

Towards a federated network of nodes hosting human genetic and phenotype data

Thomas Keane

European Genome-phenome Archive

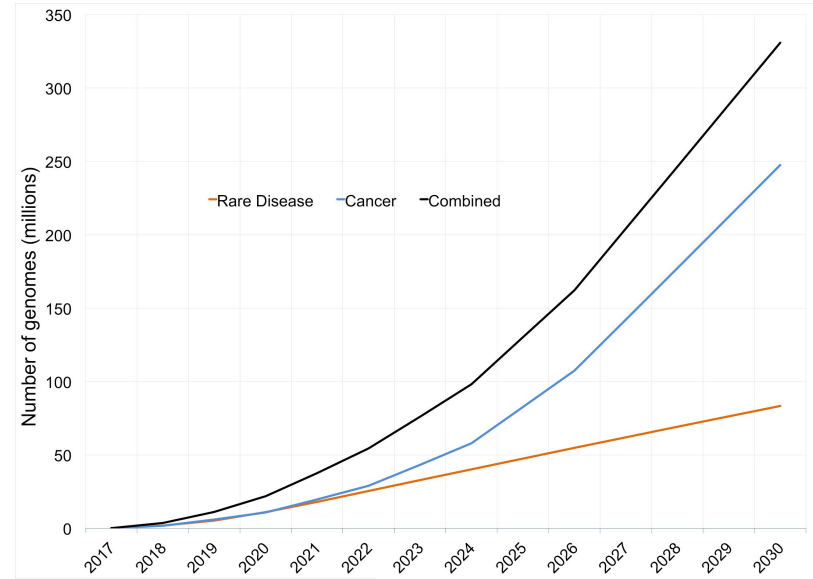
E: tk2@ebi.ac.uk

T: @drtkeane

W: <https://ega-archive.org/>

Global genetic data sharing

- Historically via International Nucleotide Sequence Database Collaboration (INSDC)
- Human genetic data sharing in research
 - dbGaP, JGA, and EGA
- Emergence of national scale cohorts
 - Jurisdictional restrictions
 - Enhanced data security
 - Political sensitivity around export



Tens of millions of genomes will be sequenced

Genomics in healthcare: GA4GH looks to 2022


© Ewan Birney, © Jessica Vamathevan, Peter Goodhand
doi: <https://doi.org/10.1101/203554>

Genomics meets healthcare

Percentage of whole genomes and exomes that are funded solely by healthcare systems

