Creating a Global Alliance to Enable Responsible Sharing of Genomic and Clinical Data

June 3, 2013

An initial draft of this White Paper was prepared for the January 28th meeting, and has since been revised substantially based on discussions at and since the meeting. A list of contributors and participants is provided at the end of this document.

Table of Contents

Introduction

The cost of genome sequencing has fallen one-million fold in the past several years, fueling an explosion of information about the genetic basis of human health and disease.

In principle, this wealth of genome sequence data should accelerate progress in biomedicine – making it possible to integrate genomic and clinical information to reveal the genetic basis of cancer, inherited disease, infectious diseases, and drug responses. Beyond research, the interpretation of individual genome sequences in clinical practice requires the widespread ability to compare each genome to a compendium of aggregated sequence and clinical data.

In practice, however, we are not organized to seize this extraordinary opportunity — nor are we on a path to do so. For the most part, data are collected and studied in silos: by disease, by institution, and by country. Existing regulatory procedures could not and did not anticipate developments in technology and the value of data aggregation. If data is to be shared responsibly we need to respect the privacy and autonomy of individuals. Tools and methods for analysis are non-standardized and incompatible. If we remain on the current path, the likely outcome will be a hodge-podge of balkanized systems — as developed in the U.S. for electronic medical records – a system that inhibits learning and improving health care.

On January 28th, 2013, fifty colleagues from eight countries met to discuss this opportunity and challenge, and how we might work together to create the conditions under which learning could take place and genomic medicine could flourish. Inspired by the example of the Internet, the World Wide Web and the Human Genome Project, we discussed development of international standards and of the information technology infrastructure needed to share and integrate data in a secure, controlled and interpretable manner, unlocking discovery while respecting patient autonomy and right to privacy. The group came to the conclusion that in order to meet the needs of the patient, research and clinical communities, it will be necessary to create:

> A **global alliance** of international partners, entrusted with the mission of enabling rapid progress in biomedicine; working together to create and to maintain the interoperability of technical standards for managing and sharing sequence data in clinical samples, developing guidelines and harmonizing procedures for privacy and ethics, and engaging stakeholders across sectors to encourage responsible and voluntary sharing of data and of methods.
>
> **Technology platforms with open standards** designed to enable secure storage; access control and controlled sharing of information at multiple levels; participant centric consent; tools for data processing that support major sequencing platforms; method to make results of analyses comparable across centers and technologies; and a computational architecture and application programming interface (API) supporting innovative “Apps” and services.

The Alliance will also include **operating entities** that instantiate the platform standards, commit to shared principles, provide services and aggregate data for users, develop tools, spark innovation, and thereby advance research and learning, application and practice.

Making genomic data and tools interoperable in a secure and trusted manner will generate a powerful **network effect**: the more data and methods can interoperate on common platforms, the more valuable to patients, researchers and health care professionals each will become.

This document focuses on the **creation of the alliance**, and is divided into three sections: *Section 1* sets the context; *Section 2* considers ethical and technical considerations that will be a major focus of the work of the alliance; and *Section 3* outlines the path forward.

Section 1: Setting the Context

Why now?

Medicine is in the midst of a revolution, fueled by the ability to inexpensively gather genome sequence information in large numbers of individuals. The cost of sequencing an individual genome is expected to reach $1,000 in coming years. Before long, it seems inevitable that millions of individuals will undergo genome sequencing. Increasingly, the key challenges are in management, analysis, integration and interpretation of data, rather than in data generation.

The opportunity. In principle, it should be possible to dramatically accelerate medical progress by learning from the world's data on genome sequences and clinical phenotypes: illuminating the biological basis of cancer, infectious diseases, inherited diseases and drug response. By aggregating and analyzing large amounts of genomic and clinical data, it should be possible to discover patterns that would otherwise remain obscure – for example, which mutations within a tumor predict treatment response, or which genetic variants explain rare childhood diseases.

Clinical interpretation of individual genome sequences will be powerfully enabled by comparison to extensive data on variation in genome sequence and phenotype. At present, it is generally not possible to predict which changes in DNA sequence lead to clinical consequences. When held against a large repository of other such data, however, robust patterns and relationships can be identified. Given the wide variety of disease endpoints, varied biogeography, and low frequencies of sequence variations, data from *millions of samples* will be needed.

The challenge. Despite the clear benefits of data integration, the scientific and medical communities are not yet organized to seize this opportunity — nor are they on a path to do so.

Currently, such data are typically analyzed in isolation, with sample sizes inadequate to make robust discoveries. Incompatible methods inhibit learning across datasets. Regulatory and ethical procedures could not anticipate, and thus were not designed to enable, widespread comparison across studies and the sharing of information. Few clinical investigators have access to the analytical infrastructure needed to perform analyses for their patients; even the most sophisticated and well-resourced medical centers find it difficult to keep pace with rapidly evolving tools and pipelines. The cause of interoperability and data aggregation has been championed at recent meetings (e.g. NHGRI - June 2012) and by pioneering individuals and organizations (e.g. P3G and Sage Bionetworks). The internet and social media offer new ways for participants, researchers and hospitals to engage with one another in an ongoing and dynamic manner (e.g. Portable Legal Consent). Nonetheless, key issues remain unsolved.

Moreover, the **window of opportunity** may soon close: at present, relatively little data have been collected, and incumbent systems have yet to be established. In the absence of an open and interoperable solution, closed, proprietary systems will by necessity be created. This would create a fundamental barrier to gaining the benefits of data aggregation and slow the understanding, diagnosis and treatment of disease.

A fork in the road. In recent decades, there have been repeated choices between closed, proprietary systems, and open networks with interoperability. In the case of medical records, the United States ended up with a fragmented system, inhibiting for decades the quality of patient care and the ability to learn from experience. Today, with incumbents established, the US medical record system is nearly impossible to change.

In contrast, the Internet, WWW and Human Genome Project are open, despite efforts to create walled gardens. Secure systems make it possible to transmit private information on the Internet (e.g. financial transactions). The resulting explosion of innovation has transformed our world.

As we enter the era of widespread genome sequencing, we face another such fork in the road.

Potential for impact on human health

These developments create new opportunities to gain insight into disease, improve prevention and early detection, define diagnostic categories, streamline clinical trials, and match patient to therapy. The impact can be rapid (e.g. targeted therapy based on genomic characterization) and longer term (discovering molecular targets, leading to new and more effective therapies).

Cancer therapy and prognosis. Targeted therapies hold enormous potential for cancer patients, with more than 800 targeted cancer drugs being tested for FDA approval. Matching drugs to patients requires widespread collection and analysis of genomic data in a dynamic and ongoing manner. Where only a small subset of individuals carry a particular genetic mutation, very large numbers of patients, beyond the scale of any single institution, will be needed.

In some cases, therapy shows limited efficacy in a broad group of patients – but a small fraction of patients show dramatic responses. The ability to jointly analyze genomic and clinical data from these rare responders can lead to new predictors of clinical response — information that can then be used to match patients carrying these predictors to drugs that will help them.

Rare genetic diseases. More than one percent of all newborn children have a developmental or genetic disease. The combination of genome sequencing and clinical phenotyping offers the surest hope for diagnosing disorders due to known genetic syndromes, and to discovering underlying causes for many of the rest.

However, most of these disorders are individually very rare, such that no single hospital will ever see enough cases to forge convincing links between genetic mutations and disease. Only by analyzing data in aggregate will it be possible to unlock the potential to diagnose these disorders that are individually rare, but collectively common. (For an example of a pioneering project, see Deciphering Developmental Disorders (DDD))

Common medical conditions. One of the most pressing challenges in the pharmaceutical industry is the high rate of failure in clinical trials: only a small fraction of drug candidates that enter the clinic emerges as a safe and effective approved drug. This high failure rate in drug development is due in large part to our ignorance of the underlying root causes of most diseases, and the limited ability of preclinical models to predict safety and effectiveness in patients.

Human genetic information offers an opportunity to improve the rate of success in drug discovery, by directly linking drug targets to clinical outcomes in humans, and by helping to stratify patients for treatment based on the underlying genetic causes.

Infectious Diseases. Sequencing technologies can be used to monitor the spread of infectious agents at unprecedented spatial and temporal resolution, speed time to diagnosis, study microorganisms that are as yet uncultureable, elucidate susceptibility or resistance to antibiotics, and reveal changes in skin and gut flora that are associated with disease states.

Achieving these benefits requires overcoming the many barriers to aggregating, responsibly sharing, and jointly analyzing genomic and clinical data at adequate scale and quality. Success will benefit the entire biomedical ecosystem (see Box: “Serving the needs of the biomedical ecosystem”).

Box 1: Serving the needs of the entire biomedical ecosystem

Many segments of the biomedical ecosystem are negatively affected by the inability to share data on genome sequence, clinical phenotype, treatment and outcome in a safe, secure and reproducible manner.

Patients with cancer and inherited disease want to learn the causes of their disease, and to find any targeted treatments that may exist. Many will choose to contribute data to help develop better solutions for their families and communities, as long as their wishes and privacy are respected.
Patients need a trusted route for the altruistic sharing of personal genetic information to accelerate progress, including ways to manage privacy and consent.

