Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

GA4GH Connect: A 5-Year Strategic Plan

Table of Contents

A Note About the Roadmap

A technically detailed 5-year Strategic Roadmap with deliverables and timelines will be released soon after GA4GH 5th Plenary. More details to come.

GA4GH Vision

Genomic and Health-Related Data Sharing in 2022

In 2022, GA4GH standards enable broad access to genomic and health related data on tens of millions of individuals. Tools based on these standards create the platform upon which federated data sites—including research, healthcare, and commercial organizations as well as individuals—use, analyse, and store the data needed to drive precision medicine. The vast majority of these data come from health care rather than research and they span individuals of many national and ethnic origins. Harmonized data governance architectures allow for wide spheres of responsible data access, allowing researchers to perform analysis on virtual cohorts of populations or using virtual analytical tools without data movement.

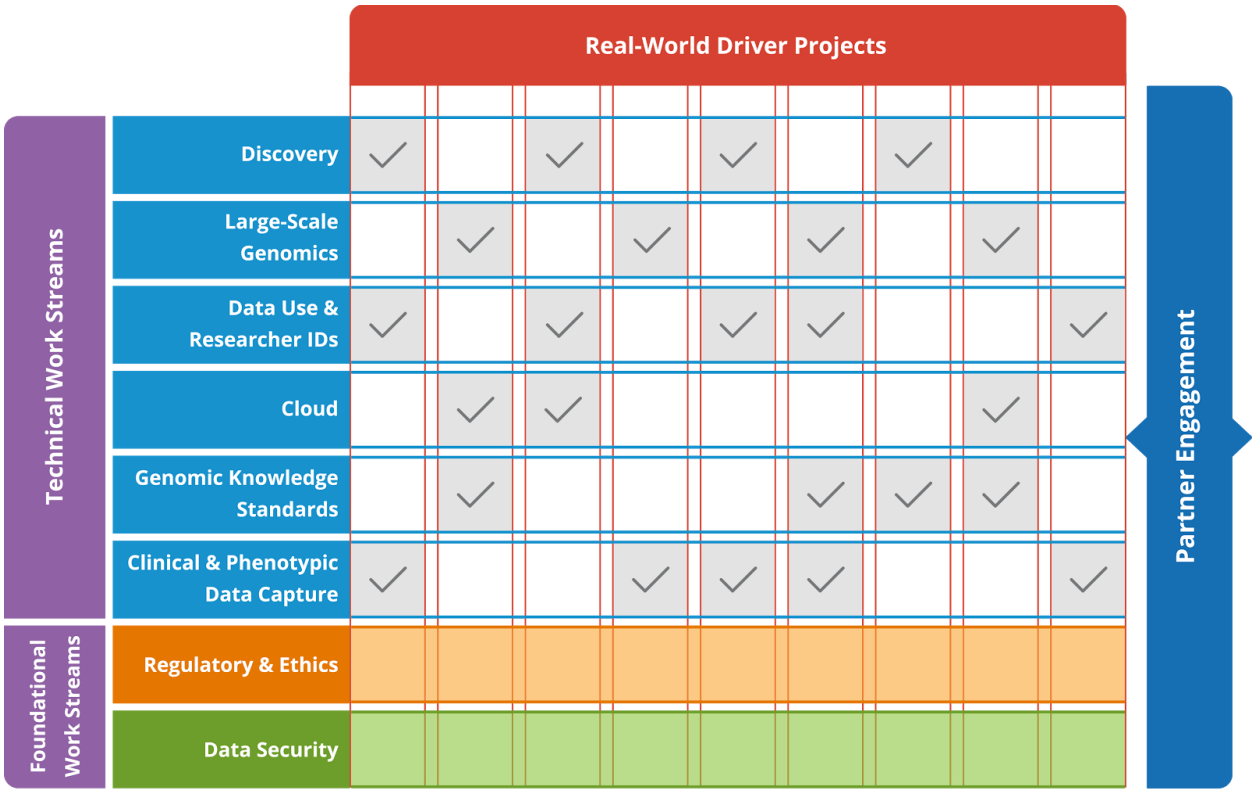
Genomics data that can be shared responsibly are shared responsibly. This means every qualified clinician, researcher, and corporate entity worldwide shares and accesses federated databases according to responsible consent and privacy governance policies. GA4GH APIs and tools are compatible with a variety of consent schema, including individual dynamic consent and broad consent models. Access permissions depend on the type and sensitivity of data requested, as well as researcher credentials. Electronic negotiation and monitoring of data access requests take place on a network of trust that is internet enabled, secure, and robust, usually allowing fully electronic access. An international Code of Conduct for Health-Related Data provides assurance that data sharing is proportionately protective, legally compliant, and addresses technical standards for identifiability, consent, privacy and security safeguards, research oversight, and sanctions for misuse.

