Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

Global Alliance for Genomics and Health

Summary of Third Plenary Meeting

June 9-11, 2015, Leiden, the Netherlands

This document serves as an interactive record of the third plenary meeting of the Global Alliance for Genomics and Health. An executive summary provides a short, high-level view of the three-day event. More detailed, modular session summaries follow, and include links (where available) to presentation videos. An appendix provides summaries of Working Group and Demonstration Projects.

Executive Summary **2**

Strategic Roadmap, 2015-2016 ..... **5**

Introduction **6**

Keynote Speaker: Challenges and opportunities, 2015-2020 ..... **7**

Actions since 2nd Plenary meeting **9**

(1) BRCA Challenge 9

(2) Matchmaker Exchange 11

(3) Beacon Project and Data Sharing APIs 13

(4) Policy: Consent and Privacy & Security 15

Emerging Projects **16**

(1) eHealth 16

(2) Clinical Cancer Genomes 17

National Initiatives **18**

Big Data Challenges and Solutions **21**

Working Group Panel Discussion **24**

Action Items **26**

APPENDIX

Executive Summary

The Global Alliance for Genomics and Health (GA4GH) held its third plenary meeting on Wednesday, June 10th in Leiden, the Netherlands, with ancillary meetings on June 9th, 10th, and 11th. More than 230 individuals attended the meeting, representing 152 organizations active in 24 countries. Key themes of the meeting included a push for implementation of tools and approaches, a need to integrate with other data sharing efforts, and the importance of global and sector diversity among GA4GH membership.

In opening remarks, Steering Committee Chair David Altshuler (Vertex Pharmaceuticals) called on attendees to envision the future of the GA4GH with a "laser-like focus on making a difference" and to come up with concrete actions for the next two years. He also announced changes to the GA4GH leadership: Tom Hudson (Ontario Institute for Cancer Research) will serve as Chair Elect until Altshuler's term ends later in 2015, while David Glazer (Google), Nazneen Rahman (Institute for Cancer Research), and Heidi Rehm (Partners Healthcare) will join the Steering Committee as Charles Sawyers and Betsy Nabel step down.

A keynote presentation from Ewan Birney (European Molecular Biology Laboratory-European Bioinformatics Institute) focused on enabling secondary use of clinical data for research through a federated approach to data sharing, increased global diversity in the GA4GH, and long-term engagement with the clinical community. He noted that in the world of human genetics, whether for rare or common diseases or cancer, "sample size is king." To reach the millions of cases and controls that will be necessary to identify a valuable second rare disease match or to understand the biological bases of cancer, federation is critical. But more than sample size, federation will also allow us to leverage things like genetic drift, an increasingly cosmopolitan global population, and environmental impacts on disease. "The Global Alliance is not only a good thing," Birney said, "but a necessary thing for us to make progress over the next decade. This is the only way I can imagine this working."

The GA4GH has made great strides in the two years since its founding, having grown to more than 330 organizational members across more than 30 countries. Some 750 individual members across 48 countries have developed dozens of priority products such as the genomics API and the *Framework for Responsible Sharing of Genomic and Health Related Data*, which are increasingly being used by organizations, clinicians, and researchers around the globe. But while GA4GH has moved quickly, Altshuler said, the field and the world are evolving faster than anticipated.

Several national initiatives have sprung up around the world to assemble massive cohorts of genomic data. It will be critical to stay relevant and to engage with these efforts, Altshuler said. To that end, an afternoon session welcomed presentations from representatives of eight regions around the globe: Latin America, Singapore, China, The Netherlands, the United Kingdom, the United States, Australia, and Canada. The presenters summarized the current and future data-sharing plans of their respective efforts and discussed ways they are leveraging GA4GH tools and approaches as well as ways GA4GH might help them overcome barriers to effective and responsible data sharing. The speakers expressed eagerness in working with GA4GH and said the Alliance could provide assistance in both general and specific areas, including: (1) taking pilot efforts to a national scale, (2) developing information dissemination technologies, (3)

implementing tools and approaches within a national health system, and (4) creating a Brazilian national database that is representative of the unique mixed population.

In addition to these national initiatives, several organizations have commenced large-scale data sharing efforts. Security Working Group (SWG) Chair Dixie Baker facilitated a session on the challenges facing Big Data, which include an overarching lack of computational capability and infrastructure to generate, maintain, exchange, analyze, and visualize large-scale data sets. Five presentations introduced solutions being developed to overcome those challenges. Virtualizing computational resources, Baker said, may be the only approach for achieving scalability and elasticity for DNA sequencing and analysis, but that approach comes with its own set of hurdles. Plenary attendees heard from a wide range of activities, including a patient advocacy effort, a commercial cloud provider, and an academic medical center.

At the last GA4GH plenary meeting (San Diego, October 2014), the importance of demonstration projects rose to the surface as well as a need for actionable solutions to the philosophical guiding principles being developed within GA4GH. A morning session at the June 2015 meeting focused on updates since the last plenary, with presentations from John Burn on BRCA Challenge, Heidi Rehm on Matchmaker Exchange (MME), David Haussler on the Beacon Project and other activities of the Data Working Group (DWG), and Graeme Laurie and Susan Wallace on the Privacy & Security and Consent Policies developed by the Regulatory and Ethics Working Group (REWG).

The BRCA Challenge held its first meeting directly after the GA4GH Plenary at the UNESCO headquarters in Paris, France. The meeting led to the development of time plans for a BRCA web portal and phone app. The MME now has three databases using its API and is working to connect them together. The journal *Human Mutation* will release an edition dedicated to the MME in October 2015 with more than a dozen papers highlighting its activities over the past 1.5 years, including three that document rare disease cases solved through matchmaking. The Beacon Project has been working with several other data sharing initiatives to create a large, interconnected, international network of datasets. Its near term goals include new development tools and use-cases to foster implementations. The REWG's Privacy & Security and Consent Policies aim to guide practical implementation of the *Framework*, which was designed to serve as a high level starting point for deliberation and action. The policies are now available on the GA4GH website.

In addition to well-established projects, new activities have been percolating in the areas of eHealth and clinical cancer genomes. A session on these projects included presentations from John Mattison (Kaiser Permanente) and Mark Lawler (Queens University Belfast). The eHealth Task Team has created a catalogue of activities to guide increased data sharing collaboration and is working to manage the ecosystem of ontologies and semantic representations. The Clinical Cancer Genomes (CCG) Task Team recently surveyed the GA4GH community to identify ways to highlight and enable collaboration between ongoing data sharing efforts. It has also produced an opinion paper arguing that the international community must overcome several challenges including lack of interoperability and reluctance to share before Big Data genomics can advance human health.

Ancillary meetings were also held in conjunction with the main plenary session. The DWG and the Clinical Working Group (CWG) held a crosscutting meeting on June 9th, during which audience members heard 15 presentations on topics ranging from large-scale clinical data sharing initiatives to more granular efforts to improve standardization and consistency across technologies. By the next plenary meeting, a newly formed planning team co-led by Lawler and Adam Margolin (Oregon Health & Science University) will identify and launch a cross-cutting global actionable cancer initiative that brings together the domain expertise and medical goals of the CWG with the data models and APIs developed across the DWG.

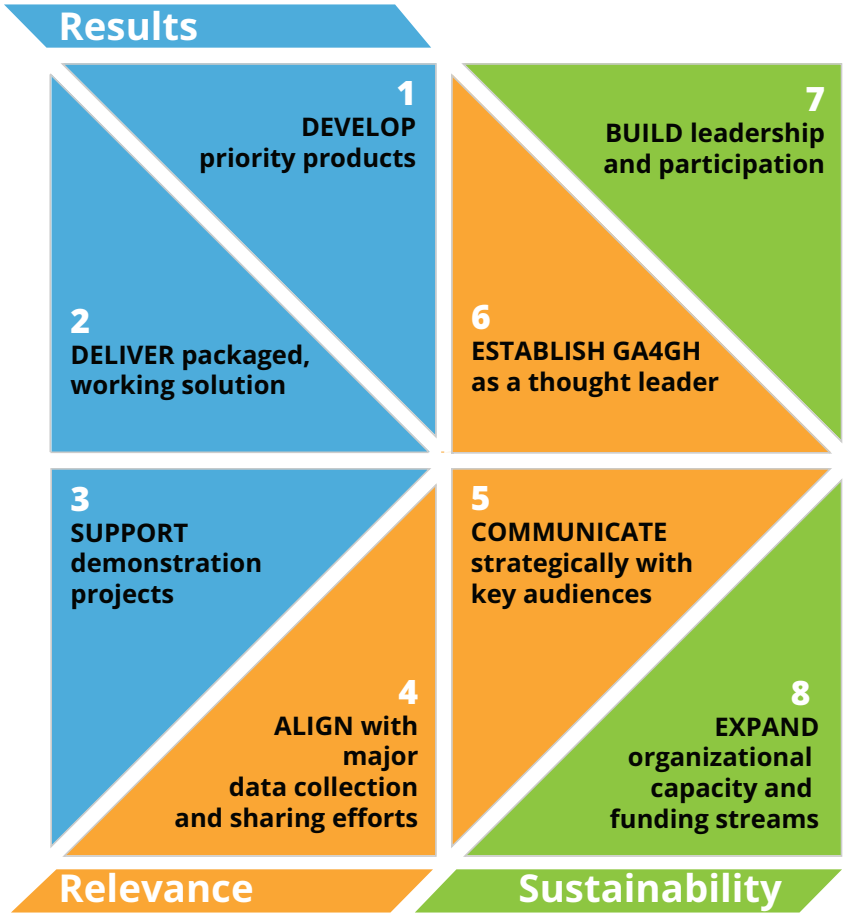
In addition to the cross-cutting cancer session, the CWG, the DWG, the SWG, and the Regulatory and Ethics Working Group (REWG), as well as MME and Beacon all held internal and several cross-cutting planning meetings, and a Beacon hack-a-thon, over the course of the three day meeting. Working Group Co-Chairs provided summaries of these meetings on June 9th, followed by an open-air panel discussion facilitated by Martin Bobrow. Key themes that arose during this discussion centered on engaging patient, physician, and disease advocacy communities, increasing global representation of the GA4GH, and implementing GA4GH tools and approaches.

Finally, a closing session on June 10th facilitated by David Altshuler identified concrete action items for the coming months. The GA4GH Strategic Road Map for 2015 and 2016 helped guide the discussion, which focused on developing priority products, delivering packaged working solutions, supporting demonstration projects, communicating with key audiences, establishing GA4GH as a thought leader, and building leadership and participation.

Some specific recommendations included (1) an initiative of patient advocacy organizational members of the GA4GH to drive increased patient engagement, (2) establishing thought leadership in the area of data sharing incentivization, (3) creating a communications and engagement working group, (4) developing ethics training tools for non-clinical researchers working with genomic data, and (5) encouraging implementation of both the technical tools and regulatory approaches being developed by the GA4GH Working Groups.

Strategic Road Map, 2015-2016

Throughout the three-day plenary meeting, discussions tied back to the GA4GH Strategic Road Map, which outlines activities for 2015 and 2016 focused on priority needs to enable data sharing in areas that provide value to members and the field, so that the effort ensures results, relevance, and sustainability. The objectives and activity directives in the Road Map provide GA4GH Members with the prioritization and framing needed to develop systems, drive concrete actions, and create targeted progress in eight key areas:



Introduction

In opening remarks, Steering Committee Chair **David Altshuler (Vertex Pharmaceuticals)** reminded the audience of the history of the Global Alliance for Genomics and Health (GA4GH) and called on them to envision its future with a "laser-like focus on making a difference."

In early January 2013, the founding members of GA4GH met with nothing but an idea: they sought to create an organization that would bring together people from different sectors, countries, and perspectives to enable genomic and clinical data sharing in order to improve health. "It is remarkable and exciting to see what has happened," Altshuler said.

Now a network of more than 750 individuals across 320 organizations and 48 countries have organized into four vibrant working groups, which are creating the tools, methods, and frameworks critical to enabling genomic and clinical data sharing. Upon this solid foundation, GA4GH members have established data sharing demonstration projects to put those ideas and people to work.

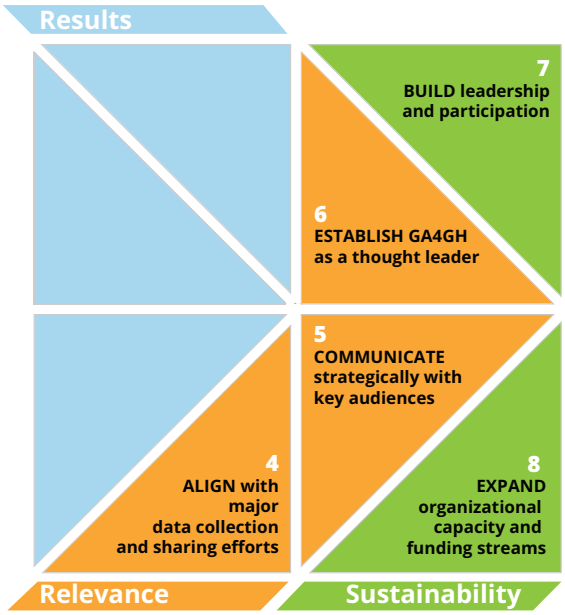
But while GA4GH has made great, quick strides, Altshuler noted, the world has moved even quicker. "We're pedaling hard and making progress but the world is speeding up," he said. The world is at an inflection point, moving from decades of talk to actually seeing genomic data making their way into clinical medicine. Several large-scale clinical genomics projects are developing or underway around the globe. In order for GA4GH to succeed—for data to be shared responsibly and effectively—it must engage strongly with these efforts, Altshuler said.

He also announced changes in the GA4GH leadership, which were established through a nominating committee led by **Martin Bobrow (University of Cambridge).** Current Steering Committee member **Tom Hudson (Ontario Institute of Cancer Research)** will serve as Chair-Elect until Altshuler steps down later in 2015. "I've known Tom for 20 years," Altshuler said, "and I can't think of better person to take this forward." Additionally, he announced three new Steering Committee members: **Heidi Rehm (Partners Healthcare), Nazneen Ramhan (Institute of Cancer Research),** and **David Glazer (Google).** GA4GH continues to search for one to two more Steering Committee members with a particular eye toward geographic diversity. Finally, Altshuler announced that founding Steering Committee members Charles Sawyers and Betsy Nabel will step down. Nabel will maintain a senior advisory role while Sawyers, a current Clinical Working Group Co-Chair, will place more focus on cancer data sharing projects.

