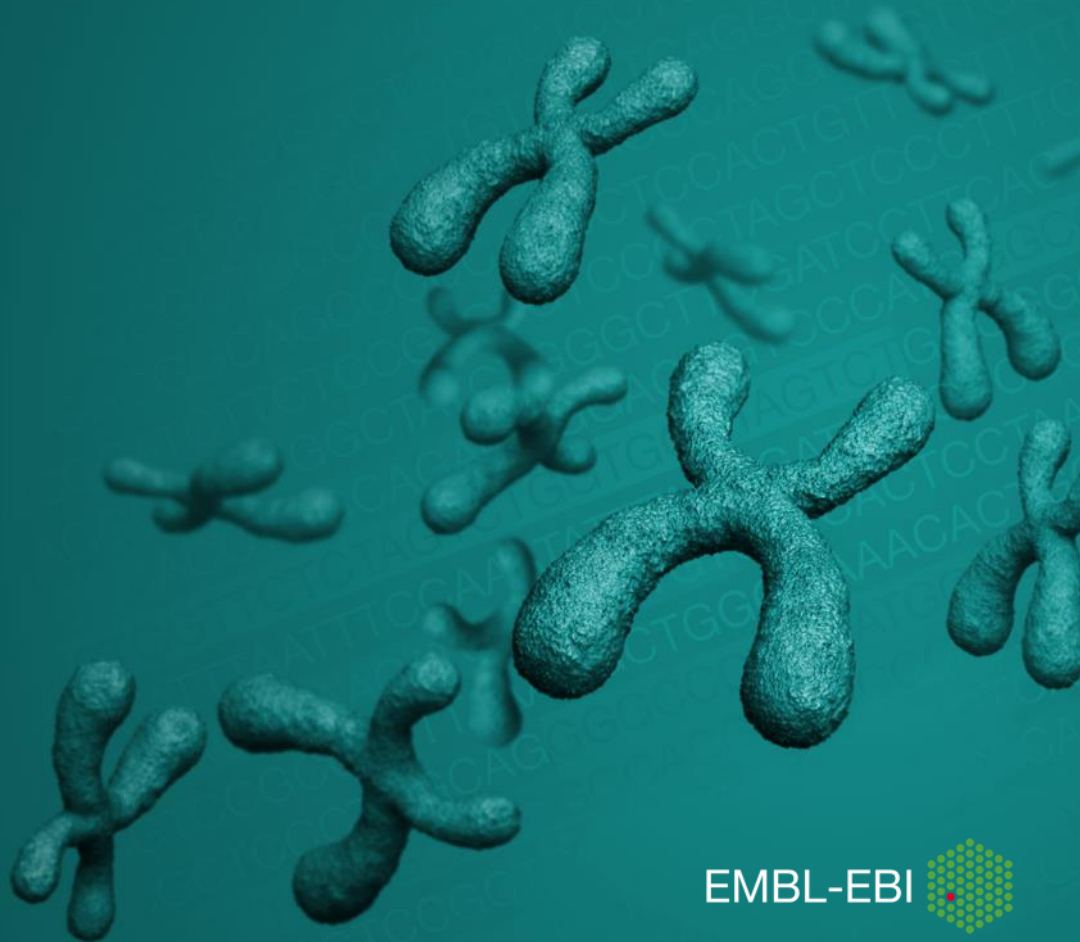


The EBI Variation Archives

Ilkka Lappalainen

Variation Archive Project Leader

www.ebi.ac.uk



EMBL

European Molecular Biology
Laboratory

Founded in 1974

International center on basic research
in molecular biology

Operates five sites across Europe
each focused on supporting a
particular science area.



The European Molecular Biology Laboratory

Heidelberg



Hamburg



Hinxton, Cambridge



Grenoble



Monterotondo, Rome



EMBL staff:

1700 people

>60

nationalities

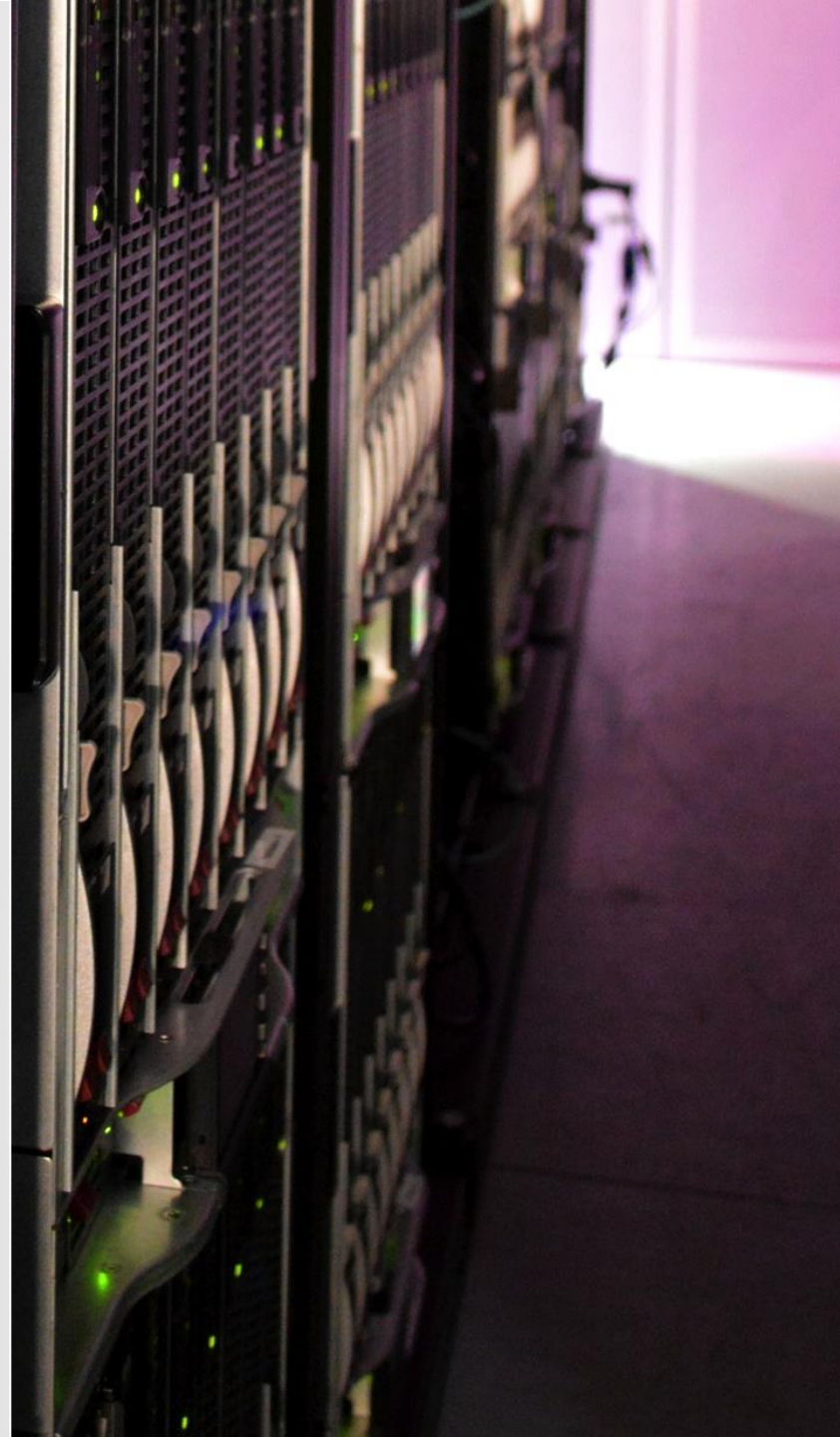
What is EMBL-EBI?

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



OUR MISSION

To provide freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress



OUR MISSION

To contribute to the advancement of biology through investigator-driven research in bioinformatics

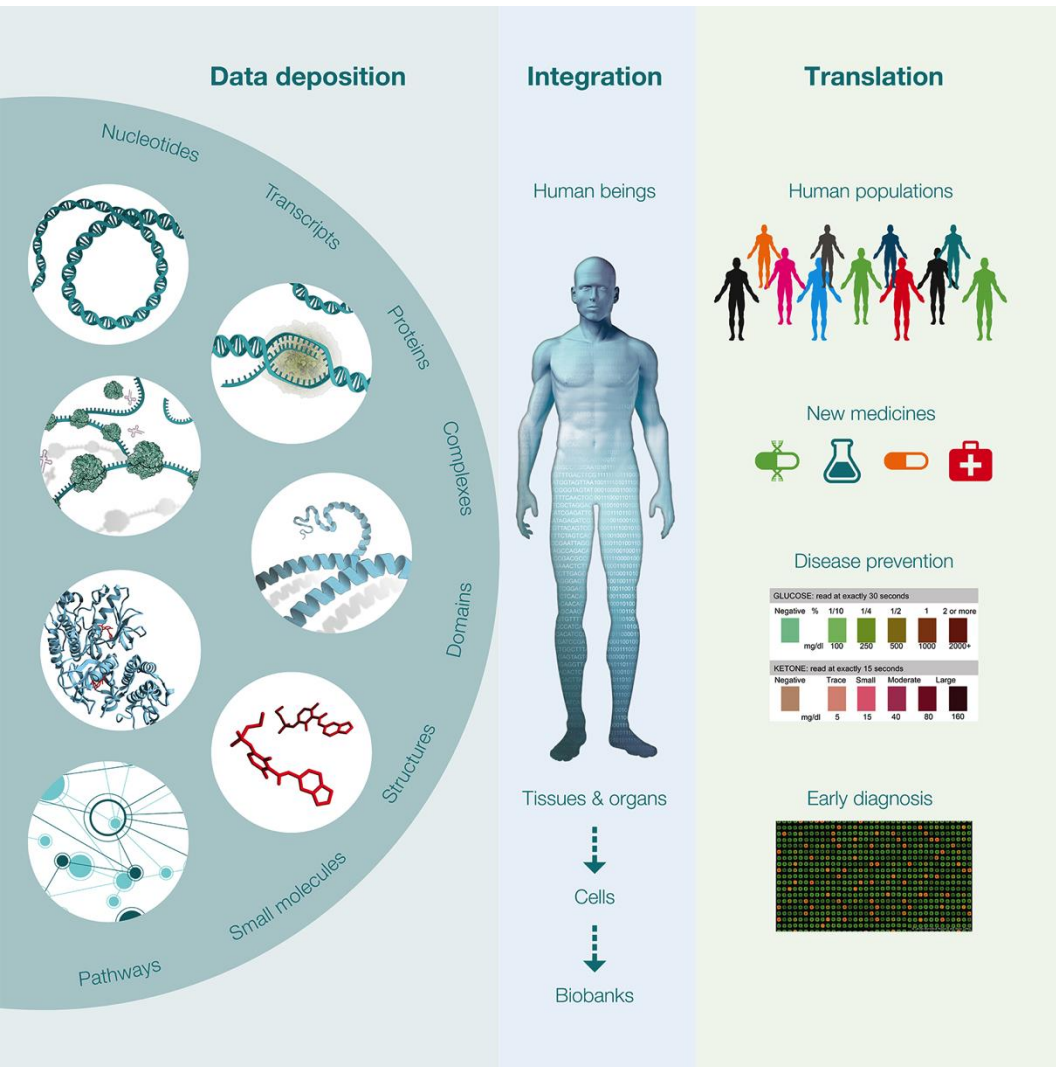


OUR MISSION

To provide advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators



Genetic Variation at the EBI



Biology is changing:

- Data explosion
- New types of data
- Emphasis on systems
- Applied biology:
 - molecular medicine
 - agriculture
 - food
 - environmental sciences.