Clinical records include genomic and phenotypic data to support a 'learning health system.' The complementarity of GA4GH standards and electronic health record (EHR) schemes, such as HL7 Frameworks, allow for secure storage and optimal use of patient/participant data. The community uses GA4GH standards to translate research findings into clinical application. Interoperability of health information facilitates real-world evidence generation and continuous learning among diverse healthcare stakeholders. Clinicians capture and transmit genotypic and phenotypic health data in standardized ways that are useful for research and clinical applications alike. In addition, health economic data demonstrate cost-effectiveness, diagnostic utility, and improvement in patient outcomes and public health.

GA4GH collaborates and coordinates with the many other global, regional, and national activities within the genomics and health ecosystem. GA4GH staffs new projects dynamically, bringing in experts from the appropriate disciplines. Research productivity as a result of concerted data sharing is demonstrated in the volume of published work citing GA4GH standards. In addition to contributing to new research findings, publication of common GA4GH standards and APIs enables improved understanding of the molecular basis of human health and fundamental biology.

New GA4GH Structure: Focused on Delivery

GA4GH Connect will refocus the organization on delivering technical standards and framing regulatory policies for genomic and health-related data sharing. Technical and Foundational Work Streams will drive policy and standards development to meet the specific needs of real-world clinical and research Driver Projects. This work will primarily focus on four key theme areas: cancer, rare disease, complex traits, and basic biology.



Strategic Approach

Policies and standards for genomic data sharing will only move us toward the future we envision if they are adopted and implemented by world-wide organizations and projects with genomic data to share. GA4GH will strategically partner with real-world genomic medicine and research initiatives ("Driver Projects") to create the policies and standards that meet their data sharing needs, as exemplars of the broader global genomics community.

The thirteen initial Driver Projects will play a crucial role in determining the policies, standards, and tools on which GA4GH focuses its development work over the next year. By embedding Driver Projects at different stages in their genomic data sharing efforts within Work Streams, there is a strong likelihood that completed policies and standards will meet the varying needs of the broader genomics community. Additional Driver Projects will be brought in for collaboration through periodic open calls to the community. With proactive partner engagement, at a national and organizational level, we will ensure that GA4GH standards are easily accessible and ready for use.

Work Stream Descriptions

Technical Work Streams

Technical Work Streams are the key production teams of GA4GH and consist of leaders in their respective sub disciplines from around the world. They work together to develop technical standards that are of most relevance to the practicing genomic medicine and research communities. These groups work to both discover and harden existing "working standards" and to create entirely new standards where needed.

Clinical & Phenotypic Data Capture: Supports the clinical adoption of genomics through establishing standard ontologies and information models to describe the clinical phenotype for use in genomic medicine and research, including the capture and exchange of information between electronic clinical systems and research.

Cloud: Helps GA4GH Driver Projects 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 data access.

Data Use & Researcher Identifiers (DURI): Facilitates and enables the harmonization of researcher identities by defining who is a bona fide researcher and one or more identity providers that respects this definition and can provide a portable electronic identity. DURI will also generate a data use ontology that will be used to both state use restrictions as well as researchers' purposes.

Discovery: Creates a unified data discovery platform to accelerate genomics and clinical data access and utilization around the world.

Genomics Knowledge Standards: Develops, adopts, and adapts standards-based components to enable the exchange of reference genomic information through common APIs, thereby enabling the downstream analysis of genomic data.

Large Scale Genomics: Creates standardized methods for accessing large-scale genomic data (reads, variants, and expression data) by file-based, API-based, cloud-based, and distributed access.

Foundational Work Streams

Foundational Work Streams provide guidance in the areas of legal regulation, ethics, and data security in genomics—both within GA4GH and more broadly. They develop policies and tools of immediate relevance to the international genomics community and support the ongoing efforts of real-world initiatives around the globe.

