Title

Facilitating a culture of responsible and effective sharing of cancer genome data

Journal

Nature Medicine, 22(5)

ISSN

1078-8956

Authors

Siu, Lillian L

Lawler, Mark

Haussler, David

et al.

Publication Date

2016-05-01

DOI

10.1038/nm.4089

Peer reviewed

HHS Public Access

Author manuscript

Nat Med. Author manuscript; available in PMC 2016 August 24.

Published in final edited form as:

Nat Med. 2016 May 5; 22(5): 464–471. doi:10.1038/nm.4089.

Facilitating a Culture of Responsible and Effective Sharing of Cancer Genome Data

Lillian L. Siu[1,*], **Mark Lawler**[2,*,^], **David Haussler**[3], **Bartha Maria Knoppers**[4], **Jeremy Lewin**[1], **Daniel J. Vis**[5], **Rachel G. Liao**[6], **Fabrice Andre**[7], **Ian Banks**[8], **J. Carl Barrett**[9], **Carlos Caldas**[10], **Anamaria Aranha Camargo**[11], **Rebecca C. Fitzgerald**[10], **Mao Mao**[12], **John E. Mattison**[13], **William Pao**[14], **William R. Sellers**[15], **Patrick Sullivan**[16], **Bin Tean Teh**[17], **Robyn Ward**[18], **Jean Claude ZenKlusen**[19], **Charles L Sawyers**[20], and **Emile E. Voest**[5]

[1]Princess Margaret Cancer Centre, University of Toronto, Toronto, Canada [2]Centre for Cancer Research and Cell Biology, Queen's University, Belfast, UK [3]UC Santa Cruz Genomics Institute, University of California, Santa Cruz, CA, USA [4]Centre of Genomics and Policy, McGill University, Montreal, Canada [5]The Netherlands Cancer Institute, Amsterdam, the Netherlands [6]The Global Alliance for Genomics and Health, Toronto, Canada and the Broad Institute, Cambridge, MA, USA [7]Gustave Roussy and Université Paris Sud, Villejuif, France [8]Patient's Advocacy Committee, European Cancer Organization, Brussels, Belgium [9]Translational Sciences, Oncology iMED, AstraZeneca, Waltham, MA, USA [10]University of Cambridge, Cambridge, UK [11]Hospital Sírio-Libanês, São Paulo, Brazil [12]Yonsei Cancer Research Institute, Yonsei University College of Medicine, Seoul, Korea [13]Kaiser Permanente, Pasadena, CA, USA [14]Roche Innovation Center Basel, Pharma Research and Early Development, Roche, Basel, Switzerland [15]Novartis Institutes for Biomedical Research, Cambridge, MA, USA [16]Advocacy for Canadian Children Oncology Network, Vancouver, Canada [17]National Cancer Centre Singapore, Singapore [18]University of Queensland, St. Lucia, Australia [19]The Cancer Genome Atlas, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA [20]Memorial Sloan Kettering Cancer Center, New York, NY, USA

Abstract

Rapid and affordable tumor molecular profiling has led to an explosion of clinical and genomic data poised to enhance diagnosis, prognostication and treatment of cancer. A critical point has now been reached where analysis and storage of annotated clinical and genomic information in unconnected silos will stall the advancement of precision cancer care. Information systems must be harmonized to overcome the multiple technical and logistical barriers for data sharing. Against this backdrop, the Global Alliance for Genomic Health (GA4GH) was established in 2013 to create a common framework that enables responsible, voluntary, and secure sharing of clinical and genomic data. This Perspective from the GA4GH Clinical Working Group Cancer Task Team highlights the data aggregation challenges faced by the field, suggests potential collaborative solutions, and describes how GA4GH can catalyze a harmonized data sharing culture.

^Corresponding Author: Mark Lawler, Centre for Cancer Research and Cell Biology, Queen's University Belfast, CCRCB, Lisburn Road 97, Belfast, BT9 7AE, United Kingdom, Phone: 44-28-9097-2931, Fax: 44-28-9097-2776, mark.lawler@qub.ac.uk.
*These authors contributed equally

INTRODUCTION

There is broad consensus that identifying aberrations in tumor DNA is key not only to a better understanding of cancer but also to improved selection of patients for specific treatments. The latter is embraced by patients and their oncologists because it holds the promise of improving therapeutic outcomes through precision medicine, and by payers and governments because of its potential to reduce healthcare costs. Several governments and government sponsored initiatives have recognized that linking and sharing clinical information and genomic knowledge are key requisites for delivering 21st century cancer care (www.whitehouse.gov/blog/2015/01/20/watch-president-obamas-2015-state-union), including the recently announced "Moonshot" effort to cure cancer endorsed by US President Obama[1]. Examples of initiatives with a clinico-genomic data sharing aspiration include the US-based Precision Medicine Initiative and the UK's 100,000 Genomes Project, both of which have cancer as a major focus of their activities[2,3]. The goal of these and other projects is to show how a genomically-informed understanding of diseases like cancer can transform patient care.

To this end, many institutions worldwide have developed cancer molecular profiling initiatives to identify relevant biological drivers and use this information to inform biomarker-guided clinical trials. These initiatives, coupled with the increased utility of Next Generation Sequencing (NGS) and its ever-reducing cost, have fueled an unprecedented expansion of genomic data generated from cancer patients. However, these efforts typically occur at an institutional level, or are compartmentalized within disease-specific activities[4,5]. Analysis and storage of annotated genomic data in such isolated "silos" prevents collective data curation and sharing, making analysis of phenotype-genotype relationship prone to inconsistent interpretations, especially for low frequency variants, due to the use of different bioinformatics algorithms. A global unified approach is required to maximize our capability to recognize biological patterns between groups of patients, whose information may currently reside in different databases or institutions and use this knowledge to drive preventative or therapeutic interventions. The benefits afforded by data aggregation are substantial and would address a number of scenarios that are currently encountered by the cancer community (Box 1).

The sharing of aggregated data has thus become a substantial rate-limiting step in developing new cancer prevention and treatment strategies. This has implications not only for patient populations with uncommon histologies or rare phenotypes, but is increasingly relevant for cancer treatment in general, given the rise in molecular stratification of patients with common malignancies into smaller groups to tailor their treatment, either through drug repurposing or innovative precision medicine protocols.