Scientific Researchers collect genetic information at a prodigious rate, but generally lack access to software tools and computational infrastructure needed to manage this magnitude of data. Investigators working in each disease solve the same set of computational and software challenges, but often do so in ways that are not interoperable. Few have access to sample sizes needed to achieve power, and thus to forge unexpected connections among diseases.
Scientific researchers need secure data storage, cutting- edge software tools, and high performance elastic computing. Interoperable tools and data will empower each scientist's work, and facilitate secure and ethical sharing.

Hospitals and health care systems increasingly need to collect, store and interpret genetic information, but it is expensive for each to individually create software tools and infrastructure. On their own, each lacks the critical mass of comparative data required to care for their own patients. The lack of norms for sharing and protecting data are slowing the uptake of genomic medicine in the clinical setting.
Hospitals need a trusted solution to manage and process genomic data, tools and services for interpretation, and a network of partners to share knowledge and information to help individual patients.

Biopharma relies increasingly on genetic information both to identify new targets for therapy, and for patient stratification and the design of clinical trials. However, companies find it difficult to obtain access to genetic and clinical information, lack internal expertise and infrastructure to perform analyses, and worry about liabilities associated with genetic data regarding informed consent and privacy.
Biopharma needs enterprise-ready software solutions and computation resources, access to public data and potential collaborators, and standards for informed consent to facilitate use of genetic data.

Clinical Trials depend on patient recruitment -- identifying patients with specific genomic alterations, and understanding their natural history, is a central challenge to trials for targeted therapies.
Clinical trialists need mechanisms to identify patients with specific genomic alterations, to follow them over time, to design efficient and powerful trials, and to invite patients to join in research. Social networks catalyzed by data sharing may provide new routes to access patients for research and trials.

Governments and Foundations increasingly require the data they fund to be made broadly available. However, they lack mechanisms to manage sharing of data, and much data sit idle in locked storage. They are often asked to fund redundant projects to develop incompatible software for genetic analysis.
Governments and foundations need a solution for storing and managing access to data, responsibly sharing data and methods, and ensuring government and philanthropic investments return the greatest yield.

Disease Advocacy Organizations exist to bring together communities of patients, provide information, and catalyze research. Many are now catalyzing genetic research projects in which patient-members provide samples and clinical information. Each faces the same challenge of developing platforms for patient interactions, and finding a solution to storing and analyzing data they collect.
Disease advocacy organizations need platforms on which to develop branded sites to interact with their members, and a turn-key and integrated solution to storing and analyzing clinical and genetic data that is economical, dependable, secure, and preserves privacy where desired.

Engaging stakeholders

The next sub-section considers this issue from a variety of stakeholder perspectives: that of public opinion, health care providers, industry, and the setting of technical standards.

Engaging stakeholders: public attitudes towards sharing of data

Research on public opinion regarding privacy and the use of genetic information informs our understanding of public support for sharing of genetic and clinical information, and provides insights as to how to govern and organize so as to promote public engagement and trust. Key themes include:

- Public attitudes related to privacy and data collection are highly varied, and thus it will be necessary to tailor approaches by regions and nationality[1].
- Survey data indicates among some respondents a willingness to participate in sharing of genetic information.
- Attitudes regarding "public" release of data (without restriction on use) vary substantially, encouraging approaches that provide participants more control of access and use.
- Protections on data use that are important to citizens need to be acknowledged and addressed.

The research discussed below has substantial limitations. Much has focused on "biobanks" rather than sharing of data from other sources, and focused on privacy of personal data more than on secure uses of data. Public attitudes may have changed since this research was done.

Varied attitudes towards sharing data

When asked in the context of a biobank, stated willingness to provide personal information varies widely across countries. In Iceland, Sweden, and Norway, a strong majority responds affirmatively to the question: "Would you be willing to provide information about yourself to a biobank?" The situation is reversed, however, in Latvia, Greece, and Lithuania, with a majority of respondents opposing. Opinions regarding sharing personal data also vary widely across the EU. While most support data exchange overall, including large majorities in countries like Cyprus, Iceland, and Finland, the rate of concern is much higher in countries such as Austria and Germany.[2]

Various factors play a role in decisions to share personal information. In one study, respondents who knew about biobanks were more willing to share information than those without prior knowledge. Age, education, and religion are associated with attitudes towards participation.

Any international effort will encounter variable attitudes on specific issues related to giving, sharing, and holding personal data. For example, when asked for their ideas about what institutions will protect data, a majority of Chinese and Japanese respondents put their trust in their governments to protect their personal information, compared to approximately half of Canadians, and only 20 percent of Brazilians[3]. Thus, generalizing about public opinions across borders should be avoided, as attitudes may be relatively specific to countries or areas.

In the United States, a majority of respondents in two different projects reported willingness to share genetic data for the purpose of scientific or medical research. One study asked "assuming that appropriate privacy protections were used, would you be willing to share your personal health information to advance medical research?" Two thirds replied in the affirmative[4].

Another study asked if participants would "contribute a DNA sample for use in current or future scientific or medical research". Three-fifths reported being "somewhat willing" or "willing"[5].

A majority of Europeans stated support for use of genetic data for disease research. One study asked: "Would you be willing…for your genetic information to go in to a national data bank for research into the origins of diseases?". Three fifths replied "yes, definitely" or "yes, probably"[6]. Another study showed a narrow plurality of support in Europe, with 46 percent of Europeans overall willing to provide personal information to a biobank, and 44 percent opposed[7].

As part of a 2007 study by the Institute of Medicine (United States), survey data was collected on attitudes towards health data and privacy[8]. Substantial majorities expressed trust in health care providers and, to a lower extent, health researchers to protect privacy and confidentiality of personal health information. Nonetheless, a majority answered that existing regulations did not go far enough to privacy, and a majority said they would agree to participate in research only if their permission was asked (38%) or their identity could never be revealed (19%).

Access to and control over data

Survey data show that individuals respond differently based on the protections in place around data usage, and on information regarding consent. In the United States, a commercial survey found the top three privacy concerns for the public were security protection, sharing of data only with consent, and the ability to delete one's personal information[9]. A significant majority of Europeans stated they believe genetic data should have special protection in the same category of information like health, religious beliefs, or ethnic origin[10].

Among the American public, research has shown a general willingness to publicly share genetic data, but this willingness decreased as the participants were offered more information and options about data use. In one study, 80 percent of participants initially consented to public data release, but after a follow up debriefing, only 53 percent chose public release of their data[11,12].

Studies have asked which institutions are considered by the public as most trusted to hold or share data. An international study showed that a minority of respondents trusted private companies or the government to protect personal data[13]. In the United States, healthcare, consumer products, and banking were most trusted with regards to privacy[14]. In some surveys, concern was expressed about sharing of data with commercial stakeholders or government, but these concerns don't appear to cause participants to withdraw or withhold their data.[15]. In the United States, security and personal control of data are more important than elements such as having clear policies[16]. In Europe, a significant majority were opposed to private insurance companies having access to people's genetic information.

This brief overview of existing studies suggests several conclusions.

- There is evidence for a general willingness to participate in data sharing.
- Substantial differences are observed based on regional, national, and demographic factors. Thus, a "one size fits all" approach is unlikely to succeed.
- Many in the public care about control over use of their data, informed consent, privacy and security. A successful solution will have to address these concerns.
- Individual choices are related to the information provided, and thus transparency is key.
- Public opinion is responsive to changing circumstances and is likely evolving rapidly.
- Public opinion research with relatively rapid turnaround times will be needed to identify the concerns of individuals and groups, and to develop approaches that remain current with attitudes as they change over time.

Engaging stakeholders: health care providers

Genome sequencing has thus far had a greater impact in research than in clinical medicine, but going forward there exist clear and compelling applications in diagnosis and stratification. Already, health care institutions are investing in laboratory and computational infrastructures to perform clinical genome sequencing. To improve human health, data resources and informatics capabilities must be enabled for use in clinical care. Moreover, as genome sequencing moves into the clinic, health care will become a primary location for collecting the phenotypic and genetic data needed to create learning systems for research and clinical care.

Challenges specific to clinical sequencing

Clinical medicine places perhaps greater demands on accuracy and reproducibility of genome sequencing than does research, and yet to date expertise and methods are concentrated in the research setting. Clinical medicine is regulated, such that tools and approaches used in clinical care may need to be approved or certified by national or local regulatory bodies. Genome sequencing is unusual (for a clinical test) in that the methods and data are rapidly evolving, and have often changed by the time evaluation for approval can be performed.

Health care delivery systems can support the clinical application of genome sequencing with available mechanisms for interacting with patients, interpreting tests, delivering results to patients, and performing clinical research. However, the standards, technologies and expertise, as well as the comparative data needed to make genome sequencing useful in clinical care, are generally lacking.

Incorporating clinical and phenotypic data

Learning will require joint analysis of genomic sequences with clinical phenotype and data on outcomes. In principle, the more extensive the phenotypic data, the greater the opportunity to learn. In practice, access to phenotypic and clinical data will vary dramatically across the community due to cultural, regulatory, ethical, historical and technical factors. Thus, genome sequencing in the clinic will necessarily involve marked heterogeneity among healthcare systems regarding the amount and detail of phenotypic and clinical data available; the sophistication of medical records systems; and the accessibility of that data for research uses.

Based on this heterogeneity, a varied and staged approach will likely be appropriate, enabled by regulatory harmonization and shared technical standards. Where extensive clinical data is available, its value will be unquestioned, experience can be gained and technical capabilities developed. But, even limited clinical data can be useful, and individuals and institutions may want to walk before they run. A staged approach recognizes the challenge in harmonizing the diversity of clinical data types, avoids delay until these challenges have been overcome, while retaining a focus on the importance of fully integrating genomic and phenotypic data in time.