Data Security: Creates technology standards and best practices for protecting data and services consistent with the GA4GH policy framework, including the development, customization, and adoption of standards for identity management, data security, privacy protection, and service assurance.

Regulatory & Ethics: Focuses on the ethical, legal and social implications of international data sharing. Building on a novel human rights framework, the REWS aims to create and harmonize forward-looking consent and privacy policies, and anticipatory data governance models.

Clinical & Phenotypic Data Capture Work Stream Vision

Motivation and Mandate

The widespread adoption of Electronic Health Records and Electronic Medical Records provides a mechanism 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 the genomics includes the request for a genomics test, the sharing of the results from the test and the representation of genomics information in clinical information systems.

This Work Stream will support the clinical adoption of genomics through establishing standard ontologies and information models to describe the clinical phenotype for use in genomic medicine and research, including the capture and exchange of information between electronic clinical systems and research.

Existing Standards

The Clinical and Phenotype Data Capture Work Stream will seek to leverage key existing standards that support clinical data capture and exchange. These include:

- Human Phenotype Ontology - The Human Phenotype Ontology (HPO) aims to provide a standardized vocabulary of phenotypic abnormalities encountered in human disease.
- SNOMED CT - a comprehensive, multilingual clinical healthcare terminology to encode the meanings that are used in health information and support the effective clinical recording of data.
- The HL7 Fast Health Interoperable Resources (FHIR) standard - a specification to enable the transfer of healthcare information over standard APIs.

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.

Representing these data as FHIR resources with standard terminologies such as SNOMED CT and HPO will enable interoperability in the health system and support data analytics in research.

Proposed Solutions

The potential solution set for this Work Stream will include:

- Further development of key terminologies, such as HPO and SNOMED CT, to support the capture of clinical phenotype information.
- 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.

Cloud Work Stream Vision

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 Solutions

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:

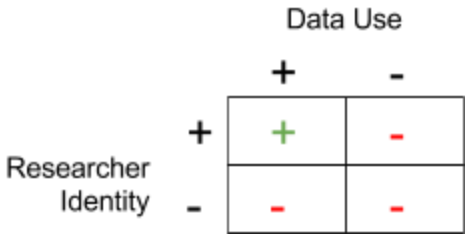
- Defining portable workflows: tool builders need to be able to package their tools for reuse. The CWS will build on existing standards to allow workflows built by one researcher to be used by many others.

- 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.

Data Use & Researcher Identifiers (DURI) Work Stream Vision



Motivation and Mandate

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

1. 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.
2. 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 and automate 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 "consent codes" and ADA-M systems of data use restrictions.
- The Library Cards and Bona Fide Researchers efforts to define researchers and their identities.
- The eRA Commons, ORCID, and EGA systems of identities.

Proposed Solutions

The DURI workstream will drive progress in the following two areas:

1. 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.
2. 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.

Discovery Work Stream Vision

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 Solutions

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 :

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

Genomics Knowledge Standards Work Stream Vision

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 Solutions

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).

Large Scale Genomics Work Stream Vision

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 everyday healthcare. 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. Perhaps the most significant challenge is to move from the traditional purely file-based approach to accessing and analysing genetic data, to one where we are presented with standardised API interfaces.

Existing Standards

Incumbent file formats for read sequencing data include SAM/BAM/CRAM and VCF/BCF for genetic variation. Large cohorts of genetic variation data are now available from multiple different resources (e.g., ExAC/gnomAD, dbSNP, 1000 Genomes, EVA, EGA).

Proposed Solutions

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 and key sequencing and bioinformatics tool developers 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.

Data Security Work Stream Vision

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

Regulatory & Ethics Work Stream Vision

Motivation and Mandate

Data sharing bridges the genomic-clinical divide, thereby enabling translational medicine. Current frameworks for privacy protections, however, may frustrate the desire of individuals to share such data.

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 REWS Work

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 Solutions

In the next five years, the REWS aims to address three pressing governance challenges for international data sharing. First, governance must serve and be informed by the individuals providing the data. The REWS is addressing this through a multilingual, international Public Attitudes Survey. Second, processes for handling individual genomic findings of clinical relevance are highly divergent across countries and sectors. A Task Team on return of results will be established to combat this variation. A third challenge is the constantly shifting winds of global data protection regulation. The REWS aims to address this through an international data sharing Code of Conduct. This Code will guide the current push to establish clear and legally binding data sharing rules.

