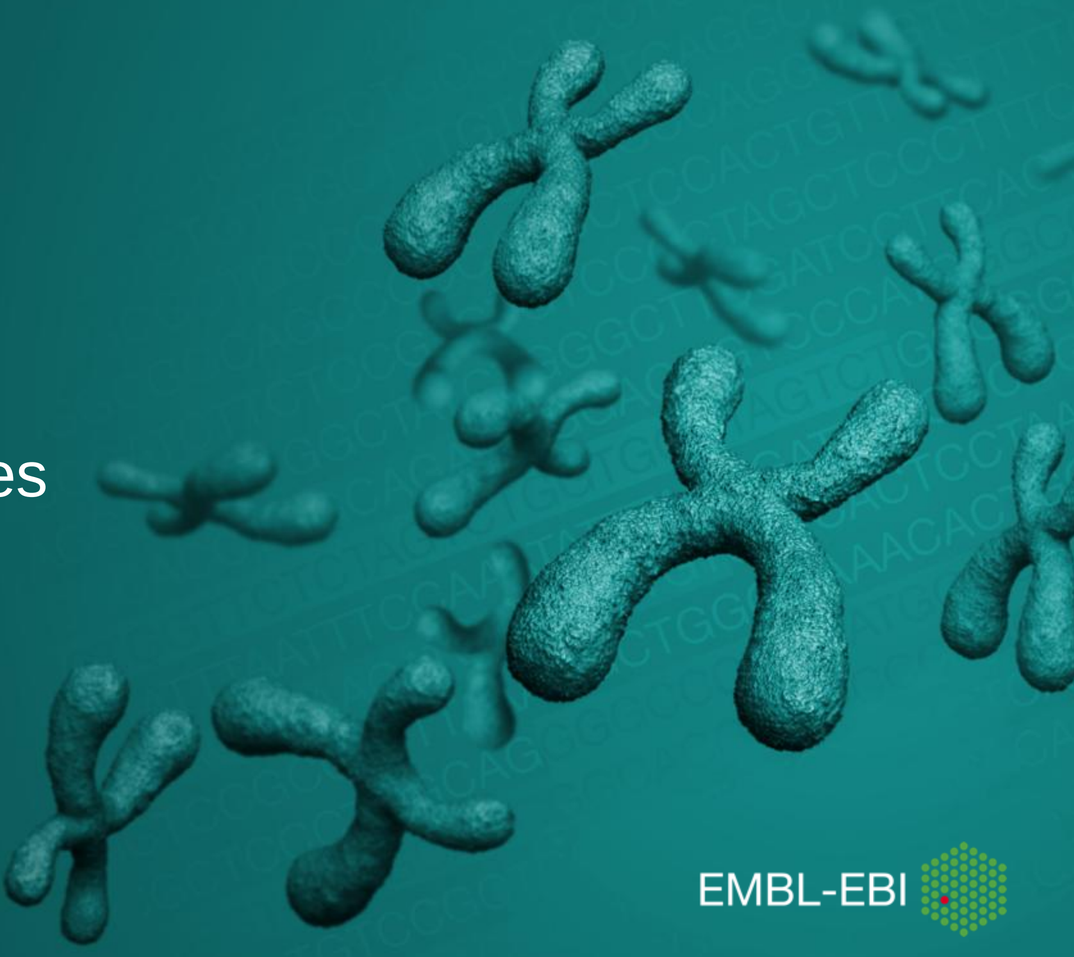
European Genome phenome Archive at the European Bioinformatics Institute

Helen Parkinson

Head of Molecular Archives



EMBL-EBI

What is EMBL-EBI?

- International, non-profit research institute
- Part of the European Molecular Biology Laboratory
- Europe’s hub for biological data services and research
- ~600 members of staff from 53 nations.

70 Petabytes of storage
>40,000 CPU Cores
20 Gbit Internet connection

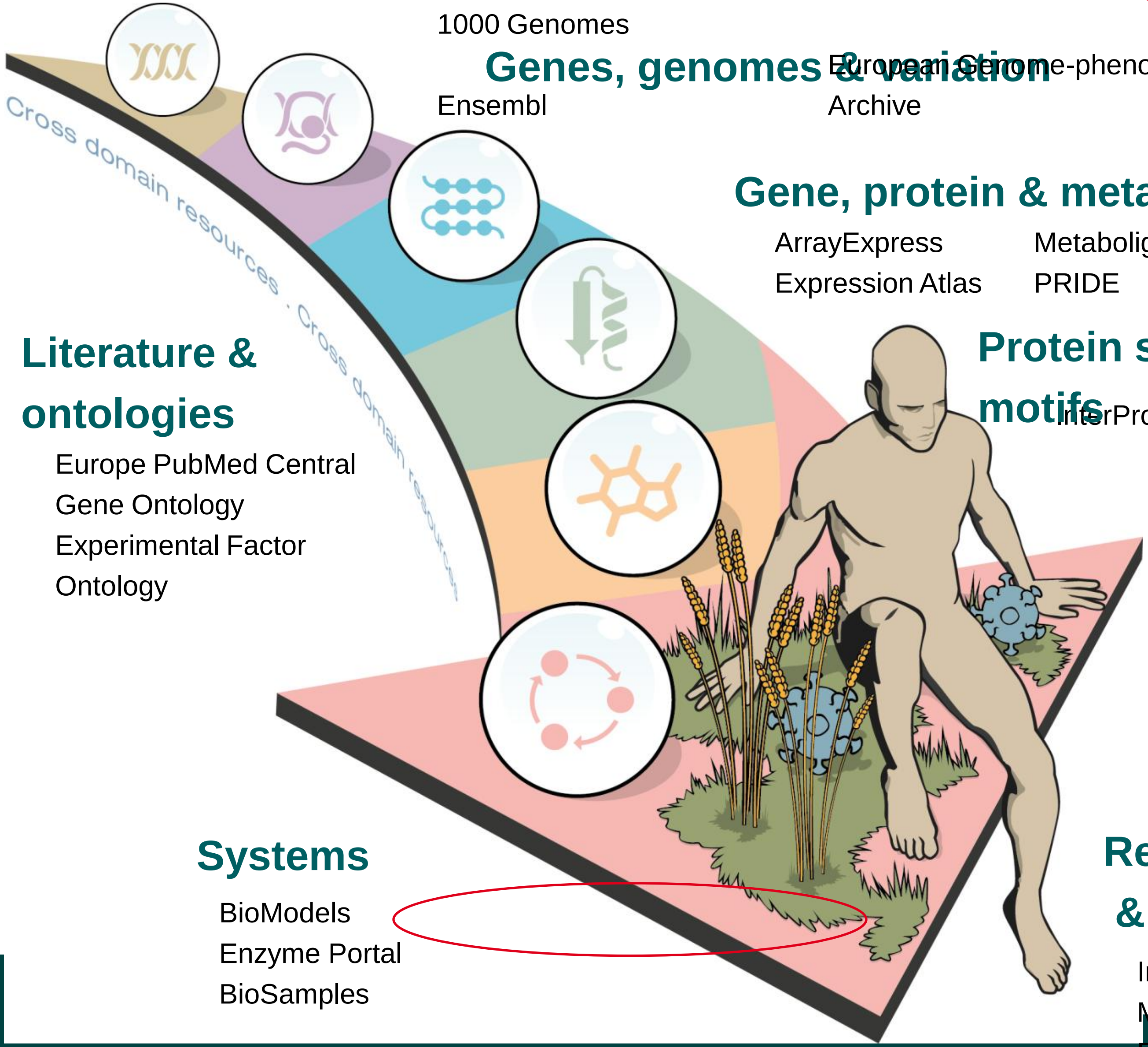
Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes
Ensembl

