Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH: getting started with standards for genomics data sharing in health

Susan Fairley
Chief Standards Officer
GA4GH
Outline

1. Why standards?
2. What is GA4GH?
3. What standards exist?
4. Where to start?
5. Getting involved
How do we advance genomics?
The 1000 Genomes Project and Standards
The 1000 Genomes Project and Standards
The 1000 Genomes Project and Standards
The 1000 Genomes Project ecosystem

Common consent across 26 populations from five continental superpopulations

Common resources: cell lines, arrays, exomes and genomic sequences

File formats (SAM and VCF) and data sharing methods

Software: samtools, FreeBayes, GATK, VCFtools, tabix, etc.

Interoperable
Together, able to address research questions and lay foundation for future projects
Where are we now?
Standards are essential

Interoperability is key to data sharing and enabling the necessary ecosystem

- Findable, Accessible, Interoperable, and Reusable Datasets
- Interoperable compute
- Interoperable tools
- Interoperable data access methods
GA4GH is a global forum
The **Global Alliance for Genomics and Health** aims to accelerate progress in genomic science and human health by developing standards and framing policy for responsible genomic and health-related data sharing.
GA4GH: Anna and her robots

Driving improvements based on real-world uses

bit.ly/GA4GH-Anna
GA4GH aims to...

- Enable **international** data sharing
- Promote sharing across the **translational continuum**
- Encourage technology-enabled **federated approaches**
- Promote **interoperability**

GA4GH achieves this by...

- **Convening** stakeholders
- **Catalyzing** sharing of data
- **Creating** harmonized approaches
- **Acting** as a clearinghouse
- **Fostering** innovation
- **Commiting** to responsible data sharing
Implementations and Starter Kit

A minimal set of standards that can support the analysis needs of numerous researchers

Example Implementations

1. **Tool Registry Service**  
e.g. Dockstore and Biocontainers

2. **Workflow Execution Service**  
e.g. TOPMed and HCA pipelines, supported in Cromwell and Toil

3. **Data Repository Service**  
e.g. Anvil Gen3 repository on Terra

4. **Task Execution Service**  
e.g. CINECA, ELIXIR Cloud and supported in Nextflow, Snakemake and Cromwell

Starter Kit

1. Lightweight and minimal implementations

2. Function as easily deployable implementations or as a reference for building an implementation

3. Accompanying test suites

4. DRS and WES already available

5. Extending across the standards
1. Customisable Cloud compute at EBI
2. Starter Kit implementation of Workflow Description Language (WDL)
3. Starter Kit implementation of the Data Repository Service (DRS) to get data
4. A “new” environment

GA4GH Work Streams

- Clinical and Phenotypic Data Capture
- Cloud
- Data Use and Researcher Identities
- Discovery
- Genomic Knowledge Standards
- Large Scale Genomics
- Regulatory and Ethics
- Data Security
Acknowledgements

GA4GH Host Institutions

Core Funders

GA4GH staff and contributors