Partner Engagement

GA4GH's Partner Engagement initiative will ensure proactive two-way dialogue at both the national and organizational level with potential end-users of GA4GH frameworks and standards. It's ultimate goal is to enable easy and immediate adoption of these tools in real-world settings.

This initiative, which will include leading representatives from national initiatives and the clinical genomics community, will have the following objectives:

1. **Foster two-way dialogue with end-users and key communities**
    a. Outreach to potential users (e.g. national initiatives, major health care providers/hospitals, EHR vendors, journal editors) to share information on GA4GH activities and tools, and promote and assist with implementation
    b. Intake ideas/suggestions/feedback from users, regarding the relevance of the GA4GH Toolkit and ease of use/implementation
    c. Host bi-annual meetings of national initiatives
    d. Convene symposia, promote tools and approaches
2. **Provide expert guidance and insight to the GA4GH Steering Committee**
    a. The Partner Engagement team will consist of key members who have clinical expertise in the four theme areas (cancer, rare disease, complex traits, and basic biology), and/or who represent diverse geographic areas
    b. This team will give input on whether we are connected with the best representative Driver Projects across each theme area, suggest additional Driver Projects to consider, and share insights on the evolving needs in different national contexts
3. **Align with related global genomics organizations** (e.g. the Global Genomic Medicine Collaborative (G2MC), HL7, and others)
    a. Ensure compatibility and harmony of approaches and messages across organizations

Immediate areas of focus will include the hosting of bi-annual meetings of established and emerging national genomics and precision medicine initiatives, following from an initial meeting jointly hosted by Australian Genomics and Genomics England in London (May 2017). By engaging with national-level projects, Partner Engagement will be well-poised to identify potential Driver Projects for collaboration with GA4GH, and to promote the uptake of GA4GH tools across global projects.

Partner Engagement will ultimately contribute to establishing GA4GH as a thought-leader and go-to organization for national initiatives and other global genomics projects. It will drive the global adoption of GA4GH tools, promote practical solutions for the integration and implementation of genomic medicine into clinical practice, and contribute to the development of an evidence base for best practice.

2017 Driver Project Descriptions

All of Us Research Program

Location: United States
Themes: Cancer, Rare Disease, Complex Traits
The All of Us Research Program of the US National Institutes of Health is a historic effort to gather data from one million or more people living in the United States to accelerate research and improve health. By taking into account individual differences in lifestyle, environment, and biology, researchers will uncover paths toward delivering precision medicine. allofus.nih.gov

Australian Genomics

Location: Australia
Themes: Cancer, Rare Disease, Complex Traits
Australian Genomics is a research collaboration of more than 70 organisations committed to integrating genomic medicine into healthcare in Australia. Our goals are to improve diagnostics, enable early intervention and support equitable access to genomic medicine. Our research is developing the knowledge to translate genomic technology sustainably into clinical practice so patients and their families benefit. australiangenomics.org.au

BRCA Challenge

Location: International
Themes: Cancer, Rare Disease
The BRCA Challenge aims to advance our understanding of the genetic basis of breast cancer, ovarian cancer and other diseases by pooling data on BRCA1/2 genetic variants and corresponding clinical data from around the world. The project has developed a publicly available portal, the BRCA Exchange, to make aggregate data accessible to all users and to facilitate expert variant pathogenicity classifications made by the ENIGMA Consortium. brcaexchange.org

Canadian Distributed Infrastructure for Genomics (CanDIG)

Location: Canada
Themes: Rare Disease, Cancer, Complex Traits, Basic Biology
The Canadian Distributed Infrastructure for Genomics, CanDIG, is a fully distributed platform that allows national-scale, privacy-maintaining analyses of locally-controlled data sets. candig.github.io

Clinical Genome Resource (ClinGen)

Location: United States
Themes: Rare Disease, Cancer
ClinGen is a National Institutes of Health (NIH)-funded resource dedicated to building an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research. clinicalgenome.org

ELIXIR Beacon

Location: Europe
Themes: Rare Disease, Cancer, Complex Traits, Basic Biology

ELIXIR Beacon enables discovery of research consented sensitive human genetic data stored in databases affiliated with ELIXIR Nodes and in the European Genome-phenome Archive (EGA). Building on the GA4GH Beacon technology, the project is developing common interface to streamline and simplify access to these resources using ELIXIR Authentication and Authorization Infrastructure. ELIXIR Beacon involves partners from eight ELIXIR Nodes: EMBL-EBI, Belgium, Finland, France, Netherlands, Spain, Sweden and Switzerland.
https://www.elixir-europe.org/beacons