Ensembl Genomes
European Genome-phenome Archive

Metagenomics portal
GWAS Catalog



Gene, protein & metabolite expression

ArrayExpress
Expression Atlas

Metabolights
PRIDE

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL
ChEBI

Systems

BioModels
Enzyme Portal
BioSamples

Reactions, interactions & pathways

Intact
Metabolights
Reactome

EBI as a use case for ‘data linkage’

… there are difficulties in defining boundaries around what is technically possible, what is legally permitted, and what should be done ethically. Indeed, linking data can also compromise the privacy of an individual's personal information, making them identifiable …

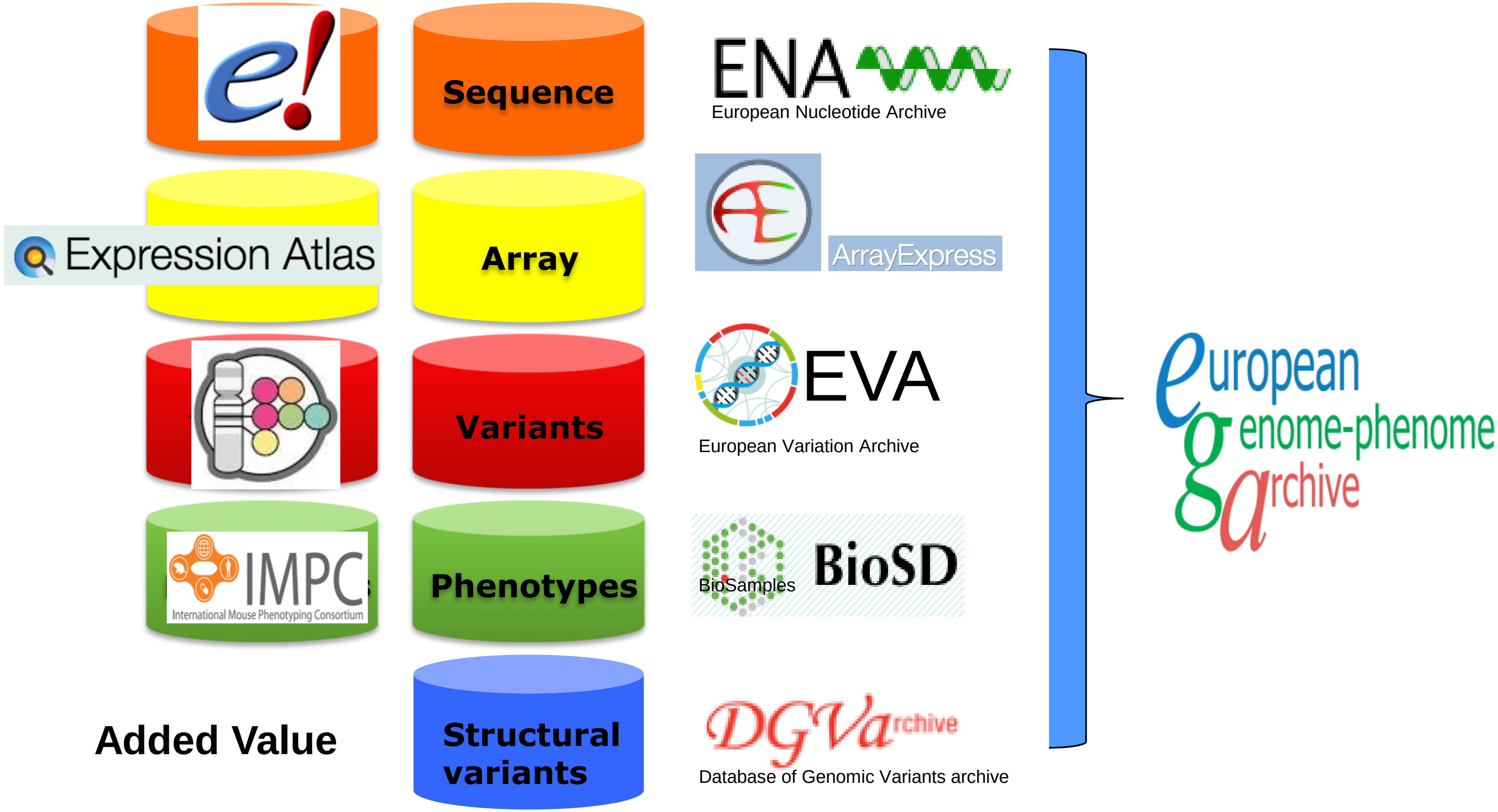
Use case

Ethical issues

Technical requirements

balancing the drive for automation against the need for human judgement

EBI’s Data Sources



Expression Atlas
IMPC
International Mouse Phenotyping Consortium
ENA
ArrayExpress
EVA
BioSamples
BioSD
DGVa rchive
european genome-phenome archive

Sequence

Array

Variants

Phenotypes

Structural variants

European Nucleotide Archive

European Variation Archive

Database of Genomic Variants archive

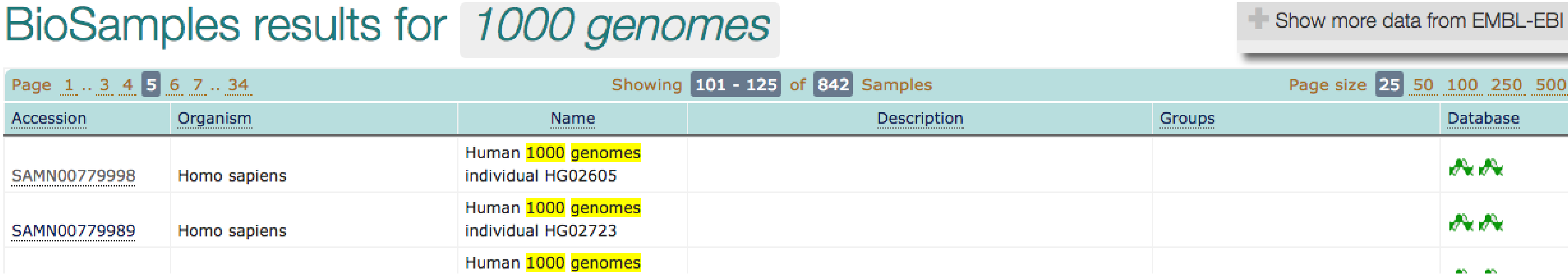
Added Value

Open & public archives

Controlled access archive

BioSamples Database at the EBI

- Repository of information about biological materials, or "samples" – 4.3 million samples, 2 million human
- Aggregates reference samples – such as HapMap, Coriell cell lines or samples from 1000 genomes – with sample data from EBI assay databases
- Enables cross linking of assay data to metadata about the samples from which they are derived
- Accepts direct submission of sample data

