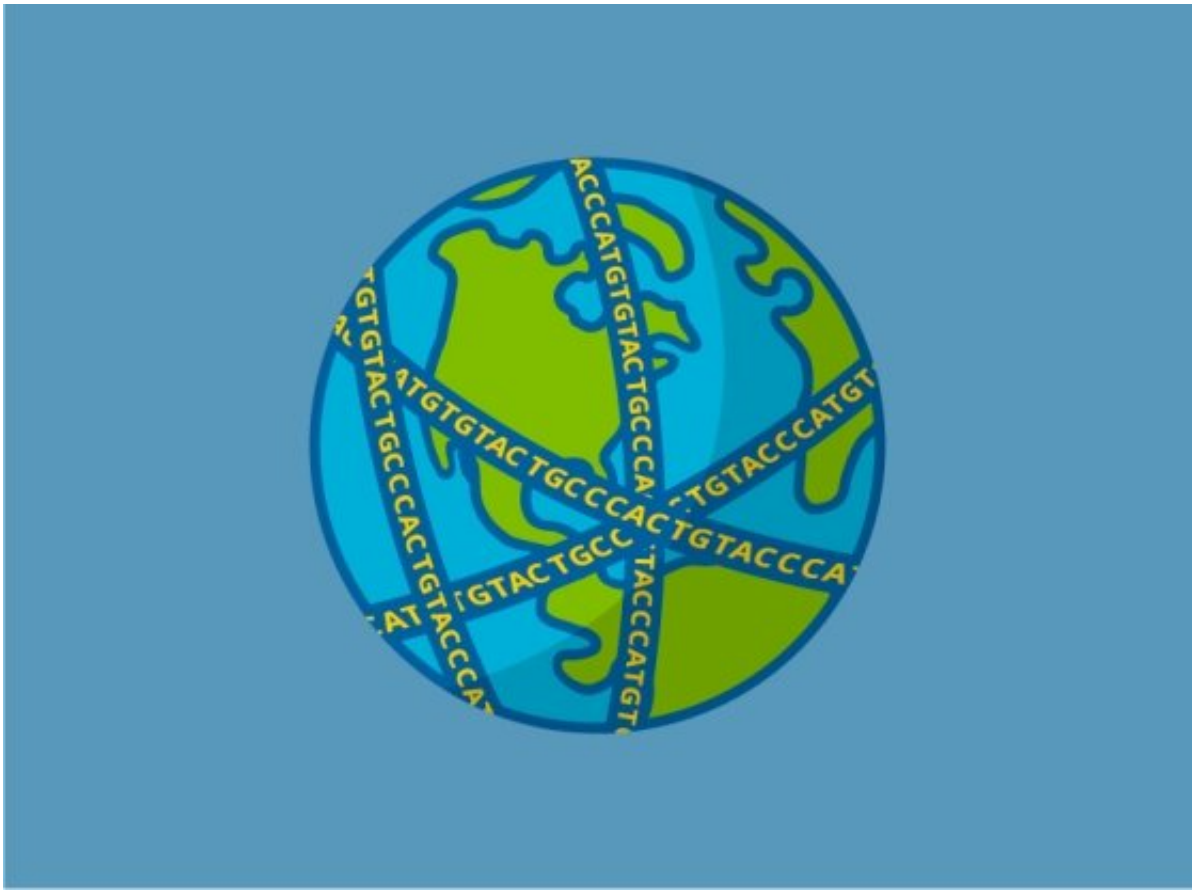
Navigating the jungle: GA4GH and a global infrastructure for seamless genomic data sharing

December 14 2021, by Tom Ulrich



Credit: Ricardo Job-Reese, Broad Communications

Over the last decade and a half, the biomedical field has witnessed an explosion in the volume and variety of genomic and health-related data.

On the one hand, this has been an incredible boon for advancing human health; as our knowledge about the genetics of health and disease grows, so do the opportunities for researchers to make remarkable gains in disease prevention, diagnosis, and treatment.

That explosion, however, has also created immense challenges. Datasets are spread across research centers, universities, health care systems, government agencies, and more, often stored in systems that cannot directly talk to each other or in formats that cannot be readily translated from one to the other. There is a lack of infrastructure that makes it difficult for researchers to share and analyze those data and turn them into knowledge that can benefit patients.

Launched in 2013 to tackle this challenge, the Global Alliance for Genomics and Health (GA4GH) is a community of 650 organizations and 1,000 individual members from more than 90 countries dedicated to creating standards, policies, and approaches that promote effective and responsible genomic and health-related data sharing. GA4GH is building common pipelines, languages, and rules so that researchers can share and analyze genomic and clinical data in a consistent way.