Altshuler closed by asking members to consider where the GA4GH should be two years from now. "I think everybody would agree we'd like to see tangible benefits to medicine, health care, and patients," he said. In order to achieve that goal, data sets will need to be brought together in an appropriate, responsible way to enable discovery, better diagnosis, and interpretation of genomic information. But, he said, we must first identify the actual steps to get there. "We need to make sure we don't, in the spirit of trying to solve all problems, not solve any," he said.

Keynote Speaker: Challenges and opportunities, 2015-2020

Ewan Birney (European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute (EBI)) delivered the keynote address in which he argued for the importance of a federated approach to data sharing, increased global diversity in the GA4GH, and long-term engagement with the clinical community.



The worlds of practicing medicine and research are different, Birney said, operating in different language, legal, financial, and regulatory contexts. Data in the clinic are often totally closed, while data in the research space are often totally open. But globally, research is becoming more relevant to healthcare, which depends on the skills and techniques researchers generate. Simultaneously, researchers are becoming more reliant on health care, which will generate a "remarkable data set over the next decade," which researchers want to use to feed back into health care. In order to do that well, Birney said, we must acknowledge the differences between these two worlds.

He used his own research as an example of how this can work: In collaboration with clinicians at Imperial College London, Birney performed genome wide association studies against a set of characteristic cardiac features as portrayed in MRIs from 1,500 healthy volunteers. "We are repositioning a medical dataset that was generated for purposes of human health to understand basic things about cardiac structure," Birney explained.

In the current paradigm, which lacks federation and harmonization between clinical and research datasets, Birney was still able to move his research forward. "Maybe people like me can just charm clinicians and there's enough there," he said. "Maybe we don't need to federate more than swapping papers and collaborating with other scientists. Maybe medical data can stay siloed." But, he said, as human geneticists across all diseases know, "sample size is king."

With the case for sample size well established for common diseases, Birney focused this part of his discussion on the value of large data sets for rare disease. Taken together, rare disease is relatively common, affecting eight percent of the human population. Every time a gene involved in a rare phenotype is identified, it becomes an amazing point of leverage into human biology, he said. "It may have been discovered with a small set of people, but it immediately tells us about other things, often common diseases or common phenotypes." But to do this well, he said, we need to screen millions upon millions of people in order to find a match between just two individuals with precisely the same rare allele with the same rare phenotype.

Sample size is also critical for effective cancer studies, since each somatic tumor is so unique. “We need gob-smacking numbers of people to understand the relationship between somatic, germline changes, and environment,” he said. “The calculations go above 10 million.”

More than enabling large sample sizes, though, a federation of data sets would also allow us to leverage things like genetic drift, human migration, and environmental factors. He also noted that infectious disease research, though a complex area with an existing dynamic scientific community, would also benefit from this effort. Thus, Birney argued that federation is critical for understanding a broad range of human diseases.

As a result, Birney said, “the Global Alliance is not only a good thing, but a necessary thing for us to make progress over the next decade. This is the only way I can imagine this working: a federated system that repurposes clinical medical data for secondary research.”

Over the last year, Birney said, GA4GH has created a forum not previously present, which ranges from low-level technical issues to big-picture clinical issues. In particular he noted that the effort to establish a reference genome graph shows how these disparate communities are working together to deliver a critical tool for the community. He noted that the success of this forum, whose kin often fail, goes to GA4GH’s persistent focus on a large but manageable scope. “My advice is to keep it in this manageable zone. It’s easy to try to bite off all of the problems in electronic health records, or all of the problems in genomics, but if we do that we become too diffuse,” he said, noting that the GA4GH could stand to expand its definition of the term “Genomics” to include RNA, proteins, metabolites, and other “omics” areas beyond the genome.

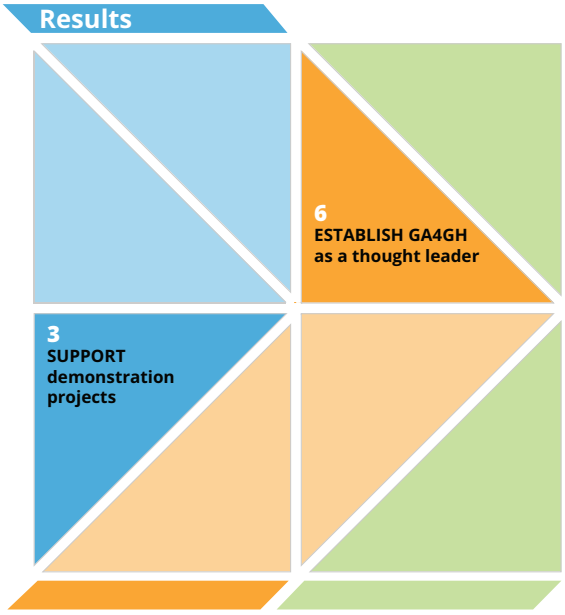
But, he said, GA4GH is still growing and still has places to go. He encouraged a push for implementation of technical products as well as long-term engagement with the clinical community. “We have to be in it for the long game,” he said. “Urgency is useful for motivation, but research and health care have been working for a long time. It is more important to achieve change in five, ten years time than pushing out an API two months earlier than deadline.”

Birney also reiterated the concern for a more global GA4GH, with the current make up too biased towards Anglophone countries. “We need to acknowledge that diversity is good and stretch out of our comfort zone to embrace that diversity,” he said. In particular, he encouraged members to look more broadly across Europe to places like Estonia, Scotland, and Denmark, which are taking unique approaches to ethical data access and regulation, as well as to the entire planet, where there is a huge diversity of talent, diseases, and thought processes.

Finally, he also discussed the various ways his organization is engaging with GA4GH, which include a joint project with the Center for Genomic Regulation in Barcelona to establish a Beacon on top of the European Genome-Phenome Archive, the development of reference and annotation APIs for Ensembl variant data, and an effort to harmonize genomic data file formats.

Actions since 2nd Plenary meeting

(1) BRCA Challenge



John Burn (Newcastle University) provided an update on the BRCA Challenge, which held its first meeting directly after the GA4GH plenary with sponsorship from AstraZeneca and hosted by UNESCO at its Headquarters in Paris, France. Next generation sequencing is revealing vast numbers of variants, Burn said, demanding an ever more pressing need to collaborate. The BRCA Challenge is an attempt to enable collaboration in the area of variant interpretation for the genes *BRCA1* and *BRCA2*. In 2003, the Human Variome Project and the International Society for Gastrointestinal Hereditary Tumors (InSiGHT) launched a Variant Interpretation Committee (VIC) for mismatch repair genes, which brought together everyone working individually in that space. The group established a database, which now receives thousands of hits each month from the clinical and molecular genetics communities alike. At the first plenary meeting of the GA4GH, it was proposed to do the same thing across the whole genome, and to start with *BRCA1* and *BRCA2*. Although not perfect genes from a bioinformatics perspective (they are big with lots of variation and noise), they are good because they are well known and already well studied—meaning there is a significant amount of data to use as a test case for integration.

The BRCA research landscape consists of many databases with unique variants, including NCBI, ClinVar, LOVD, and the French Universal Mutation Database. Federation is needed to cull all of the definitive work around the world to present an authoritative network of databases with a reliable curation system for variant annotation.

Data security is key, Burn said, as it will enable critical buy-in from the public and from patient advocacy groups. For instance, Britain has a large amount of data in the National Health System which could be beneficial to this effort, but it isn't yet being shared because patients are nervous thanks to near-daily data breaches. An Ethical, Regulatory, and Advocacy (ERA) group akin to the REWG of the GA4GH is dedicated to this topic and held a special session at the Paris meeting. In addition to the ERA group, three new subcommittees recently formed to organize and guide activities of the BRCA Challenge. These include the Evidence Gathering Group, the Classified Variant Collection group, and the Interpretation Group.

Burn identified three primary needs of the project, which are being actively worked on by the subcommittees. First, legacy data in hundreds of diagnostic laboratories with up to 20 years of experience issuing reports to families with a history of breast/ovarian cancer must be captured

and retained. Second, an API must be developed for identifying *BRCA1* and *BRCA2* variants from whole genomes and exomes. And third, data from the general population must be collected. Until 2010, Burn said, only those with a family history of breast or ovarian cancer were genotyped. According to research from Nazneen Rahman, one in every 250 people carries a pathogenic *BRCA* variant. “Once people are aware of this, they’ll want to be tested,” Burn said.

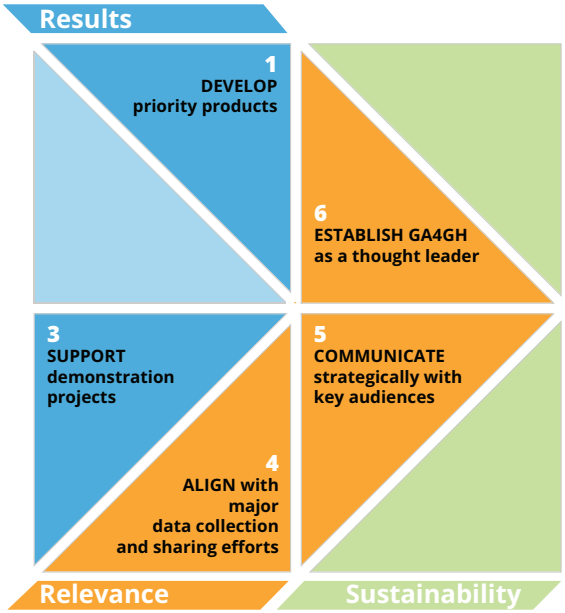
He went on to discuss the story of a Newcastle family that was given conflicting information about the pathogenicity of its variant, which Burn replicated by pretending to be the daughter of the sequenced carrier. He received four conflicting responses from regional diagnostic service laboratories that used a standard approach but interpreted the slightly confusing literature differently. They collectively spent eight hours incorrectly answering his question, Burn said. “We don’t have this kind of time or manpower. We need a reliable reference database,” he said.

A Q&A session addressed engagement with patient groups, engagement with Myriad, and the project’s five-year horizon. “What I want is a database on my phone next year, so I can go to an app and say which pathogenic variants are those,” he said. “If we can get to that point with these two genes, then the next stop is 20,000 genes.”

The BRCA Challenge meeting in Paris led to the development of time plans for such an app as well as for a web portal.

(2) Matchmaker Exchange

Heidi Rehm (Partners Healthcare) provided an update on the **Matchmaker Exchange** (MME), which also held an internal planning meeting and a cross-cutting meeting with the Beacon Project on June 11th. MME is a federated network of genomic databases whose goal is to identify causal genes for rare disease by matching phenotypes and genotypes.



While a single case is never enough to implicate a gene as causative, a second case is sometimes all that is needed to do so for a rare disease, Rehm said. At the American Society of Human Genetics meeting in 2013, a group of rare disease researchers and clinicians set out to link their respective databases, which were each currently serving as smaller, self-contained matchmakers, to help identify second cases and thus be collectively more productive for solving rare diseases.

The task of linking these databases presented a significant technical challenge, since each has distinct data schemas and content types. Some may collect variant call files without identifying candidate genes while others focus entirely on the gene. Still others may incorporate data from model organisms or human phenotype ontology terms. Furthermore, even if there is a match between cases based on genes, the phenotypic data may not be the same.

Two initial matchmaker databases—PhenomeCentral and GeneMatcher—developed an API that sifts through these differences to identify commonalities between cases. They then worked with the GA4GH Data Working Group (DWG) to align the API with GA4GH standards, and with the Regulatory and Ethics Working Group (REWG) to create a consent policy, a user agreement to enable data donation and functional access, and a set of requirements for becoming a Matchmaker service. They are currently working on security requirements for each of the databases and the matching algorithms are continually under development. A third database—Decipher—has since launched the API internally and a number of other matchmakers are actively working to do the same. The team is now working to link the Matchmakers together so that a query to one will also have the potential to return matches from the others.

There are two ways to participate in the MME: Databases can link to the Exchange by implementing the API, or a clinician with a single case can enter phenotype and/or genotype information into one of the existing Matchmakers, which will return matches from its own database as well as those to which it is linked through the Exchange. Rehm and others are working to post a decision guide on matchmakerexchange.org to help clinicians identify the most suitable database into which they should enter their case. For instance, Rehm said, if you have just a gene, you might want to enter into GeneMatcher, but if you have an entire VCF file, GEM.app might be the better choice.

In a pilot phase, the team implemented test cases into PhenomeCentral and GeneMatcher and successfully matched all those for which the case was present in both databases, demonstrating that the system works as intended. They then looked at 45 unsolved cases and identified 10 new matches. Two of these are thought to be potentially significant, two are still under consideration, and six were determined false positives (the genes but not the phenotypes were the same, emphasizing the need to optimize the matching algorithms).

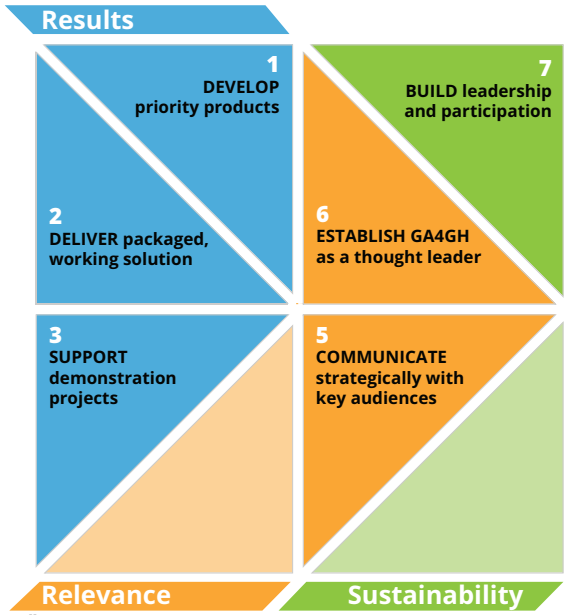
A special issue of *Human Mutation*, guest edited by Rehm, Ada Hamosh (Johns Hopkins Medicine), and Kym Boycott (Children's Hospital of Eastern Ontario), summarizes the work to date in 16 papers. These include an overview of the whole project, one paper on the API, two on the mathematics of the matching algorithms, several on the individual matchmaker systems, and three that document cases solved through matchmaking. The issue will be released in October 2015, in conjunction with the next ASHG meeting.

