https://doi.org/10.1038/s41588-024-01903-7

Defining and pursuing diversity in human genetic studies

Maili C. Raven-Adams, Tina Hernandez-Boussard, Yann Joly, Bartha Maria Knoppers, Subhashini Chandrasekharan, Adrian Thorogood, Judit Kumuthini, Calvin Wai Loon Ho, Ariana Gonzlez, Sarah C. Nelson, Yvonne Bombard, Donrich Thaldar, Hanshi Liu, Alessia Costa, Vijaytha Muralidharan, Sasha Henriques, Jamal Nasir, Aimé Lumaka, Beatrice Kaiser, Saumya Shekhar Jamuar & Anna C. F. Lewis

Calls for more diverse data in genetics studies typically fall short of offering further guidance. Here we summarize a policy framework from the Global Alliance for Genomics and Health designed to fill this gap. The framework prompts researchers to consider both what types of diversity are needed and why, and how aims can be achieved through choices made throughout the data life cycle.

Calls for more diverse data in genetics have come from the scientific literature, as well as professional societies, funders, publishers and genomics initiatives. These calls stem from the recognition that a lack of diversity in data poses two major disadvantages. First, it holds back scientific advances that could benefit everyone — for example, by reducing the power to identify causal variants[1]. Second, the lack of diverse data can contribute to health disparities — for example, by leading to different rates of return of variants of uncertain significance and differential performance of polygenic risk scores[2,3].

Yet such calls have often been unclear about what types of diversity matter, or how to act on this understanding throughout the research process[4]. In reaction to these issues, the Regulatory and Ethics Workstream of the Global Alliance for Genomics and Health (GA4GH) — a nonprofit organization that sets technical standards and frames policies to expand genomic data use to benefit human health — set out to develop a policy framework that offers actionable considerations for how genetic and genomic researchers should approach thinking about diversity in data. The framework was developed by an international working group and was iteratively refined on the basis of responses from a public comment period. The policy framework recommends that research teams systematically address four questions to address what types of diversity in data are important and plan how to achieve these types of diversity in data (Fig. 1). Here we summarize the policy framework; the policy document itself provides more details and two worked examples[5].

Defining what is meant by diverse data

A common framing to highlight the lack of diverse data in genomics research is in terms of continental ancestry categories, with a focus on improving representation of those with non-European ancestries[4]. However, one issue with this is that the term 'ancestry' is highly ambiguous, as it is used to refer to genetically-inferred categories as well as self-reported information and geography[6,7]. This ambiguity allows for conflation between these various concepts of difference. Even if the more specific term 'genetic ancestry' is consistently used, the use of continental categories obscures the heterogeneity present within, and the continuous nature of genetic variation between those categories. Their use also helps perpetuate racist ideologies by reinforcing the false idea that humans can be divided into a small number of biological types[8]. And finally, the focus on continental genetic ancestry categories risks obscuring other dimensions of difference, such as the role of geographies and environmental factors. For example, researchers can claim to have represented Latino individuals (considered as an ancestry category) when they have only recruited individuals living in the USA. Given that almost every health outcome of interest has an environmental component, diversity in the environments sampled will often be important.

Rather than framing diversity in terms of ancestry, the recommendation in the GA4GH policy framework is to think of diversity in data as a means to an end: diversity in data is important because the lack of diversity hinders us from achieving certain desired goals. The goals of research agendas differ, and so too will the types of diversity needed to achieve those goals. Beyond genetic diversity, diversity in terms of sex, gender, health status, sociocultural context, geography

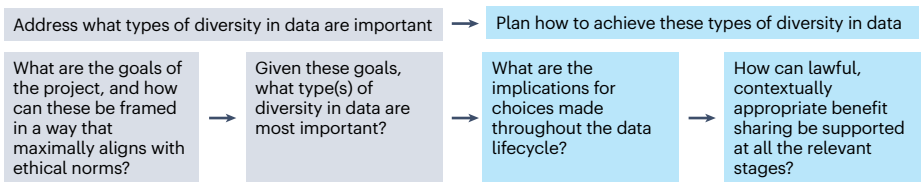

Fig. 1 | Questions genetics and genomics research teams should consider regarding diversity in data. In responding to calls for more diverse data, genetics and genomics research teams need to consider what types of diversity matter and why, and how they can achieve the benefits that those types of diversity are meant to bring through choices made throughout the data life cycle.

BOX 1

Additional considerations on defining research goals

This policy framework requires that researchers define research goals in a way that aligns with ethical norms, and then tailor the scope, conceptualization and expectations of diversity in genomic data accordingly. Such a process has many complexities.