In a recent special issue of the journal *Cell Genomics,* GA4GH members published a collection of 10 papers describing the organization's goals, principles, and work to date, including recommendations for federating data access internationally and standards for data access and oversight. Together, these commentaries and technical papers provide both a history of GA4GH and a roadmap for how its work will grow and evolve in the coming years.

We sat down with four Broad members who play key roles in the GA4GH community—software engineer and GA4GH Variant Representation Standard (VRS) project co-lead Lawrence Babb; software product manager and GA4GH Data Use co-lead Jonathan

Lawson; Broad chief data officer, Eric and Wendy Schmidt Center co-director, and GA4GH steering committee member Anthony Philippakis; and institute member and GA4GH vice-chair Heidi Rehm—to talk about the organization's progress and future.

How has the data-sharing landscape in clinical genomics changed in the last decade?

JL: It became clear in the late 2000's that genomics research was becoming more common, and that data stored in large databases could potentially be linked back to patients. Those concerns raised the need for controlled data access. The benefit there is that this brought about a strong sense of data stewardship. The problem is that many institutions went ahead and did their own thing, which led to a proliferation of different processes for data access.

At a high level, what we now see is a desire to lay down some clear pathways so that researchers can navigate that jungle of processes and access data in a controlled manner. This is one area where GA4GH has really stepped up, by helping bring consistency to how data access is granted and established. The more interoperable things can be, the easier it's going to be for a researcher to get all the data they could want or need to answer a scientific question and ultimately generate innovative scientific results that positively impact human health.

HR: There is now global recognition that we cannot support clinical genomics without widespread sharing of both genomic and health data, as well as curated knowledge. The success of voluntary knowledge-sharing in the NCBI's ClinVar and widespread use of federated platforms like Matchmaker Exchange has shown us that clinical laboratories and the rare disease community have bought into the notion that we must work together to understand causal variation and apply

genomics to the care of patients with rare disease.

What do you think are some of the most significant challenges in bringing genomic data into the clinic?

LB: Getting to the point where we can reliably represent genetic findings and knowledge with the clinical precision necessary to inform patient care will require significant effort. This is starting to happen in various areas, but there is still not yet a common foundation on which everyone in the field can reliably build approaches to using this invaluable data. We need standards, tools, and resources that rise to the level the healthcare system requires.

JL: We still have to solve the prime issues of what data is out there, how can we share it, and how can we analyze it. Governments, foundations, and corporations have spent millions of dollars to generate genomic data, often for a singular research project. After the initial project those data are essentially put in the back shelf of a warehouse, and no one knows that they're even there or how they can be used. All of this data could be extremely useful, and we need to discover the right incentive structures that would encourage data owners to engage with this problem.

AP: We're at a pivotal moment, where genomics is starting to shift from a purely research activity to one that is increasingly driven by clinical care. But while there has been a real embrace of patient-level data-sharing in basic research, this is much less clear in the clinic. For example, the amount of patient-level cancer genomic data generated through oncology care dwarfs what has been generated in the research setting. Why do we not share the genomic and clinical data of every patient with cancer, so that we can learn from the outcomes of the treatments that we're giving? We need to extend the kind of data sharing that's now common in the research setting to the clinical setting.

We also need to build the evidence base to show that knowing an individual's genomic makeup improves outcomes. For example, should patients with a high polygenic risk score (PRS) for cancer undergo more aggressive screening? Should we start those with a high PRS for coronary artery disease on a statin earlier? In order for genomics to become standard of care, we need to do the studies that will address these questions.

HR: There is still a lot of concern about data privacy and security when trying to collaborate and share across international boundaries. We need to engage with and educate the genomics community about the risks and benefits of sharing genomic data, and of allowing individuals, instead of regulators, to drive decision-making, while at the same time applying the most advanced approaches and thoughtful policies to protect individuals' rights to privacy and the security of their data.

How has GA4GH changed since it launched? What are some of its biggest successes?

LB: I have been involved with GA4GH for five years or so. In that time, GA4GH has significantly honed its organization so that it can tackle key areas that are too big or complex for any one organization, segmenting their efforts into work streams and relying on partnerships with real-world driver projects to build open solutions. GA4GH leadership's "rough consensus running code" motto helps the driver projects produce and share solutions around common problems.

JL: The community-led model has been very successful, more than I had hoped it would be when I first joined GA4GH in 2017. It's been interesting to see what it really takes to manage a community-led standard. Ultimately there needs to be arbiters, but it's often obvious when there's consensus on what needs to be done and how to do it.