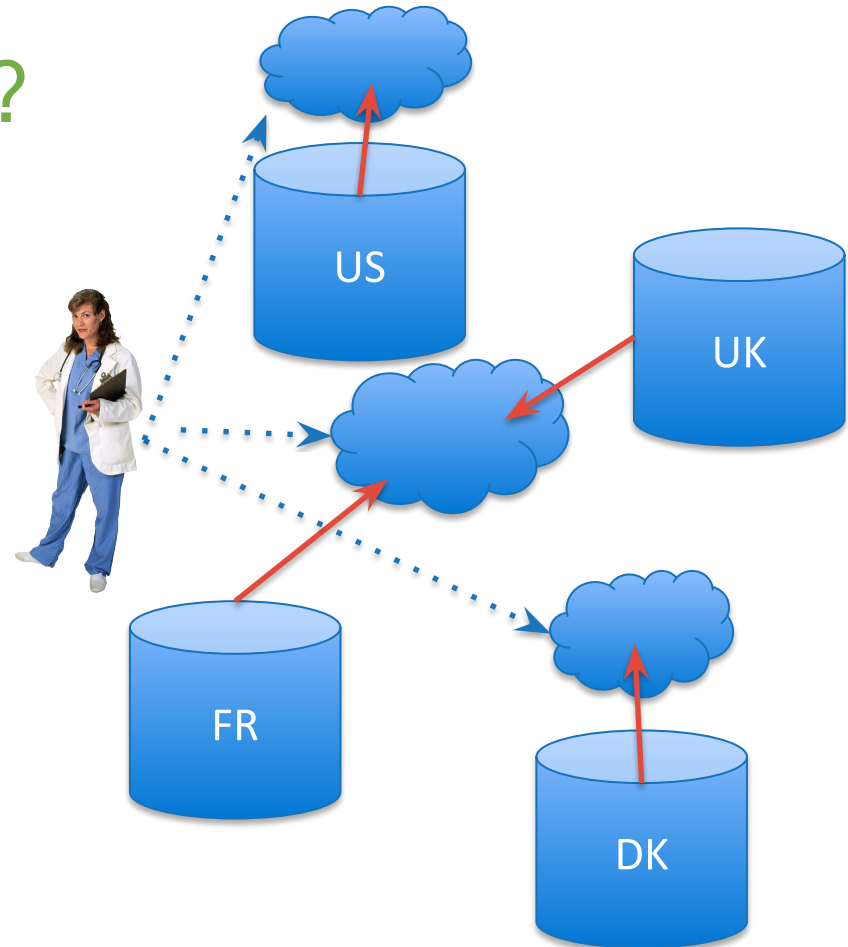


Genomics in healthcare: GA4GH looks to 2022

 Ewan Birney,  Jessica Vamathevan, Peter Goodhand
doi: <https://doi.org/10.1101/203554>

What are the challenges?

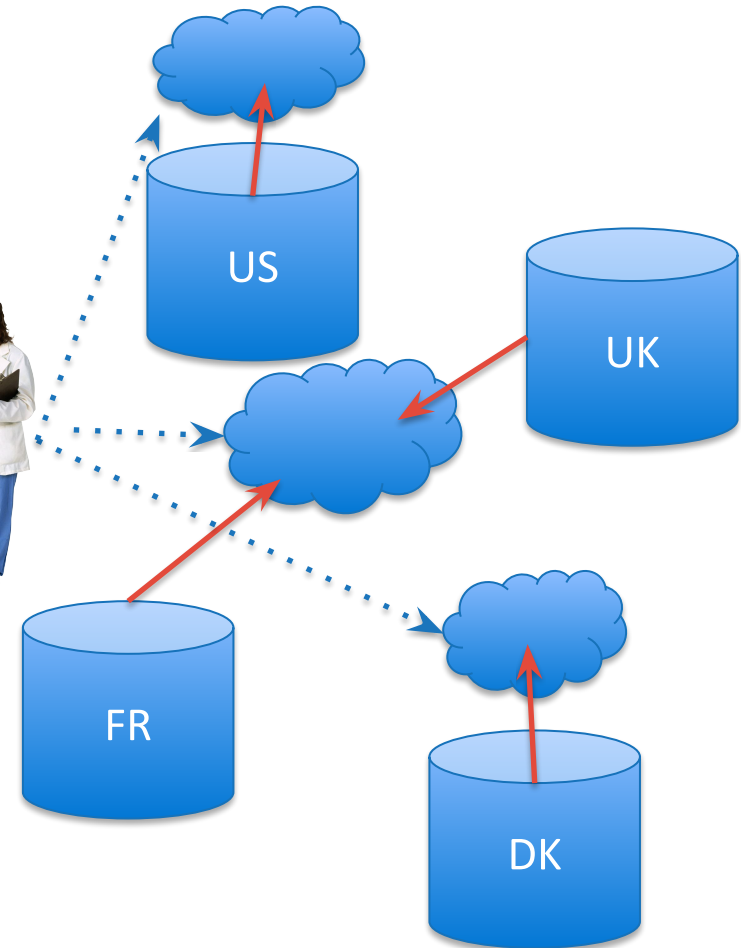
- 1) Data discoverability
- 2) Secure data sharing
- 3) Permissions automation
- 4) Interoperable researcher IDs
- 5) Portable pipelines



What are the challenges?

- 1) Data discoverability
- 2) Secure data sharing
- 3) Permissions automation
- 4) Interoperable researcher IDs
- 5) Portable pipelines

Interoperability





Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



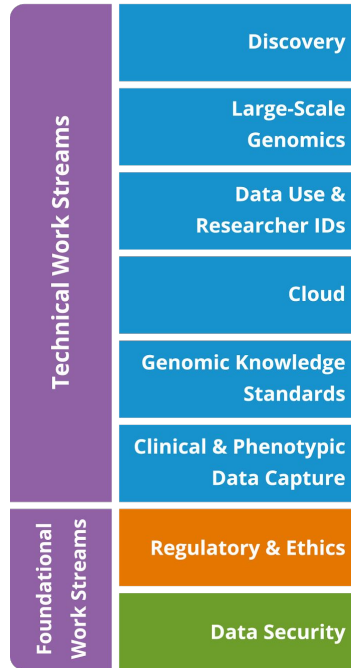
Global Alliance members include:

