A call to action to scale up research and clinical genomic data sharing

Zornitza Stark[1,2,3†], David Glazer[5†], Oliver Hofmann[1,3,5], Augusto Rendon[6], Christian R. Marshall[7], Geoffrey S. Ginsburg[8], Chris Lunt[8], Naomi Allen[9,10], Mark Effingham[10], Jillian Hastings Ward[11], Sue L. Hill[12], Raghib Ali[13,14,15], Peter Goodhand[16,17], Angela Page[16], Heidi L. Rehm[16,18,19], Kathryn N. North[1,3,20] & Richard H. Scott[6,21,22]

[1]Australian Genomics, Melbourne, VIC, Australia

[2]Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, VIC, Australia

[3]University of Melbourne, Melbourne, VIC, Australia

[4]Verily Life Sciences, South San Francisco, CA, USA

[5]University of Melbourne Centre for Cancer Research, Melbourne, VIC, Australia

[6]Genomics England, London, UK

[7]Division of Genome Diagnostics, Pediatric Laboratory Medicine Department, The Hospital for Sick Children, Toronto, ON, Canada

[8]All of Us Research Program, National Institutes of Health, Bethesda, MD, USA

[9] Nuffield Department of Population Health, University of Oxford, Oxford, UK

[10]UK Biobank, Stockport, Greater Manchester, UK

[11]Participant Panel, Genomics England, London, UK

[12]National Health Service England, London, UK

[13]Our Future Health, Manchester, United Kingdom

[14]Oxford University Hospitals NHS Trust, Oxford, United Kingdom

[15]MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom

[16]Global Alliance for Genomics and Health, Toronto, ON, Canada

[17]Ontario Institute for Cancer Research, Toronto, ON, Canada

[18]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

[19]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

[20]Murdoch Children's Research Institute, Melbourne, VIC, Australia
[21]Great Ormond Street Hospital for Children, London, UK
[22]UCL Great Ormond Street Institute of Child Health, London, UK
[†]e-mail: zornitza.stark@vcgs.org.au; dglazer@verily.com

Abstract

Genomic data from millions of individuals have been generated worldwide to drive discovery and clinical impact in precision medicine. Lowering the barriers to using these data collectively is needed to equitably realize the benefits of population data diversity and scale. We examine the current landscape of global genomic data sharing, including the evolution of data sharing models from data aggregation, through to data visiting, and for certain use cases, cross-cohort analysis using federated approaches across multiple environments. We highlight emerging examples of best practice relating to participant, patient and public engagement; evolution of technical standards, tools and infrastructure; and impact of research and healthcare policy. We outline twelve actions we can all take together to scale up efforts to enable safe global data sharing and move beyond projects demonstrating feasibility to routinely cross-analysing research and clinical datasets, optimizing benefit.

[H1] Introduction

The promise of precision medicine and targeted therapeutics coupled with advances in sequencing technologies and analytics have driven declining costs and increasing investments in genomics over the past decade, resulting in large and growing volumes of data[1]. Sequencing large population cohorts has improved our understanding of genetic variation[2], while analysis of data from well-characterized individuals has revealed the underlying molecular basis of thousands of rare and common diseases, including multiple cancer types[3-5]. In turn, this has led to unprecedented acceleration in the development of precision treatments[6,7].

Genomic testing in rare and inherited disease and cancer has been shown to increase the rate, speed and accuracy of diagnosis[8-11]. Genomic data are increasingly generated as part of clinical care, enabling personalized and stratified risk assessments, treatments and interventions. Considerable efforts are currently underway to diversify genomic data and improve equitable outcomes by focusing on historically, culturally and geographically under-represented populations[12-14]. With sequencing and analytical capabilities continuously improving, population-level genomic screening is being explored in healthy adults[15,16] and in newborns[17] in a move towards pre-emptive and preventative health care.

As a result, linked genomic, phenotypic and other metadata from millions of individuals are being generated worldwide, offering new opportunities to drive research, with feedback to patients, participants and health system planners, providing evidence-based insights to improve clinical care. We now need to lower the barriers to using these data collectively so that individuals and whole populations can benefit equitably from data diversity and scale.

