GA4GH 4th Plenary Report

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Summary

On October 16-18, 2016, stakeholders in genomics and health-related data sharing convened in Vancouver, BC, Canada, for the 4th Plenary Meeting of the Global Alliance for Genomics and Health (GA4GH). The Plenary focused on facilitating rapid uptake of GA4GH tools and solutions in real-world settings and translating the last two years of work into action. The Plenary also highlighted progress made on key GA4GH projects, provided updates on the organization, and invited input from the community on its future goals and next steps.

In the final count, the meeting attracted 355 people from 204 organizations and 26 countries — making it the most popular GA4GH meeting to date. About half of the attendees were new to GA4GH.

The three-day conference opened with working meetings of the four working groups — Data, Clinical, Security, and Regulatory and Ethics — including a joint meeting of the Clinical and Data Working Groups. Day two focused on four data-sharing demonstration projects — Matchmaker Exchange, BRCA Challenge, Beacon, and the new Cancer Gene Trust. The schedule was interspersed with breakout sessions on special topics, demonstrations, and task teams. The full group came together on day three, October 18, for a day of keynotes, talks, and discussions, including updates and highlights from the first two days. The Plenary Day ended with a reflection on the brilliant creative work by thousands of volunteers and a renewed call for GA4GH to develop the data sharing tools necessary to deliver on the promise of genomics to advance human health.

Funding for this meeting was made possible in part by 3 U54 HG 007990 - 02 S1. The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.

How to Cite

Global Alliance for Genomics and Health. Summary of the Fourth Plenary, 2016. Available at:
ga4gh.org

Sponsored by

AstraZeneca

MedImmune

EMBL-EBI

intel

VERTEX
THE SCIENCE *of* POSSIBILITY

Canada Health Infoway
Inforoute Santé du Canada

elixir

GARVAN INSTITUTE
OF MEDICAL RESEARCH

Genome BritishColumbia

GenomeCanada

inspire 2 l!ve

Melbourne Genomics
Health Alliance

Microsoft
Genomics

ONTARIO BRAIN INSTITUTE
INSTITUT ONTARIEN DU CERVEAU

SIB
Swiss Institute of Bioinformatics

Table of Contents

Opening Remarks | Peter Goodhand (Global Alliance for Genomics and Health)

Keynote: Towards a Platform for Global Health | Phil Bourne (National Institutes of Health)

Facilitating Access to Data | Chaired by Nicola Mulder (University of Capetown)

Human Genomics: ELIXIR and GA4GH Collaboration | Serena Scollen (ELIXIR)

Variant Databases: Tripartite Responsibilities | Bartha Knoppers (Centre of Genomics and Policy, McGill University)

Tooling Meta to Micro | Anthony Brookes (University of Leicester) and Anthony Philippakis (Broad Institute of MIT and Harvard)

From Discovery to the Clinic | Chaired by Kathryn North (Murdoch Childrens Research Institute, University of Melbourne)

BRCA Exchange: Answering the Challenge | Heidi Rehm (Partners Healthcare, Broad Institute, Brigham & Women’s Hospital, Harvard Medical School)

Matchmaker Exchange: The Next Generation | Kym Boycott (Children’s Hospital of Eastern Ontario)

Putting Patients on the Genomic Discovery Team | Alastair Kent (Genetic Alliance)

Keynote: Data Sharing, Public Health, and the Bioeconomy | Gerardo Jimenez-Sanchez (Harvard T.H. Chan School of Public Health)

Big Genomic Data and Industry Panel | David Glazer (Verily), Andrew Ury (ActX), Jonathan Hirsch (Syapse), Steve Lincoln (Invitae)

Data Sharing: Fast Forward Consent | Chaired by Nazneen Rahman (The Institute of Cancer Research, The Royal Marsden NHS Foundation Trust)

Open Source, Open Data | David Haussler (University of California, Santa Cruz)

A Systematic Approach to Data Sharing in Translational Medicine | Steven Jones (BC Cancer Agency)

Citizen Science, Social Engagement | Megan Doerr (Sage Bionetworks)

Closing: The GA4GH Tomorrow | Ewan Birney (EMBL-EBI, Genomics England)

Opening Remarks | Download Slides

In his opening remarks, GA4GH Executive Director **Peter Goodhand (Global Alliance for Genomics and Health)** welcomed 355 attendees from 204 organizations and 26 countries to the biggest GA4GH plenary meeting to date. Some people in the audience had attended the original small planning meeting to create GA4GH in New York City on January 28, 2013. More than 100 of the attendees of the Vancouver meeting were new to GA4GH.

Goodhand reviewed the unprecedented opportunity for genomics to transform health care worldwide — if the surge in genomic data can be shared in large enough numbers to make sense of it. From the beginning, the GA4GH was never about bigger, better, faster research, he said. Instead, it is about translating research into better health for humanity.

GA4GH set itself on a mission to help genomics and healthcare data emulate the interoperability of mobile telephones and the Internet. It formally launched in September 2014, with four working groups to tackle the major technical and societal challenge areas — security, data, clinical, and regulatory and ethics. Within those, task teams populated by more than 1,000 volunteers representing leading experts in their field have been delivering results, ranging from security and regulatory frameworks to a Genomics API. Four data sharing projects help showcase and drive the evolving tools. A new GA4GH catalogue listed nearly 100 global genomic initiatives. Representatives from more than one dozen of them traveled to Vancouver to compare approaches and align their interests.

A huge difficulty remains in lifting the knowledge from genomics research into health applications, such as precision medicine. GA4GH aims to serve a catalytic role to make the data as accessible as possible. Goodhand thanked the meeting sponsors, the passionate individual members who contribute time and energy as volunteers, and the organizational members who help in the adoption, dissemination, and support of GA4GH tools, policies, and projects. The larger community collaborates on interoperability and competes on implementation, he said.

Keynote: Towards a Platform for Global Health | Download Slides

Phil Bourne (National Institutes of Health) described several platforms to share data and bring researchers together in cost-effective and productive ways, including the NIH Commons. He pointed out lessons and touchpoints to help GA4GH move toward its next phase, complete with a clearly articulated vision and endpoint that clients and funders can endorse and support.

In the first example, Bourne discussed ASAPbio , a scientist-driven initiative to promote the productive use of preprints. A preprint is a complete manuscript deposited on a server prior to review. Preprint services, long a staple of physics research, have started to proliferate in the life sciences. Bourne described the trend as a response to a growing recognition that publishing delays are interfering with scientific progress. In traditional journals, an article may take close to 300 days between submission and publication. In a fast-moving field, the knowledge in those papers needs to be accessible to the community immediately, Bourne said. The preprint approach also allows for more informed grant review, especially from young investigators, and allows for the publication of negative data, another category of important knowledge.

Bourne predicted that a central life sciences preprint service will be developed soon by ASAPbio , thanks to about 15 global funders from governments and foundations, who will make a 5-10 year commitment. The project has buy-in from open access publishers, such as F1000, PeerJ, and PLoS, Bourne said.

