# GA4GH Strategic Roadmap

*The GA4GH Strategic Roadmap presents standards and frameworks planned for development under GA4GH Connect — a 5 year Strategic Plan aimed at aligning with the key needs of the genomic data community.*

## Approved GA4GH Standards

### Beacon

bit.ly/GA4GHBeacon

Beacon is a platform for global discovery of genomic variant sharing and discovery. A "Beacon" is defined as a web-accessible service that can be queried for information about a specific allele. A user of a Beacon can pose queries of the form "Have you observed this nucleotide (e.g. C) at this genomic location (e.g. position 32,936,732 on chromosome 13)?" to which the Beacon responds with either "yes" or "no", plus additional metadata. In this way, a Beacon allows allelic information of interest to be discovered by a remote searcher with no reference to a specific sample or patient of origin, thereby mitigating risks to patient/participant privacy.

Contributors

**Work Streams**: Discovery (primary)

**Driver Projects**: ELIXIR Beacon, Genomics England, EVA/EGA/ENA, Australian Genomics

### Data Use Ontology

bit.ly/GA4GHDUO

DUO allows data holders to semantically tag datasets with restrictions about their usage, making them automatically discoverable based on the intended usage. It enables machine readable descriptions of data access requests and data use restrictions to be matched, alleviating the need for manual review when datasets are requested by researchers.

Contributors

**Work Streams**: Data Use & Researcher Identities (primary), Data Security, Regulatory & Ethics

**Driver Projects**; EVA/EGA/ENA, Australian Genomics, All of Us

### htsget API

bit.ly/GA4GHhtsget

A key challenge for human genetics is the ability to share large volumes of genomic data between different locations to enable discovery of new genetic associations or provide supporting evidence to new findings. Today, this is largely achieved by copying and transferring large files between two services. However, this approach by definition requires a file and therefore restricts the development of novel strategies for storing and indexing genomic data. We are proposing to develop a secure standard interface for slicing and streaming sequencing data that decouples the assumption of a file at the remote location. It will build upon the incumbent sequencing file formats and use these as the on-the-wire format.

Contributors

**Work Streams**: Large Scale Genomics (primary)

**Driver Projects**: Australian Genomics, Canadian Distributed Infrastructure for Genomics (CanDIG), Genomics England, EVA/EGA/ENA, Human Cell Atlas

**Other Partners**: Wellcome Trust Sanger Institute, DNA Nexus, Verily, ELIXIR Finland, Google Cloud Platform

### refget API

bit.ly/GA4GHrefget

At its core, genetics is about examining differences in the DNA sequence across individuals or species. This API provides a framework to retrieve 'reference sequences' by a unique checksum, allowing users to retrieve such reference sequences without ambiguity from different databases and servers.

Contributors

**Work Streams**: Large Scale Genomics (primary), Genomic Knowledge Standards
**Driver Projects**: EVA/EGA/ENA, Australian Genomics

### Workflow Execution Service (WES) API

bit.ly/GA4GHWES

The ability to execute the same scientific tools and workflows in a variety of environments without modification is a key concern for researchers. WES provides a standard that allows researchers to do just this. In particular, this standard will enable disparate platforms to accept and run workflows in Common Workflow Language and Workflow Definition Language (CWL/WDL)—and possibly other formats—using a common API.

Contributors

**Work Streams**: Cloud (primary), Discovery, Data Security
**Driver Projects**: Australian Genomics, ENA/EGA/EVA, Genomics England, Human Cell Atlas, TOPMed

### Read File Formats (SAM/BAM/CRAM)

https://goo.gl/8ZpeZR

SAM, BAM and CRAM are standard formats for genomic data that require continued maintenance and development as our capability to interrogate genomic information changes with new technologies. This team will maintain and evolve the primary these file formats.

Contributors

**Work Streams**: Large Scale Genomics (primary)
**Driver Projects**: EVA/EGA/ENA, Human Cell Atlas Other Partners: Wellcome Trust Sanger Institute, Broad Institute of MIT and Harvard, 10x Genomics, Pacific Biosciences, Oxford Nanopore

## Planned Roadmap Deliverables

### Authentication and Authorization Infrastructure (AAI)

The GA4GH Authentication and Authorization Infrastructure (AAI) Profile is the GA4GH standard technical profile for authenticating the identity of individuals seeking to access data and services offered by the Driver Projects, and for authorizing access in accordance with applicable Driver Project policies. 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 & Researcher Identity (DURI) work stream.

