scientific reports

Check for updates

OPEN

Data access arrangements in genomic research consortia

Dianne Nicol[1]✉, Jane Nielsen[1] & Madeleine Archer[2]

One of the most common terms that is used to describe entities responsible for sharing genomic data for research purposes is 'genomic research consortium'. However, there is a lack of clarity around the language used by consortia to describe their data sharing arrangements. Calls have been made for more uniform terminology. This article reports on a review of the genomic research consortium literature illustrating a wide diversity in the language that has been used over time to describe the access arrangements of these entities. The second component of this research involved an examination of publicly available information from a dataset of 98 consortia. This analysis further illustrates the wide diversity in the access arrangements adopted by genomic research consortia. A total of 12 different access arrangements were identified, including four simple forms (open, consortium, managed and registered access) and eight more complex tiered forms (for example, a combination of consortium, managed and open access). The majority of consortia utilised some form of tiered access, often following the policy requirements of funders like the US National Institutes of Health and the UK Wellcome Trust. It was not always easy to precisely identify the access arrangements of individual consortia. Greater consistency, clarity and transparency is likely to be of benefit to donors, depositors and accessors alike. More work needs to be done to achieve this end.

Keywords Genomic research consortia, Genomic data sharing, Access to genomic data

The genomic research landscape is teeming with hundreds of bodies, organisations, groups and teams which use a variety of terms to describe themselves. One of the most common terms that is used to describe many of these entities is 'consortium'. This term emerged in the early 2000s to describe a particular type of research collaboration[1]. Core elements identified in the literature include: collaboration and/or partnership; pooling of resources; and efforts to reach a common/shared goal[2–7]. In this article, we rely on the following definition of genomic research consortia: 'a group of scientists from multiple institutions who have agreed to cooperative research efforts involving, but not limited to, pooling of information from more than one study for the purpose of combined analyses and collaborative projects'[5].

The overarching purpose of genomic research consortia is to ensure that data generated by consortium members are shared for research purposes[8]. Establishing mechanisms to facilitate this, though vital to the genomics research effort, is not without scientific and technical challenges. The ongoing development of standards and policies for genomic data sharing by organisations like the Global Alliance for Genomics and Health (GA4GH) is addressing many of these challenges[9]. GA4GH also recognises these challenges are intimately interconnected with the regulatory and ethical concerns that arise whenever human genomic data is shared. The GA4GH Regulatory and Ethics Workstream (REWS) uses a global human rights framework to provide guidance on the regulatory, ethical, and social implications of genomic research and data sharing, and the development of standards and policies to address them[10,11].

Funding agencies have also started to impose their own standards for the governance of genomic research funded by them. The US-based National Institutes of Health (NIH), for example formulated a policy on sharing genomic data in 2014[12], which sits alongside its more recent broader data management and sharing policy[13]. One key feature is the expectation that researchers will ensure broad and responsible sharing. Investigators and their institutions are expected to provide the NIH with a Genomic Data Sharing Plan showing how the policy will be followed, and to submit large-scale human genomic and associated data to NIH-approved data repositories. Likewise, Wellcome in the UK has requirements regarding data access, including an explicit outputs management plan[14].

GA4GH has been in existence for only eleven years and the NIH Genomic Data Sharing Policy came into effect shortly after that. Prior to their existence, the growth in genomic research consortia occurred largely organically. Although many of them are now bound by the policies of the NIH and other funding agencies,

[1]Centre for Law and Genetics, University of Tasmania, Hobart, TAS, Australia. [2]Faculty of Business and Law, Australian Centre for Health Law Research, Queensland University of Technology, Brisbane, QLD, Australia. ✉email: Dianne.nicol@utas.edu.au

and/or are starting to adopt GA4GH standards, others continue to operate under their own governance arrangements. Assessing the adequacy and appropriateness of these governance arrangements is challenging, given the sheer number and diversity of entities that use this title. There are various reasons why it is important for genomic research consortia to be transparent in their governance arrangements, not least because they host data sourced from members of the public and generally obtained using public funding. In their examination of good governance of human genomic data, O'Doherty et al.[15] describe transparency as 'a meta-function of good governance'.

The research reported in this article examined governance of data access by genomic research consortia. It included a review relevant general academic literature, and creation of a dataset of academic literature and other documentation relating to specific consortia.