Going forward, the team will work to enable hypothesis-free matching, since clinicians often may have sequencing data but still no candidate causal genes. The team wants to aggregate the datasets to look for commonalities that might not be obvious with only an individual case in hand. Additionally, they hope to enable broader sharing of at least some of the data. For instance, Decipher has already proportioned 15,000 cases to be freely searched on. They are also working with ClinGen and the Genetic Alliance to incorporate patient initiated matchmaking, which will engender a unique set of technical as well as ethical complexities. While still in the planning stages, it is expected that a group of physicians will initially receive results from any patient queries to weed out false positives and to serve as Matchmaker stewards.

In a Q&A session, Bartha Knoppers of the REWG noted that MME uses a tiered access approach, in which consent is only required when variant data and other patient details are entered. If only the gene name and structured phenotype terms are entered into the system, the MME does not consider that identifiable, but rather the “practice of medicine” as it is carried out today.

Another audience member suggested that level two access be built into future standard patient consent forms. Rehm agreed, noting that she and others are working with the NIH and the REWG to develop a single page clinical consent form for patients to easily share their data to advance human health. The only way to capture the most data is to advance clinical data sharing, Rehm said, since the clinic, more than research, represents the majority of cases.

(3) Beacon Project and Data Sharing APIs



David Haussler (University of California at Santa Cruz) provided an update on the Beacon Project and other activities of the Data Working Group (DWG), which consists of more than 500 people across for- and nonprofit entities who are all collaborating to build a working interface and a common language for genomics data sharing. The group also held internal planning meetings and crosscutting meetings with the other working groups during the three-day event.

The vital information we need to decode our health is held in silos around the globe, Haussler said, but we can change that. A world of technology awaits us if we can break through some of the social and technical issues. The DWG is developing an "Internet of genetic information that will make accessing genetic data as easy as finding a restaurant on your iPhone," Haussler said.

The general Internet consists of agreed upon "names" for the objects we want to access (urls), protocols for how to access them (http), and a universal understanding for how to semantically interpret content (html). "For genomics, we just need to specialize this for the kind of information we are transmitting," Haussler explained. The DWG has developed a means for naming objects, which doesn't reveal anything about the object; methods for requesting data, which are used in all GA4GH tools and applications; and a schema for each type of data that denotes in precise mathematical terms how to interpret the content. Now we can reliably build apps that depend on that interpretation, Haussler said. Two such apps are the Beacon Project and the Reference Human Variation Map, and Haussler provided specific updates on each.

The **Beacon Project**, led by Mark Fiume (DNAStack), pulls together the simplest of all genetic queries and makes it a ubiquitous feature available on the Internet. First identified by Jim Ostell (National Institutes of Health) at the 2014 GA4GH Plenary meeting in London, Beacons ask trusted data stewards "do you have any genomes in your database with a particular variant at a particular position." The queried database responds with either "yes" or "no." Beacons present no technical barriers, but there is a social barrier to exchanging even the simplest, most atomic unit of genetic information, Haussler said. The advantages of the Beacon approach are an open API, which enables interoperability between systems; a federated model, which negates the need for a centralized database and instead relies on trusted data stewards; and the technically simple, sufficiently primitive query, which mitigates (but does not completely eliminate) risk of identification. To date, fifteen organizations have lit 155 Beacons across 252 datasets. "This is a very dramatic uptake of a new technology that's still in its early defining stages," Haussler said.

Beacon currently has two levels of access: open and controlled. Open is intended for the anonymous Internet user looking for any record of a variant anywhere in a database. Controlled

is for those interested in a whole genome, which contains identifiable, private information and thus requires a full legal contract. A newly proposed access level is registered access for those who want to dig deeper than allowed by open access but who don't need a whole genome. It would depend on a set of agreements and would require the user to provide credentials. The DWG sees this as an important intermediate level and is working with the Regulatory and Ethics Working Group to develop it.

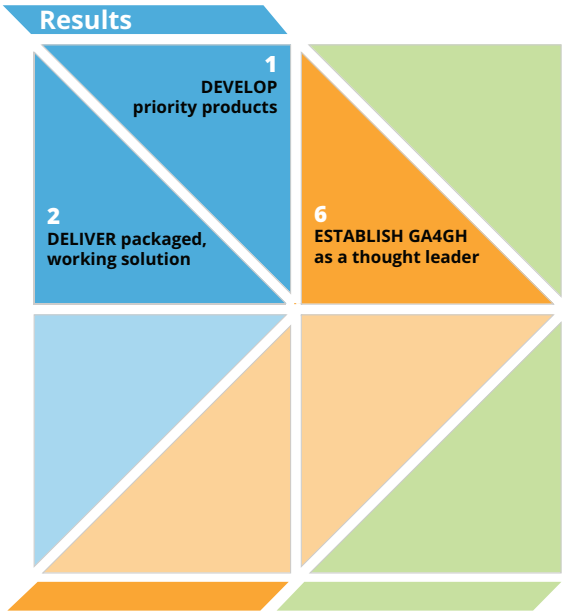
The **Reference Human Variation Map**, led by Gil McVean (University of Oxford) and Benedict Paten (University of California at Santa Cruz), is a comprehensive and unbiased representation of the human genome. The current reference genome is not representative of all human variation. "There is a huge amount of human diversity, and by leaving it out of the standard reference we have crippled ourselves in many ways," he said. A graph representation of all the variation in the human population gets around "what is essentially a giant Tower of Babel problem," he said. Various efforts to codify genetic variation and link it to phenotype each use different schemes to represent the same thing and none is definitively comprehensive. The Reference Variation Task Team is building a graph reference that will be a comprehensive "Rosetta Stone" for human genetic variation. It does not replace existing references but translates them into one comprehensive archival representation, Haussler explained.

A single dominant line on the graph denotes the current human reference genome while surrounding lines represent different individuals' genomes. "Think of it like a theme in music with beautiful variations," Haussler said. "There may be different themes throughout a region." At its core, the graph consists of individual bases of DNA. Whereas the reference genome may have a sequence of ACGGCC at a certain locus, a common variant among the human population may turn that into GAGGCC. An alternate path on the graph represents this alternative variant. Within the graph model, every base has a permanent identifier that won't change when we learn more about a locus. "How we identify a genetic variant should be durable over decades," Haussler said. "And it will be in this new structure."

These two examples offered a flavor of the many projects ongoing within the DWG, whose ultimate goal is to develop a clean, secure, Internet-based system of accessing and exchanging health information that is available to the doctor, the researcher, and the consumer alike. If we can achieve this, Haussler said, then the whole Internet will be a health learning system and no data will be left on the table.

A Q&A session focused on issues of using crowd-sourced rankings to qualify annotations to the current reference genome; the importance of healthy cohorts; the fact that most data currently left on the table is clinical data; and pseudonymizing clinical data in order to capture its richness. Finally, the distinction between Matchmaker Exchange (MME) and Beacon was addressed. The two are conceptually different, Haussler said. MME matches individuals across a multidimensional query of genetic and phenotypic attributes whereas Beacon queries only whether a database contains a genome with a specific variant. "You could think about it as a Google search," Haussler said, where Beacon is the initial high-level search and MME allows you to drill down further like clicking on a web page. Heidi Rehm (Partners Healthcare) noted that in most cases matching patients via MME requires different variants on the same gene whereas Beacon looks at the same variant. Beacon and MME are complementary systems, she said, and individual Matchmakers are working to launch their own Beacons.

(4) Policy: Consent and Privacy & Security



Graeme Laurie (University of Edinburgh) and **Susan Wallace (University of Leicester)** **presented** two policies developed by the Regulatory and Ethics Working Group (REWG) that are meant to guide practical implementation of the *Framework for Responsible Sharing of Genomic and Health-Related Data*, which was designed to serve as a high level starting point for deliberation and action. The policies aim to facilitate and operationalize the *Framework's* various principles.

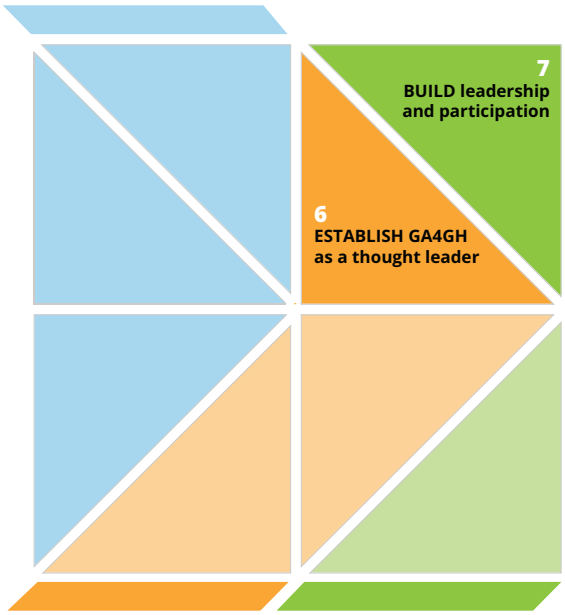
The **Privacy and Security Policy** demonstrates how an entity or individual involved in providing, storing, accessing, or managing data can and ought to promote privacy while also promoting science and sharing. It serves as a framework for mutual recognition and trust, providing immediate and common ground for interaction. The policy can also help entities assess their readiness to share data based on GA4GH objectives and aspirations and identify unmet needs which GA4GH can help them address, though it has no authority to enforce the policy's uptake. Privacy and security have different best practices, Laurie said, and while the fields overlap they are distinct. Privacy is a fundamental human value, enshrined in laws as a basis of rights and responsibilities. Security is a set of practical measures to manage risk. Users must determine which of the policy's measures are most relevant to their pursuits and identify clear lines of responsibility and accountability in the implementation of the various elements of the policy.

The **Consent Policy** provides guidance around whether existing consents allow forward data sharing and how to proceed if changes need to be made. It is intended to help in the design of prospective consents and does not override existing consent gained from a data donor. It is specifically geared toward international data sharing and offers practical measures based on a series of best practices that map to the 10 core elements of the *Framework*. The consent policy is based on five basic principles: That consent is an open, communicative, and continuing process; that there is an intention to share data across institutions, jurisdictions, and national borders with appropriate approvals; that plans for data sharing should be transparent, understandable, and accessible; that data donors have a right to withdraw participation or not participate at all with the understanding that it may not be possible to retrieve and/or destroy data once shared; and that data users and producers will abide by applicable regulations and ethical norms when seeking and conducting international data sharing.

A Q&A session addressed issues of encouraging international data sharing through robust consent policies, the dynamic state of privacy regulation in the European Union, and the distinction between data sharing and data discovery.

Emerging Projects

(1) eHealth



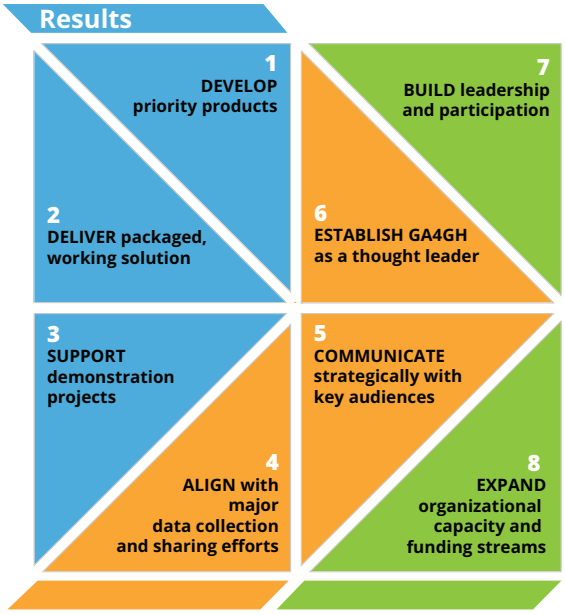
John Mattison (Kaiser Permanente) provided an update on activities of the eHealth Task Team, first inviting **Gil Alterovitz (Harvard University)** to present the eHealth Catalogue of Activities, an online listing of international genomic and clinical data sharing initiatives. A visual representation of the catalogue reveals a geographic distribution of activities that aligns well with GA4GH membership. Using co-authorship between these activities as a proxy for collaboration, the eHealth Task Team identified which organizations are working together and which are isolated and found a direct correlation between collaboration and shared governance. Increasing leadership among shared initiatives, Alterovitz proposed, may help motivate collaboration across currently isolated groups. The eHealth Task Team will continue its work on the Catalogue, expanding its functionality and publicizing it to the broader community, with the hope that it will foster increased collaboration by providing a foundation for dialogue between institutions with shared missions and activities.

Mattison discussed the unchecked growth of semantic representations, data models, ontologies, and APIs, and how to manage them for global “omics” collaborations. To constrain this growth, a byproduct of many siloed activities around the globe, Mattison proposed several recommendations, including the use of federated data enclaves with researcher credentialing protocols, “lighter” data representations, micro-consent and micro-credit to illuminate data hoarding, and data concierges to guide users of an organization’s shared dataset.

Additionally, he proposed a matrix analysis to guide vendor adoption and market penetration of the most valuable ontologies and semantic representations. Migration of data across different institutions and associated semantic environments results in unnecessary semantic degradation by virtue of multiple transformations, Mattison said. But as data sets become more federated, the community will naturally arrive at that smaller, constrained set of representations, which will also reduce semantic degradation. Finally, he posited that “ontology” is a one-word oxymoron, since there is no single right way of organizing all of the information in a given reference model.

A Q&A session addressed the best means for capturing data from clinical care systems, be it patient or vendor directed (Mattison said both are useful) as well as how to drive vendor uptake and market penetration of GA4GH standards, which Mattison said should be done through vendors’ clients.

(2) Clinical Cancer Genomes



Mark Lawler (Queens University Belfast) provided an update on activities of the Clinical Cancer Genomes Task Team, which aims to provide added value to and harmonize with ongoing projects and outputs in the clinical cancer genome space. The Task Team has submitted an opinion paper for publication, outlining the global cancer genomics landscape and highlighting GA4GH's activities and potential in this precision medicine domain. The paper argues that Big Data genomics will not efficiently advance until the international community overcomes a number of barriers, including shortsightedness, lack of interoperability, and reluctance to share and it outlines how the GA4GH vision can deliver for scientists, health care professionals, and, most importantly, patients. The Task Team is also deploying a survey to GA4GH members to identify ongoing efforts, highlight examples of best practice, and enable collaboration between data sharing initiatives in clinical cancer sequencing. Together, the survey and the opinion paper aim to delineate the importance and benefits of data sharing, discuss key challenges of merged data usability, identify best practices, develop solutions, and highlight the importance of implementation. The Clinical Cancer Genomes Task Team is also developing an Actionable Cancer Genome Initiative.

