GA4GH 2020-2021 Roadmap

Table of Contents

Executive Summary

The continued decrease in the cost of genomic sequencing has yielded research cohorts of hundreds of thousands of genomes; millions more samples are anticipated in the coming years from both research and healthcare[1].

However, major barriers in data sharing hinder effective use of the data, including: lack of data sharing mandates, difficulty in submitting data for sharing, inadequate resources for ingesting and storing data, insufficient consent for data sharing, lack of dataset interoperability due to disparate data models and terminologies, inconsistency in data-generating pipelines, and inability to address privacy issues and provide sufficient security[2].

In order to overcome these barriers, make the most use of the data, and fulfill the human right to benefit from scientific advances as stated in the *Universal Declaration of Human Rights*[3], the research and healthcare communities must come together to agree on common methods for collecting, storing, transferring, accessing, and analyzing data. Otherwise, they will remain siloed (e.g., by institution, country, disease area), locking away their potential to contribute to research and medicine. The Global Alliance for Genomics and Health (GA4GH) was established to address this need by cultivating a common framework of standards and harmonized approaches for effective and responsible genomic and health-related data sharing.

GA4GH standards aim to be interoperable with one another and the broader standards ecosystem in order to enable a future in which clinical geneticists can quickly and efficiently search across all of the relevant genomic data to reveal unanticipated gene-disease associations and solve previously impenetrable cases; clinicians can make otherwise impossible treatment decisions by accessing clinical decision support that is based on the world's best genomic knowledge; basic biologists and common disease researchers can interrogate cohorts large enough to achieve the power to detect all significant contributors to disease; and all qualified researchers—regardless of their means—can participate in genomics at a competitive pace. This

[1] Birney, Vamathevan, and Goodhand 2017
[2] Lewin et al. 2016
[3] Assembly 1948

ambition depends on data sharing across the globe as well as a federated system for searching, discovering, exchanging, and analyzing genomic and clinical data that is built on standards and interoperability frameworks embraced by the broad genomics and health community.

GA4GH has released 15 standards for APIs, data models, and more since rolling out the initial GA4GH Connect Strategic Roadmap in 2018. Collectively, these standards have been implemented or deployed by over 40 leading genomics institutions around the globe, including ELIXIR, the NIH All of Us Research Program, TOPmed, Genomics England, Australian Genomics, Illumina, Google, and Amazon Web Services. The GA4GH federated Systems Analysis Project (FASP) is an early step in demonstrating how multiple of these standards can be implemented in concert to achieve the ambitious vision described above.

All told, the promise of genomic medicine lies at a crossroads that depends on harmonization across the community to significantly enhance human health and medicine. The following Strategic Roadmap aims to enable that promise.

About this Document

The 2020-2021 Strategic Roadmap outlines strategies, standards, and policies for enabling genomic and related health data sharing. The GA4GH Strategic Roadmap is developed through both a bottom up and top down process:

The top down process is the GA4GH Gap Analysis process, which is executed every two years to identify gaps in the high level strategy of the organization. The 2020 Gap Analysis was led by Heidi Rehm (MGH/Broad Institute) and Andrew Morris (HDR UK), who communicated with more than two dozen stakeholders to identify three key community imperatives: (i) improve interoperability and alignment with external standards and between GA4GH standards, (ii) improve implementation support for technical standards, and (iii) engage more closely with healthcare and clinical standards.

This work informs the development of Part I of this document, the GA4GH 2020-2021 Strategic Roadmap.

The bottom up process consists of the GA4GH Work Streams developing and evolving their planned deliverables, informed through continuous engagement with Driver Project representatives and others from the community to identify the standards and policies most needed to enable the sharing of genomic and related health information within their local environments. This work produces Part II of this document—a list of proposed deliverables to be submitted for approval in 2020 and 2021. As a living document, Work Streams are able to add deliverables to this section at any time within the confines of the GA4GH Product Approval Process. Upon formal release in September 2020, this document includes 43 new or updated deliverables that aim to address needs in the areas of data use and researcher identities, variation representation and annotation, federated analysis, privacy and security, regulatory and ethics, discovery, clinical and phenotypic data capture, and large scale genomics.

This work informs the development of Part II of this document, the GA4GH 2020-2021 Product Roadmap.

Part I: Strategic Roadmap

Section Table of Contents

Overview of Gap Analysis Process and Outcomes

Every two years, GA4GH executes a gap analysis to identify areas of need that are not being adequately addressed by existing or planned GA4GH products, as well as feedback on current GA4GH activities and methods. The GA4GH Gap Analysis aims to (i) identify gaps that may not be identified through the Work Stream Roadmap process, (ii) identify new opportunities for GA4GH, (iii) improve internal and external interoperability and collaboration, and (iv) improve uptake and usage.

The gap analysis process includes round table discussions with groups of GA4GH Driver Project Champions and Work Stream Leads as well as a community survey. The 2020 gap analysis was led by Heidi Rehm (MGH/Broad Institute) and Andrew Morris (HDR UK), and consisted of 14 round table discussions and 59 survey responses (77% from academia, 15% from industry, and 7% from healthcare). From this work, the GA4GH community identified three strategic imperatives for GA4GH to focus on over the next two years: (i) improve interoperability and alignment with external standards and between GA4GH standards, (ii) improve implementation support for technical standards, and (iii) engage more closely with healthcare and clinical standards.

Community Imperatives Identified by Gap Analysis

1. Improve interoperability and alignment with external standards and between GA4GH standards

Significance

To effectively drive uptake, GA4GH must demonstrate how our standards work together and allow seamless support of genomic activities. . We must ensure our teams develop an interconnected suite of standards that are compatible and interoperable with each other and hardened for real-world use. We will identify alignment opportunities between GA4GH standards and support a centralized forum for discussing all ongoing GA4GH technical details.

Solutions include:

- Create documentation to support adoption that crosses deliverable boundaries and explains how to use standards together
- Ensure Driver Projects can provide opportunities to test the integrated deployment of standards

What Success Looks Like

Truly interoperable standards will enable solutions that can encompass multiple components of a pipeline and multiple platforms and use cases.

Tactics

1. Federated Analysis Systems Project (FASP)

FASP was established by GA4GH 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. Implementations of multiple GA4GH API specifications are planned as part of FASP, including Search, Passports, DRS, TRS, and WES. Tests will be run via the testbed infrastructure against a wide variety of web service implementations, showing that common API specifications facilitate interoperability. This work will involve the development of a comprehensive list of scientific use cases, as well as new web services for the test scenarios to run against. The FASP test scenarios will illustrate how having GA4GH standards solve a spectrum of challenges across the search-access-analyze workflow (e.g., datasets not discoverable, barriers to data access), driving larger scale and more powerful analyses.

2. Technical Alignment Sub-Committee (TASC)

TASC serves as a central decision-making group, including the documentation and communication of these decisions across multiple stakeholders. The TASC primarily functions to create consistency and technical alignment across the GA4GH Work Streams. For instance, the group may make internal technical recommendations that impact multiple Work Streams (e.g., best practices for namespacing, GitHub usage, common data elements), both in response to requests from Work Streams, as well as independently as proactive initiatives. The TASC also maintains and distributes technical resources to facilitate external alignment (e.g., list of technical contacts at other standards organizations). The TASC Team does not act as a product review gatekeeper, provide extensive engineering support to Work Streams, or define products.

3. SchemaBlocks

SchemaBlocks (https://schemablocks.org) is a “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics. Launched in December 2018, this community initiative aligns documentation and implementation examples provided by GA4GH members. While future products and implementations may be completely based on SchemaBlocks components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

2. Improve implementation support for technical standards

Significance

Implementations of standards, particularly those that serve the high priority needs of the community, are critical to inform development of and harden the standard, ensuring that it can solve a real-world problem. Implementations also serve to instantiate a standard by bringing awareness to the standards as well as forcing adherence through the need to enable downstream and interconnected functions.

Solutions include:

- Create tooling and testbeds
- Highlight end-to-end implementation demonstrations
- Increase facilitation, training, and support for implementations

What Success Looks Like

Driver projects and the expanding community are able to quickly adopt and implement GA4GH standards, driving broad uptake and subsequent interoperability across the community.

Tactics

We will provide training and support to developers working to implement a GA4GH standard, as described below. User support for researchers and clinicians interacting with those implementations remains the job of the individual platform developers.