Similarities between ASAPbio and GA4GH include a perceived critical mission, strong leadership, participation from leading scientists, and significant community support. ASAPbio has worked out a few other crucial components that Bourne advised for GA4GH, including an obvious endpoint, funder buy-in, champions identified within each funding body, and a vision to coalesce the larger community. For ASAPbio, the endpoint is accelerated scientific research through a human- and machine-accessible corpus of knowledge that is accessible to all.

Bourne noted that GA4GH has a possible endpoint contained within its mission statement: To establish a common framework. That common framework may serve as a natural touchpoint for NIH and other funding agencies to come together behind GA4GH as a collective and provide ongoing sustainable resources.

In an encouraging trend, NIH funding is moving from the siloed pipes of the current system to more unified platforms, Bourne said. This aligns with GA4GH thinking and may provide many opportunities. He cited Sangeet Paul Chowdry , who defines pipes as business models that enable the linear flow of value from producer to consumer, while platforms enable interactions between connected producers and consumers in an ecosystem. NIH previously promulgated pipes, Bourne noted, but the approach is not sustainable anymore, nor does it map to current science needs.

Holding up an example from the sharing economy, Bourne suggested biology research look to Airbnb, a platform that supports a trusted relationship between a consumer (paying guest) and a supplier (host). It seems to be working because everyone benefits—the hosts make more money and the guests have an authentic experience.

Likewise, biomedical research has suppliers (data providers), consumers (data consumers), and trusted platforms for exchanging services and conducting transactions where there is potential for everyone to benefit. The problem, Bourne said, is that none of these are connected in biomedical research. There is no interoperability between the layers in any meaningful way. GA4GH is trying to bring the two layers together. Bourne has crowdsourced a draft paper on this comparison, which is now under review by a journal. The analogy is not as strong as it could be, he said, in part because Airbnb was "born digital," not to mention the tremendous complexity of biomedical research.

Digital data plays a critical role in the realm of modern scholarship. Funders have become increasingly concerned about the sustainability and maintenance of the amount of data being generated, as well as how to use it effectively. For example, data is not yet valued as highly as published papers in the academic community, despite similar vetting and quality control processes. In one subtle but powerful message, it is expected the NIH will require data sources to be cited in the same way as publications and accrue citations based on their value to the community.

The bigger issue is the wasted effort and inefficiencies in the current "data-knowledge cycle," Bourne said. For example, researchers submit digital data in analogue form, and then funders spend money to have that information extracted and put back into a data resource for others to use. Bourne is writing a book about this with a working title, "Is There No Intelligent Life Down There?"

NIH is trying to move to platforms with greater integration across the data knowledge cycle. Bourne praised GA4GH for also thinking in terms of platforms. Importantly, there is no single platform for biomedical research. Many impediments stand in the way, including work practices, entrenched business models, and the sheer size of the endeavor. It can be done, Bourne believes, in an iterative and agile way, leveraging both trust in the system and incentives to use the platform.

In fact, NIH has just begun to put incentives in place to use its Commons platform, a shared virtual space where scientists can work with the digital objects of biomedical research, such as data, papers, software, metadata, and workflows. In this space, everything is treated as a digital object with an identifier, such as DOI or some other kind of handle. The platform operates under the FAIR principles (Findable, Accessible, Interoperable, and Reusable).

The Commons will begin by integrating two of the layers of scholarly workflow—data and analytics. Pilot projects now include the NCI Data Commons and public clouds, with reference data and individual data, as well as layers of APIs and software. Through the Big Data to Knowledge (BD2K) initiative, NIH is funding different projects that exercise all these layers. For example, BioCADDIE is an indexing tool, like a PubMed for data.

Another piece of the NIH Commons is the business model of cloud credits, issued to researchers to spend on major cloud computing service providers of their choice, as long as they are working in the virtually shared environment. Look for a general call to apply for credits soon, Bourne said.

The Commons has been a long time in the making, and it is early days in creating a platform for a new kind of research. One of the keys to reproducibility and sustainability is to train the next generation of data scientists working in genomics and health, who will become a new kind of digital native, Bourne said.

NIH has been in conversation about these issues with Wellcome Trust and others in Europe, and they are hoping to move this into the Pacific Rim by engaging funders in Japan and China as well. They have identified three large datasets to move into this space, TopMED , the model organism databases, and GTEx . Funders are closely cooperating to help create a seamless virtual environment across very different enterprises. Bourne advised GA4GH to consider working collectively with these funders to move some of the activities into these kind of platform environments.

In a question and answer session following the talk, Bourne acknowledged the current difficulty in the complex approvals to access data, such as dbGaP, and host it in a cloud computing environment for analysis. It's being actively looked at, he said.

Another person asked how, hidden in the vast sea of data, will people be able to find what they need. Bourne suggested a data indexing service with the equivalent of article-level metrics to help valuable data rise to the top—or something like a Yelp for data, where annotations from others in a trusted community could help guide choices. On the other hand, good and popular are not the same thing, he agreed, pointing out that the best software now available is not necessarily the most popular.

Trust loomed large in several questions. One person pointed out that data producers have professional duties toward patients, which might not be reflected in a commercial cloud environment. Bourne agreed, saying the current system works on a trust and reputation model. The idea of funders is to bring together the multiple trust and reputation layers of data producers, biological curators, software developers, and academic scientists. In response to another question, Bourne hoped that the platform will be able to benefit larger groups of participants and that problems will be addressed in a better way.

Summing up in response to a final question, Bourne said funders are working toward recognizing efforts of data providers, which are currently undervalued, and also rewarding people who annotate the data. Platforms, such as NIH Commons, could accelerate these activities.

Facilitating Access to Data

Nicola Mulder (University of Cape Town) introduced the session, noting that the underlying core principle of GA4GH is facilitating access to data and overcoming the associated challenges. The clinical, regulatory and ethics, security, and data working groups have gone a long way toward addressing accessibility, interoperability and reusability of data, she said.

Human Genomics: ELIXIR and GA4GH Collaboration | Download Slides

Serena Scollen (ELIXIR) opened her presentation with a vision of where the human genomics field is going. With whole genomic sequences relatively quick and inexpensive, it is important how we link that data, she said. The big question: How can we put together an infrastructure where we can manage, archive, share, and use that data effectively to put the scientific and medical community on an effective path for discovery?

All sorts of types of data are generated for an individual, such as genomic sequence, electronic health record, and lifestyle data from wearables. The scientific community is not very good at putting that together, let alone for linking all the data for more than 1 million people. Now that we have the data, she said, it's important to have the platforms available to discover data effectively and quickly.

This will speed discoveries and transform health. Scollen used an example from her past research, a gene (SCN9A) associated with extreme pain disorders. In 2006, researchers found that loss-of-function mutations to SCN9A cause complete insensitivity to pain. Other mutations result in a gain-of-function and cause primary ethromelalgia, extreme pain in the hands and feet. It took about 10 years from the first genetic discoveries to gather a compelling package of evidence, including identifying other mutations in this gene and functional data. This is the level of data that is required to suggest a drug target from a genetic starting point, Scollen said.

Scollen works for ELIXIR , an inter-governmental organization building the data infrastructure for life sciences across Europe. The coordinating Hub, where Scollen is based, is in Cambridge, UK. ELIXIR Nodes are dotted in each member country, with their own specific national priorities and strengths. Together, they comprise more than 160 institutions.