In addition, a diversity of environments and approaches will enable different entities to solve distinct and complementary problems. For example, some countries and integrated systems have in place advanced electronic medical records and appropriate consent. Such groups will be well positioned to advance methods for managing and harmonizing phenotype data, but may lack extensive experience with technologies for genome sequencing. Other groups may have great expertise in technical or analytical methods, but lack expertise with phenotype data. A mechanism for learning about the harmonization of clinical data alongside learning from harmonization of genome sequence will be more successful than if these two remain independent, as otherwise may be the case.

Engaging stakeholders: industry

A vibrant and innovative ecosystem will require and benefit from for-profit as well as not-for-profit organizations. Just as the non-profit World Wide Web Consortium spurred creation of innumerable commercial enterprises, interoperable platforms for medical genomics can create conditions under which private innovation can support medical advances and public benefit.

In aiming to enable both the non-profit and for-profit sectors, a level playing field is needed on which a diverse array of organizations can innovate and compete. A clear voice is needed regarding ethical and regulatory issues, as well as a set of shared technical standards that support a diversity of sequencing platforms, data transfer protocols, and distributed-computing capabilities. This combination will encourage continued innovation and reductions in cost.

It would be neither necessary nor desirable to limit the diversity of approaches that these enterprises may develop. The World Wide Web is an open ecosystem that allows and supports both non-profit solutions such as Wikipedia, and "walled gardens", such as Facebook and eBay. The freedom for individuals to choose how their data are used must include their freedom to participate in market-driven solutions offered by private enterprise.

Nonetheless, similar to the role of the World Wide Web Consortium (W3C) in maintaining standards that promote interoperability, competition, and innovation on the Web, there needs to be an authoritative group that works continuously to advance the core principles such as respect for patient and participant autonomy; collaboration across sectors and jurisdictions; technology platforms with open standards; compliance with relevant regulatory and ethical frameworks; and transparency in governance.

Engaging stakeholders: setting standards

The creation of shared, interoperable standards is a lynchpin of many fields, but has been less typical in life science. Recently, W3C, IEEE, the Internet Society, IETF and IAB created OpenStand as "a global community that stands together in support of The Modern Paradigm for Standards – an open, collective movement to radically improve the way people around the globe develop, deploy and embrace technologies for the benefit of humanity."

The Modern Paradigm is described as including: cooperation (between organizations), adherence to principles (such as due process, consensus, transparency, and balance), availability (accessible to all), voluntary adoption (success is determined by the market), and collective empowerment, meaning commitment to striving for standards that:

- are chosen and defined based on technical merit;
- provide global interoperability, scalability, stability, and resiliency;
- enable global competition;
- serve as building blocks for further innovation; and
- contribute to the creation of global communities, benefiting humanity.

As the genomic and clinical communities consider the creation of shared standards and harmonized approaches, much can be learned from examples from other fields.

A path forward

Advancing medical knowledge and improving clinical care will require the widespread ability to access genomic and clinical data in a secure and trusted manner, and to enable comparison of genetic variants and clinical features. As regulations and attitudes towards sharing data vary within and across national boundaries, and as different sectors have different needs and goals, it is necessary to enable interoperability while preserving diversity of approach and application.

In order to achieve these goals, a **global alliance** is needed to bring together researchers, health care providers, funders, disease advocacy groups, life science and technology companies, and informed citizens to enable, support and promote the responsible sharing of genomic and clinical data. This global alliance will become a trusted voice that works to:

- Advance the idea that patients have a *right to share* genomic and clinical information to advance human health, as well as to *privacy* and to *transfer* data as they choose;
- Collaborate with governments and funders to create and promulgate *policies and regulations* that allow individuals and organizations to choose to share information, while respecting and addressing their needs and those of local communities;
- Support the development of *open technology standards* (consistent with policies and regulations), the creation of reference implementations, and an innovative ecosystem that advances knowledge and health by technical enablement of information sharing;
- Provide a forum to evaluate and share *best practices* and *create collaborative projects* to accelerate the positive impact of genome sequence information on medicine;
- Support a *level playing field* that enables technology and business innovation;
- Enable *social networks* to bring together patients and families with the genomic and medical communities, facilitating access to tools, data and methods.

The global alliance will work with its partners to create technology platforms that are extensible with open standards, formats and tools that enable stakeholders to:

- Store genomic data and relevant clinical information in a *secure* and *trusted* manner;
- *Enable users to share*, while managing informed consent, regulations and privacy;
- Provide tools for *participant-centric initiatives (*such as the Portable Legal Consent) that enable portability across studies and sites, and for patients to engage with researchers;
- Support *multiple sequencing platforms* and *cloud providers*, encouraging innovation;
- Analyze variants observed in a given study in light of extensive *comparator data*;
- Collaborate as teams, enabled by provenance, file tracking, and attribution;
- Catalyze the rapid development and distribution of *third-party "apps" and services* that support data analysis, clinical interpretation and knowledge discovery; and
- Provide ways for *communities* of users (e.g. researchers or families working on a particular disease, or pharmaceutical companies looking to enroll clinical trials) to communicate and share, including creation of "safe haven" enclaves as needed.

To bring this to life, multiple **operating entities** will be needed to instantiate the platform standards, and serve the needs of specific users such as managing, storing, processing, and analyzing data; supporting queries and brokering transactions. These entities, which may be a combination of existing and new organizations, will bring together users, datasets and tools to create value for individuals and organizations. Operating entities that join the global alliance will commit in writing to support the mission, core principles and standards set by the alliance.

If successful, this effort will generate a powerful network effect, with increasing returns to scale: the more users, data and analytical methods become interoperable and networked, the more valuable each will become to patients, health care organizations, technology providers, and most importantly, to the goal of advancing medical knowledge and human health.

Section 2: Regulation, Ethics and Technology

Regulatory and ethical considerations

Context for Global Action

A discussion of global collaboration on human research participants data must start from an understanding of the relevant national and international laws, policies and procedures regarding the ethical conduct of research and clinical care in the research setting, including, among others, *informed consent, patient privacy, protection of research data including electronic data privacy,* and *research oversight*. An interoperable platform that supports storage, analysis and controlled sharing of genomic and clinical data will in each jurisdiction need to adhere to applicable regulations and standards. To the extent that platforms support data used for patient care, they will also be subject to the *laws governing the provision of clinical laboratory services*.

Legal frameworks and regulatory requirements differ substantially both within and across the US, UK, Canada, Europe and elsewhere. Privacy protections on personal data vary considerably. Some policies relate specifically to health or genetic information while others apply broadly to personal information, inclusive of health information. Such legal differences require an organized effort to enable international collaborations to share data across borders.

Thus, to the extent that the international scientific and medical communities believe that responsible sharing of data will be key to future progress, there is a *pressing need to engage the public and relevant governmental authorities on a collaborative effort to harmonize* policies, procedures and regulations across jurisdictions. This will require a firm foundation of trust from stakeholders regarding protection of privacy, autonomy and subject rights, respect for local jurisdiction, coordination of global research on Ethical, Legal, and Social Implication issues[17], and commitment to the mission of improving human health and patient care.

A non-governmental, international not-for-profit global alliance could have a major impact, by *bringing stakeholders together across jurisdictions* to share information and best practices, and by supporting local parties that work in-country to align (to the greatest extent possible) rules, regulations and procedures under which such data are managed.

Overview of Existing Ethical and Regulatory Framework for Human Subjects Research

A review of international documents such as the Helsinki Declaration[18], Belmont Report[19], European Convention of Human Rights and the Convention of Biomedicine[20], and the UNESCO Universal Declaration on the Human Genome and Human Rights[21] reveals *a core set of shared ethical principles: respect for persons, the right to self-determination, and the right to make informed decisions*. Although the principles have been translated into research oversight differently in each country, this **similarity of core principles** provides a foundation for the future alignment of procedures. Notable commonalities include the use of Ethics Review Boards, Institutional Review Boards, and Informed Consent documents. Moreover, international consortia (e.g. the International Cancer Genome Consortium: ICGC) have translated these core ethical principles into policies, procedures, tools, and, governance that facilitate interoperability.

Responsibly sharing data and information internationally will require engagement with a wide variety of national regulatory agencies responsible for health-care data collection, storage, privacy, access and use. Privacy Commissioners generally have oversight responsibility for personal health information, which is considered highly sensitive. The variety of health and research regulatory agencies includes: in the United States, the Office of Human Research Protection (OHRP), the Food and Drug Administration (FDA), and the Office of Civil Rights; in Canada, Canadian Institutes of Health Research; in France the *Comité de protection des personnes* (CPP), *Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la santé* (CCTIRS) or the *Commission nationale de l'informatique et*

deslibertés (CNIL); in the U.K, the Human Tissue Authority and the newly formed Health Research Authority; in Mexico, the Ministry of Health; in Thailand, China, and Malaysia, the Ministry of Health; in South Korea, the Ministry of Health and Welfare; and in Japan, the Ministry of Health, Labour and Welfare (MHLW) and the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Given the wide range of agencies, a federated approach is required, where a set of shared principles (developed in concert with the support of the global alliance) are advocated locally by parties that live and work in-country, and who understand best the local context.