1. Universities and research institutes (32%)
2. Academic medical centers and health systems (10%)
3. Disease advocacy organizations and patient groups (5%)
4. Consortia and professional societies (5%)
5. Funders and agencies (5%)
6. Life science and information technology companies (43%)

GA4GH Connect



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

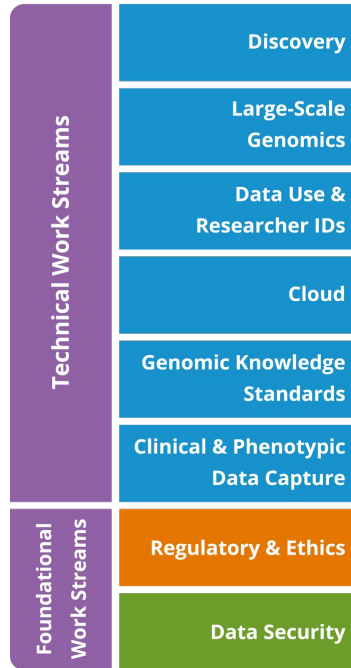


GA4GH Connect



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

Real-World Driver Projects



GA4GH Connect



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

		Real-World Driver Projects									
Technical Work Streams	Discovery	✓		✓		✓		✓			
	Large-Scale Genomics		✓		✓		✓		✓		
	Data Use & Researcher IDs	✓		✓		✓	✓				✓
	Cloud		✓	✓						✓	
	Genomic Knowledge Standards		✓				✓	✓	✓		
	Clinical & Phenotypic Data Capture	✓			✓	✓	✓				✓
Foundational Work Streams	Regulatory & Ethics										
	Data Security										

