Human Genetics (2018) 137:569–574
https://doi.org/10.1007/s00439-018-1923-y

EDITORIAL

Introduction: the why and whither of genomic data sharing

B. M. Knoppers[1,2] · **Yann Joly**[1]

Published online: 10 August 2018

Introduction

The Global Alliance for Genomics and Health (GA4GH) has estimated that, by the end of 2018, over 20% of genome and exome sequencing will be within and funded by healthcare systems for possible use in what can be termed "genomic medicine" (https://www.ga4gh.org). By 2030, it foresees that 83,000,000 rare-disease genomes will have been sequenced for diagnosis and 248,000,000 genomes will have been sequenced for cancer diagnosis (Birney et al. 2017). Faced with these overwhelming figures, the tendency is to search for technological and IT solutions to manage such data. Yet, unless such genomic data sharing is framed by common policies and the data linked to electronic medical records via harmonized and interoperable systems, it will not improve genomic variant interpretation or inform clinical decisions and targeted health care.

In 2017, the health data sharing landscape was described as an ecosystem with diverse stakeholders building the medical information commons (Deverka et al. 2017). Cook-Deegan and McGuire maintain that this medical information commons founded on both Ostrom's work on the knowledge commons and on the 1996 Bermuda Principles (HUGO 1996) require stakeholders' participation to be successful (Cook-Deegan and McGuire 2017). In 2017, the NIH announced a Data Commons Pilot Phase to explore using the Cloud to access and share FAIR biomedical big data (NIH 2017).

A current example of such an emerging ecosystem is, the "genomic commons"—the worldwide collection of publicly accessible repositories of human and non-human genomic data, built in part on the rapid public data release policies initiated by the Human Genome Project via the 1996 Bermuda Principles but also on concepts such as the "common heritage of mankind" (Knoppers 1991) and "global public goods" (HUGO Ethics Committee 2002). Today, this genomic commons serves as "an exemplar of polycentric, multi-stakeholder governance" (Contreras and Knoppers 2018).

By way of introduction then to this special issue of Human Genetics on *Genomic Data Sharing*, it is interesting and informative to compare the "Core Principles" of both the medical information commons (Deverka et al. 2017) with the "Conclusions" of the genomic commons (Contreras and Knoppers 2018) (Table 1). They reveal similar priorities:

It should be noted that, in 2017, the Organization for Economic Co-operation and Development (OECD) set forth its Recommendation on Health Data Governance. In particular, it reiterated the need for governments to "support transborder co-operation…remove barriers…[and] facilitate the compatibility or interoperability of health data governance frameworks." (Rec IV). This was followed in 2017, by UNESCO's revised recommendation on "Science and Scientific Researchers" which recognized in its preamble "the significant value of science as a common good" including "access to research results" art. 16 (v) as an ethical condition. Most importantly, it reiterated the human right to share in scientific advancement and its benefits (art. 21) (UNESCO 2017).

This largely dormant human right has its origins in the 1948 *Universal Declaration of Human Rights* (art. 27). It became art. 15 of the "legally actionable" *International Covenant on Economic, Social and Cultural Rights* in 1976. The *Covenant* has been signed and ratified by 167 countries signifying both legal actionability and accountability on the part of its signatories. Indeed, in the fall of 2018, the Committee on Economic, Social and Cultural Rights of the United Nations is due to report on the periodic national reports submitted by States parties, since the adoption of the *Covenant*. This report will include the interpretation of the

✉ B. M. Knoppers
bartha.knoppers@mcgill.ca

Yann Joly
yann.joly@mcgill.ca

[1] Faculty of Medicine, Human Genetics, Centre of Genomics and Policy, McGill University, Montreal, Canada

[2] Canada Research Chair in Law and Medicine, Montreal, Canada

Table 1 Comparison between the core principles of the Medical Information Commons and those of the Genomic Commons