Although the importance of open access to genomic information is clearly recognized[6-9], multiple technical and logistical barriers for effective data sharing persist, including data non-comparability, coding heterogeneity, difficulties in storage and transfer of large data sets, and non-standardized bioinformatics analyses. Additionally, regulatory, legal and ethics processes are not designed for global data sharing and require urgent attention. In this fragmented landscape, the Global Alliance for Genomics and Health (GA4GH) was

established in 2013 with a vision to promote responsible and effective sharing of genomic and clinical data and to transfer the benefits of this “team science” approach directly to patients (Box 2). In this Perspective, we highlight the challenges that a global clinical and genomic data sharing approach presents, suggest potential solutions, and highlight key initiatives (some sponsored by GA4GH) fostering these activities in the molecular profiling landscape (Table 1).

CHALLENGES IN DATA SHARING

Recognizing the urgent need to generate and maximize the value of high-throughput molecular data in cancer, international efforts such as The Cancer Genome Atlas (TCGA)[10] and the International Cancer Genome Consortium (ICGC)[11] were established to unify genomics-driven research efforts. However, although these initiatives were groundbreaking and laid the foundation for future opportunities, maximizing the utility of data sharing can only be achieved when its scope is extended beyond information derived from tumor samples collected at a single time point without clinical correlates (as was the case with initiatives like TCGA and ICGC). Ideally, attendant clinical data would include a longitudinal series of samples with detailed clinical, genomic and pathological information[12]. The analysis of a single tumor sample per patient can cause researchers to inadvertently ignore the phenomena of tumor heterogeneity and clonal evolution, and obscures the dynamics of disease progression at both clinical and molecular levels[13]. In most cases, longitudinal data, also including clinical comorbidities, medications and environmental exposures, are required for a granular assessment that enables the identification of correlates for favorable or poor clinical outcome, and of patient-specific *de-novo* resistance mechanisms under treatment pressure. However, longitudinal data with detailed clinical and pathological information are more difficult to harmonize and share between institutions.

Clinical Data: Challenges and Potential Solutions

Data in most electronic health record (EHR) systems are not vetted for quality assurance and are not structured in a way that readily enables easy extraction. These problems are magnified when we attempt to extract and compare data across institutions, and become significant barriers to cross-border data sharing initiatives.

In contrast to rare diseases, where initiatives such as Human Phenotype Ontology[14,15], Phenotips[16] and PhenoDB[17] have underpinned the development of a standardized human phenotypic ontology, a universally accepted lexicon is currently lacking in the cancer field. Tools such as PheWAS (developed from the Vanderbilt-led Phenome-Wide Association Study) were developed to allow unbiased interrogation of the EHR for detection of associations between a specific genetic variant and a wide range of clinical outcomes and phenotypes[18]. Although PheWAS or similar tools have thus far been mainly applied to germline genetic diseases, as in initiatives conducted by the Electronic Medical Records and Genomics (eMERGE) Network[19], these tools can potentially support similar approaches in somatic diseases such as cancer.

The tracking of longitudinal clinical outcome is crucial to linking clinical and molecular data for prognostic or predictive relevance, however efficacy outcomes (such as objective responses and time-to-disease-progression based on validated criteria, and overall survival), and toxicity information (as classified by the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events, NCI CTCAE), are not routinely captured in patient EHR outside the context of clinical trials.

Solutions to these clinical data challenges are at a less mature stage than are the solutions for genomic data, but are developing. There are emerging standards for representing data from EHRs in a way that can be shared between institutions. Leading among these is the international Fast Healthcare Interoperability Resources effort (FHIR) (www.hl7.org/fhir/), developing in conjunction with the Health Level 7 International (HL7) infrastructure. Technical tools are now emerging that use standards such as FHIR to federate data from EHRs in a functional way that can perform aggregation, cleaning and parsing of data longitudinally over time and from multiple disparate sources. It is critical that in such activities, the quality of the merged data is assured and controlled. A key example of such an effort is the American Society of Clinical Oncology (ASCO)'s CancerLinQ[20], a system that is being custom-built to gather data through direct electronic feeds from numerous oncology practices. CancerLinQ aims to measure, monitor and learn through the analysis of pooled information to improve the quality of cancer care and to provide clinical decision support.

The tools described above require well-developed and widely accepted ontologies or vocabularies to standardize the classification of diseases, such as the International Classification of Diseases (ICD) (www.who.int/classifications/icd/en/) or Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) (www.nlm.nih.gov/snomed/). Recognizing the lack of standardized phenotypic variation descriptors in malignancy, particularly in the genomic era, a Task Team of the GA4GH Clinical WG is developing approaches to support alignment and mapping across ontologies in cancer, leveraging specialist resources such as those provided by the National Cancer Institute Thesaurus[21] and the Human Phenotype Ontology[14,15].

Genomic Data: Challenges and Potential Solutions

Genomic data sharing in cancer has been successful within large research consortia such as TCGA and ICGC. Databases such as the Cancer Genomics Hub (CGHub), the European Genome-Phenome Archive, and the ICGC data portal provide cancer genomics data to researchers at a rate of multiple petabytes per month, representing the largest exchange of genomic information in any area of research[22]. New databases, such as the Genome Data Commons (GDC) of the NCI (USA) (created as part of the development of a Precision Cancer Medicine knowledge system; www.cancer.gov/news-events/press-releases/2014/GenomicDataCommonsNewsNote) and the 100,000 Genomes Project (UK)[3] are being constructed. However, these systems are not designed to handle data generated at the scale of millions of samples, as is anticipated with widespread clinical application of NGS. This is an entirely new data engineering challenge.

Most cancer genomics data generated by clinical application are held separately in silos by different medical institutions or their contractors. This makes aggregated data analysis more

difficult. Because the data sets are very large, now reaching multiple petabytes (10^{15} bytes), a simple transmission of genome sequencing data between geographically remote repositories is increasingly unfeasible.