GA4GH is the bridge that gives us a reason and framework to solve data sharing and data access problems at a global scale. If you don't coordinate globally, you're destined to only solve these problems just within the US, or just within a limited network of institutions. I talk to colleagues in Singapore and South Africa and Australia and Brazil and the UK, people who I would likely never work with in my day job, but with whom I can concentrate on these problems together. That's huge.

AP: During the past decade, GA4GH has developed an excellent set of processes for seeing the creation of standards. This wonderful collection of papers in Cell Genomics demonstrates that progress.

HR: We have successfully convinced countries, funders, and organizations of the importance of developing common standards and frameworks for genomic and health data-sharing. We have gone from a distributed grassroots effort to a well-powered operation that has figured out how to work across the global community.

These *Cell Genomics* papers represent a huge amount of work. Can you describe some of the key take-home messages?

JL: I see a prominent and recurring theme of data sharing and data access. The perspective paper focuses on data sharing, and it's a primary objective in everything that we do in GA4GH. I think this is because data sharing is a force multiplier in genomic discovery. GA4GH participants can improve data storage and analysis in their home institutions, but building out federated and standardized systems assures everyone that data will be easier to share and access—making the results of analyses more powerful and more rapidly achieved.

AP: They are a clear indication of the power that standards can play in driving progress. Take, for example, automated approaches to data use oversight. If you ask any researcher who wants to put large genomic data

to use about their greatest pain point, almost everyone would say it's the process of gaining access to data. The Data Use Ontology and Data Use Oversight System could potentially streamline this process significantly by automating many of the steps that go into validating whether a researcher's purpose is consistent with the terms of a given dataset's informed consent.

HR: One common theme is long authorship lists with many people, institutions, and countries represented, demonstrating the importance of community consensus and widespread engagement to drive the work of GA4GH.

Another is the recognition that we need broad participation in order to ensure that our work is informed by key projects and perspectives, and ultimately adopted by the entire community. The participation of our driver projects is one way in which we accomplish this. Another is through our Genomics in Health Implementation Forum, which we launched in 2020 and which is bringing large-scale genomic data initiatives such as Australian Genomics and Genomics England together to share resources and knowledge, and to support implementation of our work.

What's next for GA4GH?

LB: We're working on "light touch" process refinement efforts that will ensure the quality, consistency, and dissemination of our work. And we're bringing in contributions from additional organizations interested in participating in this massive but essential effort.

We're also continuing to bring organizations together to create and support methods by which knowledge about genetic variants can be captured and disseminated in a standard and interoperable manner, and to share lab results and their interpretations with physicians and patients

in a meaningful way. That's a building block for incorporating genomics more directly into healthcare.

JL: On the data-use side, I think we're moving from defense to offense. We're being more proactive, and talking about creating new things, instead of just solving old problems. What innovations can we build that will unlock a lot of potential? A theme in that is federation: creating standards where if I'm a researcher, every time I connect with a new database or new institution I don't have to consistently repeat the same processes over and over again.

There's this dance in policy development and technology development; they have to move together in coordination. Through GA4GH, the Broad is in a position to help engage in a lot of policy advocacy with regulators in the US and Europe and beyond. On top of that, we're also able to build the software that either fits a newly minted policy, or which proves that a policy that we're proposing is actually actionable and gives us feedback for refining it.

HR: We're starting to focus more heavily on implementation of the standards that have been approved. We want to work with our driver projects and the wider community to stress test our standards and policies and prove that they are fit for purpose, as well as stitch together multiple standards to support end-to-end workflows. I think the Broad is well-poised to continue contributing to GA4GH's work, to implement GA4GH standards, and to help demonstrate what can be accomplished in our field.

More information: Heidi L. Rehm et al, GA4GH: International policies and standards for data sharing across genomic research and healthcare, *Cell Genomics* (2021). DOI: 10.1016/j.xgen.2021.100029

Adrian Thorogood et al, International federation of genomic medicine

databases using GA4GH standards, *Cell Genomics* (2021). DOI: 10.1016/j.xgen.2021.100032

Jonathan Lawson et al, The Data Use Ontology to streamline responsible access to human biomedical datasets, *Cell Genomics* (2021). DOI: 10.1016/j.xgen.2021.100028

Alex H. Wagner et al, The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification, *Cell Genomics* (2021). DOI: 10.1016/j.xgen.2021.100027

Moran N. Cabili et al, Empirical validation of an automated approach to data use oversight, *Cell Genomics* (2021). DOI: 10.1016/j.xgen.2021.100031

Provided by Broad Institute of MIT and Harvard