Materials and methods

Preliminary reviews of information relating to governance of data access in genomic research consortia were undertaken within the authors' existing library of over 800 academic articles and other documents. A search was then undertaken on PubMed using the terms: health/genetic/genomic/consortia/consortium. Eighty-nine documents were ultimately selected and compiled for analysis in 2019. These documents were supplemented by more recent literature during 2020 to 2024. This library was not intended to be a comprehensive catalogue of every published document about governance of data access in genomic research consortia. Rather, it was intended to provide us with a sufficiently large sample to show common themes and points of difference.

An initial list of genomic research consortia was drawn from a catalogue of genomic data sharing initiatives compiled by GA4GH (it should be noted that this catalogue is no longer available on the GA4GH website). We also scoured other publicly available lists of consortia, including: a compilation of data sharing repositories available on the NIH website[16]; repositories recommended by the journal Nature; and a list of GA4GH driver projects[17]. We acknowledge that by using GA4GH and NIH as our primary sources, there might be some skewing of our datasets towards North America. However, given the dominance of the US in genomic research and the NIH in funding this research, it is inevitable that the vast majority of consortia on our list would have their origins in this jurisdiction. We also note that it was not our intention that these searches would yield a definitive list of all genomic research consortia currently in existence. Instead, our aim was to create a reasonably comprehensive dataset of entities that self-identify as genomic research consortia. We acknowledge, however, that we may have missed some consortia which may have developed interesting and novel data access arrangements.

Based on our reliance on the definition of genomic research consortia proposed by Burgio et al.,[5] we excluded entities that were otherwise included in our search results on several grounds:

- entities that do not engage in research into genomics at the molecular level (for example the Human Cell Atlas), on the basis that their primary focus is not genomics;
- repositories, archives and databases that store genomic data, such as the US-based database of Genotypes and Phenotypes (dbGaP). dbGaP is a repository established by the NIH and other government agencies to facilitate sharing of genotypic and phenotypic information[18]. Although dbGaP and other repositories do not participate in genomic research per se, consortia that deposit data in these repositories may be bound by their data access policies. In this respect their policies are relevant to our analysis;
- entities whose only role is to analyse data, on the basis that they are not engaged in collaborative research; and
- funding bodies (such as NIH and Wellcome), and bodies whose role is to develop policies and standards for data sharing (such as GA4GH), given that their role is to facilitate data sharing rather than being actively involved in genomic research. The model data sharing policies and other documents developed by these bodies are nevertheless relevant because they are often adopted by genomic research consortia.

In total, 102 consortia were identified by one of us (MA) together with another research assistant. This dataset was subsequently analysed separately by DN and JN and reduced to 98 genomic research consortia. Four entries were removed because they were shown not to be actively involved in genomic research and one was removed because the consortium had been dissolved some years prior to our project and no information was available about its governance. Although other consortia have also completed their work, information on their governance arrangements is still available (as are their datasets). One entry was split into two (ICGC and ICGC-ARGO) on the basis that the nature of the consortium (including the data access arrangements) changed on the creation of ICGC-ARGO.

A dataset of foundational articles and governance policies, guidelines, agreements and other documentation made publicly available by individual genomic research consortia was created to facilitate examination of specific governance arrangements. Where preliminary searches failed to reveal details on data access arrangements, consortium administrators were contacted and asked to provide relevant links, which resulted in some further additions to the dataset. The dataset was reviewed and analysed separately by DN and JN in 2023 for details on access arrangements. We then jointly sorted access arrangements into categories to generate a shared terminology. Further online searches of consortia websites were undertaken where the original dataset did not reveal sufficient information. A full list of consortia, together with their purported access arrangements and other demographic features, is provided in Supplementary Materials. Access to the full dataset is available on request from the authors (except for some documents where the consortium administrators requested confidentiality).

Results

Preliminary observations on terminology

The European Expert Advisory Group on Data Access expressed concern about the lack of a shared terminology in the data sharing landscape in 2015. In response, the group attempted to impose some consistency in the terms

employed to describe data sharing processes and the actors involved in them[19]. In a further attempt to address concerns about the lack of common terminology, GA4GH developed a 'Data Sharing Lexicon' in 2016[20]. Some of the more well-resourced genomic research consortia also imposed their own requirements for consistency in language and have been highly transparent in defining and explaining this. An exemplar is the ICGC[21].

Despite these attempts to introduce a shared terminology, a host of terms continue to be employed to describe consortia activities. In particular, diverse terms have been used to describe the procedures involved in granting or facilitating data access. Some of these include: controlled access[22–25], managed access[21,26,27], tiered access[22,24], unrestricted access, open access[24,27,28], registered access[21,24], authenticated, charged, exclusive and password access[29], a passport model of access[30] and more. There is a lack of clarity about the extent to which these terms are truly distinct, related, or interchangeable with one another.

