genomic data, proposed changes to the Common Rule look to obliterate the distinction between identified and de-identified data (Federal Policy for the Protection of Human Subjects 2015). The net effect of this change will be to require some kind of informed consent for any sample or data that will be used for research. While adopting a posture that seems more respectful of individual autonomy, this change could have a profound effect on the research enterprise generally, and on LSGRs in particular. The proposed rules would likely only apply prospectively, and would introduce the requirement of consent to collecting samples and data for subsequent use where one did not exist before.

Implementation of this new requirement will depend, in part, on whether participants are willing to accept the idea of blanket or broad consent. Blanket consent refers to the notion that a participant could give their consent at a single interaction, but would give permission for ongoing, open-ended use. Broad consent is similarly non-specific, but includes provisions wherein future uses are subject to some constraints (e.g., not for morally controversial topics, such as cloning). Blanket and broad consent can be compared to other approaches that require more study-specific consent, which obviously provide more information to a potential participant, but at significant cost to the research enterprise (Grady et al. 2015).

Some form of broad consent is expected to be part of the revisions to the Common Rule. Furthermore, there is evidence to suggest that participants are willing to accept such an approach. While a complete analysis is beyond the scope of this chapter, the data seems to indicate that individuals want to be asked for their permission once, but do not need to be approached to provide consent for specific subsequent uses (Wendler 2006; Chen et al. 2005). In fact, in one recent survey of various consent models for the use of stored genetic samples, potential participants viewed real-time specific consent as the least desirable option (Tomlinson et al.

2015). Unfettered blanket consent was also not widely supported, with subjects seeming to prefer the broad consent model where one-time permission is given, but when there are limits on controversial research uses, or a mechanism to withdraw at any point.

Any informed consent paradigm will involve some tradeoffs between burden on the research enterprise and participants' ability to exercise control over the use of their samples. As LSGRs proliferate, it seems untenable to continue with the status quo, where research is being conducted on samples and data without participant knowledge or consent. However, in the interest of minimizing burden on the research enterprise, careful consideration should be given to the rules that will be imposed. If implemented thoughtfully, broad consent seems like it could be an acceptable and appropriate compromise between respecting autonomy and facilitating research.

In addition to prospective consent, two additional autonomy-related concerns are raised by the proliferation of LSGRs. First, there are retrospective questions about the appropriateness of using genomic data and samples when there is inadequate or problematic evidence of consent. We term this the "grandfathering problem." When researchers seek to access genetic samples, many of which might be very old, how much evidence of high quality informed consent is required before allowing research to be conducted? For instance, perhaps a researcher retires and transfers a career's worth of samples to a biobank. Some of those samples might have been collected before modern informed consent laws and norms were in place, meaning that consent has not been documented, or is non-existent for *any* form of research with the samples. Or perhaps some of those samples were collected for a specific research purpose, and the consent form never mentioned the possibility of any sort of genetic research methodology (or mentioned only rudimentary forms of genetic analysis) suggesting that consent could be inadequate. Even more challengingly, some of those samples might have been collected from vulnerable populations (e.g., prisoners, psychiatric patients, adults lacking capacity, etc.).

Given that norms and rules evolve, we cannot simply apply today's consent standards to yesterday's samples and data. On the other hand, it seems ethically problematic to knowingly use research resources of questionable provenance. Important conceptual work will have to be done to develop an ethical framework that considers a number of relevant factors. First, we need to establish the extent to which inadequate or missing informed consent is ethically problematic in a range of scenarios. For example, having firm evidence that samples were collected from vulnerable individuals without consent raises more concerns than a mere lack of documentation of informed consent. Second, we need to decide how strongly to weigh the feasibility of obtaining additional, present-day consent for subsequent research use as a way of demonstrating respect for individual autonomy against the additional burdens placed on the research enterprise. It is appropriate to seek re-consent in certain situations, but there should be limits on the burdens imposed on the research enterprise. Finally, we need to explore the weight that we are willing to give to the unique qualities or irreplaceable scientific value that a given set of samples or data might possess.

We suggest that an appropriate balance between these three factors would allow questionable samples and data to be grandfathered only in cases where the unique scientific value outweighs the relevant ethical concerns. As one possible model, the National Human Genome Research Institute has instituted a policy stating that as of a specific date, previously collected samples can continue to be used for genomic data sharing as long as the existing consent forms are not inconsistent with such use. In order to discourage researchers from only using previously collected samples indefinitely, this rule only remains in place for five years. After that time, researchers will need a strong scientific justification to continue using samples that were not obtained with specific consent for broad data sharing.