In this Roadmap, we examine the current landscape of genomic data sharing and the actions we can all take together to move forward. While the practical, ethical and legal challenges have been discussed extensively elsewhere[18-30], we provide a summary of recent successes and examples of best practices, together with a 12-point call to action.

[H1] Genomic data sharing approaches and models

Global data sharing has been integral to advance genomic medicine in the past two decades. Data sharing models vary from pooled approaches that aggregate and harmonize data from disparate sources, to federated approaches where individual data owners retain control but allow remote analysis of data, to siloed approaches that have not yet considered access beyond their initial audience (**Figure 1**). These models aim to honour the expectations of participants and patients — that donated data will have a positive impact but will only be used as intended. Achieving this delicate balance between data accessibility and utility on one hand and data governance and security on the other has been recognized by the World Health Organization (WHO) in a recent consultation on human genome data access, use and security[31], which highlights the need for collaboration, transparency and accountability.

[H2] Aggregated databases

International efforts to aggregate data into single, harmonized, open databases have been central to achieving accurate genomic variant interpretation at scale. The latest release of gnomAD (Genome Aggregation Database)[2], already an essential tool for understanding and interpreting normal variation, contains reference population data from 807,162 individuals, totalling over a billion variants. Over 4 million clinical interpretations of variants have already been deposited by more than 2,600 submitters in the openly accessible ClinVar database[32,33], facilitating accurate and consistent assessment of disease-associated variation. More focused variant aggregation efforts for specific genes, such as the BRCA Exchange[34], within individual countries[35] or by expert groups[36] are improving clinical care by driving efforts to identify and resolve discordant interpretations. Similar data aggregation and harmonization efforts have resulted in more rigorous assessments of the validity of gene–disease relationships, for example, through the Clinical Genome Resource[37], PanelApp platforms[38] and the Gene Curation Coalition[39].

[H2] Federated discovery and analysis

In federated approaches to data sharing, individual data custodians maintain their own databases but facilitate access by adopting enabling policies and interoperable

standards. Analysis across federated datasets can be done independently on each dataset, with merging of results, or, if allowed, by temporarily merging relevant data subsets into a common analysis environment. At the simplest level, this approach can allow remote data interrogation to answer specific queries. For example, Matchmaker Exchange has enabled international connections between 11 clinical and genomic data platforms to accelerate over 1,000 novel rare-disease-gene discoveries within just a few years by matching candidate gene names[40]. Other relatively simple federated queries, such as the presence or absence of a specific variant in datasets, have been enabled by the Beacon application programming interface standard, developed by Discovery workstream members of the Global Alliance for Genomics and Health (GA4GH) [41]. The second iteration of Beacon[42] allows more complex queries that incorporate phenotypic and variant information, including queries on variant type (for example, predicted loss of function) within a specified gene[43].

[H2] Cloud-based environments

Cloud-based environments, such as trusted research environments (TREs), secure data environments (SDEs) and data commons, combine and harmonize data and provide computing infrastructure and software tools. With restrictions applied to data export, these environments create a governed resource for sharing and managing data that allows analyses to be brought to the data, a concept known as ‘data visiting’. An increasing number of programmes, including Genomics England, UK Biobank and the *All of Us* Research Program, now require use of a cloud-based environment to access their data. Combined, these three programmes alone currently have data from close to one million individuals. This is set to double by the end of 2024, with data from the recently launched ‘Our Future Health’ programme in the UK which aims to enable research, discovery and testing of more effective approaches to the prevention, earlier detection, and treatment of common chronic diseases. To achieve that, Our Future Health will recruit a diverse cohort of 5 million adult volunteers by 2027 with 1.7 million people having already joined the programme. Cloud-based environments such as TREs and SDEs are an important element of the data sharing landscape — they bring new benefits, including democratizing data access by lowering the technical barriers to

working with large-scale data; however, they also introduce new challenges, since cross-analysis must respect the policies and technology of multiple environments. Expansion of the cloud-based environment concept into access arrangements with health systems data will add another level of complexity.