1. Federated Analysis Systems Project

While FASP was established to support interoperability between standards, it will also indirectly support implementation of those standards around the globe. As groups actively work to become interoperable, they will naturally reveal pain points to implementation. This forum ensures specification conformance and a consistent approach to implementation. Additionally, the exemplar implementations that come out of FASP can be copied by other groups who may not have the capacity to start from scratch. Finally, FASP, which includes both small players and powerhouse genomics organizations, increases the likelihood of others seeing immediate benefit of implementation and demonstrates the value of working together to accomplish real world scenarios.

2. Documentation Hub

The GA4GH ecosystem of genomic standards will continue to expand as new standards and versions are released. To ensure we remain effective in disseminating our outputs clearly and comprehensively, the emerging GA4GH staff technical team will develop tools and web services to prepare and serve standards documentation. Automated documentation generation tools will consolidate API specification documentation in a single location and under a common style, enabling users to quickly access standards without having to hunt for them. Automated documentation will shorten development times, and ensure that documentation always accurately reflects the original standard. The Documentation Hub will also point to standards on GitHub and link out to different implementations that we track and endorse. This will allow researchers to browse standards and implementations and search for deployments. We will also publish how-to guides explaining how to accomplish certain use cases based on the scenarios outlined in FASP, for example, how to detect somatic mutations on cancer datasets using GA4GH services in the cloud. These guides will be hosted on our website.

3. Accessing GA4GH Resources Online

The GA4GH website (www.ga4gh.org) aims to present all GA4GH standards and frameworks in an accessible, easy-to-digest manner while also generating additional engagement between the Work Streams and potential technical contributors. All Work Stream meeting minutes and developer repositories are openly linked on the website, along with upcoming event information, conference reports, academic publications, information on becoming an organizational member and details about existing members, governance structure, and more. The website is hosted, developed, and maintained through in-kind support from the Wellcome Sanger Institute. Nearly 48,600 unique visitors explored the GA4GH website in 2019. We are also in the process of hiring a full time web developer who will ensure that the website remains up to date, is accessible in multiple languages, encourages broad participation, and makes it clear how GA4GH standards can work alone or in concert to benefit platform users (e.g., researchers and clinicians) and ultimately improve patient outcomes.

4. In-person Training Seminars

Live training seminars will provide guidance to engineers setting up systems that process genomic data using GA4GH specifications. Seminars will be developed as a series of modules, each focusing on a subset of GA4GH standards. In this way, we will be able to adapt the program to meet the specific needs of diverse audiences. Creating the practical dimensions of the courses, updating and adding modules as exposure grows, and developing supporting documentation will require significant developer time. The first iteration of the training program will take place in 2020 and will focus on the theme of ‘Setting up a Federatable Genomic Data Centre.’ Using the suite of GA4GH standards, this first seminar will demonstrate how to store and process sequenced patient data using standard pipelines, and make the data available in a manner respecting patient consent

5. Live Webinars

GA4GH virtual webinars are free and open to the public and aim to promote uptake of approved GA4GH deliverables. Webinars consist of presentations from the relevant developers as well as user experiences from Driver Projects that have already developed an implementation. Presentations are followed by question-and-answer periods during which anyone from the community who wishes may participate. Recordings of the events are made available on our website alongside other documentation about the relevant standard to help support other platform developers as they seek to implement.

6. The GA4GH ELIXIR Maturity Model

We are currently developing a maturity model (MM) with the ELIXIR genomic data projects with a focus on the development and implementation of GA4GH standards to enhance health-related data access throughout Europe. Ultimately, this model will be expanded to support GA4GH's global community to achieve the same goal. This ELIXIR::GA4GH Strategic Partnership Maturity Model consists of three parts.

1. **The Model**: A series of progressing levels, listing the GA4GH standards that are relevant to each. Each level of the model will include the GA4GH standards deployed and how that impacts the type, methods, and improvements of data access they have enabled.
2. **An Assessment Tool**: A tool to help projects determine their level of advancement in providing access to sensitive human data using GA4GH standards.
3. **An Advancement Guide**: Clear information on how to progress to the next maturity level. This will include training materials made available through collaboration with the ELIXIR Training Platform and the GA4GH Secretariat. Information about potential time, money, and other resources required will be clearly laid out.

Together these tools will provide the information needed to progress to the next level; ultimately aiding the ELIXIR network and the broader community in expanding its sensitive human data access framework through the inclusion of GA4GH standards.

3. Engage more closely with healthcare and clinical standards

Significance

GA4GH technical standards and policy frameworks aim to support a "learning health system" in which secondary use of patient data feeds into research, and the learnings from research reciprocally inform medical care. Historically, GA4GH has been well-connected to the research side of this virtuous cycle, however the diversity of the global clinical community has limited our ability to interface with the healthcare side. For GA4GH standards to be truly effective and for the organization to achieve its mission, we must overcome this limitation, which stems from a variety of origins, including, (i) the diversity of stakeholders within the healthcare community, (ii) the limited resources in healthcare to support research engagement, (iii) the regulatory need for locked down and standardized solutions means healthcare often buys rather than builds tools, and (iv) difficulty in finding the right points of engagement (eg., vendors/industry vs. clinicians).

Solutions include:

- Engage large scale initiatives focused on implementing genomics into clinical care as a cohesive, collaborative group
- Convene a broader group of clinically focused stakeholders to ensure relevance across the diverse community
- Establish formal relationships with clinically focused standards organizations

What Success Looks Like

Ability to efficiently and effectively respond to the needs of clinical stakeholders, developing standards that support the healthcare industry

Tactics

1. Genomics in Health Implementation Forum

GA4GH engagement efforts in areas impacting precision health must be high level, allowing individual initiatives to implement standards in a manner appropriate for their local context. With proactive engagement at a regional and organizational level, GA4GH aims to ensure that its standards are easily accessible and meet the disparate needs of the global community. In order to strengthen international collaboration between national genomic initiatives, GA4GH has recently formed the **Genomics in Health Implementation Forum** (GHIF) to support the implementation of GA4GH interoperability standards and frameworks as well as to identify new use cases that require GA4GH's attention. The GHIF builds on past activities of a subset of GA4GH Driver Projects—led by Australian Genomics and Genomics England—to convene thought leaders and domain experts from more than two dozen national and continent-wide genomics initiatives to promote knowledge exchange and collaboration as they pursue the common goal of advancing human health.

2. Healthcare and Clinical Advisory Group

While GHIF represents the core strategy for linking GA4GH to the world's disparate healthcare communities, its scope does not include clinical groups such as diagnostic laboratories, specialist clinicians, electronic health record vendors, and more. GA4GH will launch a Clinical Advisory Group to help ensure that GA4GH is connecting in all the right ways and with all of the right groups to effectively and comprehensively engage the complex international clinical community. This group will have five key goals: (i) identify clinically-relevant areas of focus missing from the current GA4GH roadmap, (ii) inform GA4GH standards to ensure they can be used in clinical settings, (iii) identify implementation opportunities within the clinical domain, (iv) create region- and sector-specific engagement strategies, and (v) align GA4GH's development activities with those of other clinically-focused SDOs (ie. CDISC, HL7). The group will be a loose affiliation of leaders and key stakeholders representing the diversity of the healthcare sector; It will not be a formal entity and will not have official leadership. This group will meet regularly and provide feedback to GA4GH executive committee on specific topics around its clinical engagement strategy.

3. Cross Standard Development Organizations Consortium (xSDO)

Like GA4GH, the International Organization for Standardization (ISO) and Health Level 7 (HL7) are international standards development organizations (SDOs) that actively develop standards related to genomics. Without intentional coordination to keep our respective products aligned, there is a risk of unnecessary proliferation of redundant standards, as well as the development of semantically- and syntactically-conflicting standards that will hamper large scale interoperability and introduce confusion within the adopter community. To mitigate this risk, GA4GH has committed—through its participation in the nascent Cross SDO (xSDO) consortium— to coordinate its activities and future roadmaps with those of other SDOs, including ISO Technical Committee 215 (ISO/TC215) for Health Informatics and HL7 Clinical Genomics (CG). This proactive coordination will help to ensure international coordination of genomics standards, particularly between Asia, Europe, and North America. In particular, the GA4GH Phenopackets standard has been approved as a work item in the programme for ISO/TC 215's new Sub Committee: Genomics Informatics. This work will increase the availability of standardized phenotypic information and expand the collection of use cases to develop a standard relevant to genomics communities internationally.

