# Seeking a "Race to the Top" in Genomic Cloud Privacy?

Mark Phillips, Bartha M. Knoppers, and Yann Joly
Centre of Genomics and Policy
McGill University
Montreal, Canada

***Abstract*—The relationship between data-privacy lawmakers and genomics researchers may have gotten off on the wrong foot. Critics of protectionism in the current laws advocate that we abandon the existing paradigm, which was formulated in an entirely different medical research context. Genomic research no longer requires physically risky interventions that directly affect participants' integrity. But to simply strip away these protections for the benefit of research projects neglects not only new concerns about data privacy, but also broader interests that research participants have in the research process. Protectionism and privacy should not be treated as unwelcome anachronisms. We should instead seek to develop an updated, positive framework for data privacy and participant participation and collective autonomy. It is beginning to become possible to imagine this new framework, by reflecting on new developments in genomics and bioinformatics, such as secure remote processing, data commons, and health data co-operatives.**

## I. Introduction

Many computation- and storage-intensive genomic research projects have now enthusiastically turned to cloud service providers (CSPs) to address their needs. For example, in Genome-Wide Association Studies, "the consensus view is clear: the more samples the better" [1]. These vast quantities of genomic data can no longer be stored and processed on researchers' own equipment. Nor is it easy to quickly share the data with their colleagues at other facilities. Projects like Google Genomics and Amazon Web Services are now marketing CSP services for genomic researchers to address these concerns. In this paper, we use the term "genomic cloud research" to refer to the storage or processing of genomic data in cloud-based infrastructure for the purpose of carrying out health research.It remains unclear what data privacy law has to contribute to the complex webs of relationships forming between researchers, research institutions, and CSPs. Indeed, the relationship between genomic research and data privacy[1] may have gotten off on the wrong foot.

As lawmakers modernize data privacy frameworks, they often fail to meaningfully take into account the realities of research and data privacy in the genomic cloud. This failure has been met with corresponding frustration on the part of research institutions and their proponents. The tension has only intensified in the wake of Edward Snowden's revelations of comprehensive electronic state surveillance programs, which increased privacy concerns with respect to delocalized and potentially sensitive personal information. Global regulators and policymakers have since sought to patch up privacy vulnerabilities in fits and starts. They have largely treated the contexts of health and scientific research with a uniform degree of protectionism. Meanwhile, in the genomic research context, data privacy laws ostensibly exist to protect those whose genetic data are to be studied. These people have been largely absent from the discussion altogether.

This paper suggests that a window of opportunity has opened that can now allow a new relationship to be built, a relationship that can begin to take these concerns seriously. Advances in genomics and bioinformatics themselves, such as remote computation, data commons, and data co-operatives, present the potential to imagine a means to establish frameworks and institutions that set up what we refer to here as a "race to the top", by simultaneously giving a voice to research participants, by allowing research institutions to easily understand the legal and professional obligations expected of them, and by ensuring that robust data privacy protection is put in place.

**Contributions.** We summarize the contributions of the paper as follows:

- We describe the current legal, policy, and ethical landscape of genomic cloud research.
- We suggest combining emerging institutional and technological approaches to cloud genomic research, such as data commons or health data co-operatives, secure computing. In the cloud context, we suggest these arrangements will be best-suited to driving innovations in data privacy and autonomy. We suggest personal genomic data stores as one potential privacy-protecting result.

## II. Off on the Wrong Foot: Genomic Research meets Data Privacy

There is a growing consensus that data privacy law and policy are out of step with contemporary health research practice [2, 3, 4]. The dramatic rise of genomic research over the past two decades is seen as especially ill-adapted to the traditional tools developed by the fields of data privacy law and bioethics. These fields aimed to protect research participants from the risks inherent in research projects by imposing duties on the researchers. The primary legal duties were respect for participants' autonomy, especially through informed consent requirements, as well as duties to avoid conflicts of interest, and to maintain patient confidentiality. These traditional tools vary according to the circumstances of the research, but

1 The term "data privacy" is used in this paper to encompass both data protection and privacy.


DOI 10.1109/SPW.2015.26

essentially imagine privacy as a set of standardized check boxes that are the researchers' sole responsibility.