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide
Archive

Ensembl
Ensembl Genomes

European Genome-phenome Archive
Metagenomics portal

1000 Genomes

Gene, protein & metabolite expression

ArrayExpress
Expression Atlas

Metabolights
PRIDE

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor
Ontology

Protein sequences, families & motifs

TrEMBL

Pfam

UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL

ChEBI

Reactions, interactions & pathways

IntAct

Reactome

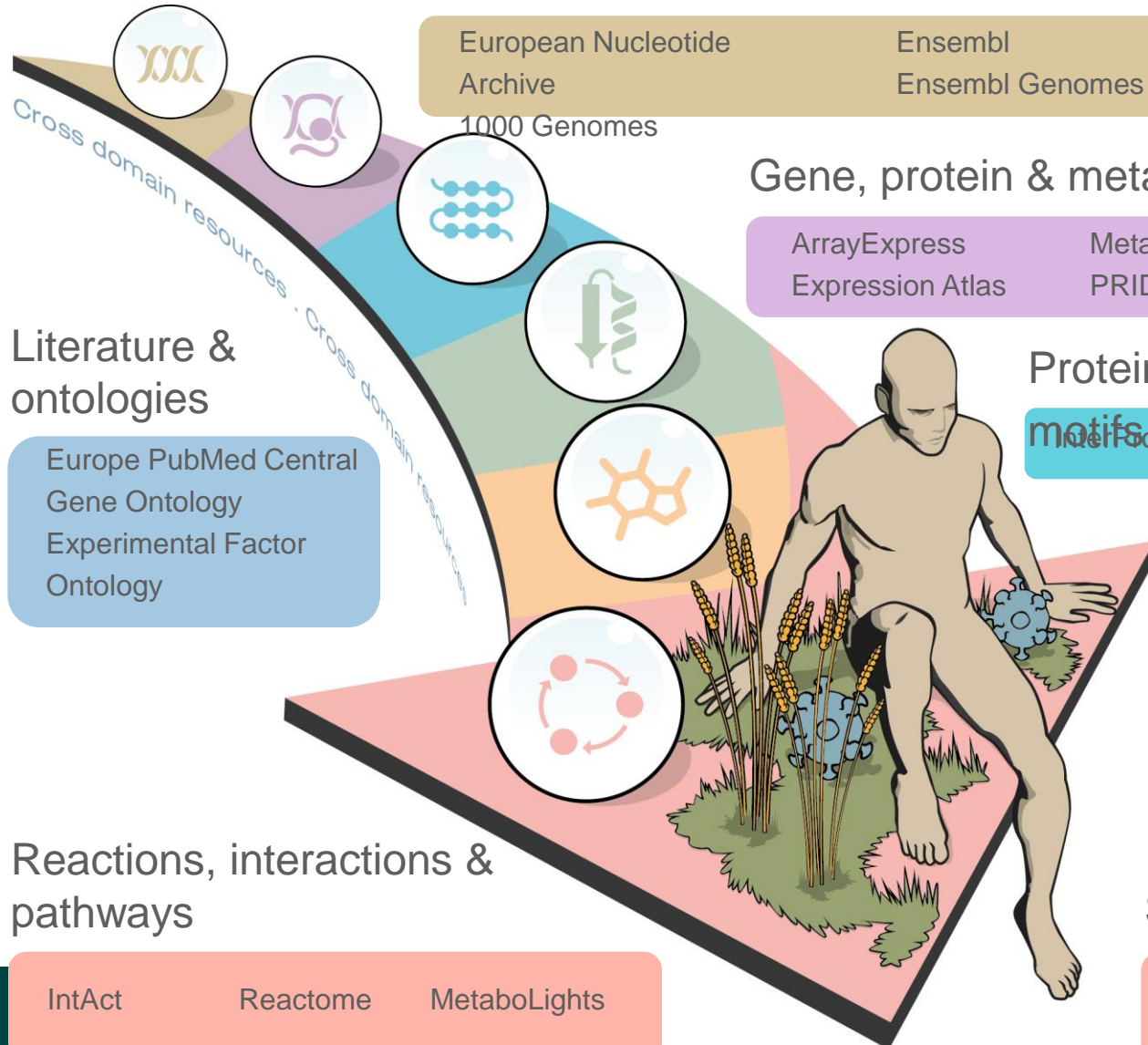
MetaboLights

Systems

BioModels

Enzyme Portal

BioSamples



EBI variation archives

European Genome-phenome Archive (EGA)

- <https://www.ebi.ac.uk/ega>
- Controlled access archive
- Accepts all experiment types from biomedical research projects



Database of Genomic Variants Archive (DGVA)

- <http://www.ebi.ac.uk/dgva>
- No controlled access mechanism – data are fully public.
- Accepts genetic structural variants from all species > 50 nt of length.



European Variation Archive (EVA)

- <http://www.ebi.ac.uk/eva>
- No controlled access mechanism – data are fully public.
- Accepts all types of variants from all species.

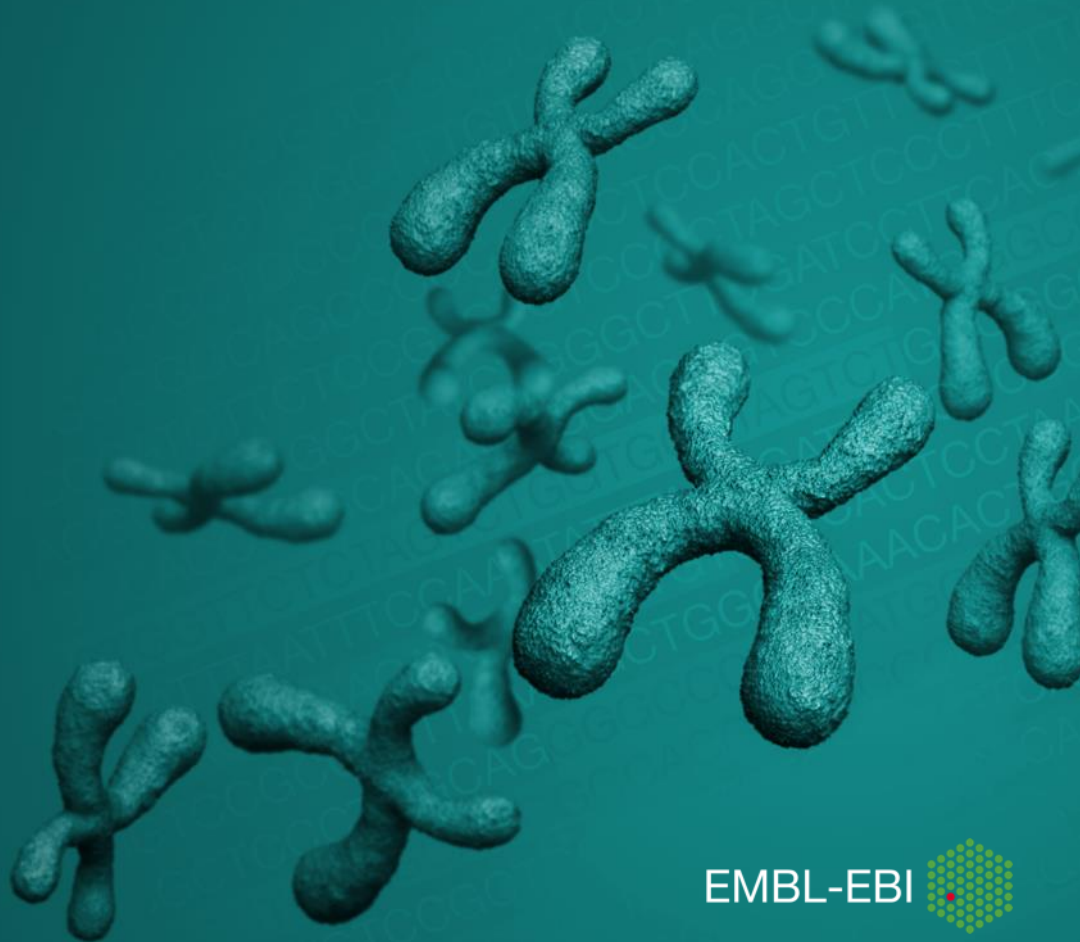


Topics today

This talk will focus on:

- What type of data each archive includes.
- How data can be retrieved from the archive.
- What are submission requirements
 - Meta data submission
 - File formats and file transfer to EBI

European Genome-phenome Archive (EGA)



EGA is provided by EBI and CRG

The EGA was created by the EBI in 2009.

In 2013, EBI and Center for Genome Regulation (CRG), Spain started working together to establish EGA as a joint venture.