Data commons are a related type of cloud-based data and analysis environment. Multiple data commons can interoperate under hybrid governance models, creating larger data meshes or ecosystems. Examples of established data commons include the NHGRI Genomic Data Science Analysis, Visualization and Informatics Lab-space (AnVIL)[44] and ELIXIR[45]. Infrastructures such as these make petabytes of curated, harmonized data available to a broad range of researchers who would otherwise lack the resources and expertise to analyse data at this scale. For example, the NCI Genomic Data Commons contains over 2.9 petabytes of curated, harmonized cancer data from more than 80 projects, is accessed by over 50,000 unique researchers a month, and has resulted in over 100 high impact publications[46]. In Europe, ELIXIR coordinates the Genomic Data Infrastructure (GDI) to support the Beyond 1 Million Genomes (B1MG) Project and enable transnational federated access to controlled-access genomic data that has been consented for secondary use across participating countries[47].

[H1] Accelerating discovery through cross-cohort analysis

Parallel analysis of data from multiple large cohorts held in separate environments is a powerful way to accelerate discovery without incurring the considerable costs of data copying or risks of data breach associated with moving large datasets. However, effective cross-cohort analysis requires the development and adoption of multiple policies, standards and tools, to generate reproducible results while preserving privacy and trust[48].

The feasibility of such approaches has recently been demonstrated by *All of Us* and UK Biobank programmes through a pilot genome-wide association study (GWAS) on circulating lipid levels, which conducted of a cross-cohort meta-analysis and compared it to pooled analysis[49]. The results of both analytical approaches correlated strongly with each other and with results obtained in external validation sets. However, the study

revealed important differences in the number and ancestry frequency distributions of variants significantly associated with the phenotype of interest. In addition, cost and complexity increased with the number of datasets used in the meta-analysis approach.

In rare disease, cross-cohort analysis is poised to accelerate the discovery of novel gene–disease associations. Unlike the predominant, two-sided model currently enabled by platforms such as the Matchmaker Exchange, which relies on the same candidate gene being independently identified by multiple researchers, a one-sided model enables the direct search for additional cases across multiple large datasets. The power of this approach was recently demonstrated by the discovery of variants in the gene *RNU4-2*, which encodes the spliceosome component U4 small nuclear RNA, as a major cause of intellectual disability[50]. Originally identified as a candidate in the Genomics England research dataset, over a hundred affected individuals were rapidly found by extending the search to US, Australian and European rare disease datasets, with absence of variation in *All of Us* and the UK Biobank serving as supporting evidence of disease association.

[H1] Current barriers and enablers to scaling up

Although the benefits of genomic data sharing are well recognized, systematically scaling up efforts beyond projects demonstrating feasibility to routinely cross-analysing research and clinical datasets is subject to a complex and evolving interplay between multiple factors.

[H2] Consent, engagement and trust

Healthcare data and genomic data are sensitive, and data use and sharing must occur in accordance with patient, participant and community expectations. Careful consideration is rightly given to the risk of privacy breach and re-identification of individuals[51]. Hence, ongoing dialogue with patients and participants and the broader public remain central to data sharing efforts. Directly involving participant communities in decisions about how their data are used and by whom is a further demonstration of commitment to their interests[52,53]. Importantly, participant consent and expectations will differ according to the setting and cohort, especially where data are shared for research

use as part of clinical care, as is the case through the partnership between the NHS Genomic Medicine Service and Genomics England. Some consent types specify that data will remain within national boundaries or be held only within a specified TRE. Consent may also specify the type of analysis that will be performed and by whom (for example, healthcare-related research) and what information will be returned to the participant or patient (for example, additional health-related findings) or to laboratories for accredited validation before clinical use. The GA4GH Data Use Ontology (DUO) provides a machine-readable representation of these different data use terms[54]. The standard can be integrated prospectively into consent processes and used to align existing data[55], facilitating data sharing at scale, including through increased automation of the oversight process[56].