As a result of this framework, lawmakers have tended either to treat all forms of medical research as equivalently dangerous, or to target individual research fields for specific rulesets (e.g. relating to stem cells, tissue, data, etc.). In each case, lawmakers have failed to meaningfully engage with the scientific, health, and privacy issues at stake. The established frameworks do not yet meet present needs, and they are not flexible enough to remain relevant to, and perhaps even contribute to shaping, technological and institutional changes yet to come. These shortcomings give rise to three different concerns.

First, genomic researchers must comply with an ever-expanding number of confusing, overlapping laws and policies regulating their activities. Even within a single country, few if any are tailored to their work. Canada alone, for example, has no fewer than twenty-nine laws aimed at comprehensively regulating privacy in the public, private, or health sectors. And these are only general privacy laws, and do not include more specific privacy laws, such as those aimed at stemming identity theft, or laws not focused on privacy that nonetheless include important privacy provisions, all of which may also bear on genomic research.

But the number of laws now confronting genomic research projects are an order of magnitude greater than simply those at the national level. Cloud-based research and cooperation across borders has become the norm. Data flows and genomic processing are carried out across multiple legal jurisdictions. A genomic research project has to comply with the data privacy laws of each jurisdiction involved, including that of any collaborating researcher and any potential site at which genomic data may be stored or managed in the cloud. A chorus of research advocates, as a result, have noted that "research activities are hobbled by thickets of laws, regulations, and guidance" [2].

Inadequate data privacy law and policy also raise a second concern. Their inhibiting effects on research do remarkably little to advance the interests of those whose genomic data is in fact being researched. The traditional approach adopted by bioethics to promote research participants' interests has been to jealously protect a single aspect of their participation: their consent. Beyond the principle of informed consent, research participants have been passive subjects in the research process.

In the rush to ensure that researchers have access to existing genomic data, some now call into question the suitability of informed consent. This protectionist measure, they argue, emerged in the context of physically invasive medical procedures. Why should we apply it to secondary research on tissue and data that were collected for an earlier research project? In these cases, the research participant may never even become aware of the existence of the secondary research project.

This approach aims to instead see the place of research participants as shaped not by risks, but instead by the social benefits that can flow to all members of society through the outcomes of health research. From this perspective, participants are better understood as partners in research projects, linked through bonds of altruism, bonds which are undermined by antiquated and ill-adapted demands put on researchers to be accountable to participants [5].

This new approach, however, ignores too many of the groups research participants and these social goals. Their data more directly benefit research institutions, whether public or private, and public health and law enforcement surveillance programs. One example of this involves routine newborn blood screenings. The screenings themselves are crucial to detecting and preventing serious, otherwise-undetectable conditions. But legal proceedings were recently instituted in Texas [6], Minnesota [7], and British Columbia [8] after the blood-sample data was allegedly transferred without the parents' consent to third parties for secondary research purposes. In at least one of these claims, the parents cited potential recipients as including pharmaceutical companies and law enforcement agencies.

Research institutions have an interest in protecting the privacy of participants to the degree necessary to allow research to continue. But the increasing commercialization of the biobanking sector [9] in the context of the concerns raised above illustrates the potential for conflicts of interest to arise between researchers and research participants. This potential is exacerbated by the entry of the health research sector into data privacy law lobbying, such as in the course of the legislative process of the forthcoming European Union *General Data Protection Regulation* (GDPR). The former EU Commissioner observed that GDPR lobbying "has been fierce – absolutely fierce" [10], and has included efforts from U.K. health research charity the Wellcome Trust to allow secondary research of pseudonymized (key-coded) data without specific consent [11]. In practice, participants will often be satisfied if they are notified of the secondary uses made and protections provided to their data, and given the option to withdraw. But recognizing this fact does not, on its own, determine optimal approach: whether and how to provide data subjects with notice-and-consent, do-not-track, or do-not-collect options will clearly have an impact on data collection and data privacy interests.

Although research participants may be ill-served by an earlier bioethical conception of individual autonomy, new concerns such as those raised here illustrate the continuing need for regulatory mechanisms to protect and promote their distinct interests.