<https://www.ebi.ac.uk/ega/>
<https://ega.crg.eu/>

ega-helpdesk@ebi.ac.uk



European Genome-phenome Archive (EGA)

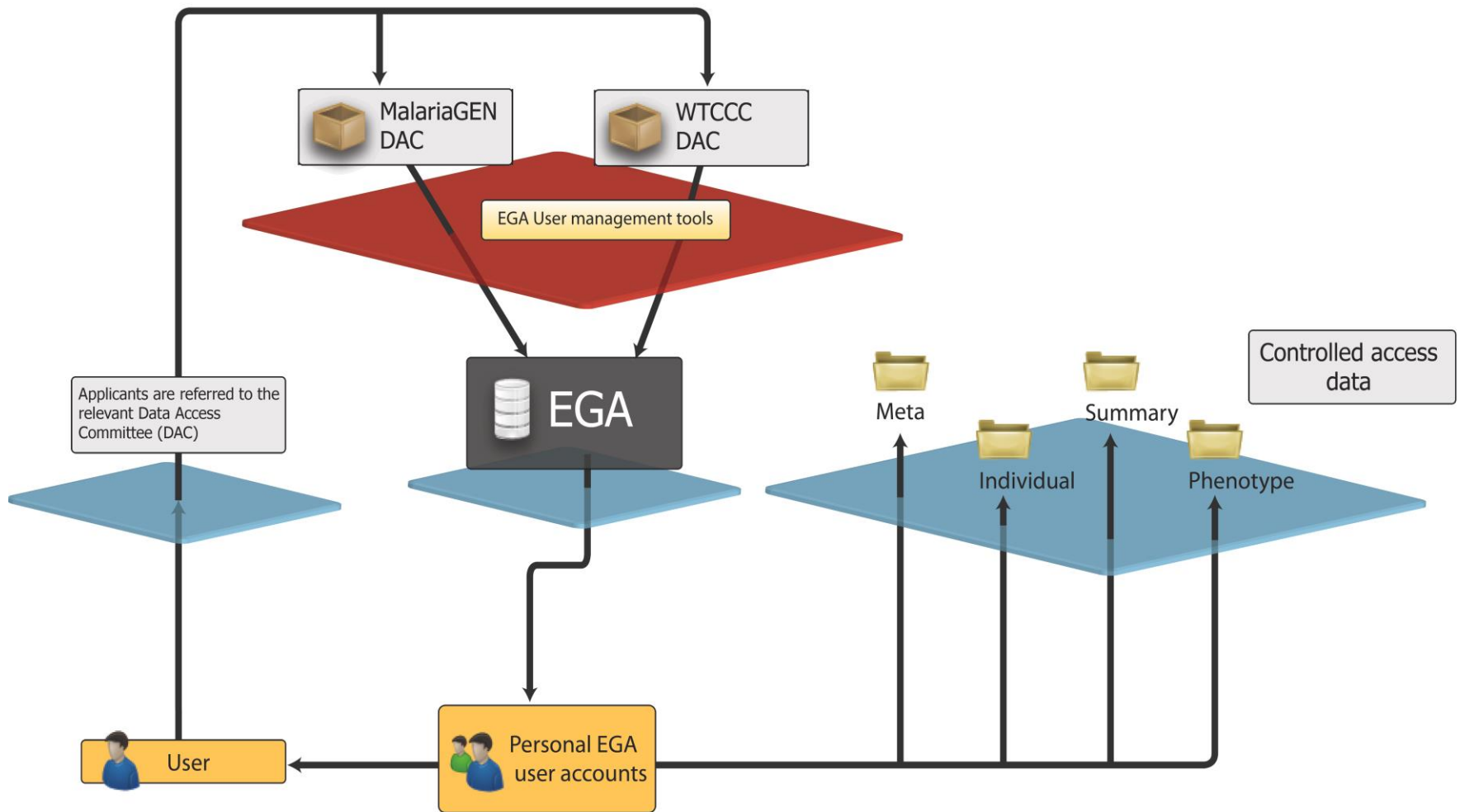
● **Primary archive for any data consented for sharing in the context of research but not for fully public distribution**

- Secure storage, management and dissemination of data – raw or processed - from biomedical research projects.
- Phenotypic data collected from the subjects.
- Submissions must be de-identified and in accordance with the informed consent.
- Data are packed into *datasets* that are governed by a Data Access Committee (DAC).
 - *Authentication* - each DAC approved individual will have a personal EGA account.
 - *Authorization* – DACs attach access permission(s) to the EGA account(s).

What does Controlled Access mean?

- Controlled access is not the same as holding data private in the archive until it is published. All EBI archives provide the later option.
- Controlled access mechanism can only be used if it is required by the informed consent.
- EGA provides tools for the Data Access Committees (DAC) to manage access to their data in our system. Once we receive authorization from the DAC it is our responsibility to make the data available for the user.

EGA works with Data Access Committees (DAC)



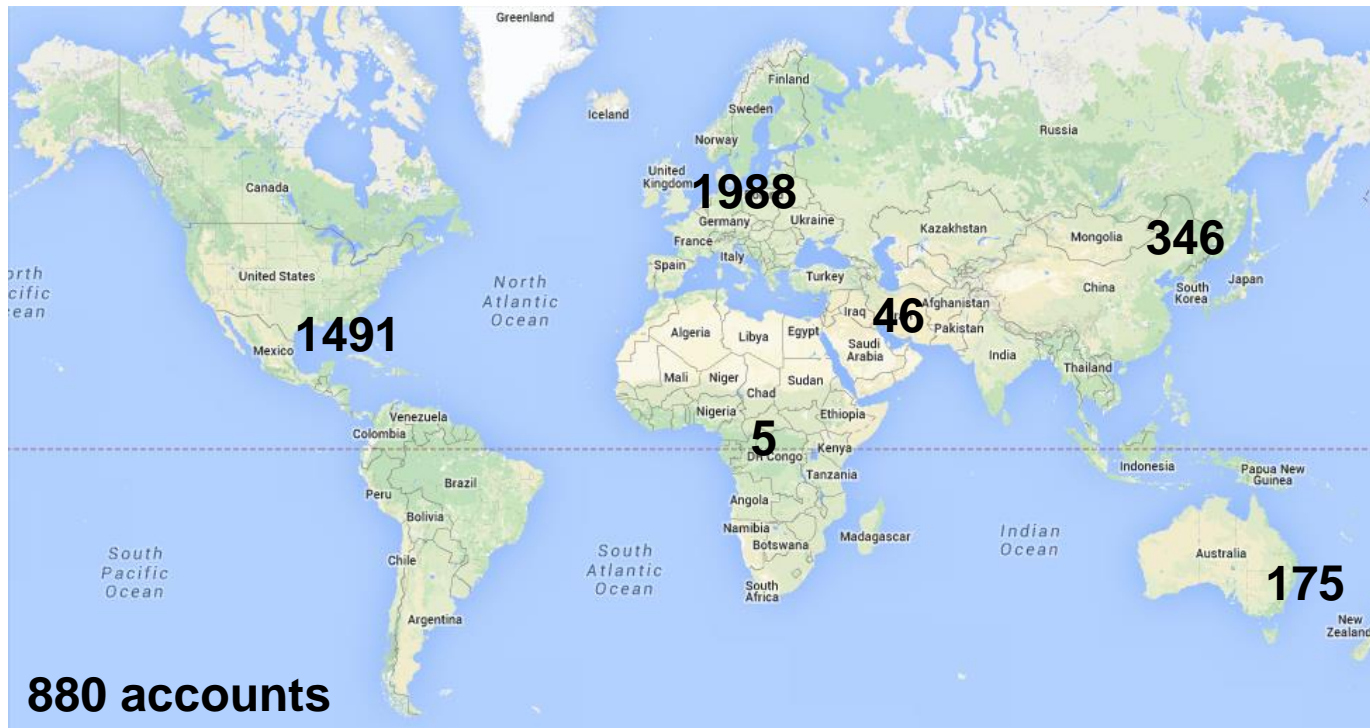
EGA is a global service

- Serving 150 institutes around the world.
- Includes projects such as International Cancer Genome Consortium, Wellcome Trust Case Control Consortium and the UK10K.



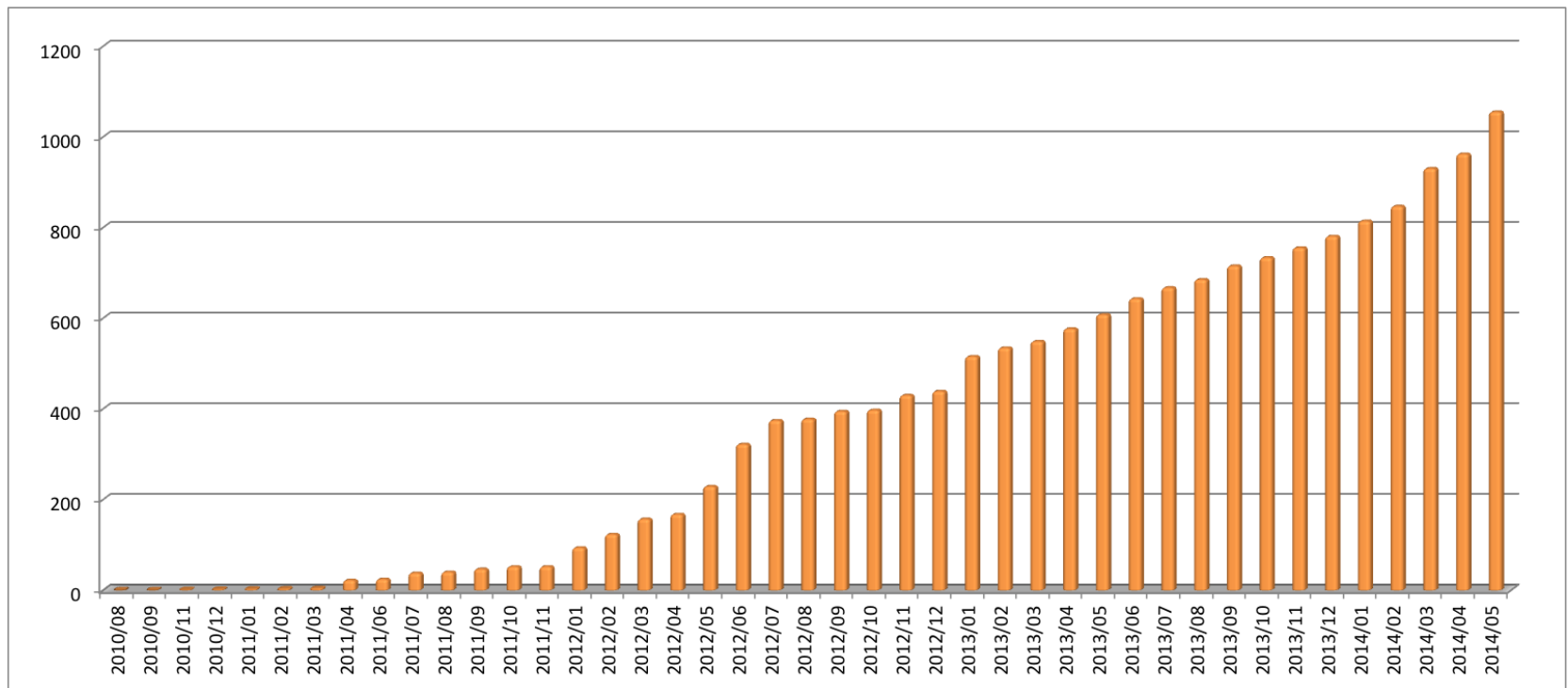
EGA is a global service

- We server more than 150 DACs with 5000 authorized users.
- Users make on average 250 contacts at our help-desk and 4700 data requests each month.