Medical information commons (Deverka et al. 2017)	*Genomic commons* (Contreras and Knoppers 2018)
Core principles • "The MIC should be a healthy 'ecosystem' of data initiatives connected through a standard approach to policy, interoperability, and collaborative work" (principle 1) • "The MIC must bring together diverse sources of data from individuals with different states of health" (principle 2) • "A participant-centric model is critical for the sustainability of the MIC" (principle 3) • "It is important to reach out to and engage under-represented populations and to investigate the feasibility and acceptability of a public health approach" (principle 4) • "Building trust is an iterative process and requires investment of efforts beyond informed consent" (principle 5) • "Regulatory policies that rely on a sharp distinction between the 'kingdom of research' and the 'kingdom of clinical care' must be reconsidered" (principle 6) • "Changes in technology and in the scale and scope of data sharing demand reconsideration of current policy frameworks related to privacy and security" (principle 7) • "Distinguishing data ownership from data access and control is critical. Notions of unitary, exclusive property rights to data run counter to building the MIC" (principle 8)	Conclusions • Liberal data access and use policies for genomic data need to be adopted and include health data • Need for data interoperability and cross-border sharing • Promotion of the international human right to benefit from the fruits of scientific research • Proportionality approach to privacy that considers the benefits of sharing • Attention that intellectual property does not undermine data sharing

right of everyone to benefit from science and its applications under article 15.

A 2018 study of these periodic reports since 1976 reveals some interesting trends (Yotova and Knoppers 2018). Of the 123 states reportedly taking specific measures to implement the right to benefit from science, 76 state parties have adopted express legislative provisions to incorporate it into their domestic laws, and 83 reported taking concrete measures to promote the dissemination of science, databases being the most commonly adopted measure. This legal analysis of State actions is comforted by the 2017 American Association for the Advancement of Science (AAAS) survey (Wyndham et al. 2017) of its members on "Giving Meaning to the Right to Science." This survey ranked health (including diagnosis–treatment/applications) and advancing knowledge as 1 and 2 out of 3462 responses.

By way of introduction to this special issue on "Genomic Data Sharing" that includes reports from six countries, we can confirm the realization of this nascent right via two illustrative international initiatives from the last decade. Indeed, the international genome commons founded on data sharing is being constructed by the scientists themselves, as evidenced by the examples of the International Cancer Genome Consortium (ICGC) (1) and the Global Alliance for Genomics and Health (GA4GH) (2).

1. International Cancer Genome Consortium:[1]

The International Cancer Genome Consortium (ICGC), launched in October 2007, generates comprehensive catalogues of genomic abnormalities (somatic mutations, abnormal expression of genes, and epigenetic modifications) in various types of cancer across the globe. It makes these data available to the entire research community as rapidly as possible, and with minimal restrictions, so as to accelerate research into the causes and control of cancer. The original consortium involved 88 project teams studying over 25,000 tumor genomes in 17 jurisdictions. This initial phase of ICGC ended successfully in May 2018 with ICGC providing data on over 20,000 primary cancers, while the remainder is currently sequenced or in the process of being uploaded for sharing on the Ontario Institute for Cancer Research (OICR) and European Genome Archive (EGA) data portals.

The next phase of the project, ICGC ARGO,[2] is pursuing the ambitious goal of analyzing the genomes of more than 100,000 research participants by 2028, and linking these data to high-quality clinical information including treatment and outcomes. These rich new data will also be shared with the research community to advance cancer research and care worldwide.

At a time and age when data repositories, access committees, and other data oversight mechanisms and structures are rapidly multiplying, the Data Access Compliance Office (DACO) of the International Cancer Genome Consortium (ICGC) presents a good example of a federated model of

[1] http://icgc.org.

[2] https://icgcargo.org/.

providing access to the controlled data of a large-scale international research organization.

DACO has been active since July 2010 providing secure access to the controlled ICGC data to over 2000 users worldwide, with no significant privacy incidents being identified or reported by users or participants. Strong central leadership and the commitment of all member projects towards open data sharing resulted in an agreement to store ICGC data in one location (the Ontario Institute for Cancer Research in Toronto) and, with a single data access office, DACO located at McGill University in Montreal, providing access through a simple uniform data access agreement.[3] This commitment to centralization, along with the development of a concise access agreement that avoided overly legalistic or technical clauses, resulted in a highly efficient access process. More concretely, in 2017–2018, new applications were processed by DACO in only 5–7 days on average compared to a period going from several weeks to a few months in more decentralized infrastructures[4] or to infrastructures using more complex access agreements.[5] Such rapid approval process is based on the premise of access agreements being satisfactorily completed. Incomplete forms are a major source of approval delays.

