

Global Alliance for Genomics and Health

Summary Report: Partner Meeting and 2014 Goals

Table of Contents

OVERVIEW	2
BACKGROUND.....	2
PARTNER MEETING SUMMARY	3
CROSSCUTTING THEMES	3
WORKING GROUPS	4
PROJECTS.....	5
USE CASE DEVELOPMENT.....	5
WORKING GROUP PRESENTATION SUMMARIES	7

OVERVIEW

The Global Alliance for Genomics and Health (Global Alliance), formed in 2013, brings together leading international organizations working in healthcare, biomedical research, disease and patient advocacy, life science, and information technology. The partners in the Global Alliance are working together to create a common framework of harmonized approaches – seeking best practices where they exist, and developing new approaches where needed – to enable the responsible, secure, and effective sharing of genomic and clinical data.

Since drafting an initial White Paper and announcing the formation of the Global Alliance in June 2013, the Alliance has made substantial effort to engage the community and progress towards understanding the existing landscape. As of April 2014, over 175 organizations have joined the Global Alliance, and four Working Groups have been established and are already working to accelerate progress in the areas of: (a) genomic data, (b) clinical data, (c) security and privacy, and (d) ethics and regulation. This report provides an update on the activities of the Global Alliance leading up to and following the first face-to-face meeting held in March 2014.

BACKGROUND

The world has changed considerably since the human genome was sequenced. The cost of genome sequencing has fallen almost one-million fold, making it possible to systematically characterize the role of genomic variation, inherited and acquired, for its relationship to human biology and disease, enabling both research and clinical practice. In parallel, changes in computing – global sharing of information through the internet, cloud computing, and interoperability among data sets – make possible greater sharing of and learning from data. The enormous volume of genomic data available and likely to come has the potential to improve patient care in many ways, such as fast-tracking diagnosis and disease gene discovery, targeting therapies based on genetic subtype, tracking new outbreaks, or identifying new markers of antimicrobial resistance.

Yet while sharing genomic data is widely acknowledged to be necessary to drive major scientific leaps in our understanding of disease, there are currently significant challenges that limit progress. The ad hoc use of different data formats and technologies in different systems, lack of alignment between approaches to ethics and national legislation across jurisdictions, and the challenges of devising secure systems for controlled sharing of data are some of the barriers that prompted the creation of the Global Alliance.

These challenges are what the Global Alliance seeks to address through plenary meetings and Working Groups, with the aim of enabling the responsible sharing of genomic and clinical data. The first meeting of partners and stakeholders in March 2014 at the Wellcome Trust was vital in providing critical input into the Alliance’s activities, priorities, and future deliverables.

PARTNER MEETING SUMMARY

The first Global Alliance Partner Meeting was hosted by the Wellcome Trust in London on March 4, 2014 and provided an opportunity for Alliance partners to reflect on progress and set concrete goals and deliverables to accomplish in the near term. Over 180 individuals participated in the plenary meeting, representing 100 organizations active in 30 countries. In addition, each of the four initial Working Groups convened a full-day meeting immediately before or after the plenary session, providing in-depth discussion of specific topics. All of the meetings were highly interactive and played a critical role in informing the next steps of the Global Alliance.

Presentation slides are available on the Global Alliance website at:

<http://genomicsandhealth.org/news-events/events/march-4th-meeting-presentations>.

The plenary session began by reviewing the mission of the Global Alliance and its mode of working. Specifically, the mission of the Global Alliance is: “To accelerate progress in human health by helping to establish a common framework of harmonized approaches to enable effective and responsible sharing of genomic and clinical data, and by catalyzing data sharing projects that drive and demonstrate the value of data sharing.” To achieve this mission, the Alliance will “convene stakeholders, catalyze sharing of data, create harmonized approaches, act as a clearinghouse, foster innovation, and promote responsible data sharing.”

Following this introductory overview, there were updates from each of the four Working Groups, interspersed with presentations from several existing data sharing initiatives by partner organizations, providing concrete examples of data sharing activities. These examples included: The International Cancer Genome Consortium; Melbourne Genomics Health Alliance; Genome Matchmaker (making it possible for labs studying rare genotypes and phenotypes to query one another for the existence of matching data); ELIXIR (a pan-European research infrastructure for biological information); and the Baylor College of Medicine Human Genome Sequencing Center.