EGA is a global service

- More than 480 studies consisting of 830 datasets available at our website.



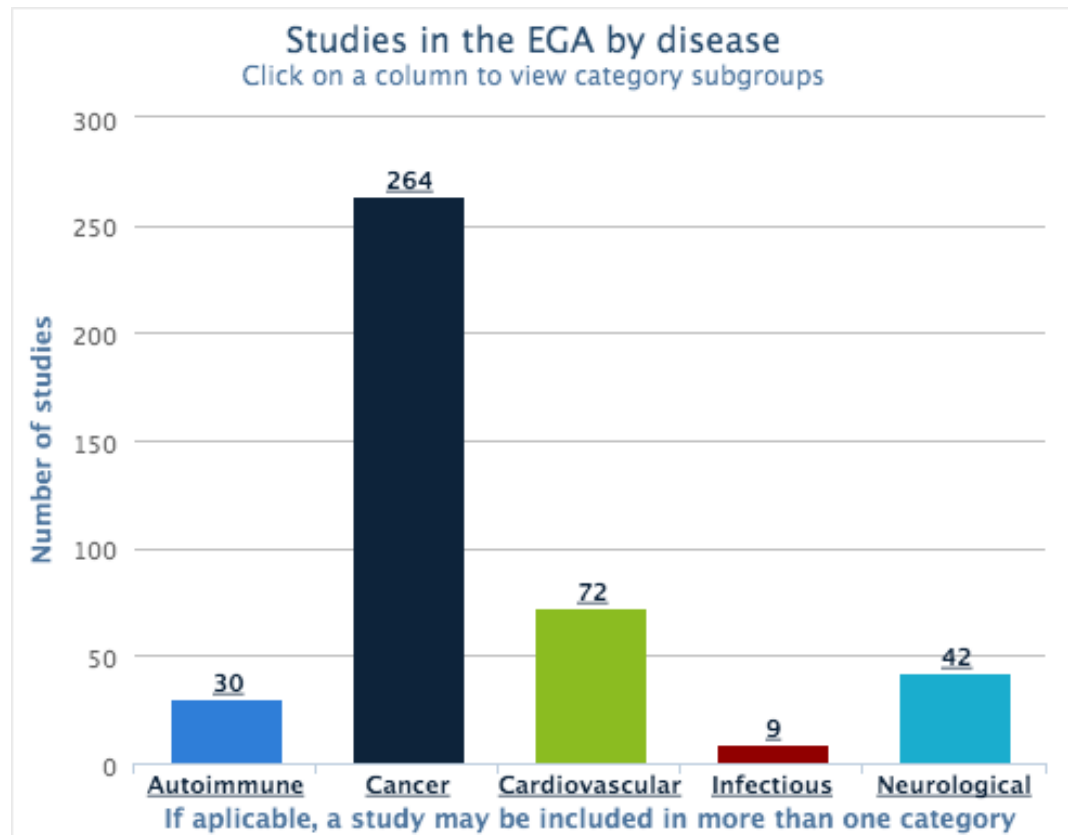
Archive growth in tera bytes of data

EGA is a global service

- More than 480 studies consisting of 830 datasets available at our website.
 - Most of the EGA data are raw data from the NGS experiments such as FASTQ, BAM or CRAM.
 - We also see now increase in VCF submissions that describe the genotypes for the studies samples.
 - We also hold a large amount of genotype data in PLINK and WTCCC formats for early array-based studies. In some cases we have the Illumina or Affymetrix raw data files as well.

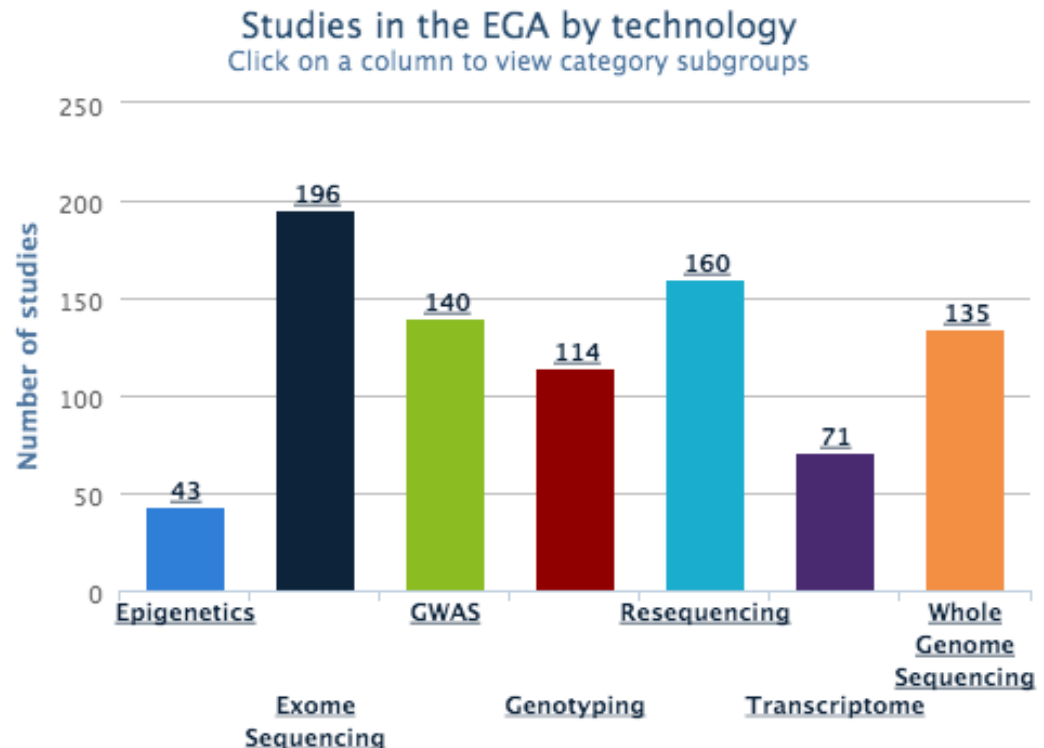
EGA is a global service

- More than 480 studies consisting of 830 datasets available at our website.



EGA is a global service

- More than 480 studies consisting of 830 datasets available at our website.



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

How can I access data stored in EGA

European Genome-phenome Archive

All

Examples: EGAS00000000001, cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Contact Us](#)

The European Genome-phenome Archive (EGA) allows you to explore **datasets** from genomic **studies**, provided by a range of **data providers**. Access to datasets must be approved by the specified **Data Access Committee (DAC)**.

Help

- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)

Studies

Studies are experimental investigations of a particular phenomenon or trait.

[Browse all studies](#)

Learn about the EGA

- [Introduction to the EGA](#)
- [How to obtain an account with the EGA](#)
- [Using your EGA account](#)

Datasets

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).

[Browse all datasets](#)


[Browse all control datasets](#)

Data Access Committees

Providers may be involved in study creation, submission and designation of Data Access Committees (DACs).

Navigation

- [Login](#)
- [Request new password](#)



<https://www.ebi.ac.uk/ega>

How can I access data stored in EGA

European Genome-phenome Archive

All [dropdown arrow]

Search

Examples: EGAS00000000001, cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Contact Us](#)

The European Genome-phenome Archive (EGA) allows you to explore **datasets** from genomic **studies**, provided by a range of **data providers**. Access to datasets must be approved by the specified **Data Access Committee (DAC)**.

Studies

Studies are experimental investigations of a particular phenomenon or trait.

[Browse all studies](#)

Learn about the EGA

- [Introduction to the EGA](#)
- [How to obtain an account with the EGA](#)
- [Using your EGA account](#)

Datasets

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).

[Browse all datasets](#)

[Browse all control datasets](#)

Data Access Committees


Providers may be involved in study creation, submission and designation of Data Access Committees (DACs).

Help

- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)

Navigation

- [Login](#)
- [Request new password](#)



<https://www.ebi.ac.uk/ega>

How can I access data stored in EGA

Pre-filter
your search
results...

All

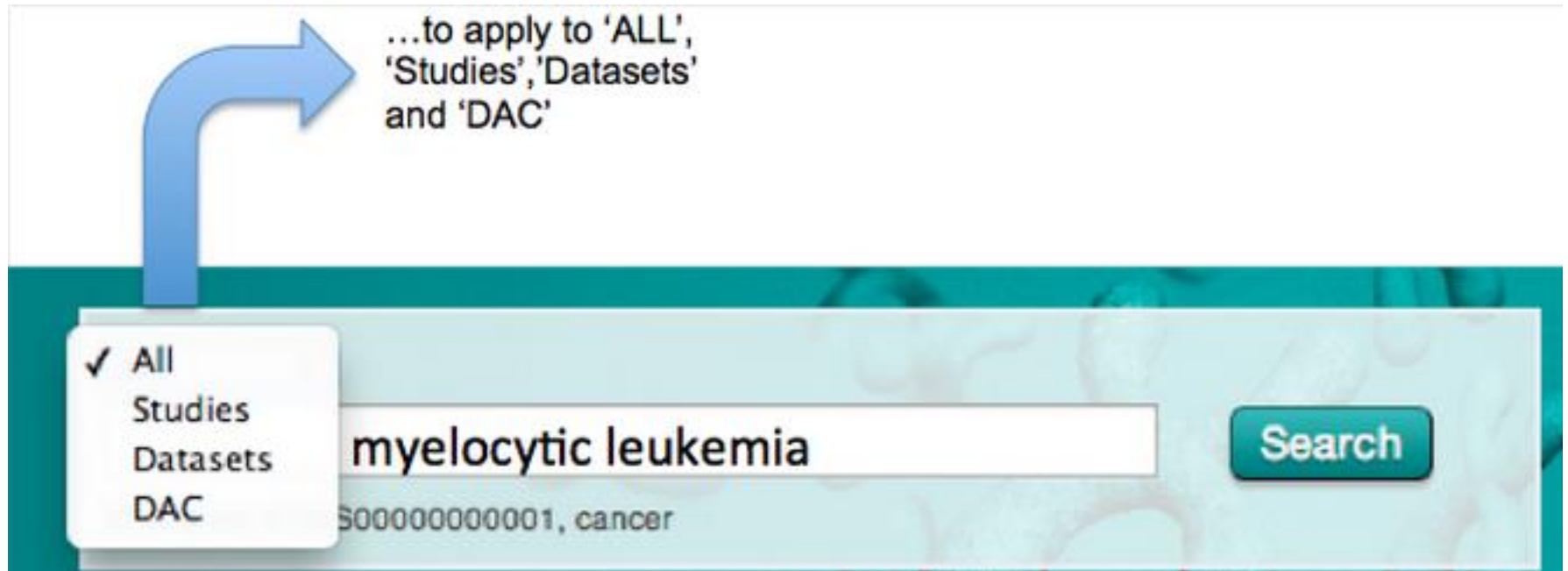
acute promyelocytic leukemia

Search

Examples: EGAS00000000001, cancer

<https://www.ebi.ac.uk/ega>

How can I access data stored in EGA



...to apply to 'ALL',
'Studies', 'Datasets'
and 'DAC'