Towards a shared terminology for access arrangements

Recognising the broad range of language that used to describe data access arrangements in genomic research consortia, we designated five major categories from our analysis: consortium access, open access, registered access, managed access and tiered access. We describe each category in more detail below.

Consortium access

Given the rationale for the establishment of genomic research consortia is to share data, we expected that members of genomic research consortia share data with other members on a reciprocal basis[8]. However, it does not necessarily follow that members will be willing to share their data outside of the consortium[31].

We use the term consortium access (CA) to describe the circumstances where consortium members only share data openly within the consortium. This puts consortium members in a privileged position relative to other researchers. There are at least three circumstances when this privileged position could fall away. The first arises when consortium-generated genomic data is made openly available to all. The second arises when all that is required to become a consortium member is registration, after which broad data access is granted without the need to participate in joint research endeavours. The third is somewhat different; rather than making data access more open, a consortium may require an application to be made by anyone accessing consortium-generated genomic data, vetted by the consortium, irrespective of whether the user is a consortium member or not.

Open access

In contrast to the closed approach of CA, open access (OA) makes data publicly available without restriction[22,27]. The theory behind OA is that unfettered access assists in the verification and replication of data, broadens opportunities to pool data and generates results without the need to collect further data, leading to better quality results and establishment of community resources[27,32]. As noted above, when a consortium opts for OA this means that consortium members do not have privileged access.

The Human Genome Project (HGP) laid the foundation for OA. The HGP commenced in 1990, and from the outset, it was a collaborative venture, both between institutions and countries. The goals of the HGP were to map the genes and sequence the genetic code for the entire human genome. In 1996 HGP participants agreed in the Bermuda Declaration that primary genomic sequences should remain in the public domain and that they should be rapidly released[33,34]. GenBank was created as a publicly accessible repository of the sequence information produced by the HGP[35].

By 2003, there was some relaxation of the high standards for OA imposed by the Bermuda Declaration. In particular, it was felt necessary to recognise the valuable input provided by data generators. The Fort Lauderdale agreement reflected this desire[34]. Further inroads were made into these pure OA principles through the HapMap Project and the Genetic Association Information Network, led by the NIH[34]. Recognition of the input of data generators, protection of the privacy of data sources and avoidance of data capture for commercialisation purposes are all reasons why it is difficult to apply OA principles to all aspects of genomic data sharing[36]

Registered access

Registered access (RA) is used here to describe access that does not fully align with OA standards[23,24]. Researchers requesting access through this mechanism are not required to sign or abide by a formalised data transfer agreement. Rather, they may simply be required to accept the terms of use through a simple online agreement, by clicking 'I agree' in a checkbox or similar administratively 'light' procedure[24]. The GA4GH Data Sharing Lexicon describes RA as 'a system of authentication and self-declaration prior to providing access to data.'[20] The benefit of RA is that it creates a simple mechanism to verify the legitimacy of the user and to bind them to contractual terms regarding future uses of the data and other matters[23,24]. In this regard, it is less burdensome than managed access, considered below.

The HapMap Project marked the first step towards this RA approach. Its goal was to make sequence information available in a publicly accessible database[37]. The licence required users accessing the database to undertake that they would not restrict others from accessing or using the early-stage data they generated from the primary data. This obligation attracted some controversy, because it marked a significant change in philosophy from OA, although others saw it as an effective safeguard against capture[38].

In some instances, consortia apply RA principles not just for data access but for consortium membership. This is another instance when there is no privileged access for consortium members who participate in joint research endeavours.

Managed access

Another broad descriptor of access arrangements is managed access (MA). The MA approach requires that data access is managed or controlled by the consortium. In some instances, consortium members are required to

comply with the same formal MA arrangements as external users, which is another example of circumstances where privileged access to data for consortium members is lost. This can arise when consortium data is not held centrally but rather is held by the organisation generating the data. Arrangements of this nature are described as federated access[39–41]. The GA4GH Beacon Project is an example of a system designed to facilitate federated access (amongst other things) [42].

By imposing restrictions on data access[24], MA arrangements involve more regulatory action by the consortium. Conditions might include who can access, what they can access, and how and under what terms access may be granted (i.e. on an internal server or downloaded) [22]. The term MA is not used ubiquitously; 'controlled access' and 'restricted access' are also common[22,24,25]. For instance, the ICGC uses controlled access to describe its access system for 'composite genomic and clinical data that are associated to a unique, but not directly identifiable, person', contrasting with its OA system for non-identifiable data[22]. Dyke et al. appear to use the terms MA and controlled access interchangeably[23]. The GA4GH Lexicon suggests that restricted access and controlled access are equivalent[20]. On this basis, we conclude that these terms are variations on the same theme—managing who can access data and on what terms, and for this reason, we refer to them collectively as MA.