After these presentations, breakout sessions that involved all meeting participants focused on identifying the top priority in each Working Group area and metrics for success. Finally, the group returned to plenary session to review progress over the course of the day, to review the priorities from the breakout groups, and to plan next steps.

The interactions at the plenary and satellites meetings were intended to provide Global Alliance partners the opportunity to share progress, offer feedback, develop new ideas, and arrive at a shared understanding of the key priorities for future action.

CROSSCUTTING THEMES

Based on the discussion during the March 4th Partner Meeting, a number of key themes emerged that will guide the Global Alliance’s approach and focus:

- **Facilitate and Connect:** At a time of rapid change, the Global Alliance will convene diverse stakeholders to share and learn from each other. Facilitating knowledge sharing will be an important role for the Global Alliance in 2014 and beyond.
- **Build on Best Practice:** The Global Alliance will not reinvent the wheel, but rather will identify, share, and disseminate best practices where they exist, and create new approaches only where needed.
- **Be Agile and Innovative:** The Global Alliance will be nimble, iterative, and innovative, rather than bureaucratic or monolithic.
- **Engage and Empower:** The Global Alliance will engage interested individuals and stakeholders, including patient and disease advocates and citizen scientists, and contribute to educating the public about sharing of genomic and clinical data, including the benefits and ways to enable responsible and secure sharing.
- **Increase Diversity:** The Global Alliance will work to increase participation of countries outside of Europe and North America including partners from developing countries, and to broaden Global Alliance membership. The Alliance will strive to work effectively with existing organizations that have brought together stakeholders for similar goals, adding to – not duplicating – their efforts.

WORKING GROUPS

Each of the Working Groups has defined goals for 2014, informed by preparation for and feedback received at the plenary and satellite meetings. While effective progress will only come if each group has focus and expertise, the issues are crosscutting: it is agreed that in many cases the Working Groups will need to collaborate with each other, the broader Alliance membership, and other stakeholders to achieve these goals.

- The **Data Working Group** is focused on the interoperability and scalability of formats and interfaces for genomic information. The group's two main tasks in 2014 are stewardship of existing file formats used to store genomic information (BAM and VCF files) and engaging the community in devising forward-looking data models and application programming interfaces (APIs) for representing, submitting, exchanging, and querying genomic data.
- The **Regulatory and Ethics Working Group** is drafting an International Code of Conduct for Genomic and Clinical Data Sharing, intended to support the establishment of a set of ethical principles and practices for research seeking to share genomic and clinical data. It will cover key issues such as consent, feedback to donors and participants, data security, privacy, access control, and sanctions.
- The **Clinical Working Group** aims to enable compatible, readily accessible, and scalable approaches for sharing clinical data and linking genomic data to facilitate patient

diagnosis and clinical trial recruitment. Recognizing how much work is already ongoing in this area, the Clinical Working Group has mapped current efforts in phenotype ontologies, data harmonization, and platforms for data sharing and is now working with existing groups to promote a coordinated approach to the sharing of phenotypic and genomic data.

- The **Security Working Group** aims to support a technology environment that provides assurance to patients, researchers, clinicians, and other stakeholders that data are shared, annotated, and interpreted only by those with appropriate authorisation to do so. The group is reviewing existing standards and technologies for performing transactions while protecting the privacy, confidentiality, and integrity of federated data sets in areas outside of genomics, such as healthcare, business, and finance.

PROJECTS

In order to root the activities of the Global Alliance in real-world problems and to demonstrate the value of interoperable approaches to data sharing, the Alliance will support specific projects. These may be initiatives undertaken by partner organizations and seen as exemplar efforts to help demonstrate needs and guide work for the Alliance through one or more Working Groups.

Examples of such projects that were discussed at the Partner Meeting and received substantial interest and support include the International Cancer Genome Consortium (ICGC), Genomic Matchmaker initiative (now referred to as “Matchmaker Exchange”), the Beacon project, the P3G-IPAC (Public Population Project in Genomics and Society International Policy interoperability and data Access Clearinghouse), and efforts to define genotype-phenotype relationships at specific genes of high interest, among others.