[H2] Technical tools and standards

From a technical perspective, the increasing maturity of cloud infrastructure, and the benefits this brings in terms of capability and cybersecurity, has increased the feasibility of more complex, large-scale analyses. However, many other technical barriers remain. Genomic, phenotypic and associated metadata, both legacy and prospective, can exist in a variety of incompatible formats, effectively precluding meaningful data sharing. GA4GH has developed a suite of standards and tools to facilitate these processes, from capturing participant consent[54], to representation of clinical[57] and genomic data[58], through to standardization of data queries and analytical approaches[48]. Widespread and consistent implementation of these standards and tools is critical for data harmonization and sharing efforts. However, data standards alone are insufficient; for example, many use cases such as rare variant burden analysis, where subtle differences in variant calling matter, require the generation of genomic data through the same analytical tool chain. Moreover, interoperability issues can arise due to differences in how standards are implemented. For example, the Observational Medical Outcomes Partnership (OMOP) common data model[59], which is widely used to structure health data, comprises data schemas and data vocabularies. Implementing the same data schemas but populating them with different vocabularies may still render data incompatible.

[H3] Research and healthcare policy

Lastly, from a policy perspective, institutional and national interests and the data policies of governments and health systems often inhibit genomic data sharing. The reasons can vary, ranging from protection of potential intellectual property to concerns regarding national security. Establishing the infrastructure and capabilities required to support genomic data sharing has considerable financial and opportunity costs. Hence, mandating data sharing at the policy level[60] or having explicit agreements through other non-mandated routes, such as those proposed by the WHO[31], provides global incentives to share data and invest in specialized data management infrastructure. These mandates have profound effects on maximizing the return on public investment in research and in clinical sequencing and delivering on the expectations of research participants and patients. The NHS Genomic Medicine Service in England is an example of how health system genomics policy and strategy has committed to sharing clinical data for research, enabling consistent, consented genomic and other data from routine clinical care to flow into the National Genomics Research Library managed by Genomics England. This approach has made data from participant groups with rare conditions, common disease (for example, COVID infection) and cancer available for research with a feedback loop to the healthcare system, improving outcomes.

The B1MG Project has developed a framework for assessing the maturity of genomic medicine practices using 49 indicators across eight domains (Maturity Level Model, MLM). Many of the indicators directly related to facilitating genomic data sharing nationally and internationally with appropriate governance and consent[61]. Frameworks such as these are important tools for self-assessment, benchmarking and service planning at the policy level, including to guide the design and delivery of supporting infrastructure such as the European GDI[61].

[H4] Equity, diversity and inclusion

Advancing genomic data sharing needs to be part of wider efforts to promote diversity and inclusion so that we achieve equitable outcomes from the implementation of precision medicine, rather than risk exacerbating inequities[62-65]. The clear moral, scientific and practical imperatives to diversify genomic datasets have shaped the

fundamental priorities of efforts such as the *All of Us* Research Program in the USA[12,66], the Our Future Health study and the Genomics England Diverse Data Initiative in the UK, and the OurDNA study in Australia. While making these datasets globally accessible to researchers and clinicians serves to democratise access, it is not a substitute for building local infrastructure and capacity or for directly engaging with global communities. Initiatives such as Human Heredity and Health in Africa (H3Africa) have played a pivotal role in this respect[67,68], and an increasing number of large-scale sequencing projects are underway in Asia, the Middle East, Africa and South America.[69-73] It is important to recognise the issues of trust that arise in many communities, notably Indigenous communities. These are based on experiences of marginalisation, exploitation and discrimination and have significant implications for data sharing. Empowering Indigenous leadership, recognising genomic data sovereignty, and supporting capacity building are imperative[74,75] and models to achieve this are starting to emerge from Canada, Australia and New Zealand[13,14,76,77].

[H1] Conclusions

We can and must do more to enable genomic data sharing. Research participants choose to donate their data to benefit science. Given the choice, most individuals undergoing clinical genomic testing do the same[78]. It is our responsibility to honour this choice while protecting their data. All of us, as funders, policymakers, health system leaders and members of the genomic data community and global genomics ecosystem, have a responsibility to act (**Box 1**).