Commonly, this degree of control is justified by the nature of the data, or to comply with conditions imposed by participant consent forms or by research ethics committees[20,27]. Dyke et al. suggest that genomic research consortia that employ MA should, by definition, have a Data Access Committee (DAC)[23]. The primary functions of a DAC are to: provide broad oversight of data access and to facilitate and expedite the data access process; review applications for access; make determinations as to access of researchers and others to data, including by reference to the qualifications and legitimacy of the researchers and the purposes specified for their research; and impose and ensure compliance with conditions upon data use[19,26,43]. Despite these relatively common features of DACs, there is a significant degree of variation in their purposes and activities[43,44].

In summary, although DACs are widely recognised as the oversight bodies to whom researchers must apply to access data, there is a lack of clarity about: the extent to which their requirements are mandatory, or universal[26,32,43,45]; the level of substantive or non-substantive oversight they wield in relation to researcher requests for access to data; and the criteria they apply to decisions about data-access[19,32,43–46]. This uncertainty calls into question the extent to which *having* a DAC ensures accountability, security and compliance with conditions and restrictions on data access.

Tiered access

It might be assumed that the four categories of data access listed above conveniently describe the gamut of access arrangements used by genomic research consortia. Our research and the work of others shows that the reality is much more complex. Tiered access (TA) is used to describe these more complex access arrangements of genomic research consortia.

The TA model provides genomic research consortia with the opportunity to segment their datasets and choose different access arrangements that best suit the consortium, its members and data sources. For example, it may be used to allow consortium members to retain privileged access up to the point of publication of their early findings. The NIH Genomic Data Sharing Policy explicitly allows delayed access for these purposes[12].

This approach may also be used to provide external researchers with OA to certain data types, but require them to go through a more formalised process for other data types. This approach can balance competing concerns relating to scientific progress and protecting donor privacy[22]. The OA model might be applied to a range of data types including: data that cannot be linked in order to identify donors; aggregate level data; or data that does not contain information that could allow donors to be identified[19,22]. By contrast, access to more sensitive data, particularly data that may allow donors to be identified, is more likely to be subjected to the controlled procedures of MA. This means that users are likely to be required to make an application to the consortium's DAC and enter into a Data Access Agreement before they are able to use the sensitive data[19,22].

The ICGC was one of the early adopters of this TA approach[21]. dbGaP also requires depositors to use a TA approach to data access by external users, applying OA to studies, study documents, phenotypic variables and genotype access, and MA (which it refers to as controlled access) to de-identified data, phenotypes and genotypes for individual study subjects and pedigrees[47].

Analysis of specific genomic data access consortia arrangements

Each of the consortia in our dataset has expressed an explicit commitment to genomic data access in one way or another. Some provide clear guidance on the ways in which they share their data, through their websites and through other documentation, making classification relatively straightforward. For a surprisingly large number of consortia, however, it took us some time to clearly and precisely identify data access arrangements, requiring extensive searching of our datasets, further online searches and email requests to consortium administrators. For 32 consortia (33% of the dataset), we had to infer data access arrangements based on: funding sources (for example, consortia funded by the NIH or Wellcome have to comply with their policies); repositories within which their data were stored (for example, dbGaP has certain requirements); nature of the consortium; types of data that the consortium generates; and other factors. Information on the main sources used to determine data access is provided for each consortium in the Supplementary Materials.

In sum, we are reasonably confident that the express and inferred data access arrangements we have recorded provide a reasonably comprehensive account of the variety of arrangements across our consortium dataset, and the prevalence of some of these arrangements within the sector as a whole. In Table 1, we list the options for genomic data access in genomic research consortia ranging from CA to highly sophisticated models using up to three access categories. It should be noted that permutations of access arrangements which were not observed in the dataset are not included in Table 1. In each case where there is at least one genomic research consortium

