GA4GH: International policies and standards for data sharing across genomic research and healthcare

Heidi I. Rehm  
*Massachusetts Institute of Technology*

Malachi Griffith  
*Washington University School of Medicine in St. Louis*

Obi L. Griffith  
*Washington University School of Medicine in St. Louis*

et al

Follow this and additional works at: [https://digitalcommons.wustl.edu/open_access_pubs](https://digitalcommons.wustl.edu/open_access_pubs)

Please let us know how this document benefits you.

**Recommended Citation**

Rehm, Heidi I.; Griffith, Malachi; Griffith, Obi L.; and et al, "GA4GH: International policies and standards for data sharing across genomic research and healthcare." Cell Genomics. 1, 2. 100029 (2021).  
[https://digitalcommons.wustl.edu/open_access_pubs/11516](https://digitalcommons.wustl.edu/open_access_pubs/11516)

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
SUMMARY

The Global Alliance for Genomics and Health (GA4GH) aims to accelerate biomedical advances by enabling the responsible sharing of clinical and genomic data through both harmonized data aggregation and federated approaches. The decreasing cost of genomic sequencing (along with other genome-wide molecular assays) and increasing evidence of its clinical utility will soon drive the generation of sequence data from tens of millions of humans, with increasing levels of diversity. In this perspective, we present the GA4GH strategies for addressing the major challenges of this data revolution. We describe the GA4GH organization, which is fueled by the development efforts of eight Work Streams and informed by the needs of 24 Driver Projects and other key stakeholders. We present the GA4GH suite of secure, interoperable technical standards and policy frameworks and review the current status of standards, their relevance to key domains of research and clinical care, and future plans of GA4GH. Broad international participation in building, adopting, and deploying GA4GH standards and frameworks will catalyze an unprecedented effort in data sharing that will be critical to advancing genomic medicine and ensuring that all populations can access its benefits.

INTRODUCTION

The Universal Declaration of Human Rights states that everyone has the right to share in scientific advancement and its benefits. In order to fully deliver the benefits from genomic science to the broad human population, researchers and clinicians must come together to agree on common methods for collecting, storing, transferring, accessing, and analyzing molecular and other health-related data. Otherwise, this information will remain siloed within individual disease areas, institutions, countries, or other...

18Canadian Center for Computational Genomics, Montreal, QC, Canada
19University of Toronto, Toronto, ON, Canada
20University Health Network, Toronto, ON, Canada
21Vector Institute, Toronto, ON, Canada
22Oregon Health and Science University, Portland, OR, USA
23European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
24PhenoTips, Toronto, ON, Canada
25Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
26University of Melbourne, Melbourne, VIC, Australia
27Vanderbilt University Medical Center, Nashville, TN, USA
28National Cancer Institute, National Institutes of Health, Bethesda, MD, USA
29Memorial Sloan Kettering Cancer Center, New York, NY, USA
30 Nationwide Children’s Hospital, Columbus, OH, USA
31The Ohio State University, Columbus, OH, USA
32National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA
33Stanford University, Stanford, CA, USA
34UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA
35Arizona State University, Tempe, AZ, USA
36Sage Bionetworks, Seattle, WA, USA
37University of Edinburgh, Edinburgh, UK
38Canadian Distributed Infrastructure for Genomics (CanDIG), Toronto, ON, Canada
39University of Utah, Salt Lake City, UT, USA
40Sidra Medicine, Doha, Qatar
41Well Cornell Medicine - Qatar, Doha, Qatar
42Addenbrooke’s Hospital, Cambridge, UK
43University of Chicago, Chicago, IL, USA
44Mayo Clinic, Rochester, MN, USA
45National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA
46Verily Life Sciences, South San Francisco, CA, USA
47Harvard Medical School, Boston, MA, USA
48Washington University School of Medicine in St. Louis, St. Louis, MO, USA
49Pryzm Health, Sydney, QLD, Australia
50Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain
51University of Colorado Anschutz Medical Campus, Aurora, CO, USA
52Johns Hopkins University, Baltimore, MD, USA
53Autism Speaks, Princeton, NJ, USA
54Duke-Margolis Center for Health Policy, Washington, DC, USA
55The University of Hong Kong, Hong Kong, Hong Kong
56BBMRI-ERIC, Graz, Austria
57Masaryk University, Brno, Czech Republic
58National Taiwan University, Taipei City, Taiwan
59École polytechnique fédérale de Lausanne (EPFL), Lausanne, Switzerland
60Cincinnati Children’s Hospital Medical Center, Cincinnati, OH, USA

(Affiliations continued on next page)
Perspective


(Author list continued on next page)
jurisdictions, locking away its potential to contribute to research and medical advances. The Global Alliance for Genomics and Health (GA4GH) is a worldwide alliance of genomics researchers, data scientists, healthcare practitioners, and other stakeholders. We are collaborating to establish policy frameworks and technical standards for responsible, international sharing of genomic and other molecular data as well as related health data. Founded in 2013, the GA4GH community now consists of more than 1,000 individuals across more than 90 countries working together to enable broad sharing that transcends the boundaries of any single institution or country (see https://www.ga4gh.org).

In this perspective, we present the strategic goals of GA4GH and detail current strategies and operational approaches to enable responsible sharing of clinical and genomic data, through both harmonized data aggregation and federated approaches, to advance genomic medicine and research. We describe technical and policy development activities of the eight GA4GH Work Streams and implementation activities across 24 real-world genomic data initiatives (“Driver Projects”). We review how GA4GH is addressing the major areas in which genomics is currently deployed including rare disease, common disease, cancer, and infectious disease. Finally, we describe differences between genomic sequence data that are generated for research versus healthcare purposes, and define strategies for meeting the unique challenges of responsibly enabling access to data acquired in the clinical setting.

HARNESSING THE GENOMIC MEDICINE REVOLUTION

As the costs associated with human genomic sequencing continue to decline, genomic assays are increasingly used in both research and healthcare. As a result, we expect tens of millions of human whole-exome or whole-genome sequences to be generated within the next decade, with a high proportion of that data coming from the healthcare setting and therefore associated with clinical information. If they can be shared, these datasets hold great promise for research into the genetic basis of disease and will represent more diverse populations than previously possible, and will be available to researchers worldwide.

GA4GH aims to enable the responsible sharing of clinical and genomic data across both research and healthcare by developing standards and facilitating their uptake. We believe that without such a consortium, the emerging utility of genomics in clinical practice will be slower, more expensive, and fragmented, with little harmonization between countries. GA4GH standards (see Table 1) allow researchers to securely and responsibly access data regardless of where they are physically located. Technical standards give researchers the confidence that someone else could reproduce their work by running the same packaged method over the same underlying data, using the same persistent identifiers. Standards also give data providers confidence that their data are being accessed in accordance with their data use policies, by researchers they have authorized, without losing control of multiple downloaded copies of the data. As a result, data providers can enable research with the assurance that their legal and ethical requirements are being upheld, while researchers benefit from the use of global data resources and tools.

As nascent genomic medicine programs emerge in many countries, we believe that federated approaches (see Federated access below), in addition to centralized data sharing where
feasible, are necessary to satisfy the goals of both the research and healthcare communities. In addition, many commercial and public organizations aim to minimize the costs and risks of the complex technical software needed to either contribute to genomic medicine or deliver genomic tools. A complex, multi-stakeholder ecosystem requires neutral and technically competent standards; these standards must be adaptable for disparate purposes and useful for the broad set of end-users: clinical, academic, commercial, and public. Finally, standards must be developed to intentionally support the global research community with specific attention to policies of equity, diversity, and inclusion to tangibly enable progress for all global communities.

GA4GH ORGANIZATION

GA4GH has partnered with 24 real-world genomic data initiatives (Driver Projects) to ensure its standards are fit for purpose and driven by real-world needs. Driver Projects make a commitment to help guide GA4GH development efforts and pilot GA4GH standards (see Table 2). Each Driver Project is expected to dedicate at least two full-time equivalents to GA4GH standards development, which takes place in the context of GA4GH Work Streams (see Figure 1). Work Streams are the key production teams of GA4GH, tackling challenges in eight distinct areas across the data life cycle (see Box 1). Work Streams consist of experts from their respective sub-disciplines and include membership from Driver Projects as well as hundreds of other organizations across the international genomics and health community.

GA4GH standards development and approval process

GA4GH Work Streams and Driver Projects have identified, and are actively developing, the technical specifications and policy frameworks they believe to be of most relevance to enable widespread data sharing, federated approaches, and interoperability across datasets to facilitate genomic research (see supplemental information for more details on the product development process); the areas of focus are outlined in Box 1, with individual products defined in Table 1 and in the 2020/2021 GA4GH Roadmap (https://www.ga4gh.org/roadmap).

Each GA4GH deliverable can be implemented on its own to enable interoperability and consistency in a single area. However, when implemented together, they support broader activities in the research and clinical domains and enable productive genomic data sharing and collaborative analyses that can leverage global datasets produced in distinct locations around the world.

Each approved GA4GH deliverable is reviewed by a panel of internal and external experts not involved in the product’s development, and then by the GA4GH Steering Committee (https://www.ga4gh.org/about-us/governance-and-leadership-2/#steering). GA4GH standards are not typically accredited by a national or international standards body, and instead follow a model inspired by the Internet Engineering Task Force (IETF; https://www.ietf.org) and the World Wide Web Consortium (W3C; http://www.w3.org). This enables a flexible and rapid response to community needs and a focus on lowering barriers to interoperability through the development and adoption of pragmatic standards. However, there are occasions when certain standards benefit from a more formal accreditation process, especially when there is a direct link into healthcare usage (see next section and Box 2).

Alignment with other standards organizations

To achieve greater international coordination and consistency of standards development, GA4GH proactively collaborates with other standards development organizations working in genomics, e.g., Health Level Seven (HL7: http://www.hl7.org), International Organization for Standardization (ISO; https://www.iso.org), Open Biological and Biomedical Ontology Foundry (OBO; http://www.obofoundry.org/). While defined work processes between GA4GH and other standards development bodies are still under development, GA4GH has initiated several pilot projects to explore mechanisms of collaboration. One such
approach is the submission of GA4GH standards to ISO’s technical committees for approval as ISO international standards. Using a product development timeline that aligns the ISO approval process with the GA4GH approval process, both communities are able to contribute to the development of a standard in a harmonized manner. These efforts expand the diversity of contributors to both organizations, leading to more robust and internationally applicable standards. Another approach, guided by HL7 working groups and experts, is the translation of GA4GH standards into HL7 Fast Health Interoperability Resources (FHIR) Implementation Guides. These implementation guides enable interoperability of GA4GH standards with clinical systems and accelerate the use of clinical data for research.

GA4GH also aims to support and interoperate with existing translational models, ontologies, and terminologies (e.g., FHIR, HGVS, OMOP, PCORnet, Human Phenotype Ontology, SNOMED CT) for clinical genetics and genomics.21–23 Before launching a new standards development project, GA4GH Work Streams are encouraged to complete a landscape analysis that both defines relevant existing standards and how they will influence the development of the new standard. Coordination activities—such as joint meetings, shared documentation, and process harmonization between GA4GH work and these health standards-focused efforts—are critical for bridging the research-clinical divide and keeping respective products aligned. This helps prevent unnecessary proliferation of redundant standards and minimizes the development of semantically and syntactically conflicting standards that could hamper large-scale interoperability and lead to confusion within the adopter community (see Box 2).

**Federated approaches**

Federated approaches—the ability to analyze data across multiple distinct and secure sites—is increasingly seen as an important strategy where data cannot be pooled for legal or practical reasons. These approaches are characterized by independent organizations hosting data in secure processing environments (e.g., clouds, trusted research environments) while adopting technical standards that enable analysis at scale.24 Application programming interfaces (APIs) can be deployed to enable researchers and portable workflows to visit multiple databases even where the data and computing environment are variably configured.25 Tools like “identity federation” can facilitate even closer integration across organizations.26–28,29

GA4GH Driver Projects and other partners are beginning to implement cloud-based workflows built on GA4GH standards that allow scientists to share, access, and interrogate data stored at disparate sites around the globe. Some concrete examples of this access pattern include (1) the Data Coordination Platform of the Human Cell Atlas, an internationally federated compute environment for analyzing single-cell data; (2) Genomics England’s secure Research Environment for approved investigators to access the 100,000 Genomes Project dataset; (3) the NHGRI Genomic Data Science Analysis, Visualization, and Informatics Lab-space (AnVIL)30 and the Gen3 Data Commons, which provide cloud-based spaces for scientists to work with large-scale genomic and genomic-related datasets and shared tools; and (4) H3ABioNet, a bioinformatics platform that serves data from the Human Heredity and Health in Africa (H3Africa) network to researchers across the continent and provides containerized workflows for analysis of the data.

Because these workflows are built on interoperable standards, they allow for secure access and efficient discovery, portability, and analysis. With more instances like these, the global community will be able to harness the power of large data and improve the reach of genomic medicine research. The federation and transparency enabled by standards will also encourage greater willingness among non-western and other underrepresented populations to share their data, affording greater diversity in the overall data available and equity in its impacts.