Genomic data and data sharing infrastructure are critical in driving discovery and generating the evidence to support the adoption of evidence-based innovations in precision medicine, thus optimising clinical impact. The global genomics community has made great strides over the last decade in the transition to a data-rich world. Genomic data generation is no longer the bottleneck, rather it is our ability to harmonise, share and derive full benefit from the data we have, both in the research and clinical settings. We now have the standards, tools and experience to move beyond projects demonstrating feasibility. With health systems increasingly mainstreaming genomics in clinical care, our attention needs to focus on scaling up efforts to enable safe data

sharing globally to unlock the value of today's and tomorrow's datasets for the benefit of future generations.

Author contributions

Z.S, D.G. and R.H.S. researched the literature, discussed the content and drafted the article. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Related links

Our Future Health https://ourfuturehealth.org.uk/

B1MG Maturity Level Model: https://www.b1mg-project.eu/resources/maturity-level-model

Genomics England Diverse Data Initiative: https://www.genomicsengland.co.uk/initiatives/diverse-data

OurDNA study: https://populationgenomics.org.au/project/australian-genetic-diversity-registry-agdr/

SING Canada: https://indigenousgenomics.com.au/ https://sing-canada.ca/

Australian Alliance for Indigenous Genomics: https://indigenousgenomics.com.au

Figure 1. Data sharing models. These include a). independent databases that allow data download, with data aggregation and analysis undertaken by individual researchers; b). cloud-based environments that allow data visiting, enable analysis and download of results; and c). federated approaches that allow multiple data sources to be accessed through different methods to enable cross-cohort analysis.

Box 1. A call to action: ten steps to systematically scale up genomic data sharing

If you're a research participant:

- Engage with those governing your cohort and researchers to ensure your expectations about data sharing are met.

- Share your ‘why’ story about what motivates you to participate in research, how your participation has affected you, and how you hope your contribution will help others.

If you’re a researcher working with genomic data:

- Speak up if access and cross-analysis of data from additional cohorts can support your research. Provide concrete examples to funders, policymakers and health system leaders to help them understand how you could further their mission if they lowered barriers to access.
- Expand your thinking to include the potential of cross-analysis. Rather than limiting yourself to hypotheses that can be explored with a single dataset, ask what you could discover by exploring multiple datasets – and then advocate to make it possible.

If you lead a cohort programme:

- Incorporate the value of data sharing into your discussions with funders and ethics boards, and your communication with participants. Design your consents and data use agreements using standards that support data sharing[54]. Highlight discoveries that were only possible because you and others chose to share.
- Grow your data sharing community. If your programme’s data are currently siloed at the hospital level, move toward sharing within your region. If they are siloed at the regional level, move towards sharing within your country. If they are siloed at the national level, move towards global sharing.
- Start a pilot. Choose a small example of sharing data between two or more programmes, and work with your funders and policymakers to test it out. Use the pilot to communicate the opportunity for expansion, and the tool and policy changes needed.
- Share best practices. Document tools and practices to encourage sharing, and help others adopt them. Adopt standards, tools and best practices shared by others. Join a GA4GH workstream or community of interest (for example, cancer, rare disease) and contribute to expanding global data sharing (https://www.ga4gh.org/get-

involved/). Review implementations of programs in your area to align on standards and interoperability.

If you fund cohort projects or influence research policy:

- Promote common infrastructure and policies across your portfolio. Invest in computational and data infrastructure that supports national and international analysis and re-use of data.
- Create incentives for cohorts that enable data sharing, and for cross-cohort and cross-country analysis projects. Make data storage costs free for cohorts that are shared. Lower barriers to access by simplifying processes.
- Encourage funding policies that create value for local and global communities. Promote diversity and inclusion through participant, citizen and community engagement and build bridges between research and healthcare.

If you lead and/or influence healthcare system policy or deliver clinical and diagnostic genomics:

- Advocate for the benefits of data sharing to be extended to clinical genomic testing as standard, to improve the speed and accuracy of diagnosis and to enable precision treatments. Fund and develop appropriate clinical consent and data management infrastructure, and support integration with the healthcare system.