Key issues

Key issues to be considered include public engagement; protection of privacy, participant-centered initiatives (PCI); regulation of access to data and results by participants, researchers, and others; sharing of data across institutions and jurisdictions; re-contact of research participants; and the governance and stewardship of repositories. In these areas there are inherent tensions between important principles. To earn and keep the public trust, they will need to be identified, discussed openly, weighed, and managed responsibly. This sensitive and complex endeavor must ultimately be guided by the basic principles of a civil society, and respect for diversity of opinions and values among individuals and countries.

Public engagement.
The guiding principle should be the ***autonomy and self-determination of individuals*** who provide data. This applies to patients receiving ***clinical care*** as well as ***research participants*** throughout the research cycle. While other considerations, such as advancing research and improving patient care, loom large, autonomy of the individual must come first. Public engagement will be needed to encourage recruitment, active involvement and commitment of research participants from different populations. This will include a clear articulation of the public good and potential benefits that depend on large-scale and inclusive involvement of the population in research. Governance that is appropriate, transparent and accountable, and that incorporates public participation and is participant-friendly, will enhance and promote public trust.

Protection of Privacy and Participant-Centered Initiatives
In a report by the Institute of Medicine (United States), health privacy was defined as "an individual's right to control the acquisition, uses, or disclosures of his or her identifiable health data." Issues of privacy are an increasing focus of attention, with technical advances posing new challenges and pending reforms in both in Europe[22] and the USA. These proposed reforms seek to make privacy protections clearer but in Europe potentially could narrow the scope of research and data sharing. At the same time, broader societal trends in social media, "recreational" genomics (e.g. ancestry websites) and disease advocacy groups are leading to more open access, especially in the domain of rare diseases.

Participant-centric initiatives (PCIs) using social media tools offer new ways to engage with research participants[23]. By enabling on-going communication, individuals can give consent to research, specify levels of personal privacy and become partners in the research process. By providing control over personal information and the potential to give on-going consent in real time, these initiatives meet international legal standards for the protection of privacy. In the US, a "Trust Framework" for protecting privacy has been developed and advocated[24]. Active engagement with the public and relevant governmental and regulatory officials will be needed to encourage the use of PCI and promote beneficial research while providing adequate privacy protections. In the long term, greater transparency in data handling, commensurate punishment for mishandling of data, and transparent governance that includes public input is needed.

Access to Data
Who gets access to data, and how efficient that access will be, is both crucial and difficult. Scientific considerations argue for widespread and facile access by researchers to vast collections of data; respect for research participants requires strict adherence to the conditions of access they have agreed to and to all applicable law. A monitoring system will be needed to assure compliance with consents, law, and best ethical practices for realizing the expressed wishes of research participants. A platform and regulatory architecture that centrally establishes and verifies researcher *bona fides* is technologically possible and could enable controlled data sharing that is both efficient and secure. Where access to electronic medical records and health administrative databases is involved, a second tier of complexity will enter.

A separate issue is the nature and amount of access to their own data provided to research participants. This will be on a landscape of shifting public expectations about active patient participation in their own care and biomedical life planning. Current best practice in research is that research participants and the general public should have access to a description in lay language of approved projects and of published results. Policies that allow the return of individual research results from genomic studies in an ethical way through the healthcare system are being developed (e.g. UK 10K) and such information exchange could be integrated into PCI approaches to augment existing clinical and research practice.

International Data Sharing
Even where local consent and ethics approval allows data sharing, providing data to researchers in other institutions and countries often requires additional approvals (even when the foreign researchers intend to use the data in a protocol approved by their own local ethics committee). Despite "safe harbor" rules, the problem is already acute in international research consortia. Centralized access mechanisms that verify researcher *bona fides* and accompanying institutional approvals (including security mechanisms) are technologically possible and offer a path forward for more efficient secure data sharing.

Recontact
Re-contact of individuals may be considered to obtain additional information or tissue, or for the purpose of returning specific results and incidental findings. Each raises distinct issues.

Ideally, when data and samples are collected, participants are informed of, or asked to consent to, being re-contacted for additional information and/or return of results. Many legacy collections are silent on this issue.

In the absence of specific consent, the IRB/ERC must consider and approve any re-contact. If the request is for gathering additional information, and the IRB does not allow re-contact, the IRB/ERC may, in some circumstances, allow access to medical records under a waiver of informed consent. It is important that for future collections of data and samples, the informed consent form and process must include details regarding if, how and for what purpose re-contact would be initiated. PCI would hopefully facilitate ongoing relationships with and communication between participants and researchers.

Governance of repositories
While the goal of secure and ethical data sharing prompts investments in interoperability, it does not require that data physically reside in a single database. Rather, it is necessary only that data be processed in manners that are comparable across repositories, and communicated using generally understood procedures with regard to security, formats, error models, annotations, and the like.

Members of the global alliance and local operating entities will need to interact with regulatory bodies in each jurisdiction to understand regulatory responsibilities (of individuals, institutions

and governments), and to feed back to the alliance as requirements whatever is needed to ensure compliance with local rules. The alliance can contribute by sharing information from other jurisdictions, and by assuring stakeholders (where it is the case) that local, national, and international regulations have been taken into account through careful policy development and system architecture. The involvement of all stakeholders is needed in the development of governance structures to ensure that they are fit for purpose and address stakeholder concerns.

If used also to support clinical practice and research that takes place in the clinical setting, technical platforms will need to be designed and administered so as to meet the appropriate clinical laboratory standards. For example, in the USA it will need to follow the requirements of the Clinical Laboratory Improvement Amendments (CLIA)[25]; in the UK, pathology and laboratory medicine services need be registered with an approved UK based laboratory accreditation body.

A key issue will be the stewardship and governance of any repositories built by operating entities: as the report of the NHGRI meeting on data aggregation held in July, 2012 writes:

> *Enhanced methods for data stewardship and governance of data repositories are needed. As data sharing increases in scope, research participants are no longer asked to consent to a single study, but rather to make their data available to a large number of researchers, likely from many different countries. Public trust in the procedures used to store and access data is essential. Transparency is needed about methods to ensure data security, policies used to approve researcher access to data, auditing methods to ensure compliance with policies, and consequences that result from any missteps or compliance failures. Effective methods to ensure public input on each of these critical areas of governance are needed. As a component of transparency, communication is needed to inform the public about the scientific value of broad data sharing, without generating false expectations about the speed with which translation to health benefits will occur.*

The Role of the Global Alliance

The global alliance should aim to be a powerful unifying voice in the attempt to harmonize the international regulatory frameworks that oversee human participants research. At present, the approach is fragmented, and lacking action is unlikely to result in learning across jurisdictions. In addition, the alliance could (along with P3G and others) facilitate international coordination of research in ethical, legal and social aspects of genomics, as called for by Kaye, Meslin, Knoppers and Juengst (Developing a Global Vision for the Future of ELSI Research).

The global alliance should convene and support an ***Expert working group on regulatory and ethical issues*** to engage with stakeholders and to develop positions on issues such as:

- Harmonizing policies, procedures, standards and codes of conduct for the storage, analysis and responsible sharing of genomic data in clinical samples.
- Developing forward-looking consent procedures that responsibly engage patients, participants, and researchers in ways that lead to the most productive research while respecting the autonomy of each participant with regard to data s/he contributes.
- Developing a public engagement strategy to ensure the recruitment, active involvement and commitment of research participants from many different populations.
- Developing best-practices in governance and transparency of data repositories to build and maintain public trust in procedures used to secure and provide access to data.
- Making recommendations on the construction of a technical architecture that will be compliant with current rules, regulations, and laws governing research conduct, and the forward-looking support of possible new PCI approaches.

Technical considerations

The technology platforms will be secure, scalable, sustainable and built on open standards, creating the technology foundation for: (i) scalable upload and storage of data from sequencing platforms together with clinical data; (ii) rapid processing with state-of-the-art generic and custom tools; (iii) management of security, privacy and user access; and (iv) downloading and controlled sharing of data and results. The platforms will have standards and application programming interfaces (APIs) to securely interact with the data and results, enabling a wide variety of operating entities to serve users, and for developers to write third-party "apps" customized to specific uses. Like the WWW, the platform standards must be globally distributed, ubiquitously available, precisely defined, and sufficiently reliable so that both for-profit and not-for-profit organizations will be enabled to securely build upon it.

Background

For any individual medical or research center, major obstacles to building a platform for storing, analyzing and sharing genomic sequence data include: (i) lack of clarity about regulatory and policy standards for managing such data; (ii) high cost of storage required by large amounts of raw data in uncompressed formats with unoptimized code; (iii) highly variable demand for compute resources; (iv) limited local expertise in large-scale elastic computing and information technology for datasets of this type; (iv) inadequate standardization of formats and of analysis tools for genomic data; and (v) lack of local access to physical, network, system and data security along with associated regulatory compliance.

Several large technology companies – such as Amazon, Google, and Microsoft – previously encountered a similar set of challenges in their own businesses. They leveraged the expertise developed in serving their own large-scale computational needs by offering secure storage and computing as an on-demand service, known colloquially as *cloud computing* (see Box: Why Cloud Computing). Scientific and medical platforms such as Europe's Helix Nebula project are being designed to take advantage of services from multiple cloud platforms originating in Europe and the US to manage large scientific data sets[26]. Realizing enormous economies of scale, cloud services have been driving down the cost of computing while enabling large-scale, dynamic efforts such as those required herein.