Ongoing engagement between these projects and Working Groups is intended to encourage a focus on the needs of projects currently advancing science and medicine, and crosscutting engagement of the Working Groups with one another and with stakeholders in the community.

USE CASE DEVELOPMENT

Given the broad remit of the Alliance and its diverse membership, it can be challenging to ensure that different members and groups have a shared understanding of goals and tasks. To articulate and communicate clearly shared goals coming out of the plenary meeting, an effort has started to develop specific use cases that can define, delineate, and guide the near-term work of the Global Alliance.

Each use case will include a specific objective and spell out the rationale for this objective, followed by the roles and specific steps required to achieve that objective. In addition to defining, communicating, and focusing Global Alliance activity, use cases can also identify interdependencies between Alliance Working Groups, projects, and partners.

Already, the initial development of use cases has highlighted the different levels of data sharing that are envisioned under different scenarios: from group queries that do not involve exchange of individual level data (e.g., whether a dataset contains information on a given genotype or phenotype, without actually revealing that information, or exchanging information on consent or security standards), to data sharing at the individual level (e.g., of limited or broad-based information on genotype and / or phenotype). Moreover, sharing of information that is performed within a given regulatory domain raises different questions than those that cross jurisdictional domains.

WORKING GROUP PRESENTATION SUMMARIES

Much of the March 4th Partner Meeting was focused on sharing current efforts and defining near-term goals for the four initial Working Groups. Active participation and feedback on the current work and goals were encouraged from meeting participants. Below is a summary of what was presented and discussed during the March 4th Partner Meeting by each Working Group. (The summaries below do not reflect revisions made in response to feedback at the meeting.)

Using the full-day Working Group satellite meetings and the discussions summarized below, each Working Group then revised and expanded their near-term priorities that will drive their efforts. Updated priority documents for each group can be found at: <http://genomicsandhealth.org/our-work/working-groups>.

DATA WORKING GROUP

David Haussler (University of California, Santa Cruz, U.S.A.), co-Chair of the Data Working Group, updated meeting participants on the Working Group's activities, including the activities of Task Teams under the Data Working Group. These Teams currently include representatives from several academic centers, the European Bioinformatics Institute (EBI) and US National Centre for Biotechnology Information (NCBI), as well as companies such as Google, Microsoft, and Amazon. The plan is to use open Internet resources and engagement of interest groups around each Task Team to facilitate broader input.

During the meeting, it was discussed that genomic data is currently stored and shared in a variety of formats – BAM, CRAM and VCF – but these file formats have considerable shortcomings in a global, increasingly cloud-based, environment, and were not designed for the clinical settings in which they are now being used. Despite the staggering amount of genomic data being produced globally, there is very little governance of data exchange formats.

The presentation described how the Data Working Group has incorporated the creators of the existing file formats (e.g., BAM, CRAM, VCF) into a special File Formats Task Team that it has been agreed that going forward, the Data Working Group will oversee and provide governance for the management of these genomic data file format standards. The Team supports BAM and CRAM for representing sequence reads and their alignment to reference genomes, and VCF for representing genetic variation in individuals.

During the presentation, it was noted that while file formats for data exchange have been useful, they are not sufficient going forward. File formats will not allow data systems to scale up from thousands of genomes to millions of genomes unless they are accompanied with formal data models and application programming interfaces (APIs) that allow large data sets of genome "reads" to be reorganized for efficient automated access along multiple dimensions. Because of this, David Haussler shared that over the next two years, a Read Store Task Team of the Data Working Group will act as an international coordinating effort in devising formal data models and APIs for representing, submitting, and exchanging, querying and analyzing genomic data.

The presentation also focused on the fact that the current Balkanized system of naming and indexing human genome variation relative to a canonical human reference genome structure will be challenged to scale to the clinical application of millions of genomes. Another Task Team has been set up on 'Reference Variation' to work with the community to develop the next generation of human genetic reference that includes known variation and will scale properly. The key activities are to identify known variants, resolve inconsistencies, and create standardized format for novel variants.