References

1. Stark, Z.*, et al.* Integrating Genomics into Healthcare: A Global Responsibility. *Am J Hum Genet* **104**, 13-20 (2019).
2. Chen, S.*, et al.* A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* (2023).
3. Baxter, S.M.*, et al.* Centers for Mendelian Genomics: A decade of facilitating gene discovery. *Genet Med* **24**, 784-797 (2022).
4. Wright, C.F.*, et al.* Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. *N Engl J Med* (2023).
5. Boycott, K.M.*, et al.* Care4Rare Canada: Outcomes from a decade of network science for rare disease gene discovery. *Am J Hum Genet* **109**, 1947-1959 (2022).
6. Tambuyzer, E.*, et al.* Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nat Rev Drug Discov* **19**, 93-111 (2020).

7. Frangoul, H., *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and beta-Thalassemia. *N Engl J Med* **384**, 252-260 (2021).
8. Sosinsky, A., *et al.* Insights for precision oncology from the integration of genomic and clinical data of 13,880 tumors from the 100,000 Genomes Cancer Programme. *Nat Med* **30**, 279-289 (2024).
9. Stark, Z., *et al.* Australian Genomics: Outcomes of a 5-year national program to accelerate the integration of genomics in healthcare. *Am J Hum Genet* **110**, 419-426 (2023).
10. Lunke, S., *et al.* Integrated multi-omics for rapid rare disease diagnosis on a national scale. *Nat Med* **29**, 1681-1691 (2023).
11. Investigators, G.P.P., *et al.* 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med* **385**, 1868-1880 (2021).
12. All of Us Research Program Genomics, I. Genomic data in the All of Us Research Program. *Nature* **627**, 340-346 (2024).
13. Reis, A.L.M., *et al.* The landscape of genomic structural variation in Indigenous Australians. *Nature* **624**, 602-610 (2023).
14. Silcocks, M., *et al.* Indigenous Australian genomes show deep structure and rich novel variation. *Nature* **624**, 593-601 (2023).
15. Buchanan, A.H., *et al.* Clinical outcomes of a genomic screening program for actionable genetic conditions. *Genet Med* **22**, 1874-1882 (2020).
16. Guzauskas, G.F., *et al.* Population Genomic Screening for Three Common Hereditary Conditions : A Cost-Effectiveness Analysis. *Ann Intern Med* **176**, 585-595 (2023).
17. Stark, Z. & Scott, R.H. Genomic newborn screening for rare diseases. *Nat Rev Genet* **24**, 755-766 (2023).
18. Bernier, A., Molnar-Gabor, F. & Knoppers, B.M. The international data governance landscape. *J Law Biosci* **9**, lsac005 (2022).
19. Bernier, A., *et al.* Reconciling the biomedical data commons and the GDPR: three lessons from the EUCAN ELSI collaboratory. *Eur J Hum Genet* **32**, 69-76 (2024).
20. Clinical Cancer Genome Task Team of the Global Alliance for, G., *et al.* Sharing Clinical and Genomic Data on Cancer - The Need for Global Solutions. *N Engl J Med* **376**, 2006-2009 (2017).
21. Knoppers, B. Broaden human-rights focus for health data under GDPR. *Nature* **558**, 189 (2018).
22. Knoppers, B.M. Framework for responsible sharing of genomic and health-related data. *Hugo J* **8**, 3 (2014).
23. Knoppers, B.M., Chase, S., Joly, Y., Zawati, M. & Thorogood, A. Responsible Processing and Sharing of Genomic Data: Bringing Health Technologies Industries to the Table. *Am J Bioeth* **23**, 33-35 (2023).
24. Knoppers, B.M., Harris, J.R., Budin-Ljosne, I. & Dove, E.S. A human rights approach to an international code of conduct for genomic and clinical data sharing. *Hum Genet* **133**, 895-903 (2014).
25. O'Doherty, K.C., *et al.* Toward better governance of human genomic data. *Nat Genet* **53**, 2-8 (2021).

