Cell Genomics

CellPress
OPEN ACCESS

Preview

I can drive in Iceland: Enabling international joint analyses

Chris Lunt[1,*] and Joshua C. Denny[1]
[1]*All of Us* Research Program, National Institutes of Health, Bethesda, MD, USA
*Correspondence: chris.lunt@nih.gov
https://doi.org/10.1016/j.xgen.2021.100034

In this issue of *Cell Genomics*, GA4GH reports key efforts to help share data across enclaves, including a framework for responsible data sharing, a data use ontology, and approaches for data use oversight. While there remains work in establishing reciprocity between data providers, we envision a future where joint analysis across enclaves is as easy as driving in different countries.

"How can I download the data?"

You cannot. For two reasons. First, in human biomedical research, we've come a long way from "big data is anything that won't fit in Excel." Genome-wide association studies (GWAS) of common diseases have exceeded one million people, and international cohorts will soon empower similarly sized studies with whole-genome sequencing data. When you've got hundreds of thousands of CRAM files representing petabytes of data, moving the data around gets expensive (and slow). The second reason is that we have an obligation to protect study participants, and that means that we must retain control of the data. So genomic research is moving to data enclaves, and now even if you have access to two great datasets, you may not be able to analyze them jointly. We believe the Global Alliance for Genomic Health (GA4GH) has an important role to play in bringing data back together so that we can realize the promise of the massive datasets being generated to advance genomic medicine. In this issue, GA4GH both presents its strategic framework and organization and reports on technology standards and developments that advance these data-sharing efforts.[1]

Data access tiers

All data providers face a conundrum: how do I maximize scientific utility while protecting the rights of the study participants? Scientific utility is the product of two things: open access and data precision. You can imagine that as an equation, $U = A \times D$. For example, to keep the same utility, if we increase access, we need to reduce data precision for those elements that enable potential re-identification. Utility can be thought of as a landscape, and data providers can choose multiple points on that landscape to address different audiences. This is how the *All of Us* Research Program chose our data access tiers.[2] We provide a "public" tier, which offers access to anyone but only gives data that is imprecise, i.e., summary statistics. We provide a "registered" tier, which limits access to an institution-approved audience but offers more precise data. Specifically, it provides access to individual participant-level data, but with changes to inhibit participant re-identification: data suppressions (for example, free-text entries and certain publicly available codes, such as homicides) and generalizations (for example, location represented as state and dates shifted backward). Within a year, we will introduce a more detailed "controlled" tier, with less row-level generalization but more access restrictions. In all cases, obvious personally identifiable information is removed.

Concerns about re-identification continue to grow, as data breaches of personal information have become so frequent that only truly gross breaches merit news coverage. All that data flowing into the dark web makes re-identification of individuals easier. Self-disclosure on social media adds potentially identifying data that otherwise would have been hidden. In addition, to engender participant trust, researchers may also institute other protections, such as restricting secondary research use cases or limiting the national sovereignty of the research audience. It takes more than just data suppressions; auditability and data exfiltration controls are also needed. Beyond the moral obligation, our ability to engage participants from groups that are underrepresented in biomedical research requires building trust. Willing participation depends on the belief that their data will not be used to harm them.

To protect against misuse of data, it is tempting to provide access only with articulated purpose and under intense scrutiny. Institutional review boards (IRBs) and data access committees can serve this purpose. If the data cannot be trusted, then you must make sure the researcher can be trusted. There is a risk that we all create our own enclaves, with our own controls and vetting processes, inadvertently denying ourselves one of the greatest sources of discovery—the convergence of evidence. Patterns may not be visible until you have sufficient overlap of multiple data sources.

Analyses across enclaves

Discovery generally comes from one of five sources. First is new tools, like microscopes, quantum theory, or deep learning. Second is new data, like novel surveys or data from underrepresented groups. Third is new researchers, who bring fresh eyes and the perspective of other disciplines. Fourth is new questions, like whether there is a connection between Colorado Brown Stain and resistance to tooth decay, which led to the discovery of the preventative power of dental fluoridation. The fifth is new overlap. Bringing together data from different sources does more than just shrink error

bars, it creates opportunities for eurekas, like observing the overlap between chimney sweeps and testicular cancer, the source of an early discovery of the environmental causes of cancer. It also allows for emergent observations of unexpected correlations, like deep learning predicting age and sex from retinal scans[3] (something humans cannot do) or the relationship between typing rate and the early diagnosis of Parkinson's disease.[4]

So how do we analyze across enclaves? Some propose federated research[5] (where data stays in its own enclave), using privacy-preserving analysis infrastructures and algorithms. These have limitations, including introduced latency, which may preclude I/O-intensive research approaches. This approach is valuable, but it is early and unproven, and we cannot wait to advance health science. Another approach is to run research in the enclaves and then join the de-identified results outside of the enclave, but that also has limitations, as important signals may be lost in the de-identification process.

Ideally, we would be able to jointly analyze unhidden data across enclaves. Data curation is often best performed at a local level with raw data access, and certain analyses may not work as meta-analyses. To do that, we must find approaches that effectively create overlaps in the security boundaries of our existing enclaves. Cloud-based environments make this easier to do technically, but the challenges are many and include finding the union of the controls of both parties, resolving the regulations and liabilities of the differing legal domains, and more.