An inherent tension with defining research goals aligned with universal ethical norms is that such norms lack context specificity, and may therefore fall short in protecting individuals and groups from harm. For example, indigenous scholars have argued that overvaluing individual consent can ignore risk to tribal participants[13].

The appropriate response to resolve this tension, and to ensure that the identification and specification of goals is responsive to social and historical context, is to utilize a truly interdisciplinary approach and to engage with the communities that are directly and indirectly impacted by the proposed research. This is a complex and resource-intensive process that must be planned and budgeted for accordingly. Other actors, including institutions and the broader research culture, have roles to play in incentivizing adoption of best practices.

More broadly, as organizations such as the International Science Council have advocated, for scientific advances to truly benefit humanity, a new engagement model is needed for how science and society interface, such that there is not only a one-way transfer of knowledge from science to policy and public (https://council.science/publications/flipping-the-science-model/). In this re-imagining, more democratic ways of deciding on the research agenda might be sought.

and many environmental factors may be important for a given project. This means that there is no single definition of diverse data.

Thinking of diversity in these terms — as a means to achieve certain ends — also clarifies that diversity is not necessarily about representativeness. The terms 'representation' and 'under-representation' are often used in combination with diversity, but it is seldom clear who a sample should be representative of (people living in a certain country?), and which dimensions of diversity are important for a dataset to match a real-world grouping (Age? Sex? Race? Ethnicity? Educational level?). Although everyone should have equal opportunities to participate in research, data that are representative in the statistical sense do not necessarily lead to better research outcomes or health equity. Indeed, sometimes individuals with certain attributes may need to be oversampled to achieve the desired goals of a project (although this might introduce its own issues, such as research fatigue). For example, some genomic discovery projects would maximize their power by oversampling individuals living in Africa. For many projects, aiming only for 'representative' data would not enable health equity goals.

Although what diversity means will vary by project, there is always a mandate for inclusive practices — it should be made as easy as possible for those people differentially situated to participate, should they wish to do so. We note that actions taken under the banner of inclusiveness must be genuine and not tokenistic, and that even genuinely inclusive practices may not by themselves achieve equitable outcomes.

If diversity in data is a means to achieve certain goals, there first needs to be clarity on what those goals are. Although researchers have intellectual freedom, freedom always comes with responsibility, and conducting research responsibly involves aligning intellectual enquiry with ethical norms[9]. The goals of any research project thus need to be framed in a way that maximally aligns with generally recognized ethical norms, including at a minimum the principles of beneficence, justice and fairness, and respect for individuals and communities. Research teams need to consider whether there are other norms that may be particularly salient for a given project — for example, any norms that matter most to the communities studied. Much of the work that centers calls for diversity is motivated by equity concerns, whether this means health equity or equitable benefits from research[10]. Equity will thus often — but not always — be the driving consideration; in these cases, diversity in data should be thought about as a means to further the end of equity.

Research teams can both call on experts from bioethics and other disciplines, and consult with the communities who will be most affected by their work to think through the norms that matter in their particular context, and to help make those norms action guiding (a process known as 'specification' in the bioethics literature)[11]. For example, rather than talking about 'achieving health equity' or 'furthering health equity', researchers should clarify exactly what they would take that to mean in their case. Clarity on the goals of the project has the added benefit of increasing the likelihood that the project actually achieves the hoped-for benefit. And conversely, failure to attend to how research could detract from ethical norms can cause harm and can contribute to mistrust in research. Box 1 presents additional considerations on defining research goals.

Clarity on the goals of the project should in turn help indicate what type(s) of diversity in data are the most important — both who should be represented in the data, and what data points are needed about them. Specificity is important here. For example, an identified need for genetic diversity does not in itself guide action. In some instances, even with clarity on the goals of the project, it may be unclear what type(s) of diversity are most important. Further work may help, including consulting the wider literature and discussing with peers from both within and outside of genetics and genomics. Indeed, by transparently acknowledging any lack of clarity, researchers can identify future needed work and approaches.

Pursuing these types of diversity in data

Having established clarity on what types of diversity in data are needed, researchers next need to consider how to achieve the desired benefits. Before considering the strategies that they can adopt to overcome any identified limitations in data available, research teams need to understand why their projects are limited by a lack of diverse data. By developing an understanding of these reasons — which will be context-specific — researchers will be in a stronger position to achieve the diversity in data their project needs to succeed. These reasons include barriers to inclusion, such as mistrust, past negative experiences of communities owing to abuse or misconduct in research, socioeconomic factors (such as transport cost and childcare coverage), language barriers, a lack of cultural understanding in study design, and a lack of diversity in those running the research. Other reasons include the legacy of historical and present-day inequities in healthcare access. For example, some groups may be systematically less likely to receive a