4. Phenopackets HL7 Fast Health Interoperability Resources (FHIR) Implementation Guide

GA4GH has been awarded a contract from the NIH National Library of Medicine (NLM) to re-develop the GA4GH-approved standard Phenopackets as an HL7 Fast Health Interoperability Resources (FHIR) Implementation Guide. FHIR is a standardized way of transmitting health data from one health information system to another through an application programming interface (API). The Implementation Guide will provide a set of rules for using FHIR resources to exchange phenotypic information. This work aims to (i) increase the availability of standardized phenotypic information, (ii) ensure that the use of FHIR for the exchange of clinical data meets the needs of genomics researchers and genomic medicine, and (iii) improve methods for clinical researchers to use electronic health records (EHRs) and other clinical data for medical research. This will enable the EHR community to extract, assimilate, and exchange genomic information from EHRs in a standard, efficient, and accurate fashion.

Part II: Product Roadmap

The true value of genomic data will only be realized when it can be responsibly shared between systems; however, this is currently difficult as each system has its own internal data models and terminologies. Similarly, the expected scale and changes in sequencing and phenotypic measurement technology will demand new standards to address a variety of use cases (e.g., structural variation, large-scale gene expression data, proteomic data, and long read genomic variation data). While many partial solutions have been developed, including clinical ontologies not specific to precision medicine, they are either applied in subtly different ways or come from legacy systems which originated in the Human Genome Project.

Individual systems around the globe will directly benefit from efficient data storage and harmonization; but to realize the full potential of each will require extensive collaboration and integration across many organizations, such as EHR vendors, patient registries, clinical research consortia, and standards bodies. To enable this federated ecosystem for data and analysis, GA4GH develops, coordinates, and provides implementation support for a suite of secure standards and frameworks for more meaningful research and patient data harmonization and sharing. The GA4GH suite of standards will be applicable across the world's accessible medical systems, knowledge bases, raw data sources, and patient-centered systems and will enable between-system comparisons.

Below, we provide an overview of the motivation, mandate, and vision for each of the eight GA4GH Work Streams and outline their planned technical and foundational deliverables for 2020-2021. These include entirely new standards, new major versions of existing approved standards, and optimization plans for existing approved standards through minor version updates. We provide expected submission dates and known implementations.

The GA4GH Product Roadmap is a living document that may be continually updated throughout the year, with Work Streams adding new proposed deliverables to their roadmaps and with proposed standards being approved by the GA4GH Steering Committee.

Section Table of Contents

Clinical & Phenotypic Data Capture

Motivation and Mandate

The widespread adoption of Electronic Health Records (EHRs), deep phenotyping methods, and improved patient generated multi-modal data provides an opportunity for information from genomics to be integrated into existing or emerging digital health infrastructure to support patient care. The existing health information infrastructure that needs to work with genomics includes the request for a genomics test, the sharing of the results from the test and the representation of genomics information in clinical and patient information systems.

This Work Stream will support the clinical adoption of genomics through establishing information models and standards to describe clinical phenotypes for use in genomic medicine and research, including the capture and exchange of information between clinical, research, and patient-centered systems.

A number of GA4GH Driver Projects are already developing the information infrastructure (forms, term lists, information models) which they are using to support the capture or sharing of information. This includes the forms which they are using to capture data on patients as they are sequenced as part of clinical demonstration projects. These examples will provide an important starting point for understanding the terminology and information models that are needed to describe a clinical phenotype to support clinical care and research.

Proposed Solution

The potential solution set for this Work Stream will include:

- Development of standard processes for defining a Reference Set of terms relevant for a particular disease or condition.
- A standard set of FHIR resources for describing a clinical phenotype.
- Standardized exchange formats for representing clinical data

Planned Deliverables

Pedigree V1

- **Type:** Data Model / Ontology
- **Expected Submission Date:** Q3 2021
- **Requesting Driver Projects:** Australian Genomics, Monarch Initiative, BRCA Exchange, Genomics England, GEM Japan

The need for high quality, unambiguous, computable pedigree and family information is critical for scaling genomic analysis to larger, complex families. Pedigree data is currently represented in heterogeneous formats that frequently result in the use of lowest-common-denominator formats (e.g., PED) or custom JSON formats for data transfer. The HL7 FHIR Family member history for genetics analysis profile supports pedigrees, but there is a need to add more data elements and rethink the data model to support a broader range of use cases. This will be accomplished by creating a minimum core dataset for family health history and developing the next version of a data model. Both will be evaluated and extended by GA4GH, extensions sent back to HL7 for inclusion into FHIR, and feedback gathered from G2MC. Standardizing the way systems represent family structure will allow patients to share this information more easily between healthcare systems and help software tools to use this information to improve genome analysis and diagnosis.

Phenopackets V2

- **Type:** Data Model / Ontology
- **Expected V2 Submission Date:** Q2 2021
- **V1 Approval Date:** 2019
- **Known V1 Implementations and Deployments:** Cafe Variome, AMED Biobank Network, RDConnect, EMBL-EBI (Biosamples), CanDIG/Epishare Metadata Service, Covidaware (Monarch Initiative/Pryzm Health)

Phenopackets is a uniform, machine-readable schema that enables the exchange of both high-level and deep phenotypic information. A phenopacket file contains a set of required and optional fields to share information about an individual, patient or sample phenotype, such as age of onset and disease severity. Phenopackets will allow phenotypic data to flow between clinics, databases, labs, and patient registries in ways currently only feasible for more quantifiable data, like sequence data, and power phenotype-driven diagnostics and computational analysis.

There are a few anticipated releases: 1) A version update to expand Phenopacket use in oncology and COVID-19 research, including better representation of time and additional fields for treatment course, exposure, and medical actions; 2) A core HL7 FHIR Implementation Guide; and 3) Integration of Variant Representation (VRS) from the GKS Work Stream.

Cloud

Motivation and Mandate

The GA4GH Cloud Workstream (CWS) helps the genomics and health communities take full advantage of modern cloud environments. Its initial focus is on ‘bringing the algorithms to the data,’ by creating standards for defining, sharing, and executing portable workflows. Standards under discussion include workflow definition languages, tool encapsulation, cloud-based task and workflow execution, and cloud-agnostic abstraction of secure data access.

Existing Standards

The CWS will build heavily on Docker for packaging of executables, and on existing text-based orchestration languages such as CWL and WDL for stitching those executables together.

Proposed Solution

The CWS will work with a variety of GA4GH Driver Projects including the NIH Genomic Data Commons, Genomics England, and other large-scale data processing efforts. These Driver Projects provide clear use cases for new standards, and deployment environments for specific implementations. As a result of our collaboration, the Driver Projects will have the ability to utilize standards to enable better tool, workflow, and data sharing with the larger community. Standards we push forward will address the following needs:

Defining portable workflows: tool builders need to be able to package their tools for reuse. The CWS will build on existing standards to allow workflows built by one researcher to be used by many others.

Sharing portable workflows: tool builders need to be able to offer their tools for others to use, and tool consumers need to be able to discover the tools they need. The CWS will support app-store-like functionality, including support for controlling access if tool builders choose.

Executing portable workflows: once a tool consumer has selected a tool, they need to be able to run it in their preferred compute environment, pointing at input and output data in their preferred storage environment. The CWS will define execution standards that will be easy for developers of existing workflow runners (e.g. Toil, Cromwell, Rabix) to support.

To ensure these standards meet the needs of GA4GH Driver Projects, the CWS will build a workflow portability testbed environment. Driver Projects will contribute one or two packaged workflows they care about, together with test input data and an output verifier. The CWS will then ask each Driver Project to run an instance of the testbed in their local environment, including contributing back patches to make it portable if needed, and to run all of the test workflows in all of their environments. Success will demonstrate the real-world usability and utility of Cloud Workstream standards.

Planned Deliverables

Cloud Testbed

- **Type:** Technical Toolkit
- **Expected Submission Date:** Q3 2021

GA4GH web services require compliance testing to ensure they fully conform with GA4GH API specifications. The GA4GH Testbed infrastructure will serve as a common platform to configure, schedule, and launch compliance tests against any web service implementing a GA4GH API. Test results will be loaded into a reporting service, which will display information on which implementations passed and failed certain test scenarios. The testbed infrastructure will extend not only to compliance tests, but also system performance benchmarking tests, and end-to-end tests, such as the Federated Analysis Systems Project (FASP), which simulate researcher use cases. The testbed will provide a comprehensive picture of how well our specifications are being implemented and identify when certain features are not implemented correctly. This will drive uptake by making it easier for developers to debug, maintain, and update their deployments, all while ensuring they remain conformant with GA4GH standards.

Task Execution Service API

- **Type:** API
- **V1 Expected Submission Date**: Q2 2021
- **Requesting Driver Projects:** ELIXIR Cloud, GeL, AGHA

Every compute environment has a different API for the batch execution of tasks. For example, each of the three major cloud vendors provides this service, but using completely different APIs. By providing a common interface that abstracts over their differences, compute engines can quickly move from one compute system to the next.