The aggregation problem is made more difficult by significant heterogeneity in the procedures for data collection, storage and representation. Problems of data size can be overcome by sharing only the mutation and gene expression information from the clinical samples, and not the raw data produced by sequencing machines. However, lack of consensus in the mutation calling process, the methods for gene expression quantification, and even in the data formats used to express this information hampers current aggregation efforts. Non-standardized *ad-hoc* functional annotation and lack of consensus on the clinical significance of genomic variants between institutions further limit their universal applicability to guide improved patient care[8]. In particular, there are no widely accepted definitions of driver mutations in cancer and "clinically actionable" results[23,24]. This represents a very significant barrier to the integration of clinical genomics into health care delivery.

Solutions presented by the GA4GH—Responding to these diverse challenges, collaborative efforts including GA4GH-enabled initiatives have proposed[25] and/or are implementing the following solutions. Recognizing the need to link diverse genomic data repositories, the GA4GH Data WG, in collaboration with initiatives such as the US National Institutes of Health (NIH) Big Data to Knowledge (BD2K) Center in Translational Genomics[26], are pioneering new standards and methods for sharing genomic information. They are developing a universal Application Programming Interface (API) that will facilitate creation of a global cohesive genome informatics ecosystem which maximizes data sharing at scale. Specific Task Teams within the GA4GH Data WG are implementing particular functionalities within the GA4GH API to allow more expressive and universal representation of genetic variation, gene and transcript expression, annotation of genomics features, and relationships between genotype and phenotype.

Furthermore, in order to facilitate harmonized mutation calling between institutions, ICGC and TCGA invited the cancer genomics and bioinformatics communities to work together to identify the best pipelines for the detection of mutations in DNA sequencing reads for cancer genomes[27]; this has led to the establishment of the ICGC-TCGA Dialogue for Reverse Engineering Assessments and Methods (DREAM) Somatic Mutation Calling Challenge ("the SMC-DNA Meta-Pipeline Challenge"), a crowd sourced benchmark of somatic mutation detection algorithms[28]. The Benchmarking Task Team of the GA4GH Data WG is working closely with the DREAM teams to identify the most effective algorithms for widespread use by the scientific/clinical community. The need to identify and share the best data analysis pipelines has also stimulated considerable work on so-called containerized computation in genomics, in which code that executes different programs for data processing, analysis and interpretation is more easily exchanged between different institutions and different computing environments. This would be analogous to the strategy used by the company Docker, in which shipping containers across the world's ports are standardized so that one set of machinery is sufficient at any port to handle any shipping container. In containerized computing, one type of packaging for data and programs allows

analysis of the data on all computing systems. The Containers and Workflows Task Team of the GA4GH WG is devoted to this area. Furthermore, containerized code that has been battle-tested in the DREAM challenges and by large consortium efforts is now being applied to clinical NGS analysis in cancer – a strategy proposed by groups such as the Next-Generation Sequencing Standardization of Clinical Testing II (Nex-StoCT II) informatics workgroup for the analysis of germline variations in disease[29].

With regard to the lack of a consensus on what constitutes an "actionable" mutation, GA4GH is driving the Actionable Cancer Genome Initiative (ACGI)[30]. The main goals of the ACGI are to identify a list of "actionable" genes in different cancers with canonical targetable mutations as well as rare variants of uncertain significance, and aggregate data related to these aberrations, their evidence-based curated actionability calls and phenotypic/clinical information (including longitudinal data), in a searchable format to enhance patient care.

Data Warehousing and Data Access Challenges

One question that members of GA4GH have considered in detail is whether the world's genomic and clinical information will reside in a single physical database, or be made available through a federated network spanning a series of interlinked data repositories in many countries. While both approaches have their supporters, GA4GH is investigating how a federated model (which may involve a relatively small number of large databases) can be organized so as to fulfil data warehousing requirements, while supporting improved data access for data consumers. In a federated system, some data are likely to be on commercial clouds (e.g. Amazon, Google, Microsoft or one of the 30 cloud providers in the Helix Nebula Marketplace associated with Europe's Helix Nebula Project; www.helix-nebula.eu/) and the rest on government clouds, private clouds, or other dedicated systems. The recent decision by the US NIH to allow private and commercial cloud computing solutions to be applied to the storage and analysis of the vast genomic data that is housed within its repository, the database of Genotypes and Phenotypes (dbGaP)[31], is a timely one. It opens up competition between different cloud solutions in genomics. Creation of a competitive market for such cloud solutions will enable secure and organized data storage at low cost, with sufficient elasticity to provide a dynamic platform that ensures rapid and efficient analysis of large datasets[32] (www.genomicsandhealth.org/working-groups/our-work/cloud-security).

Optimized interoperable technical standards are needed for analyzing data that are distributed across multiple sites as suggested above. Beyond the data harmonization challenges, there are also significant technical challenges in ensuring coordinated version control, data uniqueness and integrity, location transparency, harmony and efficiency in access procedures, privacy and security requirements, and maintaining compliance with institutional and legal regulations at regional, national and international levels.

Furthermore, in conjunction with the GA4GH's API-based standards efforts (as well as its file-based standards efforts), the Containers and Workflows Task Team of the GA4GH's Data WG is developing mechanisms that will allow a computational procedure to be ported to different institutions and run locally with reliably consistent results and minimal

customization required. This allows a single institution to perform complex analysis of large datasets at remote sites. On the other hand, when only smaller data items are needed from a remote site, these can be obtained by a simple Internet query, again using the API. A mechanism for queries of this type is being developed by the GA4GH Beacon Project (www.genomicsandhealth.org/work-products-demonstration-projects/beacon-project-0) (Table 1). Having a range of solutions like these to data aggregation and analysis problems is critical to the success of a federated system.

Ethical, Regulatory and Security Challenges

Even if technical challenges are addressed, global data sharing will require a significant shift in the traditional ethics framework, especially given the worldwide diversity in legal and regulatory requirements. We may have reached the limits of the current informed consent procedures (Box 3). Broad consent is a practical overarching solution, although this practice can be contentious, if not accompanied by proper governance[33]. Since informed consent is usually conducted as a once-off activity, respect for individual autonomy demands ongoing oversight to respect the trust of participants to the use of data and samples for “future unspecified research”. To ease these concerns, new variations of consent documents have been proposed (e.g. tiered or dynamic consent and open consent)[33-37] and novel governance models have been suggested[38]. What is important is that the consent model chosen corresponds to the nature of the study. Thus, a broad consent is particularly suited for longitudinal studies (e.g. UK Biobank; www.ukbiobank.ac.uk/) and open consent for those wishing to put their genome data in the public domain (e.g. Personal Genome Project; www.personalgenomes.org/).