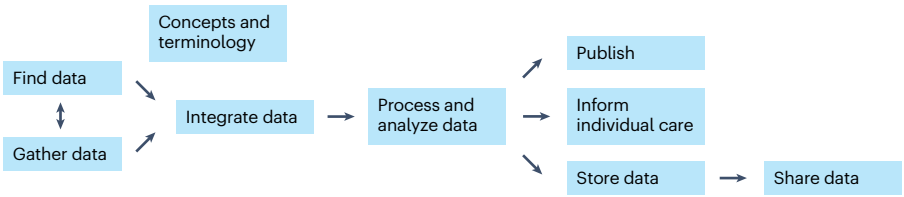

Fig. 2 | Research teams can adopt strategies throughout the data life cycle that impact the goals of utilizing diverse data. Such actions will be specific to each project, but some examples that are common to all projects include: careful use of concepts and terminology, and in particular avoiding conflations between different dimensions of difference (such as between genetically inferred groupings and self-identified labels); including all participants in analysis; carefully choosing where and how to host and share data; and proactively considering potential harmful uses of data. Strategies taken at one stage of the data life cycle will impact strategies that can be taken at other stages (indicated by the arrows). Data sharing includes plans for data governance.

diagnosis, which would have a direct impact if having that diagnosis was a criterion for inclusion. There may also be concern for risks of participating in research, such as the risk of discrimination in some insurance markets, which disproportionately affect some groups. In some cases, genomics researchers inherit a lack of trust in genomics research. Previous projects have been viewed as aimed at extracting information deemed valuable to others, with no attention to benefits to the communities sampled. One example is the Human Genome Diversity Project, deemed the 'vampire project' by the World Congress of Indigenous Peoples[12], and another is when DNA from members of the Havasupai Tribe were used without consent for research on schizophrenia and migration[13]. Finally, in the early 2000s and even 2010s, best practice in genome-wide association studies was viewed as focusing on genetically 'homogeneous' samples, resulting in a self-reinforcing cycle of tools and resources optimized for those genetically similar to individuals from European reference populations.

Who is recruited into research studies is key to considerations of diversity, and much has been written about the need to build trust with communities studied, to avoid 'helicopter science', and to identify and act on barriers for research participation. But beyond subject recruitment, researchers make decisions throughout the data life cycle that affect whether the ends that diversity is a means to achieve are met (Fig. 2). For example, decisions about exactly which data points about research participants are shared and how can be as impactful as the recruitment strategies adopted. Decisions about how the underlying continuous genetic similarity is analyzed can hugely impact the types of conclusions that are made about the relevance of genetic diversity in ways that have both scientific and ethical implications. Furthermore, decisions made about how to communicate and contextualize results impact how ideas about human difference are taken up not only in academia, but also in clinical practice, health policy and lay understanding.

Researchers can identify strategies at each stage of the data life cycle to help ensure that their attention to diversity helps to achieve the desired ends. As they do so, they should pay attention to how lawful, contextually appropriate benefit sharing can be supported at all relevant stages[14,15]. Some strategies that researchers should adopt are common to most projects, and the policy document[5] includes a table listing many such strategies. Research teams will also be able to identify and develop — ideally in consultation with the communities who will be most affected by their work — strategies that are suited to the unique nature of their project. Researchers may be able to draw on the principles and practices pioneered by indigenous communities, such as a focus on data stewardship and control[13].

Whereas individual research teams need to think critically about what diversity in data means for their project, there are different mandates for funders and others who shape the overall research agenda. We posit that the same considerations should shape thinking by these stakeholders — that is, the need to start with the overall goals of the research they are supporting and establish clarity on expectations of diversity in data. The entire genomics community needs to heed the rallying call to care about global equity, including contextually appropriate equitable benefit sharing and sustainable capacity building.

Having acknowledged the problems that a lack of diversity in data has generated, the fields of genetics and genomics have a window of opportunity to act to ensure that these problems are overcome to the extent that our collective resources allow. To capitalize on this opportunity, individual research teams have work to do: to identify what types of diversity matter (and why) and to then act on this understanding throughout the research process. We hope that they will find the GA4GH policy framework useful as a starting place for this important work.