**GENOMICS IN HEALTHCARE**

The process of sequencing a genome is essentially the same in any setting, but the scale and quality control of production,31 as...
well as the regulation and dissemination of the resulting data, can be quite different in healthcare compared to research.32,33 “Research genomes” contain de-identified data and therefore are often openly shared with other researchers, including for funding and publishing requirements (for NIH policy, see web resources), frequently with managed access, e.g., via the European Genome-phenome Archive (EGA), the Japanese Genome-phenome Archive (JGA), or the database of Genotypes and Phenotypes (dbGaP). Researchers worldwide will draw on these openly shared genomic datasets for their own studies, increasing the amount of knowledge derived from each genome.34 However, while such research genomes are more readily available, these datasets usually do not include the type or extent of longitudinal, standardized, or interoperable clinical data needed for genomic medicine.29,30

Healthcare-based research and testing have an entirely different financial, legal, and social landscape, with the structure, provision, and regulation varying by country, covering the full spectrum from state-run to private schemes.7 In each system, the cost of an assay in healthcare—genomics included—is often considered in light of its benefits to the health of an individual and the cost of an assay in healthcare—genomics included—is often considered in light of its benefits to the health of an individual and the cost of an assay in healthcare—genomics included—is often considered in light of its benefits to the health of an individual and the cost effectiveness within the healthcare system—especially if it is cost effective—the only limit to its deployment is the number of patients who will potentially benefit. In practice, however, there are logistical, financial, regulatory, educational, scientific, and clinical-based hurdles to overcome before a genomic test becomes a routine clinical offering. In addition, barriers to healthcare access will likely remain impediments to large-scale implementation in many countries.

The current case for implementing genomics in healthcare can be presented in four broad disease areas: rare disease, cancer, common/chronic disease, and infectious disease. In the following sections we outline the case for healthcare-funded sequencing in each disease area. We also highlight challenges to implementation in each area and GA4GH deliverables aimed at overcoming these issues.

**Rare disease**

Arguably, the rare disease space has seen the most successful deployment of genomics in healthcare, with many reporting diagnostic rates of at least 20%–30%, and health economic studies demonstrating cost-effectiveness and diagnostic utility.36–41 Clinical geneticists have used single-gene or small gene panel tests since the early 1990s to support diagnosis and some treatment decisions for many of these diseases. The cost of assaying broader genomic regions—including exome and genome sequencing—has fallen considerably, with a substantial impact on rare-disease diagnosis and discovery research.42,43 However, with more than 10,000 rare diseases44 affecting more than 300 million patients worldwide,45 diagnosing and discovering treatments for many of these diseases has been challenging. As such, the rare disease community has embraced data sharing in order to facilitate global knowledge exchange and improve patient diagnostic rates, understand disease progression, and augment care strategies.41

To further enable progress, clinical and research laboratories and health systems must support several key activities to effectively identify, diagnose, and eventually treat the genetic causes of rare disease: (1) aggregate genomic and phenotypic data, needed for discerning population allele frequencies in disease and non-disease populations and implicating new genes in rare disease; (2) catalog the validity of gene-disease associations using consistent annotation models and terminologies;46 (3) collectively build knowledge bases to understand variant pathogenicity; (4) define the natural histories of rare diseases to predict disease progression and enable a foundation upon which to develop clinical trials; and (5) monitor treatment efficacy...
of emerging therapeutics. GA4GH standards and policies already enable and will continue to build upon these activities. For example, the Matchmaker Exchange—a rare disease gene discovery platform which has benefited from GA4GH guidance on API-based data exchange formats as well as consent45 and data security policies46,49—illustrates the power of bringing practicing clinicians and researchers together, as cases from across the globe are necessary to build evidence to confirm new gene-disease relationships.58

GA4GH promotes knowledge sharing in ClinVar, a database which has accelerated improvements in variant classification across the clinical laboratory community.50 Additional methods are now being deployed to move beyond manual submission of variant classifications to a centralized database; such advances will enable more timely access to siloed laboratory knowledge and evidence-based variant classification. Real-time sharing with ClinVar—facilitated by APIs and with entries linked to rich, case-level data—will be needed to scale our understanding of the more than 750 million variants so far identified in the human genome (e.g., within gnomAD; https://gnomad.broadinstitute.org). The Variation Representation (VRS)18 and Variant Annotation (VA) specifications aim to support the exchange of variant data, Phenopackets and Pedigree representation to support the use of standardized clinical and family history data, as well as new APIs (e.g., Beacon v2 API and Data Connect API) to enable the identification of data for further access and analysis. The aim is for these standards to support a more global and federated approach to rare disease data and knowledge sharing that will be critical to advancing diagnosis and treatment of rare diseases.

Cancer
One in five men and one in six women worldwide will have a cancer diagnosis in their lifetime.61 This risk is 2- to 3-fold greater in higher-resource countries,63 with estimates as high as one in two people in the UK for example.58 An altered somatic genome is a consistent hallmark of cancer, often associated with specific pathogenic mutations.53 In some individuals with hereditary cancer syndromes, germline variants can disrupt cancer-related pathways and increase the risk of developing a “heritable” malignancy.54-56 Characterizing a cancer by sequencing a patient’s tumor genome alongside their germline genome has resulted in profound insights into molecular mechanisms of malignant transformation and discovery of potential therapeutic targets.57,58 Tumor/normal sequencing has demonstrated applications in disease monitoring69 as well as diagnosis,60 prognosis,61 and therapeutic response prediction,62 both at initial presentation63 and disease recurrence.64

Applying cancer genomics in the clinic is more complicated than that for rare diseases. For cancer patients, treatment strategy time frames are commonly measured in weeks and incorporating genomic information within such an urgent turnaround time is logistically challenging to integrate into clinical decision making.65 Additionally, while the use of genomics for diagnosis and improved symptom management can lead to substantial improvements for rare disease patients and their families, application of genomics in cancer treatment is more complex and may include dual assessment of both somatic and germline genomes to determine heritable cancer risk and the assessment of the evolving tumor genome due to changing selective pressures in response to targeted therapies. Cancer genomic information is most useful if it informs treatment options, yet development of systems that match patients to appropriate clinical trials would be needed to fully realize the benefits of genomic tumor data where estimates of clinical trial enrollment in patients with cancer stands at ~8%.66 Genetic information is increasingly important in clinical decision making through routine clinical sequencing assays and molecular tumor boards.67 The heterogeneity of cancer as a disease—of each individual tumor and of any concurrent or subsequent manifestation, such as metastasis or recurrence—adds many layers of complexity to genomic analysis.68 To address this complexity, it is important to analyze somatic and germline variation data together to understand their contribution to cancer risk.69

Most of the same standards and workflows important for rare disease apply to tumor sequencing, including data storage and compression standards (e.g., CRAM), variation representation (e.g., VCF and VRS), analysis (e.g., cloud-based workflows), and linkage to patient records (e.g., Phenopackets). However, discovery of oncogenic driver mutations also requires significant coordination and standardization to track outcome data (e.g., progression and response to treatment), a key element in determining the clinical significance of variation found in cancer patients.70 As such, many groups have created knowledge bases to annotate cancer genomic variation associated with evidence of pathogenicity or relevant treatment options; however, these knowledge bases can have limited levels of interoperability. In 2014, a GA4GH task team launched the Variant Interpretation for Cancer Consortium (VICC), which standardizes and coordinates clinical somatic cancer curation efforts and has created an open community resource to provide the aggregated information.71 Moving forward, major oncogenic resources are now working with GA4GH on the harmonization of variant interpretation evidence, through refinement and adoption of standards such as the Beacon API, the Data Use Ontology (DUO),9 VA, and VRS. Additionally, these standards are being implemented across multiple GA4GH Driver Projects (see Table 2) that capture genomic data and/or diagnostic variant interpretation across the longitudinal evolution of cancer.

Common/chronic disease
“Common disease” is a catchall phrase describing a vast spectrum of diseases that have complex environmental and genetic etiologies. Accurate prediction of common diseases from genetics has been a topic of study since the inception of human genetics, yet genomic information is still not widely used in clinical practice for this purpose. The discovery of a large number of genetic susceptibility loci (polygenic architecture) supported the common-disease common-variant hypothesis72 and has led to the generation of polygenic risk scores summarizing common disease risk.73 Studies are now beginning to demonstrate the clinical benefits of applying polygenic risk scores in practice through stratification of the population for deploying disease management strategies.74-76 As the assay of choice moves from genotype arrays to sequencing, there will be integration between common disease and rare disease applications; this is
already the case for certain diseases such as susceptibility to breast cancer\textsuperscript{75} or heart disease.\textsuperscript{77} When such genomic information can be used clinically for common diseases, it will be more justifiable to sequence entire populations. Population-scale sequencing is in place already in some countries (e.g., Iceland) and is likely to become more commonplace in the next two decades.

To support the discovery of the genetic causes and contributors to common disease across all populations, researchers must be able to identify and access aggregated data from large-scale cohort population studies from diverse backgrounds, carried out by multiple distinct sites such as biobanks in the UK (UK Biobank, Generation Scotland), China (China Kiddey Biobank), the US (NIH All of Us Research Program), and Japan (Tohoku Medical Megabank, Japanese BioBank); and whole population cohorts in Iceland (deCODE), Estonia (Estonian Genome Project), and Finland (FinnGen). Doing so requires the data to be harmonized across all sites using common data models and terminologies. Furthermore, since genomic datasets of this scale are too large to download and manipulate at individual sites, researchers must be able to bring analytical tools to the data, regardless of their location.

Protocols are needed to deploy these tools consistently and effectively across distinct federated sites. GA4GH products support this critical type of biological study across the typical research life cycle from data discovery to analysis: (1) identify and access datasets relevant to a disease study (e.g., GA4GH Passports, DUO, multiple data discovery APIs), (2) access secure genotype and phenotype information on patients with related traits (e.g., Phenopackets, Data Repository Service [DRS] API, VRS, VA), and (3) remotely run analytical methods on data of interest (e.g., Task Execution Service [TES], Workflow Execution Service [WES] API, Itisget API\textsuperscript{17}), avoiding the need for inter-jurisdictional transfers and disparate regulatory requirements.

**Infectious disease**

Genomics can be used to identify the infectious agents of disease with more confidence and precision than ever before, and at increasing speed, allowing treatments that can quickly resolve infections\textsuperscript{76–80} as well as identifying the evolution of new species that may evade antibiotics, antivirals, and vaccines. The main challenges to deployment of genomics in infectious disease care are managing cost and logistics, tracking disease progression and its characterization, achieving precise phenotypic prediction (e.g., antibiotic resistance), and harmonizing historical knowledge bases from non-genomic-based assays to integrate with contemporary genomic tests. The COVID-19 pandemic tested this infrastructure, with diagnostic testing becoming widespread, viral genomic sequencing enabling tracking of strains, and human genome sequencing of symptomatic individuals contributing to a better understanding of the basis of COVID-19 disease severity.\textsuperscript{81}

Infectious disease genomic research and surveillance primarily rely on sequencing bacterial and viral pathogens and the organisms in which they are carried and transmitted. These genomes vary greatly in size, content, and associated metadata, so the standards and APIs created for human genomic data may be insufficient for infectious disease data. However, while the specific data standards needed to advance pathogen genomics differ from those in human genomics, there is still considerable overlap in the mechanics of sharing the data.

Through a variety of strategic alignments with organizations such as the Public Health Alliance for Genomic Epidemiology (PHAGE; https://pha4ge.org), the International COVID-19 Data Alliance (ICODA; http://www.icoda-research.org), and the European COVID19 data portal (http://www.covid19dataportal.org), GA4GH is working to ensure that the species-agnostic elements of genomic data sharing standards are transferred into the infectious disease community. In addition, some GA4GH standards have begun to explore how they should adapt to support infectious disease data; for example, the Phenopackets standard was improved to support case-level presentation for infectious diseases in 2020 in response to the COVID-19 pandemic. In addition, recently launched initiatives such as large-scale tuberculosis sequencing in several countries,\textsuperscript{82} rapid identification of Ebola and Zika virus strains,\textsuperscript{83} and tracing hospital outbreaks using genomics\textsuperscript{84,85} demonstrate a vibrant, functional interface between research, public health institutions, and clinical practice.

**CHALLENGES TO SECONDARY USE OF CLINICALLY ACQUIRED DATA**

We envision the global clinical and research communities collaborating seamlessly in the context of practicing healthcare\textsuperscript{86,87} to enable a true “learning healthcare system” (LHS). The LHS concept has existed for over a decade,\textsuperscript{88,89} however, implementation is still in its infancy, facing several barriers.\textsuperscript{90} Some useful implementations are found across medicine,\textsuperscript{91–94} including genomic medicine.\textsuperscript{95} Increasing numbers of institutions and countries have begun biobanks, in many cases connected to their healthcare system (see Common/chronic disease above), providing fertile grounds on which to bring healthcare data—including clinical genomic data—into research.

To enable these efforts to reach their full potential, disparate systems must be able to share genomic and clinical data, requiring the community to overcome key challenges, particularly in the areas of infrastructure development, patient and physician incentives, ethics and regulation, privacy and security, and socio-cultural expectations (see Box 3). We believe these challenges can be overcome—but only if the genomics and healthcare communities commit to broad-based advocacy and coordinated efforts worldwide.