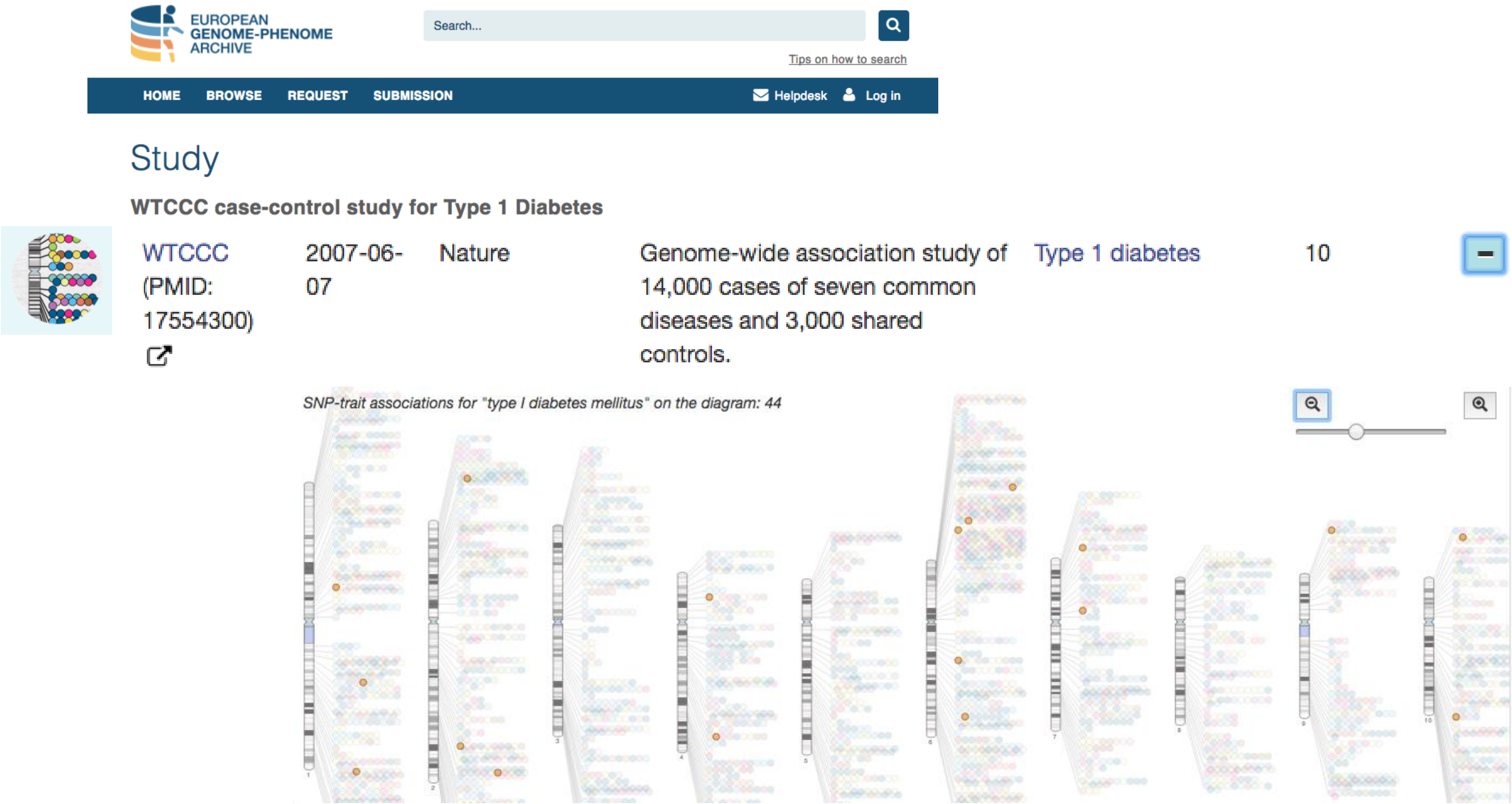


BioSamples results for *1000 genomes*

Page 1 .. 3 4 5 6 7 .. 34 Showing 101 - 125 of 842 Samples Page size 25 50 100 250 500

Accession	Organism	Name	Description	Groups	Database
SAMN00779998	Homo sapiens	Human 1000 genomes individual HG02605			
SAMN00779989	Homo sapiens	Human 1000 genomes individual HG02723			
		Human 1000 genomes			

Summary vs. consented data



Study

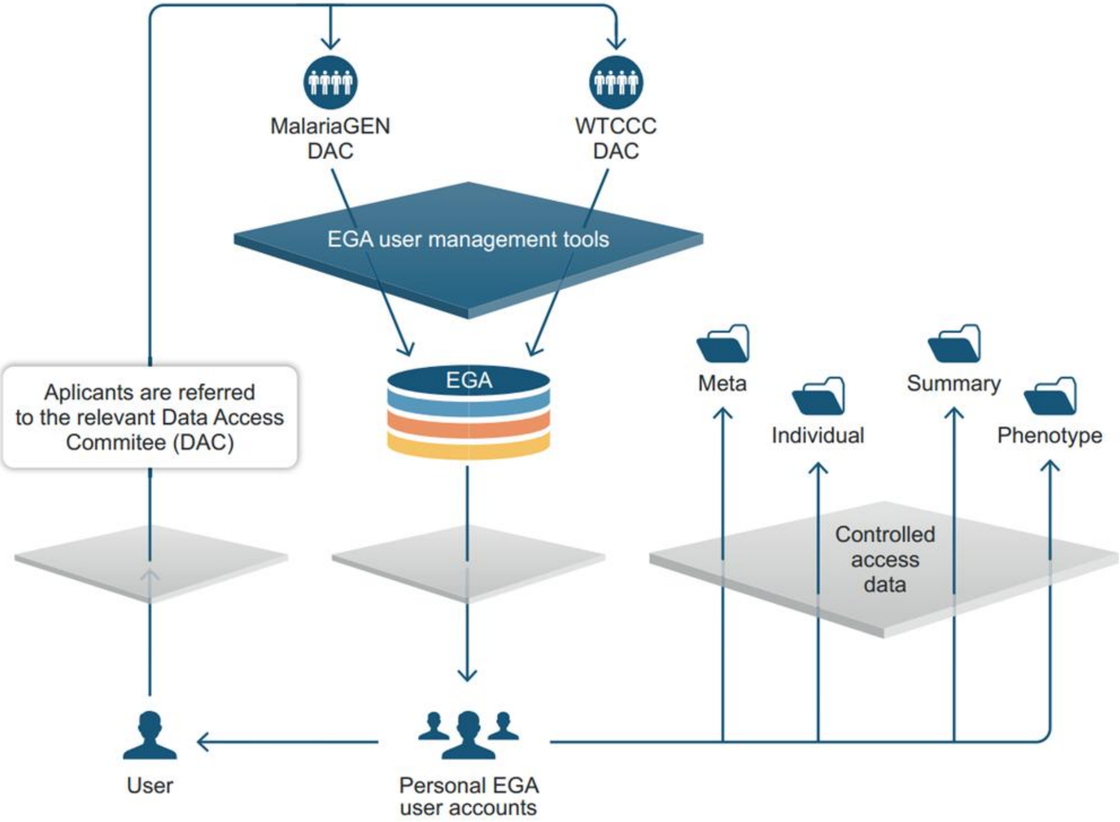
WTCCC case-control study for Type 1 Diabetes

WTCCC (PMID: 17554300)	2007-06-07	Nature	Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.	Type 1 diabetes	10

SNP-trait associations for "type I diabetes mellitus" on the diagram: 44

What is the EGA?

The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable genetic and phenotypic data resulting from biomedical research projects..


MalariaGEN DAC
WTCCC DAC
EGA user management tools
EGA
Aplicants are referred to the relevant Data Access Commitee (DAC)
Meta
Individual
Summary
Phenotype
Controlled access data
User
Personal EGA user accounts

Data is provided by research centers and health care institutions.