Approved Deliverables

Data Repository Service API

- **Type:** API
- **V1 Approval Date:** 2019
- **Known V1 Implementations and Deployments**: Cavatica, iRODS Consortium, Terra, Cancer Genomics Cloud (CGC)

The Data Repository Service (DRS) API, a standard for building data repositories and adapting access tools to work with those repositories, works with other approved APIs from the GA4GH Cloud Work Stream to allow researchers to discover algorithms across different cloud environments and send them to datasets they wish to analyze. The API allows data consumers to access datasets regardless of the repository in which they are stored or managed.

Tool Registry Service API

- **Type:** API
- **V1 Approval Date**: 2019
- **Known V1 Implementations and Deployments:** DNAstack, Terra

The Tool Registry Service (TRS) is a standard API for exchanging tools and workflows to analyze, read, and manipulate genomic data. The TRS API is one of a series of technical standards from the Cloud Work Stream that together allow genomics researchers to bring algorithms to datasets in disparate cloud environments, rather than moving data around. TRS gives researchers access to far more tools than they can presently use, and allows developers to register their products so that they are visible on a multitude of sites, expanding their audience reach. The API also provides a set of requirements for tool and workflow registries to implement TRS.

Workflow Execution Service API

- **Type:** API
- **V1 Approval Date:** 2018
- **Known V1 Implementations and Deployments:** Broad Institute, TOPMed, Human Cell Atlas, All of Us, Australian Genomics, Genomics England, ELIXIR

Portable tools — the ability to execute a single analysis in a variety of environments — allow researchers to work with more data from more sources, and tool builders to support more researchers and more use cases. The Workflow Execution Service (WES) API provides a standard for exactly that. This API lets users run a single workflow (defined using CWL or WDL) on multiple different platforms, clouds, and environments, and be confident that it will work the same way. The API provides methods to request that a workflow be run, pass parameters to that workflow, get information about running workflows, and cancel a running workflow.

Data Use and Researcher Activities (DURI)

Motivation and Mandate

At a concrete level, data from human subjects has two axes of access control:

Researcher Identity: These specify the collection of researchers that may access the dataset at any given time, and the credentials they must supply. For example, it may be the case that only researchers that are members of a consortium may access a dataset for the first year after generation.

Data Use: When human subjects are consented as participants in a study, the informed consent form specifies appropriate restrictions on secondary data use. For example, it may stipulate that the data may be used only for “cancer research in a non-profit setting.” Similarly, data owners may place additional restrictions on data use.

Each of these axes is independent—a researcher may have access to a dataset, but be unable to utilize it because her research purpose is inconsistent with the data use restrictions. Similarly, a researcher’s purpose may be entirely consistent with the data use restrictions but, because she is not a member of the consortium, she may not be able to access it.

The mandate of the Data Use and Researcher Identities (DURI) Work Stream is to create those standards required to facilitate both of these axes of access control.

Existing Standards

Important work has been done within GA4GH and beyond along both of these axes, including:

- The Library Cards and Bona Fide Researchers efforts to define researchers and their identities.
- The eRA Commons, ORCID, and EGA systems of identities

Proposed Solution

The DURI Work Stream will drive progress in the following two areas:

Establish researcher identities – The world is in need of i) a consistent definition of who a bona fide researcher is in the physical world, ii) one or more identity providers that respect this definition and provide identities in the virtual world that travel with the researcher across various data sharing repositories.

Specify a data use ontology – This ontology will be used to both state the secondary data use restrictions for datasets, as well as researchers’ purposes for wishing to access them. By expressing them in an ontology, it becomes possible to compute whether a given researcher’s purpose is consistent with a given data use restriction.

Approved Deliverables

Data Use Ontology

- **Type:** Data Model / Ontology
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments:** NIH All of Us Research Project, Broad Institute, National Cancer Institute, TOPMed, European Genome-Phenome Archive

DUO allows semantic tagging of datasets with restrictions and permissions about their usage, making them automatically discoverable based on the users' authorization level of users, or their intended data uses.

Machine-readable Consents

- **Type:** Guide
- **V1 Submission Date**: 2020
- **Known Implementations & Deployments:** Australian Genomics, Broad Institute

To maximize ethical data sharing, integration, and re-use while respecting data subjects' autonomy. Adopt data sharing consent language that unambiguously maps to DUO to facilitate data discovery, facilitate data access request submissions and approvals, This is a cross workstream product in collaboration with the Regulatory and Ethics group.

GA4GH Passports

- **Type:** Data Model/Ontology, Protocol
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments:** ELIXIR, Google Cloud Platform

GA4GH Passport specification aims to support data access policies and procedures within current and evolving data access governance systems. This specification defines Passports and Passport Visas as a standardized method of communicating the data access authorizations that a user has based on either their role (e.g. researcher), affiliation, or access status.

Planned expansion on v1.0 includes additional role values, connections with other GA4GH APIs, and additional guidance as required by Drivers adopting the standard.

- DURI, Discovery, Cloud X-WS use case implementations leveraging Passport and DUO standards

Discovery

Motivation and Mandate

We are in an era of abundant genomic information fueled by steadily decreasing sequencing and processing costs and service platforms that ease analysis. These critical resources are spread throughout the world and are increasingly challenging to aggregate for a multitude of reasons, including scale, regulatory differences, and data harmonization across information arising from diverse origins. We believe a solution to this challenge is to facilitate the discovery and utilization of these varied data sources and services via standard APIs and context-aware user interfaces. The Discovery Work Stream aims to create a unified data discovery platform to make it easier to find and use data, tools, and infrastructure for genomics and clinical analysis.

Existing Standards

Organizations such as the Matchmaker Exchange, the Beacon project, BRCA Exchange, and many others approach fragmented and diverse data sources by locally aggregating, harmonizing, and redistributing processed data through web-based user interfaces and standardized APIs. Unfortunately, each has its own data sharing formats and sharing nuances. These cause difficulties and inefficiencies to the consumer in gaining synergistic value by cross referencing and utilizing these invaluable resources. Further, diverse datasets arising from different sequencing and processing technologies as well as overlapping samples add to interpretation challenges.

Proposed Solution

The Discovery Work Stream proposes a unified interface that acts as a facade to a varied dynamic collection or registry of data sources and services, forming an interconnected ‘Internet of Genomics Data and Services.’ The network’s data sources and services can be crawled and indexed, exposing a single standardized API endpoint that a unified web interface can aggregate and present in a context-aware, meaningful manner. To achieve this, the Work Stream will design a suite of standards that:

- are easy to implement with a community-maintained reference implementation.
- reflect the context of the data that it shares.
- reflect the nuances in data sharing preference.
- leave room to include information from meta-sites, such as DUOS, to help with usage.

Planned Deliverables

Beacon API v2

- **Type:** API
- **V1 Approval Date:** 2018
- **V2 Expected Submission Date:** Q3 2021

- **Requesting Driver Projects:** ELIXIR, ENA/EGA/EVA, VICC, BRCA Exchange, Genomics England, SPHN, AGHA, Autism Speaks, EUCANCan

Evolving Beacon from discovering/sharing variant data to (safely) discover and share also entities related to variant and genomic diagnoses.

Discovery Search API

- **Type:** API
- **V1 Approval Date:** 2018
- **V2 Expected Submission Date:** Q3 2021
- **Requesting Driver Projects:** ELIXIR Beacon, ENA/EGA/EVA

The GA4GH Search API enables a search engine for genomic and clinical data by providing specification for query language across genomic, phenotypic, and clinical data that can be used to implement, for example, Beacons and Matchmakers, but also other applications (e.g. diagnostics, pharmacogenomics, family analysis).

Approved Deliverables

Service Registry/Service Info

- **Type:** API
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments:** ELIXIR, Autism Sharing Initiative

The Service Info/Registry provides a digital network infrastructure for a proposed “Internet of Genomics”. The registry lists GA4GH services (e.g. Beacons, DRS, etc.) or other registries (e.g., Matchmaker Exchange) that have been registered to it. The Service Info/Registry allows for dynamic registration and on-demand discovery of online GA4GH APIs (data, tools, services) to enable their realtime discovery and use.

Genomic Knowledge Standards (GKS)

Motivation and Mandate

Genomic data analysis and interpretation is at the heart of enabling genomic data to improve human health. Many developed analyses require locating interesting and potentially causative changes in genomic sequence before attempting to categorize, rank, and prioritise potential leads by intersecting patient data with known reference data sets. All analysis methods develop their own solutions to access reference genomic sequence, find and use baseline reference genomic annotation (e.g. genes, variations, regulatory regions, expression), integrate and find equivalence with other resources, model data, and distribute the results of said analysis to downstream consumers—be they human or computational. In addition, the provenance of annotation can be unclear and associated metadata may be unstructured. Results may not be directly comparable between two resources due to ambiguity in data representation, semantics, and provenance.