The experience of ICGC DACO serves as a strong argument that, if controlled access is to be considered an acceptable substitute to completely open access for more sensitive genomic and clinical data, the process must be both centralized (via a few or a unique access committee) and streamlined (a single concise access agreement) as much as permitted by legal and ethical norms. It bears noting that the core reason justifying the controlled access process is the need to reasonably protect the privacy of human research participants (Toronto International Data Release Workshop Authors 2009). Clauses serving other purposes (e.g. IP protection, data embargos, legal disclaimers, etc.) are from this standpoint superfluous and should be removed if they create more than minimal additional delays or are a source of complexity for the data users.

The ICGC DACO case illustrates, as well, that choices regarding the type of data access processes and agreements to be used for a research organization should ideally be agreed upon and enforced from the earliest stage of project development. This is because it is much easier to impose common practices and standards on a group of researchers when they have a strong incentive to collaborate, for example a need to jointly apply for a grant or the desire to join a successful research consortium, rather than when they have already joined and benefited from their membership. The broad objectives, research topics, and security challenges may change over a repository's lifetime. However, unlike laws, access agreements and practices can be modified through simple, transparent, and deliberative processes to respond to emerging needs and issues.

An example of a transformative change that impacted ICGC policies was the important decision taken by ICGC, in late 2015, to use commercial cloud repositories to store and analyze its vast amount of data in collaborative projects and quality-control exercises. Given that ICGC members had amassed over two petabytes of genomic and clinical data in the first 5 years of ICGC, cloud computing seemed like the only practical and secure way to move forward. The cloud also promised a secure storage environment where data usage could be audited more rigorously to prevent and address cases of misuses. The limited experience in cloud computing for health research, in 2015, clearly demonstrated that the benefits of moving the data to the cloud, in terms of money and time, would be very substantial (Stein et al. 2015).

Large commercial providers were keen to work with ICGC offering to store its data for free, or at very competitive rates, to demonstrate their capabilities as an incentive for the research community to use their services in the future. However storing data in commercial clouds entailed new challenges related to, for instance, data security, portability, control, and deletion. The regulatory protection offered by cloud providers through contractual arrangements developed, mainly, for financial data storage, was suboptimal. For example, these contracts allowed cloud providers to shift data from one geographical location to another at will, change contractual terms with limited notification, and included extensive liability and non-responsibility clauses (Dove et al. 2015). However, a strong incentive for all stakeholders to advance genomic research through cloud-based services facilitated negotiations that led to contractual agreements better tailored for the storage and analysis of genomic and health data. A few more legalistic clauses were added to the DACO access agreement to reflect the specific environment where ICGC data would be hosted and so as to conform to the national privacy regulations of members' projects (ICGC, Application for Access to ICGC Controlled Data).

[3] The only outlier here being the United States: The Cancer Genome Atlas (TCGCA) agreed to a reciprocity agreement with ICGC where both TCGA and ICGC data sets would be hosted (mirrored) and accessible in the two repositories.

[4] For example, in the decentralized IHEC, it took several weeks for the faster DACs and months for the slower ones (IHEC Consortium, statistics compiled in the context of the EpiMap project by David Bujold and Guillaume Bourque, McGill University MUGQIC, 2018).

[5] For example, the China Kadoorie Biobank has a two-step process: registration (can take up to 2 weeks before approval) and Data Access Request (biobank takes 4–6 to review the agreement and respond).

2. Global Alliance for Genomics and Health (GA4GH):[6]

The GA4GH was founded in 2013 to accelerate progress in genomic science and human health by developing standards and framing policy for responsible genomic and health-related data. It seeks to both enable international data sharing across the translational continuum via technology-enabled federated approaches to data sharing (i.e., bringing analysis to the data) and to promote interoperability by prospectively harmonizing ethical and legal norms. This non-profit Alliance accelerates the potential of research and medicine to advance human health by bringing together 500+ leading organizations working in healthcare, research, patient advocacy, life science, and information technology. The GA4GH community is working together to create frameworks and standards to enable the responsible, voluntary, and secure sharing of genomic and health-related data. All work of the GA4GH builds upon its Framework for Responsible Sharing of Genomic and Health-Related Data (Knoppers 2014).