Partner Engagement

2017 Driver Projects



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



All of Us Research Program
United States



Australian Genomics
Australia



BRCA Challenge
International



CanDIG
Canada



ClinGen
United States



ELIXIR Beacon
Europe



ENA / EVA / EGA
Europe



Genomics England
United Kingdom



Human Cell Atlas
International



ICGC-ARGO
International



Matchmaker Exchange
International



Monarch Initiative
International



**National Cancer Institute
Genomic Data Commons**
United States



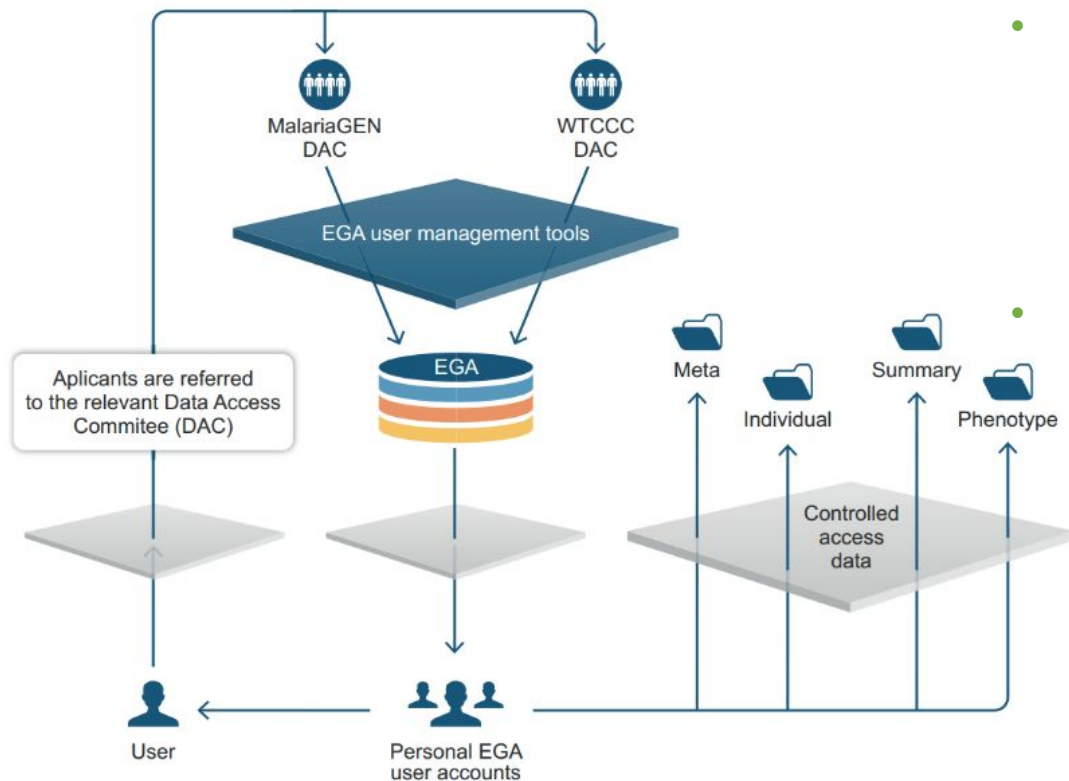
TOPMed
United States



**Variants Interpretation
for Cancer Consortium**
International

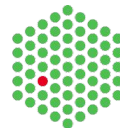


European Genome-phenome archive (EGA)



- Permanent secure archiving, accessioning, and sharing of all types of potentially identifiable genetic and phenotypic data
- Dataset access is controlled by the local Data Access Committees (DACs)

EMBL-EBI



European Genome-phenome archive (EGA)

How the EGA is managed

The EGA was launched in 2008 by the EBI



In 2012, EBI and CRG started working together to establish EGA as a joint venture, which has grown in the context of ELIXIR



European Genome-phenome archive (EGA)

By the numbers

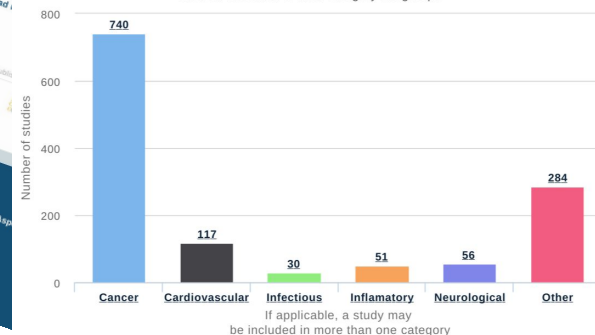
- 1,698 studies
- 3,591 datasets
- 777 data providers
- >10,000 requestors