Existing Standards

VMC (Variation Modelling Collaboration) is a specification, now at version 0.1, for modelling simple variation and was developed by members of the Variation Annotation Task Team (VATT). FHIR (Fast Healthcare Interoperability Resources) is a specification to enable the transfer of healthcare information over standard APIs. In addition a number of GA4GH standards for modelling ontologies, genomic annotation and RNA quantification have been developed as part of the schema/reference/compliance suite of applications.

Proposed Solution

The Genomic Knowledge Standards Work Stream (GKSWS) aims to develop, adopt, and adapt standards-based components to enable the exchange of reference genomic information through common APIs, thereby enabling the downstream analysis of genomic data. It will focus on developing specifications related to genomic sequence, annotation, and associated metadata/provenance.

GKSWS will engage with GA4GH Driver Projects, including analysis tool developers/consumers (VICC, GEL) and reference data providers (ClinGen, Ensembl), to ensure that standards-based solutions to data access and exchange are developed based on real-world use-cases whilst also being applicable to more generalized scenarios. GKS will work closely with other GA4GH Work Streams (Large Scale Genomics, Discovery) in areas of common interest to move standards into production (VMC), and we will partner with external standards development organizations to leverage existing specifications and to ensure GKSWS-developed standards are suitable to healthcare environments (HL7, FHIR).

Planned Deliverables

Variant Annotation

- **Type:** Data Model/Ontology
- **Expected V1 Approval Date:** Q2 2021
- **Requesting Driver Projects:** ClinGen, VICC, Genomics England, BRCA Challenge, Monarch Initiative

The VA Specification (VA Spec) will define extensible data models to support representation of diverse kinds of statements made about genetic variation, and the evidence and provenance supporting these statements. The specification will include information models, message exchange schema, and a formal framework for defining custom extensions to the core model. A more detailed description of these components can be found here.

Approved Deliverables

Variant Representation

- **Type:** Data Model/Ontology, Protocol
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments**: ClinGen, VICC, BRC Exchange

The VR specification is a standardised extensible model for computational variation representation. The specification includes a data model, message schema, and methods for normalization of variants and computing identifiers. Any updates to concepts within the specification include updates to all of the previously stated aspects.

The VR roadmap will extend our 1.0 release to define and model classes of variation in support of complex and aggregate variation.

Large Scale Genomics (LSG)

Motivation and Mandate

High-throughput sequencing projects continue to produce data at a massive and still accelerating scale. The generated information is having an impact on everything from basic science to healthcare provision. The challenges for large scale genomics are clear. The vast quantities of raw sequencing data and derived results mean that we must continue to develop highly efficient and standardised formats and API interfaces to store, access, and analyse sequencing reads, genetic variation, and gene expression information. We must continue to adapt and evolve genomics formats as new sequencing assays (e.g. single-cell, methylation/base modifications), platforms (e.g. long reads), and increased scale (e.g. population scale variation - UK Biobank) are created to ensure the data remains interoperable. Perhaps the most significant challenge is to move from the traditional purely file-based approach to storing, accessing, and analysing genetic data, to one where we are presented with standardised API interfaces.

Existing Standards

There are a variety of incumbent file formats for read sequencing data include SAM/BAM/CRAM, VCF/BCF for genetic variation, and tabular formats for expression and genomic ranges (e.g. BED). Large cohorts of genetic variation data at increasing scale are now available from multiple different resources (e.g., ExAC/gnomAD, dbSNP, 1000 Genomes, UK Biobank, EVA, EGA).

Proposed Solution

As genome sequencing becomes integrated into national and regional healthcare initiatives, it is not realistic to assume that all human genetic and phenotypic data will be stored in a small number of large repositories. Carrying out queries remotely across these repositories opens up the possibility of making new disease associations without the need to physically download all of the data to a single location. Reliably processing and managing information at this scale requires robust software architecture and widely supported standards. The Large Scale Genomics Work Stream engages sequencing vendors, key sequencing and bioinformatics tool developers, and population scale driver projects to ensure that the primary data formats and libraries are evolved and adapted to meet this need. It also coordinates closely with a variety of Driver Projects to support adoption and implementation of APIs for access to large scale projects or databases. The guiding principles of this workstream will be:

- Engage with driver projects and the wider genomics community to identify requirements and use-cases.
- Build on existing standards to ensure a gradual transition to new standards.
- Engage with key community software tool maintainers to drive adoption of standards.
- Engage with key large data repositories to drive community adoption.
- Metric for workstream success will be adoption of standards.

Approved Deliverables

Genetic Variation Formats (VCF)

- **Type:** File Format
- **V4 Approval Date:** 2017
- **Known Implementations & Deployments (selected):** GATK, htslib, samtools, HTSJDK, ENA-EGA-EVA, GEM Japan, AGHA, Genomics England,

VCF is a text file format (most likely stored in a compressed manner). It contains meta-information lines, a header line, and then data lines each containing information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

This project focuses on improvements to the VCF format (e.g. representation of structural variants, scaling to population size collections of genotypes) as well as looking into potential future formats to handle the increasing scale of large sequencing projects more efficiently.

Read Data Formats (SAM/BAM/CRAM)

- **Type:** File Format
- **V3 Approval Date:** 2017
- **Known Implementations & Deployments (selected):** EMBL-EBI, Australian Genomics, Broad Institute, H3Africa, Wellcome Sanger Institute, Genomics England, Illumina, TOPMed, SciLifeLab

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. The BAM file format is binary equivalent of the SAM format. CRAM is a related compressed columnar file format that uses optionally differences to a genomic reference to reduce storage cost. The vast majority of sequencing data produced worldwide is stored in the using the SAM representation, underscoring its importance as a key format for both research and healthcare genomics.

This project describes ongoing additions to SAM/BAM to improve efficiency and support new sequencing platforms. For CRAM the primary focus is on improved efficiency, random access for long reads and CRAM file sizes for short read data (with an expectation of 15-20% reduction). Additionally improved indexing, and especially index documentation.

htsget API

- **Type:** API
- **V1 Approval Date:** 2017
- **Known Implementations & Deployments:** Australian Genomics, Genomics England, Broad Institute, European Nucleotide Archive, DNAnexus, European Genome-Phenome Archive, Google Cloud Platform, Wellcome Sanger Institute, University of Oxford

htsget is a protocol for secure, efficient and reliable access to sequencing read and variation data. This project describes ongoing improvements to the protocol and implementations to support 'two-dimensional' slicing of very large variant (VCF) datasets, e.g., N=1,000,000 WGS. Such datasets need to be accessed efficiently not only by genomic range, but also by subsets of the cohort -- potentially arbitrary client-determined subsets. To support the more-complex queries involved, this will imply evolving the basic protocol to support “POST” HTTPS requests in addition to the “query string” request format used currently.

Reference Sequences (Refget)

- **Type:** API
- **V1 Approval Date:** 2018
- **Known Implementations & Deployments:** Broad Institute, European Genome-Phenome Archive, Amazon Web Services

Currently, refget API focuses on single reference sequences uniquely identified by their checksums. The refget API service offering will be extended to reference sequence collections, e.g. genome assemblies, and to reverse lookup of reference sequence. Like individual reference sequences, reference sequence collections would be associated with unique identifiers computationally derived from the set of sequences themselves. The reference sequence collections are envisioned to make it easier to share and exchange commonly used and semantically meaningful sets of sequences. The reverse lookup of reference sequences would provide an alternative to checksum based services, allowing commonly used sequence names such as 'Chromosome 1' from human genome assembly 'GRCh37.p13' to be used. As the reverse lookup of individual sequences typically requires the reference sequence collection to be also specified, we envisage that the reverse lookup service may also be extended to reference sequence collections.

Both systems require development of a transmission format to communicate these data/concepts between clients. An additional API will be developed to support ambiguous reverse lookup systems developed from said formats.

Genetic Data Encryption (Crypt4GH)

- **Type:** File format
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments:** European Genome-Phenome Archive, htslib, HTSJDK

Crypt4GH is a file format that can be used to store data in an encrypted and authenticated 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.

RNA Data (RNAget) V1

- **Type:** API
- **V1 Approval Date:** 2020
- **Known Implementations & Deployments:** CanDIG, California Institute of Technology, Centre for Genomic Regulation (CRG), ENCODE

The RNAget API describes a common set of endpoints for search and retrieval of processed RNA data. This currently include 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.