Many partners offered feedback that, based on their experience, developing APIs requires both cooperation and competition. Meeting participants agreed that cooperation will be vital in creating a common API, but once this application interface is agreed upon, competition between technical groups will be important in creating the most efficient technological solution. The phrase "open interface, competition on implementation" was used to capture this spirit.

In addition to standards, it was relayed that the Task Teams are each developing reference implementations and publicly (openly) available benchmark data sets so as to support evaluation of methods, e.g., for genetic variant calling. The intention is that benchmarking be available to any developer any time. One attractive possibility is to use cloud services so that in addition to comparative evaluation based on accuracy, different implementations can be evaluated in terms of time and memory requirements within the same run time environment, and so that all results are reproducible because executable code is stored.

Finally, David Haussler shared that two more Task Teams are envisioned to launch in the coming year: one to create an API for the representation of gene expression and the epigenetic state of DNA in a tissue sample, and one for metadata, which is general information about a sample such as tissue type, including how, when and where information was extracted from that sample, such as the name of the sequencing center.

REGULATORY AND ETHICS WORKING GROUP

At the meeting, Bartha Knoppers (McGill University, Montreal, Canada) and Kazuto Kato (Osaka University, Osaka, Japan), co-Chairs of the Regulatory and Ethics Working Group, outlined the group's initial focus on developing an International Code of Conduct for Genomic and Clinical Data Sharing.

The co-Chairs discussed how such a code is intended to articulate a set of ethical principles for research and the sharing of genomic and clinical data, covering key issues such as consent, privacy, feedback to patients and research participants, data deposit and access and sanctions. The Code would define protections against those who misuse such data that are volunteered for the public good.

The Regulatory and Ethics Working Group is creating this Code of Conduct through collaboration with multiple international consortia, including H3Africa, the Biobank Standardisation and Harmonisation for Research Excellence project (BioSHaRE), the International Cancer Genome Consortium (ICGC), and the International Rare Disease Research Consortium (IRDIRC).

The co-Chairs noted that funders, research consortia, and clinicians all operate within their own national frameworks and rules, but global data-sharing requires meta-level governance based on a universally shared human rights framework for addressing international problems. In addition to bioethics norms, there is a need for robust tools for the governance of genomics research. Such tools can responsibly steer the sharing and stewardship of research discovery and the integration of genomic data with clinical data for genomic medicine, namely an International Code of Conduct.

The presentation shared how an International Code of Conduct would:

- Provide a robust ethics framework for the integration of genomics and clinical data;
- Speak to groups and institutions, not just individuals;
- Urge action by governments, industry, funders, physicians and researchers to create an environment for the responsible sharing of data; and
- Foster responsible genomic research by offering stronger protection in critical areas, namely privacy, anti-discrimination, and procedural fairness.

To develop the Code, the Working Group is drawing from different instruments, such as the International Declaration of Human Rights, which includes the right of access to the benefits of scientific research for all and the right for scientists to be recognized for their work.

The presenters described how an International Code of Conduct is timely, as both Europe and major economies such as India, Japan and China, are setting up new data protection legislation laws. An International Code of Conduct could inform these laws while they are being created.

The co-Chairs also shared that the Working Group is developing the idea of a privacy safe haven. While genomics experts are accustomed to sharing information between themselves, sharing data internationally by clinicians offers new challenges and such "havens" require structure and legitimacy.

Finally, it was discussed that in order for data to be shared across jurisdictions, international ethics review processes need to be streamlined including criteria for international recognition of the possible equivalency of such processes. The goal would be that ethics review undertaken in one country that meets such criteria could be recognized by other participating countries as substantially equivalent.

CLINICAL WORKING GROUP

Kathryn North (Murdoch Childrens Research Institute, Melbourne, Australia), co-Chair of the Clinical Working Group, presented the work and goals of the Clinical Working Group, explaining that the biggest focus initially has been to map existing endeavors—so as to build on current efforts, rather than “re-inventing the wheel.”

The presentation focused on the mandate of this group, which is to provide information to the genomics community on standards for representing phenotypic data and linking it to genotypic information. The initial focus will be on rare genetic disorders and cancer, later expanding to more complex disorders.

The first areas the group focused on are: (a) phenotype ontology (recording and sharing patient data in standardized manner); (b) data harmonization; and (c) biomedical informatics and data extraction.