By volume

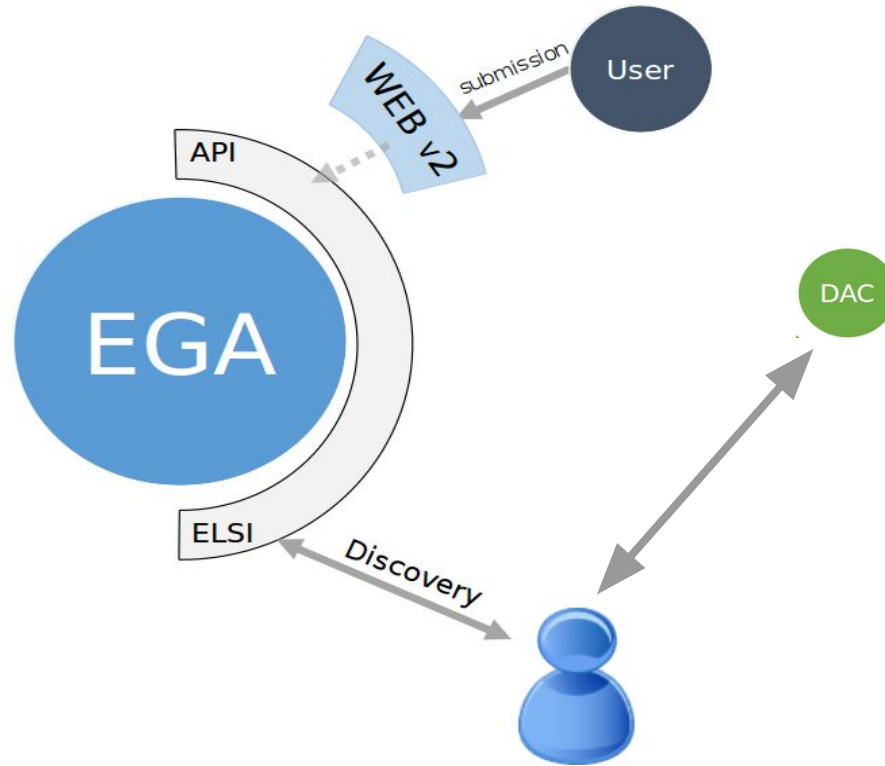
- 3.5 Petabytes



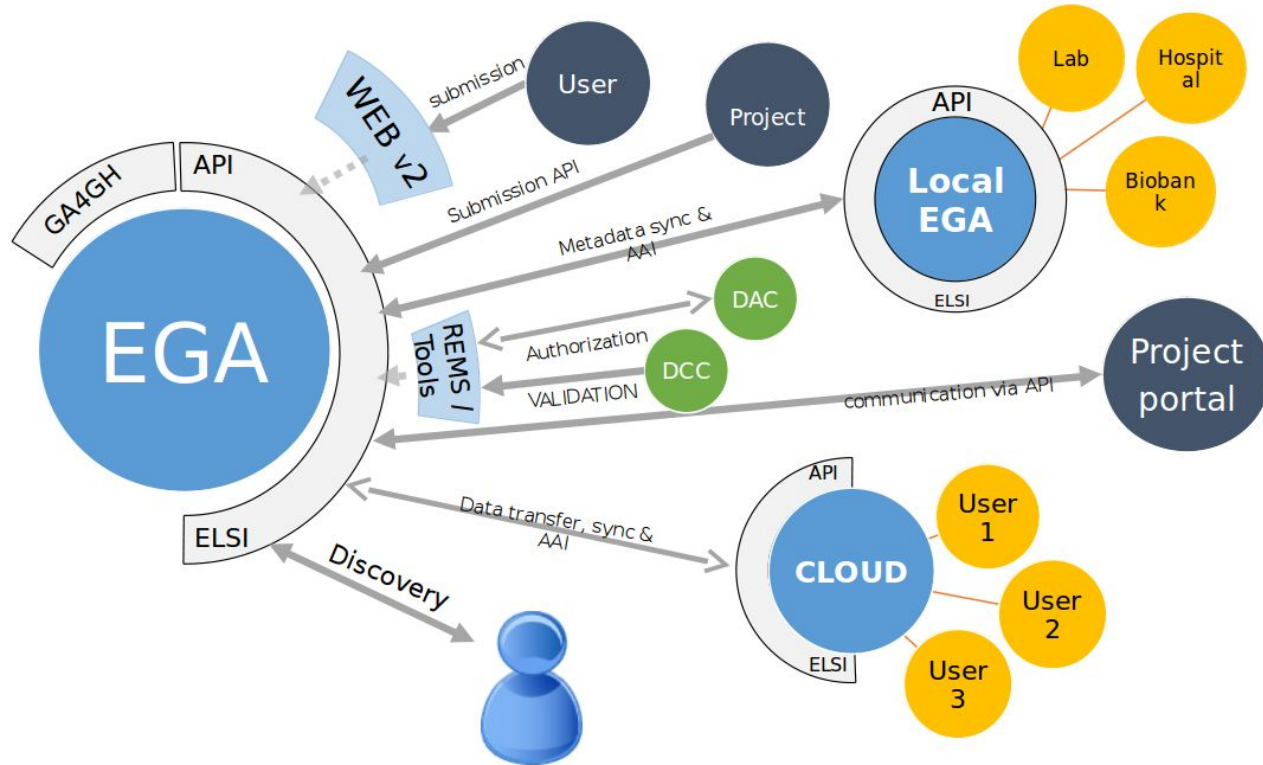
Studies in the EGA by disease
Click on a column to view category subgroups



EGA (circa 2015)