Acronym	Title	Description	Exemplar	Number of Consortia
CA	Consortium Access	Pre- and post-publication data sharing • only with consortium members, with no evaluation, no terms	German Cancer Consortium (DKTK)	7
OA	Open Access	Pre- and post-publication data sharing: • with both consortium members and external researchers, with no registration, no evaluation, no terms	Human Genome Project	5
RA	Registered Access	Pre- and post-publication data sharing: • with both consortium members and external researchers, with registration (either to access the data or to join the consortium), but no evaluation other than identity verification, no terms	Evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA)	1
MA	Managed Access	Pre- and post-publication data sharing: • with both consortium members and external researchers, with registration, evaluation and terms (usually determined by a DAC)	International Cerebral Palsy Genetics Consortium (ICPGC)	8
TA (CA-OA)	Tiered Access (Consortium Access and Open Access)	Pre- and post-publication data sharing For some data • only with consortium members, with no evaluation, no terms For other data (possibly only post-publication) • with no registration, no evaluation, no terms	Genome in a Bottle Consortium (GIAB)	5
TA (CA-RA)	Tiered Access (Consortium Access and Registered Access)	Pre- and post-publication data sharing • only with consortium members, with no evaluation, no terms Additional post-publication data sharing for external researchers • with external researchers, with registration, but no evaluation other than identity verification, no terms	Glioma Longitudinal Analysis Consortium (GLASS)	2
TA (CA-MA)	Tiered Access (Consortium Access and Managed Access)	Pre- and post-publication data sharing • only with consortium members, with no evaluation, no terms Additional post-publication data sharing for external researchers • with external researchers, with registration, evaluation and terms (usually determined by a DAC)	Malaria Genomic Epidemiology Network (MalariaGEN)	14
TA (CA-OA-MA)	Tiered Access (Consortium Access, Open Access, and Managed Access)	Pre- and post-publication data sharing For some data • only with consortium members, with no evaluation, no terms For other data (possibly only post-publication) • with no registration, no evaluation, no terms Additional post-publication data sharing for external researchers • with external researchers, with registration, evaluation and terms (usually determined by a DAC)	Most consortia governed by NIH policies	41
TA (OA- MA)	Tiered Access (Open Access and Managed Access)	Pre- and post-publication data sharing For some data • with no registration, no evaluation, no terms Additional post-publication data sharing for external researchers • with both consortium members and external researchers, with registration, evaluation and terms (usually determined by a DAC)	The Brazilian Initiative on Precision Medicine (BIPMed)	10
TA (OA-RA-MA)	Tiered Access (Open Access Registered Access and Managed Access)	Pre- and post-publication data sharing For some data • with no evaluation, no terms For other data (possibly only post publication) • with registration, but no evaluation other than identity verification, and no terms for some data • with registration, evaluation and terms for other data (usually determined by a DAC)	The Variant Interpretation for Cancer Consortium (VICC)	1
TA (RA-OA)	Tiered Access (Open Access and Registered Access)	Pre- and post-publication data sharing For some data • with no registration, no evaluation, no terms Additional post-publication data sharing for external researchers • with both consortium members and external researchers, with registration (either to join the consortium or access the data), but no evaluation other than identity verification, no terms	International HapMap Project	1
TA (RA-MA)	Tiered Access (Registered Access and Managed Access)	Pre- and post-publication data sharing For some data • with registration, but no evaluation other than identity verification, and no terms for some data For other data (possibly only post publication) • with registration, evaluation and terms for other data (usually determined by a DAC)	International Multiple Sclerosis Genetics Consortium (IMSGC)	3

Table 1. Options for data sharing in genomic research consortia.

using the stated access options, we have identified an exemplar and listed the total number of consortia in our dataset which appear to use this arrangement, noting again our caution that the actual arrangements are more difficult to accurately identify for some consortia than others.

Discussion

This study is testament to the growth of genomic research since the HGP. It illustrates that genomic research consortia are providing the essential engine rooms for this research effort, by providing access to large and diverse sets of genomic data. The Supplementary Materials provide ample evidence of volume and diversity of genomic datasets, with some consortia being highly disease specific, others focusing more generally on classes of disease (for example, cancers) and large-scale population-wide initiatives.

This study also illustrates an increasing sophistication in data access arrangements. The simple dichotomy between OA and closed CA has been replaced with more tiered approaches, with TA (CA-OA-MA) being most prevalent (with 41 consortia manifesting this structure). Some data access arrangements are relatively uncommon. Complete OA, for example, is unusual, seen only rarely post-HGP. Far more commonplace is the open sharing of summary results, generally post-publication. Somewhat unanticipated also was the lack of reliance on RA as a mechanism for controlling the movement of data. Even within a TA structure, mechanisms that provided greater control (through MA) appear to be favoured over systems that simply require a process of registration.