The founding principles of GA4GH's Framework for Responsible Sharing of Genomic and Health-Related Data seek to activate the right of everyone to share in scientific advancement and its benefits as outlined in the *Universal Declaration of Human Rights*, 1948 (article 27). This right to benefit was made legally actionable in the 1976 *International Covenant on Economic, Social and Cultural Rights* (United Nations OHCHR 1976). One hundred and sixty-seven countries have signed and ratified this *Covenant*. In particular, the GA4GH Framework advocates for:

- respect for the data sharing and privacy preferences of participants;
- transparency of governance and operations;
- accountability to best practices in technology, ethics, and public outreach;
- inclusivity by partnering and building trust among stakeholders;
- collaboration to share data and information to advance human health;
- innovation to develop an ecosystem that accelerates progress;
- agility to act swiftly to benefit those suffering with disease; and,
- independence by structure and governance.

The GA4GH Framework is available in 13 languages, as is its "Your DNA Your Say" public survey of attitudes and perceptions towards genetics. More importantly, GA4GH policies "activate" the human right to benefit from science and its applications by asking why, if data are consented to for sharing, researchers, and institutions are not held accountable if they refuse or neglect to do so. They provide an ethico-legal roadmap for the use of legacy data that may not have a specific consent for international data sharing as well as a more proportionate weighing of privacy risks based on real evidence. The regulatory and ethics workstream has developed policies on consent, privacy, and accountability, as well as a data sharing lexicon and ethics review recognition matrix (GA4GH 2017a). In addition, the security workstream of GA4GH also contributes standards for identity management and data privacy and security, and is working on breach notification.

In 2017, GA4GH restructured itself focussing on the need for standards, tools, and policy framing via 15 driver projects (GA4GH 2017b). These driver projects are supported by GA4GH's technical workstreams including genomic knowledge standards, data use and researcher IDs, large-scale genomics, clinical and phenotypic data, the cloud, and discovery. There are two foundational workstreams: security, and regulatory and ethics (REWS). They both work with the technical workstreams and driver projects. In 2018, the REWS focussed its attention on international data sharing in the context of research involving pediatric (Rahimzadeh et al. 2018) and dementia/aging populations (Thorogood et al. 2018).

Even prior to this restructuring, the GA4GH had a solid international reputation.[7] It had developed a standardized application programming interface (API). This API offers a defined protocol to allow disparate technology services of institutions around the globe to communicate with one another to exchange genotypic and phenotypic information. The API and the Framework were used in three demonstration projects spearheaded by GA4GH members. The first was the Beacon Project, an open technical specification for sharing genetic variant data sets collected from large-scale population-sequencing projects, clinical diagnostic settings, and variant curation efforts available to the community. A beacon is a Web-accessible service that allows data sets to be queried for the presence or absence of a specific allele. A user of a beacon can ask it questions of the form, "Have you observed this nucleotide (e.g., C) at this genomic location (e.g., position 32,936,732 on chromosome 13)?" to which the beacon must respond with either "yes" or "no." Beacon allows data discovery without exposing identifiable information, because it does not require data generators to share fully described data representations or annotations.

[6] https://www.ga4gh.org.

[7] This description of the early activities of the GA4GH is abstracted from Global Alliance for Genomics and Health (GA4GH) (2016), a federated ecosystem for sharing genomic, clinical data. Science 352:1278–1280.

The Beacon Network has enabled the discovery of genetic variants from over 500,000 subjects, and has been queried over one million times (https://beacon-network.org/#/).

This was followed by the BRCA Challenge aiming to advance understanding of the genetic basis of breast, ovarian, and other cancers that are driven by germline variants in BRCA1 and BRCA2. The Challenge built the BRCA Exchange, a publicly accessible Web portal that provides a simple interface for patients, clinicians, and researchers to access curated, expert interpretations of BRCA1/2 genetic variants, as well as supporting evidence (http://brcaexchange.org).