Data Security

Motivation and Mandate

An international consortium federating large volumes of sensitive clinical and genomic data across virtual computing environments presents formidable challenges in assuring data confidentiality, data integrity, service availability, and individual privacy. The fact that healthcare data are a leading target for cyber-security attackers exacerbates these challenges.

GA4GH and its partners must implement defense in depth to protect the high-value data we rely upon to accelerate the acquisition and application of biomedical knowledge. A key mandate of the Data Security Work Stream is to help assure that the standards produced by the Technical Work Streams have been developed within a sound risk-management framework.

Existing Standards

Some of the security challenges GA4GH faces call for innovative application of well-established security standards and protocols, such as identity federation on a global scale, using OpenID Connect; distributed authorization using OAuth 2.0; transmission protection using Transport Layer Security (TLS), and data encryption using symmetric encryption algorithms such as Advanced Encryption Algorithm (AES). Other challenges require solutions still emerging from security research, such as privacy-preserving data linkage, homomorphic encryption, and quantum key distribution.

Risk management is central to the Data Security Work Stream’s standards-development process, which seeks to leverage industry standards and best practices wherever possible, including GA4GH-specific profiles of existing standards.

To enable GA4GH and its partners to effectively prevent and respond to breach attacks requires a layered and proactive scheme to identify potential threats and vulnerabilities, continuously monitor the use of data and services, detect potential attacks, and collectively respond to potential breaches. The Data Security Work Stream will work with the Driver Projects to broadly apply breach-response methods currently in use to collaboratively protect collective data assets.

Proposed Solution

The remit of the DSWS includes, but is not limited to, identity management, access authorization and control, privacy-preserving computation, non-repudiation, accountability, service continuity, and breach detection and response. High-priority needs include:

- Standard templates to support “gatekeeper” function
- Standard profiles of OAuth 2.0 and OpenID Connect standards for authorizing access and federating authentication across GA4GH (incorporating vocabulary being developed by Data Use and Researcher Identities (DURI) work stream)

- Standard operating procedure for collaboratively detecting and responding to breaches

Planned Deliverables

Risk Assessment Methodology for Software Stacks

- **Type:** Guide
- **Expected Submission Date:** 2021
- **Requesting Driver Projects:** Foundational

In the community of genomics, many groups lack training in security assessments and the followup of security best practices. This deliverable will be multiple parts:

1. A how-to on assessing risk aligned to a known framework. This will include both formal alignment as well as “colloquial” alignment so that a typical software developer can use it to assess their product.
2. Methodology in assessing what information needs to be protected at what levels. Data is all different and having “rules” for data classification is helpful.
3. If U24 funding is approved, starting up a group that would actually do assessments and provide for a stream of open source tooling to increase the automation of these assessments.
    a. This would be a group that would also train other groups and encourage using self-service tools.

Cloud Security and Privacy

- **Type:** Guide
- **Expected Submission Date:** 2021
- **Requesting Driver Projects:** Foundational

The Cloud has gained the attention of many GA4GH projects, and it is becoming increasingly used for large-scale distributed computing services. The Cloud Work Stream emerged to focus on API standards to make it easier to send the algorithms to the data in such environments, and run full workflows on the cloud.

The use of Cloud services (and outsourced services in general), poses multiple legal, ethical and technological challenges in terms of data transfer and processing, for which GA4GH should develop appropriate specific guidelines and recommendations for a secure and privacy-conscious use of Cloud services. An example issue to be addressed is the recommended policy for cryptographic key management.

Approved Deliverables

Authentication and Authorization Infrastructure (AAI)

- **Type:** Guide
- **V1 Approval Date:** 2019
- **Known Implementations & Deployments:** ELIXIR, Google Cloud Platform

The GA4GH Authentication and Authorization Infrastructure (AAI) Profile is a technical profile for managing and authenticating the identity of users, and for authorizing access requests for data and services offered through the Driver Projects. The GA4GH AAI Profile is based on the IETF OAuth 2.0 standard, and the OpenID Connect identity layer based on OAuth 2.0, and incorporates the researcher identity vocabulary and data-use ontology developed by the Data Use and Researcher Identity (DURI) work stream.

Regulatory & Ethics (REWS)

Motivation and Mandate

Data sharing bridges the genomic-clinical divide, thereby enabling translational medicine. Current frameworks for research ethics governance, information governance, and data privacy protections, however (and among other things), may frustrate the desire of individuals to share such data, both within and across jurisdictions.

The internationally-recognized human right of everyone to benefit from the progress of science and its applications can serve to break open current barriers. The primary role of the Regulatory and Ethics Work Stream (REWS) is to “activate” this human right—to promote forward-looking data governance through the Framework for Responsible Sharing of Genomic and Health Related Data, harmonized across countries, sectors, and institutions.

Existing Standards

The REWS continues to elaborate on the Framework with policies and tools found in the REWS Toolkit on consent, privacy and security, accountability, and ethics review equivalency. The REWS also plays an important support role within the GA4GH. It regularly assesses the ethical and regulatory implications of GA4GH Work Streams and work products. It liaises with Driver Projects to identify common and emerging issues and to harmonize real-world governance.

Proposed Solution

In the next five years, the REWS aims to reduce the gap between the development of policies and actual implementable gene protocols and associated applications by responding to direct needs that were cited by the GA4GH community such as consent tools for clinical sequencing to enable use of data and health information in research. As part of the GA4GH work on harmonization of consent, the REWS will develop generic consent clauses for genomic research, large scale initiatives, clinical WGS and a typology of familial consent clauses. Recommendations from the multilingual, international Public Attitudes Survey (Your DNA Your SAY) will be incorporated in these clauses. These tools will help researchers when drafting consent forms so they can use language matching cutting-edge GA4GH international standards. Further, in collaboration with DURI, the REWS participated in the creation of a Machine Readable Consent Guidance document, associated with DUO, so researchers have the opportunity to articulate consent clauses in a machine readable way to promote efficient and accurate data sharing. It also permits aligning legal and technological terminology. In a sustained effort to provide best practices across sectors and jurisdictions, the REWS has established a Data Access Committee Review Standards (DACReS) team. The purpose of this initiative is to draft procedural standards and guidance to improve consistencies in Data Access Committee reviews, as well as their quality and effectiveness at ensuring adequate research data protections. The community also asked for a standardized approach to handling return of results since policies are highly divergent across sectors and jurisdictions. The REWS has contributed to completing a survey of stakeholder perspectives and is now in the process of drafting an aspirational policy on Genomic Return of Results in Research Studies. Noteworthily, the latter receives support from the GHIF community. Another way to reduce the gap between theory and practice in a healthcare setting while ensuring that governance serves the individuals providing the data is achieved through the multilingual, international Public Attitudes Survey (Your DNA Your Say). The Participant Patient Public engagement policy further explores solutions to build and sustain trusting relationships between participants, patients, publics and researchers. Finally, the Standard Genomic Data Licenses and Agreements led by GEM Japan marks the first step for a GA4GH intellectual property suite of external tools which will be based on values of interoperability and openness that enables sustainable data science and innovation.

Planned Deliverables

Consent Toolkit

- **Type:** Guide
- **Expected Submission Date:** Q4 2021
- **Requesting Driver Projects:** Genomics England, All of Us, Australia Genomics, Wellcome Genome Campus (Your DNA Your Say)

We envision a future where genomic variant testing will be part of standard medical care. The REWS Toolkit aims to contribute to the discussion in genomic medicine on consent, familial interests and data sharing. Guidance on the integration of genetics in medical care serves the needs of both clinicians and patients. How will their genomic data be shared (or not) with family members, health professionals or researchers?