GA4GH has the opportunity to foster, nurture, and drive a global actionable cancer initiative, Lawler said, and he announced intentions to establish a cross-cutting cancer driver project, which will be a joint activity between the GA4GH Clinical and Data Working Groups. This joint activity will create an authoritative approach that defines clinical relevance of the somatic cancer genome. The group will also seek to establish the validity of actionable cancer panels, define optimal standards for technical practices, develop curation and annotation approaches to somatic variant “Big Data,” facilitate regulatory and reimbursement pathways, and establish flexible, clear principles to justify variant test selection, delivering clinical cancer diagnostic, prognostic or predictive tools that are useable, billable, payable, and sustainable.

A lively Q&A session addressed the best first steps in an area as complex as somatic cancer, overcoming the challenge of working with data that is currently fragmented, aligning with regulators to drive uptake, and leveraging crowdsourcing as a tool for integrating currently siloed data.

National Initiatives



Facilitated by **Mark Guyer (US Precision Medicine Initiative)**, the participants in this session were asked to introduce their national initiatives and answer three questions: (1) how are you leveraging or planning to leverage GA4GH in your national initiative? (2) how do you plan to share data or support data sharing? and (3) have you identified major barriers that GA4GH is currently not trying to address and/or are their challenges that are unique to your jurisdiction that GA4GH should be aware of?

Morris Swertz introduced **Genome of the Netherlands**, a project of the Dutch arm of the federated European Biobanking and Biomolecular Research Infrastructure (BBMRI), which seeks to harmonize and enrich existing biobanks around Europe in order to integrate and provide easy access to data. BBMRI-NL consists of 900,000 samples across 200 Dutch biobanks, representing about five percent of the total Dutch population. Four of the 12 broad-based projects were specifically relevant to genomics, including Genome of the Netherlands (GoNL), which was the first population-specific whole genome sequencing study in the world. GoNL led to the discovery of many new variants, knowledge which is being put to immediate clinical use. Two follow-up studies, Biobank Integrative OMICS Studies (BIOS) and the Society of Clinical Genetic Diagnostic Laboratories (VKGL), sought to integrate additional “omics” data into the GoNL database and to allow clinicians to readily query existing knowledge on a particular variant, respectively. The team is now focusing on ways to improve dissemination of the information captured in these projects, with a particular focus on technological challenges. They are eager to synergize with GA4GH in developing and implementing new APIs and connecting with those that are already emerging. ⇧

Tim Hubbard introduced **Genome England** and the **100,000 Genomes Project**, a project of the National Health Service focused on treating rare disease and cancer. In addition to generating new clinically beneficial treatments through whole genome sequencing and enabling future research, it is also meant to stimulate activity in the UK economy through genomics-related spinoffs. Eleven Genome Medical Centers across 75 hospitals have been established to collect consents, DNA samples, and phenotypic data when a patient presents with a disease that cannot be diagnosed by an existing test. Illumina performs sequencing on the sample and all data are fed into a centralized Data Centre for analysis by Commercial Interpretation Services, which deploy their own algorithms within the Data Centre as virtual machines. A clinical report feeds back to the NHS for verification. Genome England is a member of GA4GH and expects to implement its APIs within its infrastructure as they become broadly adopted as standards. Since the data are not consented to be redistributed outside of the Data Centre, the project will likely interface with GA4GH through the Beacon Project as well as Matchmaker Exchange at its lowest level of consent. Hubbard noted that GA4GH could provide more

guidance on implementing its tools and approaches within a health system, pointing to the Global Genomic Medicine Collaborative **(**G2MC) as another group focused on this issue. ⇧

Paul Lasko introduced the **CARE for RARE** (formerly the Canadian FORGE Project) as well as the **International Rare Disease Research Consortium (IRDiRC).** CARE for RARE aims to provide diagnoses and new therapeutics for pediatric rare disease and to work with patient organizations and other stakeholders to bring genomics into Canadian healthcare. It has led to the identification of causative mutations for 146 rare diseases, half of which were not previously linked to any rare disease, as well as diagnoses for 500 patients. The CARE for RARE database is called PhenomeCentral and is an integral part of the Matchmaker Exchange. While the project leads to clinical diagnoses for patients, it is a research project and thus data are consented for international sharing. It is fully supportive of a federated model of data sharing.

IRDiRC was launched in 2011 to facilitate international collaboration and data sharing among rare disease researchers. It has two goals: to catalyze the development of diagnostics for most rare diseases and to catalog 200 new therapies by 2020. To date, it has linked more than 3,000 genes to rare disease and identified 144 new rare disease therapies. Matchmaker Exchange started within IRDiRC, but took on a higher level of activity with the help of GA4GH. IRDiRC is also involved in GA4GH's Machine Readable Consent Task Team and believes continuing collaborations between the two groups will be beneficial. Toward that end, Canada launched a funding opportunity in 2014 called *Sharing Big Data for Health Care Innovation - Advancing Objectives of the Global Alliance for Genomics and Health*, aimed at specifically strengthening the Canadian contribution to GA4GH. An announcement is forthcoming on its outcome. ⇧

Kathryn North presented two Australian initiatives to integrate genomics into everyday healthcare. First, the **Melbourne Genomics Health Alliance** (MGHA) is a collaboration of 10 research and healthcare organizations, which compared the clinical utility and cost effectiveness of whole exome sequencing (WES) compared to standard clinical practice as a first tier assay for germline and somatic conditions. MGHA members have developed a series of shared tools, including ethics and consent protocols, a clinical bioinformatics pipeline, and a clinical genomics data repository. In a pilot study, MGHA showed that WES resulted in a 54% diagnostic rate for childhood syndromes, compared to 20% for standard practice. The approach is now being expanded across the entire state of Victoria. Second, the **Australian Genomic Health Alliance (AGHA)** is a collaborative effort of 41 academic, diagnostic, and genetic services organizations across Australia, which has applied for funding from the National Health and Medical Research Council to develop a national approach to genomic medicine. The proposed program aims to develop a federated genomic data repository that will link phenotypic and genomic data with electronic health records, using the data sharing approaches promoted by the GA4GH. The group also plans to use the GA4GH regulatory and ethical standards for consenting patient data. The biggest challenge facing Australia in this effort is identifying ways to successfully translate the state-based approach into a national, federated system. ⇧

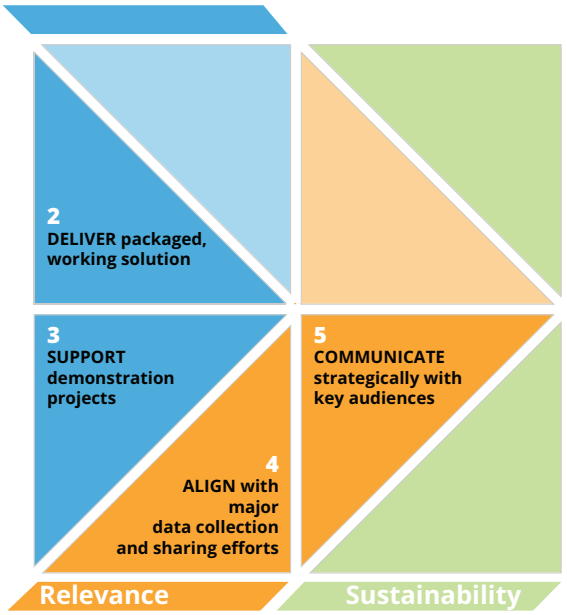
Iscia Lopes-Cendes introduced several activities in Latin America. **The Latin American Collaborative Study on Congenital Malformations** is a clinical and epidemiological investigation launched in 1967 across Latin American hospitals, which has recently begun incorporating genomic information into its database. **The Study Group on Hereditary Tumors** is working with two other international genomics initiatives: the Collaborative Group of the Americas on Inherited Colorectal Cancer and the International Society for Gastrointestinal

Hereditary Tumors (InSiGHT). Two Brazilian projects that could benefit from collaboration with GA4GH are **The National Institute of Population Medical Genetics (iNaGeMP)**, which was established in 2009 to study rare diseases via clinical, epidemiological, family history, and genomic data, and **The Brazilian Epidemiological and Biobank Stroke Study**, a two-year prospective population-based study of stroke across 5 cities using genomic and clinical imaging data. It expects to collect 2,400 patient samples and 5,000 control samples per year. Samples collected in a pilot study are now undergoing genomic analysis at the University of Campinas (UNICAMP) under Lopes-Cendes' direction. Next, Lopes-Cendes introduced the **School of Human and Medical Genetics**, an educational initiative established in 2005 with the Latin American Network of Human Genetics, which has enrolled hundreds of students across every Latin America country as well as several Caribbean nations, with particularly strong representation from Brazil, Mexico, Argentina, and Colombia. Finally, Lopes-Cendes highlighted a **collaborative project with GA4GH** to address the fact that the Brazilian population is not well represented in international genomic databases. The country's population has a complex genomic background for which other Latin American countries, including those with mixed populations such as Mexico, cannot be used as proxies. Lopes-Cendes and her team are producing a database of molecular profiles based on SNP arrays and whole exome sequencing. They are currently preparing an environment to publicly share the data, are integrating them into LOVD, and will soon light a Beacon on top of them. ⇧

Bin Tean Teh introduced various **precision medicine activities in Singapore**, which span basic science, translational research, and clinical applications in medicine. Three years ago, Singapore launched POLARIS, a pilot genomic medicine program to identify barriers to the clinical implementation of genomics. This proof-of-concept multi-institutional collaboration to translate Singapore research to improve health has established an infrastructure of CAP-certified genomics laboratories; Ministry of Health-compliant software for analyzing genomic data and patient reporting, compatible with local EMR systems; community standards for regulatory and ethical applications of genomics; and a recently launched Precision Medicine Institute. Other recent precision medicine efforts include genetic services to test for the *TGFB1* mutation for genetic eye disease and multi-gene assays for sudden cardiac death. Over the next decade, the country will build a medical database consisting of whole genome sequencing data from 5,000 healthy volunteers integrated with serum metabolites, immunophenotyping data, cardiac imaging data, and EMR data. Finally, a centralized facility for clinical genomics will perform rapid data analysis and visualization in concert with a clinical follow up mechanism. Singapore's precision medicine leaders are eager to work with GA4GH to scale their efforts. ⇧

Mark Guyer closed the session with an introduction to the US **Precision Medicine Initiative** (PMI) and the **Big Data 2 Knowledge** (BD2K) program. BD2K was established to advance basic and translational science by facilitating and enhancing the sharing of research-generated data. It resulted in the creation of an open digital ecosystem to accelerate biomedical research and its application to human health. The Precision Medicine Initiative is a proposed program with three main components: (1) a near-term focus on cancer, (2) a long-term aim to generalize to the full range of human disease, and (3) to advance the nation's regulatory framework in order to implement precision medicine in practice. The second component will require the creation of a research cohort of at least 1 million participants that is representative of the American population. This will likely involve several existing cohorts. The NIH fully supports GA4GH data sharing goals and approaches and intends for the Precision Medicine Initiative to be consistent with its principles and objectives. ⇧

Big Data Challenges and Solutions



Dixie Baker (Martin, Blanck and Associates) facilitated the session, beginning with an overview of the challenges relating to Big Data. The overarching challenge, she said, is the lack of computational infrastructure needed to securely generate, maintain, transfer, analyze, and visualize large-scale data sets, and to integrate the various types of "omics" data with each other and with clinical data. Virtualizing computational resources, she said, may be the only approach for achieving the kind of scalability and elasticity needed for DNA sequencing and analysis. But virtualization raises new security and privacy challenges associated with the blurring of physical boundaries, "Big Data" technologies that render everything "discoverable," and compliance with national laws and institutional policies. "The only thing we can be sure of is that the challenges will intensify and build over time," Baker said. The session's five speakers shared some of the solutions their organizations are taking to overcome those challenges. ⇧

Jun Wang (Beijing Genomics Institute) discussed BGI's efforts to realize its Million Genomes Project, announced two years ago to build a strong national "multi-omics" database in three to five years. CompleteGenomics (a BGI company) currently has capacity to sequence 10,000 genomes per year using its newly released Revolocity platform, which has a current upper limit of 30,000 genomes per year but is being expanded to reach 300,000 and has an ultimate goal of 1 million genomes per year. Achieving that capacity will require significant investment, for which BGI is exploring unique business models. The company anticipates that if More's law continues to apply, the cost of sequencing a genome will drop to $1 genome by 2019. BGI hopes to work together with GA4GH to develop an open research platform on top of the Revolocity system and aims to create a network of multi-omics data akin to the Internet. Wang said that cultural differences mean different issues of data privacy and ownership in China. As a pilot project, the company has collected a comprehensive set of multi-omics data on the Chinese wheat crop, foxtail, and is analyzing it in a controlled, machine learning environment with promising results. ⇧

Angel Pizarro (Amazon) presented Amazon Web Services (AWS), a global provider of on-demand, pay-by-the-hour, commitment-free "cloud" computing infrastructure. The ability to easily share data and applications across institutional boundaries and the ability to publish preconfigured resources to the community at large are key to collaboration. Before the cloud, population-scale science was limited to institutions with the most computational (and financial) resources. With AWS, users provision only the size of compute needed for a given project. Data can be put into the cloud and access can be requested and granted on a temporary basis. AWS operates 11 regions worldwide; once data are persisted in a region, they remain within that

region. This enables AWS to comply with jurisdictional laws restricting the physical location of data. As an example, the National Database for Autism Research (NDAR) is storing its phenotype data on the AWS cloud. Credentialed collaborators that want to use NDAR data can bootstrap it into their own AWS environment, within which they use their own analytical processes or leverage AWS's open-source ecosystem of preconfigured genome analysis pipelines. For data security, AWS uses a shared responsibility model similar to that described in the GA4GH *Security Technology Infrastructure*. AWS uses a shared responsibility model for data and service security. It does not use customer data, but provides a suite of service offerings for the user to control all of its own risk mitigation strategies. At the end, a customer can have a third party audit their solution for compliance with industry standards. ⇧