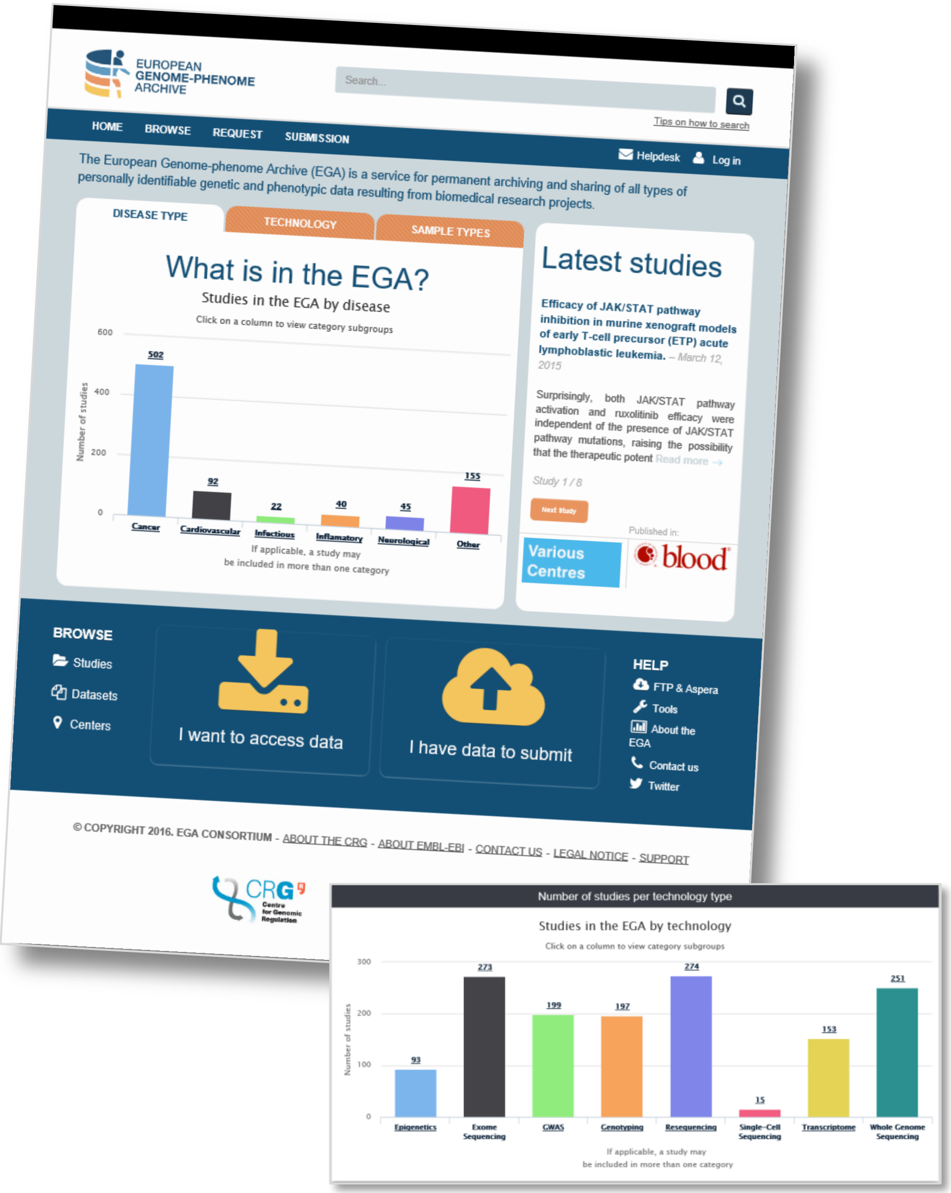
Access is controlled by Data Access Committees.

Data requesters are researchers from other research or health care institutions.

Data and services and Federated between CRG and EBI

EGA Numbers

- 911 Studies (31% growth in 2015)
- 1,966 Datasets (46% growth)
- 492 Data providers (92% growth)
- >8,000 Data requesters (20% growth year on year)
- 200 helpdesk tickets per month



The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

What is in the EGA?

Studies in the EGA by disease

Click on a column to view category subgroups

If applicable, a study may be included in more than one category

Latest studies

Efficacy of JAK/STAT pathway inhibition in murine xenograft models of early T-cell precursor (ETP) acute lymphoblastic leukemia. – *March 12, 2015*

Surprisingly, both JAK/STAT pathway activation and ruxolitinib efficacy were independent of the presence of JAK/STAT pathway mutations, raising the possibility that the therapeutic potent Read more →