ELIXIR is working with GA4GH and the user community to develop standards and tools to simplify searches and requests for identifiable data across international and national borders.

Scollen is co-leading the GA4GH Beacon project, a successful experiment to see if people would share some level of data. ELIXIR will be funding and implementing the Beacon technology in its nodes and with partners to bring European data forward to be discoverable on an international scale. It has also developed a three-tiered access system for different data levels: Public (accessible to Internet users), registered (accessible to bona fide researchers), and controlled (authorized and signed agreement needed). The next steps fall into security, deployment, and data sources. It's possible to identify people if you have the same sequence, by putting in a very large number of different requests and amassing data, so ELIXIR will be monitoring and preventing this activity.

The partnership between ELIXIR and GA4GH is important to facilitate data sharing internationally, beyond European borders. It's not simple to put an infrastructure in place, and there will be many problems to solve, chunk by chunk. Once the infrastructure is in place, Scollen said, we will be able to expedite discoveries in this field. This will lead to a better understanding of disease, generation of pharmacological hypotheses, and patient stratification for precision medicine approaches.

In a questions and answer session, Scollen noted that Beacon searches can be conducted more broadly than a simple nucleotide change, but for other alterations. She acknowledged there is an ongoing discussion at the Plenary and beyond about what constitutes a "bona fide researcher" to qualify for registered access.

Variant Databases: Tripartite Responsibilities | **Download Slides**

Bartha Knoppers (Centre of Genomics and Policy, McGill University) addressed the evolving responsibility for shared variant databases, a topic she acknowledged may instill more fear than positive passion.

For the those in the audience who were new to GA4GH, Knoppers outlined and invited people to get involved in the Regulatory and Ethics Working Group (REWG), which she co-chairs with Kazuto Kato (Osaka University). The most controversial thing the REWG did at the creation of GA4GH was to completely flip the traditional ethics policy approach, Knoppers said. Instead of building a foundation on the presumption that research is harmful, it took a human rights approach based on the right of citizens to benefit from science and its application. This right to science has its origins in article 27 of the 1948 UN Universal Declaration of Human Rights , (later elevated to an International Covenant and signed and ratified in 1966 by 164 countries) reaffirming the right to science and the right to recognition:

1. Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits.
2. Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.

In its first three years, GA4GH has concentrated on the right to benefit from science. The second part, the right of contributors, such as scientists with intellectual curiosity, to be recognized, is why attribution is also important in data sharing, she said. The GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data, translated into 13 languages, has simple principles and short procedures to help in interpretation. Additional policies are available on the REWG web page.

Genetics are changing the classical duties for physicians, researchers, and the health care learning system, Knoppers said. GA4GH has a best-practices accountability policy for the different stakeholders that also incorporates the need for public variation databases. She discussed a three-way sharing of responsibility, since published December 15, 2016 in Genetics in Medicine. To what extent does a lab have the duty for quality control and quality assurance measures? To what extent does a database have the duty to update, curate, and annotate? To what extent does a physician have to diagnose and stay up to date with the changing genetic knowledge?

As a case study, she reviewed a request to REWG by the BRCA Exchange t ea mto look at the question of liability, which Knoppers prefers to call “responsibility.” The team is considering a third tier for the BRCA Exchange web portal. The first tier is a publicly available expert-reviewed space for clinicians and patients, with data that has been curated and interpreted by a standardized expert review process and expert panel, as specified by the Clinical Genome Resource (ClinGen). A second public space will provide access to variant databases around the world. The data contained here come directly from data submitters and can contain conflicting information about specific BRCA1/2 variants. A third space, which is still under development, plans to provide secure storage for individual-level variant evidence. This evidence can be used by experts to annotate and curate variants without compromising the requisite security and privacy of the individuals whose data is being analyzed.

Knoppers related a pending U.S. case, Williams vs Athena, in which a boy with symptoms of epilepsy was treated and died at age 3. He was found to have Dravet Syndrome, a form of epilepsy for which the usual anticonvulsant agents are contraindicated. The lawsuit alleges Athena failed to provide an accurate genetic test result and failed to update its variant classifications. When this variant database issue comes to court, the sharing of responsibility among the different partners is going to be an important issue, she said.

The latest REWG contribution has been a response to the U.S. FDA’s request for comments on a policy to regulate next-generation sequencing (NGS) tests. A central question is what public variant databases the FDA will recognize as providing evidence of clinical validity for NGS test submissions. In a comment drafted by REWG Coordinator Adrian Thorogood, GA4GH recommends that the FDA work to establish an international recognition process. If the FDA accepts the offer, GA4GH and its partners will work with the FDA to set standards for the computational representational variation used to exchange data among recognized variation databases.

In a question and answer session, one person pointed out the gray areas of variants and their interpretations. For example, it takes time after publishing a paper linking a variant and its effect before it’s accepted—or rejected, if it turns out to be wrong. And the information guiding clinical treatments of people with hereditary diseases is often not published. Knoppers replied that courts make decisions based on practice and professional norms at the time, not hindsight. The FDA guidance also looks at the issue of updates.

Another audience member raised a fourth aspect of responsibility: The patients themselves, who are supposed to be controlling their own health information, particularly in the arena of personalized medicine.

Tooling Meta to Micro | Download Slides (Anthony Brookes) Download Slides (Anthony Philippakis)

In a tag-team presentation, **Anthony Brookes (University of Leicester)** and **Anthony Philippakis (Broad Institute of MIT and Harvard)** discussed separate tools they hope to bring together in the service of data sharing.

Brookes opened by identifying a fundamental challenge: to disentangle possession of data from control of data. The idea is to allow data generators to share data while still maintaining a sense of control over its use and without overly-limiting the activities of data users. It's like a good negotiation process, he said, but it can be automated to speed up discovery and sharing of data.

He discussed the first product that has emerged from a year-long effort by the Automatable Discovery and Access (ADA) task team (co-chaired with John Wilbanks and coordinated by Emily Kirby with support from P3G and Intel).

Many rules govern what can and cannot be done with data, including consent, legal and policy requirements, and institutional and research requirements. The information exists in various complicated and often paper-based forms, ranging from spreadsheets to an investigator's notes. Because it's not standardized, it's open to interpretation and difficult for both custodians and potential users of the data to know if and how they can use it for different purposes, such as discovery.

So the ADA team created the first version of a matrix called ADA-M, a standard, machine-readable structure to represent the complicated assortment of data-use conditions and permissions. Computationally, that equates to generating metadata about the datasets, Brookes said. Others might call it a license for data sharing under certain terms.

The developers deliberately struck a balance between a rigid and flexible structure. They worked for something granular enough to capture diversity but with few enough elements to remain useable and to fit both retrospective and prospective data sets. They have produced specific standardized guidance and support documents, plus software to help people enter a pertinent set of use conditions into the ADA-M (called an ADA-M 'profile') with content validation and in a standardized file type (JSON or key:value text formats).

Potential uses of ADA-M include: To describe a dataset's accessibility; to enhance discovery with extra information (such as with Beacon); to enable automated data sharing if ADA-M profiles align and conditions are met; and even to support the Ethics Review and Equivalency task team's efforts to encourage greater regularization and optimization of parameters and conditions for IRBs.