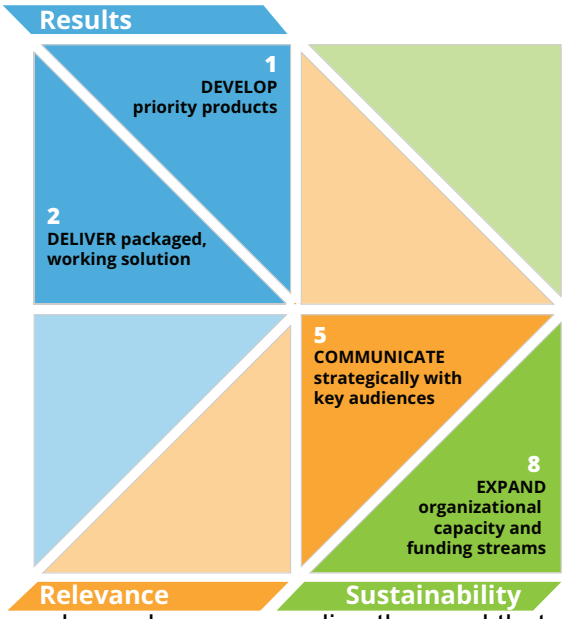
Mathew Pletcher (Autism Speaks) presented MSSNG, an initiative to generate an open access database of 10,000 whole genomes and associated phenotypes from families with autism. The general goal of MSSNG is to accelerate understanding of the genetic underpinnings of autism and its specific goal is to introduce more categorical granularity into the diagnosis. As a patient advocacy group, Autism Speaks intends to provide a community portal to connect donor families to their data and, ultimately, to other families with a shared genetic subtype. MSSNG aims to interface with a diverse set of stakeholders, including academic researchers, large and midsize pharmaceutical and clinical centers, as well as non-traditional users such as entrepreneurs, the diagnostic industry, educators, parents, data donors, and non-academic researchers. MSSNG has so far generated more than 3,000 whole genome sequences and is on track to reach its goal of 10,000 by the beginning of 2016. The data will be freely available to credentialed researchers in the Google cloud environment. A web portal will be launched in July 2015 to support simple queries while the Google platform allows command line access for more in depth investigations of the dataset. Autism Speaks is taking on all the costs associated with data hosting and analysis. It is working with the Public Population Project in Genomics and Society (P3G) to establish an access policy to reduce obstacles to data sharing while still protecting patient privacy and honoring consent. Current MSSNG data have not been consented for broad sharing, so Autism Speaks has begun a process of re-consent and going forward will use a universal consent, currently in development. A Q&A session touched on the expected number of data access requests, the breadth of associated clinical phenotype data, and issues of credentialing researchers in a tiered consent model. ⇧

Richard Gibbs (Baylor College of Medicine) provided an update on work performed on data from the CHARGE project, in a collaboration between BCM, DNANexus, and Amazon Web Services (AWS) to perform the largest ever cloud-based genomic analysis. The group has expanded its scope to also include sequencing data from Alzheimer's patients from the Alzheimer's Disease Sequencing Project. Since last year, BCM has also grown its clinical datasets, thanks to the opening of a routine diagnostic lab that brought in more than 6,000 new cases. The group has made its data available to researchers through a series of portals based on selective access. Gibbs said that much of the success of the collaboration can be attributed to the technical solutions developed by the GA4GH community and requested a continued focus on refining the model. He also requested better solutions to integrating clinical data into the research arena, pointing to Matchmaker Exchange as a good, but as yet not readily scalable, option. He also noted that redundancy between the Matchmaker Exchange and the Beacon project needs to be addressed. Gibbs identified individual privacy, and HIPAA compliance, as the CHARGE project's biggest challenge. He noted that the collaborations would be well served by an easier, more straightforward approach to data access and consent. The consortium has

been working on this in a cohort of pancreatic cancer patients, for whom data are now freely available for querying via the GA4GH API on the DNANexus website. A Q&A session addressed the favorability of federation versus a centralized hub of data, new analytical methods for interpreting the non-coding parts of the genome, protection of patient privacy, and the process of re-consenting patients after testing. ⇧

Arcadi Navarro (Centre for Genomic Regulation) presented the European Genome Phenome Archive (EGA), a federated collection of 1,300 datasets across Europe, spanning 73 studies from 308 data providers, including International Cancer Genome Consortium (ICGC), IRDiRC, UK10K, and the Wellcome Trust Sanger Institute. EGA includes data that may identify individuals, so the 1.7 Perabytes of information it contains need to be kept secure, despite heterogeneous patient consents. Access to EGA data is controlled in accordance with Data Use Agreements. Furthermore, the data must be able to be redistributed across a variety of databases, each housed in different countries with varying regulatory contexts. The EGA, a project of the European Bioninformatics Institute (EMBL-EBI) and the Centre for Genomic Regulation (CRG), faces similar issues to those being addressed by GA4GH, including how to successfully federate. Luckily, EGA isn't the first group to face this problem, Navarro said, and then he displayed a picture of Tony Stark in his Ironman suit. Taking a cue from the great, if fictional, inventor, the EGA team set out to create a series of modular parts that interact with one another in a standardized way. Each can work independently, but can also be wrapped into a standard interface, Navarro said. The external services are the same, but the internal organization of "EGA 2.0" is based on micro-services modules. GA4GH can help the group tackle federation, Navarro said, particularly in the areas of secure computing clouds and federated authentication systems. In the second half of his presentation, Navarro addressed the importance of robust metadata for effective redistribution. Sharing relies on good metadata, he said, but there is currently no standardized or proscriptive approach to collecting that information. An EGA task force has been established to address this issue, and will likely engage Beacon and Matchmaker Exchange to do so. ⇧

Working Group Panel Discussion



Ancillary meetings took place on June 9th, June 11th, and the evening of June 10th, which included several crosscutting sessions as well as internal meetings of each of the four Working Groups, Matchmaker Exchange, Beacon Project, and several of the Task Teams (see appendix for details). Martin Bobrow facilitated a closing session in the afternoon of June 9th, during which Working Group Chairs provided summaries of discussions thus far followed by a panel discussion with Paul Flicek and Dixie Baker (SWG Co-Chairs), Kathryn North (CWG Co-Chair), David Haussler (DWG Co-Chair), Kazuto Kato (REWG Co-Chair), and Ewan Birney (plenary keynote speaker).

Bobrow noted that the landscape for data sharing has changed considerably over the past year. This is due in large part, he said, to GA4GH members who are spreading the word that genomic and clinical data will not improve health unless we share and analyze them in groups rather than "in our own little siloes."

A summary of the ensuing discussion follows.

- There has been much activity around cancer. What are we doing in the rest of medicine?
  - The goal is to make the tools and the general principles of ethics, consent, and security applicable across a range of disorders.
  - We are beginning to engage with the infectious disease community to identify ways to build on existing data sharing efforts using our principles and tools.
  - We could begin to catalyze projects in other disease areas, perhaps in a single area within neuroscience, such as dementia or epilepsy.

- For each panel member, what is one thing that would help us be more global?
  - By expanding into the infectious disease space.
  - By reaching out to non-English speaking countries in the developed world, perhaps through Task Teams proactively seeking input from those regions.
  - By producing useful tools that become the standards for clinical practice.

- Is it time to present our APIs and other tools at clinical professional meetings and engage clinicians in the design?
  - We want a few hearty physicians to join us in the building process, but we don't yet have a complete API that a physician could immediately upload and use to improve patient care.

  - A number of upcoming papers, including a whole edition of *Human Mutation*, will highlight activity in the cancer space.
  - If we choose the right products to push early that have an impact, then we don't have to a have a fully-tooled system for everybody.

- How would you pitch a medical center overseeing a large clinical research project to encourage them to apply the GA4GH data sharing framework to their efforts?
  - You are intelligent people: You know where the future is going and that is toward a lot more molecular measurement of patients. And you know that the clinicians in your hospital will have to make informed decisions using those measurements. The power to make better decisions comes from a clinical research base pooled worldwide. So to invest in the GA4GH is to do two things: First, it is to help your clinicians get their heads around the future of medicine and, second, it is to be altruistic, to contribute to that pool from which you'll benefit in the future.
  - Heads of hospitals and governments also need to hear the economic argument before they will begin to change policy.
  - In the act of sharing data, we often discover inconsistencies and lack of standards inherent in our data. Realizing where data structures and terminologies differ allows harmonization that benefits both the submitter as well as the community at large.

- Disease advocacy groups need to talk more with patients about the value of data sharing and to incorporate GA4GH into those discussions.
  - Patients and public enrollment are incredibly important. It is not enough to talk amongst ourselves, we also need to look outwards.

- Many physicians are driven by two things: (1) compliance with regulations and professional licensing and (2) liability concerns. We need to promote the fact that data sharing is both legal and ethical and part of the responsibility of providing the best care.

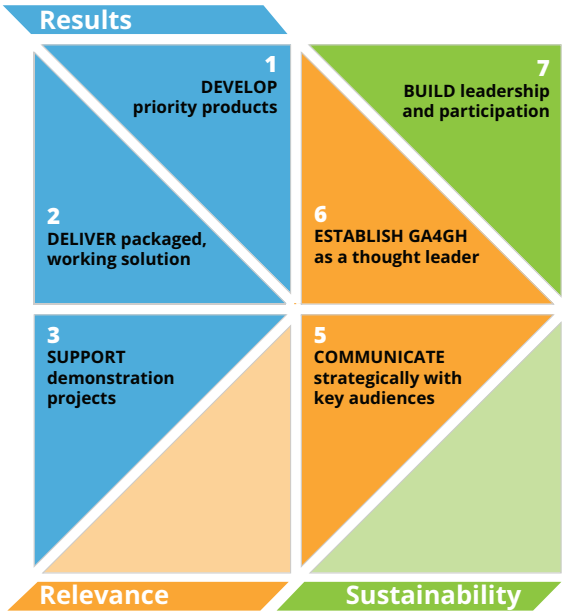
- Can we leverage the pharmaceutical industry and its budget for genomic research?
  - The pharmaceutical industry is going through a complicated change: Their model does not generate enough drugs that get licensed at the rate needed to support them. But the industry is beginning to realize the benefit of studying genetic and other molecular measurements. We should all work with our local biotech and pharmaceutical industries to make public/private partnerships that leverage patient cohorts that are in line with the GA4GH framework.

- How can we think about not only sharing data but also sharing the outputs of data in order to engage regions or countries with less capacity?
  - By making things as easy and accessible as possible. Large-scale sequencing data should be immediately converted into smaller, more useable forms.

- 23andMe has a dataset of roughly 1 million individuals, many of whom have agreed to share their results for research purposes. We should look at how commercial and entrepreneurial genomics efforts are tackling this problem by cutting through middlemen and going straight to where the data are. One way to change behavior might be through leveraging patient power.

Action Items

David Altshuler led a closing session on June 10th to discuss action items that emerged during the first two days of meetings, inviting audience members to identify the primary challenges that the GA4GH is not currently taking on but should. Altshuler used the Road Map that emerged after the last Plenary to guide the discussion.



1. Develop priority products
   - Needed products:
     - Ethical training and a code of conduct on privacy, ontology, and ethics for non-clinical bioinformatics researchers working with clinical genomic data.
     - Thought processes, structures, APIs, and databases developed for the clinic as have been developed for research, and more interaction between the two.
     - Strategies and tools to standardize data curation efforts.
   - Products in development:
     - Accountability Task Team to address issues of researcher authenticity and validation in the context of increasing trust rather than a culture of caution.
     - A common lexicon to enable data sharing between entities that use disparate means of patient data anonymization, whether through de-identification, pseudonymization, encryption, or other approaches.

2. Deliver packaged, working solutions
   - Engineers are needed to deliver easy-to-use software.
   - A downloadable start-up-kit which provides clear guidelines for helping any group take GA4GH tools and approaches from principles to implementation
   - To encourage organizations to mention the *Framework* as the philosophical position guiding new consent forms and to build new consent forms upon the generic GA4GH template. We have seen two such consents emerge thus far.
   - Demonstrated uptake is one good way for measuring success. We need to continue thinking about how to make our products usable, for if they are not used they will not have impact.

3. Support demonstration projects
   - We need to help scale successful existing projects and connect them to other similar groups, as was done for ClinGen via Matchmaker Exchange.
   - A catalog of examples to immediately demonstrate clinical value of medical genetics efforts and to incite new ideas, applications, and development.

4. Align with major data collection and sharing efforts
   - No comments

5. Communicate strategically with key audiences
   - Engagement with patients and healthy volunteers, not just advocacy groups, perhaps leveraging the rapidly growing DIY community.
   - More representation of vulnerable groups. We have established a Paediatric Task Team (launched June 10th) and will soon establish another focused on aging and cognitive impairment.
   - More communication and engagement with everyday physicians—not just clinical geneticists, but cardiologists, gastroenterologists, and others who are the lynchpins between patients, payers, and hospitals.
     - To help clinicians realize the value of what we're doing, we need to convince them through demonstration. When the BRCA Challenge and Matchmaker Exchange lead to diagnoses, society, physicians, and patients will notice.
   - Proposal: A joint project between advocacy organizations within GA4GH membership (i.e., IRDiRC, Genetic Alliance, Global Genes, and others) to drive patient engagement.
   - To take what we're learning in academic medicine and implement it in the rest of the healthcare system, which is where the majority of care is delivered.
     - We have Quest Diagnostics and LabCorp, but we need more patient diagnostic companies. We also need to engage with organizations such as Flatiron Health, which process clinical transactions.

6. Establish GA4GH as a thought leader
   - GA4GH could act as standard-setting body to establish a much-needed robust nomenclature for joining variants across knowledge sources.
   - GA4GH should serve as the international face of the genomics community and act as a lobbying organization to spur change among government funding agencies and develop incentives that support a share-first, publish-later model.
     - Our best chance for changing the world's incentive structures or physicians' belief systems is through demonstration. When we show we can make discoveries and improve health, then we will see more widespread sharing.

7. Build leadership and participation
   - After first defining our message, we could establish a communication and engagement Working Group to provide platforms for communicating the value demonstrated by flagship projects and driving engagement from other communities.