This has already been successfully modeled through the Clinical Genome Resource (ClinGen; a GA4GH Driver Project), where healthcare providers, clinical laboratory staff, and researchers work together to develop standards for gene and variant curation, share underlying evidence, and then apply that evidence through a consensus-driven process to classify genes and variants which are made freely accessible to the broader community to support both research and clinical care.\textsuperscript{96,97}

**Developing clinical data standards**

Much of the clinical data contained within healthcare are not encoded in a standardized format.\textsuperscript{99} Multiple electronic health record (EHR) vendors exist today and are highly proprietary in their technical structures, making standardization across EHRs and
with downstream research systems difficult. Although data recorded in EHRs often use standardized clinical terminologies (e.g., ICD, SNOMED CT), the intent of these systems is generally to present clinical information on individuals to healthcare providers and, in some regions, facilitate billing practices. This presents a challenge for secondary users, where it is difficult to make accurate, population-scale conclusions, often requiring extensive efforts to understand practices and generate useful research data. In order to promote adoption of standardized formats in research and ultimately within EHRs, GA4GH is developing standardized information models (e.g., Phenopackets, Pedigree) to describe clinical phenotypes and family histories. Standardizing the representation of phenotype and pedigree information will allow patients, care providers, and researchers to share this information more easily between healthcare and research systems and enable software tools to use this information to improve genome analysis and diagnosis.

### Incentivizing and facilitating data sharing in healthcare

Resource limitations for healthcare providers and patients also impact their ability to share valuable clinical data. Some healthcare institutions (e.g., NHS England [https://www.england.nhs.uk/genomics/nhs-genomic-med-service]), Dana-Farber Cancer Institute [http://www.dana-farber.org/for-patients-and-families/becoming-a-patient/preparing-for-your-first-appointment/checklist-for-new-adult-patients], Danish healthcare[105] have built layered consent procedures into the regular routine of medical practice. Others support parallel biobanking efforts to separately consent patients for research. Still others have built this into their operations as an inherent part of the healthcare system. Further incentives can be built if providers can experience the direct benefits of research. For example, the clinical laboratory genetic testing industry largely participates voluntarily in data sharing through ClinVar, in part because they directly benefit from accurate variant interpretation. Several laboratories also joined when the US insurance industry began requiring submission as a condition of test reimbursement. However, despite progress in the sharing of variant knowledge, additional incentives and infrastructure are needed to support access to case-level results (e.g., variants interpreted for a patient indication) as well as full sequencing data, along with rich clinical phenotypes. Currently, most genetic test results are returned through PDF-based reports or accessed through external portals outside the medical system. Although standards exist for the exchange of genetic test results (see, for example, HL7’s guide in the web resources),[110] robust standards that capture highly detailed, discrete genomic data are still under development. Adoption of those standards has been motivated by the implementation of downstream clinical decision support,[111–113] but more incentives and infrastructure will be needed.

To date, GA4GH has worked on maintaining and evolving standardized file formats for raw and annotated genomic data (SAM, BAM, CRAM, VCF/BCF); individual variant representation and interpretation (VRS, VA); and transmission of individual phenotype data and interpreted results (Phenopackets), all of which are critical for the evolving use of genomics in healthcare systems—particularly clinical laboratory workflows to share genomic data and genetic testing results. Future areas of development include better representation of structural variants, unambiguous representation of complex multi-allelic loci, and research into new, more scalable formats for storing and exchanging genetic variation. Population-scale sequencing programs in which healthcare systems share clinical genomic data for research are unlikely to allow large-scale aggregation of data to migrate beyond national boundaries, but federated analysis—in which analytical algorithms or queries are brought to the data in its location without data egress—is feasible and is a major area of focus of GA4GH’s standards development.

### Ethics and regulation

Ethical considerations for patients and populations, together with responsible regulation, are essential for healthcare-funded genomics, which involves complex national regulation and legislation. Different countries and institutions have individual values and policies that relate to allowing access to personal information, with some embracing more open regulatory norms.

---

**Box 3. Major barriers hindering secondary use of clinically acquired data**

Here we outline some of the major challenges to achieving the broad goal of responsible sharing of genomic and related health data. This includes setting up the infrastructure to support the flow of data from clinical practice into research, as well as establishing data-access and accountability mechanisms that are appropriate to research settings. These need to be consistent with the legal frameworks of the healthcare setting, and respectful of the rights of the individual data donor including their privacy, the security of their data, and their autonomy with regard to research participation.

1. Inconsistency and lack of version control in data-generating pipelines
2. Lack of dataset interoperability due to disparate data models and terminologies
3. Inadequate infrastructure for ingesting and storing data
4. Difficulty or lack of resources for enabling access to data
5. Insufficient consent for data sharing and lack of resources to support the consent process
6. Data privacy and security issues, as well as real and perceived regulatory issues
7. Challenges to ensuring patients understand how their data are used and have sufficient autonomy around data sharing participation
8. Differences in priorities, experiences, and trust levels concerning data sharing between different population groups and stakeholders
9. Lack of incentives in the clinical care system for prioritizing data sharing and research
10. Lack of data-sharing mandates
Box 4. GA4GH Regulatory & Ethics Toolkit

The GA4GH Regulatory and Ethics Work Stream (REWS) develops ready-to-use policy guidance to support responsible, international genomic and health-related data sharing. Here, we list central components of the GA4GH Regulatory & Ethics Toolkit. The REWS also reviews all GA4GH technical standards for any regulatory or ethics issues that may be relevant.

**Policy Frameworks:** GA4GH has developed five policy guidance documents (or “Frameworks”) that build on the Framework for Responsible Sharing of Genomic and Health-Related Data, each aiming to address a specific area of responsible data sharing:

- **Data Privacy & Security Policy Framework:** provides principled and practical guidance for processing data in a way that protects the security, integrity, and availability of data and services, and the privacy of individuals, families, and communities whose data are processed (https://www.ga4gh.org/wp-content/uploads/GA4GH-Data-Privacy-and-Security-Policy_FINAL-August-2019_wPolicyVersions.pdf).
- **Cloud Privacy & Security Policy Framework:** provides principled and practical best practices for sharing data in a way that protects and promotes the confidentiality, integrity, and availability of data and services, and the privacy of individuals, families, and communities whose data are shared (https://www.ga4gh.org/wp-content/uploads/Privacy-and-Security-Policy.pdf).
- **Policy Framework for Clinically Actionable Genomic Research Results:** provides a reference point for managing the return of clinically actionable research results that recognizes the importance of the accountability and transparency of genomic researchers toward participants (https://www.ga4gh.org/wp-content/uploads/2021-Policy-on-Clinically-Actionable-Genomic-Research-Results.pdf).


**Machine Readable Consent Guidance (MRCG):** The MRCG provides instructions for researchers to integrate standard data-sharing language into consent forms in a way that can be translated into a computable language. Machine-readable consent language can be attached to datasets and stored in their descriptive data using DUO terms. Researchers can then search for datasets that have been consented for their research purposes (https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf).

**Data Access Committee Review Standards (DACReS):** DACReS is a set of procedural standards for data access committees that facilitate consistency, effectiveness, and robustness of reviews for data access requests to genomic and health-related data.

**Engagement Framework:** This framework enables researchers and others to robustly design engagement with various public and patient audiences implicated in genomic data sharing. Through reflexive questions centered around themes of fairness, context, heterogeneity, and the recognition of tension, the framework facilitates critical inquiry into stakeholder engagement (https://www.ga4gh.org/wp-content/uploads/GA4GH-Engagement-policy_V1.0_July2021-1.pdf).

**GDPR Briefs:** These monthly briefs answer important questions about the impact of the European General Data Protection Regulation on various aspects of international health research and genomic and health-related data sharing (https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/gdpr-forum/).

and systems on data collection, access, and sharing, and others being more restrictive. Nevertheless, most systems have some mechanism for researchers to access both research and clinical data. The GA4GH Regulatory and Ethics Work Stream (REWS) develops ready-to-use policy guidance to support responsible, international genomic and health-related data sharing. In Box 4, we list central components of the GA4GH Regulatory & Ethics Toolkit, including policies, consent tools, and data access guidance. The REWS also reviews all GA4GH technical standards for consideration of any regulatory or ethics issues that may be relevant.

The first REWS product was the GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data, which is built on the human right to benefit from scientific progress and its applications, as well as privacy, non-discrimination, and procedural fairness. It provides guidance for the responsible sharing of human genomic and health-related data, including personal health data and other types of data that may have predictive power in relation to health. The Framework has now been translated into 14 languages and has been used to inform local data sharing approaches around the globe, including, for example, the World Economic Forum, the Academy of Science of South Africa, DNA.Land, Health Data Research UK, and the Horizon-2020 CORBEL project. Keeping the fundamental human right to benefit from science at the heart of clinical and genomic data sharing ensures a universal approach to balancing the benefits and potential risks. We believe that most healthcare system actors can ultimately
participate in responsible, worldwide data sharing while remaining compliant with applicable laws and institutional policies.

**Privacy and security**

Federating large volumes of sensitive clinical and genomic data across internationally distributed virtual computing environments presents formidable challenges in assuring data integrity, service availability, and individual privacy. Some of these challenges call for innovative application of well-established security standards, frameworks, and protocols—such as identity federation on a global scale—and some GA4GH standards already do so (e.g., crypt4GH, Authentication & Authorization Infrastructure [AAI] / Passports). Another crucial challenge is to enable secure, privacy-preserving federated analysis, wherein researchers can extract information without having to transfer raw data. This evolution is key to foster inter-institutional and international collaboration and will be a strong incentive to improve ontology homogeneity. Several technical solutions are available, either based on hardware devices or on software algorithms. The former are computationally efficient, but require trusting a vendor and are prone to side-channel attacks. The latter are computationally slower, but are mathematically proven and are a better response to GA4GH expectations. Recent results have demonstrated the effectiveness of a software-based approach (a combination of homomorphic cryptography and secure multiparty computation called “Multi-party Homomorphic Encryption” or MHE); these approaches have been positioned with respect to the GDPR. One of the major strengths of MHE is that partial aggregates can be considered to be anonymized and not just pseudonymous, in the sense of GDPR, and thus potentially obviating the need for data transfer and use agreements (DTUAs).

**Societal challenges**

Societal challenges of allowing access to genomic data within the healthcare ecosystem include maintaining public trust, overcoming differences in objectives and methods between research and healthcare, and breaking down unproductive divides between disciplines. Our vision for healthcare data ecosystems is one in which vetted researchers around the world can, with appropriate oversight and policy enforcement, gain access to human health data for the benefit of patients. GA4GH has defined the core elements of responsible data sharing, including transparency, accountability, recognition, and attribution as well as sanctions for misuse which form a framework to respect and maintain the trust of participants. In particular, the GA4GH Engagement Framework (see Box 4) further assists researchers in designing and understanding engagement with public, patient, and participant stakeholders through the central themes of fairness, context, heterogeneity, and the recognition of tensions. Through the implementation arm of GA4GH, the Genomics in Health Implementation Forum described below and other engagement efforts, GA4GH is tackling the broader societal implementation issues including education and engagement of the public, healthcare providers, and regulators in order to build trust within the community. The GA4GH “Your DNA, Your Say” survey, an effort to gather international public attitudes toward genomic data sharing, has provided an evidence base for understanding which factors are important to maintaining public trust in the generation and sharing of genomic data, as well as how concerns differ according to geography. These findings help ensure that GA4GH’s work can enhance the public trust in a global context upon which the future of genomics depends.

**CONNECTING STANDARDS FOR IMPLEMENTATION**

With more than 30 GA4GH standards approved, and dozens of production-ready implementations of those standards deployed around the world, GA4GH is now shifting its focus toward demonstrating how standards can work together to provide seamless support of genomic activities. Interconnected standards that are compatible and interoperable with each other and are hardened for real-world use will enable solutions for federated analyses across platforms and use cases. To drive this effort, GA4GH has established the Federated Analysis System Project (FASP), which aims to demonstrate how GA4GH APIs, when used in concert, can support real-world, scientific use cases (see https://www.ga4gh.org/genomic-data-toolkit/2020-connection-demos/). A key outcome of FASP is a series of scripts that represent working examples of clients accessing real-world GA4GH-compatible services to solve a spectrum of challenges across the search-access-analyze workflow. The scripts illustrate how these services have adopted GA4GH standards to solve challenges, such as dataset discoverability and controlled data access, in order to drive larger scale and more powerful analyses.

By developing working implementations of GA4GH standards that are pressure tested in real world scenarios, the FASP team has identified specific areas of improvement within the standards. As a result of this work, new features will be added to existing GA4GH specifications to further facilitate secure, real-world federated data sharing and analysis. Most notably, the group is working toward a standardized solution for using a GA4GH Passport to access a controlled access dataset from a Data Repository Service (DRS), while fulfilling robust security requirements, such as preventing escalation of privilege. These efforts will be critical to support access to valuable datasets across the globe.