GA4GH standards

Josh and I learned to drive a car in our home states (Michigan and Kentucky). How is it that we're allowed to drive a car in Iceland? Standards and reciprocity. The cars are the same, the fuel is the same, the layout of the roads and street signs are close enough to make the translation without training. The governments have reviewed each other's standards and agreed to accept the driver's licenses of the other. Imagine if every potential driver had to negotiate with the country each time they wanted to drive. Because that's the state of the art for data access now. GA4GH is leading the way by setting international standards and frameworks for sharing of human biomedical data.[1]

One of the key technical standards, reported in this issue, is the GA4GH Passport.[6] The NIH's Researcher Auth Service (RAS), designed to create a common system for authorizing users, is built using the GA4GH Passport specifications. A researcher will be able to log in with eRA Commons, NIH, or Login.gov credentials. These same GA4GH Passport specifications have also been adopted in many international research programs and institutes. RAS also uses the GA4GH Data Repository Service standard, which will help repositories to communicate with each other. This is necessary, but not sufficient.

The biosafety level (BSL) standard allows researchers to share samples without both parties auditing each other: "Are you operating a BSL3 facility? Yes? Good, we can share samples." In a similar fashion, the GA4GH can define standards of data protection that are tied to the participant consent. On top of that, we build a network of reciprocity.

Creating data protection standards is an opportunity to advance the mission of the GA4GH. When the GA4GH wrote the Framework for Responsible Sharing of Genomic and Health-Related Data,[7] Dr. Bartha Maria Knoppers drew on the United Nations Universal Declaration of Human Rights, Article 27, which states that everyone has the right "to share in scientific advancement and its benefits." Data protection standards not only protect the interests of the study participants, they can also advance greater equity of access by reducing the barriers to access. A "data passport" model eases access across repositories.[6] We can also create standards for granting access to strongly de-identified data without requiring a hypothesis, thus allowing a researcher to develop a line of inquiry that may not be visible until one has seen the data. In this way, we can provide a space for young researchers to develop their skills using real data.

"Share wisely, share widely," is an aphorism our program uses, recognizing a fundamental tension that all data generators share. Programs will find different ways to balance scientific utility, open access, and data precision. The GA4GH has the opportunity to advance a set of standards that allows us to recognize equivalences, working with efforts like the GA4GH Data Use Ontology (DUO),[8] Data Use Oversight System (DUOS),[9] and the GA4GH Variation Representation Specification (VRS).[10] If we succeed at that, but rely on pairwise agreements between data providers, we will still exclude a broader audience of less wealthy countries. We must create a network of reciprocity and work with groups like the International Hundred-thousand Cohort Coalition (IHCC).

One day researchers will be able to pull data from around the world into an ephemeral workbook, based on a set of internationally recognized credentials, enabling discoveries that will benefit us all. See you on the highways of Iceland!

REFERENCES

1. Rehm, H.L., Page, A.J.H., Smith, L., Adams, J.B., Alterovitz, G., Babb, L.J., Barkley, M.P., Baudis, M., Beauvais, M.J.S., Beck, T., et al. (2021). GA4GH: international policies and standards for data sharing across genomic research and healthcare. Cell Genomics *1*, 100029-1–100029-33.
2. Denny, J.C., Rutter, J.L., Goldstein, D.B., Philippakis, A., Smoller, J.W., Jenkins, G., and Dishman, E.; All of Us Research Program Investigators (2019). The "All of Us" Research Program. N. Engl. J. Med. *381*, 668–676.
3. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., and Webster, D.R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. *2*, 158–164.
4. Adams, W.R. (2017). High-accuracy detection of early Parkinson's Disease using multiple characteristics of finger movement while typing. PLoS ONE *12*, e0188226.
5. Thorogood, A., Rehm, H.L., Goodhand, P., Page, A.J.H., Joly, Y., Baudis, M., Rambla, J., Navarro, A., Nyronen, T.H., Linden, M., et al. (2021). International federation of genomic medicine databases using GA4GH standards. Cell Genomics *1*, 100032-1–100032-5.
6. Voisin, C., Linden, M., Dyke, S.O.M., Bowers, S.R., Reinold, K., Lawson, J., Li, S., Ota Wang, V., Barkley, M.P., Bernick, D., et al. (2021). GA4GH Passport standard for digital identity and access permissions. Cell Genomics *1*, 100030-1–100030-12.
7. Knoppers, B.M. (2014). Framework for responsible sharing of genomic and health-related data. HUGO J. *8*, 3.

8. Lawson, J., Cabili, M.N., Kerry, G., Boughtwood, T., Thorogood, A., Alper, P., Bowers, S.R., Boyles, R.R., Brookes, A.J., Brush, M., et al. (2021). The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genomics *1*, 100028-1–100028-9.

9. Cabili, M.N., Lawson, J., Saltzman, A., Rushton, G., O'Rourke, P., Wilbanks, J., Rodriguez, L.L., Nyronen, T., Courtot, M., Donnelly, S., and Philippakis, A.A. (2021). Empirical validation of an automated approach to data use oversight. Cell Genomics *1*, 100031-1–100031-6.

10. Wagner, A.H., Babb, L., Alterovitz, G., Baudis, M., Brush, M., Cameron, D.L., Cline, M., Griffith, M., Griffith, O.L., Hunt, S.E., et al. (2021). The GA4GH Variation Representation Specification: A Computational Framework for variation representation and Federated Identification. Cell Genomics *1*, 100027-1–100027-11.