✓ All
Studies
Datasets
DAC

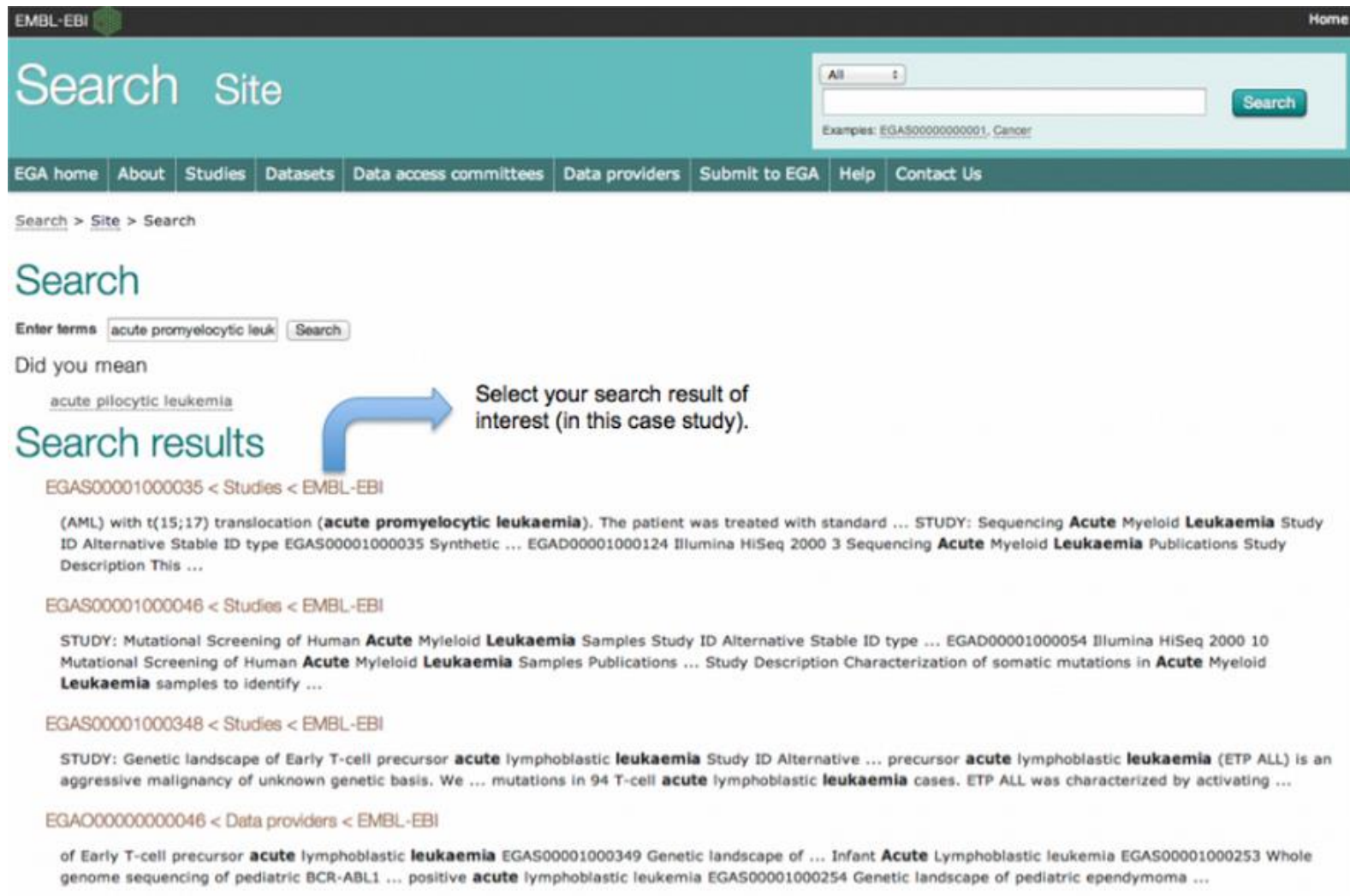
myelocytic leukemia

S000000000001, cancer

Search

<https://www.ebi.ac.uk/ega>

How can I access data stored in EGA



EMBL-EBI Home

Search Site

All

Examples: EGAS00000000001, Cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Help](#) [Contact Us](#)


[Search](#) > [Site](#) > [Search](#)

Search

Enter terms

Did you mean [acute pilocytic leukemia](#)

Search results

 Select your search result of interest (in this case study).

[EGAS00001000035](#) < [Studies](#) < [EMBL-EBI](#)

(AML) with t(15;17) translocation (**acute promyelocytic leukaemia**). The patient was treated with standard ... STUDY: Sequencing **Acute** Myeloid **Leukaemia** Study ID Alternative Stable ID type EGAS00001000035 Synthetic ... EGAD00001000124 Illumina HiSeq 2000 3 Sequencing **Acute** Myeloid **Leukaemia** Publications Study Description This ...

[EGAS00001000046](#) < [Studies](#) < [EMBL-EBI](#)

STUDY: Mutational Screening of Human **Acute** Myeloid **Leukaemia** Samples Study ID Alternative Stable ID type ... EGAD00001000054 Illumina HiSeq 2000 10 Mutational Screening of Human **Acute** Myeloid **Leukaemia** Samples Publications ... Study Description Characterization of somatic mutations in **Acute** Myeloid **Leukaemia** samples to identify ...

[EGAS00001000348](#) < [Studies](#) < [EMBL-EBI](#)

STUDY: Genetic landscape of Early T-cell precursor **acute** lymphoblastic **leukaemia** Study ID Alternative ... precursor **acute** lymphoblastic **leukaemia** (ETP ALL) is an aggressive malignancy of unknown genetic basis. We ... mutations in 94 T-cell **acute** lymphoblastic **leukaemia** cases. ETP ALL was characterized by activating ...

[EGAO00000000046](#) < [Data providers](#) < [EMBL-EBI](#)

of Early T-cell precursor **acute** lymphoblastic **leukaemia** EGAS00001000349 Genetic landscape of ... Infant **Acute** Lymphoblastic leukemia EGAS00001000253 Whole genome sequencing of pediatric BCR-ABL1 ... positive **acute** lymphoblastic leukemia EGAS00001000254 Genetic landscape of pediatric ependymoma ...

How can I access data stored in EGA

European Genome-phenome Archive

All Search
Examples: EGAS00000000001, cancer

EGA home About Studies Datasets Data access committees Data providers Submit to EGA Contact Us

STUDY: Sequencing Acute Myeloid Leukaemia

Study Description


This project will aim at sequencing and analysing three samples from a patient with acute myeloid leukaemia (AML) with t(15;17) translocation (... [Show More](#))

Study ID	Alternative Stable ID	type
EGAS00001000035		Synthetic Genomics

Data provider(s)

- Wellcome Trust Sanger Institute
- Wellcome Trust Sanger Institute Cancer Genome Project

Who archives the data?



This study includes 1 datasets:

Click on a Dataset ID in the table below to learn more, and to find out who to contact about access to these data

Dataset ID	Technology	Type	Samples	Description
EGAD00001000124	Illumina HiSeq 2000		3	Sequencing Acute Myeloid Leukaemia

Publications

How can I access data stored in EGA

European Genome-phenome Archive

All Search

Examples: EGAS00000000001, cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Contact Us](#)

DATASET: Sequencing Acute Myeloid Leukaemia

Dataset ID	Technology	Samples
EGAD00001000124	Illumina HiSeq 2000	3

No access to download

Please log in before attempting to download data from the EGA. If you do not have an EGA account and want to request access, contact information for the DAC responsible for access to this data is on the right under the heading 'Who controls access to this dataset'.