However, consenting challenges remain. For example, implementing the rights of the individual to withdraw their archived data in an international study is not possible if data are anonymized[39]. In addition, the protection of privacy is even more challenging, given the unique identifying nature of genetic information[40-43]. Germline data collection (whether preplanned or incidentally detected) increases the complexity, if individuals have been promised to be re-contacted for notification of such findings. Complicating matters further, traditional national or institutional review boards may not have the expertise to assess the risks and compliance associated with international data sharing projects, and oversight systems to date are in many cases not adequately equipped for privacy breaches[44].

GA4GH has approached these complex issues under a fundamental human rights perspective, proposing and adopting a Framework for Responsible Sharing of Genomic and Health-Related Data that emphasizes both the right of all citizens to benefit from the advances of science and of scientists to be recognized for their work[45]. This approach is complementary to and bolsters traditional bioethics principles, but employing a legal human rights perspective embeds responsible clinico-genomic data sharing within a recognized and endorsed international legal framework, thus providing the environment for the ethics principles espoused by this Framework to be recognized and adhered to by all stakeholders. This Framework can also foster responsible data sharing more strongly than a traditional bioethics approach by offering legal protection in several areas, such as privacy; anti-discrimination and fair access; and procedural fairness[44].

Addressing the complex issues already highlighted requires an ethics and regulatory framework that fosters cross-border collaborative open data sharing. Given the significant legal and ethical variations between different countries/jurisdictions, harmonization is urgent. To that end, implementation of the GA4GH Framework[46] will enable responsible data sharing, while respecting individual rights. Adopting and adhering to the following GA4GH-enabled principles, policies and tools within the Framework provides a blueprint for addressing the complex ethical, legal and security issues outlined above.

As an overarching enabler to the activities and human rights aspirations outlined above, GA4GH has created and adopted the *"Framework for Responsible Sharing of Genomic and Health Related Data"*[46] to underpin all aspects of its genomic and clinical data sharing activities going forward. As part of this *Framework,* the GA4GH Regulatory and Ethics WG produced a *GA4GH Consent Policy*[47], balancing the need to respect the autonomous decision-making rights of the individual patient with the promotion of the common good of international genomic and health-related data sharing. Three consent tools[48] have been created for GA4GH by the Public Population Project in Genomics and Society: International Policy interoperability and data Access Clearinghouse (P3G-IPAC; www.p3g.org/ipac): (A) *Legacy Consent and International Data Sharing:* allows the adequacy of previously collected "legacy" consents (consents taken at the time of the study which may not have envisaged the complexity of future scientific use(s) of samples or data) to be addressed. (B) *Clauses for International Data Sharing:* provides advice/templates for researchers wishing to add clauses/addendums on international data sharing to their existing consent document(s). (C) *Generic International Data Sharing Prospective Consent Form*: provides an adaptable consent template for international data sharing for prospective studies.

GA4GH's Regulatory and Ethics WG has prepared a *Privacy and Security Policy*[49] that requires a proportionate approach involving the weighing of the real risks and benefits as well as a concordance of terms such as pseudonymized; de-identified; coded, etc. to address the "Babel" of nomenclature. A "safe harbor" mechanism for privacy protection[44] in cross-border sharing, which elucidates the criteria for mutually agreed-upon data protection principles has been published by GA4GH's Regulatory and Ethics WG[25]. To specifically address privacy and security mechanisms, the GA4GH Security WG has created a *Security Infrastructure Policy Paper*[50] which documents the standards and implementation practices for protecting the privacy and security of shared genomic and clinical data. Where the data are highly phenotypic, that is, with sufficient data elements that either alone or in combination with other information could serve to re-identify an individual, a form of controlled access approach may be the most appropriate, such as the one utilized by the Data Access Compliance Office in ICGC[7]. GA4GH is however considering the potential of a registered system of access for less sensitive data, as an intermediary tier between closed and open access.

Ensuring ethics compliance and responsible conduct by researchers is also extremely important. To this end, the GA4GH's Regulatory and Ethics WG has developed an accountability policy[51]. Additionally, the ethics associated with commercial usage and sale of aggregated anonymized data is unclear and will require consideration. Education of the

cancer community at large to ensure responsible use and sharing of clinical and genomic information is crucial[52].

CONCLUDING REMARKS

Responsible and effective sharing of genomic and clinical data that are generated from biospecimens is becoming increasingly important for patients (including cancer patients), allowing research discoveries to be rapidly applied for their benefit. Patients are actively pursuing approaches that ensure their rights to share information for the overall benefit of citizens and societies[53,54]. A European survey of 811 cancer patients, conducted in 2012, revealed that over 91% of patients wanted their samples to be retained for future research, with a significant number of patients also indicating that they would participate in biomarker testing to allow personalization of their treatment[55]. More recently, a survey of 100 breast cancer patients indicated that over 75% of patients would share de-identified data with researchers not involved in their care while 60% of patients were prepared to share identified data also[56]. As a follow up to these studies, GA4GH, in collaboration with a number of institutions and prominent patient advocacy groups, are currently developing a survey to measure specific attitudes of cancer patients to the sharing of genomic/clinical data.

Cancer patients are emphasizing that they are no longer passive recipients but increasingly active participants in both high quality research and its clinical adoption; these principles are enshrined in the European Cancer Patients Bill of Rights[57,58] which was launched in the European Parliament on World Cancer Day 2014. Significant challenges in relation to the privacy of data exist, particularly in Europe in the context of both the Clinical Trials Directive (www.ec.europa.eu/health/human-use/clinical-trials/directive/index_en.htm) and the recently approved General Data Protection Regulation[59]. However, patients are increasingly recognizing the value of genomics research and its clinical translation[60] and the need for responsible data sharing[61]. That said, issues such as discrimination, not only in terms of access to optimal quality care (including precision cancer care) but also in relation to socio-economic factors such as employment rights and availability of affordable insurance, must be adequately addressed with clear patient education and input[61,62], otherwise patients' enthusiasm for participating in genomics research and acting as advocates for responsible data sharing may waiver. Cancer patients generally have a positive attitude to sharing their data (www.free-the-data.org/); we need to ensure that this happens in a timely, responsible and effective manner, so that the value of this data in improving health care can be realized as rapidly as possible.