The use categories include countries, types of organizations, types of research and clinical purposes, types of restrictions, and terms of institutional agreement requirements. Brookes encourages people to test it out and provide feedback.

Next, Philippakis talked about a parallel project to automate access to data, which aims to take the same ideas, plus consent codes developed by Stephanie Dyke, reduce it to practice in the form of working code, and conduct a pilot study to show proof of concept.

First, he outlined current practice. When data is deposited to a repository, such as dbGaP, it comes with a data use letter generated by an IRB that summarizes what the informed consent form says about appropriate and secondary use of data. On the other side, a data requestor fills out a form outlining the research purpose for using the data, which is evaluated by a data access committee (DAC) to see if the two are compatible. Sometimes the answer is yes, but other times it is no, suchas a commercial researcher seeking a dataset that is available only for non-profit research.

The human-mediated approach has served the data-sharing community thus far, but will present a scalability challenge as more data become available and more people make requests. At the Broad, for example, a routine request for data sets that have certain characteristics can take months of emails among researchers and project managers.

Working with Broad compliance officer Stacy Donnelly, Philippakis considered how to automate controlled access to datasets—even if only within the Broad. They settled on two axes of control: Who are you and what are you doing with the data? For example, there may be a dataset that is only available to members of a consortium before publication. Or someone may want to use it for a purpose incompatible with the consent form.

The goal is to make data use restrictions machine readable, instantiated in the access control lists (ACLs, or “ackles”) that live on every file server. So began a long journey to make the data request machine-readable. First, Philippakis and Donnelly collected every data use letter they could find and went through them one at a time to figure out how to build an ontology. Four categories captured about 90 percent of the 200 data use limitation letters: Diseases, commercial use, special populations, and future use for methods development, aggregate statistics, and controls.

Then came two important questions: How to operationalize the structuring of data use so that it can be used by researchers, and how to validate this new idea to show it actually works.

Philippakis and his team developed a piece of software call DUOS. The three-part platform allows both a data use letter and a research purpose request to be written with the same ontology. Then the DAC can run an algorithm to check the consistency. To validate DUOS, they consulted an IRB for approval to test the standard protocols against those generated by the software. The trial began in June, testing the structuring of the data use letter and data access request, as well as the matching algorithm result.

After the talk, the first questioner asked about the most exotic clauses on the 10 percent of data use letters that could not be represented easily in the DUOS project. For DUOS, Philippakis replied that one of the big operational challenges are terms like “diabetes and related diseases.” If someone is diabetic, almost any other disease is informed by the diabetes, so that permission is marked only as diabetes, to err on the side of less access but not risking a breach of consent. DUOS is not working with time restrictions, such as available for use before or after a certain date. Other things are logically inconsistent, such as a neurological disease dataset that can also be used for cancer research as a secondary purpose.

In ADA-M, Brookes has also run into similarly nuanced issues in field tests. If a dataset has a restriction for nonprofit use, does that also include a nonprofit project in a company or on a profit project in a nonprofit organization? Some dependencies are difficult to capture, such as a dataset with unfettered use in one country, but for certain disease research in another country.

In response to a comment about how this will measure how often ludicrously restricted consent forms prevent good research, Philippakis said that they have created check boxes, rather than asking people to fill in text, which results in clearer data use descriptions.

A question arose about international data sharing, citing a recent opinion survey finding more constraints there than with commercial requests. In practice, Philippakis said they’ve seen so few that they have not developed a geographic ontology.

Another person asked about the GA4GH role in standardizing consents to maximize use of data. Philippakis said an easy solution now is to consent for general research, and much of the problem-solving revolves around legacy datasets that have been around for 40 or 50 years. Brookes raised the prospect of healthcare data access for research. A recommendation in the United Kingdom proposes a simple checkbox, yes or no, for patients to consent their data for research. It contrasts with the granular permissions in some research datasets.

Philippakis agreed with another audience member who hoped that these automated solutions are a necessary, if long, bridge to a simpler world, where people need only to tick off a few consent boxes. A shorter or “ideal” list would be a longer term goal as broader forms of consent rely on the strength of supporting structures, agreed Stephanie Dyke (McGill University), who has been leading the development of a list of GA4GH Consent Codes , which now number about 20. They are based on Philippakis’ research, as well as on NIH guidance and validation with the European Genome-phenome Archive (EGA) at

European Bioinformatics Institute (EMBL-EBI). As a second stage, she suggested it would be interesting to produce a more "current" set of Consent Codes to reflect today's consent expectations in different parts of the world.

Finally, an audience member affiliated with the NIH Trans-Omics for Precision Medicine (TOPMed) Program noted they run across use cases never considered previously and consult the original investigators for interpretation of how to work with their study datasets. Philippakis responded that TOPmed would be a great test bed because of its rich set of cohorts with unfortunately complex data use restrictions with which the community is struggling.

From Discovery to the Clinic

Session chair **Kathryn North (Murdoch Childrens Research Institute, University of Melbourne)** opened the session with an overview of the GA4GH Clinical Working Group (CWG). The group helps guide GA4GH data-sharing solutions toward direct clinical benefits and assesses what needs to happen to implement genomic medicine into clinical practice.

As with many of the national initiatives around the world, the starting focus has been on gene discovery in rare disease and on targeted therapies in cancer, North said. In rare disease, extensive work in human phenotype ontologies led by Peter Robinson and Melissa Haendel has led to tools for recording clinical phenotypes in standardized and computable ways that can be linked with genomic data. A huge amount of related work remains in cancer—to monitor the longitudinal progression of disorders and also to take into account the outcomes in relation to the therapy. Another area of activity led by John Mattison is to liaise with the purveyors of electronic medical records to use a common language, moving toward a time where phenotypes will not need to be separately recorded and instead can be mined from the medical record.

North cited a meeting the previous day with leaders of many of the national genomics and precision medicine initiatives around the world, including England, Canada, USA, Australia, Japan, Africa, France, and Brazil. They came together to compare their approaches, look for ways to harmonize data collection, and learn from each other about how they're using that data in practice and interacting with patients. A catalogue of international genomic data initiatives is available as a living document on the GA4GH web site. North invited people to add new initiatives to the list.

The final role of the CWG has been to incubate several demonstration projects that are driving the use of clinical and genomic data sharing. The session showcased two examples, the BRCA Challenge and the Matchmaker Exchange, as well as the patient perspective.

BRCA Exchange: Answering the Challenge | Download Slides

Heidi Rehm (Partners Healthcare, Broad Institute, Brigham & Women's Hospital, Harvard Medical School) presented updates about the GA4GH BRCA Challenge demonstration project. The main goal of the BRCA Challenge's web portal, the BRCA Exchange , is to improve the care of patients at risk for breast, ovarian, and other cancers by creating a decentralized global public repository to assist in the analysis of BRCA1 and BRCA2 variants. BRCA1/2 gene testing is one of the most commonly offered clinical genetic tests with well-known medical implications, but no single public database harbors all of the publicly available data, Rehm said.