Maili C. Raven-Adams[1], Tina Hernandez-Boussard[2], Yann Joly[3], Bartha Maria Knoppers[3], Subhashini Chandrasekharan[4], Adrian Thorogood[5], Judit Kumuthini[6], Calvin Wai Loon Ho[7,8,9], Ariana Gonzlez[10,11], Sarah C. Nelson[12], Yvonne Bombard[13,14], Donrich Thaldar[15], Hanshi Liu[3], Alessia Costa[16], Vijaytha Muralidharan[2], Sasha Henriques[16], Jamal Nasir[17], Aimé Lumaka[18,19], Beatrice Kaiser[2,20], Saumya Shekhar Jamuar[21,22] & Anna C. F. Lewis[23,24]

[1]Nuffield Council on Bioethics, London, UK. [2]Department of Medicine, Stanford University, Stanford, CA, USA. [3]Centre of Genomics and Policy, Victor Phillip Dahdaleh Institute of Genomic Medicine, McGill University, Montreal, Quebec, Canada. [4]National Institutes of Health, Bethesda, MD, USA. [5]The Terry Fox Research Institute, Vancouver, British Columbia, Canada. [6]African Biobanks and Longitudinal Epidemiologic Ecosystem, Ibadan, Nigeria. [7]Faculty of Law, Monash University, Melbourne, Victoria, Australia. [8]Centre for Medical Ethics and Law, University of Hong Kong, Hong Kong, China. [9]PHG Foundation, University of Cambridge, Cambridge, UK. [10]Genoox, Tel Aviv, Israel. [11]Bioethics Institute, Medical Science Department, Pontifical Catholic University (UCA), Buenos Aires, Argentina. [12]Department of Biostatistics, University of Washington, Seattle, WA, USA. [13]Institute of Health Policy, Management and Evaluation, University of Toronto, Ontario, Canada. [14]Genomics Health Services Research Program, St Michael's Hospital, Unity Health Toronto, Ontario, Canada. [15]School of Law, University of KwaZulu-Natal, Durban, South Africa. [16]Connecting Science, Wellcome Genome Campus, Hinxton, UK. [17]Life Sciences, University of Northampton, Northampton, UK.

[18]Centre for Human Genetics, University of Kinshasa, Kinshasa, Democratic Republic of the Congo. [19]African Rare Disease Initiative, https://www.ardi.africa/. [20]Global Alliance for Genomics and Health, https://www.ga4gh.org/. [21]SingHealth Duke–NUS Institute of Precision Medicine, Singapore, Singapore. [22]Department of Paediatrics, KK Women's and Children's Hospital, Singapore, Singapore. [23]Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA. [24]Harvard Medical School, Boston, MA, USA.
✉e-mail: aclewis@bwh.harvard.edu

Published online: 09 September 2024

References

1. Wojcik, G. L. et al. *Nature* **570**, 514–518 (2019).
2. Manrai, A. K. et al. *N. Engl. J. Med.* **375**, 655–665 (2016).
3. Martin, A. R. et al. *Nat. Genet.* **51**, 584–591 (2019).
4. Hardcastle, F. et al. *Camb. Prism Precis. Med.* **2**, e1 (2023).
5. Global Alliance for Genomics & Health. *Diversity in Datasets Policy v2.4;* https://go.nature.com/3Z87nui (2024).
6. Dauda, B. et al. *Front. Genet.* **14**, 1044555 (2023).
7. Mathieson, I. & Scally, A. *PLoS Genet.* **16**, e1008624 (2020).
8. Bliss, C. *Hastings Cent. Rep.* **50**, S15–S22 (2020).
9. Lewis, A. C. F. et al. *Perspect. Biol. Med.* **66**, 225–248 (2023).
10. Jooma, S., Hahn, M. J., Hindorff, L. A. & Bonham, V. L. *Ethn. Dis.* **29**, 173–178.
11. Beauchamp, T. L. & Childress, J. F. *Principles of Biomedical Ethics* 7th edn, Ch. 1, 17–19 (Oxford Univ. Press, 2012).
12. Greely, H. T. *Nat. Rev. Genet.* **2**, 222–227 (2001).
13. Garrison, N. A. et al. *Annu. Rev. Genomics Hum. Genet.* **20**, 495–517 (2019).
14. Bedeker, A. et al. *BMJ Glob. Health* **7**, e008096 (2022).
15. Thaldar, D. & Shozi, B. *J. Law Biosci.* **10**, lsad018 (2023).

Acknowledgements

The Diversity and Datasets taskforce acknowledges the contributions from other members of the Regulatory and Ethics Workstream, members of the public and other commenters who attended our many meetings. In particular, we acknowledge M. Afolabi, S. H. Chen, M. Doerr, J. S. Hsu, Z. Lombard, M. Mackintosh, A. Saadat and S. Singh. A.C.F.L. is supported by the NHGRI (1K99HG012809). S.S.J. is supported by National Medical Research Council, Singapore Clinician Scientist Award (NMRC/CSAINVJun21-0003)

Competing interests

A.C.F.L. owns stock in Fabric Genomics. T.H.-B. reports consulting fees from Grai-Matter and Paul Hartmann AG outside the submitted work. Y.B. owns stock in Genetics Adviser. S.S.J. is a co-founder of Global Gene Corporation. The other authors declare no competing interests.

Additional information

Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.