This Section is informed by many discussions among the participants at the meeting on January 28, 2013, meeting and other experts, as well as at a meeting held in Santa Cruz, California on December 20, 2012 that included researchers from The Broad Institute, UC Berkeley, UC San Francisco, and UC Santa Cruz along with senior technical representatives from Amazon Web Services, Google, and Microsoft. The industry participants responded enthusiastically to the idea of a global alliance, indicating that common, shared computing standards are emerging that would promote interoperability. They suggested that it would *greatly facilitate progress* if they could interact with an organization that spoke with authority on standards, formats, APIs, and regulatory procedures. One such company volunteered to provide heavily discounted compute and storage if there were such a body.

At the meeting on January 28, it was resolved that no one cloud vendor would be exclusively engaged; rather that requirements would be specified so that platforms could be run on several services over multiple continents. In particular, it was agreed that data storage would need to be decentralized to comply with regulations specifying that some data must remain within specific national boundaries.

Below, we discuss key technical issues derived from the above requirements: (1) Scalable, cost-effective and distributed storage; (2) Rapid analysis; (3) Security; (4) Privacy; (5) Standards, APIs and Benchmarking; (6) Handling of clinical data; (7) Prototypical efforts.

Scalable, cost-effective and distributed storage

In order to support the needs of the biomedical ecosystem, the platform standards must enable analysis and storage of hundreds of thousands to millions of genomes, each with associated clinical information. One million whole genomes generated today, after compression, constitute ≈ 100 petabytes of data (1 petabyte = 1 million gigabytes). This is large but not unprecedented: YouTube has >1,000 petabytes of video, increasing by 100 petabytes every 2 months.

Through both discussions with information technology providers (offering cloud computing) and by independent estimates, we anticipate that with investment in compression and computational efficiencies, the cost of active data storage for one million whole genome datasets could be reduced to ≈$50/genome/year by 2014[27]. Using *archival* storage, which provides rapid archiving of data but a longer data retrieval time, the storage cost could drop by 10x, and is likely to drop further. In contrast, current storage costs in Cancer Genomics Hub (CGHub), which built its own storage infrastructure, are about $100/genome/year for primary storage and backup at its capacity of 50,000 genomes (including security, compliance, development, maintenance and operating costs).

Cloud storage is distributed for reasons of efficiency, elasticity and reliability. Major vendors already operate storage facilities in many different countries in order to provide service that complies with local regulations. Storage standards are widely adopted, facilitating transfers between different cloud services from both small and large vendors. Thus, a decentralized approach is preferred.

In summary, major providers of cloud computing can readily support the scale of data needed. Engaging with these providers offers major benefits in costs and scalability.

Box 2: Why Consider Cloud Computing?

Low cost: Cloud providers reduce cost by (1) employing commodity hardware, (2) purchasing in bulk, (3) building secure data warehouses in economical locations, and (4) deploying software which allows a handful of people to efficiently monitor each warehouse.

Bulk storage: Newer archival storage systems offered by several cloud providers can save 90% over an institution's in-house storage costs.

Efficient processing: By distributing the data across many low-cost computers, efficient processing can be performed without transferring data over the network. In this way, large volumes of data can be processed in ways not previously possible.

Security: Cloud-based encryption, firewalls, extensive auditing capabilities and other features have been developed to support a huge volume of online commerce. These measures are available to help secure health and genomic data.

Elasticity: The variability and uncertainty in compute and storage needs of platforms makes cloud-based solutions the most cost effective. Cloud cost is proportional to use and scales with demand over short time periods.

Rapid analysis

The platforms will need to enable rapid generation of analysis results (in particular for clinical samples) on schedules that may be unpredictable. It will need to support benchmarking jobs and large-scale comparisons of many thousands of genomes within a reasonable time. Achieving acceptable turnaround on a variety of different kinds of tasks such as these, while maintaining reasonable cost will require an ***elastic service***. For example, on the existing global commercial cloud services, it is possible to request a large number of computers for a short period of time and then release them, paying only for use. To realize the benefit of elastic cloud computing, current sequence processing pipelines must be redesigned to leverage existing frameworks, such as MapReduce, Hadoop[28], and Spark[29], that utilize a large number of computers in parallel. Many of these tools were developed by cloud service providers to perform their own computational tasks and then made broadly available.

The infrastructures offered by cloud providers are well suited to the large-scale and dynamic computational needs of platforms.

Security

Above all else, the standards must enable security and engender trust. Protecting information against unauthorized access is of primary concern for human participants data, whether managed in an on-premise data center or using infrastructure hosted in the cloud. This requires effective security controls on all aspects of a computational system, from control of physical access, through all aspects of network, system and data security.

Experience shows that few in-house data centers are as well protected as commercial computing facilities. The global Cloud Security Alliance provides best practices for cloud security internationally, participating in Europe's Helix Nebula science cloud, for example[30].

The US government has laid out a suite of risk-based security controls (FISMA[31]) that serves as a basis for the assessment of the security of IT systems[32] mandated for governmental organizations. Amazon Web Service and Google App Engine are accredited under FISMA[33].

As a result of the cloud providers' dedication to security, many organizations with data of high financial value have made the decision to store their data in the cloud. For example, the US NASDAQ financial exchange stores the private data on its largest traders in the "Fin" cloud on Amazon Web Services[34]. The Chief Technology officer of the Central Intelligence Agency in the US has said that cloud services could be leveraged to build a repository that is even more secure than CIA's own system[35].

Thus, after first expressing concern, many organizations have realized that the largest security risk is often human and internal, and that it can be safer to have a trusted and highly experienced third party secure data remotely. With respect to the privacy of personal data, the credit card industry has successfully promulgated a suite of data security standards specifically addressing the need to transmit and store personally sensitive information on systems connected to the public Internet. These standards underlie much of today's online economy, and are cloud-compatible.

Security controls implemented by major cloud providers, if matched by similarly rigorous controls in the development and operation of the platform standards, can meet requirements.

Privacy and access control

Perhaps the most critical aspect of security in the context of genomic data, as with medical record data, is privacy. The primary way privacy is ensured is by limiting data access to authorized users and by auditing all use. To achieve the goals of the global alliance, it will be necessary that patients, doctors, clinical trial investigators, researchers and medical centers have fine-grained control over who gets access to different parts of the data, and must be able to change their elections at any time. *Fine-grained access control* with audit requires appropriate data disaggregation, tagging, provenance, and versioning. Further, with a global focus, the platform standards will have to address the diversity of international data privacy regulations as discussed in the preceding section**.**

While privacy issues entail considerable effort, from an informatics perspective, there are well-established practices for putting both auditing and flexible access control in place[36].

Managing identities and consent for participants and researchers

Current systems for informed consent operate at a single moment in time and at a single institution. In the service of protecting privacy of research participants, they unintentionally segregate research data into legal and ethical silos, create barriers to discovering relationships across diseases and studies, and eliminate the possibility of asking questions unimagined when the study was designed.

Participant centric initiatives (PCI) such as the Portable Legal Consent are being developed in which consent to participate in research is a state associated with an individual rather than an aspect of a single study. Such an approach can in principle facilitate participant autonomy, making it possible for the consent to be transferred across organizations, updated based on new information and the wishes of the participant, or withdrawn.

PCI approaches could be powerfully enabling for an interoperable, distributed system for sharing genomic and clinical data, but achieving this goal will require specific technical capabilities. These could include standardized approaches to enable Institutional Review Boards (IRBs) to craft their own consent forms while retaining interoperability, and to transfer such information across institutions. In order for participants to remain connected to their data across research sites, a system of unique identifiers for participants (equivalent to IP addresses in the internet) may be needed. Similarly, in order to ensure that terms of data use are adhered to in a federated model, a system of researcher registration and unique identifiers, audit trails, will likely be needed. Transparency of governance would be enhanced by a web site that provides readily accessible summaries of past and current uses of the data.

Standards, APIs and Benchmarking

The global alliance will establish a technical working group to establish and enhance standards such as BAM (http://samtools.sourceforge.net/SAM1.pdf) and VCF (http://vcftools.sourceforge.net/) in genome storage and low-level genome analysis; to develop and promulgate standards and an application programming interface (API) for organized, higher-level access to the data for many different types of users; and to develop and manage a reference implementation in conjunction with the definition of standards. Close partnership with an initial set of operating entities who agree to be platform development partners and the broader community will be key, as standards are best developed along with a working implementation in an agile software development cycle. It will be important not to hold data hostage while standards are developed.

The platform standards should focus on a small number of critical, low-level genome analysis tasks and not attempt to build specifications or mechanisms for all types of genome interpretation and use cases. Building a powerful API can unleash the creativity of a world of developers to tackle the more sophisticated challenges of genome interpretation and medical translation. The platform standards should exist not to usurp the processes of engineering innovation and scientific discovery, but rather should enable them.

Benchmarking will be critical to develop robust systems, to identify best-in-class software products, and to encourage rapid software development and technical improvements. In the US auto industry, the development of an EPA gas mileage standard led to steady improvements in auto efficiency. In medicine, the development of care standards by the Joint Commission enables improved the standard of care across US medical centers[37]. Developing appropriate standards for sequence analysis and clinical data interpretation, in a similar fashion, will help identify algorithms and processes that provide the most accurate and informative results.