- **Consent Clauses for Genomic Research**: The first part of the Toolkit consists of updating the GA4GH 2015 genomic research consent clauses. A categorisation of generic consent clauses in genomic research is available to researchers in the format of a compendium of approved template clauses (2020). Please see the appropriate section below.
- **Consent Clauses for Large Scale Initiatives**: The second part of the Toolkit consists of collecting and analysing large scale initiatives' consent forms for biobanking and population studies. This catalogue of consent clauses will be distilled and standardised into generic consent clauses for large scale initiatives and will enable researchers to build consent forms matching their needs. The trends observed in the consent clauses in biobanking and population studies will also lead to a narrative analysis to be submitted for publication.
- **Familial Consent Clauses**: The third part of the Toolkit consists of generic familial consent clauses that cross the research and clinical settings as well as biobanking and population studies. The familial nature of genomics has implications for consent forms such as possible direct communication of genetic test results to family members and greater familial involvement. Thus, this part of the Toolkit focuses on creating a typology of familial clauses from their descriptive to their normative nature. The typology was published in *the American Journal of Bioethics*:

    - Knoppers, Bartha Maria, and Kristina Kekesi-Lafrance. "The Genetic Family as Patient?." The American Journal of Bioethics 20:6 (2020): 77-80. doi: 10.1080/15265161.2020.1754505. Generic consent clauses are under development.
- **Clinical (WGS) Consent Clauses**: As whole genome sequencing enters routine medical care, other issues emerge such as the inclusion of WGS data in medical records, recontact options for research and anonymized deposit in variant databases. For this fourth part of the Toolkit, consent forms for genetic testing in the clinic will be collected and analysed. Their distillation will result in generic clinical consent clauses for whole genome sequencing that can be customized according to local and national requirements. The trends observed in clinical consent clauses will lead to a narrative analysis to be submitted for publication.
- **Consent Clauses for Rare Disease Research**: Finally, model consent clauses for rare disease research were presented to the international genomic community in 2019 on behalf of the IRDiRC-GA4GH Model Consent Clauses Task Force through the publication of:
    - Nguyen, Minh Thu, … Knoppers, B. M., et al. "Model consent clauses for rare disease research." BMC Medical Ethics 20:1 (2019): 55. doi: 10.1186/s12910-019-0390-x.

Standard Genomic Data Licenses and Agreements

- **Type:** Policy Framework, Guide
- **Expected Submission Date:** Q4 2021
- **Requesting Driver Projects:** BRCAExchange, VICC, Genomics England, Human Cell Atlas, Monarch Initiative, GEMJapan, IHEC

The time is ripe to consider licensing issues, as biomedical big data expands across research domains, sectors, and countries; as data scientists and ML/AI-companies seek to integrate and analyse multiple datasets to train and commercialize algorithms; and as technologies like APIs change the dynamics of data control and sharing.

The aim of the new forum will be to discuss international best practices for licensing data and other scientific resources. The initial focus will be on intellectual property considerations (namely copyright and database rights), in the context of publicly available biological and biomedical knowledge-bases. The forum will also explore metrics, educational tools, and guidance to help data scientists, stewards, repositories, and aggregators select a license, with a view towards interoperability and openness that enables sustainable data science and innovation.

Initially, a forum will be established that will serve as an informal, networking opportunity consisting of a series of presentations from our members or external groups working on data licensing issues.

The forum could later expand to look at:

- Standard licenses for public access data.

- Bilateral data transfer agreements (i.e. data use or data access agreements) in the genomics or biomedical big data context.
- Licensing issues relating to data sharing, data citation, data privacy, research ethics, and international contracting practices.
- Policy and infrastructure needed to achieve standardization of data licensing practices, including funding agency and institutional support, or an international host organization to promote standard licenses, analogous to the Open Source Initiative.

Participant Patient Public Engagement Policy

- **Type:** Guide, Policy Framework
- **Expected Submission Date:** Q4 2020
- **Requesting Driver Projects:** Foundational

Building and sustaining trusting relationships between participants, patients, publics and researchers is foundational to successful genomic science: without participants there are no data and no samples. Maintaining public trust is an ongoing imperative for GA4GH's research, clinical and industry communities and there is wide recognition of the value of participants and public knowledge in shaping the policy and governance in genomics ('how' publics see Trust, is being evaluated by the Your DNA Your Say Survey). In this fast-moving field, where rapid technological development may move more quickly than our ability to appreciate its societal impacts, strong relationships with publics and participants promises to support the alignment of scientific endeavor with societal values and norms.

While building relationships and engaging with participants and the public is undoubtedly key to maintaining trust, there is less clarity about how to most effectively use and evaluate the range of engagement methods available, to what purpose and to what effect in diverse settings across the globe.

The aim of this project is to collect and analyze existing engagement practices and policies in the GA4GH Genomics in Health Implementation Forum (GHIF) as well as Driver projects and other relevant projects worldwide to identify gaps and potential for further innovation. This will enable us to develop a guide to engagement for genomics research which accounts for diversity and a GA4GH policy framework for good practice.

These activities would then be combined in an academic paper that would present the results of our research and provide best practices for institutions who are seeking guidance.

GDPR Forum

- **Type:** Ongoing Project
- **Requesting Driver Projects:** Foundational

Forum on the General Data Protection Regulation (EU) 2016/679 (GDPR) co-lead by Michael Beauvais (McGill University) and Fruzsina Molnar-Gabor (Heidelberg University). Launched in August 2018 as part of the REWS, the purpose of the Forum is to discuss and disseminate information regarding European data protection law – specifically the GDPR – and its impact on international genomic and health-related data sharing. Monthly one-page GDPR Briefs provide concise and timely guidance on specific topics relevant to health researchers, institutions, ethics bodies, and data sharing initiatives internationally. The Forum has been very successful as demonstrated by these charts. As of September 2020, the Forum has published 27 Briefs, the full list of which can be found here.

Responsible Data Sharing to Respond to the COVID-19 Pandemic: Ethical and Legal Considerations

- **Type:** Guide, Other: Manuscript/Article
- **Expected Submission Date:** Q4 2020
- **Requesting Driver Projects:** Foundational

As an initiative of REWS, the document Responsible Data Sharing to Respond to the COVID-19 Pandemic: Ethical and Legal Considerations examines the intersection of law and ethics with COVID-19-related research. The document aims to specifically detail changes to ethico-legal frameworks, as well as give guidance regarding pre-existing provisions that may be particularly useful in the context of COVID-19-related research (e.g., the availability of consent waivers where there is great public interest in the research).

With much of the initial dust having settled, the document will be edited for clarity and consistency, before being archived as a community resource.

Participant Values Survey (Your DNA Your Say)

- **Type:** Ongoing Project
- **Requesting Driver Projects:** Foundational

Global online survey gathering public attitudes towards genomic data sharing using approachable, innovative films. Now have 37,000 completed surveys from 22 countries in 15 languages.

The study outcomes can be incorporated into GA4GH tools such as the Consent Toolkit (Consent Clauses for Genomic Research, Consent Clauses for Large Scale Initiatives, Familial Consent Clauses and Clinical (WGS) Consent Clauses) and the Participant Patient Public Engagement Policy.

Return of results - Survey of Stakeholder Perspectives

- **Type:** Policy Framework
- **Expected Submission Date:** Q4 2020
- **Requesting Driver Projects:** Foundational

Stakeholder views on the return of results from genomic research: a systematic review of quantitative and qualitative studies.

The main research question is: What are stakeholders’ views and experiences regarding return of individual results from genomic research?

Follow on questions:

- Do stakeholders want to return/receive ‘individual findings’ and if so, on what basis?

- What types of results do clinicians and health providers want returned to their patients?
- Is there a sufficient evidence base for informing policy on return of individual results in genomic sequencing in the research context? If not, what are the under-researched areas?

Two questions not ultimately considered due to limitations in the data collected and synthesised during the review:

- How does this differ for secondary sharing of genomic data (i.e. data collected previously but used to answer new research questions)?
- What are stakeholders' views on return of clinically significant findings, where informed consent for individual return of results was not sought at the time of data collection?

The final review included 184 peer-reviewed publications reporting empirical findings from 2005 to April 2020. The two most represented stakeholder groups were participants from clinical research studies and members of the public. Other stakeholder groups included: participants in biobanks; patients; researchers; health care professionals; and members of IRBs. Around two thirds of the papers were based on studies conducted in North America, primarily involving White/non-Hispanic populations with higher than average levels of income and education.

The evidence showed a high desire for return of results from genomic research from those who would receive them, and a general willingness to provide such results by those who would do so, although there is an acknowledgement that the focus should be on returning results that are reliable and clinically relevant. The outcomes from returning results were generally positive and there was little evidence of harm from doing so.

Return of results - Genomic Return of Results in Research Studies (Policy)

- **Type:** Policy Framework
- **Expected Submission Date:** Q4 2021
- **Requesting Driver Projects:** Genomics England, All of Us, Australia Genomics

There is an emerging international consensus that some results should be returned to participants in research studies. This ethical consensus is supported by laws and regulations across a wide variety of jurisdictions. The resources needed to return results have been falling.

The first thing this policy aims to do is to set an aspiration for what should be aimed for. The starting proposition for this is the following: Any research study that provides genomic sequencing of human beings should budget for and implement a return of results policy to look for, and offer to return, at least a minimum list of actionable genomic findings, subject to participant consent. If resources within the research enterprise or the surrounding healthcare system are such that this is not achievable, there should be a clear statement as part of the consent that resources are not available but that returnable findings will be identified and returned if resources become available.