This dataset is featured in 1 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
EGAS00001000035	Sequencing Acute Myeloid Leukaemia

Who controls access to this dataset

For each dataset that requires access control, there is a corresponding Data Access Committee (DAC) who determine access permissions. Data access requests are reviewed by the relevant DAC, not by the EGA. If you need to request access to this data set, please contact:

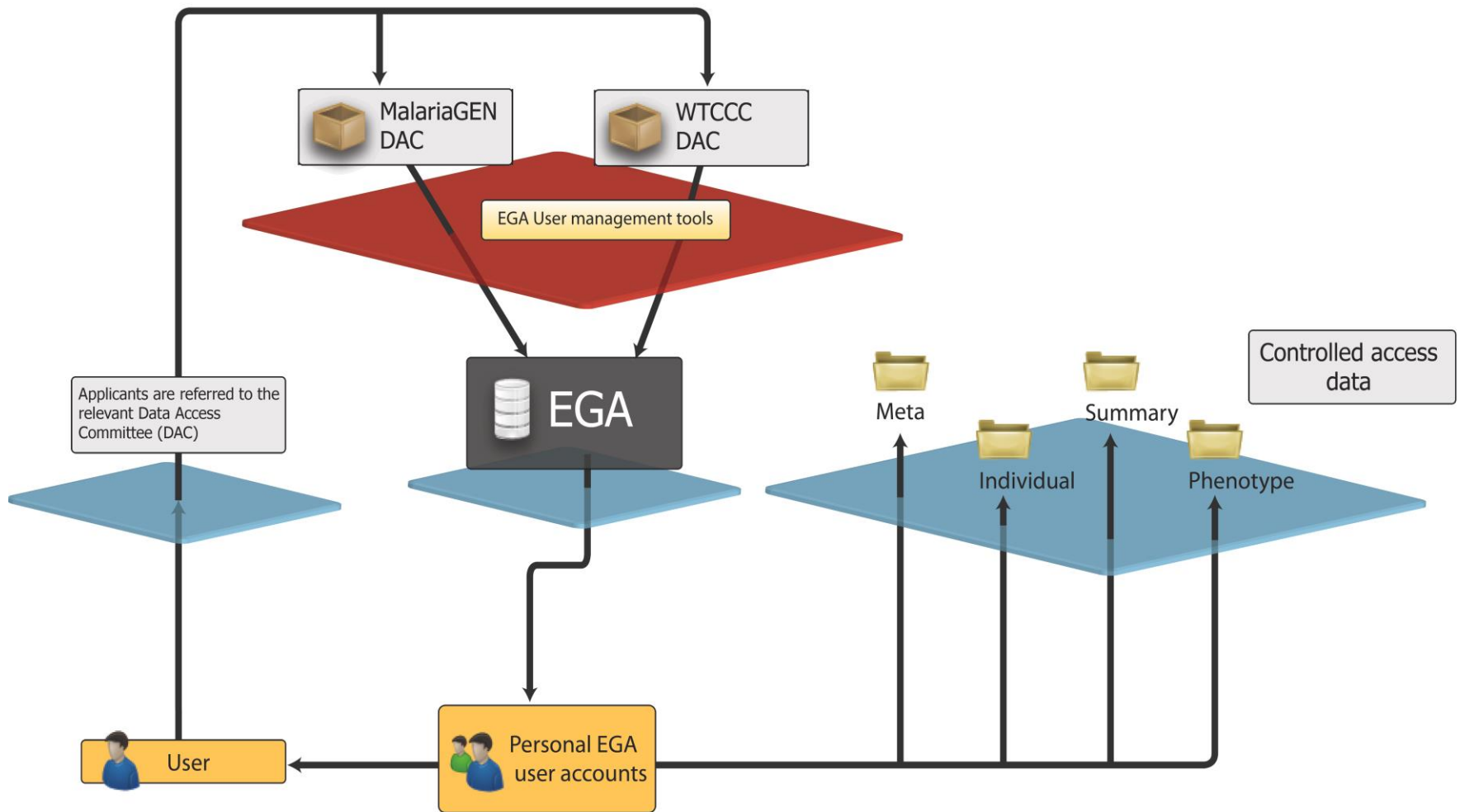
WTSI CGP Data access committee

Access information:
<http://www.ebi.ac.uk/ega/dacs/EGAC00001000000>

Contact Person: Giselle Kerry

Email: gh2@sanger.ac.uk

Data Access applied directly from DAC



How can I access data stored in EGA

European Genome-phenome Archive

All

Examples: EGAS00000000001, cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Contact Us](#)

The European Genome-phenome Archive (EGA) allows you to explore **datasets** from genomic **studies**, provided by a range of **data providers**. Access to datasets must be approved by the specified **Data Access Committee (DAC)**.

Help

- [Users FAQ](#)
- [Submitters FAQ](#)
- [Using your EGA account](#)
- [Contact Us](#)
- [EGA mailing list](#)

Studies

Studies are experimental investigations of a particular phenomenon or trait.

[Browse all studies](#)

Learn about the EGA

- [Introduction to the EGA](#)
- [How to obtain an account with the EGA](#)
- [Using your EGA account](#)

Datasets

The EGA archives a large number of datasets, the access to which is controlled by a Data Access Committee (DAC).

[Browse all datasets](#)


[Browse all control datasets](#)

Data Access Committees

Providers may be involved in study creation, submission and designation of Data Access Committees (DACs).

Navigation

- [Login](#)
- [Request new password](#)



<https://www.ebi.ac.uk/ega>

How can I access data stored in EGA

European Genome-phenome Archive

All

Examples: EGAS00000000001, cancer

[EGA home](#) [About](#) [Studies](#) [Datasets](#) [Data access committees](#) [Data providers](#) [Submit to EGA](#) [Contact Us](#)

Login

Login with Local Account

Username *
Enter your European Genome-phenome Archive username.

Password *
Enter the password that accompanies your username.

Login with Federated Identity

Use a suggested selection:

Certia Oy

Or enter your organization's name

[Allow me to pick from a list](#) [Help](#)

How can I access data stored in EGA

European Genome-phenome Archive

EGA home | About | Studies | Datasets | Data access committees | Data providers | Submit to EGA | Contact Us

DATASET: Sequencing Acute Myeloid Leukaemia

Datasets description

Dataset ID	Technology	Samples
EGAD0000100012

Who controls access to this dataset

For each dataset that requires access control, there is a corresponding Data Access Committee (DAC) who determine access permissions. Data access is not the responsibility of the EGA. If you need to request access to this data set, please contact:

WTSI CGP Data access committee

Access information:
<http://www.ebi.ac.uk/ega/dacs/EGAC0000100000>

Contact Person: Giselle Kerry

This dataset is featured in 1 studies

Studies are experimental cancer research projects listed below to find out more.

Study ID	Study Title
EGAS00001000035	Sequencing Acute Myeloid Leukaemia

Label:

Enter a label (maximum 25 characters) to help you recognise this request easily. If you do not enter a label, one will be created automatically.

Select packets to download

Packet name	Technology	Type/Sample	Format	Released	Updated
<input checked="" type="checkbox"/> EGAN00001001822	Illumina HiSeq 2000	Sequence		2013-12-13 03:10:35	2013-12-13 03:10:35
<input checked="" type="checkbox"/> EGAN00001001824	Illumina HiSeq 2000	Sequence		2013-12-13 03:10:35	2013-12-13 03:10:35
<input checked="" type="checkbox"/> EGAN00001001823	Illumina HiSeq 2000	Sequence		2013-12-13 03:10:35	2013-12-13 03:10:35

Click on 'Request selected packages'.

How can I access data stored in EGA

European Genome-phenome Archive

EGA home | About | Studies | Datasets | Data access committees | Data providers | Submit to EGA | Contact Us

DATASET: Sequencing Acute Myeloid Leukaemia

What happens now?

Your download requests are now ready to download by using the Secure EGA download client.

The progress of your download requests can be monitored using your EGA open requests status page.

Contact the ega-helpdesk if you have any issues regarding data download.

Your download requests are then passed to the Secure EGA download client ready to be downloaded.

This dataset is featured in 1 studies

Contact Person: Giselle Kerry

Click on 'Request selected packages'.

Study ID	Study Title
EGAS00001000035	Sequencing Acute Myeloid Leukaemia

Secure EGA Downloader

Version: 0.1.37
[Contact the EGA Helpdesk](#)

Download / Slice Download File Slice Decrypt / Index

EGA User Name: jeff@ebi.ac.uk EGA User Password: ***** Logged in. Logout Download Protocol: Standard FTP

Local Encryption Key: ***** Verify Key: ***** Slice to Download: Invalid Region.

Local Download Directory: /Users/jeff/Desktop

Remote Files

Download	Name	Size	Date	Progress
<input checked="" type="checkbox"/>	172TR.CEL.gpg	4.31 MB	08-Oct-2013	Download Complete
<input checked="" type="checkbox"/>	110TR.CEL.gpg	4.41 MB	08-Oct-2013	Download Complete
<input checked="" type="checkbox"/>	761TR.CEL.gpg	4.40 MB	08-Oct-2013	Download Complete
<input checked="" type="checkbox"/>	680TR.CEL.gpg	4.45 MB	08-Oct-2013	Download Complete
<input checked="" type="checkbox"/>	785TR.CEL.gpg	4.28 MB	08-Oct-2013	87%
<input type="checkbox"/>	876TR.CEL.gpg	4.32 MB	08-Oct-2013	0%
<input type="checkbox"/>	029TR.CEL.gpg	4.43 MB	08-Oct-2013	0%
<input type="checkbox"/>	175TR.CEL.gpg	4.50 MB	08-Oct-2013	0%
<input type="checkbox"/>	016TR.CEL.gpg	4.52 MB	08-Oct-2013	0%
<input type="checkbox"/>	166TR.CEL.gpg	4.41 MB	08-Oct-2013	0%
<input type="checkbox"/>	186-01-8TR.CEL.gpg	4.32 MB	08-Oct-2013	0%
<input type="checkbox"/>	100TR.CEL.gpg	4.43 MB	08-Oct-2013	0%
<input type="checkbox"/>	032TR.CEL.gpg	4.36 MB	08-Oct-2013	0%
<input type="checkbox"/>	019-01-1TR.CEL.gpg	4.66 MB	08-Oct-2013	0%
<input type="checkbox"/>	009TR.CEL.gpg	4.14 MB	08-Oct-2013	0%
<input type="checkbox"/>	040TR.CEL.gpg	4.42 MB	08-Oct-2013	0%
<input type="checkbox"/>	184-01-5TR.CEL.gpg	4.32 MB	08-Oct-2013	0%
<input type="checkbox"/>	039TR.CEL.gpg	4.31 MB	08-Oct-2013	0%
<input type="checkbox"/>	022TR.CEL.gpg	4.39 MB	08-Oct-2013	0%
<input type="checkbox"/>	012TR.CEL.gpg	4.37 MB	08-Oct-2013	0%
<input type="checkbox"/>	290TR.CEL.gpg	4.38 MB	08-Oct-2013	0%
<input type="checkbox"/>	157TR.CEL.gpg	4.44 MB	08-Oct-2013	0%
<input type="checkbox"/>	030TR.CEL.gpg	4.35 MB	08-Oct-2013	0%

Select All 4 files selected. Total size: 17.40 MB Display Files: Request Dataset Filter