Platforms will support storage and analysis of data from all major sequencing technology manufacturers. Hence, special efforts will be required to set standards by incorporating consultation from a variety of manufacturers, and to manage compatibility. Standards related to data access, auditing and data provenance will be needed to understand and address the impact of data quality (e.g. malfunctioning sequencers), and to improve data collection and analysis techniques. The implementation of open standard APIs, versioned resources, and standard hardware configurations will foster the development of best practices and an open source community as it has in other IT sectors.

The goal of standards, APIs and benchmarking is to stand on each other's shoulders rather than each other's toes.

Information from clinical trials and clinical care

Platforms will need different approaches to support data used in clinical trials, or if ultimately used to support clinical care. For example, clinical trial data may need to remain closed under the regulations governing the trial until such time as it is released; to support this use case, the platform standards could provide electronic attribution and provenance to genetic and clinical data stored therein. Many parties are working to federate data from electronic medical records (EMR), and accommodate many different formats for clinical data; the alliance would engage with such efforts, rather than duplicating them. The alliance might lead by example in creating a unified international standard for genomic data, and work with others to add clinical information to that standard. This path will take considerable time, and demands a flexible database design. In the short run, much of the clinical data will be unstructured. As repositories with large amounts of unstructured data abound, it will be highly desirable to leverage advanced IT methods from other sectors – generally labeled "Big Data" techniques – to mine any unstructured clinical data that is used for research purposes.

Platforms should take advantage of the latest Big Data techniques in approaching clinical data, not wait for it to be completely standardized.

In summary, the technical challenge of building platforms to meet the needs of the global alliance at acceptable cost is substantial but surmountable. Distributed ("cloud") computing solutions that have been successful in other sectors may be adapted to this task as in other Big Data problems. Economic drivers will ensure the continued improvement of global cloud computing, and the resources employed for this development far exceed the funds we might hope to raise to develop an alternative platform. Therefore we should avail ourselves of existing methodology, expertise and infrastructure in Big Data analysis and cloud computing.

Box 3: Prototypical efforts

Several groups have built systems that demonstrate the practicality of some of the distributed (cloud) computing concepts proposed above:

The European Bioinformatics Institute has developed an in-house cloud for the public sequence repository where users can access large datasets without performing large downloads. They have a system for restricted access data sets such as the European Genome-Phenome Archive (EGA) using federated identity management and excellent experience working in a multi-language, multi-national setting.

The US National Center for Biotechnology Information manages dbGaP to store genomes generated in biomedical research, which like the EGA, is extensively used in biomedical research. They are studying possible cloud implementations.

The **Beijing Genomics Institute** has developed five bio-cloud computing centers in different locations that store and process genomes.

CGHub is the system for storage of genomic data generated by the large projects of the US National Cancer Institute. It is designed to address large storage needs but currently not implemented on the cloud.

Bionimbus is pioneering a collaborative infrastructure for working with genomic data within the Open Science Data Cloud.

The Broad Institute has instantiated its analysis pipeline for both germline and cancer somatic data on commercial cloud environments, integrating these tools with access management to enable oversight of user access and IRBs.

The AMP Lab at UC Berkeley has developed and is deploying its genome analysis pipeline on commercial cloud environments.

Illumina has created a cloud-based environment for sequence analysis called Basespace and its machines upload data directly to the cloud.

Other sequencing companies and many informatics start-ups provide tools for sequence analysis within a cloud environment.

Section 3: Next Steps

Next steps

A number of conclusions can be drawn from the preceding sections. First, the integration of genome sequence with clinical phenotype offers great potential for improvement in medical knowledge and human health. Second, varied perspectives and diverse regulatory protocols mandate a flexible rather than a "one size fits all" approach. Third, the benefits of data sharing and integration may be catalyzed by building public support on a foundation of harmonizing regulatory procedures and interoperable technology platforms with open standards.

In order to realize these benefits, we propose the creation of a **global alliance**: a non-profit organization that is inclusive of varied stakeholders, that convenes working groups to develop shared policies and procedures, that speaks clearly to the public as to the benefits and challenges of data sharing, that is effective in its operations, and is international in scope.

In order to develop technology standards and maximize interoperability, the global alliance will work with platform development partners and the broader community to create and then to manage **interoperable technology platforms with open standards**. The uses and users of platforms will be many and varied, but to ensure interoperability, shared standards must be rapidly established, widely used, and effectively managed.

The platform standards will be brought to life by a wide variety of operating entities that develop needed components and services, and that work with users to aggregate and analyze data, and to advance medicine and clinical care. These will be focused organizations – either existing or newly created – that contribute to and demonstrate the value of the platform standards, bringing world-class technical capabilities and operational effectiveness to serve the needs of the research and clinical communities.

The global alliance will be international and not-for-profit, and emerge as a *trusted and authoritative voice* to ensure that the field of genomic medicine develops to: (a) serve the needs of society, rather than solely commercial or academic interests; (b) support current and future technologies for generation of data and for distributed computing, (c) encourage innovation and diversity, engaging both non-profit and for profit stakeholders; and (d) remain an open resource to the research and clinical community, rather than serving the needs any particular entity.

Just as the World Wide Web and the Human Genome Project spurred the creation of innumerable and unanticipated applications, interoperable technology platforms with open standards for responsible data sharing will lead to the growth of an information-based ecosystem for the biomedical sciences.

Launching the Global Alliance

Purpose: the global alliance will bring together a wide range of partner organizations and experts who together develop, evaluate, and ultimately endorse *policy solutions* that balance facilitation of data integration with protection of privacy and the autonomy of individuals. The alliance will develop as a trusted voice in the international community on matters of sharing and privacy of genomic and clinical data. The alliance will use its role as a convener, and the public trust that it engenders, to ensure that the highest standards of ethical behaviour are followed, and that open technical standards are developed and deployed. The global alliance will promote the harmonization of regulatory frameworks across organizations and jurisdictions.

The global alliance will create an expert and trusted body that speeds progress in biomedicine by facilitating responsible and effective integration of genomic and clinical data.

Composition: Once established, the global alliance will consist of partner organizations as well as individuals. Organizations will be both non-profit and for-profit, and drawn from relevant sectors including biomedicine (e.g. health care delivery; academia; disease advocacy; pharmaceutical research and development, payers), genomic technology (e.g. sequencing platform companies), and information technology (e.g. data storage; internet security protocols and standards; large-scale elastic computing; search and data mining; social media).

Over the winter and spring of 2013; the conceptualization of the alliance has been led by a Organizing Committee with the close involvement and input of the participants at the January 28th meeting. These participants were drawn from the non-profit sectors (research, health care providers, funders, disease advocates), although informal discussions have gathered input from a number of for-profit parties. In the middle of 2013 we envision a phase in which interested organizations from all sectors will be offered the opportunity to sign a Letter of Intent to join as founding partner organizations of the alliance, ultimately signing a Memorandum of Understanding that establishes the alliance, defines the relationships and responsibilities of the parties, and provides a clear and transparent approach for other interested parties to join in the future.

Structure: We recommend that the global alliance be constituted (at least, initially) as a non-profit alliance, rather than as a stand-alone corporate entity. This structure enables the alliance to form rapidly, and avoids prematurely needing to engage in the complexity and rigidity that a corporate structure might demand. Precedent supports this model: unincorporated organizations such as the World Wide Web Consortium (W3C) and the International Cancer Genome Consortium (ICGC) have successfully coordinated activity across jurisdictions and sectors, with implementation performed by partner organizations that are engaged and supportive of core principles. The members of the alliance will be linked by a Memorandum of Understanding (MOU) and other forms of agreements that set out duties and obligations between various parties (partner organizations, host institution(s), operating entities, technology providers, etc.).

Governance: Over the planning and launch phase, the global alliance will be led by a transitional steering committee, growing out of the organizing committee of the January 28th meeting. Once the MOU is drafted and signed, we envision that the founding partner organizations will constitute an executive committee, with members drawn from partner organizations and different sectors (e.g. funding agencies, research institutions, healthcare organizations, disease advocacy and community-based organizations, life science and technology companies). The executive committee will be small enough (between 8-12 members) to be effective, and include individuals with diverse perspectives and commitment to the mission of the alliance. The executive committee will establish processes for selecting a chair, defining term limits, rotation of members, etc.

Organization and administration: The alliance will be comprised of a small full-time staff and a number of expert working groups. One or more "host" organizations will be selected to provide an administrative home for the alliance. (This is modeled on the W3C, which has four host organizations: MIT, ERCIM, Keio University, and Beihang University.) The "host" organization(s) will provide administrative and financial services on a contract basis.

Day-to-day management will be delegated to a full-time executive director who reports to the chair of the executive committee, and attends meetings as a non-voting member. The executive director will supervise an office and staff that support alliance functions such as organization of meetings, the standard setting process, and communication among the partners and with the broader community.

The global alliance will have distinct divisions, each with dedicated staff:

- Regulation, Law and Ethics
- Technology Standards and Platform
- Clinical and Phenotypic data
- Public Engagement

While these functions need be part of the same organization, their personnel and expertise is sufficiently distinct that each will need a strong leader and knowledgeable staff. The alliance will convene expert working groups (charged by the executive committee) that bring in world-leading experts in each of these areas. These working groups will work closely with and be supported by the staff of the alliance to develop perspectives, policies, standards and activities in their area. Other expert working groups will be formed, as needed.