The ultimate goal is to create a one-stop shop for all available data on these two genes—including an array of knowledge ranging from clinical data to biological mechanisms—to facilitate clinical-grade variant classification. The team has succeeded in aggregating most of the publicly available variant data, with a few exceptions. The interpreted variants are also deposited in ClinVar so they are integrated with variants from all genes.

The BRCA Exchange web portal has two tiers. One tier displays expert-reviewed variant classifications only; the other tier displays all assertions of pathogenicity for aggregated variants from the original data submitters, including conflicting assertions.

The project uses GA4GH tools and resources, including data format standards, Application Programming Interfaces (APIs), security consultations, and consent guidance. The data can be viewed on the BRCA Exchange web site and is also available for download and unrestricted use. The BRCA Challenge's expert curation effort is led by members of the ENIGMA Consortium , a long-standing multidisciplinary international consortium to classify germline variants associated with hereditary breast cancers.

The BRCA Challenge demonstration project also held a half-day meeting on October 17. Rehm presented highlights from that meeting. One pilot project aims to gather family history information, a valuable type of data to support interpretation of variants, from clinical testing laboratories.

Another presentation from the 17th described the ongoing aggregation of data by the Brazilian Initiative on Precision Medicine (BIPMed). Some of the contributed data will be made public, enabling their addition to the BRCA effort. To the north, some of the 22 laboratories that have been sharing data in the Canadian Open Genomics Repository (COGR) are specifically participating in the BRCA data-sharing effort, working to aggregate all of the variants and moving toward consensus interpretations. They will use the same underlying infrastructure as the BRCA Exchange to launch a web site to display and distribute the data.

Finally, a moving presentation from Anna Kuwada and her mother offered a patient perspective on the value of tools such as the BRCA Exchange and ways that those tools could benefit patients.

One major outcome of BRCA Exchange has been a change in mindset, Rehm said. People recognize the utility of sharing data to improve what everyone can do. People compare, discuss, and usually resolve differences in interpretation based on the richer shared data. A major challenge has been that everyone's data is formatted differently in different systems and sometimes interpreted differently using different terminology. The process of sharing in centralized places has motivated people to evolve their data structures to common standards in part due to the efforts of GA4GH.

In the questions and answers following her talk, Rehm said a similar effort will be needed for each of the other 20,000 genes in the genome, but she thinks it will be easier for subsequent genes that can use the same infrastructure and approaches. As for the next genes to prioritize in this way, Rehm said that ClinGen project leaders are taking a thoughtful look at where the most clinical testing is being done, as well as areas with the most difficulty and discordance in interpreting variance. But at the end of the day, a motivated person or group with a favorite gene, a community, and the requisite infrastructure is what makes things happen, she said.

Matchmaker Exchange: The Next Generation | Download Slides

Kym Boycott (Children's Hospital of Eastern Ontario) revisited the origins of Matchmaker Exchange (MME) and laid out some next steps for the rare disease patient matching platform. There are 7,000 known rare diseases and likely thousands more, Boycott said. Most affect children and often have no definitive diagnosis, treatment, or cure. The advent of whole genome and exome sequencing cannot by itself fulfill the rare disease community's hopes of finding a clear cause of a rare disease in a particular patient or family, but a potentially transformative impact could be made if another patient could be found with a similar phenotype and genotype. The idea of a genomic matchmaker arose in several countries, Boycott said, as "a kind of dating service for lonely exomes." Each lab, group, or organization went about building datasets in different ways, but they were small and siloed.

Three years ago, several rare disease databases teamed up to create the MME to make it easier for clinicians to find answers for unsolved patients. They established a federated network, keeping the data under local control, but connected point-to-point with an API that allows each database to talk to the others. Seven database nodes are now connected.

Each node is an entry point into the system, Boycott said. It works by asking if there are any other patients with similar symptoms and a matching candidate gene. After a patient is entered into one of the nodes in the MME, the user can check any of the connected nodes for matches without needing to deposit their data again, which increases the ease and efficiency of the matching process.

Boycott related a recent successful match that discovered a novel cause for a disease. A Canadian clinician working through PhenomeCentral discovered in GeneMatcher a second child with intellectual disability and epilepsy and with a compound heterozygous mutation in the same candidate gene. The cohort has grown to six children, with subsequent functional work now completed and about to be published. More information and other success stories can be found in a special October 2015 issue of Human Mutation .

Looking ahead, the MME needs more cases to be more effective, Boycott said. As reported in Human Mutation , a team calculated that to identify another 2,000 disease genes, MME will need 50,000 to 250,000 cases connected in the network. MME currently connects about 28,000 cases.

The MME team also wants to increase the specificity of the matching, because the false-positive rate may become burdensome as data-sharing services scale up. Strategies include incorporating the mode of inheritance and more phenotypic data, as well as probabilistic approaches to score the likelihood of a real match.

Boycott shared that the MME also wants to track successes more systematically, which may include metrics of how many times someone made contact with MME, how many matches were made, how many matches were meaningful, and how many were published.

To expand its connections, the MME will be working with clinical labs for more systematic sharing to capture the increasing numbers of patients being sequenced in this setting. There is also growing interest in patient-led matching. Patient-led data submissions are facilitated as part of two MME-connected databases, but they are not yet shared in the MME system. Efforts are underway to move in this direction, including a patient guideline on how to use the MME.

So far, the MME requires each user to have a hypothesis about possible candidate genes, called two-sided hypothesis matching. The MME team would like to expand to other matching approaches. The next step is one-sided hypothesis matching, where someone with a candidate gene and phenotype can search unsolved datasets for patients with similar phenotypes and likely-deleterious variants, without those variants necessarily having been identified as candidates.

Ultimately, Boycott envisions a computer in the cloud running through all the unsolved cases and coming up with matches. An European team is launching an independent pilot project next year for autosomal recessive intellectual disability. An audience member observed that HLA matching for bone marrow transplants are a long-standing precedent and asked what is needed for a computer to emulate the success of old-fashioned case-matching between colleagues over dinner. The clinical phenotype, Boycott answered.

Responding to a question about whether clinical labs should return results for genes not yet causally associated with phenotypes, Boycott noted that three large U.S. labs now include secondary reports where they flag a compelling candidate gene. She predicted Canada will adopt that approach. It ultimately allows for discovery with more research. Another audience member noted that matches are made in real time in the MME, so they are not stored and will not trigger an alert if matching data is subsequently entered. At this time, a requester needs to ask again at another time to find newer relevant cases.

Boycott acknowledged the concerns of another audience member that in some countries, such as the U.S., genomic test results are being returned outside of the purview of medical geneticists, who are specialists that tend to be most comfortable with one foot in the clinic and one foot in the discovery camp. Untrained patients or busy physicians may lack the specialized knowledge and follow-up mechanisms to cope with uncertain test results. In fact, clinicians may actively oppose having the extra information and uncertain results due to the extra responsibility and potential for harm. A mechanism must be envisioned to facilitate discovery in this situation if genomic medicine is to make its full impact in rare diseases, Boycott added.