GA4GH is committed to engaging with key stakeholders including cancer patients, researchers and health care professionals to establish a globally effective genomic and clinical data sharing ecosystem that addresses the diverse challenges we have articulated in this Perspective (Figure 1). GA4GH's success in fostering "a coalition of the willing" within the international community, allied to its ability to develop and implement technical informatics solutions within a harmonized, secure ethical and legal Framework can help deliver a powerful, globally accessible clinico-genomic platform and foster an associated philosophy that supports data-driven advances for patients and societies.

References

1. McCarthy M. US president endorses “moonshot” effort to cure cancer. B M J. 2016; 352:i213.
2. Collins FS, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015; 372:793–795. [PubMed: 25635347]
3. Siva N. UK gears up to decode 100,000 genomes from NHS patients. Lancet. 2015; 385:103–104. [PubMed: 25540888]
4. Meric-Bernstam F, et al. Feasibility of Large-Scale Genomic Testing to Facilitate Enrollment Onto Genomically Matched Clinical Trials. J Clin Oncol. 2015; 33:2753–2762. [PubMed: 26014291]
5. Andre F, et al. Comparative genomic hybridisation array and DNA sequencing to direct treatment of metastatic breast cancer: a multicentre, prospective trial (SAFIR01/UNICANCER). Lancet Oncol. 2014; 15:267–274. [PubMed: 24508104]
6. Walport M, Brest P. Sharing research data to improve public health. Lancet. 2011; 377:537–539. [PubMed: 21216456]
7. Joly Y, Dove ES, Knoppers BM, Bobrow M, Chalmers D. Data sharing in the post-genomic world: the experience of the International Cancer Genome Consortium (ICGC) Data Access Compliance Office (DACO). PLoS Comput Biol. 2012; 8:e1002549. [PubMed: 22807659]
8. Rehm HL, et al. ClinGen--the Clinical Genome Resource. N Engl J Med. 2015; 372:2235–2242. [PubMed: 26014595]
9. National Institutes of Health Genomic Data Sharing Policy. Aug 27. 2014 https://gds.nih.gov/PDF/NIH_GDS_Policy.pdf
10. Cancer Genome Atlas Research, N. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–1120. [PubMed: 24071849]
11. International Cancer Genome, C. International network of cancer genome projects. Nature. 2010; 464:993–998. [PubMed: 20393554]
12. Aparicio S, Caldas C. The implications of clonal genome evolution for cancer medicine. N Engl J Med. 2013; 368:842–851. [PubMed: 23445095]
13. Alizadeh AA, et al. Toward understanding and exploiting tumor heterogeneity. Nat Med. 2015; 21:846–853. [PubMed: 26248267]
14. Kohler S, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014; 42:D966–974. [PubMed: 24217912]
15. Groza T, et al. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. Am J Hum Genet. 2015; 97:111–124. [PubMed: 26119816]
16. Girdea M, et al. PhenoTips: patient phenotyping software for clinical and research use. Hum Mutat. 2013; 34:1057–1065. [PubMed: 23636887]
17. Hamosh A, et al. PhenoDB: a new web-based tool for the collection, storage, and analysis of phenotypic features. Hum Mutat. 2013; 34:566–571. [PubMed: 23378291]
18. Denny JC, et al. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. Nat Biotechnol. 2013; 31:1102–1110. [PubMed: 24270849]
19. McCarty CA, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics. 2011; 4:13. [PubMed: 21269473]
20. Schilsky RL, Michels DL, Kearbey AH, Yu PP, Hudis CA. Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ. J Clin Oncol. 2014; 32:2373–2379. [PubMed: 24912897]
21. Sioutos N, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. Journal of biomedical informatics. 2007; 40:30–43. [PubMed: 16697710]
22. Stein LD, Knoppers BM, Campbell P, Getz G, Korbel JO. Data analysis: Create a cloud commons. Nature. 2015; 523:149–151. [PubMed: 26156357]
23. Van Allen EM, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med. 2014; 20:682–688. [PubMed: 24836576]