Another ongoing demonstration project is the Matchmaker Exchange (MME), a collaborative effort of consortia, including members of the International Rare Diseases Research Consortium and related laboratories in the rare-disease space, where the majority of cases studied lack a clear etiology after the initial analysis. To facilitate discovery, researchers in the rare-disease community have established a series of platforms that allow users to identify cases with phenotypes and disrupted genes in common. MME was established to connect rare-disease databases, such that a query to one would enable searches of the others, without having to deposit data into each one. With input from the REWS, MME has developed a two-tiered informed-consent policy to define the type of consent needed for using MME including when no consent is needed. If the data are associated with a unique or sensitive phenotype or with sequence-level data, consent from the patient is required to share it for research purposes. However, if only standard phenotype terms and candidate gene names are used, consent to clinical care already allows for consultation for clinical matchmaking. Still, challenges remain in balancing discovery with privacy and data protection (https://www.matchmakerexchange.org).

Conclusion

These two international initiatives serve as illustrations of the more "applied" context for policy and tool-making in the biomedical, genome scientific community. They illustrate the fact that, in spite of the differences in the regulatory approaches of the six countries examined in this special issue on "Genomic Data Sharing," ambitious international endeavours can both foster and achieve such sharing. The record of national data protection legislation and its integration across the jurisdictions reveals incredible complexity (if not contradiction and confusion). Indeed, the fact that genomic data are sensitive medical data has resulted in an overlap of additional protections accompanying personal and medical data protection generally. China reveals a particularly intimidating thicket of overlapping data regulations with very few transfers to third countries seemingly possible (e.g., cyber-security law). The other Asian country included in this series, South Korea, is in the midst of a profound reform of its privacy legislation that could result in a framework better suited for international data sharing initiatives.

The United States is also moving towards more favorable data sharing initiatives, for example, by legitimizing the concept of broad consent in its revised common rule. However, it still presents a fragmented data protection regime. Oversight is also distributed across a range of bodies, including institutional review boards and data access committees. Canada's regulatory framework is cautiously favorable to open science and international collaborations. Genomics researchers in this country have been heavily involved in open science projects and initiatives in recent years (HapMap Project, IHEL, GA4GH, ICGC, Toronto Statement on Data Sharing) and their successes. There have been no data breaches or misuse. Yet, the federated political structure of the country has prevented the establishment of one-stop Canadian data access repositories for clinical data until now.

Australia offers another example of a patchwork approach to privacy involving different sources of duties such as the common law, legislation, ethical guidelines and codes of practice. A welcome recent development is the NHMRC consultation into revisions to the human genomics sections of the *National Statement on Ethical Conduct in Human Research*. The proposed changes provide advice to researchers seeking to share genomic information, including obligations to minimise the potential for future re-identification, guidance on potential return of research findings, and requirements for ethical review of transfer agreements for sharing genomic information. Germany seems to have emerged from its traditional genetic exceptionalism to appreciate the need for genomic data sharing, albeit with some caution.

In short, this special issue on international genomic data sharing largely bears witness to the role and success of the self-regulatory mechanisms and standards set up by the scientists themselves. While the aim of genomic data sharing may hopefully be to "activate" the human right of everyone to benefit from science and its applications, ultimately, international scientific initiatives are based on good faith and mutual trust, the underlying philosophy being that genomic databases are global public goods.

Acknowledgements The authors would like to thank both the WYNG Foundation and Genome Canada/Quebec and the CIHR for its funding and Sophie Béland for her invaluable assistance.