Putting Patients on the Genomic Discovery Team | Download Slides

Alastair Kent (Genetic Alliance) talked about the necessity of patients on the discovery team. The world is on the cusp of an unprecedented capacity to generate data, but patient perspectives are essential to turn those massive isolated facts into information, into knowledge, in opportunities for action, into action, and into interventions, Kent asserted. He appealed to the future patient in the academics, regulators, and other stakeholders in the room.

Most rare and many complex diseases are not as well characterized as they could be. Experience gathered from the patient helps define not just the syndrome, but defines the questions being asked and the problems being solved. For example, patients want researchers to concentrate on the aspects of a disease that interfere most with their lives, rather than use scarce resources to address less important issues around the condition.

Patients also want a say in the risk-benefit calculation. Not all benefits are equally valuable and not all risks are equally acceptable. As an example, Kent pointed to a new medicine for Duchenne muscular dystrophy (DMD) just licensed through the European Medicines Agency and the US Food and Drug Administration, which looked at a hard endpoint of a six-minute walk test as a clinical trial endpoint. That’s a huge event in the disease, but in some cases a wheelchair will provide more mobility. What really matters to the boys with DMD is the ability to use their hands, allowing them to interact with mobile phones and keypads, broadening their social engagement. Unfortunately, there was no six-minute hand-function test.

As research partners, patients can add value by attracting crucial resources at many key steps—recruit fellow research volunteers, validate outcomes, help generate data-rich registries, build scientific collaborations and joint ventures, advocate for translational research, contribute and raise money, influence the regulatory environment, help with the transition into clinical use, and convince healthcare systems that it’s worth committing the resources.

Patients want to share data in a responsible and useful way, Kent said. In fact, they may expect information to flow between doctors and scientists. Instead, they often find themselves being the channel of communication, and they get fed up telling their stories again and again to people who should be talking to each other. Later, he added that creating the confidence to share information comes not from perfectly secure environments but instead from a system that is policed with sanctions to prevent or mitigate the risk of abuse.

Gone is the old hierarchical model where scientists invented, industry created, doctors prescribed, and patients complied, whether it worked or not to improve their lives. Patients bring a different expertise. By working together, patients can help generate innovative solutions and improve outcomes.

In a comment, Marc Williams, director of the Genomic Medicine Institute of the Geisinger Health System in Pennsylvania, noted that the U.S. Patient-Centered Outcomes Research Institute (PCORI), created by the Affordable Care Act, requires evidence of patient partnerships, up to and including having patients as investigators on research projects. He said PCORI has money to spend in the rare disease area. His own work on a PCORI project was eye opening. Patients helped inform the design and then partnered with the team through the completion of the project. Now for IRB approval, investigators have to defend why patients are not involved as investigators.

In response, Kent raised the prospect of capacity building in patient organizations and making the system more permeable to patient perspectives. Patients may have to take time off work, while others on the project are paid, and they often don't share a common language with the research community.

Keynote: Data Sharing, Public Health, and the Bioeconomy | Download Slides

Chair: Robert Cook-Deegan (Arizona State University)

In his keynote talk, **Gerardo Jimenez-Sanchez (Harvard T.H. Chan School of Public Health)** recommended that local and regional communities be considered in strategies to advance genomics research and move it into clinical settings around the world.

From his experience as founding director of the National Institute of Genomic Medicine, Jimenez-Sanchez enumerated the possibilities and lessons for engaging other emerging economies. As the Human Genome Project was wrapping up at the turn of the century, it was clear that genomics could not simply be imported into a country, especially one with ancestries as diverse as Mexico, where at least 65 different ethnic groups speak their own language. The Institute embarked on a Mexican HapMap project, creating interest and enthusiasm with extensive social engagement among a population largely lacking formal education.

Teams traveled to remote rural areas to talk about the double helix and review the consent forms in native languages. Importantly, teams returned to share results in a series of comic books, information kits, and community meetings. Each participant received the same information packet that had been presented to Mexico's president in the ceremony announcing results. The experience contrasts with foreign groups grabbing indigenous blood samples and securing patents without consent or adequate compensation.

Guided by national health needs, Jimenez-Sanchez and his colleagues teamed up with Eric Lander at the Broad Institute to investigate the genomics of type 2 diabetes, which is two times more prevalent among Mexicans than among non-Hispanic US whites. They secured funding from a Mexican business magnate. The researchers found a risk haplotype occurring in 28 percent of the general Mexican population and in 48 percent of indigenous peoples—but completely missing in African genomes and present in only 2 percent of European genomes.

In emerging economies, genomics may not be a high priority, but strong personal leaders with scientific training and global experience (in the case of the Mexican HapMap and diabetes studies, Mexico's secretary of health) may be able to seize opportunities and make rapid progress. Human genomics belongs to a larger virtuous cycle of innovation that can pull together ideas and people across sectors for other benefits, including products and services that can maximize the impact, Jimenez-Sanchez said. He called on GA4GH to engage with scientific and health leaders in emerging and developing economies to address specific and common barriers to bringing genomics to advance public health.

Through the lense of global public health, the benefits of genomic technologies also may extend beyond human health to veterinary medicine, agriculture, environment, food, the biotechnology industry, and general economic benefits emanating from those activities, Jimenez-Sanchez said. He urged GA4GH to engage with the Organisation for Economic Co-operation and Development (OECD), a global group that promotes policies to improve the economic and social well-being of people around the world. An upcoming OECD session at the February 5-7, 2017 Human Genome Meeting in Barcelona will feature opportunities and challenges in personalized medicine.

Big Genomic Data and Industry Panel | Download Slides

Session chair **David Glazer (Verily)** opened the industry panel with a reminder of the high initial industry interest in GA4GH at its founding by nonprofit members. More than 40 percent of GA4GH members now hail from industry. Presentations explored ways that the Global Alliance can help enable the growth, health, and opportunities for business and, reciprocally, what the respective businesses are doing that may contribute to the goals and missions of GA4GH.

Andrew Ury (ActX) described ActX as a business that builds genomic decision support into the electronic health record (EHR). The company sees, and wants to overcome, a last-mile barrier that prevents genomics from reaching the clinic. For genomics, clinicians need computer decision support. The way to make that practical is to build it into the point-of-care in the electronic medical record, a tool most doctors in the developed world are using but which needs enhancement to record, analyze, and act on genomic information, he said.

ActX built a web service that can be built into EHRs, and is now partnering with most major US EHRs. ActX takes care of storage, processing, and content. Individual institutions can customize the service for drug-genome interactions, actionable risk alerts, and a built-in genomic profile.

The standards-based interoperability goals of GA4GH are critical, especially the ability to exchange genomic data from valid sources, Ury said. Standards alone are not enough, he said, based on his experience working with other standards organizations. Plug-and-play interoperability also needs a locked down implementation guide about how the standard is used. All the details it takes to make a standard work are important, too. Ury requested continued enhancement of GA4GH variant call formats (VCF), including more genetic use cases and demographics. ActX is also interested in BAM and consent standardizations.

In support of interoperative data sharing, ActX can enable turning VCFs into precision medicine through a decision support platform adopted by most major EHRs, with genomic decision support fully integrated into the EHR and normal workflow. Prescriptions are checked as they are written against adverse reactions, efficacy, and dosing. Separate alerts are sent for genomic risks that the doctor and patient can do something about.