The third and final concern raised by ill-adapted data privacy laws and policies poses challenges to *both* researchers and research participants. In the cloud context, each risks losing direct control not only of their data, but of the architectures containing those data. Despite the benefits cloud computing offers to genomic research, it nonetheless places the data outside the direct control of both researcher and research participant. Because networked computing forms part of the "intellectual commons" similar to public works that make other creative and scientific activity possible, cloud computing risks representing the enclosure of a commonly held resource, as "[u]sers might be baited and hooked into Cloud service reliance" and "could lose the ability to actively defend against the monopolies of computing capacity and (illegitimate)

content control" [12]. As the Electronic Frontier Foundation has emphasized, "Architecture is Politics" [13].

A tangible illustration of the helplessness of research institutions when faced with the capricious decision of a CSP is illustrated by the abrupt shuttering of Nirvanix, a previously prominent cloud provider, in 2013. Nirvanix "closed its doors and left over 1,000 customers with only two weeks to save their [hosted] data" elsewhere [14]. The dilemma, of course, is that CSP clients turn to cloud storage in the first place because of their incapacity to store or process Big Data on their own [15].

The risks posed by cloud computing to researchers and research participants, as well as to the relationship between the two, must then necessarily be addressed.

## III. A "Race to the Top"

It may be tempting to imagine that the risks posed by genomic cloud computing will push toward a kind of "race to the bottom". The term has been used in other fields when they became increasingly internationalized and borderless. The concept's origin is commonly attributed to a 1933 decision of U.S. Supreme Court Justice Brandeis [16]. It has more recently been used to describe an erosion of environmental and labor regulation, for example, as sovereign countries compete with one another in attempts to provide the most attractive conditions for corporate multinational investors.

Some advocates of genomic research have seemed to suggest that cloud computing must now result in a "race to the bottom" as far as privacy protection. Cloud computing, the argument goes, irresistibly heralds completely open data flows and reduced privacy expectations, and efforts to achieve stronger protections than basic de-identification are destined to be fruitless. Data storage locations may change without customers even knowing, let alone consenting, and trying to resist this locationlessness is thus as useless as trying to stop a rainstorm. But experience has shown, to the contrary, that data privacy, rather than being subject to a "race to the bottom", if anything follows the opposite course. Although cloud computing clearly does increase the borderlessness of data, the internationalization of computing seems as often as not to be characterized by pulls toward stronger awareness of the need for data privacy protection and advances toward it.

For example, the adoption in 1995 by the European Union of the GDPR's forerunner, the *Data Protection Directive* was met with neither widespread international disinvestment in the EU, nor with a breakdown in international data transfer. Instead, the *Directive* prompted countries like the United States and Canada to adopt their own strengthened data privacy instruments to allow continued compatibility, namely the U.S.–EU Safe Harbor framework [17] and Canada's PIPEDA privacy statute [18].

A similar race to the top may begin to be established with respect to CSPs. Article 18 of the draft GDPR, for example, would enshrine a right to data portability, which would lighten the vulnerability of CSP clients, including research institutions. CSPs able to provide their clients with convincing assurances of data portability would also enjoy distinct advantages over competitors in the eyes of potential clients fearing a repeat of the sudden collapse of Nirvanix.

## IV. Localization Restrictions

A more recent illustration of the willingness of industry and legal actors to address novel privacy concerns has emerged with respect to data localization restrictions: laws that prohibit data from crossing borders. These restrictions have been seen as increasingly attractive to those who wish to keep sensitive data out of jurisdictions where they might become subject to state seizure, especially under the provisions of the USA PATRIOT Act.

In keeping with the idea of a "race to the top" in data privacy, Germany's move to tighten localization restrictions in the wake of the Snowden revelations, for example, were not met with Germany's isolation. Instead, Amazon Web Services, among the biggest CSPs, announced that it would provide cloud services based in Frankfurt, in order to provide its customers "with the assurance that your content will stay within the EU" [19].

Despite this ability and willingness on the part of CSPs and legislators to accommodate increased data privacy, given the borderlessness of the Internet itself, localization has largely become an anachronistic and ineffectual solution to privacy concerns. The world's fixation on the USA PATRIOT Act has served to distort its perception of the risks of surveillance specific to the cloud.