26. Phillips, M., *et al.* Genomics: data sharing needs an international code of conduct. *Nature* **578**, 31-33 (2020).
27. Rahimzadeh, V., Knoppers, B.M. & Bartlett, G. Ethical, Legal, and Social Issues (ELSI) of Responsible Data Sharing Involving Children in Genomics: A Systematic Literature Review of Reasons. *AJOB Empir Bioeth* **11**, 233-245 (2020).
28. Siu, L.L., *et al.* Facilitating a culture of responsible and effective sharing of cancer genome data. *Nat Med* **22**, 464-471 (2016).
29. Strom, B.L., Buyse, M.E., Hughes, J. & Knoppers, B.M. Data Sharing - Is the Juice Worth the Squeeze? *N Engl J Med* **375**, 1608-1609 (2016).
30. Suver, C., Thorogood, A., Doerr, M., Wilbanks, J. & Knoppers, B. Bringing Code to Data: Do Not Forget Governance. *J Med Internet Res* **22**, e18087 (2020).
31. Organization, W.H. WHO principles for human genome data access, use and sharing. Draft. (2024).
32. Harrison, S.M., *et al.* Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. *Genet Med* **19**, 1096-1104 (2017).
33. Rehm, H.L., Harrison, S.M. & Martin, C.L. ClinVar Is a Critical Resource to Advance Variant Interpretation. *Oncologist* **22**, 1562 (2017).
34. Cline, M.S., *et al.* BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet* **14**, e1007752 (2018).
35. Tudini, E., *et al.* Shariant platform: Enabling evidence sharing across Australian clinical genetic-testing laboratories to support variant interpretation. *Am J Hum Genet* **109**, 1960-1973 (2022).
36. Rivera-Munoz, E.A., *et al.* ClinGen Variant Curation Expert Panel experiences and standardized processes for disease and gene-level specification of the ACMG/AMP guidelines for sequence variant interpretation. *Hum Mutat* **39**, 1614-1622 (2018).
37. Rehm, H.L., *et al.* ClinGen--the Clinical Genome Resource. *N Engl J Med* **372**, 2235-2242 (2015).
38. Stark, Z., *et al.* Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. *Am J Hum Genet* **108**, 1551-1557 (2021).
39. DiStefano, M.T., *et al.* The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. *Genet Med* **24**, 1732-1742 (2022).
40. Boycott, K.M., Azzariti, D.R., Hamosh, A. & Rehm, H.L. Seven years since the launch of the Matchmaker Exchange: The evolution of genomic matchmaking. *Hum Mutat* **43**, 659-667 (2022).
41. Fiume, M., *et al.* Federated discovery and sharing of genomic data using Beacons. *Nat Biotechnol* **37**, 220-224 (2019).
42. Rambla, J., *et al.* Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. *Hum Mutat* **43**, 791-799 (2022).
43. Rodrigues, E.D.S., *et al.* Variant-level matching for diagnosis and discovery: Challenges and opportunities. *Hum Mutat* **43**, 782-790 (2022).