Jonathan Hirsch (Syapse) explained that Syapse wants to democratize access to precision medicine through a software platform for health systems. The platform’s four elements include integrated clinical and molecular data, clinical decision support (now focused on oncology), clinical workflow and care coordination, and a learning health system (a compilation of clinical, molecular, treatment, and outcomes data across all patients).

Meaningful data standards “with teeth,” or the power to make people obey them, are crucial for companies working toward the future of precision medicine, Hirsch said, as represented on the panel. For example, physicians are still receiving 20-page lab reports as faxes or in PDF form, rather than as structured variant data, turning structured genomic information into unstructured information. This is unacceptable for sharing data across health systems and research environments, said Hirsch, who asked everyone to stop using fax machines for genomic information.

Industry wants to focus on innovating around clinical delivery and decision support, he said. They don't want to focus on the basics of interoperability, but their interoperability solutions may help inform GA4GH work. Hirsch pledged to work with the standards that GA4GH develops. Industry also needs scalable standards for its extensive range of real-world use cases.

In fact, many people in the genomics and health care community are looking to data-sharing leaders to help them understand how to scale the process. In addition to technical interoperability, other GA4GH initiatives will help make it easy and safe for a health system to say “yes” to data sharing, such as standardized consents and de-identification.

Steve Lincoln (Invitae) provided a perspective from a medium-sized reference laboratory. Invitae has made a strategic decision to share all the data it can within the law, in contrast to the historical norms for labs. Data hoarding fundamentally retards the practice of precision medicine, reducing the quality and effectiveness of care, Lincoln said. Conversely, getting data out there under review by the entire community improves confidence in the results, helps physicians learn how to use genomic and genetic data, and will help payers realize the value of these tests in routine patient care.

Lincoln called out the BRCA Challenge, Cancer Gene Trust, and Matchmaker Exchange as important GA4GH efforts, as well as the registered access model to share richer data than now can be made public. The technical mechanisms of standards and formats are crucial, he said, but the most important thing to do is to encourage everyone to share their data. In fact, he advised publicly calling out those who are not sharing data, such as some major laboratories. That also includes those who say they are sharing data, but are really not, which he called “share-washing.” When the actual practices are transparent, then the physicians and insurance companies can choose to do business with those who are doing the right thing for the community.

David Glazer (Verily) has been working with Google genomics and engaging with GA4GH for three years. When Google restructured, Glazer moved the genomics team into Verily, a life sciences company in the Alphabet family of companies. Verily builds software (and hardware?) to store and analyze genomics data and its relationships to disease (?need sentence to describe what Verily’s business is)

From GA4GH, Verily needs three equally important things: A model policy framework for consent standards, researcher data access standards, and data security guidelines; a portable technical framework for standard ways to package tools, for the tools to access data, and to curate data; and ongoing moral leadership to “do the right thing,” Glazer said. All three of them are equally important and necessary. GA4GH is uniquely positioned to reinforce globally that it is the right thing in most circumstances to share.

Not everyone wants to share anything, and not everyone needs to share everything, Glazer said, but it’s important to take away the excuses and confusion, so that to share or not to share becomes a deliberate choice. It should be easy for chief information, data, or medical officers to browse and select options from a toolkit of the globally accepted ways to share data, secure in the knowledge that they are doing things the right way. Until now, he said, individual businesses have had to invent sharing rules and procedures from scratch, and it’s been painful.

Meanwhile, in its work with the U.S. government (NCI Cloud Pilots, Precision Medicine Initiative) and nonprofit partners (MSSNG), and in its own commercial projects (the longitudinal Baseline Study), Verily is supporting the evolving GA4GH standards, as well as supporting “commons” initiatives. Publishing all the data in exactly the same format with exactly the same tools will make life easier for everyone, Glazer said. In the meantime, Verily is working with other open source tools likely to become part of the standard, such as the Broad Institute’s Genome Analysis Toolkit .

In the question and answer session, an audience member reflected on the natural affinity between GA4GH and the panelists’ business models of managing, transmitting, and validating knowledge. He then questioned what benefits diagnostic or pharmaceutical companies would see from sharing data after making a huge investment and needing to protect the return on that investment.

Lincoln said that it takes courage for a company to say, “We think the pie will be bigger, and therefore our slice will be bigger.” In fact, while many people involved in GA4GH believe that to be true, he said, it hasn’t been proven yet. Another reason for a company to share its genomic assets is to avoid being surprised by unexpected results in clinical trials. Also, he said, data sharing can help health care businesses strike a balance between doing the best job for shareholders and aiming for the best care of the patient. For example, Lincoln said he consults with other lab directors to help make sense of puzzling results, even as their respective sales teams are fiercely competing for business.

Hirsch added that the health system controls many of these relationships and should be determining the rules. For example, if a pharmaceutical maker invests in data generation, there should be provisions for the data to be reused and shared for the good of those patients and beyond.

Glazer applauded global data sharing for everyone’s good, but also said it was legitimate for a business that invested in data to decide to not make all of it available to everyone, especially if they are making the world a better place. He said his moral indignation is reserved for people pretending to do things for the greater good (“share-washing”), but actually working against the interests of patients. The point is to have everyone using GA4GH standards and tools, so data is only one boardroom decision away from being shared as widely as desired, and can even be implemented the next day with a one-line code change.

Further discussion centered on how new technologies fit into data-sharing and interchangeable standards, such as a novel sequencing chemistry. There is no silver bullet about how to do this, said Glazer, who advised people to engage early, at a stage when they are still proving that the new technology works. The ongoing effort to create a sustainable sharing environment will be better able to accommodate new kinds of knowledge. As more “omics” information becomes clinically useful, several of the panelists predicted electronic medical records systems will be forced to support interoperability around clinical decisions and workflows. To get there from here, several U.S. initiatives, including the Cancer Moonshot, are considering mandating electronic exchange of data related to genomic and genetic information, Hirsch said. He also suggested that CMS could also condition reimbursements for testing and subsequent therapy upon electronic receipt and storage of data.

One audience member suggested that industry can help by leading efforts to lower costs, such as for sequencing, removing barriers, and promoting wider use of the technologies. True, but generating the data still dwarfs the cost of storing the data for the next five years, Glazer said. Yet, it is in industry’s interest to participate, share, and add value to that overall effort. Anything that lowers the friction — shrinking costs, standards, interoperability — generates more money available for the fundamentals of taking care of patients with better science.

In the clinical setting, the cost of a one-time molecular diagnostic test is miniscule to the monthly therapeutic drug regime that may follow, Hirsch noted. An argument can be made for increasing the cost of diagnostic tests, especially for those that drive down other health care costs, such as a contraindicated drug, he said. In terms of costs and data sharing, Lincoln added, health care systems actually paid for the data not being shared by some laboratories using proprietary tests.

Another audience member observed that there are many mixed models of payment for data generation, including researchers that invest their time and reputation, publicly funded projects, and patient-powered biobanks. Glazer agreed, adding that the standards, policies, and tools at the heart of the GA4GH effort will make it easier to share data in all those models.