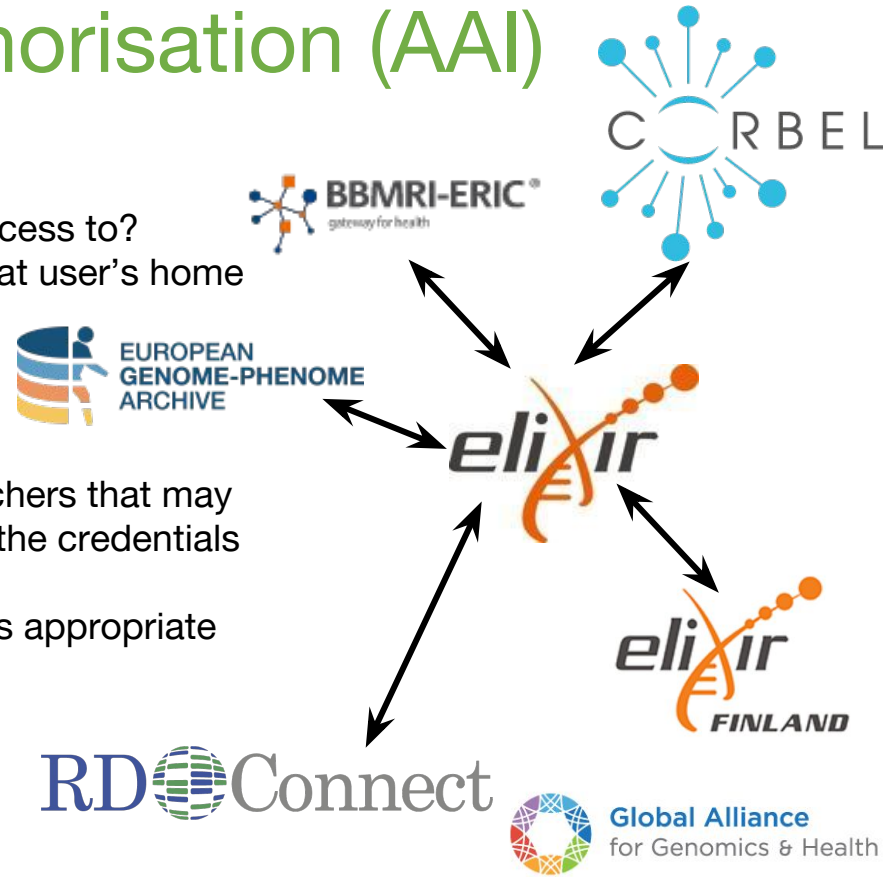


A federation of EGA nodes



Authentication and Authorisation (AAI)

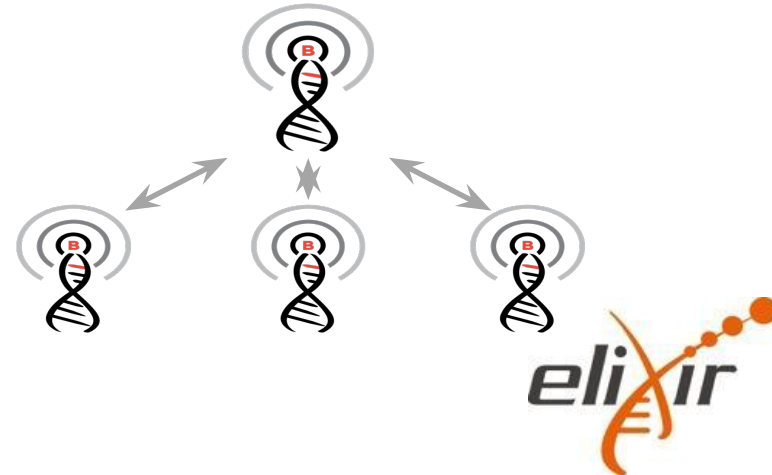
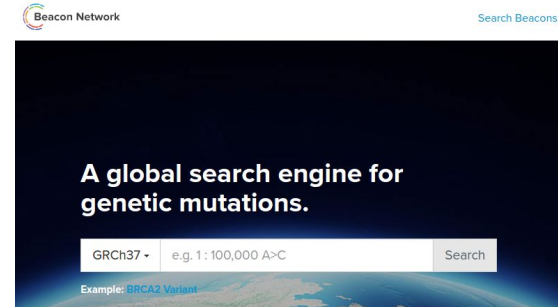
- Fundamental requirements
 - Who are you? What are your credentials?
 - What resources have you been granted access to?
 - Securely share user information managed at user's home organisation with remote services
- Two axes of human data sharing
 - **Researcher Identity:** Collection of researchers that may access the dataset at any given time, and the credentials they must supply.
 - **Data Use:** Informed consent form specifies appropriate restrictions on secondary data use.
- GA4GH security working group



Data Discovery

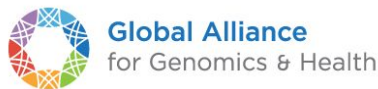
- Discovery by genotype
 - Beacon Network
 - Do you have evidence for this allele in your database?
 - Open, registered, and controlled access

- Consent codes
 - Groups datasets together in terms of consent agreements
 - Researchers can search by access requirements
 - Data Use Ontology (DUO)
 - Automate dataset application / access



Data Discovery

- Discovery by genotype
 - Beacon Network
 - Do you have evidence for this allele in your database?
 - Open, registered, and controlled access
- Consent codes
 - Groups datasets together in terms of consent agreements
 - Researchers can search by access requirements
 - Data Use Ontology (DUO)
 - Automate dataset application / access