**GA4GH Starter Kit**

To date, GA4GH has primarily focused on overcoming the challenges of enabling interoperability within new initiatives built on a foundation of cloud infrastructure. However, an additional—and potentially more significant—challenge is bringing high-performance computing (HPC) infrastructures that are not already focused on cloud interoperability into the federated network envisioned by this community.

While more ambitious goals are on the horizon for connecting and extending GA4GH standards (e.g., discovery of datasets; matching requests, analyses, and datasets; describing phenotypes; reporting on variants), FASP has shown through its real-world demonstrations of access across distributed but interoperable datasets that the initial groundwork for federated analysis is now in place. The Data Repository Service (DRS) allows data
custodians to make controlled access data available at multiple sites; the Workflow Execution and Task Execution Services (WES & TES) allow researchers to encapsulate and run analyses on those data; and AAI and Passports allow for federated authorization and authentication, streamlining the data access process for both researchers and data custodians.

In 2021, GA4GH has begun to develop the GA4GH Starter Kit, a set of open source reference implementations (for example, code bases that demonstrate the standards working in practice), to help ensure existing HPC environments can interoperate with the wider GA4GH network. These resources consist of “plug-and-play” code that any institution (cloud-based or HPC) can use to quickly achieve GA4GH compatibility and will facilitate the progressive movement of established large-scale systems toward interoperability. In addition, a testing suite will be developed to ensure deployments of both reference and non-reference implementations are compliant to their respective GA4GH specifications.

Genomics in Health Implementation Forum

Once standards have been piloted in real-world Driver Project settings and shown to enable true federated analysis in FASP, they can begin to be promoted more broadly in the research and clinical genomics communities. Launched in 2020, the Genomics in Health Implementation Forum (GHIF) brings together a group of national-scale genomic data initiatives to share resources, experiences, and best practices for implementing GA4GH standards, as well as broader experience in rolling out national and international data sharing activities. GHIF aims to support more accurate data interpretation and disease diagnosis plus other innovative solutions across healthcare through global cooperation in data sharing and clinical implementation of genomics.

Broad uptake of GA4GH standards among GHIF members—which include both GA4GH Driver Projects as well as other national and multi-national initiatives (see https://ga4gh.org/implementation for full list)—will provide strong evidence that GA4GH standards are supporting the community’s actual data sharing needs.

Implementation of GA4GH policies and standards throughout the scientific and healthcare communities will allow researchers to access data across the globe—a critical step toward answering otherwise impenetrable questions about disease and basic human biology. As the volume of genomic and health-related data grows exponentially around the world, researchers, clinicians, and bioinformaticians have a responsibility to make that data appropriately accessible and to use it to realize benefits for all humans everywhere. The promise of genomic medicine lies at a crossroads that depends on harmonization across the global community to significantly enhance human health and medicine. We believe that GA4GH, by embracing collaborative innovation and knowledge exchange, is well poised to meet this challenge.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2021.100029.

ACKNOWLEDGMENTS

We acknowledge all current and past members of the GA4GH Work Streams, Steering Committee, Strategic Advisory Board, and Secretariat. We also acknowledge the members of the Human Genetics Society of Australasia Education, Ethics & Social Issues Committee who contributed to the development of the clauses for pediatric consent to genetic research.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the institutions with which each individual is affiliated.

B.P.C. acknowledges funding from Abigail Wexner Research Institute at Nationwide Children’s Hospital; T.H. Nyrönen acknowledges funding from Academy of Finland grant #31996; A.M.-J., K.N., T.F.B., O.M.H., and Z.S. acknowledge funding from Australian Medical Research Future Fund; M.S. acknowledges funding from Biobank Japan; D. Bujold and S.J.M.J. acknowledge funding from Canada Foundation for Innovation; L.J.D. acknowledges funding from Canada Foundation for Innovation Cyber Infrastructure grant #34860; D. Bujold and G.B. acknowledge funding from CANARIE; L.J.D. acknowledges funding from CANARIE Research Data Management contract #RDM-090 (CHORD) and #RDM2-053 (ClinDIG); K.K.-L. acknowledges funding from CanSHARE; T.L.T. acknowledges funding from Chan Zuckerberg Initiative; T. Burdett acknowledges funding from Chan Zuckerberg Initiative grant #2017-171671; D. Bujold, G.B., and L.D.S. acknowledge funding from CIHR; L.J.D. acknowledges funding from CIHR grant #404896; M.J.S.B. acknowledges funding from CIHR grant #SBED-163124; M. Courtot and M. Linden acknowledge funding from CINECA project Europe Horizon 2020 grant #825775; D. Bujold and G.B. acknowledge funding from Compute Canada; F.M.-G. acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – NFDI 1/1 “GHGA – German Human Genome-Phenome Archive; R.M.H.-S. acknowledges funding from Duke-Margolis Center for Health Policy; S.B. and A.J.B. acknowledge funding from EJP-RD EU Horizon 2020 grant #825575; A. Niewielska, A.K., D.S., G.I.S., J.A.T., J.R., M.A.K., M. Baudis, M. Linden, S.B., S.S., T.H. Nyrönen, and T.M.K. acknowledge funding from ELIXIR; A. Niewielska acknowledges funding from EOSC-Life EU Horizon 2020 grant #824087; J.-P.H. acknowledges funding from ETH Domain Strategic Focal Area “Personalized Health and Related Technologies (PHRT)” grant #2017-201; F.-M.-G. acknowledges funding from EU-CAN EU Horizon 2020 grant #825835; B.M.K., D. Bujold, G.B., L.D.S., M.J.S.B., N.S., S.E.W., and Y.-J. acknowledge funding from Genome Canada; B.M.K., M.J.S.B., S.E.W., and Y.-J. acknowledge funding from Genome Quebec; F.M.-G. acknowledges funding from German Human Genome-Phenome Archive; C. Voisin acknowledges funding from Google; A.J.B. acknowledges funding from Health Data Research UK Substantive Site Award; D.H. acknowledges funding from Howard Hughes Medical Institute; S.B. acknowledges funding from Instituto de Salud Carlos III; S.-S.K. and K.T. acknowledge funding from Japan Agency for Medical Research and Development (AMED); S. Ogishima acknowledges funding from Japan Agency for Medical Research and Development (AMED) grant #20kk0205014h0005; C.Y. and K. Kosaki acknowledge funding from Japan Agency for Medical Research and Development (AMED) grant #JP18kk0205012; GEM Japan acknowledges funding from Japan Agency for Medical Research and Development (AMED) grants #19kk0205014h0004, #20kk0205014h0005, #20kk0205013h0005, #20kk0205012h0005, #20kk040501h0003, and #19kk040501h0104; J.R. acknowledges funding from La Caixa Foundation under project #LCF/PR/GEN13/50260009; R.R.F. acknowledges funding from Mayo Clinic Center for Individualized Medicine; Y.J. and S.E.W. acknowledge funding from Ministere de l’Innovation du Quebec for the Can-SHARE grant #141210; A.M.H., M.C.M.-T., J.O.J., H.E.P., and P.N.R. acknowledge funding from Monarch Initiative grant #R24OD011883 and Phenomics First NHGRI grant #1R1H1HG010860; A.L.M. and E.B. acknowledge funding from MRC grant #MC_PC_19024; P.T. acknowledges funding from National University of Singapore and Agency for Science, Technology and Research; J.M.C. acknowledges funding from NHGRI; A.H.W. acknowledges funding from NHGRI awards K39HG010157, R00HG010157, and R35HG011949; A.M.-J., K.N.,
<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Beacon API</strong></td>
<td><a href="https://app.swaggerhub.com/apis/ELIXIR-Finland/ga-4_gh_beacon_api_specification/1.0.0-rc1">https://app.swaggerhub.com/apis/ELIXIR-Finland/ga-4_gh_beacon_api_specification/1.0.0-rc1</a></td>
<td>API</td>
<td>data custodians, researchers (via research infrastructures), identity provider services</td>
<td>The Beacon protocol defines an open standard for genomics data discovery. It provides a framework for public web services responding to queries against genomic data collections, for instance from population-based or disease-specific genome repositories. Beacon is designed to (1) focus on robustness and easy implementation, (2) be maintained by individual organizations and assembled into a federated network, (3) be general-purpose and able to be used to report on any variant collection, (4) provide a boolean (or quantitative) answer about the observation of a variant, and (5) protect privacy, with queries not returning information about single individuals. A new version of the API will include support for more granular control based on a user’s identity authorization and will enable discovery of cohorts, cases (patients), biological samples, and genomic variants and associated knowledge. More details can be found on the Beacon Project website.</td>
</tr>
<tr>
<td><strong>Data Connect</strong></td>
<td><a href="https://github.com/ga4gh-discovery/data-connect">https://github.com/ga4gh-discovery/data-connect</a></td>
<td>API</td>
<td>data custodians, researchers, and API &amp; tool developers</td>
<td>Data Connect is a specification for discovery and search of biomedical data, which provides a mechanism for describing data and its data model, and for searching data within the given data model. The primary container for data in Data Connect is the table. Tables contain rows of data, where each row is a JSON object with key/value pairs. The table describes the structure of its row objects using JSON Schema (<a href="https://json-schema.org/">https://json-schema.org/</a>). Row attributes can take on any legal JSON value, e.g., numbers, strings, booleans, nulls, arrays, and nested JSON objects. The API supports browsing and discovery of data models and table metadata, listing table data, and optionally searching table data using arbitrarily complex expressions including joins and aggregations. The query language is SQL with domain-specific functions to facilitate informative typing of the result fields. Data publishers can wrap existing data storage and retrieval systems in the Data Connect API or may choose to publish data directly as static files in the Data Connect JSON format. Data consumers can use Data Connect via graphical data discovery and exploration built upon the API, via command line tools (interactively or in batch workflows), and directly as an API in custom analysis programs. More information can be found in the specification (<a href="https://github.com/ga4gh-discovery/data-connect/blob/master/SPEC.md">https://github.com/ga4gh-discovery/data-connect/blob/master/SPEC.md</a>).</td>
</tr>
<tr>
<td>Relevant standards</td>
<td>URL</td>
<td>Type</td>
<td>Target user</td>
<td>Purpose</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----</td>
<td>------</td>
<td>-------------</td>
<td>---------</td>
</tr>
</tbody>
</table>
| Data Use Ontology | http://purl.obolibrary.org/obo/duo.owl | Data Model / Ontology | data custodians, researchers, DACs | The Data Use Ontology (DUO) is a hierarchical vocabulary of terms describing data use permissions and modifiers, in particular for research data in the health/clinical/biomedical domain. The GA4GH DUO standard allows large genomics and health data repositories to consistently annotate their datasets, ensuring a shared, machine readable, representation of data access conditions, and making them automatically discoverable based on a researcher’s authorization level or intended use. Reference implementations are available at  
  - Broad’s FireCloud - Data Library  
  - Broad’s DUOS (Data Use Oversight System) - Data Catalog  
  - European Genome-Phenome Archive. DUO is based on the OBO Foundry principles (http://www.obofoundry.org/principles/fp-000-summary.html) and developed using the W3C Web Ontology Language. DUO can be browsed online via the Ontology Lookup Service or Ontobee. It has been registered with the OBO Foundry (http://obofoundry.org/ontology/duo.html). |
| GA4GH Passports | https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md | API / Data Model | data custodians, researchers, DACs, clinicians, API and tool developers | The GA4GH Passport specification aims to support data access policies within current and evolving data access governance systems. This specification defines Passports and Passport Visas as the standard way of communicating a user’s data access authorizations based on either their role (e.g., researcher), affiliation, or access status. Passport Visas from trusted organizations can therefore express data access authorizations that require either a registration process (for the Registered Access data access model) or custom data access approval (such as the Controlled Access applications used for many datasets). |
| Service Info | https://github.com/ga4gh-discovery/ga4gh-service-info | API | API and tool developers | Service discovery is at the root of any computational workflow using web-based APIs. Traditionally, this is hard-coded into workflows, and discovery is a manual process. Service Info provides a way for an API to expose a set of metadata to help discovery and aggregation of services via computational methods. It also allows a server/implementation to describe its capabilities and limitations. Service-info is described in GA4GH OpenAPI specification, which can be visualized using Swagger Editor (https://editor.swagger.io/?url=https://raw.githubusercontent.com/ga4gh-discovery/ga4gh-service-info/develop/service-info.yaml). |