Data Sharing: Fast Foward Consent

Chair: Nazneen Rahman (The Institute of Cancer Research, The Royal Marsden NHS Foundation Trust)

Open Source, Open Data | Download Slides

David Haussler (University of California, Santa Cruz) reviewed the technical process and progress of GA4GH toward data sharing in translational medicine. The evolving vision centers on fostering agreement on how genetic variants and other data elements are represented, so that an application programming interface (API) can define how the data are exchanged, transferred, and manipulated. The API can be used in public data sets, and also provide appropriate levels of authorized access to the private data sets hosted by medical research institutions, patient advocacy registries, hospitals, health IT companies, national health services, and other constituents.

The code is all open source and lives in Github, a collaborative web space where hundreds of people worldwide have contributed code under the constant evaluation of the most active GA4GH contributors. Still in development toward the 1.0 version, the GA4GH Genomics API is a constellation of objects making up an ecosystem. Haussler urged individuals and organizations to get involved in developing the API further. Task teams tackle specific technical issues, and the API supports demonstration projects, such as Beacons, Matchmaker Exchange, the BRCA Challenge, and the Cancer Gene Trust.

The underlying bricks and mortar of the API are data standards, which started when the 1000 Genomes Project built BAM and VCF, two file formats that became the de facto world standard for genetic reads and variants. GA4GH volunteers are maintaining those and also building an abstract version of that data so that it can be articulated in the more flexible API format.

A new project aims to create a single ethnically diverse reference using all the world's genomes, rather than an individual genome from one particular ethnic group, as is the case for the current reference genome. Another project answers a surprisingly basic question about what constitutes a genetic allele in precise machine-readable form. Subtle differences exposed by Reece Hart and his Variation Modeling Collaboration revealed the potential for ambiguity and confusion between the major constituencies storing and exchanging genetic variants. This must be solved before using the shared data in clinical practice, Haussler said. A recent addition was the streaming API to transfer large data files efficiently.

One task team aims to bring code to the data, addressing the expectation that code may move more freely, because of privacy restrictions and cross-border genomic data restrictions. Code would move in containers known as Dockers, connected in processes called workflows—all requiring an API that allows the complex code to not only run reproducibly in different data centers, but to understand layers of authorization and access.

A new project, Variant Interpretation for Cancer Consortium, led by investigators at Washington University in St. Louis, U.S., and Universitat Pompeu Fabra in Spain, is uniting a number of the best databases for actionable interpretation of somatic cancer variants to create a one-stop, carefully reviewed and authoritative place to find the information. As of the Plenary, GA4GH counted 92 international genomic data initiatives. GA4GH wants the world to collaborate on the interface, allowing all constituents to communicate, and also wants to encourage competition on implementation.

A Systematic Approach to Data Sharing in Translational Medicine | Download Slides

Steven Jones (BC Cancer Agency) discussed the experience of integrating genomics analysis with cancer treatment in British Columbia and introduced a new data-sharing infrastructure called the Canadian Distributed cyber-Infrastructure for Genomics (CanDIG).

In British Columbia, a single hospital, the BC Cancer Agency, registers, follows, and treats all the cancer patients in the province. It has moved a fairly extensive bioinformatics pipeline into the clinic in the context of clinical trials. It deploys a full genomic and transcriptomic analysis for more than one patient a day. Analyses and tools are tested on a continual learning curve. In the most important and time-consuming steps, data is interpreted by PhD scientists, who then translate to clinicians a synopsis of what somatic or underlying germline variants might mean for their patients. The interpretations can be greatly aided by shared information about somatic cancer mutations, such as from CIViC and TCGA .

In one example of a second recurrence of colon cancer metastasis to a woman's spine, researchers found two unusual genes among more than 1,000 bad actors common in cancers. The two genes were in a molecular pathway that could be targeted by a high blood pressure drug. With that therapeutic insight, the cancer disappeared for more than a year. It may be a rare genotype, but such knowledge of the clinical outcome could help other patients, if shared. "Genomics gives us this wonderful international currency of DNA sequences, which are readily shareable," Jones said.

Toward that end, CanDIG will initially link three major cancer centers in Canada. CanDIG will use the GA4GH Genomics API to access data, such as tumor genetics, and to distribute metadata, such as treatment history and outcomes. The participating sites will control access to data, provide data, and dispense user requests for data. Many questions remain about who can use the data and where it resides. In addressing these constraints, project leaders have contributed authentication code back into the GA4GH server. The $5 million project is led by Michael Brudno (SickKids, University of Toronto).

Citizen Science, Social Engagement | Download Slides

Megan Doerr (Sage Bionetworks) urged the genomic data-sharing community to broaden its perspective and enlist laypeople to collect and share data. The widespread and diffuse nature of phenotypic data is particularly well suited to a network of citizen scientists, she said.

Outside the clinic, data can be most easily harnessed and quantified by mobile phones, a ubiquitous and powerful tool saturated with sensors, and owned by more than one-third of the world's adults overall, and even more in developing nations and some other countries. mPower , a Parkinson's disease study using mobile phones, has provided unexpected lessons. In the first six months, participants generated 7500 responses. A specially designed phone app returned results and allowed people to make observations about their own data (for example, memory test results are better after coffee in the morning or before wine at night).

The researchers' excitement about their study was tempered by a reality check. After 7 weeks, 90 percent of people stopped using the app. When they asked for feedback, researchers learned that the application wasn't as fun as expected on a mobile phone. In fact, a person's results could be a depressing reminder of the disease progression. The research team is retooling their application, and reconsidering how to enhance the self-administered consent form to more completely prepare people for the experience.

Part of balancing benefits with potential harm is fulfilling the promise to turn the torrent of data into useful information, Doerr said. That means expanding access to the data beyond the organization and a small community of credentialed scientists to others who can bring fresh viewpoints and tools for analysis. For the mPower data, 75 people on five continents have requested access so far, ranging from famous Parkinson's disease researchers to people in industry to a high school student in Bangalore. One of the take-home lessons for GA4GH is the need to put the great ideas into practice and test them to move the field forward.

Closing: The GA4GH Tomorrow

GA4GH chair-elect **Ewan Birney (EMBL-EBI, Genomics England)** closed the meeting by video. He thanked departing chair Tom Hudson (AbbVie) for his energy and enthusiasm in developing and growing GA4GH and looked ahead to the next decade of opportunity for data sharing in genomics and health.

Healthcare institutions will be collecting multiple measurements of hundreds of millions of people, including genomes and other molecular features, because of the importance to patient care, Birney said. These high-dimensional data sets have huge value to researchers as well, enabling humans to become the most-studied model organism on earth in efforts to understand, prevent, and treat disease.

Healthcare data quite rightly have different rules from research data. GA4GH must establish a system that allows researchers appropriate ethical access and supplies many different standards, such as security and clinical metadata, that together create an ecosystem that serves both research and healthcare needs.

The last few years has seen brilliant creative work by thousands of volunteers. Next, GA4GH must make data sharing tools that work in actual practice and can be deployed in the big engines that drive research and health care, Birney said. The job is best done by people who are thinking about genomics deeply and practically and who understand the practical importance of delivering this system. GA4GH will be initiating a strategic plan that challenges the working groups and task teams to deliver on this promise to human health, Birney said.