One thing that is clear is that consortia continue to be committed to making their datasets as open as possible. The Variant Interpretation for Cancer Consortium[48] provides a good example. Although it describes itself as an 'open consortium', consortium documents set out a commitment to 'share… at least a minimal set of data elements for cancer variant interpretations including: gene symbol, variant name, cancer subtype (tumor type and organ), clinical implication (drug sensitivity, drug resistance, adverse response, diagnostic, or prognostic), source (e.g., PubMed identifier) and curation group.' Justifiable concerns around the privacy of data donors and primary researcher interests no doubt motivate the adoption of a limited definition of 'open' access in this and other consortia.

The dominance of the TA(CA-OA-MA) model is not unexpected, given the mandates from funders like the NIH and Wellcome. This arrangement appropriately recognises the value in making some information freely available, whilst acknowledging that there is a need to provide appropriate protections to donors and their families. It further acknowledges the valuable contributions made by consortium members, by allowing them to publish the results of their initial analyses before data is made more broadly available. As consortia move towards more federated access models, it seems likely that tiered access arrangements of this nature will become the norm.

Table 1 illustrates the intricacies of data access arrangements in genomic research consortia. Despite calls for standardisation, the level of diversity is striking. These differences are, unsurprisingly, more pronounced with consortia unaffiliated with major funders, such as the NIH or Wellcome. This demonstrates that, absent mandated adherence to policy arrangements, consortia tend to adopt access arrangements that best suit the interests of members. Beyond this, what is more striking is the effort it took us to precisely identify data access arrangements for many of the consortia in our dataset. As noted earlier, transparency is recognised as one of the fundamental requirements for good governance in genomics[15].

Transparency and accuracy in articulation of data access arrangements is important for a number of reasons. For consortium members, clear parameters around access are essential to ensure that they are fully cognisant of arrangements in respect of the data they contribute to the consortium. Whether data is truly openly available, for example, may influence a decision as to whether or not to contribute to a particular consortium. For data users, certain access structures undoubtedly complicate the process of requesting data. Managed access provides a level of protection to consortium members and donors, but is more administratively burdensome[22]. A lack of transparency around the process for requesting access is likely to constitute a significant disincentive for users.

The adoption of consistent data access arrangements by consortia funded by the NIH and other bodies is an important step in simplifying the genomic data sharing landscape. We are aware that some of the other consortia in our dataset are also in the process of updating their access arrangements to more closely reflect developing norms within the sector. Nevertheless, the analysis presented in this article highlights the need for more work to achieve a shared, consistent terminology for access arrangements within genomics consortia and for greater transparency about these access arrangements. Not only is this important for consortium members and users of consortium data, but it is also critical in recognising the valuable contributions made by donors.

Data availability

A list of all of the genomic research consortia in the dataset is provided in the supplementary materials. The full set of consortia-related documents is available from the authors on request.

Received: 4 April 2024; Accepted: 9 September 2024
Published online: 17 September 2024

References

1. Mittleman, B., Neil, G. & Cutcher-Gershenfeld, J. Precompetitive consortia in biomedicine—how are we doing?. *Nat. Biotechnol.* **31**, 979–985 (2013).
2. Leonelli, S. (2009) Centralising labels to distribute data: The regulatory role of genomic consortia. In *The Handbook of Genetics and Society: Mapping the New Genome Era* (eds. Atkinson, P., Glasner, P., Lock, M.) 469–485 (Routledge, 2009).
3. Österle, H. & Otto, B. Consortium research. *Bus. Inf. Syst. Eng.* **2**, 283–293 (2010).
4. Wallace, S. E. The needle in the haystack: international consortia and the return of individual research results. *J. Law. Med. Ethics* **39**, 631–639 (2011).
5. Burgio, M. R. *et al.* Collaborative cancer epidemiology in the 21st century: the model of cancer consortia. *Cancer Epidemiol. Biomarkers Prev.* **22**, 2148–2160 (2013).
6. Papadaki, M. & Hirsch, G. Curing consortium fatigue. *Science Trans. Med.* **5**, 200fs35 (2013).
7. Teare, H. J. A. *et al.* The governance structure for data access in the DIRECT consortium: an innovative medicines initiative (IMI) project. *Life Sci. Soc. Policy* **14**, 1–17 (2018).
8. Burton, P. R. *et al.* Policies and strategies to facilitate secondary use of research data in the health sciences. *Int. J. Epidemiol.* **46**, 1729–1733 (2017).
9. Rehm, H. L. *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genomics* **1**, 100029. https://doi.org/10.1016/j.xgen.2021.100029 (2021).
10. Knoppers, B.M. (2014) Framework for responsible sharing of genomic and health-related data. *HUGO J.* **8**, Article 3 https://doi.org/10.1186/s11568-014-0003-1 (2014)