8. Expand organizational capacity and funding streams
   - No comments

Appendix

Crosscutting Sessions **2**
Cancer Projects (CWG, DWG). 2
CWG Morning Session Feedback 2
DWG Morning Session Feedback. 2
eHealth + Consent Task Team (CWG, REWG) 3
References/Graph and Benchmarking (DWG, REWG) 4
Provenance/Ontologies/BRCA Evidence Gathering Group (DWG, CWG) 5
Security Working Group, Internal Meetings **6**
Regulatory and Ethics Working Group, Internal Meetings **12**
Clinical Working Group, Internal Meetings **15**
Data Working Group, Internal Meetings **17**
Demonstration Project Meetings **18**
Matchmaker Exchange Meeting 18
BRCA Challenge Meeting 18
Beacon Use Cases 18
Beacon Meeting & Hackathon 18
Beacon and Security Working Group (DWG, SWG) 19
MME + Beacon 19
Action Items, Detailed Discussion **20**

Crosscutting Sessions

Cancer Projects (CWG, DWG): *A crosscutting discussion between DWG and CWG members to identify cancer use cases as appropriate models of tools developed by DWG.*

- The session included presentations and discussion from 15 groups, including the Genotype 2 Phenotype Task Team of the GA4GH DWG, the Treehouse Childhood Cancer project, GENIE, the Clinical Cancer Genome Task Team of the GA4GH CWG, Cancer Core, the Pacific Pediatric Neuro-Oncology Consortium, BRCA Challenge, the Phenotype Ontology Task Team, and the GA4GH Security Working Group.

CWG Morning Session Feedback: *A discussion of the morning's Cancer Projects and specific participation by the CWG, to establish goals, deliverables, member participation, and timelines*

- **Summary:**
  - Mapping initiatives are critical to determining what we do next.
  - Cancer is a complex area, so we need to clearly identify who does what and to do things in achievable "bite size chunks."
  - Working Groups need to develop data sharing start-up kits (for ethics, codes of conduct, securing data, clinical phenotypes, etc.) in order to enable project leaders around the globe to work together to create interoperable solutions for clinical cancer.
  - Our efforts will only have long-term value if they become a daily part of routine clinical care. GA4GH-backed international standards will help us influence policy in our individual countries and will help make the tools billable when used for clinical testing
- **Outcome:**
  - By the next Plenary meeting, a planning group led by Mark Lawler (CWG) and Adam Margolin (DWG) will identify and launch a cross-cutting global actionable cancer initiative that brings together the domain expertise and medical goals of the CWG with the data models and APIs developed across the DWG and that interfaces with many of the existing groups described below. It will also bring in expertise of SWG and REWG and may focus on AML to start.

DWG Morning Session Feedback: *A summary of the morning's discussions and other recent developments across the Data Working Group.*

- **Summary:**
  - Some of the biggest high profile projects in the cancer space are interested in sharing data in a global context and could benefit from working the DWG and collaboratively implementing the DWG APIs.
    - In particular:
      - Actionable Cancer Genome Initiative
        - Determining the actionable genes for cancer treatment and best technical practices
        - DWG to help develop a database of curated mutations that are well annotated
        - Similar to ClinGen structurally, but for somatic cancer mutations
        - Bringing together ongoing siloed efforts
      - GENIE

        - Consortium of 8 institutions in the US, Canada, and Europe
        - The community needs a federated model of clinical data sharing
        - Institutes have invested in siloed clinical sequencing and want to share their data
        - All 8 institutes have signed a data use agreement
        - SAGE bionetworks in Seattle is aggregating mutation calls made on thousands of patients
        - GENIE is working through data access policies and retrieval of clinical follow-up data
        - Clinical follow-up associations will be shared
        - Some questions on when follow up would be widely available
        - Other Project DWG can be involved
    - Other projects to work with DWG:
        - Cancer Core Europe
        - Prostate Cancer
        - BRCA
        - Next phase of ICGC
        - Athena
        - Orion (out of Florida Moffatt cancer center)
  - DNA Data → RNA data → Variant Calling → Variant Annotation → Decision support with G2P APIs and potentially overlay analytics for predictive modeling.
- **Outcome:**
  - DWG will make a reference implementation to demonstrate APIs for these various projects are all working together with a real world example.

eHealth + Consent Task Team (CWG, REWG): An open, cross-cutting workshop to address issues of Pedigree Consent in eHealth.

- Select an RDF-compliant approach to family history representation (GA4GH should attempt to identify and foster adoption of an existing initiative, with FHIR-based HL7 work a good candidate) Corollary: GA4GH should not try to craft this standard autonomously.
- Debate about centralizing FHx and Pedigree information vs. distributing it within individual health records. Some consensus that localization within health record would better support both human and machine readable access for supporting appropriate care.
- Understand opportunity for ingestion of FHx data currently residing in large commercial databases with direct to consumer marketing, e.g. Genealogy.com, Genie.com. If we import these data we prefer tight-binding provenance metadata, compliant with open standards, business model issues, and ethical considerations.
- Create defined test case exemplars to test full cycle of acquisition, validation and then application of pedigree and family history data to index case. Must carefully consider ethics and consent associated with propagation of discovery to family members, vis-à-vis ethics of right to NOT know, and defined consent and right to disclose to patient. Open Notes technologies force the convergence of ethical/consent/disclosure policies and practices.

- Provenance metadata is critical to understanding and managing data integrity throughout all data lifecycles.
- We must catalogue and address international variations in policy and values as they impact capture, use, and disclosure of FHx and pedigree information for both clinical and research purposes.
- We must support revocability of both pedigree data and previously exercised consent options.
- We must clearly articulate duties and life cycles for both clinical and research uses.
- We must identify three components to the FHx solution set:
    - A core data set for family history: source 2007 per Grant Wood
    - A functional and efficient UX
    - Semantic interoperability of primary data set and metadata, including provenance
- Approach EHR vendors NOW with our intent to develop consensus on a single model and solicit their input and direct participation in the process. We have a unique opportunity to capture interoperability BEFORE disparate implementations emerge in different EHR vendors.
- As these data are available both for human readability and machine readability for decision support, and as the corpus of actionable information expands, we will witness a dramatic expansion in the scope of discussion between all healthcare practitioners and their patients. This shift from centralized expert-based discussion within genetic counseling professionals to the larger healthcare team will incur a profound need to educate ALL healthcare providers who will soon engage in these discussions, which were previously isolated to genetic counselors. This significant educational burden (ethics/consent/CDS/family history/pedigree) will increasingly traverse the entire care delivery provider population.
- We must engage the legal community to advise us of the impacts of case law on the set of use cases and test cases.

References/Graph and Benchmarking (DWG, REWG): A joint meeting of the Reference Variation and Benchmarking teams to present progress in graph genomes in both the Human Genome Variation Map Project and in the graph based variant comparison tool.

- Benedict Paten presented on a graphs pilot, where DWG researchers are creating a reference format incorporating variation and comparing construction methods. The focus has been pared down to the driver project of interesting regions of genomes.
- A contest has been created and the first round was taken in on June 1st. The results will be published by September and the goal is to scale graphs to the whole genome by end of 2015. The graph reference server code is available on GitHub and supports endpoints,
- The Reference Variation Task Team is attempting to create an API for sharing variant data, but the main API needs to have more adoption over the next six months. It is likely that large-scale community buy-in will be necessary if this is to replace traditional genome modeling.
- If the reference implementation were put into commercial clouds, there would be room to use it for promotional purposes and for outside organizations to write interesting scripts and digitalizations. Researchers also want to have people sign up and commit to using the reference implementation of the API in the next few months.
  Any development process of this scale needs to be handled well or it risks creating a

large, complex (unimplementable) schema. There was a proposal to use the GA4GH GitHub repository, taking "forks" from the master that would remain under the control of each task team. The development of these topic branches will be made quicker by not applying the normal +1 rules.
There will soon be a fully sustainable compliance suite to facilitate this pace of progress, together with full documentation. It is clear that any proposed change should be accompanied by proposed implementation, as only with implementation can the consequences of the change be understood.

- There is to be a general push towards implementation driven software applications, which creates a need for a clearer management of API releases, establishing a proper cadence. This will allow people to build real applications based on that API, developed through GitHub discussions.
  Kevin Jacobs and Peter Krusche from the Benchmarking group shared graph based variant comparison tool. The next step is to make compatible with RefVar API, working to reconcile different tools and to set up accounting methods for false positives and false negatives

Pro*venance/Ontologies/BRCA Evidence Gathering Group (DWG, CWG):* A crosscutting session of software developers and clinicians to consider ways establish clear ontological terms and combine these with data provenance information to present in a clear, intuitive format that is most useful to the medical establishment.

Attendees of this meeting discussed:

- What ontologies are and how to use them in G2P association
- How to think about **provenance** for evidence and how to build algorithms for the different types of evidence and sources
- **Attribution** falls along evidence change
- A **new task team**, involving all working groups, was proposed to develop strategies for attribution across GA4GH and to incentivize contribution of data and also give credit to software development activities
- How to represent **ontologies in AVRO** schema
  - Some strategies to manage the proliferation of ontologies
  - Proposed group of roving ontology experts to attend Task Team calls

Security Working Group, Internal Meetings

Dixie Baker welcomed, and presented an overview of the GA4GH Security Technology Infrastructure document, emphasizing the value of Trust to the success of GA4GH. Cooperating to build systems, processes and workflows that both individuals and establishments can have confidence in.

Software Security Task Team update

Knox Carey presented an overview of the ways in which the Software Security Task Team is providing practical guidance for building software applications capable of operating in compliance with the GA4GH Security Technology Infrastructure. The goal is to help software implementers by providing specific recommendations, profiles, and reference implementations. Knox highlighted specific security support his team is providing to current GA4GH stakeholder projects, such as Beacon and Matchmaker, to assure that the team's work is maximally relevant to GA4GH.

During the discussion, the need for standard metrics for characterizing the level of assurance associated with different identity proofing methods was highlighted.

Data Access Agreement

- Tendency to develop into an arms race amongst lawyers
- Some are almost impossible for users to comply
- GA4GH should aim for something reasonable
- Crisp, short, containing the legally relevant terms
- Look to existing federated bodies for examples
- Describe what is needed for access at different levels and how it will work
- Consequences of inappropriate behavior
- Audited (record of actions)

Cloud Security Task Team Update

James Williams presented an update of the work of the Cloud Security Task Team, whose objective is to provide insight and guidance for purchasers, providers, and users of cloud services

James presented the results to date of a stakeholder survey (those who use or plan to use cloud services) -- 6 respondents to date (5 of them Application Service Providers), others to follow.

Security oversight including:

- Legal regulations
- Standards to work to
- Maintain audit logs
- Dissemination of information regarding attacks

Genomic Privacy

Jean Louis Raisaro presented an overview of current academic research relating to genome privacy and security.

- Unprecedented concerns, both real and media based
- Referenced 2013 Paper "Identifying personal genomes by Surname inference"
- See also Genomeprivacy.org (meeting San Jose, May 2015)

GA4GH Need for Security Oversight and Incident Response

Steven Newhouse led the group in a discussion about the importance of, need for, and potential methods for providing the GA4GH with security oversight and incident response (slides)

- Need to launch a new task team to prepare for (inevitable) breach of security
- Knowledge gained through previous work with European Grid Infrastructure (EGI) Federated network -- Security Collaboration among Infrastructures
- Developed a mature security model, including
- Policy Group
- Lead Security officer
- Risk Advice Team (RAT)
- 24/7 coverage, support, resources, help
- Best practise (encourage / expect site admin to work within procedures)
- Establish a chain to the cause of the incident
- Ultimate - ban an individual
- Use cases
    - Data flows A to B ⇒ incident ⇒ report back
    - Sharing incident-response reports
- GA4GH approach may be more complicated
- decentralised, collaborative (across international borders)
- What resources are shared / common
- How best to contact individuals at short notice?
- Traceability (identify cause)
- Audit?

Volunteers to participate in new task team: Paul Flicek, Steven Newhouse, (Co-Chairs), Ian Poynter, Caroline Rivett.

DB: The REWG is launching a new TT on Accountability. (A very early version of the Infrastructure Document called on each Institute to have two people available in case of a crisis, this was dropped at the time but the point remains that Infrastructure needs a point of contact.)

In sharing of data:

- What are the risks?
- What level of protection is required?
- How the hardware / software operate
- How to identify if a site / node / machine is being attacked
- Detection / reporting mechanism
- Disseminate information of an breach to others

If a user has acquired data and is later attacked, there is currently no duty to inform the original data owner. Should there be?
Given the volume of data, easier to take the analysis tool to the data and only pass back the results.

Major Outcomes

- Agreement on need for metrics for identity proofing levels of assurance
- Agreement on need to launch a task team to develop recommendations on security oversight and incidence response

Next Steps

- Form a new task team "Incident Response Task Team" – several members indicated their willingness to participate (Paul Flicek, Steven Newhouse (Co-Chairs) Caroline Rivett, Ian Poynter).
- Software Security Task Team will seek to identify metrics for associating levels of assurance with different identity-proofing methods

Privacy & Security Policy

Discussion around the Privacy and Security Policy expanding the details of the GA4GH Framework

- Recently developed by an REWG Task Team
- Aim to take a high level principles into practical security and privacy policy
- A living document
- Multiple drafts and revisions, although potentially still has gaps
- Call for volunteers to review GA4GH Security Technology Infrastructure vis-a-vis Privacy and Security Policy and to identify any changes that need to be made. (James Williams to head this up)
- Privacy: Protected under Human Rights (and other) Legislation
- Security: protection of data, individuals, business, etc. (Risk Management)
- GA4GH cannot offer guarantee, managed within abilities.
- (Perfect security means to lock data away and show no one, essentially is to have no data)
- Best Practice: develop approach to guide GA4GH work across various teams.

GA4GH Framework → P&S Policy → Operational Rules → Security Technology Infrastructure → Implementations and Operations

Set out responsibility for:

- Data provider
- Service Provider
- Access Provider
- Data Steward (key role)
- Data Access Committee (DAC)

Includes glossary, definition of terms

- Especially where one role is given a different term in different establishments

Encoding of data as early as possible / practical.

Discussion

Knox: Is raw, anonymized data covered by the policy?
Dixie: Covered by patient consent form (differing levels; open, granular, restricted, controlled, time-bound, etc.).

K: Policy requires users to agree to not attempt to re-identify donors. Tracking to pick up any attempts? Develop technical solutions?
D: Beacon currently does not track access attempts. Should be aware of? Especially multiple, concurrent, multiple attempts?
P: Technology carries a risk; people can always choose to behave badly.
D: Build in technology to pick that up, silently.
P: EBI does not track users, not able to maintain that. (But can stop IP addresses that threaten servers.) UCSC have hit-rate limit (pick up attempts to harvest data) and will control access.
TH: Possible to define granularity to detect abuse of service.