24. Andre F, et al. Prioritizing targets for precision cancer medicine. Ann Oncol. 2014; 25:2295–2303. [PubMed: 25344359]
25. Kosseim P, et al. Building a data sharing model for global genomic research. Genome Biol. 2014; 15:430. [PubMed: 25221857]
26. Paten B, et al. The NIH BD2K center for big data in translational genomics. Journal of the American Medical Informatics Association : JAMIA. 2015
27. Boutros PC, et al. Global optimization of somatic variant identification in cancer genomes with a global community challenge. Nat Genet. 2014; 46:318–319. [PubMed: 24675517]
28. Ewing AD, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. Nat Methods. 2015; 12:623–630. [PubMed: 25984700]
29. Gargis AS, et al. Good laboratory practice for clinical next-generation sequencing informatics pipelines. Nat Biotechnol. 2015; 33:689–693. [PubMed: 26154004]
30. Lawler M, et al. All the World's a Stage: Facilitating Discovery Science and Improved Cancer Care through the Global Alliance for Genomics and Health. Cancer Discov. 2015; 5:1133–1136. [PubMed: 26526696]
31. Tryka KA, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. 2014; 42:D975–979. [PubMed: 24297256]
32. Global Alliance for Genomics and Health: White Paper: creating a global alliance to enable responsible sharing of genomic and clinical data. Jun 3. 2013 https://genomicsandhealth.org/about-the-global-alliance/key-documents/white-paper-creating-global-alliance-enable-responsible-shar
33. Kaye, J. The tension between data sharing and the protection of privacy in genomics research. In: Mascalzoni, D., editor. Ethics Law and Governance of Biobanking. Vol. 14. Springer Netherlands; 2015. p. 101-120.
34. Haga SB, Beskow LM. Ethical, legal, and social implications of biobanks for genetics research. Adv Genet. 2008; 60:505–544. [PubMed: 18358331]
35. Wolf LE, Lo B. Untapped potential: IRB guidance for the ethical research use of stored biological materials. IRB. 2004; 26:1–8. [PubMed: 15449408]
36. Lunshof JE, Chadwick R, Vorhaus DB, Church GM. From genetic privacy to open consent. Nat Rev Genet. 2008; 9:406–411. [PubMed: 18379574]
37. Lolkema MP, et al. Ethical, legal, and counseling challenges surrounding the return of genetic results in oncology. J Clin Oncol. 2013; 31:1842–1848. [PubMed: 23589552]
38. O'Doherty KC, et al. From consent to institutions: designing adaptive governance for genomic biobanks. Soc Sci Med. 2011; 73:367–374. [PubMed: 21726926]
39. Zika E, Schulte In den Baumen T, Kaye J, Brand A, Ibarreta D. Sample, data use and protection in biobanking in Europe: legal issues. Pharmacogenomics. 2008; 9:773–781. [PubMed: 18518854]
40. Homer N, et al. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. PLoS Genet. 2008; 4:e1000167. [PubMed: 18769715]
41. Lin Z, Owen AB, Altman RB. Genetics. Genomic research and human subject privacy. Science. 2004; 305:183. [PubMed: 15247459]
42. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y. Identifying personal genomes by surname inference. Science. 2013; 339:321–324. [PubMed: 23329047]
43. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014; 15:409–421. [PubMed: 24805122]
44. Knoppers BM. International ethics harmonization and the global alliance for genomics and health. Genome Med. 2014; 6:13. [PubMed: 25031613]
45. Knoppers BM, Harris JR, Budin-Ljosne I, Dove ES. A human rights approach to an international code of conduct for genomic and clinical data sharing. Hum Genet. 2014; 133:895–903. [PubMed: 24573176]
46. Global Alliance for Genomics and Health: Framework for responsible sharing of genomic and health-related data. Sep 10. 2014 https://genomicsandhealth.org/node/6611
47. Global Alliance for Genomics and Health: Consent Policy. May 27. 2015 https://genomicsandhealth.org/consent-policy-pdf-27-may-2015

48. GA P3G-IPAC Consent Tools. Aug 6. 2014 https://genomicsandhealth.org/ga-p3g-ipac-consent-tools
49. Global Alliance for Genomics and Health: Privacy and Security Policy. May 26. 2015 https://genomicsandhealth.org/privacy-and-security-policy-pdf-26-may-2015
50. Global Alliance for Genomics and Health: Security Infrastructure: standards and implementation practices for protecting the privacy and security of shared genomic and clinical data. Mar 12. 2015 https://genomicsandhealth.org/security-infrastructure-version-11
51. Global Alliance for Genomics and Health: Accountability Policy. Feb 10. 2016 https://genomicsandhealth.org/ga4gh-accountability-policy
52. Christensen KD, et al. Are physicians prepared for whole genome sequencing? a qualitative analysis. Clinical genetics. 2016; 89:228–234. [PubMed: 26080898]
53. Pillai U, et al. Factors that may influence the willingness of cancer patients to consent for biobanking. Biopreserv Biobank. 2014; 12:409–414. [PubMed: 25496153]
54. Mancini J, et al. Consent for biobanking: assessing the understanding and views of cancer patients. J Natl Cancer Inst. 2011; 103:154–157. [PubMed: 21173381]
55. Tejpar S, et al. Awareness and understanding of stratified/personalized medicine in patients treated for cancer: a multi-national survey. Ann Oncol. 2012; 23(Supplement 9) Abstract 1382.
56. Rogith D, et al. Attitudes regarding privacy of genomic information in personalized cancer therapy. Journal of the American Medical Informatics Association : JAMIA. 2014; 21:e320–325. [PubMed: 24737606]
57. Lawler M, et al. A Bill of Rights for patients with cancer in Europe. Lancet Oncol. 2014; 15:258–260. [PubMed: 24503530]
58. Lawler M, et al. A catalyst for change: the European cancer Patient's Bill of Rights. Oncologist. 2014; 19:217–224. [PubMed: 24493667]
59. Proposal for a regulation of the European Parliament and of the council on the protection of individuals with regard to processing of personal data and on the free movement of such data (General Data Protection Regulation). Jan 25. 2012 http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf
60. Husedzinovic A, Ose D, Schickhardt C, Frohling S, Winkler EC. Stakeholders' perspectives on biobank-based genomic research: systematic review of the literature. Eur J Hum Genet. 2015
61. McGuire AL, et al. To share or not to share: a randomized trial of consent for data sharing in genome research. Genet Med. 2011; 13:948–955. [PubMed: 21785360]
62. Green RC, Lautenbach D, McGuire AL. GINA, genetic discrimination, and genomic medicine. N Engl J Med. 2015; 372:397–399. [PubMed: 25629736]
63. Lacombe D, et al. European perspective for effective cancer drug development. Nat Rev Clin Oncol. 2014; 11:492–498. [PubMed: 24935010]
64. Conley BA, Doroshow JH. Molecular analysis for therapy choice: NCI MATCH. Semin Oncol. 2014; 41:297–299. [PubMed: 25023344]
65. Schilsky RL. Implementing personalized cancer care. Nat Rev Clin Oncol. 2014; 11:432–438. [PubMed: 24687035]
66. Eggermont AM, et al. Cancer Core Europe: a consortium to address the cancer care-cancer research continuum challenge. Eur J Cancer. 2014; 50:2745–2746. [PubMed: 25263570]
67. American Association for Cancer Research: Project GENIE (Genomics Evidence Neoplasia Information Exchange). http://www.aacr.org/Documents/GENIE_InfoGraph.pdf