European Nucleotide Archive (ENA) / European Variation Archive (EVA) / European Genome-phenome Archive (EGA)

Location: Europe
Themes: Rare Disease, Cancer, Complex Traits, Basic Biology
The core mission of the European Genome-phenome Archive (EGA), European Variation Archive (EVA), and European Nucleotide Archive (ENA) is to provide the basic infrastructure to enable global public sharing of genetic data. The EGA is a service and database for permanent archiving and sharing of genetic and phenotypic human data resulting from biomedical research projects. It provides the necessary security required to control access in accordance with participant consent, providing access only to authorised researchers and clinicians. The ENA captures and presents information relating to experimental workflows that are based around nucleotide sequencing. The ENA is a partner in the International Nucleotide Sequence Database Collaboration (INSDC) to provide worldwide exchange and replication of all public nucleotide sequences. The EVA is an open-access database of all types of genetic variation data from all species. ebi.ac.uk/ena, ebi.ac.uk/eva, ega-archive.org

Genomics England

Location: United Kingdom
Themes: Cancer, Rare Disease, Complex Traits
Genomics England is a company owned by the UK Department of Health and was set up to deliver the 100,000 Genomes Project. This flagship project is sequencing 100,000 whole genomes from NHS patients and their families. We aim to bring benefit to patients, create an ethical and transparent programme based on consent, enable new scientific discovery and medical insights and kickstart the development of a UK genomics industry. The project is focusing on patients with rare diseases, and their families, as well as patients with common cancers.
genomicsengland.co.uk

International Cancer Genome Consortium for Accelerating Research in Genomic Oncology (ICGC-ARGO)

Locations: International
Themes: Cancer
The ICGC-ARGO project is an international initiative to sequence the germline and tumor genomes of thousands of participants. ICGC-ARGO will link genomic data to clinical and health information across the cancer spectrum. Researchers, scientists, policymakers, and clinicians will be able to work with patients, healthcare providers, and others through a shared knowledge-base to improve disease prevention, detection, diagnosis, prognosis, and intervention. A series of cloud-based Regional Data Processing Centres will accept raw genomic reads and clinical submissions from participating institutions, running a uniform suite of analytic tools. icgcmed.org

Matchmaker Exchange (MME)

Location: International
Themes: Rare Disease
The 'Matchmaker Exchange' project was launched in October 2013 to find genetic causes for patients with rare disease. This involves a growing federated platform (Exchange) to facilitating the matching of cases with similar

phenotypic and genotypic profiles (matchmaking) through standardized application programming. matchmakerexchange.org

National Cancer Institute Genomic Data Commons (NCI GDC)

Locations: United States
Themes: Cancer
The National Cancer Institute (NCI) Genomic Data Commons (GDC) is a data sharing platform that promotes precision medicine in oncology. The GDC contains some of the largest and most comprehensive cancer genomic datasets, including The Cancer Genome Atlas (TCGA) and Therapeutically Applicable Research to Generate Effective Therapies (TARGET). These datasets have been harmonized using a common set of bioinformatics pipelines so the data can be directly compared. As a growing knowledge system for cancer research, the GDC enables researchers to submit data, harmonize the data for import into GDC, as well as query and download high-quality data. gdc.cancer.gov

Monarch Initiative

Location: International
Themes: Rare Disease, Cancer, Complex Traits, Basic Biology
Monarch isn't just another data aggregator, it is driven to truly integrate biological information using semantics and present it in a novel way. Their niche is the use of computational reasoning to compare phenotypes both within and across species, with the ultimate goal of improving biomedical research. https://monarchinitiative.org.

Variant Interpretation for Cancer Consortium (VICC)

Location: International
Themes: Cancer
The Variant Interpretation for Cancer Consortium (VICC) brings together the leading institutions that are independently developing solutions for cancer variant interpretation. The VICC will develop and refine standards for describing and tiering variant evidence, promote the adoption of relevant existing ontologies and APIs, and create a federated query service able to interrogate associations between cancer gene alterations and clinical actions based on evidence amassed from all participating institutions worldwide. https://cancervariants.org.