(Continued on next page)
<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Registry</td>
<td><a href="https://github.com/ga4gh-discovery/ga4gh-service-registry">https://github.com/ga4gh-discovery/ga4gh-service-registry</a></td>
<td>API</td>
<td>API and tool developers</td>
<td>Service registry is a GA4GH service providing information about other GA4GH services, primarily for the purpose of organizing services into networks or groups and service discovery across organizational boundaries. Information about the individual services in the registry is described in the complementary Service Info specification (see above). The Service Registry specification is useful when dealing with technologies that handle multiple GA4GH services. Common use cases include creating networks or groups of services of a certain type (e.g., Beacon Network searches networks of Beacon services across multiple organizations, a workflow can be executed by a specific group of Workflow Execution Services, or Data Connect search on biomedical data is federated across a set of nodes), or a certain host (e.g., an organization provides implementations of Beacon, Data Connect, and Data Repository Service APIs, or a server hosts an implementation of refget and htsget APIs).</td>
</tr>
<tr>
<td>htsget(^\text{12})</td>
<td>samtools.github.io/hts-specs/htsget.html</td>
<td>API</td>
<td>API and tool developers, researchers</td>
<td>htsget is a data retrieval API that bridges from existing genomics file formats to a client/server model with the following features: (\text{incumbent data formats (BAM, CRAM, VCF) are preferred initially, with a future path to others.} ) (\text{Multiple server implementations are supported, including those that do format transcoding on the fly, and those that return essentially unaltered filesystem data.} ) (\text{Multiple use cases are supported, including access to small subsets of genomic data (e.g., for browsing a given region) and to full genomes (e.g., for calling variants).} )</td>
</tr>
<tr>
<td>refget(^\text{13})</td>
<td>samtools.github.io/hts-specs/refget.html</td>
<td>API</td>
<td>API and tool developers, researchers</td>
<td>refget (<a href="https://w3id.org/ga4gh/refget">https://w3id.org/ga4gh/refget</a>) is an API and mechanism for generating identifiers for reference sequences and retrieving sequences via API. The refget identifier is derived from sequence content directly and therefore does not rely on a central issuing authority. This allows downstream clients to unambiguously refer to a reference sequence and to retrieve said sequence. The refget API can also provide subsequences and metadata pertaining to the checksum identifier. A refget server can host any number of reference sequences of any type, e.g., genomic DNA or protein sequences. The refget protocol is a fundamental building block of the CRAM specification. An OpenAPI description of this specification is available and describes the 1.0.0 version (<a href="https://github.com/samtools/hts-specs/blob/master/pub/refget-openapi.yaml">https://github.com/samtools/hts-specs/blob/master/pub/refget-openapi.yaml</a>). Implementors can check if their refget implementations conform to the specification by using our compliance suite (<a href="https://github.com/ga4gh/refget-compliance-suite">https://github.com/ga4gh/refget-compliance-suite</a>). A summary of all known public implementations is available from our compliance report website.</td>
</tr>
<tr>
<td>Relevant standards</td>
<td>URL</td>
<td>Type</td>
<td>Target user</td>
<td>Purpose</td>
</tr>
<tr>
<td>--------------------</td>
<td>-----</td>
<td>------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>Task Execution Service (TES)</td>
<td><a href="https://github.com/ga4gh/task-execution-schemas">https://github.com/ga4gh/task-execution-schemas</a></td>
<td>API</td>
<td>API and tool developers, researchers, academic institutions</td>
<td>The Task Execution Service (TES) API is a standardized schema and API for describing and executing batch execution tasks. A task defines a set of input files, a set of containers and commands to run, a set of output files, and some additional logging and metadata. TES servers accept task documents and execute them asynchronously on available compute resources. A TES server could be built on top of a traditional HPC queuing system, such as Grid Engine, Slurm, or cloud style compute systems such as AWS Batch or Kubernetes.</td>
</tr>
<tr>
<td>Tool Registry Service (TRS)</td>
<td><a href="https://github.com/ga4gh/tool-registry-service-schemas">https://github.com/ga4gh/tool-registry-service-schemas</a></td>
<td>API</td>
<td>API and tool developers, researchers, academic institutions</td>
<td>The GA4GH Tool Registry (TRS) API aims to provide a standardized way to describe the availability of tools and workflows. In this way, multiple repositories that share Docker-based tools and workflows (based on Common Workflow Language [CWL], Workflow Description Language [WDL], Nextflow, or Galaxy) can consistently interact, search, and retrieve information from one another. The end goal is to make it much easier to share scientific tools and workflows, enhancing our ability to make research reproducible, shareable, and transparent. To access the specification, users can:</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>• view the human-readable Reference Documentation • explore the specification in the Swagger Editor • preview documentation from the gh-openapi-docs for the development branch at <a href="https://ga4gh.github.io/tool-registry-service-schemas/preview/develop/docs/index.html">https://ga4gh.github.io/tool-registry-service-schemas/preview/develop/docs/index.html</a></td>
</tr>
<tr>
<td>Workflow Execution Service (WES)</td>
<td><a href="https://github.com/ga4gh/workflow-execution-service-schemas">https://github.com/ga4gh/workflow-execution-service-schemas</a></td>
<td>API</td>
<td>API and tool developers, researchers, academic institutions</td>
<td>The Workflow Execution Service (WES) API describes a standard programmatic way to run and manage workflows. Having this standard API supported by multiple execution engines will let people run the same workflow using various execution platforms running on various clouds/environments. Key features include: (1) ability to request a workflow run using CWL or WDL; (2) ability to parameterize that workflow using a JSON schema; and (3) ability to get information about running workflows.</td>
</tr>
</tbody>
</table>

Securely access genotype and phenotype information on patients with related traits

| Authentication & Authorisation Infrastructure (AAI) | https://github.com/ga4gh/data-security/blob/master/AAI/AAIConnectProfile.md | Guide | API and tool developers | The GA4GH Authentication & Authorisation Infrastructure (AAI) specification profiles the OpenID Connect (OIDC) protocol to provide a federated (multilateral) authentication and authorization infrastructure for greater interoperability between genomics institutions in a manner specifically applicable to (but not limited to) the sharing of restricted datasets. In particular, this specification introduces a JSON Web Token (JWT) syntax for an access token to enable an OIDC provider (called a Broker) to allow a downstream access token consumer (called a Claim Clearinghouse) to locate the Broker’s /userinfo endpoint as a means to fetch GA4GH Claims. This specification is suggested to be used together with others that specify the syntax and semantics of the GA4GH Claims exchanged. |

(Continued on next page)
Table 1. Continued

<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloud Security and Privacy Policy v1.0</td>
<td><a href="https://docs.google.com/document/d/1cBTwtetnsvO2vU3HVwLTLaC9H_ya4MjZUa_g_xzOBg/edit">https://docs.google.com/document/d/1cBTwtetnsvO2vU3HVwLTLaC9H_ya4MjZUa_g_xzOBg/edit</a></td>
<td>Guide</td>
<td>anyone handling sensitive data in a cloud infrastructure.</td>
<td>An increasing number of GA4GH projects rely on Cloud services to pursue their goals, and the GA4GH Cloud Work Stream is working on several products to make the GA4GH community take full advantage of the Cloud paradigm. However, the use of the Cloud poses significant security and privacy challenges that need to be carefully evaluated and addressed. The purpose of the Cloud Security and Privacy Policy is to outline a common security technology framework that can be used to systematically assess the products developed by the CWS from a security perspective. Product developers and reviewers can leverage the information contained herein to identify requirements, threats, and countermeasures related to the products they are working on, thus facilitating the production of secure standards.</td>
</tr>
</tbody>
</table>

CRAM\(^{14}\) | samtools.github.io/hts-specs/CRAMv3.pdf | File Format | API and tool developers, researchers | The CRAM file format holds DNA sequencing records. It has the following major objectives:  
- Significantly better lossless compression than BAM  
- To permit simple and lossless transformations to and from BAM files  
- Support for controlled loss of data  
The first two objectives allow users to take immediate advantage of the CRAM format while offering a smooth transition path from using BAM files.  
The third objective supports the exploration of different lossy compression strategies and provides a framework in which to effect these choices.  
Data in CRAM is stored in a columnar fashion, with each column being compressed with either a general-purpose compressor or a custom method. If aligned, sequences may be stored as differences against a reference sequence, which is optionally stored within the CRAM file.  
External references may be either a local file or obtained remotely via the refget API. Data may be retrieved either as whole alignment records, or selectively only for the fields (columns) required. |

(Continued on next page)
By its nature, genomic data can include information of a confidential nature about the health of individuals. It is important that such information is not accidentally disclosed. One part of the defense against such disclosure is to, as much as possible, keep the data in an encrypted format. The Crypt4GH specification describes a file format that can be used to store data in an encrypted state. Existing applications can, with minimal modification, read and write data in the encrypted format. The choice of encryption also allows the encrypted data to be read starting from any location, facilitating indexed access to files. The format has the following properties:

- **Confidentiality**: Data stored in the file are readable only by holders of the correct secret decryption key. The format does not hide the length of the encrypted file, although it is possible to pad some file structures to obscure the length.
- **Integrity**: Data are stored in a series of 64 kilobyte blocks, each of which includes a message authentication code (MAC). Attempts to change the data in a block will make the MAC invalid; it is not possible to recalculate the MAC without knowing the key used to encrypt the file. The format only protects the contents of each individual block. It does not protect against insertion, removal, or reordering of entire blocks.
- **Authentication**: The format does not provide any way of authenticating files.

Crypt4GH may be used with any data file or stream, but one usage is encryption of BAM, CRAM, VCF, and BCF data within the htsget API while still retaining full random access.
<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Security Infrastructure Policy (DSIP)</td>
<td><a href="https://github.com/ga4gh/data-security/blob/master/DSIP/DSIP_v4.0.md">https://github.com/ga4gh/data-security/blob/master/DSIP/DSIP_v4.0.md</a></td>
<td>Policy</td>
<td>data protection authorities</td>
<td>The Data Security Infrastructure Policy (DSIP) describes the data security infrastructure recommended for stakeholders in the GA4GH community. It is not meant to be a normative document, but rather a set of recommendations and best practices to enable a secure data sharing and processing ecosystem. However, it does not claim to be exhaustive, and additional precautions other than the ones collected in the policy might have to be taken to be compliant with national/regional legislations. As a living document, the DSIP will be revised and updated over time, in response to changes in the GA4GH Privacy and Security Policy, and as technology and biomedical science continue to advance.</td>
</tr>
<tr>
<td>Machine Readable Consent Guidance (MRCG) v1.0</td>
<td><a href="https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf">https://www.ga4gh.org/wp-content/uploads/Machine-readable-Consent-Guidance_6JUL2020-1.pdf</a></td>
<td>Guide</td>
<td>researchers, institutional review boards/research ethics committees (international and national), research ethics policy makers, data generators, funding agencies</td>
<td>The Machine Readable Consent Guidance (MRCG) provides standardized consent clauses and supporting information to enable the development of consent forms that map unambiguously to the GA4GH Data Use Ontology (DUO). Integrating DUO into consent forms thereby facilitates data discovery and data access requests and approvals, maximizing data sharing, integration, and re-use while respecting the autonomy of data subjects. MRCG implementations include the Broad Data Use Oversight System (DUOS) and the Australian Genomics dynamic consent participant platform, CTRL.</td>
</tr>
<tr>
<td>Pedigree V1</td>
<td><a href="https://github.com/GA4GH-Pedigree-Standard/pedigree">https://github.com/GA4GH-Pedigree-Standard/pedigree</a></td>
<td>Data Model / Ontology</td>
<td>clinicians, researchers, API and tool developers, data generators, EHR vendors</td>
<td>Family health history is an important aspect in both genomic research and patient care. The GA4GH pedigree standard is an object-oriented graph-based model to represent family health history and pedigree information. It is intended to fit within the structure of other standards like HL7 FHIR and Phenopackets and enable the computable exchange of family health history as well as representation of larger, more complex families. Computable representation of family structure will allow patients, physicians, and researchers to share this information more easily between healthcare systems and help software tools use this information to improve genomic analysis and diagnosis. The draft model can be found on Github along with a Family History Relations Ontology and draft FHIR implementation guide. A draft recommendation for a minimal dataset of family health history (<a href="https://docs.google.com/document/d/1UAiSLB5E_QePRlVdPrpOfpXn6VQJXK2eQByEmtGY/edit?usp=sharing">https://docs.google.com/document/d/1UAiSLB5E_QePRlVdPrpOfpXn6VQJXK2eQByEmtGY/edit?usp=sharing</a>) was developed as a foundation of these efforts.</td>
</tr>
</tbody>
</table>

(Continued on next page)
<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenopackets</td>
<td><a href="http://phenopackets.org">http://phenopackets.org</a></td>
<td>Data Model / Ontology</td>
<td>data generators, data custodians, researchers, clinicians, API and tool developers</td>
<td>The Phenopacket specification is an open machine-readable schema that supports the global exchange of disease and phenotype information to improve our ability to diagnose and conduct research on all types of diseases, including cancer and rare disease. A Phenopacket links detailed phenotypic descriptions with disease, patient, and genetic information, enabling clinicians, biologists, and disease and drug researchers to build more complete models of disease. Version 2 of the standard, released in June 2021, expands on the previous version to include better representation of the time course of disease, treatment, and COVID-19 and cancer-related data. The schema, as well as source code in Java, C++, and Python, are available from the phenopacket-schema GitHub repository.</td>
</tr>
<tr>
<td>RNAget</td>
<td><a href="https://ga4gh-rnaseq.github.io/schema/docs/index.html">https://ga4gh-rnaseq.github.io/schema/docs/index.html</a></td>
<td>API</td>
<td>Data generators, data custodians, researchers, tool developers</td>
<td>The RNAget API describes a common set of endpoints for search and retrieval of processed RNA data. This currently includes feature level expression data from RNA-seq type assays and signal data over a range of bases from ChIP-seq, methylation, or similar epigenetic experiments. By using these common endpoints, data providers make it easier for client software to access their data with minimal or no modifications to underlying code. This improves interoperability with other compliant data providers and makes it easier for investigators to retrieve and compare data from multiple sites. For the software developer, these common endpoints and patterns make it easier to access multiple compliant server sites with the same client software. This reduces development time which may have otherwise been spent writing parsers and custom request generators. Using the API, it becomes much easier to write software to conduct comparisons, data mingling, or other analyses on data retrieved from multiple, potentially geographically dispersed data servers. The OpenAPI description of the specification can be used with code generators like OpenAPI Generator. The testing and compliance page includes a list of example server implementations which can be used as is or as a starting point. A custom solution can be implemented to link the API endpoints and queries to a local data backend (of any desired type) serving the data.</td>
</tr>
</tbody>
</table>