For example, in a recent decision of the privacy commissioner of the Canadian province of Saskatchewan, the commissioner cited the USA PATRIOT Act as its reason for recommending heightened privacy protections for government employee data outsourced to be stored in the cloud [20]. But this approach to the problem of state surveillance neglects that the Snowden revelations were by no means restricted to concerns about the US. Instead, Snowden revealed that the members of "Five Eyes"—an intelligence alliance between all of the most powerful English-speaking countries, including Canada—had all assisted each other in spying on one another's citizens as well as those of numerous other countries. Cloud or no cloud, Saskatchewan's government employee data was always at risk of being harvested. A sense of security achieved through localization restrictions in fact simply masks these underlying concerns, which must largely be addressed closer to the source than localization restrictions permit.

This challenge is daunting, but the genomic research community is beginning to produce the institutional and technological frameworks that may be able to overcome them.

## V. Democratizing Genomic Privacy

These new frameworks spring from an increased commitment on the part of health researchers to "democratize" genomic privacy.

Genomic researchers, of course, have been particularly active in this respect. Projects like Bionimbus and the Genomic Data Commons in the United States, the Cancer Genome Collaboratory in Canada, and the Global Alliance for

Genomics and Health at the international level are each setting out to reverse the effects of digital enclosure. Each project combines secure remote computation with the beginnings of community cloud infrastructure controlled by and for research institutions themselves. These institutions need no longer be stuck with third-party commercial cloud providers. Even where researchers choose to continue to use commercial providers, the cloud research projects promote a "race to the top" by opening up new sets of features that the commercial providers will have to match to retain their customer base.

But although cloud researcher projects set in motion a process of genomic privacy democratization, its polity remains limited. Researchers, academics, and technicians each have their place, but those whose data are studied remain relatively voiceless, unless research institutions choose to integrate responses to their concerns.

What interest do genomic research participants rightly hold in their data, or the economic or informational proceeds that flow from it? Genomic data has, at times, taken on significant economic value, as is now clearly recognized by pharmaceutical companies, among many others. Participants seem to have an interest in their data that is stronger than those of private interests, perhaps even of the state. Indeed a prominent line of philosophy stretching back to John Locke holds that humans beings control and have domain over their own bodies. This is especially so given that although it is possible to significantly limit the risks posed to research participants, "[w]e can't guarantee zero risk if we want to share any useful data" [21].

But we may no longer be stuck with having to choose between individualistic or potentially coercive approaches to data privacy. The aggregated economic value of genomic data may itself, in the near future, allow for the building of institutions that are directly responsive to the collective decisions of the people whose data is studied. This process might extend the emerging idea of a "genomic commons" so that it might even empower data subjects themselves. The people studied in genomic research are, truth be told, growing so rapidly in number that they may soon come to approximate the entire population.

Writers have highlighted the importance of achieving "the structural incorporation of participant interests into governance via participant bodies," which is necessary because "the collective, public nature of a biobank's constitution and stored material in the form of data and samples militates strongly in favor of the active involvement of contributors in decisions regarding the allocation and stewardship of biobank resources" [22]. Winickoff, for example, has suggested a shareholder model in which participants hold a measure of decision-making power [23].

But it has also now become possible to imagine genomic research bodies controlled entirely by participants, although computer scientists, bioinformaticians, and others would undoubtedly be retained by the organizations to play key roles in their functioning. Such "health data co-operatives" could allow for meaningful exercise of democratic interests by data contributors themselves, as recently proposed by writers such as Hafen, Kossman, and Brand [24]. Within the existing genomic research paradigm, even the basic duties imposed by privacy law, of providing access to one's own personal data, is not a given [25]. The health data co-operative approach might instead allow research participants the framework within which to practically achieve a critical mass of collective autonomy [26], opening up the potential for participants to, in fact, more directly participate.

A health data cooperative would not necessarily need to aim to gather the genomic data of the entire population, but instead a number of cooperatives might each collect data from participants with shared identities, such as those united by geographic proximity or a shared genetic disease. The later is already a basis upon which patient advocacy groups are organizing [27]. This would allow for decision-making along the lines of each specific identity. Individual co-operatives could then join together in one or more federations, as Hafen, Kossman, and Brand themselves suggest [24].