Decisions about exactly what to return to research participants will be project specific. Aspects of the process by which a project's return of results policy is developed and implemented are generalizable. The second aim of this policy is to capture these as guidelines and points to consider.

Finnish and Dutch Translation of the GA4GH Framework

- **Type:** Other: Translation
- **Expected Submission Date:** Q4 2020
- **Requesting Driver Projects:** Foundational

The Global Alliance for Genomics and Health (GA4GH) has developed a Framework for Responsible Sharing of Genomic and Health-Related Data which 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. This Framework is cited by many international expert organizations such as the World Economic Forum and the UK Digital Strategy. It has also proven to be useful to many genomics national initiatives.

We wish to provide two additional translations of the Framework to the current 13 translations (Arabic, Chinese, French, German, Greek, Hindi, Italian, Japanese, Japanese Interlinear, Korean, Portuguese, Russian, Spanish). We believe a Finnish and Dutch translation would be highly valuable and highly relevant for the international genomics community. We have contacted legally trained volunteers that can act as translators and validators and have found volunteers for both languages.

Data Access Committee Review Standards (DACReS)

- **Type:** Policy Framework, Guide
- **Expected Submission Date:** Q3 2021
- **Requesting Driver Projects:** Foundational

Executive Summary. The Regulatory and Ethics Work Stream (REWS), in association with Data Use and Research Identities Work Stream (DURI), propose to develop procedural standards for data access committees that facilitate consistency, effectiveness, and robustness of reviews for data access requests to genomic and health-related data.

Background. Both inappropriately restrictive and overly permissive policies governing access to genomic and associated clinical data challenge underlying principles of research ethics. Whereas the former prevents authorized persons from accessing data needed to advance genomic research, the latter places research participants at increased privacy risk given the inherent identifiability and sensitivity of genomic data. Data access committees (DAC) represent one institutional safeguard charged with applying rules meant to ensure an ethically permissible balance between data protection and utility in the research context. There are no standard review criteria, however, for how to operationalize these procedures across DACs. Moreover, access processes can differ where commercially valuable data or intellectual property is at stake, where data privacy laws impose jurisdiction-specific obligations, or where the data requested involve indigenous peoples. Lack of procedural standards and guidance can therefore invite inconsistencies in DAC reviews, compromising their quality and effectiveness at ensuring adequate research data protections.

Policy Objectives. The DACReS team proposes to develop procedural standards for DACs using a similar policy development approach as the ERR with the following specific aims:

- Aim 1. Review current policies, processes and practices for reviewing access requests among publicly funded genomic-phenotypic research datasets e.g., those held by NIH dbGap and EGA
- Aim 2. Identify sector-specific issues in reviewing data access requests, including where health and genomic data collected for clinical purposes are repurposed for research and where data are proprietary or have commercial value.
- Aim 3. Outline best practices/processes that mediate access of researchers to secure computing environments and datasets by applying a model-to-data approach.

Approved Deliverables

Framework for Responsible Sharing of Genomic and Health-Related Data

- **Type:** Policy Framework
- **Approval Date:** 2014

The GA4GH Framework for Responsible Sharing of Genomic and Health-Related Data provides a principled and practical framework for the responsible sharing of genomic and health-related data. It contains foundational principles and core elements for responsible data sharing and is guided by human rights, including the right to benefit from the progress of science, as well as privacy, non-discrimination, and procedural fairness.

Consent Policy

- **Type:** Policy Framework
- **V2 Approval Date:** 2019

The GA4GH Consent Policy aims to guide the sharing of genomic and health-related data in a way that respects autonomous decision-making while promoting the common good of international data sharing.

Data Privacy and Security Policy

- **Type:** Policy Framework
- **V2 Approval Date:** 2019

The GA4GH Data Privacy and Security Policy aims to guide the sharing of genomic and health-related 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.

Ethics Review Recognition Policy

- **Type:** Policy Framework
- **V2 Approval Date:** 2020

The GA4GH Ethics Review Recognition Policy aims to inspire confidence in the adequacy of an ethics review from another jurisdiction's ethics review system on the basis of equivalent requirements and the quality of the ethics review performed as part of that system. Recognizing the diversity of legal and ethical approaches and being responsive to emerging issues, this Policy encourages the reduction of duplicative ethics reviews through recognition approaches that enable ethics committees to accept the review of another ethics committee.

Machine Readable Consent Guidance

- **Type:** Guide
- **V1 Approval Date:** 2020

The GA4GH Machine Readable Consent Guidance provides instructions for researchers to integrate standard data sharing language into consent forms in a way that is able to be translated to a computable language. Machine readable consent language is able to be attached to datasets and stored in their descriptive data using DUO terms. Researchers can then search for datasets that have been consented to for their research purposes.

Copyright Policy

- **Type:** Policy Framework
- **V1 Approval Date:** 2020
- **Known Implementations & Deployments:** Foundational

The GA4GH Copyright policy ensures that the GA4GH has clear rights under copyright law to adapt and utilize individual and institutional contributions, incorporate them into GA4GH standards, and to distribute those standards widely. The policy encourages open and collaborative participation of institutions and individuals in standards development, and aims to recognize those who contribute.

Consent Toolkit: Consent Clauses for Genomic Research

- **Type:** Guide
- **V1 Approval Date:** 2020

The GA4GH Consent Clauses for Genomic Research provides researchers with sample phrases addressing various consent elements. Each sample clause can be adapted to fit different research and legal contexts, making this guidance useful to genetics and precision medicine studies around the globe.

COVID-19 Research: Navigating the European General Data Protection Regulation

- **Type:** Guide, Other: Manuscript/Article
- **V1 Approval Date:** 2020

This popular manuscript gives an in-depth look at the intersection of COVID-19-related research and the GDPR.

Becker, Regina, Adrian Thorogood, Johan Ordish, and Michael J.S. Beauvais. “COVID-19 Research: Navigating the European General Data Protection Regulation.” *Journal of Medical Internet Research* 22, no. 8 (2020): e19799. https://doi.org/10.2196/19799.

Appendix: Summary of Deliverables

Bold items are new to the 2020-2021 Roadmap; approved items will continue to be updated and maintained over time.

Clin/Pheno

- **Pedigree V1**
- **Phenopackets V2**
- Phenopackets V1 - *approved*

Cloud

- **Cloud Testbed**
- **Task Execution Service API**
- Data Repository Service API - *approved*
- Tool Registry Service API - *approved*
- Workflow Execution Service API - *approved*

Data Security

- **Risk Assessment Methodology for Software Stacks**
- **Cloud Security and Privacy**
- Authentication and Authorization Infrastructure (AAI) - *approved*

Data Use & Researcher Identities

- Data Use Ontology - *approved*
- Machine-readable Consents - *approved*
- GA4GH Passports - *approved*

Discovery

- **Beacon API V2**
- **Discovery Search API**
- Beacon API V1 - *approved*
- Service Registry/Service Info - *approved*

Genomic Knowledge Standards

- **Variant Annotation**
- Variant Representation - *approved*

Large Scale Genomics

- **GA4GH BED**
- Genetic Variation Formats (VCF) - *approved*
- Read Data Formats (SAM/BAM/CRAM) - *approved*
- htsget API - *approved*
- Reference Sequences (Refget) - *approved*
- Genetic Data Encryption (Crypt4GH) - *approved*
- RNA Data (RNAget) V1 - *approved*

Regulatory & Ethics

- **Consent Toolkit**
- **Standard Genomic Data Licenses and Agreements**
- **Participant Patient Public Engagement Policy**
- **GDPR Forum**
- **Responsible Data Sharing to Respond to the COVID-19 Pandemic: Ethical and Legal Considerations**
- **Participant Values Survey (Your DNA Your Say)**
- **Return of results – Survey of Stakeholder Perspectives**
- **Return of results – Genomic Return of Results in Research Studies (Policy)**
- **Finnish and Dutch Translation of the GA4GH Framework**
- **Data Access Committee Review Standards (DACReS)**
- Framework for Responsible Sharing of Genomic and Health-Related Data - *approved*
- Consent Policy - *approved*
- Data Privacy and Security Policy - *approved*
- Ethics Review Recognition Policy - *approved*
- Machine Readable Consent Guidance - *approved*
- Copyright Policy - *approved*
- Consent Toolkit: Consent Clauses for Genomic Research - *approved*
- COVID-19 Research: Navigating the European General Data Protection Regulation - *approved*