(Continued on next page)
### Table 1. Continued

<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAM and BAM(^7)</td>
<td>samtools.github.io/hts-specs/SAMv1.pdf</td>
<td>File Format</td>
<td>researchers</td>
<td>SAM, or Sequence Alignment/Map format, is a format for storing primary DNA sequencing records. These are typically aligned and sorted by genomic coordinate, but unaligned data can also be represented. SAM is a TAB-delimited text format consisting of a header meta-data section and an alignment section. The BAM format is a binary serialization of SAM for more efficient access. SAM and BAM support full random access, selected by genomic region. The SAMtags document defines the optional per-record annotations. These are also used by the CRAM specification.</td>
</tr>
<tr>
<td>Variant Annotation</td>
<td><a href="https://github.com/ga4gh/va-spec">https://github.com/ga4gh/va-spec</a></td>
<td>Data Model / Modeling Framework</td>
<td>API and tool developers</td>
<td>Variant annotations are structured data object that holds a central piece of knowledge about a genetic variation, along with metadata supporting its interpretation and use. A given variant annotation may describe knowledge about its molecular consequence, functional impact on gene function, population frequency, pathogenicity for a given disease, or impact on therapeutic response to a particular treatment. The GA4GH VA-Specification will define an extensible data model for representation and exchange these and other diverse kinds of variant annotations. It will provide machine-readable messaging specifications to support sharing and validation of data through APIs and other exchange mechanisms. It will also provide a formal framework for defining custom extensions to the core model - allowing community-driven development of VA-based data models for new data types and use cases. A more detailed description of these components can be found online.</td>
</tr>
</tbody>
</table>

\(^{7}\)SAM and BAM are widely used in bioinformatics for storing, aligning, and analyzing DNA sequencing data. SAM is a text-based format, while BAM is its binary version for efficiency.
## Table 1. Continued

<table>
<thead>
<tr>
<th>Relevant standards</th>
<th>URL</th>
<th>Type</th>
<th>Target user</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variation Representation(^1)</td>
<td><a href="https://vrs.ga4gh.org">https://vrs.ga4gh.org</a></td>
<td>Data Model &amp; terminology</td>
<td>data generators, API and tool developers, data custodians</td>
<td>Maximizing the personal, public, research, and clinical value of genomic information will require that clinicians, researchers, and testing laboratories exchange genetic variation data reliably. The Variation Representation Specification (VRS, pronounced “verse”) — written by a partnership among national information resource providers, major public initiatives, and diagnostic testing laboratories — is an open specification to standardize the exchange of variation data. The primary contributions of VRS include (1) terminology and an information model, (2) a machine readable schema, (3) conventions that promote reliable data sharing, (4) globally unique computed identifiers, and (5) a Python implementation (available at vrs-python) that demonstrates the above schema and algorithms and supports translation of existing variant representation schemes into VRS for use in genomic data sharing. It may be used as the basis for development in Python, but it is not required in order to use VRS. The machine-readable schema definitions and example code are available online at the VRS repository. Readers may wish to view a complete example before reading the specification. For a discussion of VRS with respect to existing standards, such as HGVS, SPDI, and VCF, see “Relationship of VRS to existing standards,” an appendix to the specification documentation.</td>
</tr>
<tr>
<td>VCF/BCF(^1)</td>
<td><a href="https://samtools.github.io/hts-specs/VCFv4.3.pdf">samtools.github.io/hts-specs/VCFv4.3.pdf</a></td>
<td>File Format</td>
<td>researchers</td>
<td>The variant call format (VCF) is a generic format for storing DNA polymorphism data such as single nucleotide polymorphisms (SNPs), insertions, deletions, and structural variants, together with rich annotations. VCF may hold data for multiple samples within the same file. The specification contains the header meta-data fields, a series of mandatory columns describing the variants, and details of the optional annotations which are either per-site or per-sample. VCF and its binary counterpart, BCF, is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome.</td>
</tr>
</tbody>
</table>

The GA4GH Toolkit outlines a suite of secure standards and frameworks that will enable more meaningful research and patient data harmonization and sharing. This suite addresses a variety of challenges across the data sharing life cycle and is applicable across the world’s accessible medical and patient-centered systems, knowledgebases, and raw data sources. All standards are subject to the GA4GH Copyright Policy ([https://www.ga4gh.org/wp-content/uploads/GA4GH-Copyright-Policy-Updated-Formatting.pdf](https://www.ga4gh.org/wp-content/uploads/GA4GH-Copyright-Policy-Updated-Formatting.pdf)) and should be made available under an open source license such as the Apache 2.0 license for software.
<table>
<thead>
<tr>
<th>Driver Project</th>
<th>URL</th>
<th>Location</th>
<th>Thematic area*</th>
<th>Current size</th>
<th>Data type(s) collected</th>
<th>Data hosting model(s)</th>
<th>Data access model(s)</th>
<th>Implementations / deployments of GA4GH standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>All of Us Research Program</td>
<td><a href="https://allofus.nih.gov/">https://allofus.nih.gov/</a></td>
<td>US</td>
<td>RD, Ca, CT</td>
<td>100k whole-genome sequences (planning for 1 million)</td>
<td>WGS, WES</td>
<td>centralized</td>
<td>cloud</td>
<td>CRAM, DRS (forthcoming), htsget (forthcoming), Passports (forthcoming), TRS (forthcoming), and WES (forthcoming)</td>
</tr>
<tr>
<td>Australian Genomics</td>
<td><a href="https://www.australian">https://www.australian</a></td>
<td>Australia</td>
<td>RD, Ca, CT</td>
<td>13,500 whole-genome sequences across all pilots</td>
<td>WGS, WES, panels, phenotype</td>
<td>centralized</td>
<td>cloud</td>
<td>Beacon V1, CRAM, Crypt4GH, DRS (forthcoming), DUO, htsget, MRCG (forthcoming), Passports (forthcoming), refget WES (forthcoming)</td>
</tr>
<tr>
<td></td>
<td>genomics.org.au/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism Sharing Initiative</td>
<td><a href="https://www.autism">https://www.autism</a> sharinginitiative.org/</td>
<td>international</td>
<td>CT</td>
<td>11,316 whole-genome sequences (estimating 15k by 2025)</td>
<td>WGS</td>
<td>distributed</td>
<td>federated analysis</td>
<td>AAI (forthcoming), Beacon V1 (forthcoming), CRAM (forthcoming), Data Connect, DRS (forthcoming), DUO (forthcoming), Passports (forthcoming), Service Registry / Info, TRS (forthcoming), WES (forthcoming)</td>
</tr>
<tr>
<td>BRCA Exchange</td>
<td><a href="http://www.brca">http://www.brca</a> exchange.org</td>
<td>international</td>
<td>RD, Ca</td>
<td>66,657 variants</td>
<td>genetic variant pathogenicity assertions and supporting evidence</td>
<td>centralized</td>
<td>public</td>
<td>Beacon V1, VA (forthcoming), VRS, WES (forthcoming)</td>
</tr>
<tr>
<td>CanDIG</td>
<td><a href="https://www.distributed">https://www.distributed</a></td>
<td>Canada</td>
<td>RD, Ca, CT, Bio</td>
<td>1,700 data records</td>
<td>WGS tumor/normal and whole transcriptome for cancer; WGS for COVID; clinical phenotype</td>
<td>distributed</td>
<td>federated analysis</td>
<td>Beacon V1, CRAM, DRS, DUO, htsget, Phenopackets, refget (forthcoming), RNAGet, Service Registry / Info (forthcoming), VRS (forthcoming), WES (forthcoming)</td>
</tr>
<tr>
<td></td>
<td>genomics.ca/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClinGen</td>
<td><a href="https://www.clinical">https://www.clinical</a></td>
<td>US</td>
<td>RD</td>
<td>2,077 unique genes with at least one curation and 2,417 unique variants with at least one curation</td>
<td>genetic and experimental evidence</td>
<td>centralized</td>
<td>public</td>
<td>VA (forthcoming), VRS</td>
</tr>
<tr>
<td></td>
<td>genome.org/</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*(Continued on next page)*
<table>
<thead>
<tr>
<th>Driver Project</th>
<th>URL</th>
<th>Location</th>
<th>Thematic area*</th>
<th>Current size</th>
<th>Data type(s) collected</th>
<th>Data hosting model(s)</th>
<th>Data access model(s)</th>
<th>Implementations / deployments of GA4GH standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELIXIR</td>
<td><a href="https://elixir-europe.org/">https://elixir-europe.org/</a></td>
<td>Europe</td>
<td>RD, Ca, CT, Bio</td>
<td>23 national nodes hold a variety of data types and run multiple services, some listed within this table (e.g., EGA). For a list of ELIXIR Core Data Resources, see <a href="https://elixir-europe.org/platforms/data/core-data-resources">https://elixir-europe.org/platforms/data/core-data-resources</a></td>
<td>distributed</td>
<td>download (also exploring Cloud)</td>
<td>AAI, Beacon V1, Crypt4GH, DRS, DUO, htsget, Passports, Phenopackets, refget, RNAGet, Service Registry / Info, TES, TRS, WES</td>
<td></td>
</tr>
<tr>
<td>ENA / EVA / EGA</td>
<td><a href="https://www.ebi.ac.uk/ena">https://www.ebi.ac.uk/ena</a>,</td>
<td>Europe</td>
<td>RD, Ca, CT, Bio</td>
<td>EGA - 700k data records</td>
<td>EGA - WGS, WES, RNaseq, epigenetics, genotyping, transcriptome, single-cell seq, healthy and disease cohorts</td>
<td>distributed</td>
<td>download (also exploring Distributed Cloud)</td>
<td>Crypt4GH, htsget AAI, Passports, DUO</td>
</tr>
<tr>
<td>EpiShare</td>
<td><a href="https://epishare-project.org/">https://epishare-project.org/</a></td>
<td>international</td>
<td>Bio</td>
<td>~2,800 data records</td>
<td>FASTQ, CRAM/BAM, bigwig, bigbed for epigenomics experiments</td>
<td>distributed</td>
<td>federated analysis</td>
<td>CRAM (forthcoming), DRS, DUO, htsget (forthcoming), Phenopackets, RNAGet, Service Registry / Info, WES</td>
</tr>
<tr>
<td>EUCANCan</td>
<td><a href="http://www.eucancan.com">http://www.eucancan.com</a></td>
<td>international</td>
<td>Ca</td>
<td>data from 35 different sources including human, model, and non-model organisms</td>
<td>whole-genome, whole-exome, and whole-transcriptome sequence data</td>
<td>distributed</td>
<td>Cloud and federated analysis</td>
<td>AAI (forthcoming), Beacon V1 (forthcoming), CRAM (forthcoming), Data Connect (forthcoming), DRS (forthcoming), Passports (forthcoming), Phenopackets (forthcoming), Service Registry / Info (forthcoming), TES (forthcoming), TRS (forthcoming), VRS (forthcoming), WES (forthcoming)</td>
</tr>
<tr>
<td>European Joint Programme on Rare Disease (EJP RD)</td>
<td><a href="https://www.ejprarediseases.org/">https://www.ejprarediseases.org/</a></td>
<td>Europe</td>
<td>RD</td>
<td>&gt;130,000 data records across several resources hosting genomic human data, mainly the EGA, DECIPHER and the RD-Connect Genome-Phenome Analysis Platform</td>
<td>a mix of WGS, WES, plausibly pathogenic variants and phenotypic information</td>
<td>distributed across centralized resources</td>
<td>download and Cloud analysis</td>
<td>AAI (forthcoming), Beacon V1, CRAM, Crypt4GH, DRS (forthcoming), DUO, htsget, Passports, Phenopackets, Service Registry / Info, TES, TRS, WES</td>
</tr>
</tbody>
</table>