Contributors

**Work Streams**: Data Security (primary), Clinical & Phenotypic Data Capture, Discovery, Data Use & Researcher Identities, Genomic Knowledge Standard

**Driver Projects**: ELIXIR Beacon, Others

## Breach Response Protocol

The Breach Response Protocol is a jointly developed strategy, and supporting processes, through which the GA4GH Driver Project community can collaboratively protect itself and effectively respond to and recover from security breaches. This deliverable will be a flexible Best Practices document which will allow genomic data sharing organizations to (i) monitor for and detect breaches, (ii) ascertain whether a breach involves one or more GA4GH standards, (iii) collaboratively share information regarding breaches that involve GA4GH standards, and (iv) support response to and recovery from breaches.

Contributors

**Work Streams**: Data Security (primary), Regulatory & Ethics

**Driver Projects**: Australian Genomics, All of Us Project, CanDIG, ClinGen, BRCA Challenge, ELIXIR Beacon, EVA/EGA/ENA, NCI Genomic Data Commons, Genomics England, Human Cell Atlas, TOPMed, ICGC-ARGO, Matchmaker Exchange, Monarch Initiative, VICC

## Clinical Data Exchange (FHIR/Phenopackets)

While ontologies and terminologies provide the standard data concept definitions for capturing clinical information, an information model is required to successfully exchange that information between clinical information systems and with related information systems. This standard will provide information models with different levels of complexity to enable high level clinical phenotype information as well as deep clinical phenotype information to be exchanged.

Contributors

**Work Streams**: Clinical & Phenotypic Data Capture (primary), Discovery, Genomic Knowledge Standards

**Driver Projects**: Australian Genomics, Monarch Initiative, ELIXIR Beacon, Matchmaker Exchange, Variant Interpretation for Cancer Consortium (VICC), BRCA Challenge, ENA / EVA / EGA, Clinical Genome Resource (ClinGen)

## Data Registry Service (DRS) API

The ability access (read+write) data across multiple clouds is a key concern for researchers, especially as large, multi-institution projects leverage cloud resources in multiple environments. This API standard will create a common way to refer to data and access it regardless of cloud or platform, making it easier to do work across projects and environments.

Contributors

**Work Streams**: Cloud (primary), Discovery, Data Security, Data Use & Researcher Identities, Large Scale Genomics

**Driver Projects**: Australian Genomics, EVA/EGA/ENA, Genomics England, Human Cell Atlas, TOPMed

## Genetic Variation File Formats

VCF is a standard format to represent genomic variation. It requires maintenance and updates to represent new genomic information in an unambiguous manner. In addition to maintaining and evolving VCF, this team will investigate and research new more scalable formats for storing and exchanging genetic variation.

Contributors

**Work Streams**: Large Scale Genomics (primary), Genomic Knowledge Standards

**Driver Projects**: NCI Genomic Data Commons, ENA/EVA/EGA, VICC, ClinGen Other Partners: Wellcome Trust Sanger Institute, Broad Institute of MIT and Harvard

## International Participant Values Survey

The multilingual International Participant Values Survey, or "Your DNA, Your Say," explores how people around the world feel about the collection, use, and sharing of genetic and health data for research such as attitudes about genetic exceptionalism, reasons for sharing or not, and what perceived benefits or harms are involved.

Contributors

**Work Streams**: Regulatory & Ethics (primary)

**Driver Projects**: TBD

## Phenotype Representation for Genomic Medicine and Research

Defining what constitutes a “phenotype” for both the clinical and scientific communities as well as an evaluation of clinical ontologies used within these definitions. The complexity of implementing genomics into healthcare will require accurate data to be captured about the patient’s genomic information to support clinical decision making and clinical and medical research.

Contributors

**Work Streams**: Clinical & Phenotypic Data Capture (primary), Discovery, Genomic Knowledge Standards, Data Use & Researcher Identities (DURI)

**Driver Projects**: Australian Genomics, Monarch Initiative, Variant Interpretation for Cancer Consortium (VICC), Clinical Genome Resource (ClinGen), Genomics England, ELIXIR Beacon, Matchmaker Exchange, ENA / EVA / EGA

## Pedigree Representation

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 need for high quality, unambiguous, computable pedigree and family history information is critical for scaling genomic analysis to larger, complex families.