The final autonomy-related concern exacerbated by the proliferation of LSGRs relates to the right to withdraw from research. Enrolling in research is not just a one-time decision; it is a well-established principle of research ethics that participants have the right to withdraw from participation at any time. In the context of actual physical participation in research, this is conceptually straight-forward as an individual can choose not to show up or to leave the study premises. But in the context of LSGRs, where data are being shared widely throughout the research community, withdrawal can be difficult or impossible. LSGRs should be designed such that individuals retain some ability to pull their information back should they choose. However, once the data has been widely shared, absolute eradication of data might not be feasible. LSGRs should prompt a re-examination of what the right to withdraw from research actually entails, and should encourage construction of consent forms that manage participant expectations accordingly.

5.3 Justice

There are two primary justice concerns arising out of LSGRs. The first relates to the unfortunate lack of diversity in genomic medicine. While genomic research has been presented as an important tool for unlocking the potential of genomic medicine, research efforts thus far have focused almost exclusively on people of European descent. For example, as of 2011, less than 10 % of participants included in genome-wide association studies ("GWAS") were not of European decent (Rotimi 2012). In the U.S., one study found that 92 % of GWAS participants were white, and only 3 % were African-American (Haga 2010). The worry is that without a broader racial and ethnic focus, researchers will develop a skewed understanding of which variants are relevant to human disease. Genotype-phenotype associations will be less generalizable for underrepresented populations, meaning that the majority of medical benefits will flow to an already advantaged segment of our global population. As Carlos Bustamante and colleagues stated:

> It is tempting to focus on populations that are motivated, organized, medically compliant and otherwise easy to study. But by failing to develop resources, methodologies and

> incentives for underserved people, we risk perpetuating the health disparities that plague the medical system. Those most in need must not be the last to receive the benefits of genetic research. (Bustamante et al. 2011).

In order to avoid exacerbating health care inequality, LSGRs need to focus on engaging and recruiting under-represented populations.

LSGRs also run the risk of creating group harms. Beyond individual re-identification, there is a concern that through aggregating a sufficient amount of genetic information, and allowing it to be compared to other available databases, LSGRs may permit inferences about groups of people that could be considered harmful on a number of levels. First, there is a risk that genetic information could be mobilized to stigmatize or discriminate against individuals due to their perceived membership in a particular group. Often described as a "group-mediated harm to individuals," this kind of harm can arise in situations when a group is associated with increased genetic risk for having a particularly stigmatizing disease or trait (Hausman 2007). Genetic information also can cause harms to groups themselves where such groups have "structures, leadership, causal capacities, and interests that are distinct from and not reducible to the interests of their members" (Hausman 2008). An evolutionary genetics study reporting migration patterns, for example, could present results that differ from group lore thereby undermining the group but not necessarily harming its members. There are many ways in which this kind of group harms can be expressed, including loss of status in the majority society, self-stigmatization, and dignitary harms to the community (Freeman et al. 2006).

LSGRs pose a particular risk of creating both kinds of group harms because even though data contained in genomic repositories are not associated with personal information, racial and ethnic information is often retained (Hausman 2008). Furthermore, research has made it possible to infer ancestry about a given individual with high reliability, particularly when that individual is from a structured group whose genetic material has been relatively isolated. This means that as genomic data is shared widely, research might produce associations between racial or ethnic groups, and certain traits or medical predispositions. One such example arose in New Zealand in 2006, when researchers reported a variant of the "warrior gene"—associated with traits such as aggression, violence, and impulsivity—as being "strikingly overrepresented" in New Zealand Māori. A lead researcher was quoted as saying that "obviously" the findings meant that Māori men were "going to be more aggressive and violent and more likely to get involved in risk-taking behavior like gambling." (AAP 2006). The claim generated widespread media attention, and led to immediate opposition from Māori and other commentators (Crampton and Parkin 2007).

The fact that certain population groups can have higher frequencies of certain genotypes based on historical patterns of migration, isolation, and other features of population genetics warrants vigilance about the potential for group-mediated harms from genetic research (Hartl and Clark 2007). Even though the individual participants might have agreed to take part in research, current models of informed

consent and promises of privacy do not offer protection from these kinds of group-mediated harms. Because of this, LSGRs present wider-ranging threats than those raised by typical research.

Given these concerns, the question is whether or how policy-makers should impose governance structures on LSGRs to minimize risks to groups. To date, there has been some consideration of group harms, at least in the context of the NIH GWAS data sharing policy which required data access committees (DACs) to ensure that proposed research did not pose a risk of creating group harms. It is not clear whether that policy has been effective, and the more recent NIH genomic data sharing policy has dropped concerns about group harms entirely. While a formal review body might not be necessary, other governance options might mitigate worries about group harms. LSGRs could consider requiring that researchers seeking access to data agree to specific limits on data usage when conducting analyses with sensitive data (e.g., race, ethnicity, geography). Alternatively, researchers could stipulate that their results will not unduly impact any specific group in a foreseeably adverse way, placing the burden on the investigator to consider the ramifications of their findings.

