Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

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. The Roadmap will be updated annually with new deliverables and timelines.

2018 - Q4

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).

WORK STREAMS

Discovery, Cloud, Large Scale Genomics

DRIVER PROJECTS

ELIXIR Beacon, EVA/EGA/ENA

Beacon

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.

WORK STREAMS

Discovery

DRIVER PROJECTS

ELIXIR Beacon, Genomics England, EVA/EGA/ENA, Australian Genomics

Data Object Service (DOS) 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.

WORK STREAMS

Cloud, Discovery, Data Security, Data Use & Researcher Identities, Large Scale Genomics

DRIVER PROJECTS

Australian Genomics, EVA/EGA/ENA, Genomics England, Human Cell Atlas, TOPMed

Data Use Ontology

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.

WORK STREAMS

Data Use & Researcher Identities, Data Security, Regulatory & Ethics

DRIVER PROJECTS

EVA/EGA/ENA, Australian Genomics, All of Us

htsget Streaming API

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.

WORK STREAMS

Large Scale Genomics

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

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.

WORK STREAMS

Regulatory & Ethics

DRIVER PROJECTS

TBD

Reference Sequence Retrieval API

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.

WORK STREAMS

Large Scale Genomics, Genomic Knowledge Standards

DRIVER PROJECTS

EVA/EGA/ENA, Australian Genomics

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.

WORK STREAMS

Data Use & Researcher Identities, Data Security, Regulatory & Ethics

DRIVER PROJECTS

ELIXIR Beacon, EVA/EGA/ENA

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.

WORK STREAMS

Large Scale Genomics

DRIVER PROJECTS

Human Cell Atlas, NCI Genomic Data Commons, ICGC-ARGO

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.

WORK STREAMS

Discovery, 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.

WORK STREAMS
Cloud, 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).

WORK STREAMS
Cloud, 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.

WORK STREAMS
Cloud, Discovery, Data Security

DRIVER PROJECTS
Australian Genomics, ENA/EGA/EVA,Genomics England, Human Cell Atlas, TOPMed

Variant Submission

The deliverable for this project will be a document that describes a series of variants and variant attributes to allow for computational analysis and swapping of variant information between organizations.

WORK STREAMS
Discovery, Genomic Knowledge Standards

DRIVER PROJECTS
ClinGen, VICC, ENA/EGA/EVA, Monarch Initiative

Workflow Execution Service (WES) API

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.

WORK STREAMS
Cloud, Discovery, Data Security

DRIVER PROJECTS
Australian Genomics, ENA/EGA/EVA, Genomics England, Human Cell Atlas, TOPMed

2019 – Q3

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.

WORK STREAMS

Regulatory & Ethics, Data Use & Researcher Identities

DRIVER PROJECTS

All of Us Project, Genomics England, Australian Genomics

2019 - Q4

Cloud Access Policy

Access processes, requirements, and conditions vary substantially across jurisdictions, projects, and data types, presenting barriers to researcher access and the establishment of data sharing networks. Even data that is publicly accessible is subject to diverse data-use agreements that restrict reuse and redistribution. Moreover, cloud computing, federated networks, and APIs are also fundamentally changing the nature of access to data.This ethical-legal framework will aim to harmonize policies and requirements for access to cloud-based genomic and health-related data across research and clinical contexts.

WORK STREAMS

Regulatory & Ethics, Data Use & Researcher Identities

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

Information Models for Clinical/Genomic Data Exchange

While ontologies and terminologies provide the standard definitions for capturing clinical information, a standardised "information model" is required to successfully exchange that information between disparate computers. This standard will enable the exchange of both deep and high level clinical phenotype information.

WORK STREAMS

Clinical & Phenotypic Data Capture, Data Use & Researcher Identities, Discovery, Genomic Knowledge Standards

DRIVER PROJECTS

Australian Genomics, Monarch, VICC, ClinGen, Genomics England, ELIXIR Beacon, Matchmaker Exchange, EVA/EGA/ENA

Phenotype and Disease Ontology Recommendations

This policy will recommend standard ontologies and terminologies for capturing the clinical phenotype for use in genomic medicine and research, as well as harmonisation policies to enable advanced, machine-readable use of those terminologies. The recommendations will cover both high level clinical phenotyping (e.g., at the disease/disorder level) and deep phenotyping (e.g., family history, clinical relevance) for both primary (clinical) and secondary (research) use.

WORK STREAMS

Clinical & Phenotypic Data Capture, Discovery, Genomic Knowledge Standards, Partner Engagement Initiative

DRIVER PROJECTS

Australian Genomics, Monarch, AllofUs, ELIXIR Beacon, ClinGen, Matchmaker Exchange, VICC, BRCA Challenge

Phenotype Standards Implementation

Consistent descriptions of clinical information across the many relevant sources—such as scientific and medical journals, labs, and database providers—will make it easier to make use of all available information when treating patients. In conjunction with the Partner Engagement group, the CPDC team will work with 3rd party information providers to identify a set of recommendations for associating standardised metadata with content in a way that best supports genomic medicine.

WORK STREAMS

Clinical & Phenotypic Data Capture, Discovery, Genomic Knowledge Standards, Partner Engagement Initiative

DRIVER PROJECTS

Australian Genomics, Monarch, AllofUs, ELIXIR Beacon, ClinGen, Matchmaker Exchange, VICC, BRCA Challenge

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.

WORK STREAMS

Genomic Knowledge Standards, Clinical & Phenotypic Data Capture, Discovery

DRIVER PROJECTS

ClinGen, VICC, Genomics England, BRCA Challenge, Monarch Initiative

Variant 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.

WORK STREAMS

Genomic Knowledge Standards

DRIVER PROJECTS

ClinGen, ELIXIR Beacon, Genomics England, Monarch Initiative

2020 - Q4

International Code of Conduct for Data Sharing

The health research community must inform the interpretation of privacy laws around the world, such as the new European Union General Data Protection Regulation (2018), to ensure they allow for responsible data sharing. The REWS will advise on efforts to develop a EU Code of Conduct for health-related data, and will provide guidance on international harmonization when the Code is finalized.

WORK STREAMS

Regulatory & Ethics

DRIVER PROJECTS

ENA/EVA/EGA, ICGC-Argo

Internet of Genomics

We envision the Internet of Genomics to be a set of complementary services that together enable secure, federated, global search, discovery, exchange, and analysis of genomics and clinical data. The goal of this long-term, "moonshot" project is to deliver critical connective infrastructure to enable the exchange of genomics and clinical information via the resulting application platform.

WORK STREAMS

Discovery, Clinical & Phenotypic Data Capture, Data Use & Researcher Identities, Genomic Knowledge Standards

DRIVER PROJECTS

Genomics England, Australian Genomics, ICGC-ARGO, MME, ELIXIR Beacons, Monarch Initiative, EGA

Date TBD

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.

WORK STREAMS

Data Security, Clinical & Phenotypic Data Capture, Discovery, Data Use & Researcher Identities, Genomic Knowledge Standard

DRIVER PROJECTS

ELIXIR Beacon, Others

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.

WORK STREAMS

Large Scale Genomics, Genomic Knowledge Standards

DRIVER PROJECTS

NCI Genomic Data Commons, ENA/EVA/EGA, VICC, ClinGen

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.

WORK STREAMS

Data Security, 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

Ongoing

Read File Formats (SAM/BAM/CRAM)

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.

WORK STREAMS

Large Scale Genomics

DRIVER PROJECTS

EVA/EGA/ENA, Human Cell Atlas