Cancel Download

Status: Downloading file 5 of 5 File: 1837 KB/s

Log in/select transfer protocol



Create key



Select destination directory



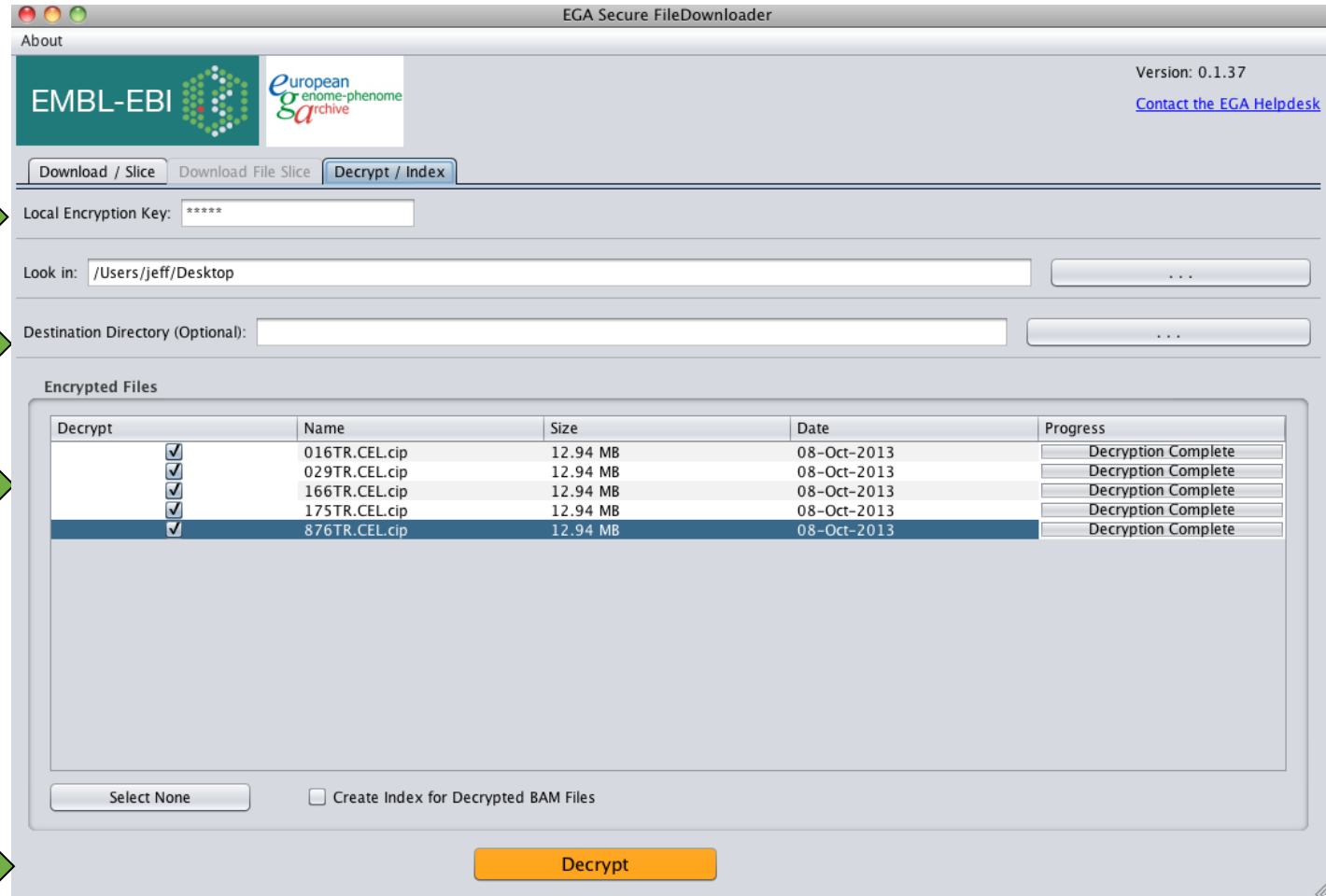
Filter and select files



Download!



Secure EGA Downloader



https://www.ebi.ac.uk/ega/about/your_EGA_account/secure_EGA_download_client

Secure EGA Downloader

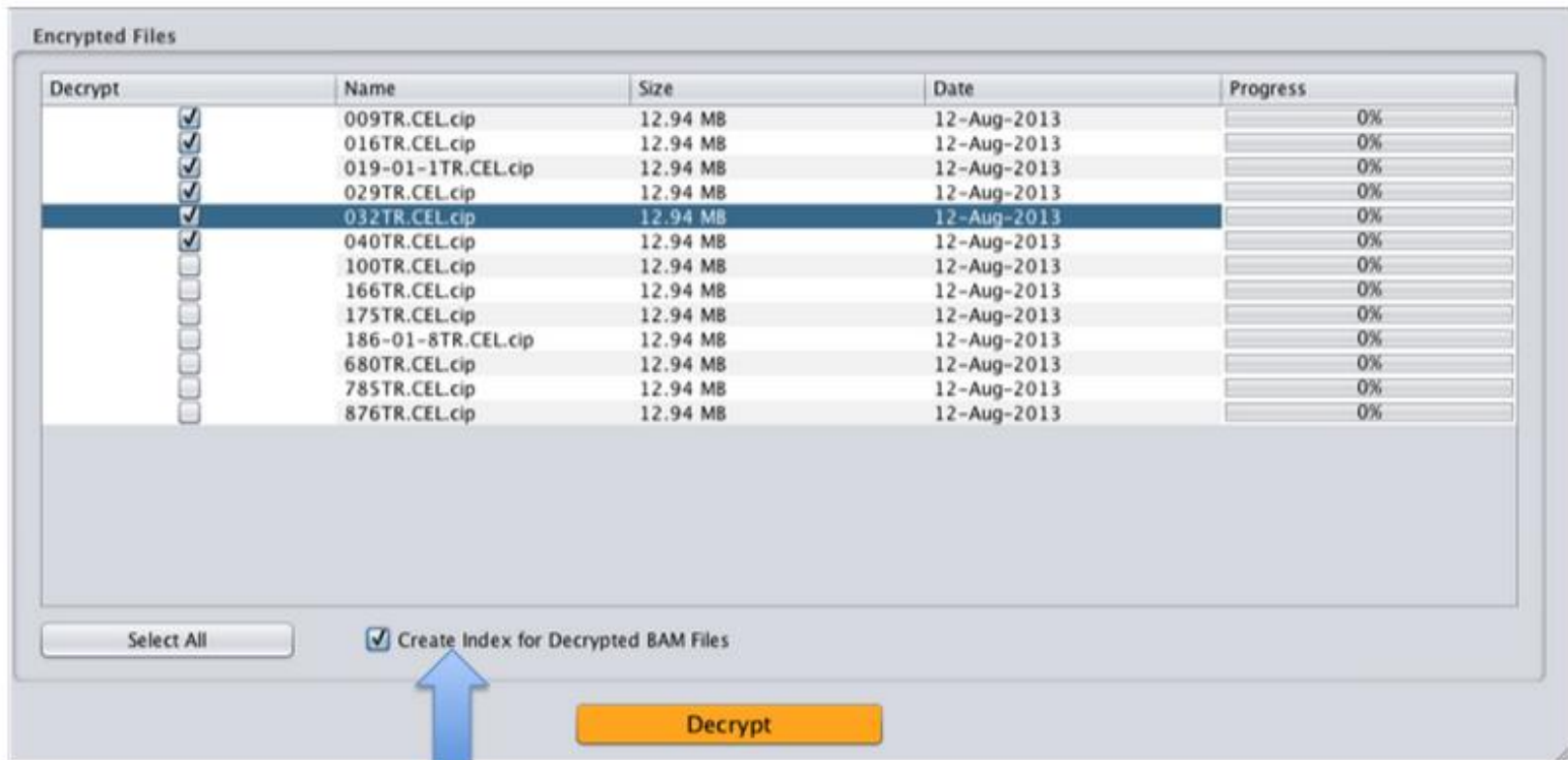


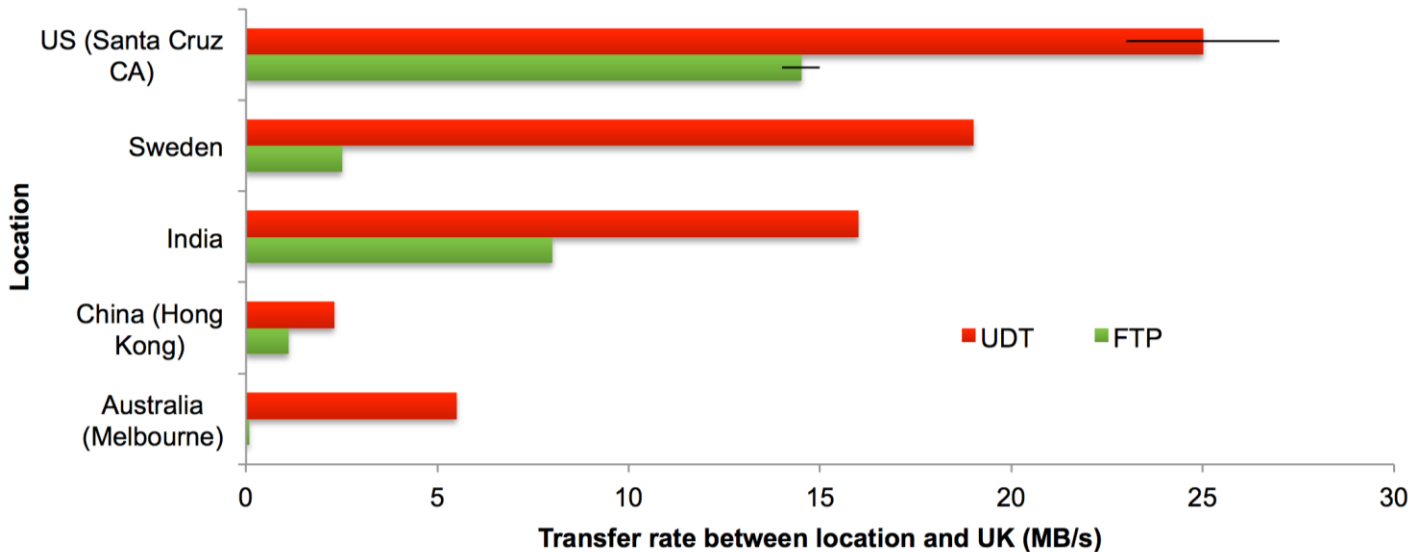
Fig 4 Initiating decryption

Generate indexes
for each decrypted
bam file

https://www.ebi.ac.uk/ega/about/your_EGA_account/secure_EGA_download_client

Data transfer optimization

Webin-UDT



<https://github.com/enasequence/webin-data-streamer-UDT>

Data Submissions to EGA

● **Should I submit to EGA or use fully public data resources?**

- Defined by the informed consent
- It is possible to use EGA and other archives at the EBI?
- Approval documentation for a submission
- Establishing Data Access Committee (DAC) or authorizing data access approval process for an existing DAC