Second, unlike other private health entities, cooperatives need not be driven by a profit motive. Although their existence would depend on a measure of revenue-generation, based on, for example, researchers paying for access to their valuable data or public subsidy, their core purpose need not be to generate revenue for their members. Instead, they could pursue innovation aimed at the collective autonomy and fullest possible participation of their membership in the decisions about their data.

The co-operatives would clearly need to consult with and develop relationships with researchers, regulators, information technology experts, and others, but their institutional integrity would require independence with respect to their governance functions. Hafen, Kossman, and Brand propose a "digital data repository in a cloud solution provided by a trusted Swiss cloud-computing provider" [24]. Cooperatives, alone or as a federation, might instead follow the path of building community clouds, for even greater autonomy.

Such co-operatives open the door to previously unimagined data privacy practices. It is not inconceivable, for example, that in the foreseeable future, increased computing capacity and secure remote computation could allow data contributors themselves to hold the only extant copy of their genomic data on their own devices (perhaps even mobile devices), in a personal data store. The risk that individuals would inadvertently reveal their sensitive data could be mitigated or eliminated by the support provided by their co-operatives. These organizations would provide the necessary technological infrastructure to keep the data secure. A person might, for example, opt-in or out of allowing particular research projects to allow to run secure remote computational operations on their data. The data co-operatives could provide individual members with relevant information about each researcher or research organization. For example, their decision might be affected by knowing more about the strength of researchers' own security measures, about the goals and potential benefits of the research, or about the benefit to the co-operative based on their participation.

A bilateral consent framework was recently developed by Erlich et al. [28], which envisions a system that would assign a reputation value to individual researchers based on aggregated

participant "reviews". This type of arrangement is one approach through which health data co-operatives might assist informed participant decisionmaking. The approach could foster a sense of collective autonomy through dynamic consent [27], allowing the autonomy principle to be renewed and adapted instead of abandoned. It is difficult to imagine a system that would provide greater protection to, for example, state surveillance, than allowing participants themsleves to keep the only intelligible copy of their data.

Whether such personal genomic data stores are desirable or indeed feasible in the foreseeable future is open to question. The concept simply illustrates the overarching issue: if participants have no institutional role in determining the orientation of the development of genomic research, these possibilities will never be explored, developed, or implemented. The institutional forces driving genomic cloud computing will instead cause the field to develop by simply taking into account the minimum privacy protections necessary to ensure that participants continue their involvement.

## VI. Conclusion

Genomic cloud research has the means to develop along a path that is not only tolerable but welcomed by researchers and research participants alike. Yet, for the moment, we remain instead trapped in the thickets of maladjusted privacy norms. Participant participation, ironically, verges on non-existence.

Although the frustration of genomic researchers and bioinformaticians with the current morass of data privacy norms is understandable, there is a need to adjust our approach to data privacy laws and policies if we are to arrive at a coherent legal framework for research that protects and promotes the core values of research participants and researchers alike.

The strategy of aggressively pruning the thickets of data privacy law and research participant autonomy to enable free flows of research data should also encourage developing positive normative frameworks. To do so, it should help to create research entities whose very structures provide the necessary incentives to allow socially beneficial, responsible, ethical, and accountable genomic research to move forward, and as much of it as possible [29].

## References


[1] Mark I McCarthy et al, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," Nature Reviews Genetics, vol. 9:5, pp. 356–69. May 2008.

[2] W.M. Lowrance, Privacy, Confidentiality, and Health Research. Cambridge, UK: Cambridge University Press, 2012.

[3] Mark Taylor, Genetic Data and the Law: A Critical Perspective on Privacy Protection. Cambridge: Cambridge University Press, 2012.

[4] Institute of Medicine, Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington, DC: National Academies Press, 2009.

[5] Ma'n Zawati, "There will be sharing: population biobanks, the duty to inform and the limitations of the individualistic conception of autonomy," Health Law Journal, vol. 21, pp. 97–140.

[6] Beleno v. Texas Dept. of State Health Serv. Case 5:2009cv00188. U.S. District Court for the Western District of Texas in San Antonio, 3 March 2009.