Developing the platform standards: One of the critical functions of the global alliance will be to ensure that technology platforms are rapidly developed with initial reference implementation, broadly used and widely implemented, and are advanced and supported as the technology and science require. It will do this by convening an expert working group on technical standards, which will work closely with **platform development partners** drawn from the initial operating entities and other experts as needed to create an API and reference implementation.

The global alliance will manage a process that promotes the development of open high-quality standards, and will promulgate these standards for voluntary use by interested parties.

Relationships with operating entities: The global alliance will publish its core principles, policies and standards, with implementation and the provision of services left to operating entities. As technology platforms will have open standards, and the goal is to spark innovation, it is anticipated that many types of organizations (both existing and newly formed) may choose to develop tools or offer services. The alliance may choose to monitor whether and how operating entities are using the platform standards, and to publish recommendations of which entities conform to recommended practice. Moreover, only operating entities that join the alliance, and agree to support these principles, policies and standards, will have the chance to participate in the governance, standard setting and policy making of the alliance. Over the first year, it will be critical that the alliance work hand-in-hand with the first operating entities (as well as other stakeholders) to simultaneously develop the standards and platforms alongside the first reference implementation.

Because the platforms will be geographically distributed across jurisdictions, and due to the need and desire for local management and control, it is likely that many operating entities will ultimately be required. Operating entities will need to be well-managed and resourced enterprises, and offer all or some of the following: (a) technical capabilities for storage, processing, analysis and/or controlled sharing of information, (b) datasets provided by their partners and users; (c) management of data access and compliance with local informed consent and data use provisions; (d) an interface between users, cloud providers and external software developers; (e) hosting of portals that allow users (researchers, clinicians, patients) to access data and results, (f) management of withdrawals of data (as requested).

While any party will be free to borrow from the policies and standards promulgated by the alliance, the alliance will have a responsibility to ensure that some operating entities are fully compliant with and support the full set of core principles, policy advisories, and technical standards. To further this goal, *organizations will be allowed to join the alliance as operating entities only if they agree to a set of core principles and technical standards, and to subject themselves to review by the alliance.*

Funding: the global alliance will be funded by a variety of means that might include philanthropic support, grants from research and other funding agencies, and/or member dues.

Box 4: Incentives for responsible sharing of data

If technical challenges can be solved, and regulatory procedures harmonized, will participants, researchers and health care providers deposit data, develop tools, and share information? To create a vibrant and growing ecosystem, voluntary participation will be key.

Altruism can and must be a major incentive: the desire of stakeholders to advance medical knowledge for their families, communities and patients. The alliance should promote and advance altruistic sharing of data, while protecting the rights of those who choose to share. A second driver could be self-interest, as a growing repository of accessible information and analytic methods will *create a network effect*, with increasing utility with increasing numbers of users. *Economic* incentives could include preferential pricing (made possible by philanthropic donations and in-kind contributions) to those who contribute and share data.

Careful thought will be needed to encourage sharing by a wide variety of stakeholders while avoiding coercion or disadvantaging any segment of the ecosystem.

Draft mission statement, goals and core principles

Mission:

We are a global alliance of healthcare providers, research institutions, disease advocacy organizations, life science and information technology companies - dedicated to improving human health by maximizing the potential of integrating genome sequence and clinical information, while respecting and enabling the autonomy of participants.

Our Core Principles are:

- Respect – respecting the data sharing preferences, including privacy and the right to share data, as well as the autonomy of research partners,
- Transparency –maintaining transparent governance and operations, and communications with research participants and partners
- Accountability –developing, evaluating, communicating, and implementing best practices in matters of technology, ethics, and public outreach.
- Inclusivity – partnering, sharing, and building trust between stakeholders.
- Collaboration – sharing data and information to advance human health.
- Innovation –a vibrant ecosystem that draws on technological breakthroughs and advances to accelerate progress in life science and clinical medicine.
- Agility –acting swiftly to enable the aggregation of genomic and clinical data for the benefit of those affected by cancer and inherited disease

In order to achieve this mission and consistent with our principles we will:

- Support the ability of patients, clinicians and researchers to choose to share information in order to increase knowledge and improve patient outcomes.
- Seize the opportunity to learn from the integration of genomic and clinical data made possible by the arrival of low-cost genome sequencing.
- Drive the development and adoption of technology standards to effectively manage, protect, analyze and voluntarily share genomic data, clinical information, and analysis tools.
- Promote harmonization of regulatory frameworks, lower barriers to data sharing by developing policies for informed consents, align guidelines across jurisdictions, while respecting privacy and engaging individuals, families, and communities.
- Work closely with allied operating entities that commit to our core principles and implement interoperable information technology platforms to enable secure sharing and learning from genomic and clinical information.
- Lead by example, demonstrating how the international community can work together to share technology, data, and learning to make discoveries that advance human health and well being that would otherwise be impossible.

Participants, Contributors and Acknowledgements

Participants of the January 28th Meeting in New York City

Organizing Committee

David Altshuler	Broad Institute of Harvard and MIT, MGH
Peter Goodhand	Ontario Institute for Cancer Research
David Haussler	HHMI/University of California, Santa Cruz
Thomas Hudson	Ontario Institute for Cancer Research
Brad Margus	A-T Children's Project
Betsy Nabel*	Brigham and Women's Hospital
Charles Sawyers	HHMI / Memorial Sloan-Kettering
Michael Stratton*	Wellcome Trust Sanger Institute

Participants

Wylie Burke*	University of Washington
Martin Bobrow	University of Cambridge
Michael Boehnke	University of Michigan
Greg Brandeau	Former Chief Technology Officer, Disney and Pixar
Fabien Calvo	Institut National du Cancer
Vicki Chandler	Gordon & Betty Moore Foundation
Lynda Chin	MD Anderson Cancer Center
Dr. Guy Cochrane	EMBL-EBI
Francis Collins	National Institutes of Health (US)
Bob Darnell	Rockefeller University, New York Genome Center
Kay Davies	University of Oxford
Sue Desmond-Hellmann	University of California, San Francisco
James R. Downing	St. Jude Children's Research Hospital
Michael Dunn	Wellcome Trust
Sean Eddy	HHMI Janelia Farm Research Campus
Tom Freedman	Freedman Consulting, LLC
Stephen Friend	Sage Bionetworks
Richard A Gibbs	Baylor College of Medicine
Todd Golub	Broad Institute of Harvard and MIT, DFCI
Hank Greely	Stanford University School of Law
Leif Groop	Lund University
Mark Guyer	National Human Genome Research Institute (US)
Karin Jegalian	Science Writer, Freelance
Jane Kaye*	University of Oxford
Karen Kennedy	Wellcome Trust Sanger Institute
Bartha Knoppers	McGill University
Eric Lander	Broad Institute of Harvard and MIT
David J. Lipman	National Center for Biotechnology Information
Pierre Meulien	Genome Canada
Nicky Mulder	University of Cape Town
Arcadi Navarro	CRG (Centre de Regulació Genòmica)
Pearl O'Rourke*	Partners Healthcare
Andy Palmer	Koa Lab
Aarno Palotie	Wellcome Trust Sanger Institute
Dave Patterson	University of California, Berkeley
Anthony Philippakis	Brigham and Women's Hospital
Herman A. Taylor, Jr	Jackson Heart Study

Sharon Terry	Genetic Alliance
Marc Tessier-Lavigne	Rockefeller University
Harold Varmus	National Cancer Institute (US)
Mark Walport	Wellcome Trust
John Wilbanks	Sage Bionetworks
Barbara Wold	Caltech

*Unable to attend in person (some participated by phone).

Acknowledgements

The following contributed ideas and suggestions that shaped the proposal described above:

Richard Barker	Centre for Advancement of Sustainable Medical Innovation
John Bell	University of Oxford`
Tim Berners-Lee	World Wide Web Consortium
Ewan Birney	European Bioinformatics Institute
Carlos Bustamante	Stanford University
Don Chalmers	University of Tasmania
Stephen Chanock	National Cancer Institute (US)
Mark Daly	Massachusetts General Hospital, Broad Institute
Mark DePristo	Broad Institute of Harvard and MIT
Mark Diekhans	University of California at Santa Cruz
Peter Donnelly	Oxford University
Richard Durbin	Wellcome Trust Sanger Institute
Paul Flicek	European Bioinformatics Institute
Gaddy Getz	Massachusetts General Hospital, Broad Institute
Ted Goldstein	University of California at Santa Cruz
Eric Green	National Human Genome Research Institute (US)
Tim Hubbard	Wellcome Trust Sanger Institute
George Komatsoulis	National Cancer Institute (US)
Kazuto Kato	Osaka University
Zak Kohane	Children's Hospital, Boston
Dominic Kwiatkowski	University of Oxford
Daniel MacArthur	Massachusetts General Hospital, Broad Institute
Elaine Mardis	Washington University in St. Louis
David Margulies	Children's Hospital, Boston
Mark McCarthy	Oxford University
Jill Mesirov	Broad Institute of Harvard and MIT
Jeff Murray	University of Iowa / Gates Foundation
Deborah Peel	Patient Privacy Rights
Margaret Sleeboom-Faulkner	University of Sussex
Taylor Sittler	University of California at San Francisco
Louis Staudt	National Cancer Institute
John Todd	University of Cambridge
Matthew Trunnell	Broad Institute of Harvard and MIT
Henry Yang (Yang Huanming)	BGI

Author contributions

Section 1 - Setting the Context:

David Altshuler, John Bell, Todd Golub, Peter Goodhand, Tom Hudson, David Haussler, Karen Kennedy, Eric Lander, Todd Golub, Brad Margus, John Wilbanks, Charles Sawyers, Anthony Philippakis, Martin Bobrow, Lynda Chin, Sharon Terry, Tom Freedman, Kyra Jennings, Sarah Olinger, Jane Kaye, and Bartha Knoppers

Section 2 – Regulatory, Ethical and Technical considerations:

Wylie Burke, Hank Greely, Jane Kaye, Bartha Knoppers, Pearl O'Rourke, Barbara Wold, Stacey Donnelly, Elizabeth Lawler, David Altshuler, Peter Goodhand, John Wilbanks

David Haussler, Gaddy Getz, David Patterson, Taylor Sittler, Matthew Trunnell, David Altshuler, Bill Bolosky, Mark DePristo, Mark Diekhans, Ted Goldstein, Jamie Kinney, Anthony Philippakis, Paul Flicek

Section 3 – Next steps:

Tom Hudson, David Altshuler, Peter Goodhand, Brad Margus, Betsy Nabel, Charles Sawyers, Martin Bobrow, Karen Kennedy, Eric Lander, Todd Golub, Francis Collins, Harold Varmus, Sharon Terry, David Patterson, Kay Davies, Pearl O'Rourke, Barbara Wold, and Pierre Meulien

Overall editing and integration
Peter Goodhand, Marian Orfeo, and David Altshuler

References

[1] **"The Globalization of Personal Data Project: An International Survey on Privacy and Surveillance"** *The Surveillance Project, Queen's University,* Chan, Yolande E.; Stalker, L. Lynda Harling; Lyon, David; Pavlov, Andrey; Sharpe, Joan; Smith, Emily; Trottier, Daniel; Zureik, Elia, Eric Brousseau Graphic Arts, November 2008. http://www.sscqueens.org/projects/gpd

[2] **"Biotechnology"** *European Commission,* Conducted by TNS Opinion and Social, co-ordinated by Directorate General Research, October 2010, Pg. 148, 152. http://ec.europa.eu/public_opinion/archives/ebs/ebs_341_en.pdf

[3] **"The Globalization of Personal Data Project: An International Survey on Privacy and Surveillance"** *The Surveillance Project, Queen's University,* Chan, Yolande E.; Stalker, L. Lynda Harling; Lyon, David; Pavlov, Andrey; Sharpe, Joan; Smith, Emily; Trottier, Daniel; Zureik, Elia, Eric Brousseau Graphic Arts, November 2008, Pg. 5. http://www.sscqueens.org/projects/gpd

[4] **"Results of a New Public Opinion Poll"** *Research America: An Alliance for Discoveries in Health,* In partnership with JZ Analytics, December 2012, Pg. 11. http://www.researchamerica.org/poll_history

[5] **"Technology Optimism or Pessimism: How Trust in Science Shapes Policy Attitudes toward Genomic Science"** *The Brookings Institution,* Hochschild, Jennifer; Crabill, Alex; Sen, Maya, December 2012, Pg. 8. http://www.brookings.edu/research/papers/2012/12/genomic-science

[6] **"Europeans and Biotechnology in 2005: Patterns and Trends"** *European Commission,* Gaskell, George; Stares, Sally; Allansdottir, Agnes; Allum, Nick; Corchero, Cristina; Fischler, Claude; Jurgen, Hampel; Jackson, Jonathan; Kronberger, Nicole; Mejlgaard, Niels; Revuelta, Gemma; Schreiner, Camilla; Torgersen, Helge; Wagner, Wolfgang, July 2006. Pg 51. http://ec.europa.eu/public_opinion/archives/ebs/ebs_244b_en.pdf

[7] **"Biotechnology"** *European Commission,* Conducted by TNS Opinion and Social, co-ordinated by Directorate General Research, October 2010, Pg. 147. http://ec.europa.eu/public_opinion/archives/ebs/ebs_341_en.pdf

[8] "**How the public views privacy and health research"** Survey conducted by Alan Westin (Columbia University) and Harris Interactive. http://patientprivacyrights.org/media/Westin_IOM_Srvy_Rept_2008_.doc.

[9] "**2012 Most Trusted Companies for Privacy"** *Ponemon Institute,* Independently conducted by Ponemon Institute LLC, January 28, 2013, Pg. 11. http://www.ponemon.org/library/2012-most-trusted-companies-for-privacy-1

[10] **"Attitudes on Data Protection and Electronic Identity in the European Union"** *The European Commission,* Conducted by TNS Opinion and Social at the request of Directorate-General Justice, Information Society & Media and Joint Research Centre, co-ordinated by Directorate General Communication, June 2011, Pg. 194. http://ec.europa.eu/public_opinion/archives/ebs/ebs_359_en.pdf

[11] **"DNA data sharing: research participants' perspectives"** *Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas*, 2008 Jan;10(1):46-53. doi: 10.1097/GIM.0b013e31815f1e00, McGuire AL; Hamilton JA; Lunstroth R; McCullough LB; Goldman A., Pg. 1. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2767246/pdf/nihms-151711.pdf

[12] **"To share or not to share: a randomized trial of consent for data sharing in genome research"** *Center for Medical Ethics and Health Policy, Baylor College of Medicine, Houston, Texas,* 2011 Nov;13(11):948-55. doi:0.1097/GIM.0b013e3182227589, McGuire AL; Oliver JM; Slashinski MJ; Graves JL; Wang T; Kelly PA; Fisher W; Lau CC; Goss J; Okcu M; Treadwell-Deering D; Goldman AM; Noebels JL; Hilsenbeck SG., Pg. 2, http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3203320/pdf/nihms301465.pdf

[13] **"The Globalization of Personal Data Project: An International Survey on Privacy and Surveillance"** *The Surveillance Project, Queen's University,* Chan, Yolande E.; Stalker, L. Lynda Harling; Lyon, David; Pavlov, Andrey; Sharpe, Joan; Smith, Emily; Trottier, Daniel; Zureik, Elia, Eric Brousseau Graphic Arts, November 2008, Pg. 13, http://www.sscqueens.org/projects/gpd

[14] "**2012 Most Trusted Companies for Privacy"** *Ponemon Institute,* Independently conducted by Ponemon Institute LLC, January 28, 2013, Pg. 1, http://www.ponemon.org/library/2012-most-trusted-companies-for-privacy-1

[15] **"Ethical and practical challenges of sharing data from genome-wide association studies: The eMERGE Consortium experience"** *Genome Research,* 2011 July; 21(7): 1001–1007, Provided courtesy of Cold Spring Harbor Laboratory Press on the US National Library of Medicine National Institutes of Health, McGuire, Amy L.; Basford, Melissa; Dressler, Lynn G.; Fullerton, Stephanie M.; Koenig, Barbara A.; Li, Rongling; McCarty, Cathy A.; Ramos, Erin; Smith, Maureen E., Somkin, Carol P.; Waudby, Carol; Wolf, Wendy A.; Clayton, Ellen Wright, Pg. 5. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3129243/pdf/1001.pdf

[16] "**2012 Most Trusted Companies for Privacy"** *Ponemon Institute,* Independently conducted by Ponemon Institute LLC, January 28, 2013, Pg. 11. http://www.ponemon.org/library/2012-most-trusted-companies-for-privacy-1

[17] Kaye et al, Science 2012: 336 pp. 673-674.

[18] World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects. Amended 2008. http://www.wma.net/en/20activities/10ethics/10helsinki/

[19] The Belmont Report | HHS.gov." 2011. <http://www.hhs.gov/ohrp/humansubjects/guidance/belmont.html>

[20] Council of Europe - ETS no. 164 - Convention for the Protection of ..." 2003. <http://conventions.coe.int/Treaty/en/Treaties/Html/164.htm>

[21] <http://portal.unesco.org/en/ev.php-URL_ID=13177&URL_DO=DO_TOPIC&URL_SECTION=201.html>

[22] http://ec.europa.eu/commission_2010-2014/reding/pdf/m13_4_en.pdf

[23] Kaye J. et.al Participant-centric Initiatives *Nature Review Genetics* 2012

[24] http://patientprivacyrights.org/trust-framework/

[25] CLIA - <http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfclia/search.cfm>

[26] http://helix-nebula.eu/

[27] "A Million Cancer Genome Warehouse | EECS at UC Berkeley." 2012. 17 Jan. 2013 <http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-211.html>

[28] http://hadoop.apache.org/

[29] http://spark-project.org/

[30] https://cloudsecurityalliance.org/, http://helix-nebula.eu

[31] The Federal Information Security Management Act of 2002 and associated standards.

[32] ISO/IEC 27001 is a similar security framework commonly adopted by commercial organizations.

[33] http://aws.amazon.com/about-aws/whats-new/2011/09/15/aws-fisma-moderate/, http://www.google.com/enterprise/apps/government/benefits.html?section=security

[34] http://ir.nasdaqomx.com/releasedetail.cfm?ReleaseID=709164

[35] http://www.informationweek.com/government/cloud-saas/cia-cloud-solving-our-petascale-data-pro/231901640

[36] Boxwala, Aziz A., et al. "Using statistical and machine learning to help institutions detect suspicious access to electronic health records." *Journal of the American Medical Informatics Association* 18.4 (2011): 498-505.

[37] http://www.jointcommission.org/standards_information/standards.aspx