Study 1 / 8

© COPYRIGHT 2016. EGA CONSORTIUM - ABOUT THE CRG - ABOUT EMBL-EBI - CONTACT US - LEGAL NOTICE - SUPPORT

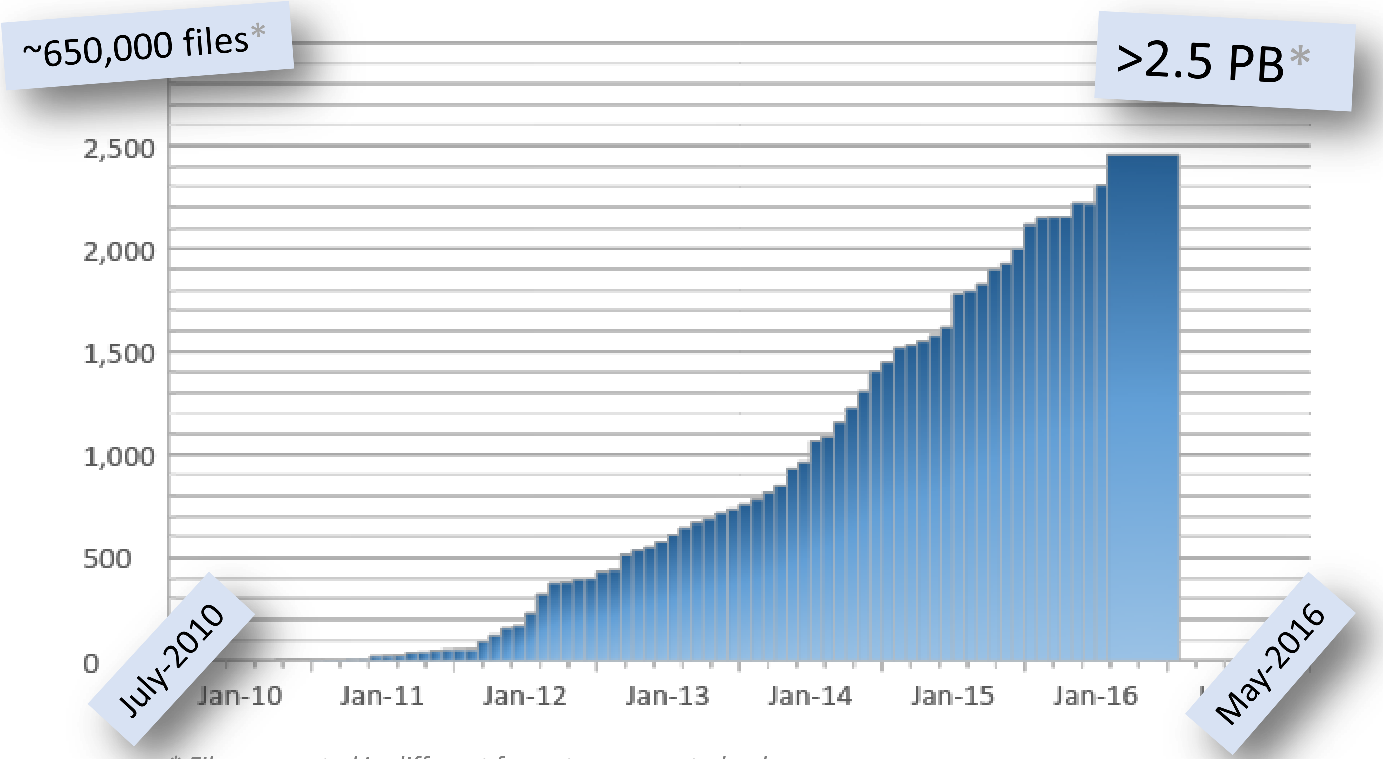
Number of studies per technology type

Studies in the EGA by technology

Click on a column to view category subgroups

If applicable, a study may be included in more than one category

The EGA contains a growing amount of data



* *Files encrypted in different formats are counted only once*

International Cancer Genome Consortium

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

UK 10K

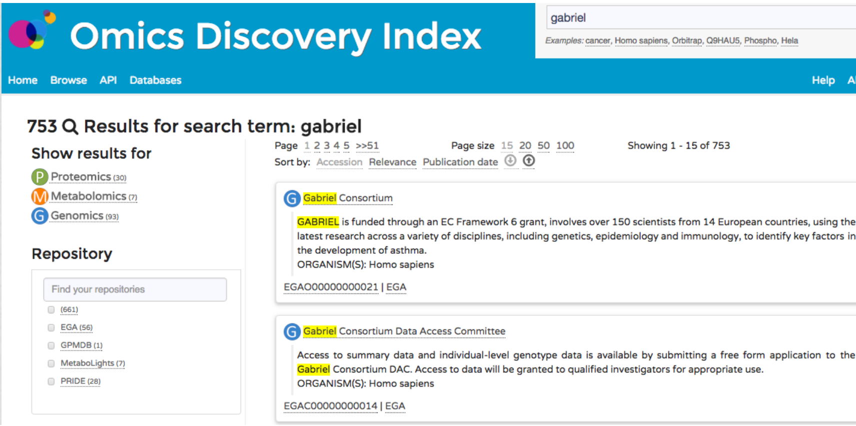
wellcome trust sanger institute

Data Sharing and Access

- EGA operates at two sites EBI and CRG, Barcelona
- All data hosted at EBI – highly accessed data stored at CRG
- Study level meta data is shared between dbGAP (USA) and in future DDBJ
- Study level data is shared in the BD2K funded Data Discovery Index
- GA4GH Beacon provides variant le

Since January 1, 2015

- 1921245 files distributed
- 4.5 PB in Data Volume distribu


Omics Discovery Index
gabriel
Examples: cancer, Homo sapiens, Orbitrap, Q9HAU5, Phospho, Hela
Home Browse API Databases
Help A

753 Results for search term: gabriel

Show results for

- Proteomics (30)
- Metabolomics (7)
- Genomics (93)

Repository

Find your repositories

- □ (661)
- □ EGA (56)
- □ GPMDB (1)
- □ MetaboLights (7)
- □ PRIDE (28)

Page 1 2 3 4 5 >>51   Page size 15 20 50 100   Showing 1 - 15 of 753

Sort by: Accession Relevance Publication date

Gabriel Consortium

GABRIEL is funded through an EC Framework 6 grant, involves over 150 scientists from 14 European countries, using the latest research across a variety of disciplines, including genetics, epidemiology and immunology, to identify key factors in the development of asthma.
ORGANISM(S): Homo sapiens

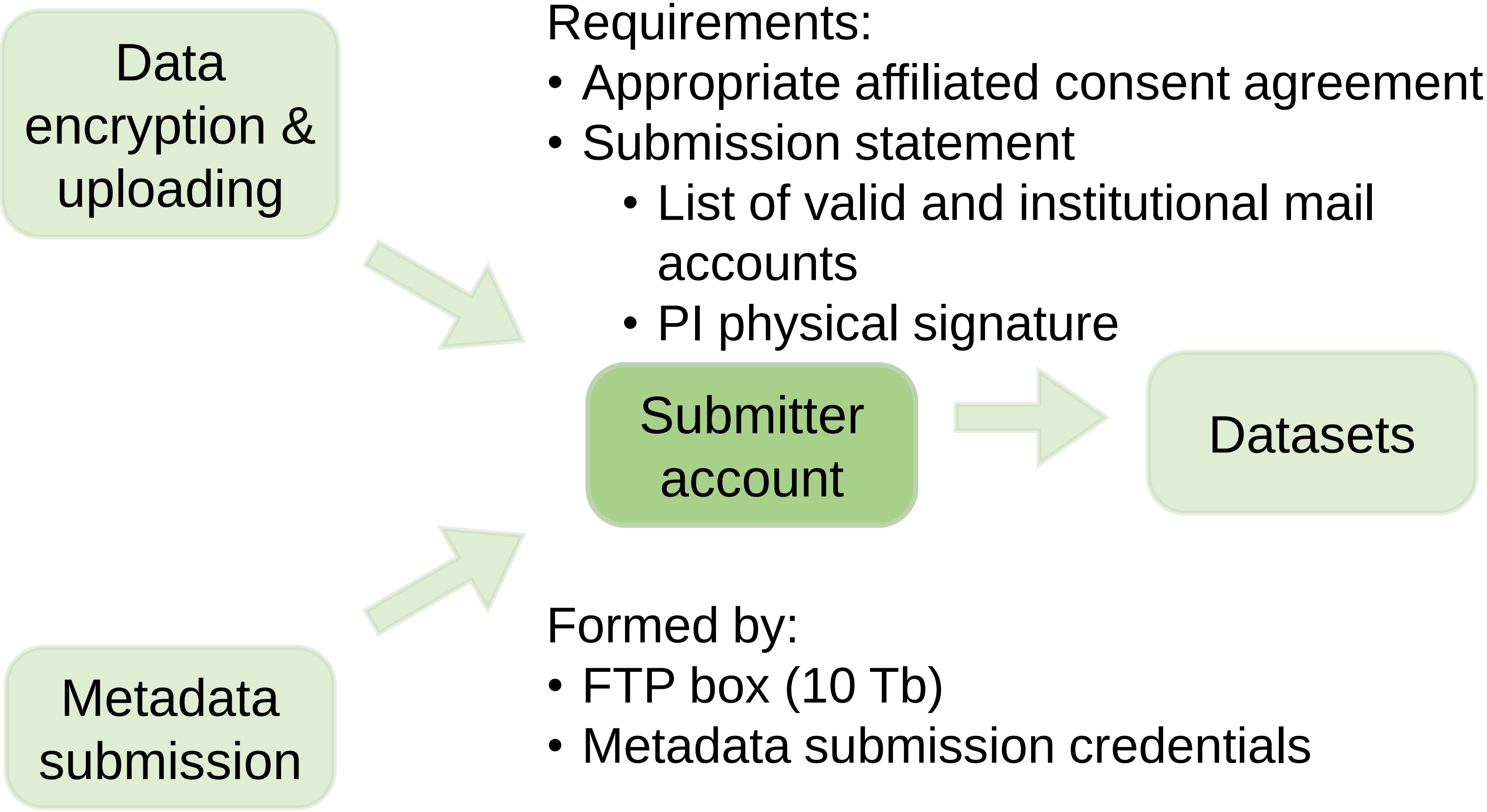
EGAO00000000021 | EGA

Gabriel Consortium Data Access Committee

Access to summary data and individual-level genotype data is available by submitting a free form application to the Gabriel Consortium DAC. Access to data will be granted to qualified investigators for appropriate use.
ORGANISM(S): Homo sapiens

EGAC00000000014 | EGA

Submitter accounts



Data encryption & uploading

Metadata submission

Submitter account

Datasets

Requirements:

- Appropriate affiliated consent agreement
- Submission statement
  - List of valid and institutional mail accounts
  - PI physical signature

Formed by:

- FTP box (10 Tb)
- Metadata submission credentials

DAC Demographics

- 322 DACs
- Oversight of some 1966 datasets

- Created 8056 individual user accounts

- Our biggest DAC is Wellcome Trust Sanger Institute – which has 664 datasets

- WTSI has 3 DACs in top ten with4 K users accessing them

- DACs creating >5 accounts per month provided with secureID key

Consent Codes

Consent Codes: Upholding Standard Data Use Conditions

Stephanie O. M. Dyke, Anthony A. Philippakis, Jordi Rambla De Argila, Dina N. Paltoo, Erin S. Luetkemeier, Bartha M. Knoppers, Anthony J. Brookes, J. Dylan Spalding, Mark Thompson, Marco Roos, Kym M. Boycott, Michael Brudno, Matthew Hurles, [...], Stephen T. Sherry [view all]

Published: January 21, 2016 • http://dx.doi.org/10.1371/journal.pgen.1005772

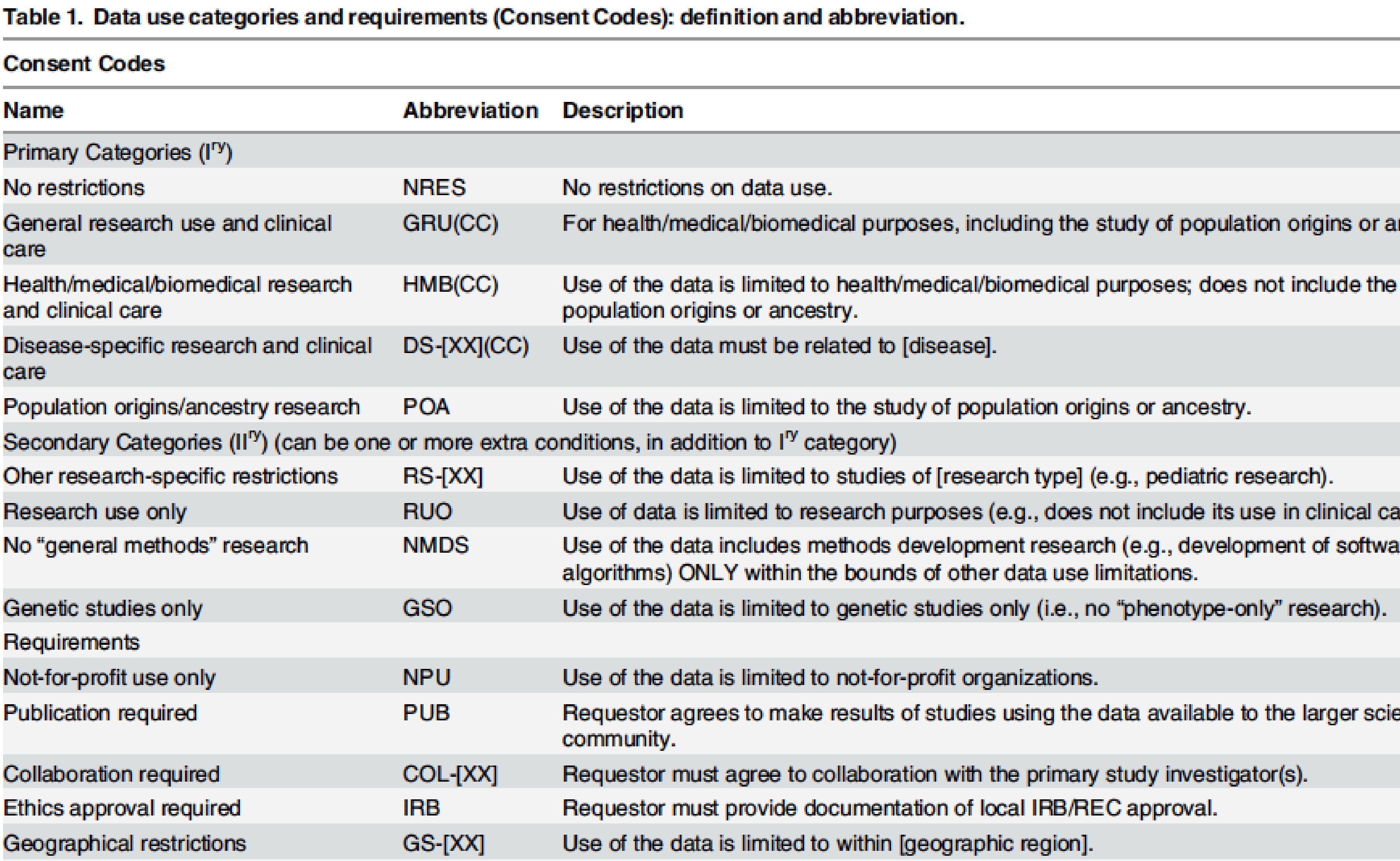
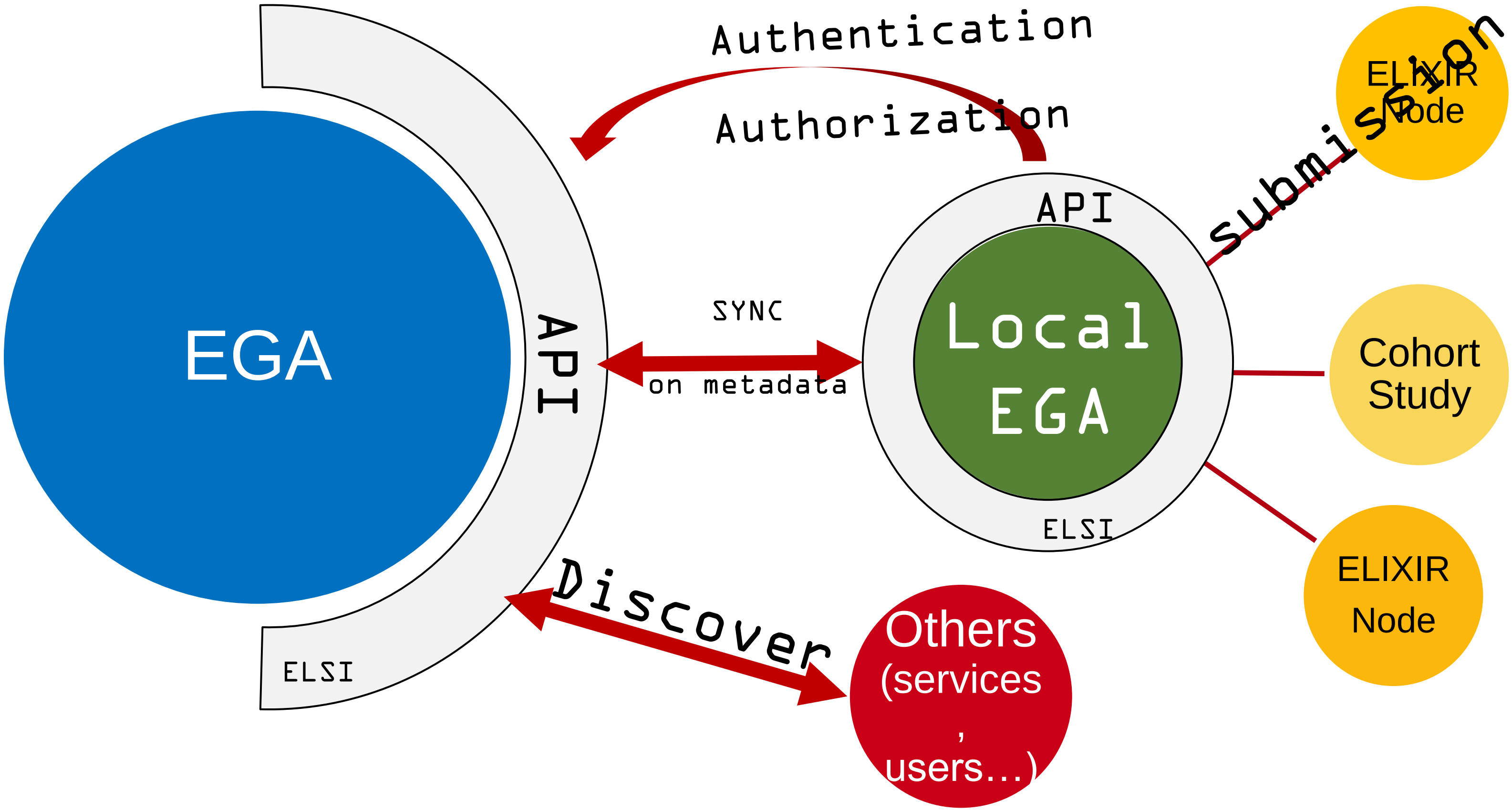


Table 1. Data use categories and requirements (Consent Codes): definition and abbreviation.