11. Joly, Y., Dove, E., Knoppers, B.M. & Nicol, D. The GA4GH Regulatory and Ethics Work Stream (REWS) at 10: An interdisciplinary, participative approach to international policy development in genomics. In *The Law and Ethics of Data Sharing in Health Sciences. Perspectives in Law, Business and Innovation*: (eds. Compagnucci, M.C., Minssen, T., Fenwick, M., Aboy, M. & Liddell. K.) 13–32 (Springer, 2024).
12. NIH, *Genomic Data Sharing Policy* (2014). Available at: https://sharing.nih.gov/genomic-data-sharing-policy (accessed 4 September 2024).
13. NIH, *Policy for Data Management and Sharing* (2023). Available at: https://oir.nih.gov/sourcebook/intramural-program-oversight/intramural-data-sharing/2023-nih-data-management-sharing-policy (accessed 4 September 2024)
14. Wellcome, *Data, Software and Materials Management and Sharing Policy* (undated). https://wellcome.org/grant-funding/guidance/policies-grant-conditions/data-software-materials-management-and-sharing-policy (accessed 4 September 2024)
15. O'Doherty, K. C. *et al.* Toward better governance of human genomic data. *Nat. Genet.* **53**, 2–8 (2021).
16. NIH, *NIH-Supported Data Sharing Resources* (undated). Available at: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html (accessed 4 September 2024)
17. GA4GH, *Driver Projects* (undated). Available at: https://www.ga4gh.org/our-community/driver-projects/# (accessed 4 September 2024)
18. Mailman, M. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
19. Expert Advisory Group on Data Access, *Governance of Data Access* (2015). Available at: https://wellcomecollection.org/works/zzr2qqpk/items (accessed 4 September 2024)
20. GA4GH, *Global Alliance for Genomics and Health Data Sharing Lexicon* (2016). Available at: https://www.ga4gh.org/document/archived-data-sharing-lexicon/ (accessed 4 September 2024)
21. International Cancer Genome Consortium, *Goals, Structure, Policies and Guidelines, International Cancer Genome Consortium* (2008). Available at: https://www.riken.jp/medialibrary/riken/pr/topics/2008/20081119_1/20081119_1_en.pdf (accessed 4 September 2024)
22. Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M. & Chalmers, D. Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). *PLoS Comput. Biol.* **8**, e1002549. https://doi.org/10.1371/journal.pcbi.1002549 (2012).
23. Dyke, S. *et al.* Registered access: a 'Triple-A' approach. *Eur. J. Hum. Genet.* **24**, 1676–1680 (2016).
24. Joly, Y., Dyke, S. O. M., Knoppers, B. M. & Pastinen, T. Are data sharing and privacy protection mutually exclusive?. *Cell* **167**, 1150–1154 (2016).
25. Learned, K. et al. Barriers to accessing public cancer genomic data. *Sci. Data* **6,** Article 98, https://doi.org/10.1038/s41597-019-0096-4 (2019)
26. Kaye, J. & Hawkins, N. Data sharing policy design for consortia: challenges for sustainability. *Genome Med,* **6,** Article 4, https://doi.org/10.1186/gm523
27. Muddyman, D., Smee, C., Griffin, H., Kaye, J. & UK10K Project. The UK10K Project. Implementing a successful data-management framework: the UK10K managed access model. *Genome Med.* **5,** Article 100, https://doi.org/10.1186/gm504 (2013).
28. Joly, Y., Zeps, N. & Knoppers, B. M. Genomic databases access agreements: legal validity and possible sanctions. *Hum. Genet.* **130**, 441–449 (2011).
29. Fortin, S., Pathmasiri, S., Grintuch, R. & Deschênes, M. Access arrangements' for biobanks: a fine line between facilitating and hindering collaboration. *Public Health Genomics* **14**, 104–114 (2011).
30. The All of Us Research Program Investigators. The "All of Us" research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
31. Villanueva, A. G. *et al.* Characterizing the biomedical data-sharing landscape. *J. Law Med. Ethics* **47**, 21–30 (2019).
32. Murtagh, M.J. et al. Better governance, better access: practising responsible data sharing in the METADAC governance infrastructure. *Hum. Genomics* **12,** Article 24 **,**https://doi.org/10.1186/s40246-018-0154-6 (2018)
33. Bentley, D. R. Genomic sequence information should be released immediately and freely in the public domain. *Science* **274**, 533–534 (1996).
34. Cook-Deegan, R., Ankeny, R. A. & Maxson Jones, K. Sharing data to build a medical information commons: From Bermuda to the Global Alliance. *Annu. Rev. Genomics Hum. Genet.* **18**, 389–415 (2017).