(Continued on next page)
<table>
<thead>
<tr>
<th>Driver Project</th>
<th>URL</th>
<th>Location</th>
<th>Thematic area*</th>
<th>Current size</th>
<th>Data type(s) collected</th>
<th>Data hosting model(s)</th>
<th>Data access model(s)</th>
<th>Implementations / deployments of GA4GH standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEnome Medical Alliance Japan (GEM Japan)</td>
<td><a href="https://www.amed.go.jp/en/aboutus/collaboration/ga4gh_gem_japan.html">https://www.amed.go.jp/en/aboutus/collaboration/ga4gh_gem_japan.html</a></td>
<td>Japan</td>
<td>RD, Ca, CT</td>
<td>24k WGS (aiming for 100k)</td>
<td>whole-genome sequencing, whole-exome sequencing, gene expression, panels, phenotypic</td>
<td>centralized</td>
<td>download (also exploring Cloud)</td>
<td>Beacon V1 (forthcoming), CRAM, DUO, Phenopackets (forthcoming)</td>
</tr>
<tr>
<td>Human Cell Atlas</td>
<td><a href="https://www.humancellatlas.org">https://www.humancellatlas.org</a></td>
<td>International</td>
<td>RD, Ca, CT, Bio</td>
<td>1,300 donors</td>
<td>single-cell sequencing</td>
<td>centralized</td>
<td>public and Cloud</td>
<td>AAI, DRS, DUO (forthcoming), Passports (forthcoming), TES, TRS, WES</td>
</tr>
<tr>
<td>Human Heredity and Health in Africa (H3Africa)</td>
<td><a href="https://h3africa.org/">https://h3africa.org/</a></td>
<td>Africa</td>
<td>CT, Bio</td>
<td>75,000 participants (across all projects)</td>
<td>whole-genome sequencing, whole-exome sequencing, gene expression, microbiome, imaging, phenotypic, environmental/lifestyle</td>
<td>centralized</td>
<td>download</td>
<td>AAI (forthcoming), Beacon V1, CRAM, Crypt4GH, Data Connect (forthcoming), DUO, Passports (forthcoming), Phenopackets (forthcoming), VRS (forthcoming)</td>
</tr>
<tr>
<td>International Cancer Genome Consortium (ICGC) Accelerating Research in Genomic Oncology (ARGO)</td>
<td><a href="https://www.icgc-argo.org">https://www.icgc-argo.org</a></td>
<td>International</td>
<td>Ca</td>
<td>100k Genomes</td>
<td>WGS, WES, RNA-Seq, phenotype</td>
<td>distributed</td>
<td>Cloud and federated analysis</td>
<td>AAI (forthcoming), Beacon V1, CRAM, Passports (forthcoming), TRS, WES</td>
</tr>
<tr>
<td>Matchmaker Exchange</td>
<td><a href="https://www.matchmakerexchange.org">https://www.matchmakerexchange.org</a></td>
<td>International</td>
<td>RD</td>
<td>&gt;109K cases</td>
<td>WGS, WES</td>
<td>distributed</td>
<td>federated analysis</td>
<td>AAI (forthcoming), Beacon V1, CRAM, htsget, Phenopackets</td>
</tr>
<tr>
<td>Monarch Initiative</td>
<td><a href="https://monarchinitiative.org/">https://monarchinitiative.org/</a></td>
<td>International</td>
<td>RD, Ca, CT, Bio</td>
<td>N/A</td>
<td>gene, genotype, variant, disease, and phenotype data across many species in the tree of life, from over 30 data sources</td>
<td>centralized</td>
<td>public cloud</td>
<td>DUO (forthcoming), Passports (forthcoming), Phenopackets, VRS</td>
</tr>
<tr>
<td>National Cancer Institute Cancer Research Data Commons (NCI CRDC)</td>
<td><a href="https://datascience.cancer.gov/data-commons">https://datascience.cancer.gov/data-commons</a></td>
<td>US</td>
<td>Ca</td>
<td>~100,000 data records (includes GDC)</td>
<td>whole-genome sequencing, whole-exome sequencing, gene expression, panels, phenotypic, biospecimen, imaging, proteomics</td>
<td>centralized</td>
<td>Cloud and federated analysis</td>
<td>CRAM, DRS, DUO (forthcoming), Passports (forthcoming), Service Registry / Info, WES</td>
</tr>
</tbody>
</table>

(Continued on next page)
Table 2. Continued

<table>
<thead>
<tr>
<th>Driver Project</th>
<th>URL</th>
<th>Location</th>
<th>Thematic area*</th>
<th>Current size</th>
<th>Data type(s) collected</th>
<th>Data hosting model(s)</th>
<th>Data access model(s)</th>
<th>Implementations / deployments of GA4GH standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>National Cancer Institute Genomic Data Commons</td>
<td><a href="https://gdc.cancer.gov">https://gdc.cancer.gov</a></td>
<td>US</td>
<td>Ca</td>
<td>83,700 cases</td>
<td>WGS, WXSS, panel, RNA-seq, miRNA-seq, methylation array, genotyping array, diagnosis slides, tissue slides, ATAC-seq, scRNA-seq, Also clinical (phenotypic) and biospecimen information</td>
<td>centralized download and Cloud</td>
<td></td>
<td>AAI (forthcoming), CRAM (forthcoming), DRS (forthcoming), DUO (forthcoming), Passports (forthcoming), Phenopackets (forthcoming), TES (forthcoming), TRS (forthcoming), VRS (forthcoming), WES (forthcoming)</td>
</tr>
<tr>
<td>Swiss Personalized Health Network (SPHN)</td>
<td><a href="http://sphn.ch">http://sphn.ch</a></td>
<td>Switzerland</td>
<td>RD, Ca, CT, Bio</td>
<td>24 health data projects across Switzerland</td>
<td>clinical phenotypic, clinical routine, omics (genomic, transcriptomic, proteomic, etc), cohort, and imaging data and expert variant curation</td>
<td>distributed federated analysis</td>
<td></td>
<td>Beacon V1, DRS (forthcoming), htsget (forthcoming), Phenopackets, TES (forthcoming), WES (forthcoming)</td>
</tr>
<tr>
<td>Trans-Omics for Precision Medicine (TOPMed)</td>
<td><a href="https://topmed.nhlbi.nih.gov">https://topmed.nhlbi.nih.gov</a></td>
<td>US</td>
<td>RD, Ca, CT, Bio</td>
<td>180k whole genome sequences (233k by 2025), 96k panels</td>
<td>WGS, RNA-seq, metabolome, methylome (MethylationEPIC “850K”), proteome (SomaScan and Olink), longitudinal epidemiology studies, disease-studies, environmental/lifestyle, imaging</td>
<td>centralized cloud</td>
<td></td>
<td>AAI (forthcoming), CRAM, DRS, DUO, Passports (forthcoming), Service Registry / Info (forthcoming), TRS, WES</td>
</tr>
<tr>
<td>Variant Interpretation for Cancer Consortium (VICC)</td>
<td>cancervariants.org</td>
<td>international</td>
<td>Ca</td>
<td>24,366 evidence items</td>
<td>genetic and experimental evidence</td>
<td>centralized public</td>
<td></td>
<td>Beacon V1, Service Registry / Info, VA (forthcoming), VRS</td>
</tr>
</tbody>
</table>

GA4GH Driver Projects are external genomic data initiatives that have committed to both contributing to the development of genomic data sharing standards as well as piloting their use in real world practice. Abbreviations: RD, rare disease; Ca, cancer; CT, complex traits; Bio, basic biology.
D.P.H., O.M.H., T.F.B., and Z.S. acknowledge funding from NHMRC grants #GNT113531 and #GNT200001; D.L.C. acknowledges funding from NHMRC Ideas grant #1188098; A.B.S. acknowledges funding from NHMRC Investigator Fellowship grant #APP177524; J.M.C. and L.D.S. acknowledge funding from NIH; A.A.P. acknowledges funding from NIH Anvil; A.V.S. acknowledges funding from NIH contract #HHSN2682018000022 (TOPMed Informatics Research Center); S.U. acknowledges funding from NIH ENCODE grant #UM1HG009443; M.C.M.-T. and M.A.H. acknowledge funding from NIH grant #U1U5CA211044; R.J.C. acknowledges funding from NIH grants #U1U5CA210092 and #U1U5CA2023196; M.G. acknowledges funding from NIH grant #R00HG007940; J.B.A., S.L., P.G., E.H.R., and L.S. acknowledge funding from NIH grant #U24HG011025; K.P.E. acknowledges funding from NIH grant #U2RC-160010; J.A.E. acknowledges funding from NIH NCATS grant #U24TR002306; M.M. acknowledges funding from NHGRI grant #U24HG008384; B.J.W. acknowledges funding from NIH NHGRI grant #UM1HG009443A; M. Cline acknowledges funding from NH NHLBI BioData Catalyst Fellowship grant #5118777; M.M. acknowledges funding from NIH NHLBI BioData Catalyst Program grant #1OT2HL142478-01; N.C.S. acknowledges funding from NIH NHGMS grant #R35-GM128636; M.C.M.-T., M.A.H., P.N.R., and R.R.F. acknowledge funding from NIH NLM grant #5M97019P00280; E.B. and A.L.M. acknowledge funding from NIH; R.G. acknowledges funding from Project RiisCAT VEIS; S.B. acknowledges funding from RD-Connect, Seventh Framework Program grant #305444; J.K. acknowledges funding from Robertson Foundation; S.B.; and A.J.B. acknowledge funding from Solve-RD, EU Horizon 2020 grant #779257; T.S. and S. Oesterle acknowledge funding from Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health Network (SPHN), supported by the Swiss State Secretariat for Education, Research and Innovation SEI; S.J.M.J. acknowledges funding from Terry Fox Research Institute; A.E.H., M.P.B., M. Cupak, M.F., and J.F. acknowledge funding from the Digital Technology Supercluster; D.F.V. acknowledges funding from the Australian Medical Research Future Fund, as part of the Genomics Health Futures Mission grant #76749; M. Baudis acknowledges funding from the BioMedIT Network project of Swiss Institute of Bioinformatics (SIB) and Swiss Personalized Health Network (SPHN); B.M.K. acknowledges funding from the Canada Research Chair in Law and Medicine and CIHR grant #SSD-163124; D.S., G.J.S., M.A.K., S.B., S.S., and T.H. Nyrden acknowledge funding from the EU Horizon 2020 Beyond 1 Million Genomes (B1MG) Project grant #651724; P.F., A.D.Y., F.C., H.S., I.U.L., D. Gupta, M. Courtot, S.E.H., T. Burdett, T.M.K., and S.F. acknowledge funding from the European Molecular Biology Laboratory; Y.J. and S.E.W. acknowledge funding from the Government of Canada; P.G. acknowledges funding from the Government of Canada through Genome Canada and the Ontario Genomics Institute (OGi-206); J.Z. acknowledges funding from the Government of Ontario; C.K.Y. acknowledges funding from the Government of Ontario, Canada Foundation for Innovation; C. Viner and M.M.H. acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (grant #RGPIN-2015-03948 to M.M.H. and Alexander Graham Bell Canada Graduate Scholarship to C.V.); K.K.-L. acknowledges funding from the Program for Integrated Database of Clinical and Genomic Information; J.K. acknowledges funding from the Robertson Foundation; D.F.V. acknowledges funding from the Victorian State Government through the Operational Infrastructure Support (OIS) Program; A.M.L., R.N., and H.V.F. acknowledge funding from Wellcome (collaborative award); F.C., H.S., P.G., and S.E.H. acknowledge funding from Wellcome Trust grant #108749/Z/15/Z; A.D.Y., H.S., I.U.L., M. Courtot, H.E.P., P.F., and T.M.K. acknowledge funding from Wellcome Trust grant #202134/Z/16/Z; A.M., J.K.B., R.J.M., R.M.D., and T.M.K. acknowledge funding from Wellcome Trust grant #206194; E.B., P.F., P.G., and S.E.H. acknowledge funding from Wellcome Trust grant #220544/Z/20/Z; A. Hamosh acknowledges funding from NIH NHGRI grant U41HG006627 and U54HG006542; J.S.H. acknowledges funding from National Taiwan University #91F701-45C and #109T098-02; the work of K.W.R. was supported by the Intramural Research Program of the National Library of Medicine, NIH. For the purpose of open access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. H.V.F. acknowledges funding from Wellcome Grant 200990/A/16/2 ‘Designing, developing and delivering integrated foundations for genomic medicine’.