Consent Codes

Name	Abbreviation	Description
Primary Categories (I^{ry})		
No restrictions	NRES	No restrictions on data use.
General research use and clinical care	GRU(CC)	For health/medical/biomedical purposes, including the study of population origins or a
Health/medical/biomedical research and clinical care	HMB(CC)	Use of the data is limited to health/medical/biomedical purposes; does not include the population origins or ancestry.
Disease-specific research and clinical care	DS-[XX](CC)	Use of the data must be related to [disease].
Population origins/ancestry research	POA	Use of the data is limited to the study of population origins or ancestry.
Secondary Categories (II^{ry}) (can be one or more extra conditions, in addition to I^{ry} category)		
Oher research-specific restrictions	RS-[XX]	Use of the data is limited to studies of [research type] (e.g., pediatric research).
Research use only	RUO	Use of data is limited to research purposes (e.g., does not include its use in clinical ca
No "general methods" research	NMDS	Use of the data includes methods development research (e.g., development of softwa algorithms) ONLY within the bounds of other data use limitations.
Genetic studies only	GSO	Use of the data is limited to genetic studies only (i.e., no "phenotype-only" research).
Requirements		
Not-for-profit use only	NPU	Use of the data is limited to not-for-profit organizations.
Publication required	PUB	Requestor agrees to make results of studies using the data available to the larger scie community.
Collaboration required	COL-[XX]	Requestor must agree to collaboration with the primary study investigator(s).
Ethics approval required	IRB	Requestor must provide documentation of local IRB/REC approval.
Geographical restrictions	GS-[XX]	Use of the data is limited to within [geographic region].

Federated ‘Local’ EGA



“Offers solutions if data cannot leave the submitter facilities”

Local EGA

- Docker based Virtual Machine (infrastructure independent)
  - Database container
  - Data Archive container
  - API Access container
  - FTP server container
- Allows FTP upload of files, re-encryption, archiving, and distribution via downloader
- https://github.com/elixir-europe/human-data-local-ega

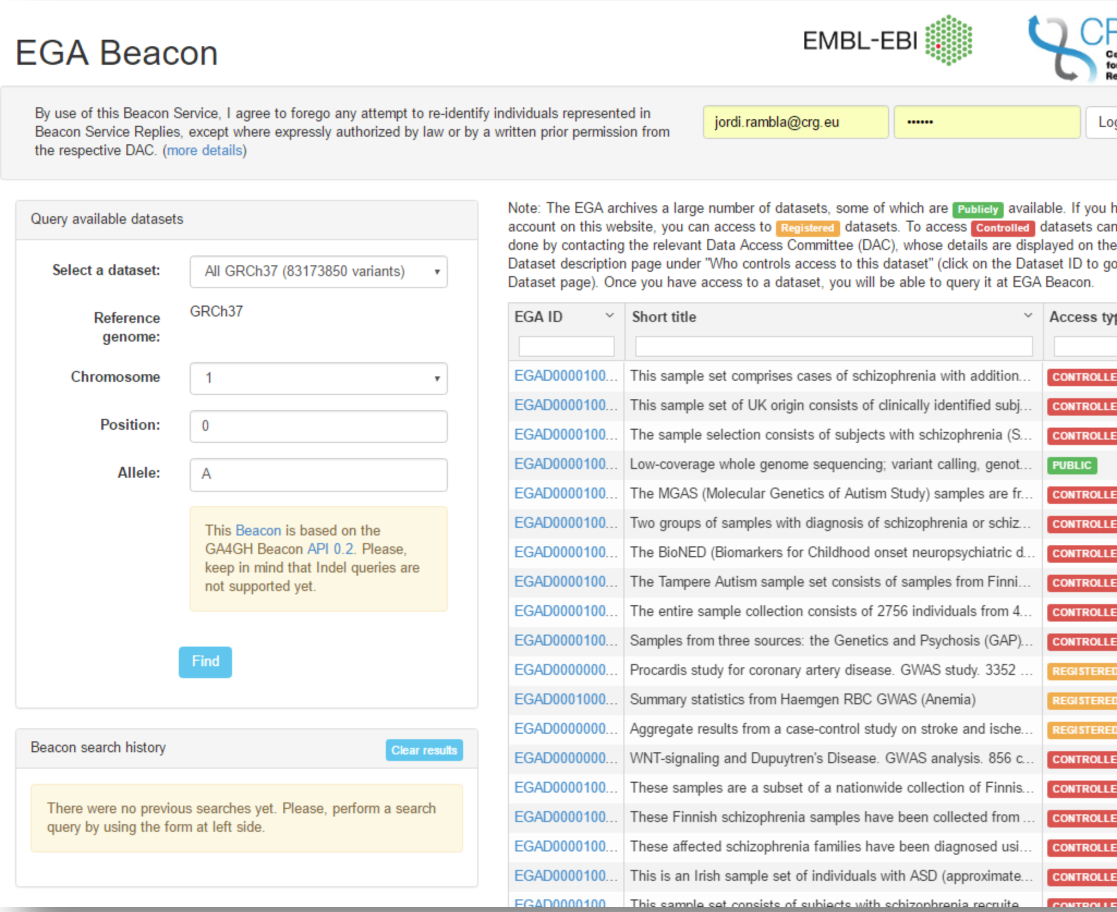
EGA Future

- Consent codes at data acquisition – transparency of consent
- EGA Local – for use as a country level EGA
- Shared meta data back to core for data discovery
- AAI Access – authentication, access, identification provides access to multiple resources with a common user id and AAI profile
- EGA in the cloud
- Addressing DAC scaling issues
- Improved usability of data – visualisation/slicing/query/input to analysis tools
- Improved streaming technology via GA4GH
- Improved submission tools - efficiency and helpdesk, content

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

The EGA Beacon

- GA4GH Beacons are a discovery service:
  - which datasets include genomes with allele of interest?
  - 60 Beacons, 160 datasets, 25 organisation
- ELIXIR Objectives:
  - Provide ELIXIR reference implementation
  - Provide an example on capacity build across ELIXIR Nodes
- ELIXIR pilot project with partners from the Netherlands, Sweden, Finland, France, Belgium



EGA Beacon

By use of this Beacon Service, I agree to forego any attempt to re-identify individuals represented in Beacon Service Replies, except where expressly authorized by law or by a written prior permission from the respective DAC. (more details)

jordi.rambla@crg.eu
••••••

Query available datasets

Select a dataset: All GRCh37 (83173850 variants)

Reference genome: GRCh37

Chromosome: 1

Position: 0

Allele: A

This Beacon is based on the GA4GH Beacon API 0.2. Please, keep in mind that Indel queries are not supported yet.