Apple ResearchKit Presentation

Steve Hershman presented an overview of the Apple ResearchKit framework and the Asthma Health Research App LifeMap developed for Mt. Sinai (slides)

ResearchKit is intended to encourage health / medical research apps.
Utilize the functionality of iPhone, etc., to gather data.

- Interpret and feed back to the individual.
- Facilitate large-scale studies.
- Open-source
- Secure
- Low threshold (easy to sign up)

Asthma App (also available: Parkinson's, Diabetes, Breast Cancer, Cardiovascular Disease)

- Gathers data
- Location (to within 100 feet)
- Environmental (weather, air quality)
- Personal (height, weight, age)
- Daily data (self-completed, daily survey, high attrition rate)
- Aim to identify hot-spots

- E-consent built in.
- No data saved until consent form "signed" (i.e., no record of dropout while working through the questions)
- Consent document summarized across 12 screens (1 - 2 sentences each screen, click next to move on)
- 3 screen quiz (True / False; assess whether people have read and understood. Or have a reasonable level of intelligence)

- Data are uploaded from iPhone to independent server, encrypted (location = double encryption)
- App has second password to deny access if your iPhone is lost / stolen.

NB. Google has not moved into eHealth (concerns about vulnerability) focused more on fitness and training tools.

FitBit, similar tools, and other medical apps are all expanding the data gathered. (Apple Watch has a built-in pulse detector, currently inactivated)

Discussion

Consent

- Consent documents tend to be substantial (need to be comprehensive)
- Legalistic language (impenetrable)
- A one page summary of a ten page, legally tight document ?
- Aim is to give people the information to enable them to make a choice ("consent" = verb not noun)

Re-identification

- Genuine risk of identifying an individual from anonymized data
- 4 pieces of information can generally be enough
- A signed agreement to not attempt re-identify needs to be backed up / enforced

(The problem of re-identification is not unique to GA4GH; ANNALS, AAPSS, 659, May 2015)

Proposal for a Privacy Task Team

- Draw up pragmatic approaches to be picked up by other teams:
  - Beacon
  - BRCA Ch.
  - MME
  - Elixir
- Tool kit, reference implementation (more than just a document, something more tangible)
- Use existing tools to test / break
- Demonstration Proj? (eg. 5 Institutions sharing real patient data across international borders)
- Implementing best practise would quickly make a difference (a lot of low hanging fruit)
- Possibly within the remit of existing team (eg. to demo an implementation for OpenID Connect = a deliverable?)

Beacon, tiered access.

Level	*Description*	*Details reqd.*	*Who*
Tier 1	*Public (anonymous)*	*email*	*Anyone*
Tier 2	*Registered (The beacon knows who you are.)*	• *email@research.establishment* • *Verified id (inc. postal address)* • *Blacked listing*	*Researcher Doctor Personal enquirer*
Tier 3	*Controlled (The Beacon knows who you are, and that you are entitled to*	• *Turn up in person* • *Passport* • *Letter from establishment (headed)* • *Signed agreement*	*Researcher Doctor*

	access.)	• *Req. institute to take action on inappropriate behaviour.* • *White listing*	

1. *Open up possibility of hospitals getting involved*
2. *Interoperability of beacons*
3. *Authenticate across a number of beacons (trust a third party to verify id.)*

Success is powerful, but achieving a long-term goal is even more so.
Establishing a method of sharing consent-limited data is a big ask.

Future Planning/Next Steps

- *As part of the on-going Software Security Task Team, define (or select an existing standard that defines) levels of assurance associated with methods for identity-proofing individuals, servers, and organizations, for incorporation into Security Technology Infrastructure. Draw upon the work of Internet2 (Michael Sullivan) as input into this task. (Lead: Knox Carey)*

- *Review the Security Technology Infrastructure with respect to the new GA4GH Privacy and Security Policy; identify gaps and recommend actions. (Lead: James Williams)*
- *Launch a new Task Team to define an approach for engaging the GA4GH community in collective Security Oversight and Incident Response, and for planning an exercise to test the approach against an envisioned scenario. (Task team members include Paul Flicek, Steven Newhouse, Ian Poynter, Caroline Rivett and others to be recruited.)*

- *Launch a new Task Team to review and write a paper (for journal submission) regarding current research efforts on privacy-preserving technologies. (Potential task team members include Jean-Pierre Hubaux, Jean-Louis Raisaro, Yaniv Erlich, and Gunnar Ratsch; need Lead.)*

- *As part of the on-going Software Security Task Team, develop and demonstrate a consumer-facing app (possibly on ResearchKit) that enables a consumer to retrieve an EHR data set and to transmit it to a server for use in research. Use de-identified data from Memorial Sloan-Kettering Cancer Center (offered by Gunnar Ratsch), and leverage the authorization profile developed by the Argonaut Project. (Lead: Knox Carey; Data source: Gunnar Ratsch)*

- *Given that the SWG will need to define technical standards for associating computable consents with shared data, reach out to the REWG and offer 1-2 SWG representatives to work on the new Task Team addressing computable consents.*

- *Encourage organizations to make public statements that their systems are compliant with the GA4GH Security Technology Infrastructure. Collect and publicize these commitments. This is an outreach task, with support and perhaps leadership from Kyra Jennings and Tom Freedman.*

Regulatory and Ethics Working Group, Internal Meetings

Summary

After opening introductions by REWG Chair Bartha Knoppers and Co-Chair Kazuto Kato (and attendee introductions), there was brief discussion regarding the informal REWG meeting held at the European Society of Human Genetics meeting in Glasgow on 6th June. Approximately 20 people attended the meeting, and there was good discussion of the REWG's current and future projects, including the topics of public engagement, the spectrum of data access (open, registered, controlled), and commonalities/differences between data discovery and data sharing.

Major Work in 2015

- The REWG's ***Framework for Responsible Sharing of Genomic and Health-Related Data*** is now available in 10 languages (thanks to all the volunteers for the translations/validations), and the **Consent Policy** and **Privacy and Security Policy** have been finalized and are now available on the GA4GH website. An article exploring the attributes of data safe havens, discussed at the REWG meeting at the second GA4GH plenary meeting in San Diego in October, is forthcoming in the journal *Bioinformatics* (lead author: Paul Burton).
- Ted Dove and David Townend (Co-Chairs) provided an update on the **Data Protection Regulation Task Team**. They thanked Minae Kawashima and Natsuko Yamamoto for keeping them up-to-date on developments in Japan's data protection statute, and noted that the proposed EU General Data Protection Regulation is nearing a vote on a final reading in the European Council on June 15-16. If it passes, trilogue negotiations will begin with the European Commission and European Parliament and likely run until the end of 2015, leading to possible final agreement and passage in 2016 (the Regulation would not come into effect until 2018 at the earliest). There is cautious optimism that the Regulation will be favourable to health research, and the strict European Parliament version from March 2014 will not get carried forward.
- David Townend (Chair) provided an update on the **Ethics Review Equivalency Task Team**. The Task Team will be meeting at the Brocher Foundation in Hermance, CH at the end of June to spend 2.5 days working through the key issues to determine whether it may be possible to identify elements of mutual recognition between research ethics committees (RECs) across jurisdictions for multi-site genomics-related studies. A meeting attendee questioned whether there was training for RECs in other parts of the world. David responded that EURECNET has regular webinars for procedural issues and substantive issues, to involve lawyers, ethicists and the broader community. Another attendee mentioned H3Africa is working on harmonization of RECs (a bottom-up model) and a training module. An attendee wondered whether accreditation could be a possible approach. David says that accreditation is something to consider, but that there is a fear of 'ethics imperialism'. Mutual recognition based on common attributes is what the Task Team is looking for at this stage, though accreditation could be considered in the future.
- Adrian Thorogood (Coordinator) provided an update for the **Accountability Policy Task Team**. The Task Team is Co-Chaired by Mats Hansson and Calvin Ho. The Task Team's first face-to-face meeting was held on 11th June. They are limiting the scope of

the Policy to accountability for data sharing issues, keeping within the remit of the GA4GH and REWG. It was discussed that to enable data sharing, the focus should equally be on stimulating incentives and not just imposing sanctions. The Task Team intends to complete its Policy in autumn 2015.

- Jan Friedman and Martina Cornel (Co-Chairs) provided an update for the **Paediatric Team Task Team**. The focus at the first stage will be on **genomic newborn screening**. A reading list has been composed and distributed to the Task Team members. The goal is to produce one or two tools over the next year. The Task Team may well consider other topics such as **pre-natal screening**, but this is a future topic to explore as it is not within the Task Team's original terms of reference. A sub-Task Team could be created for this topic and others (e.g. paediatric cancer).
- Tony Brookes (Co-Chair) and Emily Kirby (Coordinator) provided an update for the **Machine Readable Consent Task Team**. The Task Team is Co-Chaired by Tony and John Wilbanks. The goals are still open to shaping. There is no standard for machine readable consent. The Task Team might start with a white paper, and/or create something informatics-driven, and seek a group to pilot it. There should be some principles at play: what happens in the Task Team should be consistent with how we are trying to move with consent generally, paper-based or otherwise. An attendee questioned whether the focus would be on clinical consent or research consent; Tony responded that this remains to be seen, but that the Task team should perhaps focus on the area that is most acute. Another attendee noted that HO7 started a draft at a machine readable consent intended for clinical use. Tony added that while it is great to explore the topic theoretically, he would prefer to dive in and make a go for a machine readable consent, perhaps starting with research consent. The Task Team is not discussing interference with the consent process, but rather what happens with the consent after, i.e. how it is used. The Task Team has obtained funding for a workshop from IRDiRC for a meeting in autumn 2015.
- Stephanie Dyke, Liaison to the GA4GH Demonstration Projects, provided an update on the GA4GH **Demonstration projects**. The **Beacon project** is having ongoing discussion about data use conditions (possibly linked with the Machine Readable Consent Task Team). There is ongoing 'registered access' discussion with a possible link to **BRCA Challenge** and **Matchmaker Exchange**. BRCA Challenge held its own meeting June 12-13 at UNESCO in Paris. An upcoming project within BRCA Challenge will explore liability concerns with provision of access to variant databases with different access levels and reliance by physicians for clinical care. Stephanie noted that the **eHealth Task Team** and Consent Task Team met on 9th June to discuss pedigree consent. All Demonstration projects are looking for volunteers from the REWG. Anyone interested should contact Stephanie, Bartha Knoppers, and/or the Demonstration project leaders.

Future Planning/Next Steps

- REWG Coordinator Ted Dove will be stepping down as REWG Coordinator as of 1st July 2015. Ted thanked everyone in the REWG, especially Chair Bartha Knoppers and Co-Chair Kazuto Kato, for their leadership, collegiality and inspiration since the Working Group was formed in late 2013. Ted will turn to focus on his PhD at the University of Edinburgh but will continue to be involved in the REWG and GA4GH. As of 1st July 2015, the new REWG Coordinator will be Adrian Thorogood, a lawyer based at the Centre of

Genomics and Policy at McGill University. Adrian is currently the Coordinator for the Accountability Policy Task Team.

- Bartha Knoppers will ask the UNESCO / International Bioethics Committee (IBC) whether they would consider endorsing the *Framework* itself as consistent with the declarations the IBC has put forward since 1997. The OECD is creating a new committee to come up with a Council recommendation on the use of health-related data; Bartha will speak with them as well to see how the GA4GH can work with them.
- The REWG accomplished a number of items in 2014 and the first half 2015. How much can we do going forward? The REWG meeting concluded with a 'wish list' of next steps for the second half of 2015 and 2016:
- A **Lexicon** of data sharing-related terms to be openly published on the GA4GH website.
- A new Task Team to explore circumstances in which an individual should or should not be able to have **access to his or her own health-related personal data** (e.g. raw genetic data). The Task Team will be led by Anna Middleton and another Co-Chair.
- A new Task Team to explore data sharing issues in **aging/cognitive impairment** genetic research. Among other things, the Task Team will set out principles and policies for research on people with cognitive impairment.
- A new Task Team to explore the **spectrum of data access** (from open to registered to controlled). The Task Team will look at what a registered access level means, i.e. what its conditions/attributes would be as distinct from open access or controlled access, and when one can (or should) move from one system to another.
- A new Task Team to explore **liability issues** about provision of access to variant databases with different access levels, including the potential identifiably of allelic information.
- The creation of a fully functional and interactive **REWG webpage** on the GA4GH website.
- REWG members to work with the Clinical Working Group's Clinical Actionable Cancer Genome project to create a **Start-up Kit**.
- The REWG as a whole needs to think about the increasing flow between clinical and non-clinical data, and the integration of all types of data, and the implications from a regulatory and ethical perspective.

Clinical Working Group, Internal Meetings

eHealth

John Mattison initiated the eHealth discussion following a productive cross-cutting session on Pedigree Consent between members of the eHealth task team and the Regulatory and Ethics Working Group. The group began with introductions and determined that nine countries were represented in the discussion. Several members of the eHealth task team then proceeded to give short presentations on both novel and ongoing work in the eHealth space.

- Gil Alterovitz presented ongoing work to identify synergy between workflows enabling federated queries of health data. Current efforts are working to outline a common minimum set of standards between Syapse, an software company enabling the use of genomics in medical records and care, and HL7, an organization dedicated to developing open source standards for sharing electronic health information. These efforts also involve collaboration from members of similar initiatives in both England and Scotland. The discussion focused on the emerging concept of mobile health or "mHealth", which requires secure infrastructure for federated queries that preserve patient privacy. Discussion concluded with input from members of REWG.
- Gil then presented a brief update on the Catalogue of Activities, which he also went on to present in greater detail during the plenary talks on June 10. Since a prior iteration of the catalogue is currently online, the call was made for input and recommendations for the upcoming version, which will be published after internal review from the membership of the eHealth task team and the Clinical Working Group.
- Tina Hambuch presented an overview of the Illumina Clinical Laboratory, which focuses on understanding the challenges and needs of customers who use Illumina sequencers. In particular, their work evaluates undiagnosed disease, for which pedigree information in combination with genome sequencing data is often crucial. Discussion then focused on determining the proper ontology/ies for this work and ways to incorporate these efforts into the GA4GH, such as through the Family History/Pedigree Consent work group or the DWG's variant annotation task team. PhenoDB, an analysis tool enabling secure, searchable access to case level data to enable interpretation of WGS and WES for clinical and research use, was suggested as a potential useful tool for some of these needs.
- Rodrigo Barnes of Aridhia presented an update on the federated genomic analysis being done by Genomics England at the Genomic Analysis Centre in the UK. Discussion focused on the use case of NHS customers sending their data to Genomics England, and the ongoing work to incentivize development of tools for federated queries by keeping the cost of initial and continued implementation low. Genomics England is currently working to model these concepts using public test data at multiple sites.
- Ada Hamosh from Johns Hopkins presented on different pedigree tools that are in development to enable better sharing of genomic and clinical data in health and medical records. These included PROBAND at CHOP, an iPad drawing tool that connects to Phenotips. Work is ongoing to enable this program and others, such as PhenoDB, into Epic's workflow. Discussion focused on how much independence to give individual clinicians to enter data elements of particular interest to them, and how to best use technology such as iPads.