VIEWPOINTS

Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke^{1*}, Anthony A. Philippakis², Jordi Rambla De Argila^{3,4}, Dina N. Paltoo⁵, Erin S. Luetkemeier⁶, Bartha M. Knoppers¹, Anthony J. Brookes⁷, J. Dylan Spalding⁸, Mark Thompson⁹, Marco Roos⁹, Kym M. Boycott⁹, Michael Brudno^{10,11}, Matthew Hurles¹², Heidi L. Rehm^{2,13}, Andreas Matern¹⁴, Marc Fiume¹⁵, Stephen T. Sherry¹⁶

The screenshot shows the European Genome-Phenome Archive (EGA) website. The main content area displays the dataset "WTCCC1 project samples from 1958 British Birth Cohort". A table lists the dataset ID (EGAD000000000001), technology (Athyraetix 500K), and number of samples (1504). Below this, there is a section for "Who controls access to this dataset" which identifies the Wellcome Trust Case Control Consortium Data Access Committee as the contact person. A "Downloads" section indicates that access is restricted and requires collaboration. A "Terms of Use" section includes a link to consent codes and a "Collaboration required" callout box. At the bottom, a table lists 10 studies associated with this dataset, each with a Study ID, Title, and Study Type (GWAS).

Dataset ID	Technology	Samples
EGAD000000000001	Athyraetix 500K	1504

Who controls access to this dataset
For each dataset that requires controlled access, there is a corresponding Data Access Committee (DAC) who determine access permissions. Access to actual data files is not managed by the EGA. If you need to request access to this data set, please contact:
Contact person: DAC contact
Wellcome Trust Case Control Consortium Data Access Committee
Email: datasharing [at] sanger [dot] ac [dot] uk
More details: EGAC000000000001

Downloads
You don't have access to the download section.

Terms of Use
See the full list of consent codes here.
HMB(CC) DUO NMDL COL-13 15

This dataset is featured in 10 studies
Studies are experimental investigations of a particular phenomenon, e.g. case-control studies on a particular trait or cancer research projects reporting matching cancer normal genomes from patients. Click on one of the Study IDs below to find out more.

Study ID	Study Title	Study Type
EGAS000000000001	WTCCC case-control study for Bipolar Disorder	GWAS
EGAS000000000002	WTCCC case-control study for Bipolar Disorder - Combined Controls	GWAS
EGAS000000000003	WTCCC case-control study for Coronary Artery Disease	GWAS
EGAS000000000004	WTCCC case-control study for Coronary Artery Disease - Combined Controls	GWAS
EGAS000000000005	WTCCC case-control study for Coronary Artery Disease, Hypertension, T2D - combined cases	GWAS
EGAS000000000006	Genomewide Association Study of Inflammatory Bowel Disease	GWAS
EGAS000000000007	Genomewide Association Study of Inflammatory Bowel Disease - Combined Controls	GWAS
EGAS000000000008	WTCCC case-control study for Inflammatory Bowel Disease, T1D and RA - combined cases	GWAS
EGAS000000000009	WTCCC case-control study for Hypertension	GWAS
EGAS000000000010	WTCCC case-control study for Hypertension - Combined Controls	GWAS

Secure Data Transfer

- Secure access to sensitive human data
 - Sequencing read and variant data
- Use-cases
 - Real-time secure streaming to project portal
 - Inspection of underlying read data in a read viewer (e.g. IGV)
 - Bring the analysis pipelines to the data
 - Standardised API for accessing the genetic data
- Htsget protocol
 - v1.0 launched in October



Key partnerships

- ELIXIR

- EXCELERATE (2015-2019)
- Implementation studies
 - TrAIT: EGA backend for TranSMART (2016), IMI Oncotrack (2016), Real-time RD-visualisation (2017), Beacon (2017)



- GA4GH

- Large Scale Genomics
- Data Use and Researcher ID
- Discovery
- Clinical data and phenotype capture
- Data Security
- Regulatory and Ethics



Projects

- AMP-T2D
 - Federated browsing, searching, and analysis of human genetic information linked to type 2 diabetes and related traits
- RD-Connect
 - Platform connecting databases, registries, biobanks and clinical bioinformatics
- UK Biobank
 - Sharing genetic data from hundreds of thousands of individuals
- CORBEL



EGA Team

EMBL-EBI