A. Phenotype ontology:

To date, the group has engaged with the following groups to develop a broad overview of the field: PhenoTips and PhenomeCentral, recording and sharing of detailed patient data developed by Michael Brudno (University of Toronto), Human Phenotype Ontology lead by Peter Robinson (Charité Universitätsmedizin in Berlin) and GeneMatcher and PhenoDB tools developed by Ada Hamosh and Nara Sobreira (Johns Hopkins University School of Medicine).

Key challenges that were discussed are: to identify commonality across all ontologies, to expand ontologies in some areas that are underdeveloped, and to reduce barriers such as the lack of electronic health record (EHR) development or IT support to implementation.

B. Data harmonization:

Genome-wide studies require the study of large numbers of individuals measured carefully for a variety of phenotypes. This often cannot be achieved at a single site or through a single study. The ability to share data requires focused efforts in data harmonization across studies and from different sites to allow co-analysis, direct comparison and replication of results. It generates analytical power because of its ability to increase sample size. To date the group has heard an overview of data harmonization and key initiatives from Paul Burton (University of Bristol) Isabel Fortier (McGill University Health Centre, a principal investigator for the Maelstrom Research initiative. Gaps that need to be addressed include broader communication of the value of harmonization and the development of resources (such as a dedicated website) that allow for knowledge transfer of how to harmonize data and provide tools, advice and examples of successful harmonization.

C. Biomedical informatics, and extraction of electronic health records and data:

The ability to extract data out of existing records including electronic health records will be critical in enabling efficient and longitudinal collection of phenotype data. To date, the group has had an overview of ASCO's CancerLinQ presented by Lillian Siu (University of Toronto), phenotyping in electronic medical health records (the eMERGE consortium) by Dan Roden (Vanderbilt University) and an overview on advances in health database technology by George Hripcsak (Columbia University). The group will continue by exploring validated methods to extract phenotypic information from EHRs with varying architectures, exploring efforts that involve a larger catalogue of validated phenotypes and validation of automated methods for phenotype development and data extraction from EHRs.

The presentation closed with a summary of the future goals of the Working Group. The group has identified the following key needs: establishing compatible, readily accessible, and scalable approaches for sharing clinical data and linking genomic data; and facilitating genotype-based clinical trial recruitment for rare diseases and cancer. The longer-term goal is to incorporate phenotypic data, linked with genomic data into the electronic health record as a secure and

individualized record of all relevant data for a specific patient that can then be linked more broadly to international data sharing initiatives.

SECURITY WORKING GROUP

Paul Flicek (European Bioinformatics Institute, UK), and Dixie Baker (consultant to Genetic Alliance, US), co-Chairs of the Security Working Group, outlined some of the key issues for the Security Working Group. A core principle of security, they said, is that security that is transparent tends to be most effective. Institutions also need to have a pragmatic attitude about the realities of security breaches.

During the presentation, it was emphasized that the goal of the Security Working Group is to identify and support data technology solutions that provide assurance to patients, researchers, clinicians and other stakeholders that genomic data are shared, annotated, and interpreted only by those with appropriate authorisation.

The co-Chairs shared that core issues that the Working Group is focusing on include: identity management; access management; data integrity and immutability; audit and non-repudiation; availability; federated data-sharing models; and processing requirements.

They emphasized that defining and communicating among data-sharing partners that equivalent security practices and technologies are being used will be critical to creating this transparent and pragmatic environment. To this end, driving the work of this group will be the definition of use cases that define: what types of data will be shared; whether sharing networks will involve a few repositories or many; whether data is stored on site by or on shared “cloud” services; who will credential individual users and identity-proof individuals and what level of assurance is required; and how global sharing policies will interact with national law.

The presentation noted that several relevant standards and technologies for protecting the privacy, confidentiality, and integrity of large data sets, and for federating databases across globally distributed entities already exist in areas beyond genomics, such as healthcare, business and finance. The group will review these technologies and standards, and road-test various security mechanisms.

The co-Chairs said that by 2015 the group plans to publish guiding principles, standards, and a technical framework that fulfil a set of requirements derived from use cases designed to enforce the policy within an environment that represents relevant technical, practical and community perspectives of genomic data sharing.