Beacon search history

There were no previous searches yet. Please, perform a search query by using the form at left side.

Note: The EGA archives a large number of datasets, some of which are Publicly available. If you h
account on this website, you can access to Registered datasets. To access Controlled datasets can
done by contacting the relevant Data Access Committee (DAC), whose details are displayed on the
Dataset description page under "Who controls access to this dataset" (click on the Dataset ID to go
Dataset page). Once you have access to a dataset, you will be able to query it at EGA Beacon.

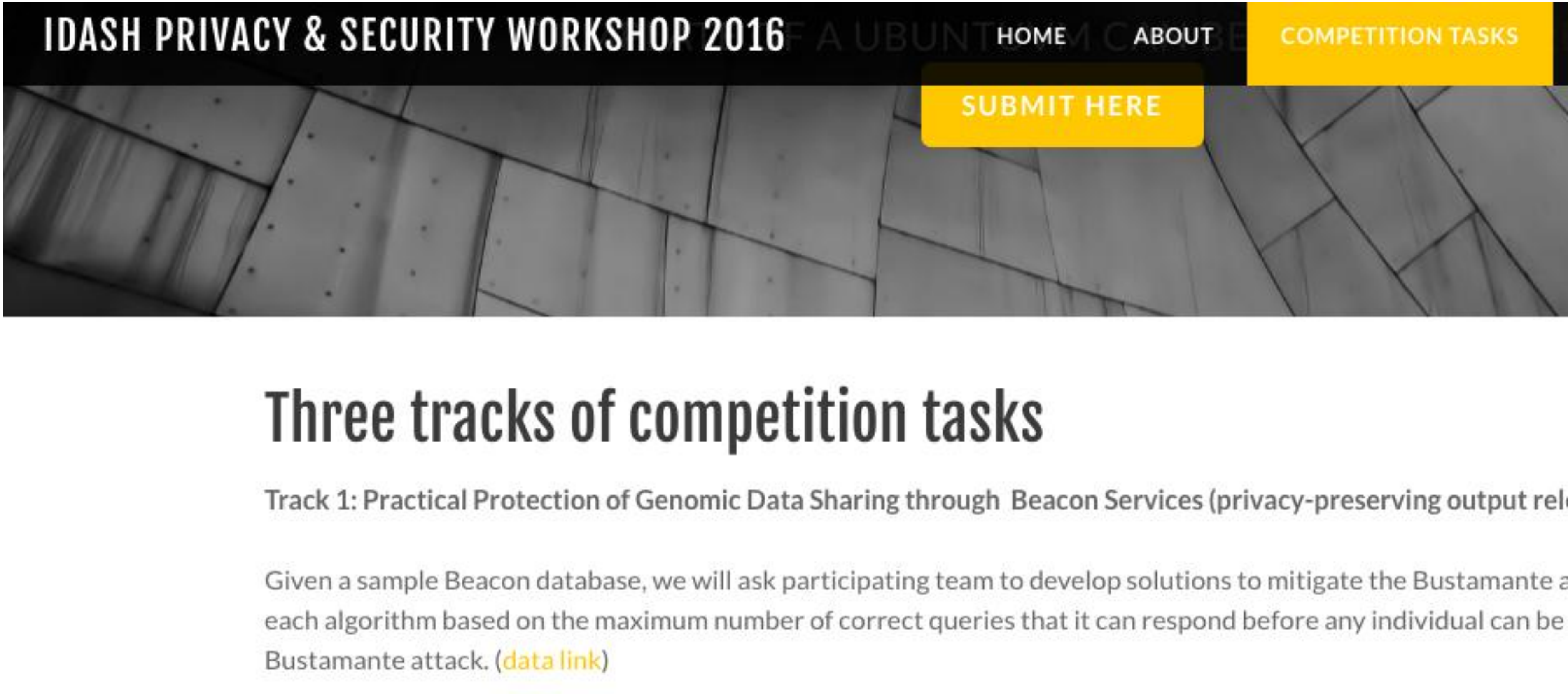
EGA ID	Short title	Access ty
EGAD0000100...	This sample set comprises cases of schizophrenia with addition...	CONTROLLE
EGAD0000100...	This sample set of UK origin consists of clinically identified subj...	CONTROLLE
EGAD0000100...	The sample selection consists of subjects with schizophrenia (S...	CONTROLLE
EGAD0000100...	Low-coverage whole genome sequencing; variant calling, genot...	PUBLIC
EGAD0000100...	The MGAS (Molecular Genetics of Autism Study) samples are fr...	CONTROLLE
EGAD0000100...	Two groups of samples with diagnosis of schizophrenia or schiz...	CONTROLLE
EGAD0000100...	The BioNED (Biomarkers for Childhood onset neuropsychiatric d...	CONTROLLE
EGAD0000100...	The Tampere Autism sample set consists of samples from Finni...	CONTROLLE
EGAD0000100...	The entire sample collection consists of 2756 individuals from 4...	CONTROLLE
EGAD0000100...	Samples from three sources: the Genetics and Psychosis (GAP)...	CONTROLLE
EGAD0000000...	Procardis study for coronary artery disease. GWAS study. 3352 ...	REGISTERED
EGAD0001000...	Summary statistics from Haemgen RBC GWAS (Anemia)	REGISTERED
EGAD0000000...	Aggregate results from a case-control study on stroke and ische...	REGISTERED
EGAD0000000...	WNT-signaling and Dupuytren's Disease. GWAS analysis. 856 c...	CONTROLLE
EGAD0000100...	These samples are a subset of a nationwide collection of Finnis...	CONTROLLE
EGAD0000100...	These Finnish schizophrenia samples have been collected from ...	CONTROLLE
EGAD0000100...	These affected schizophrenia families have been diagnosed usi...	CONTROLLE
EGAD0000100...	This is an Irish sample set of individuals with ASD (approximate...	CONTROLLE
EGAD0000100	This sample set consists of subjects with schizophrenia recruite	CONTROLLE

https://ega.crg.eu/beacon_web/#/

Beacon Challenges

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure, Carlos D. Bustamante



Three tracks of competition tasks

Track 1: Practical Protection of Genomic Data Sharing through Beacon Services (privacy-preserving output rel

Given a sample Beacon database, we will ask participating team to develop solutions to mitigate the Bustamante a each algorithm based on the maximum number of correct queries that it can respond before any individual can be Bustamante attack. (data link)

Summary

EBI has data linkage use cases for both open and controlled access data

Likely that secure and summary data access mode will persist

Ethical issues - largely determine the shape of our implementation

Technical requirements – federation, automation of access control, AAI, scaling

Ethics and policy around secure cloud sharing of EGA data

Balancing the drive for automation against the need for human judgement

Acknowledgements

GWAS/BioSamples

EGA @ EBI and CRG

GWAS – EMBL Core funds, NHGRI Grant Number 1U41HG007823-01

BioSamples – EMBL Core Funds**,** EBiSC- IMI agreement n° 115582, resources of which are composed of financial contribution from the European Union’s Seventh Framework Programme (FP7/2007-2013) and EFPIA companies’ in kind contribution.
HipSci is funded by the Wellcome Trust and MRC

EGA –EMBL Core Funds, ELIXIR-EXCELERATE is funded by the European Commission within the Research Infrastructures programme
of Horizon 2020, grant agreement number 676559.
The RD-Connect Project is funded under the European Commission. Seventh Framework Programme, grant number 305444,
European Union’s Horizon 2020 research and innovation programme under grant agreement No 654248.
Accelerating Medicines Partnership