6 Conclusion

The capacity to utilize big data represents a substantial shift in the research landscape; our ability to collect, store, share and aggregate data in such expansive ways is a monumental opportunity, but will surely also present significant ethical challenges. While existing policies and procedures may need to be modified to better protect subjects, some scholars have gone further, suggesting that fundamentally new standards of practice should be developed to deal with the unique ethical concerns created by LSGRs (Gymrek et al. 2013). Our analysis suggests, however, that caution is warranted before any major policies are implemented. Much attention has been directed at privacy concerns raised by LSGRs, but perhaps for the wrong reasons, and perhaps at the expense of other relevant concerns. We do not think that there is yet sufficient evidence to motivate enactment of major policy changes in order to safeguard welfare interests, although there might be some stronger reasons to worry about subjects' non-welfare interests. We also believe that LSGRs raise genuine concerns about autonomy and justice. Big data research, and LSGRs in particular, have the potential to radically advance our understanding of human disease. While these new research resources raise important ethical concerns, any policies implemented concerning LSGRs should be carefully tailored to ensure that research is not unduly burdened.

References

AAP. (2006, August 8). Warrior Gene" Blamed for Maori Violence. National Nine News.

Bollier, D., & Firestone, C. M. (2010). *The promise and peril of big data*. Washington, DC: Aspen Institute.

Broadstock, M., Michie, S., & Marteau, T. (2000). Psychological consequences of predictive genetic testing: A systematic review. *European Journal of Human Genetics, 8*(10), 731–738.

Bustamante, C. D., Francisco, M., & Burchard, E. G. (2011). Genomics for the world. *Nature, 475*(7355), 163–165.

Chen, D. T., Rosenstein, D. L., Muthappan, P. G., Hilsenbeck, S. G., Miller, F. G., Emanuel, E. J., et al. (2005). Research with stored biological samples: What do research participants want? *Archives of Internal Medicine*, *165*, 652–655.

Citizens' Council on Health Care. (2009). State by state government newborn blood & baby DNA retention practices. Retrieved at http://www.cchfreedom.org/pdf/50_States-Newborn_

Claes, P., Hill, H., & Shriver, M. D. (2014). Toward DNA-based facial composites: Preliminary results and validation. *Forensic Science International: Genetics, 13*, 208–216.

Crampton, P., & Parkin, C. (2007). Warrior genes and risk-taking science. *New Zealand Medical Journal, 120*, U2439.

Dunn, C. K. (2012). Protecting the silent third party: The need for legislative reform with respect to informed consent and research on human biological materials. *Charleston Law Review, 6*, 635–684.

Eiseman, E. (2000). Stored tissue samples: An inventory of sources in the United States. In National Bioethics Advisory Commission (NBAC), *Research involving human biological materials: Ethical issues and policy guidance*. Rockville, Maryland: NBAC.

Federal Policy for the Protection of Human Subjects. (2015). Retrieved at https://www.federalregister.gov/articles/2015/09/08/2015-21756/federal-policy-for-the-protection-of-human-subjects

Freeman, W. M., Romero, F. C., & Kanade, S. (2006). Community consultation to assess and minimize group harms. In E. A. Bankert & R. J. Amdur (Eds.), *Institutional review board management and function* (2nd ed.). Sunderland, MA: Jones and Bartlett.

Geetter, J. S. (2011). Another man's treasure: The promise and pitfalls of leveraging existing biomedical assets for future use. *Journal of Health and Life Science Law, 4*, 1–104.

Genetic Information Non-Discrimination Act Charges. (2014). Retrieved at http://www.eeoc.gov/eeoc/statistics/enforcement/genetic.cfm

Genomic Data Sharing. (2014, August 27). Retrieved at https://gds.nih.gov/

Grady, C., Eckstein, L., Berkman, B. E., Brock, D., Cook-Deegan, R., Fullerton, S. M., et al. (2015). Broad consent for research with biological samples: Workshop conclusions. *American Journal of Bioethics, 15*(9), 34–42.

Green, R. C., Roberts, J. S., Cupples, L. A., Relkin, N. R., Whitehouse, P. J., Brown, T., & Farrer, L. A. (2009). Disclosure of APOE genotype for risk of Alzheimer's disease. *New England Journal of Medicine, 361*(3), 245–254.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science, 339*(6117), 321–324.