- John Mattison finished the time with a discussion on the topic of federated databases, on which he has been drafting a white paper outlining many needs and other crucial information in collaboration with members of the GA4GH, in particular Ed Conley. The premise of this paper states that data elements will transform between federated and centralized models as data evolve within local settings, and in those transformations, data will be lost or degraded. There are, and will be, so many databases that they cannot be fully harmonized. When we do mappings and there are degradations, there will need to be an evolution of database schemas that can support these changes in data storage.

Data Working Group, Internal Meetings

AVRO Objects (DWG): *A working meeting around the software engineering techniques by which the AVRO IDL may be used to better represent object relational hierarchies within the Genomics API.*

Summary

During the Leiden meeting, various members of the DWG had a lively discussion about our current methods of work. The goal of the session was to try to bring as much community comment as possible to flush out what has worked, what hasn't and why we think it did not, and what the group will do about issues that we identified. The point of the discussion was specifically not to commit the larger group to a specific solution without input from the broader community.

Identified Issues

- A lack of clarity about what the specification is actually specifying, and what is the "source of truth". For example, are we specifically requiring AVRO JSON format (which has some differences from a standard JSON translation of the AVRO IDL Schema) as the source of truth and thus require that AVRO JSON be the on-the-wire format
- A lack of beginner documentation/guidance/software library support to enable them to move from beginner status to capable implementer.
- There is a string desire from the community to be able to have access to the data model, and hence data, separate from the methods defined in the specification.
- There are serious issues being encountered with incompatibility between different AVRO language bindings.
- There is a current lack of tooling around our schema and process to enable developers of the schema to test out proposed schema changes before putting to the larger group. In particular an implementation and validation framework is seen as crucial to further development of the schema and reference implementation
- We should have an index of implementations out there for reference.
- Be clearer about what our goals are. Explicit identification of use cases and have we met them.
- There is a strong desire for a revamp release process.
- There is a strong desire for adoption of real development branching models. For example, the current Reads API is lacking for some RNA-Seq use cases, and it would be handy if they could fork a specific release of the Reads API and develop what they need there, with an eye towards merging the necessary changes at a later date to bring it back into mainline. Similar story can be told for graph genomes team.
- We need to identify how much work would be really needed to shift away from AVRO to a proposed protobuf based workflow.

Demonstration Project Meetings

Matchmaker Exchange Meeting: *An open, working meeting of Matchmaker Exchange to discuss the MME API, technical aspects of MME, practical aspects of MME, and long term goals of the project.*

- During this meeting, attendees received an update on existing matchmakers (PhenomeCentral, GeneMatcher, and Decipher), an update on the technical aspects of the MME API, including supporting implementation and defining barriers, and the long term goals of the MME and next steps. Attendees discussed overlap between Beacon and Matchmaker Exchange, whether data deposition should be a requirement for access, and whether a federated approach is most appropriate.

BRCA Challenge Meeting: *A meeting held after the Plenary at the UNESCO Headquarters in Paris, France, in collaboration with the Human Variome Project and Astra Zeneca.*

- The BRCA Challenge for the Global Alliance for Genomics and Health held a two-day meeting at the UNESCO headquarters in Paris from June 12-13, in partnership with the Human Variome Project (HVP) and UNESCO. The meeting began on the first day with plenary talks on the need for the Challenge, as well as regulatory/ethical issues and data issues associated with BRCA variant data sharing and interpretation. Next, representatives presented from large member BRCA variant databases, including UMD, ClinVar, LOVD, and DECIPHER. The day ended with overviews of the working groups in the Data Collection and Interpretation Subcommittee and gave attendees an opportunity to get involved in the Challenge. The second day began with summaries from the previous day's working group sessions as well as from the recently held ENIGMA and CIMBA meetings, and ended with scientific discussions on the topics of modeling pathogenicity of variants of uncertain significance, and contribution of modifiers to pathogenicity risk estimates for BRCA variants. Afterwards, the Steering Committee met to discuss next steps, during which time plans for a BRCA web portal and phone app were developed.

Beacon Use Cases: *Presentations on current and future beacon use cases, and discussion of how these will shape next steps for APIs, security, and regulatory frameworks.*

- Stephanie Dyke leads the development of categorizing data use conditions/limitations.
- Needed because data sets come with conditions and restrictions on data use, so we must understand where they are coming from and how they are being represented in order to use them most effectively and disseminate widely.
- This work is being done for Beacon, but is intended to serve as a model for other demonstration projects as well.

Beacon Meeting & Hackathon: *To encourage members of other groups to learn about the project and to light beacons using our reference implementations, and a debrief from Use Case discussions to formulate future action items.*

- Overviews of Beacon's work with dbGaP, the European Genome-phenome Archive, and ELIXIR, which will launch in September 2015 and will work to integrate Beacon across the international ELIXIR Data Nodes.

- Demonstration of lighting a Beacon (which is now available on YouTube).
- Common theme of this and other meetings was leveraging the Security Infrastructure
- Have built working reference implementation of a secured access Beacon
- Discussions brought up a number of issues:
    - Revealed gaps in our thinking about how we will do authorization
    - Need to think about who will be our main identity provider
- Proposal to allow more complex queries based on Variant Annotation.
    - Build links to Variant API for more powerful search.
    - Will create proposal and circulate to G2P, Matchmaker, Variant Annotation teams

Beacon and Security Working Group (DWG, SWG): *Demonstrating how each Beacon can provide secured access to data by allowing authorized, tiered access through OAuth2.0 and OpenIDConnect.*

- William Knox Carey, chair of the Software Security Task Team, explained that his team would work with Beacon to utilize the Security Infrastructure most effectively and understand the needs coming from the field around authentication and authorization.
- Demonstration of the newly working reference implementation of a secured access Beacon
- Need to work with different task teams across the DWG and the new REWG Machine Readable Consent team

MME + Beacon: *To clarify differences between Beacon and Matchmaker and find ways to coordinate and build off the combined work.*

- Beacon can be viewed as a 'bottom up' effort, providing a yes/no response for any case of a variant or allele and can allow more casual access
- Matchmaker Exchange can be viewed as a 'top down' effort, providing genotypic/phenotypic matches to clinicians based on specific rare diseases case. It is a closed network.
- The two projects complement each other well
    - Beacon API is currently being used by at least one Matchmaker partner with the intention of rolling this out across others.
    - Could align eventually, need to have communication now in development

Matchmaker	Beacon
Links similar patient phenotypic and genotypic profiles (Top down)	Identifies the occourance of a single Allele (Bottom up)
Closed network	Open, anonymous access
Aimed at clinician to establish contact with others (e.g. rare diseases)	Initially a research tool to test the willingness of establishments to share data

Action Items, Detailed Discussion

Below is a detailed summary of the closing session facilitated by David Altshuler on June 10th.

1. Develop priority products

Name (Org): Bioinformaticians and non-clinical researchers working with clinical genomic data need ethical training and a code of conduct with minimum information about privacy, ontology, and ethics.

> ***DA:*** *This is a great, specific suggestion. We could also add doctors who are not clinical geneticists to this list.*

Peter Yu (ASCO): Routine clinical care is where data reside, but there is no structure. Research has structure but is inefficient. How do we merge these two imperfect systems to create one healthcare system that serves both purposes better and more rapidly? We need to develop the same thought processes, structures, APIs, databases, on the clinical side as we have for the research side and to create more interaction between the two.

> ***DA:*** *While there is a virtuous cycle between the clinical and research sides, there is also a sharp boundary between them. We need to hone in on whether and how to break down that barrier to create a shared learning system, which is not necessarily something that every organization in the room wants. But we have constructed a room in which that conversation can begin.*

Melissa Haendel (Monarch Initiative): GA4GH has done amazing job rallying volunteer effort to curate this data, but we need to make sure we are coordinating efforts. There are many algorithmic and technical solutions for combining machine-learning methods with manual efforts and for evaluating manual efforts to ensure consistency. GA4GH can help us come up with strategies and tools to standardize curation efforts, which are so valuable to everything we do.

Bartha Knoppers (McGill University): REWG is establishing an Accountability Task Team to address issues of researcher authenticity and validation in the context of increasing trust rather than a culture of caution. We are also building a common lexicon to enable data sharing between entities that use disparate means of patient data anonymization, whether through de-identification, pseudonymization, encryption, or other approaches.

2. Deliver packaged, working solutions

Jerome Kelleher (Wellcome Trust Sanger Institute): We need engineers in order to deliver actual software that is easy and a pleasure to use.

> ***DA:*** *All of the tools and approaches of the GA4GH (including software but also consents, for example) will benefit from clear guidelines for taking them from principles to actual implementation. One idea is a downloadable start-up kit that will allow any group to easily use our tools. How that actually happens might be the "shift" we remember from this meeting, it could be through funders or through companies that want to participate. Though we won't solve this problem today, it does need to be a focus: if we don't deliver useable tools and approaches, people will go elsewhere.*

Bartha Knoppers (McGill University): We need to encourage organizations to mention the *Framework* as the philosophical position guiding new consent forms and to build new consent forms upon the generic template provided by the GA4GH. We have seen two such consents emerge thus far.

> ***DA:*** *We have discussed how to measure our success; demonstrated uptake is one good way. We need to continue thinking about how to make our products usable, for if they are not used they will not have impact.*

3. Support demonstration projects

Heidi Rehm (Partners Healthcare): Instead of not "reinventing the wheel," perhaps we should be making the wheel faster. GA4GH has allowed us to build a massive international collaborative effort that that would not have been possible through the smaller-scale, US-based ClinGen, despite existing overlaps. We need to help scale successful existing projects and connect them to other similar groups.

> ***DA:*** *Baylor provides a good test case. They are using our tools and can say what is working (the technical tools) and where more needs to be done ("Matchmaker on steroids"). This is another way of demonstrating value. It is also a way to incentivize sharing: People, especially in the diagnostic and clinical space, will not leave after one patient sample, but will have another patient tomorrow, next week, next year. If we allow them to do their job better, they will participate.*

Jacqui Beckmann (Swiss Institute of Bioinformatics): We need a catalog of examples of the demonstrated clinical value of medical genetics efforts. This might also help incite new ideas, applications, and development by revealing gaps in our efforts.

> ***DA:*** *Yes, building a clinically-focused case to show value would be useful.*

4. Align with major data collection and sharing efforts

No comments

5. Communicate strategically with key audiences

Terry Vrijenhoek, UMC Utrecht: We need to engage patients and healthy volunteers, not just advocacy groups. Could we leverage the rapidly growing DIY community?

> ***DA:*** *From the beginning, we aimed to be patient-focused, but the leadership for such an endeavor has not yet risen to the surface. We have some thin interactions, including through member organizations such as Global Genes, but there should be more.*

Bartha Knoppers (McGill University): The more vulnerable groups are not well represented in GA4GH, so we have established a Pediatric Task Team (launched June 10th) and will soon establish another focused on aging and cognitive impairment.

Mark Batlett (Geneix): We need to communicate and engage more with everyday physicians—not just clinical geneticists, but as cardiologist, gastroenterologists, and others. They are the lynchpins between patients, payers, and hospitals.

> ***DA:*** *To help clinicians realize the value of what we're doing, we need to convince them through demonstration. When the BRCA Challenge and Matchmaker Exchange lead to diagnoses, society, physicians, and patients will notice.*

Paul Lasko (IRDiRC): Perhaps to drive patient engagement, the various advocacy organizations in GA4GH membership, who are powerful advocates for data sharing, can work jointly on something. These include IRDiRC, Genetic Alliance, Global Genes, and others.

> ***DA:*** *Yes. There is interest, but it hasn't taken flight yet. How do we make it more real?*

Steve Lincoln (InVitae): Most GA4GH clinical/translational representation is in academic medical centers. But 85% of health care is delivered in community health centers. Can we take what we're learning in academic medicine and implement it in the rest of the system? Note that doctors in community health settings don't get paid to share data.

> ***DA:*** *This is critical. We have Quest Diagnostics and LabCorp, but we need more patient diagnostic companies so that we don't win the battle but lose the war. We also need to engage with the many organizations that have huge amounts of data on (at least the American) population, because they process clinical transactions, such as Flatiron Health.*

6. Establish GA4GH as a thought leader

Stanley Letovsky (LabCorp): GA4GH could act as standard-setting body to establish a much-needed robust **nomenclature for joining variants** across knowledge sources.

> ***DA:*** *The Variant Annotation Task Team of the Data Working Group is already working to harmonize existing formats, but it will also be important to lay out and work toward a more perfect solution for the future.*

Deniz Kural (Seven Bridges Genomics): People don't share not because they lack tools or software, but because the incentive structures do not support a share first, publish later model. Many people recognize the value of correlating their data with others' shared data to, less so the value of sharing their own data once they've already done their analysis. GA4GH should serve as the international face of the genomics community and act as a lobbying organization to spur change among government funding agencies.

> ***DA:*** *This may be one area where GA4GH could act as a thought leader, but we need to have good things to say. Our best chance for changing the world's incentive structures or physicians' belief systems is through demonstration. Years ago, people weren't openly sharing human genetics data, but big consortium research arose because researchers realized they couldn't make discoveries unless they worked together. Now, researchers and clinicians are already sharing, but they are frustrated by a lack of tools, standardized consents, etc., thus they have the incentive to come together to solve problems. When we show we can make discoveries and improve health through activities like the demonstration projects, then we will see more widespread sharing.*

7. Build leadership and participation

Mark Lawler (Queens University Belfast): As we move into the implementation phase, we should consider a communication and engagement Working Group (or something similar) that provides platforms for actually communicating the value demonstrated by flagship projects and bringing more people in.

> ***DA:*** *We are working with a world-class communications and strategy consulting team that is ready to take on this task. We first need to figure out our message. Is it about the need for funding, software, incentives, education?*

8. Expand organizational capacity and funding streams

No comments