Box 1

Hypothetical Examples Illustrating Importance of Data Sharing

	Hypothetical Examples
1	A group from country A employed a targeted panel to assess a selected set of hotspot mutations in 50 genes and published these results based on analysis of 1,000 colorectal cancer patients. A group in country B, capable of whole exome sequencing, has identified that one of these hotspots, in the presence of another specific mutation, may adversely affect clinical outcome in colorectal cancer. The country B group would like to collaborate with the country A group to determine if remaining DNA samples and clinical data can be shared to confirm their hypothesis.
2	A research group from country A reported favorably on a variant in gene X which may predict response for drug Y. Their findings indicated that the variant is significantly associated with progression-free survival at 6 months after initiation of treatment. A group from country B investigated the same gene variant for the same drug, but found no statistical relations at 3 and 24 months after treatment initiation. A research group from country C now wants to perform a meta-analysis to determine if the findings of both trials are in agreement, but it requires the original and individualized clinical and genomic data from the groups in countries A and B.
3	Two large studies have recently been published suggesting that the use of drug Z may confer a better prognosis in breast cancer among patients with a common somatic variant. There are many large oncology practices worldwide that capture the use of drug Z in their patients' electronic health records. Can this information be collected, integrated and provide a reliable validation of this finding?
4	One patient out of 500 patients in a clinical trial in country A responded to Drug X, and this patient's tumor is known to harbor a rare germline variant. A large institution in country B is currently considering running a similar trial for the same drug. The sharing of the details of such incidental findings would have a significant bearing on the new trial.

Box 2

Description of GA4GH

GA4GH is a not-for-profit worldwide alliance of over 380 international stakeholders from 38 countries with a current focus on rare diseases and cancer[30], and an emerging interest in infectious disease. GA4GH operates through a series of Working Groups (WG): Data WG, Regulatory and Ethics WG, Security WG, and Clinical WG, developing initiatives, policies, recommendations and Application Program Interfaces (APIs) that promote and harmonize responsible and effective data sharing. While GA4GH produces recommendations, it does not seek to enforce data standards, but rather persuades potential stakeholders of the added value of a collaborative data sharing culture. The first plenary meeting of GA4GH stakeholders took place in March 2014 at the Wellcome Trust in London (United Kingdom), with subsequent conferences in San Diego (USA) and Leiden (the Netherlands). This Perspective was undertaken following discussions during the third GA4GH plenary meeting in Leiden, and in response to a number of international data sharing initiatives (including the recently launched American Association of Cancer Research (AACR)'s transatlantic data sharing project *GENIE* (**G**enomics **E**vidence **N**eoplasia **I**nformation **E**xchange)[67].

Box 3

Consent Models for Biomedical Research

In all the situations indicated below, the principles of respect for the individual "data donor" and ethically responsible data sharing are implicit to the process.

Specific Consent

In this situation, consent is limited to data generated from a particular research protocol applied in a specific disease type

Specific and "Related Conditions" Consent

This consent process adds the possibility for the consent for use of research data to be extended to other related disease domains

Tiered Consent

Here, a series of options is provided, with the research participant being able to indicate consent for one, some or all of the options indicated

Dynamic Consent

This is a continuous consent process, with the opportunity for the participant to indicate their consent (or lack of consent) for their data to be used in an evolving series of research studies which have developed over time from the original research protocol

Broad Consent

This consent indicates consent for future unspecified research studies, whose ethical principles are ensured through oversight from an independent Research Ethics Committee

Open Consent

For an open consent model, all future biomedical research is indicated, with resulting research data becoming accessible to other researchers

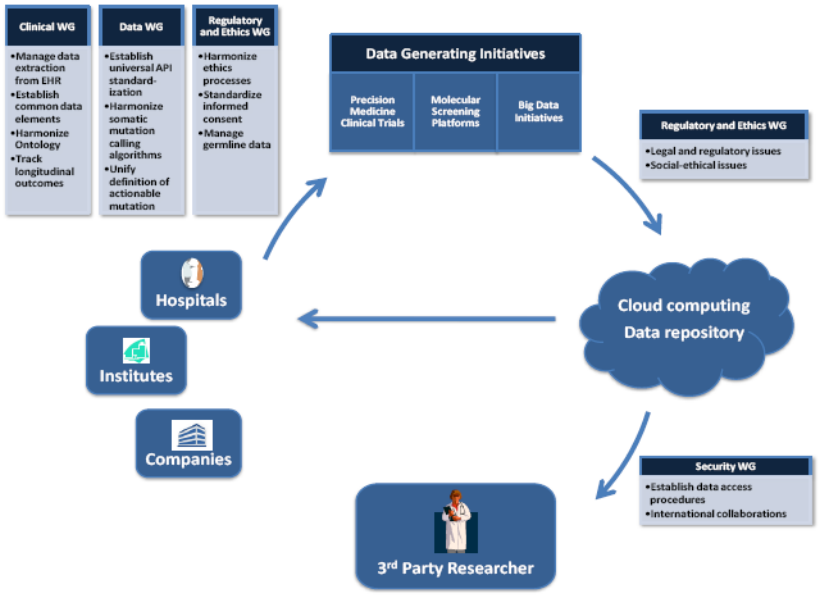

Figure 1.

Data Sharing Vision as Facilitated by GA4GH: Through its Working Groups, each of which focuses on particular data sharing challenges e.g. Clinical Working Group – Establish common data elements; Data Working Group – Establish universal API Standardization; Regulatory and Ethics Working Group – Harmonize ethics processes; Security Working Group – Establish data access procedures GA4GH provides guidance to facilitate responsible, effective and secure data sharing. Groups such as hospitals, institutes and pharmaceutical companies conducting data generating initiatives are encouraged to share clinical and genomic information under the framework developed by GA4GH, including collaborations with third party researchers via robust access procedures.