DECLARATION OF INTERESTS

H.L.R., K.N., N.M., and E.B. are members of the Cell Genomics Advisory Board. E.B. is a paid consultant to Oxford Nanopore Technologies and DoveTail Inc, both in the field of genomics. A.A.P. is a Venture Partner at GV, and has received funding from Alphabet, Microsoft, Intel, IBM, and Bayer; he is on the Novartis - Data 42 External Advisory Board and the Additional Ventures SAB. D. Glazer is on the NIH Advisory Committee to the Director, the IDCA Organizing Committee, and the Vanderbilt Biomedical Science Advisory Board. F.-G. is co-editor of the GA4GH GDPR and International Health Data Forum. J.O.J. is a consultant to Congenica Ltd. J.-P.H. is a co-founder of start-up Tune Insight (http://www.tuneinsight.com); he was on the Scientific Advisory Board of Sofia Genetics from 2012 to 2018. M.F.L. is on the boards of DNAexans, Amazon Web Services, and Google. M.N.C. is an employee of Foundation Medicine and equity holder of Roche. P.F. is a member of the Scientific Advisory Boards of Fabric Genomics, Inc, and Eagle Genomics, Ltd. R.C.G. has advised the following companies: AIA, Genomic Life, Grill, Humanity, Kneed Media, Plumcare, UnitedHealth, Verily, and VibrantHealth; and is co-founder of Genome Medical, Inc. R.M.C.-D. is on the Genome Quebec Scientific Advisory Committee, thebRCEXChange Steering Committee, and the Human Pangenome Reference Consortium ELSI Committee. R.M.H.-S. is the Chief Data Governance Officer at the National Alliance Against Disparities in Patient Health. S.S.J. is a co-founder of Global Gene Corporation Pte Ltd. A. Hamosh serves as the PI ofOMIM and is on the Steering Committee of the Matchmaker Exchange. A.V. is a member of the Data Science Committee at the Novo Nordisk Foundation, a panel member of the European Research Council Synergy Grants Committee, member of the Scientific Advisory Board of the Barcelona Institute for Global Health (ISGlobal), member of the Scientific Advisory Board of the Institut Universitaire du Cancer de Toulouse, Vice-Chair of the Scientific Committee of IMI – Innovative Medicines Initiative, member of the Scientific Committee of IMI – Innovative Medicines Initiative, member of the Scientific Committee of the Programmes Transversaux set up by the Institut National de la Santé et de la Recherche Médicale (INSERM), member of the Scientific Advisory Board of the Institute of Genetics and Development of Rennes, member of the Turing Centre for Living Systems (CENTUR) Scientific Advisory Board, member of the Institute Curie bioinformatics program Scientific Advisory Board (chair), member of SAB of the Biology Department UPF Barcelona, member of SAB of the Barcelona Biomedicine Research Institute (IRB Barcelona), member of the IntePRO database SAB, and member of the Swiss Institute of Bioinformatics SAB. R.K.H. is an employee of MyOme, Inc, former employee of Invitae, and received funding from the Broad Institute.

WEB RESOURCES

Australian Genomics, https://www.australiangenomics.org.au
Beacon Project, https://beacon-project.io/
Beacon Project API protocol, https://app.swaggerhub.com/apis/EULXIR-Finland/ga-4_g beacon_api_specification/1.0.0-rc1
BRCA Exchange, https://brcaexchange.org/
Broad Data Use Oversight System (DUOS), https://duos.broadinstitute.org/dataset_catalog
Broad’s FireCloud - Data Library, https://portal.firecloud.org/#library
GA4GH Tool Registry (TRS) API, https://github.com/ga4gh/tool-registry-service-schemas/

GA4GH RNAget testing and compliance, https://github.com/ga4gh-rnaseq/schema/blob/master/testing/README.md

GA4GH VRS Python implementation, https://github.com/ga4gh/vrs-python/
GA4GH VRS relationship to existing standards: https://vrs.ga4gh.org/en/stable/appendices/relationships.html#relationships
GA4GH VRS repository, https://github.com/ga4gh/vrs
GA4GH Workflow Execution Service (WES) API, https://github.com/ga4gh/workflow-execution-service-schemas

International Organization for Standardization (ISO), https://www.iso.org/
International COVID-19 Data Alliance (ICODA), https://icoda-research.org/
International Organization for Standardization (ISO), https://www.iso.org/home.html

Internet Engineering Task Force (IETF), https://www.ietf.org/
Matchmaker Exchange, https://www.matchmakerexchange.org/

GA4GH Cloud Security and Privacy Policy, https://docs.google.com/document/d/1cBTwtetnsvO2vU3HVwLTLaC9H_ya-4MjZUa_g_x0Bg/edit

GA4GH Data Repository Service (DRS) API, https://github.com/ga4gh/data-repository-service-schemas

GA4GH Discovery Work Stream, https://ga4gh-discovery.github.io/doi:

GA4GH minimum dataset for family health history, https://docs.google.com/document/d/1UAtSLEQ_7ePrLvDPrpoFpiXn6VQExKL2eQBxEmf0Y/edit?usp%20=sharing


GA4GH RNAget testing and compliance, https://github.com/ga4gh-rnaseq/schema/blob/master/testing/README.md


GA4GH Variant Annotation repository, https://github.com/ga4gh/ga4gh-va-spec

Health Level Seven (HL7), https://www.hl7.org/
Human Heredity and Health in Africa (H3Africa), https://h3africa.org/
International Cancer Genome Consortium (ICGC) Accelerating Research in Genomic Oncology (ARGO), https://www.icgc-argoe.org/
International COVID-19 Data Alliance (ICODA), https://icoda-research.org/
International Organization for Standardization (ISO), https://www.iso.org/home.html

Internet Engineering Task Force (IETF), https://www.ietf.org/
Matchmaker Exchange, https://www.matchmakerexchange.org/

National Cancer Institute Genomic Data Commons (NCI GDC), https://gdc.cancer.gov/

Open Biological and Biomedical Ontology Foundry (OBO): obofoundry.org/
OpenAPI description of refget v1.0.0, https://github.com/samtools/hts-specs/blob/master/pub/refget-openapi.yaml
Phenopackets, http://phenopackets.org/
Phenopackets GitHub repository, https://github.com/phenopackets/phenopacket-schema
PLINK, http://zzz.bwh.harvard.edu/plink/data.shtml
Public Health Alliance for Genomic Epidemiology (PHA4GE), https://pha4ge.org/

REFERENCES


118. DIGITAL INNOVATION HUB PROGRAMME PROSPECTUS APPENDIX (2020). PRINCIPLES FOR PARTICIPATION (HDR UK).


Supplemental information

GA4GH: International policies and standards for data sharing across genomic research and healthcare

GA4GH Product Development & Approval Processes

* For the most recent and detailed version of the GA4GH product development and approval process, please refer to our website: ga4gh.org/how-we-work/ga4gh-product-approval

Table of Contents:

- Introduction
- Product Proposal Process
- Development Process
- Product Approval Process
- Glossary

Introduction

This document outlines the processes by which the Global Alliance for Genomics and Health (GA4GH) develops and certifies technical standards and policy frameworks emerging from its work streams (WSs). GA4GH intentionally does not prescribe an in-depth and strict set of do's and don'ts. Each product comes with its own community and working practices. To enforce the same working practice across all products would be counterproductive and only increase the administrative burden of initiating new work or bringing external products into GA4GH. However, the organization does require that teams meet two specific milestones:

1) **Proposal**: before work on a new deliverable begins, the relevant Work Stream must present a proposal to the GA4GH Steering Committee, outlining the scope and expected impact of the new deliverable and the stakeholders that will be involved / affected.

2) **Approval**: before a new major version of a deliverable is released, it is submitted for approval and undergoes review by GA4GH committees as outlined below.

Product Proposal Process

Products are proposed and developed in response to the needs of the GA4GH Driver Projects (DPs)—real world genomic data initiatives that have committed to contributing to GA4GH standards development and implementation—or the wider genomics and health community. To maximize return on effort, WSLs prioritize development based on the number of requesting sources, leading to a limited number of deliverables with a greater adoption potential. WSSs are
encouraged to include a broad landscape analysis in their initial prototyping activities; the specific details of this exercise is left to the discretion of the WSLs.

If WSLs decide to formally develop an initial prototype into a completed product, they work with the WS Manager (WSM) to create a Product Proposal slide deck to present to the GA4GH Steering Committee (SC). The groups that explicitly request a product are considered the “source projects” for that product. Each proposed product must be associated with at least two source projects. The product proposal deck and presentation at SC serves to notify other WSLs as well as DP Champions (DPCs) of the nature of the product being developed and to identify potential areas of overlap. In conjunction with the Product Proposal deck, the WSM also supports the team in completing the Data Security Product Proposal Questionnaire (reviewed by the Data Security Work Stream [DSWS] to ensure relevant security risks are considered) and the Regulatory and Ethics Questionnaire (reviewed by the Regulatory and Ethics Work Streams [REWS]). If significant issues are raised by anyone on SC or during the DSWS or REWS reviews, amendments may be made to the initial Product Proposal. Following these actions, the product is incorporated into the GA4GH Roadmap.

Development Process

GA4GH consists of eight WSs—two foundational and six technical. Foundational Work Streams provide guidance across GA4GH in the areas of regulation, ethics, and data security in genomics. Technical Work Streams develop GA4GH’s technical standards and consist of leaders in their respective sub-disciplines from around the world. Technical WSs work to both discover and harden existing “working standards” and to create entirely new standards where needed. WS membership is voluntary and open to anyone. GA4GH welcomes additional interested individuals and organizations and encourages them to actively contribute across the organization by subscribing to emails, joining teleconferences, and attending face-to-face meetings.

GA4GH Subgroups are established within WSs to provide a focused space to develop and maintain specific deliverables. These groups also serve to guide future development activities and efforts to integrate the deliverable with other GA4GH products. Integration efforts also make use of Driver Projects, the Technical Alignment Subcommittee (TASC), and the Federated Analysis Systems Project (FASP) as ways of aligning and integrating deliverables.

Development work is carried out publicly and may use a system capable of tracking requests for changes, authorship, and history. In general, work such as program code, schemas, and documentation takes place in GitHub or Google Drive. Each WS and subgroup may have a unique approach to accepting changes, additions, and revisions to a standard, although most will recommend these be requested on GitHub trackers beginning with a pull request. See the GA4GH How We Work document for more details on the development process.

GA4GH specifications should meet the requirements of the source Driver Projects which requested them, or have had some feedback from the community in which it is intended to be
used. Development teams are encouraged to consider RFC 7282 from the Internet Engineering Task Force (IETF) to guide the decision making process.

Foundational Work Streams are contacted with Security, Regulatory, or Ethical concerns that arise during the development process.

Product Approval Process

When product leads and WSLs feel a GA4GH product is ready for a major version release, it is Submitted for Approval using the GA4GH Product Approval Submission Form. The GA4GH secretariat then notifies SC, as well as DSWS and REWS, which review the completed product using the DSWS and REWS review questionnaires. A Product Review Committee (PRC) is assembled, consisting of three members nominated by the submitter: (i) WSL from a different TWS, (ii) a member of a third Technical WS, and (iii) a representative from one of the product’s source DPs.

The review committee may give a response of “Accept”, “Reject”, or “Changes Requested”. All three members must agree unanimously for the committee to give a positive assessment. The committee should deliver their response one month after the specification is submitted.

If one of the three review bodies (REWS, DSWS, PRC) does not unanimously approve the product, requested upgrades will be communicated to the submitting WSMs. An upgraded product can be sent to the reviewers directly. This cycle can be repeated until the product passes the review body requirements.

Once all three bodies have approved the product, it is sent to the GA4GH SC for Approval. This must be done two weeks prior to the Steering Committee meeting at which it is to be assessed.

At the SC meeting, one WSL presents the product and the PRC may be called upon to explain the PRC decision. If SC votes unanimously to approve the product it becomes an Approved GA4GH Deliverable. If one or more SC members reject the product, the product may be required to make changes and then re-submit as described above. If the issues are minor enough, to allow the product to be re-considered in a single expedited review.

Approving New Versions

Minor and Patch updates to GA4GH products may take place without the Product Review Committee being re-convened. The approval of a new major version requires a new Product Approval submission, as described above.

Implementations

Each specification should have at least two associated implementations. These do not need to be officially maintained reference implementations. For a client-server model there should be at least two server deployments on separate implementation code bases interoperating successfully with two clients. These are viewed as best efforts and viewed as a way to ensure a robust product is developed. These implementations are not written or managed by WSs, but
rather by the Driver Projects and the community. Ideally, they are real world implementations of standards inside real-world environments that use real-world data. WSs facilitate interoperability testing between these implementations, which may lead to specification updates.

The Federated Analysis Systems Project (FASP) is a cross-WS initiative that aims to show that GA4GH APIs, when used in concert, can facilitate real-world, scientific use cases by conducting genomic analysis in the cloud. FASP aims to simulate how a researcher would search, access, and analyze genomic data within the GA4GH ecosystem via end-to-end test scenarios involving multiple Driver Projects. The FASP team also runs test scenarios against a wide variety of web service implementations, showing that common API specifications facilitate interoperability. This work involves the development of a comprehensive list of scientific use cases, as well as new web services for the test scenarios to run against. Learnings from FASP test scenarios are also fed back to the WSs and may lead to additional specification updates.

**Retirement Process**

WSLs may retire a product by submitting a request using the [GA4GH Product Retirement Form](https://example.com). Any attribution or mention of GA4GH will be removed if appropriate. Products will be updated to point to their replacement if they have been superseded by a newer GA4GH Approved Product. GA4GH may choose to fork a product if the Secretariat deem a need to.

**Glossary**

- DP = Driver Project
- DPC = Driver Project Champion
- DSWS = Data Security Work Stream
- FASP = Federated Analysis Systems Project
- IETF = Internet Engineering Task Force
- PRC = Product Review Committee
- REWS = Regulatory & Ethics Work Stream
- SC = Steering Committee
- TASC = Technical Alignment Sub-Committee
- WS = Work Stream
- WSL = Work Stream Lead
- WSM = Work Stream Manager
Supplemental Information for GA4GH: international policies and standards for data sharing across genomic research and healthcare

Author Contributions


**Formal Analysis:** A.Nisselle, D.F.V., D.L.C., J.M., R.J.M.


Supplemental Information for GA4GH: international policies and standards for data sharing across genomic research and healthcare