35. National Centre for Biotechnology Information, *GenBank overview* (undated). Available at: https://www.ncbi.nlm.nih.gov/genbank/ (accessed 4 September 2024)
36. Gold, E.R. & Nicol, D. Beyond open source: patents, biobanks and sharing. In *Comparative Issues in the Governance of Research Biobanks: Property, Privacy, Intellectual Property, and the Role of Technology.* (eds. Pascuzzi, G., Izzo, U. & Macilotti, M.) 191–208 (Springer, 2013).
37. Chokshi, D.A. & Kwiatkowski D.P. Ethical challenges of genomic epidemiology in developing countries. *Genomics, Society and Policy 1,* Article 1 https://doi.org/10.1186/1746-5354-1-1-1 (2005).
38. Kent, A. Patients and IP – Should we care?. *SCRIPT-ed* **3**, 260–264 (2006).
39. Chaterji, S. *et al.* Federation in genomics pipelines: Techniques and challenges. *Brief. Bioinform.* **20**, 235–244 (2019).
40. Stark, Z. *et al.* Integrating genomics into healthcare: A global responsibility. *Am. J. Hum. Genet.* **104**, 13–20 (2019).
41. Thorogood, A. *et al.* International federation of genomic medicine databases using GA4GH standards. *Cell Genomics* **1**, 100032. https://doi.org/10.1016/j.xgen.2021.100032 (2021).
42. Fiume, M. *et al.* Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol.* **37**, 220–224 (2019).
43. Shabani, M., Thorogood, A. & Borry, P. Who should have access to genomic data and how should they be held accountable? Perspectives of Data Access Committee members and access. *Eur. J. Hum. Genet.* **24**, 1671–1675 (2016).
44. Lawson, J., Rahimzadeh, V., Baek, J. & Dove, E. S. Achieving procedural parity in managing access to genomic and related health data: A global survey of data access committee members. *Biopreserv Biobank.* **22**, 123–129 (2023).
45. Shabani, M., Dyke, S. O., Joly, Y. & Borry, P. Controlled access under review: improving the governance of genomic data access. *PLoS Biol.* **13**, e1002339. https://doi.org/10.1371/journal.pbio.1002339 (2015).
46. Cheah, P.Y. & Piasecki, J. Data access committees. *BMC Med. Ethics* 21 Article 12 https://doi.org/10.1186/s12910-020-0453-z (2020).
47. Tryka, K.A. et al. (2013) The Database of Genotypes and Phenotypes (dbGaP) and PheGenI. In: The NCBI Handbook [Internet]. 2nd edition. (National Center for Biotechnology Information (US), 2013). Available at: https://www.ncbi.nlm.nih.gov/books/NBK154410/ (accessed 4 September 2024)
48. VICC, Variant Interpretation for Cancer Consortium (undated). Available at: https://cancervariants.org/index.html (accessed 4 September 2024)

Acknowledgements

The authors acknowledge and thank other members of the Centre for Law and Genetics at the University of Tasmania for their encouragement and advice throughout this project and thank the Centre's research assistants for their assistance with data collection, storage and preliminary analysis. We particularly acknowledge and thank Rachel Hay, who made a significant contribution to the early data collection and preliminary analysis, and Hugh Oxbrough, who assisted with finalising the consortium documentation.

Author contributions

DN and JN conceptualised the project and contributed equally to data collection post-2021, data analysis and writing of the manuscript. MA undertook data collection and preliminary analysis. She wrote summary of the literature and reviewed the final manuscript.

Funding

This research was funded through the Australian Research Council's Discovery Project scheme, DP180100269.

Declarations

Competing interests

Dianne Nicol is co-lead of the Regulatory and Ethics Workstream of the Global Alliance for Genomics and Health. The authors have no other competing interests.

Ethics approval

This project received ethics approval from the University of Tasmania Human Research Ethics Committee (Ethics Project ID: 18556). Ethics approval was not required for the aspect of the project reported in this article as all information was publicly available. Other aspects of this project (not reported in this article) involved interviews with consortium administrators for which ethics approval was granted.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-72653-z.

Correspondence and requests for materials should be addressed to D.N.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.