Table 1

Key Cancer Molecular Profiling and Big Data Initiatives, Including GA4GH-Enabled Initiatives

Examples of Initiatives	Description^	How is This Initiative Enabling Data Sharing?
National and International Molecular Screening Platforms		
SPECTA (**S**creening **P**atients for **E**fficient **C**linical **T**rial **A**ccess)[63]	*SPECTA* represents a pan-European collaboration involving over 40 clinical centers in 16 European countries, with initial focus on colorectal cancer, but now expanding to various other tumor types	■ Molecular screening platform matching patients' genomic profiles to potential clinical trials
Precision Medicine Clinical Trials		
NCI-MATCH (National Cancer Institute's **M**olecular **A**nalysis for **T**herapy **Ch**oice)[64]	A complex basket trial evaluating a new or existing agent against a specific molecular aberration across tumor types. *NCI-MATCH* will be available at over 2,400 clinical sites across the US	■ Enables sharing of clinical and genomic data to facilitate access to innovative targeted therapies
TAPUR (**T**argeted **A**gent and **P**rofiling **U**tilization **R**egistry)[65] and **DRUP** (**DRU**g Rediscovery **P**rotocol)	In these two trials, genomic analysis identifies a particular abnormality which allows patients access to a molecularly targeted agent already shown to be effective against this "actionable" mutation in at least one cancer	■ Inter-continental parallel data sharing approach to facilitate patient access to "approved" treatments
Big Data Initiatives		
TCGA (**T**he **C**ancer **G**enome **A**tlas)[10]	An NCI/NHGRI collaboration that has generated comprehensive catalogues of the key genomic changes in major types and subtypes of cancer	■ Supports open access to genomic data
ICGC (**I**nternational **C**ancer **G**enome **C**onsortium)[11] and **ICGCMed** *	A consortium created to coordinate large-scale comprehensive molecular characterization of 50 different tumor types and/or subtypes. *ICGCMed* is the next generation project of *ICGC* with a stated aim to link genomic data with longitudinal clinical data	■ Supports open access to data ■ Links genomic and clinical data / outcomes (ICGCMed)
CCE (**C**ancer **C**ore **E**urope)[66]	A consortium of 6 European cancer centers that share a common translational genomic platform to conduct next generation clinical trials	■ Establishes a European virtual e-cancer hospital
100,000 Genomes Project[3]	A project supported by Genomics England to sequence 100,000 whole genomes with focus on rare diseases, cancer and infectious diseases	■ Enables sharing of clinical and WGS data for clinical actionability
CancerLinQ[20]	An ASCO-led initiative to create a data informatics system that will collect, analyze and learn from complete EHR, with the primary goal to improve the quality of care provided to patients with cancer	■ Enables sharing of clinical and genomic data ■ Addresses data security and access issues
The Cancer Genome Collaboratory	A Canadian NSERC, Genome Canada, and CIHR-supported initiative to make ICGC data available for cloud computing in a community cloud infrastructure (www.genomecanada.ca/en/cancer-genome-collaboratory)	■ Enables sharing of clinical and genomic/NGS data ■ Co-locates compute with big data sets.
GENIE (**G**enomics **E**vidence **N**eoplasia **I**nformation **E**xchange)[67]	An AACR-enabled trans-Atlantic initiative to integrate genomic profiles and longitudinal clinical outcome at 7 different cancer centers in the US, Canada and Europe	■ Enables sharing of clinical and genomic/NGS data

Examples of Initiatives	Description^	How is This Initiative Enabling Data Sharing?
		■ Addresses data security and access issues
GA4GH-enabled Data Sharing Initiatives		
BRCA Challenge *	A global initiative to pool data on *BRCA1/2* genetic variants and corresponding clinical data (www.genomicsandhealth.org/work-products-demonstration-projects/brca-challenge-0)	■ Creates a curated catalogue, *BRCA Exchange*
Beacon Project *	A simple online web service that allows users to query an institution's databases to determine whether they contain a genetic variant of interest (www.genomicsandhealth.org/work-products-demonstration-projects/beacon-project-0)	■ Enables sharing of genetic data
Matchmaker Exchange *	A collaborative effort to facilitate matching of cases with similar phenotypic and genotypic profiles through standardized APIs (www.genomicsandhealth.org/work-products-demonstration-projects/matchmaker-exchange-0)	■ Establishes federated platforms through standardized APIs
Other Data Sharing or Harmonization Initiatives		
BD2K (**B**ig **D**ata **t**o **K**nowledge)[26] *	A trans-NIH program to support the development of innovative approaches and tools to maximize and accelerate the integration of data science into biomedical research.	■ Develops new methods and standards for sharing genomic information
eMERGE (**E**lectronic **Me**dical **R**ecords and **Ge**nomics)[19]	A national network that combines DNA biorepositories with EHR systems for large scale, high-throughput genetic research	■ Finds solution to link EHR data to genomic data
GDC (**G**enome **D**ata **C**ommons)	An interactive knowledge system to store, analyze and distribute cancer genomics data generated by NCI and other research organizations (www.cancer.gov/news-events/press-releases/2014/GenomicDataCommonsNewsNote)	■ Enables sharing of clinical and genomic data ■ Finds solution to data warehousing
Helix Nebular Project	A European partnership between information technology providers and research centers that aims to develop a science cloud to meet the growing demand for computing power (www.helix-nebula.eu)	■ Enables sharing of clinical and genomic data through cloud computing
HL7 (**H**ealth **L**evel **7** International) and **FHIR** (**F**ast **H**ealthcare **I**nteroperability **R**esources)	An international collaboration dedicated to provide frameworks and standards for the exchange, sharing and integration of electronic health information (www.hl7.org/fhir/)	■ Develops standards for sharing EHR data
ICGC-TCGA DREAM (**D**ialogue for **R**everse **E**ngineering **A**ssessments and **M**ethods)[28] *	An international effort to improve standard methods for identifying cancer-associated mutations and rearrangements in WGS data	■ Standardizes WGS and bioinformatics algorithms ■ Identifies best pipelines for harmonizing mutation calling

* Collaborations with GA4GH

^ Description of each initiative is adapted from related publication or website.

AACR = American Association for Cancer Research; API = Application Programming Interface; ASCO = American Society of Clinical Oncology; CIHR = Canadian Institutes of Health Research; EHR = electronic health record; EORTC = European Organisation for Research and Treatment of Cancer; GA4GH = Global Alliance for Genomics and Health; NCI = National Cancer Institute; NGS = next generation sequencing; NHGRI = National Human Genome Research Institute; NIH = National Institutes of Health; NSERC = Natural Sciences and Engineering Research; WG = Working Group; WGS = Whole Genome Sequencing