44. Schatz, M.C., *et al.* Inverting the model of genomics data sharing with the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space. *Cell Genom* **2**(2022).
45. Harrow, J., *et al.* ELIXIR: providing a sustainable infrastructure for life science data at European scale. *Bioinformatics* **37**, 2506-2511 (2021).
46. Heath, A.P., *et al.* The NCI Genomic Data Commons. *Nat Genet* **53**, 257-262 (2021).
47. Saunders, G., *et al.* Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat Rev Genet* **20**, 693-701 (2019).
48. Rehm, H.L., *et al.* GA4GH: International policies and standards for data sharing across genomic research and healthcare. *Cell Genom* **1**(2021).
49. Deflaux, N., *et al.* Demonstrating paths for unlocking the value of cloud genomics through cross cohort analysis. *Nat Commun* **14**, 5419 (2023).
50. Chen, Y., *et al.* De novo variants in the RNU4-2 snRNA cause a frequent neurodevelopmental syndrome. *Nature* (2024).
51. Gymrek, M., McGuire, A.L., Golan, D., Halperin, E. & Erlich, Y. Identifying personal genomes by surname inference. *Science* **339**, 321-324 (2013).
52. Hammack-Aviran, C., *et al.* Integrating participants as partners in research governance and operations: an approach from the All of Us Research Program Engagement Core. *BMJ Open* **13**, e068100 (2023).
53. Hastings Ward, J., *et al.* Research participants: critical friends, agents for change. *Eur J Hum Genet* **30**, 1309-1313 (2022).
54. Lawson, J., *et al.* The Data Use Ontology to streamline responsible access to human biomedical datasets. *Cell Genom* **1**, None (2021).
55. Lawson, J., Ghanaim, E.M., Baek, J., Lee, H. & Rehm, H.L. Aligning NIH's existing data use restrictions to the GA4GH DUO standard. *Cell Genom* **3**, 100381 (2023).
56. Cabili, M.N., *et al.* Empirical validation of an automated approach to data use oversight. *Cell Genom* **1**, 100031 (2021).
57. Jacobsen, J.O.B., *et al.* The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat Biotechnol* **40**, 817-820 (2022).
58. Wagner, A.H., *et al.* The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification. *Cell Genom* **1**(2021).
59. Hripcsak, G., *et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* **216**, 574-578 (2015).
60. Ross, J.S., Waldstreicher, J. & Krumholz, H.M. Data Sharing - A New Era for Research Funded by the U.S. Government. *N Engl J Med* **389**, 2408-2410 (2023).
61. Schmitt, T., Poirel, H.A., Cauet, E., Delnord, M. & Van den Bulcke, M. Unlocking the genomic landscape: Results of the Beyond 1 Million Genomes (B1MG) pilot in Belgium towards Genomic Data Infrastructure (GDI). *Health Policy* **143**, 105060 (2024).
62. Madden, E.B., *et al.* Advancing genomics to improve health equity. *Nat Genet* **56**, 752-757 (2024).

63. Fatumo, S., *et al.* A roadmap to increase diversity in genomic studies. *Nat Med* **28**, 243-250 (2022).
64. Baynam, G., *et al.* Advancing diagnosis and research for rare genetic diseases in Indigenous peoples. *Nat Genet* **56**, 189-193 (2024).
65. Skantharajah, N., *et al.* Equity, diversity, and inclusion at the Global Alliance for Genomics and Health. *Cell Genom* **3**, 100386 (2023).
66. All of Us Research Program, I., *et al.* The "All of Us" Research Program. *N Engl J Med* **381**, 668-676 (2019).
67. Consortium, H.A., *et al.* Research capacity. Enabling the genomic revolution in Africa. *Science* **344**, 1346-1348 (2014).
68. Choudhury, A., *et al.* High-depth African genomes inform human migration and health. *Nature* **586**, 741-748 (2020).
69. Fatumo, S., *et al.* Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat Genet* **54**, 531-536 (2022).
70. Wong, E., *et al.* The Singapore National Precision Medicine Strategy. *Nat Genet* **55**, 178-186 (2023).
71. Mbarek, H., *et al.* Qatar genome: Insights on genomics from the Middle East. *Hum Mutat* **43**, 499-510 (2022).
72. Gurdasani, D., *et al.* Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002 e1036 (2019).
73. Ziyatdinov, A., *et al.* Genotyping, sequencing and analysis of 140,000 adults from Mexico City. *Nature* **622**, 784-793 (2023).
74. Fox, K. The Illusion of Inclusion - The "All of Us" Research Program and Indigenous Peoples' DNA. *N Engl J Med* **383**, 411-413 (2020).
75. Hudson, M., *et al.* Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nat Rev Genet* **21**, 377-384 (2020).
76. Te Aika, B., *et al.* Aotearoa genomic data repository: An ahuru mowai for taonga species sequencing data. *Mol Ecol Resour* (2023).
77. Robertson, S.P., *et al.* Genomic medicine must reduce, not compound, health inequities: the case for hauora-enhancing genomic resources for New Zealand. *N Z Med J* **131**, 81-89 (2018).
78. Martyn, M., *et al.* Secondary use of genomic data: patients' decisions at point of testing and perspectives to inform international data sharing. *Eur J Hum Genet* (2024).