Contributors

**Work Streams**: Clinical & Phenotypic Data Capture (primary), Discovery, Genomic Knowledge Standards

**Driver Projects**: Australian Genomics, Monarch Initiative, All of Us Research Program, ELIXIR Beacon, Clinical Genome Resource (ClinGen), Matchmaker Exchange, Variant Interpretation for Cancer Consortium (VICC), BRCA Challenge

## Researcher Identity and Bona Fide Status

In a future where human genomics and health data is stored in a federated network of public clouds there will be a need to tightly control and monitor which users access this data. At the same time it is important to enable smooth process and remove friction and artificial barriers between researchers and insights they can glean from the data. This system will allow researchers and other users to establish identity and credentials claims with regards to their professional identity to acquire access across datasets.

Contributors

**Work Streams**: Discovery (primary), Cloud, Large Scale Genomics
**Driver Projects**: ELIXIR Beacon, EVA/EGA/ENA

## Return of Results Policy

This document will aim to inform research policy makers and projects about what to consider when deciding whether to tell participants about genomic findings relevant to their health. It will include international ethical, legal, and policy guidance around return of clinically relevant individual findings (e.g., individual research results, incidental findings) and generated by whole genome/exome sequencing to research participants and will consider developments in data sharing practices.

Contributors

**Work Streams**: Regulatory & Ethics (primary), Data Use & Researcher Identities
**Driver Projects**: All of Us Project, Genomics England, Australian Genomics

## RNASeq Expression Matrix

Expression results when we have billions of cells. There will be huge matrices to be represented and users should be able to access these without a need for huge amounts of memory.

Contributors

**Work Streams**: Large Scale Genomics (primary)
**Driver Projects**: Human Cell Atlas, NCI Genomic Data Commons, ICGC-ARGO

## Search

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).

Contributors

**Work Streams**: Discovery (primary), Cloud, Large Scale Genomics
**Driver Projects**: ELIXIR Beacon, EVA/EGA/ENA

## Service Registry Prototype

The Service Registry Prototype provides a digital network infrastructure for a proposed "Internet of Genomics". The registry will list GA4GH services (e.g. Beacons, DOS, etc.) or other registries (e.g., Matchmaker Exchange) that have been registered to it. The Service Registry will allow for dynamic registration and on-demand discovery of online GA4GH APIs (data, tools, services) to enable their realtime discovery and use.

Contributors

**Work Streams**: Discovery (primary), Cloud, Large Scale Genomics
**Driver Projects**: ELIXIR Beacon, EVA/EGA/ENA

## Task Execution Service (TES)

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.

Contributors

**Work Streams**: Cloud (primary), Discovery, Data Security
**Driver Projects**: TBD

## Testbed & Interoperability Demonstration

This project aims to demonstrate that workflows can be exchanged between Driver Project sites and used reproducibly, using preliminary versions of the GA4GH Cloud APIs (TES, TRS, WES, and DOS).

Contributors

**Work Streams**: Cloud (primary), Data Security
**Driver Projects**: Australian Genomics, ENA/EGA/EVA,Genomics England, Human Cell Atlas, TOPMed

## Tool Registry Service (TRS)

The portable exchange of tools and workflows is key to scientific reproducibility. The TRS standard, and implementation in Dockstore.org, is designed to robustly address this need.

Contributors

**Work Streams**: Cloud (primary), Discovery, Data Security
**Driver Projects**: Australian Genomics, ENA/EGA/EVA,Genomics England, Human Cell Atlas, TOPMed

## Variant Annotation: Data Model

This common data model will guide the linkage of annotations and structured clinical interpretations to variant data. It will include support for current clinical lab standards (e.g., ACMG/AMP), clinical phenotypes (disease/disorder), clinical relevance and context, and associated metadata.

Contributors

**Work Streams**: Genomic Knowledge Standards (primary), Clinical & Phenotypic Data Capture, Discovery

**Driver Projects**: ClinGen, VICC, Genomics England, BRCA Challenge, Monarch Initiative

### Variation Representation: Data Model/Specification

This specification will a standardised extensible data model and message schema specification for the representation of variants. It will build heavily on the work of the Variant Modeling Consortium and will expand that schema to include support for structural and complex variants.

Contributors

**Work Streams**: Genomic Knowledge Standards (primary)

**Driver Projects**: ClinGen, ELIXIR Beacon, Genomics England, Monarch Initiative