[7] Bearder, *et al* v State of Minnesota, *et al*. Fourth Judicial Circuit, County of Hennepin District Court; 24 November 2009.

[8] L.D. v. Provincial Health Services Authority, 2011 BCSC 628; 2012 BCCA 491.

[9] Timothy Caulfield et al., "A review of the key issues associated with the commercialization of biobanks," J. Law and the Biosciences, vol. 1:1, pp. 94–110. 2014.

[10] Matt Warman, "EU privacy regulations subject to 'unprecedented lobbying'," The Telegraph., 8 Feburary 2012.

[11] Cynthia O'Donoghue, "EU research group condemns EU regulation for restricting growth in life sciences sector," Global Regulatory Enforcement Law Blog, 20 February 2014.

[12] David Lametti, "The cloud: boundless digital potential or enclosure 3.0?" Virginia J.L. & Tech., vol. 17:3, pp. 190–243. 2012.

[13] Electronic Frontier Foundation, "Frequently asked questions", online: eff.org/pages/frequently-asked-questions-about-building-for-a-digital-future.

[14] Tom Coughlin, "Nirvanix provides cautionary tale for cloud storage," Forbes, 30 September 2013.

[15] Edward S. Dove et al., "Genomic cloud computing: legal and ethical points to consider" Eu. J. Hum. Gen., 24 September 2014, doi: 10.1038/ejhg.2014.196.

[16] Liggett v. Lee, 288 U.S. 517, 557–60 (1933) (Brandeis J, dissenting).

[17] United States Department of Commerce, "Safe harbor privacy principles," online: export.gov/safeharbor/eu/eg_main_018475.asp.

[18] See the Attorney General of Canada's comments to this effect at para. 25 of its factum in Information and Privacy Commissioner of Alberta v. United Food and Commercial Workers, Local 401, online: scc-csc.gc.ca/factums-memoires/34890/FM040_Intervener_AGCanada.pdf.

[19] Amazon Web Services, "Now open – AWS Germany (Frankfurt) region – EC2, DynamoDB, S3, and much more," 23 October 2014, online: aws.amazon.com/blogs/aws/aws-region-germany.

[20] Public Service Commission (Re), 2013 CanLII 55439 (SK IPC).

[21] Khaled El Emam and Luk Arbuckle, Anonymizing Healh Data: Case Studies and Methods to Get You Started. North Sebastopol, CA: O'Reilly, 2013.

[22] Edward S. Dove, Yann Joly, and Bartha M. Knoppers, "Power to the people: a wiki-governance model for biobanks," Genome Biology, vol. 13:5, no. 158.

[23] David E. Winickoff, "From benefit sharing to power sharing: partnership governance in population genomics research," in Jane Kaye and Mark Stranger, eds, Principles and Practice in Biobank Governance. Farnham, Surrey: Ashgate, 2009, pp. 54–66.

[24] E. Hafen, D. Kossman and A Brand, "Health data cooperatives: citizen empowerment," Methods of Information in Medicine, vol. 2, pp. 82–86. 2014.

[25] Jeantine E. Lunshof, George M. Church, and Barbara Prainsack, "Raw personal data: providing access," Science Policy Forum, vol. 343:6169, pp. 373–74. 2014.

[26] For a helpful reconceptualization of autonomy in collective contexts, especially administrative institutions, see Jennifer Nedelsky, Law's Relations: A Relational Theory of Self, Autonomy, and Law. Oxford: Oxford University Press, 2011, pp. 118–26..

[27] Russell L. Bromley, "Financial stability in biobanking: unique challenges for disease-focussed foundations and patient advocacy organizations," Biopreservation and Biobanking, vol. 12:5, pp. 294–99. 2014.

[28] Yaniv Erlich et al., "Redefining genomic privacy: trust and empowerment," PLOS Biology, vol. 12:11, e1001983. 2014.

[29] Global Alliance for Genomics and Health, Framework for responsible sharing of genomic and health-related data. 2014.. http://genomicsandhealth.org/about-the-global-alliance/key-documents/framework-responsible-sharing-genomic-and-health-related-data.