References

Accelerating Research in Genomic Oncology (ARGO) https://icgcargo.org/

Birney E, Vamathevan J, Goodhand P (2017) Genomics in healthcare: GA4GH looks to 2022. BioRxiv. https://doi.org/10.1101/203554

Contreras JL, Knoppers BM (2018) The genomic commons. Annu Rev Genom Hum Genet 19:1.1–1.25. https://doi.org/10.1146/annurev-genom-083117-021552

Cook-Deegan R, McGuire AL (2017) Moving beyond Bermuda: sharing data to build a medical information commons. Genome Res 27:897–901

Deverka PA, Majumder MA, Villanueva AG et al (2017) Creating a data resource: what will it take to build a medical information commons? Genome Med 9:1–5. https://doi.org/10.1186/s13073-017-0476-3

Dove ES, Joly Y, Tassé AM, Knoppers BM, Public Population Project in Genomics and Society (P3G) International Steering Committee, International Cancer Genome Consortium (ICGC) Ethics and Policy Committee (2015) Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet 23:1271–1278

Global Alliance for Genomics and Health (GA4GH) (2016) A federated ecosystem for sharing genomic, clinical data. Science 352:1278–1280

Global Alliance for Genomics and Health (GA4GH) (2017a) Ethics review recognition policy. http://www.ga4gh.org/docs/ga4ghtoolkit/regulatoryandethics/GA4GH-Ethics-Review-Recognition-Policy.pdf. Accessed 14 June 2018

Global Alliance for Genomics and Health (GA4GH) (2017b) Strategic roadmap. http://www.ga4gh.org/docs/Strategic-Roadmap-print.pdf. Accessed 14 June 2018

Global Alliance for Genomics and Health (GA4GH) https://www.ga4gh.org

HUGO (1996) Summary of principles agreed at the first international strategy meeting on human genome sequencing (Bermuda, 25–28 February 1996). http://www.casimir.org.uk/storyfiles/64.0.summary_of_bermuda_principles.pdf. Accessed 14 June 2018

HUGO Ethics Committee (2002) Statement on human genomic databases. http://www.hugo-international.org/Resources/Documents/CELS_Statement-HumanGenomicDatabase_2002.pdf. Accessed 14 June 2018

International Cancer Genome Consortium (ICGC) Application for access to ICGC controlled data (version 2.3). http://icgc.org/daco

International Cancer Genome Consortium (ICGC) http://icgc.org

Knoppers BM (1991) Human dignity and genetic heritage: a study paper prepared for the law reform commission of Canada. The Commission, Ottawa

Knoppers BM (2014) Framework for responsible sharing of genomic and health-related data. HUGO J 8:1–6. https://doi.org/10.1186/s11568-014-0003-1

National Institutes of Health (NIH) (2017) NIH data commons pilot phase research opportunity announcement. https://commonfund.nih.gov/commons/faqs. Accessed 15 June 2018

Rahimzadeh V, Schickhardt C, Knoppers BM et al (2018) Key implications of data sharing in pediatric genomics. JAMA Pediatr 172:476–481

Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO (2015) Data analysis: create a cloud commons. Nature 523:149–151

Thorogood A, Mäki-Petäjä-Leinonen A, Brodaty H, Dalpé G, Gastmans C, Gauthier S, Gove D, Harding R, Knoppers BM, Rossor M, Bobrow M (2018) Consent recommendations for research and international data sharing involving persons with dementia. J Alzheimer Assoc. https://doi.org/10.1016/j.jalz.2018.05.011

Toronto International Data Release Workshop Authors (2009) Prepublication data sharing. Nature 461:168–170

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2017) Recommendation on Science and Scientific Researchers. http://unesdoc.unesco.org/images/0026/002636/263618e.pdf. Accessed 14 June 2018

United, Nations, Office of the High Commissioner - Human Rights (OHCHR) (1976) International covenant on economic, social and cultural rights. https://www.ohchr.org/EN/ProfessionalInterest/Pages/CESCR.aspx. Accessed 15 June 2018

Wyndham JM, Vitullo MW, Kraska K, Sianko N, Carbajales P, Nuñez-Eddy C, Platts E (2017) Giving meaning to the right to science: a global and multidisciplinary approach (Report prepared under the auspices of the AAAS Scientific Responsibility, Human Rights and Law Program and the AAAS Science and Human Rights Coalition). https://mcmprodaaas.s3.amazonaws.com/s3fs-public/reports/Right_to_Science_Report.pdf. Accessed 14 June 2018

Yotova R, Knoppers BM (2018) The right to benefit from science and big data. Eur J Int Law (**Submitted**)