Haga, S. B. (2010). Impact of limited population diversity of genome-wide association studies. *Genetics in Medicine, 12*(2), 81–84.

Hartl, D. L., & Clark, A. G. (2007). *Principles of population genetics* (4th ed.). Sunderland, MA: Sinauer Associates.

Hausman, D. M. (2007). Group risks, risks to groups, and group engagement in genetics research. *Kennedy Institute of Ethics Journal, 17*, 351–369.

Hausman, D. (2008). Protecting groups from genetic research. *Bioethics, 22*(3), 157–165.

Heshka, J. T., Palleschi, C., Howley, H., Wilson, B., & Wells, P. S. (2008). A systematic review of perceived risks, psychological and behavioral impacts of genetic testing. *Genetics in Medicine, 10*(1), 19–32.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., & Craig, D. W. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics, 4*(8), e1000167.

Hudson, K. L., Rothenberg, K. H., Andrews, L. B., Kahn, M. E., & Collins, F. S. (1995). Genetic discrimination and health insurance: An urgent need for reform". *Science, 270*(5235), 391–393.

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931–945.

Joly, Y., Feze, I. N., & Simard, J. (2013). Genetic discrimination and life insurance: A systematic review of the evidence. *BMC Medicine, 11*, 25–40.

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement*. Cambridge: Cambridge University Press.

McKinnon, W., Banks, K. C., Skelly, J., Kohlmann, W., Bennett, R., Shannon, K., & Wood, M. (2009). Survey of unaffected BRCA and mismatch repair (MMR) mutation positive individuals. *Familial Cancer, 8*(4), 363–369.

Meiser, B. (2005). Psychological impact of genetic testing for cancer susceptibility: An update of the literature. *Psycho-Oncology, 14*, 1060–1074.

National Science Foundation. (2010). Core techniques and technologies for advancing big data science and engineering program solicitation. Retrieved at http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.htm

Olson, J. (2014, January 14). Minnesota must destroy 1 million newborn blood samples. *Star Tribune*.

Otlowski, M., Taylor, S., & Bombard, Y. (2012). Genetic discrimination: International perspectives. *Annual Review of Genomics and Human Genetics, 13*, 433–454.

Peters, S. A., Laham, S. M., Pachter, N., & Winship, I. M. (2014). The future in clinical genetics: Affective forecasting biases in patient and clinician decision making. *Clinical Genetics, 85*(4), 312–317.

Pollitz, K., Peshkin, B. N., Bangit, E., & Lucia, K. (2007). Genetic discrimination in health insurance: current legal protections and industry practices. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, *44*(3), 350–368.

Prince, A. E., & Berkman, B. E. (2012). When does an illness begin: Genetic discrimination and disease manifestation. *The Journal of Law, Medicine & Ethics, 40*(3), 655–664.

Roberts, J. S., Christensen, K. D., & Green, R. C. (2011). Using Alzheimer's disease as a model for genetic risk disclosure: Implications for personal genomics. *Clinical Genetics, 80*(5), 407–414.

Roser, M. A. (2009, December 23). State agrees to destroy more than 5 million stored blood samples from newborns. *Statesman*.

Rothstein, M. A. (2010). Is deidentification sufficient to protect health privacy in research? *The American Journal of Bioethics, 10*(9), 3–11.

Rotimi, C. N. (2012). Health disparities in the genomic era: The case for diversifying ethnic representation. *Genome Medicine, 4*(8), 65–68.

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Molecular Systems Biology, 8*(1), 612.

Schadt, E. E., Woo, S., & Hao, K. (2012). Bayesian method to predict individual SNP genotypes from gene expression data. *Nature Genetics, 44*(5), 603–608.

Scutti, S. (2014, July 24). The government owns your DNA. What are they doing with It? *Newsweek*.

Suter, S. M. (2014). Did you give the government your baby's DNA? Rethinking consent in newborn Screening. *Minnesota Journal of Law Science and Technology, 15*, 729–790.

Tomlinson, T. (2009). Protection of non-welfare interests in the research uses of archived biological samples. In K. Dierickx & P. Borry (Eds.), *New challenges for biobanks: Ethics, law, governance*. Intersentia: Antwerp.

Tomlinson, T., De Vries, R., Ryan, K., Kim, H. M., Lehpamer, N., & Kim, S. Y. (2015). Moral concerns and the willingness to donate to a research biobank. *Journal of the American Medical Association, 313*(4), 417–419.

Waldo, A. (2010, March 16). The Texas newborn bloodspot saga has reached a sad—and preventable—conclusion. *Genomics Law Report*.

Wendler, D. (2006). One-time general consent for research on biological samples. *British Medical Journal, 332*(7540), 544–547.