● **Submission to EGA consists of two actions:**

- File upload - supported file formats
- Meta data submission

Data Submissions to EGA

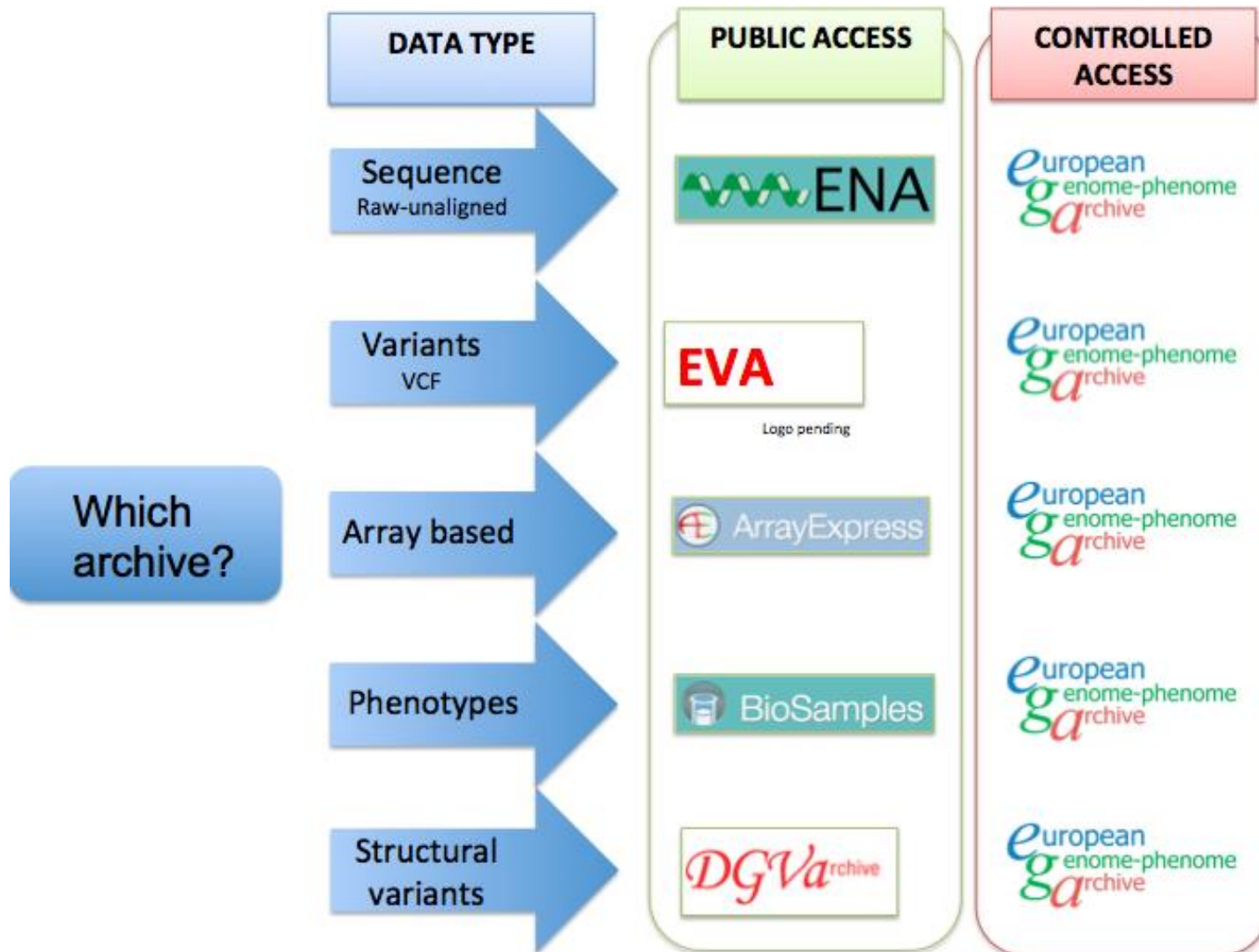
● **Should I submit to EGA or use fully public data resources?**

- Defined by the informed consent
- It is possible to use EGA and other archives at the EBI?
- Approval documentation for a submission
- Establishing Data Access Committee (DAC) or authorizing data access approval process for an existing DAC

● **Submission to EGA consists of two actions:**

- File upload - supported file formats
- Meta data submission

Using EBI archives



The Nasopharyngeal Microbiome and respiratory disease in African Children

● ***“A study of microorganisms in the upper respiratory track in children”***

- Study includes raw sequencing data of various bacteria and virus genomes (and potentially their variants).
- Each child may be sampled a number of time. For each sample time we need to be able to link discovered microorganisms to the correct individual.
- There may be phenotypic attributes of these children that cannot be fully published.

<http://www.h3africa.org/consortium/projects/16-projects/76-the-nasopharyngeal-microbiome-and-respiratory-disease-in-african-children>

The Nasopharyngeal Microbiome and respiratory disease in African Children

● ***“A study of microorganisms in the upper respiratory track in children.”***

- Study includes raw sequencing data of various bacteria and virus genomes (and potentially their variants). **ENA**
- Each child may be sampled a number of time. For each sample time we need to be able to link discovered microorganisms to the correct individual. **EGA**
- There may be phenotypic attributes of these children that cannot be fully published. **EGA**

<http://www.h3africa.org/consortium/projects/16-projects/76-the-nasopharyngeal-microbiome-and-respiratory-disease-in-african-children>

African Collaborative Center for Microbiome and Genomics Research (ACCME)

● ***“Focuses on understanding the associations between high risk HPV infection, vaginal microenvironment, HPV genomics, germline and somatic mutations in the etiology of the cervical cancer”.***

- Deep phenotypic information, genome sequence data and germline mutations discovered from the studied women.
- Somatic mutations discovered from the cancer samples.
- Bacterial and virus genomes from screening vaginal microenvironment.

<http://www.h3africa.org/consortium/projects/16-projects/81-african-collaborative-center-for-microbiome-and-genomics-research-accme>

African Collaborative Center for Microbiome and Genomics Research (ACCME)

● ***“Focuses on understanding the associations between high risk HPV infection, vaginal microenvironment, HPV genomics, germline and somatic mutations in the etiology of the cervical cancer”.***

- Deep phenotypic information, genome sequence data and germline mutations discovered from the studied women. **EGA**
- Somatic mutations discovered from the cancer samples. **EVA**
- Bacterial and virus genomes from screening vaginal microenvironment. **ENA**

<http://www.h3africa.org/consortium/projects/16-projects/81-african-collaborative-center-for-microbiome-and-genomics-research-accme>

How to initiate submission to EGA?

Contact



Receive



Upload



Document



- Email to ega-helpdesk@ebi.ac.uk
- Establish a submission account with us.
- Download tools, guidelines and examples from <https://www.ebi.ac.uk/ega/submission>



Marc



Jeff

How to initiate submission to EGA?

Contact



Receive



Upload



Document



- Submission account is always assigned to an institute – it is not a personal account.
- More than one person can operate submission accounts – this will only impact project meta data management as all data files have been encrypted prior data upload.



Marc



Jeff

EGA submission statements

- ◆ Informed consent signed by project participants requires controlled-access mechanism for data dissemination.
- ◆ Submission is compliant to the local laws and regulations.
- ◆ Submitter is authorized to upload the data to the EGA on behalf of the project.

https://www.ebi.ac.uk/ega/submission/data_access_committee/policy_documentation

Example of the submission statement

To whom it may concern,

This document refers to the submission account, **<ega-box-xx>**, which will be used to submit data and metadata to the European Genome phenome Archive (EGA) for the purpose of controlled access for individuals approved by a Data Access Committee (DAC).

Please be advised that **<FULL NAME and INSTITUTIONAL EMAIL ADDRESS>** is authorised to upload data and metadata to the EGA for archiving and distribution as part of your submission process.

We can confirm that this submission is consistent with the informed consent of the participants of the study or has been granted ethical approval and is in accordance with the applicable laws and regulations.

We understand that should any information referenced in this document be subject to change, an updated Submission statements document should be provided to the EGA.

Sincerely,

<Representative of study, e.g. Principal Investigator>

Jeff Almeida-King 7/10/13 4:15 PM

Comment [2]: Provided by the EGA at the start of the submission process. The format should be: **ega-box-xx**.

Jeff Almeida-King 1/20/14 5:23 PM

Comment [3]: All individuals uploading data files and metadata **MUST** be named.

Jeff Almeida-King 6/4/10 10:16 AM

Comment [4]: Individual must have the authority to underwrite the statement. In most cases, the PI associated with the study is sufficient.

https://www.ebi.ac.uk/ega/submission/data_access_committee/policy_documentation

Application form and Data Access Agreement

- **Research title and short description**
- **Personal details of all applicants and relevant publication history**
- **Accept terms and conditions**
 - How data must be stored, transferred and what type of analysis are allowed
 - Publication policy
 - Intellectual property rights
 - What happens to the local copy of the data once the project is no longer active?
 - Analysing data from more than one dataset controlled by the DAC – preventing study participant identification.


Data Access Agreement

Data Access Agreement (DAA)


Please find below links to examples of Data Access Agreements (DAA) used by existing Data Access Committees (DACs).

The Data Access Agreement is a contract made between user and Data Access Committee. The agreement should be drafted by the DAC and includes, but is not limited to, details of data use, publication embargoes and storage.

Completion of a DAA by the applicant/s should form part of the application process to the DAC.

[Wellcome Trust Case Control Consortium DAA](#) 

[Wellcome Trust Sanger Institute Cancer Genome Project \(UK- Academic\)](#) 

[Wellcome Trust Sanger Institute Cancer Genome Project \(US - Corporate\)](#) 

https://www.ebi.ac.uk/ega/submission/data_access_committee/policy_documentation#DAA
http://www.uk10k.org/data_access.html

Examples of Data Access Application forms

Data access application form

Please find below links to examples of Data access application forms used by existing Data Access Committees (DACs).

The Data access form should be drafted by the DAC, for the purpose of capturing the necessary information from a user wishing to access data.

Completion of a Data access application form by the applicant/s should form part of the application process to the DAC.

[MalariaGen Data access form](#)

[Wellcome Trust Case Control Consortium Data access form](#)

https://www.ebi.ac.uk/ega/submission/data_access_committee/policy_documentation#DAAF

Data Submissions to EGA

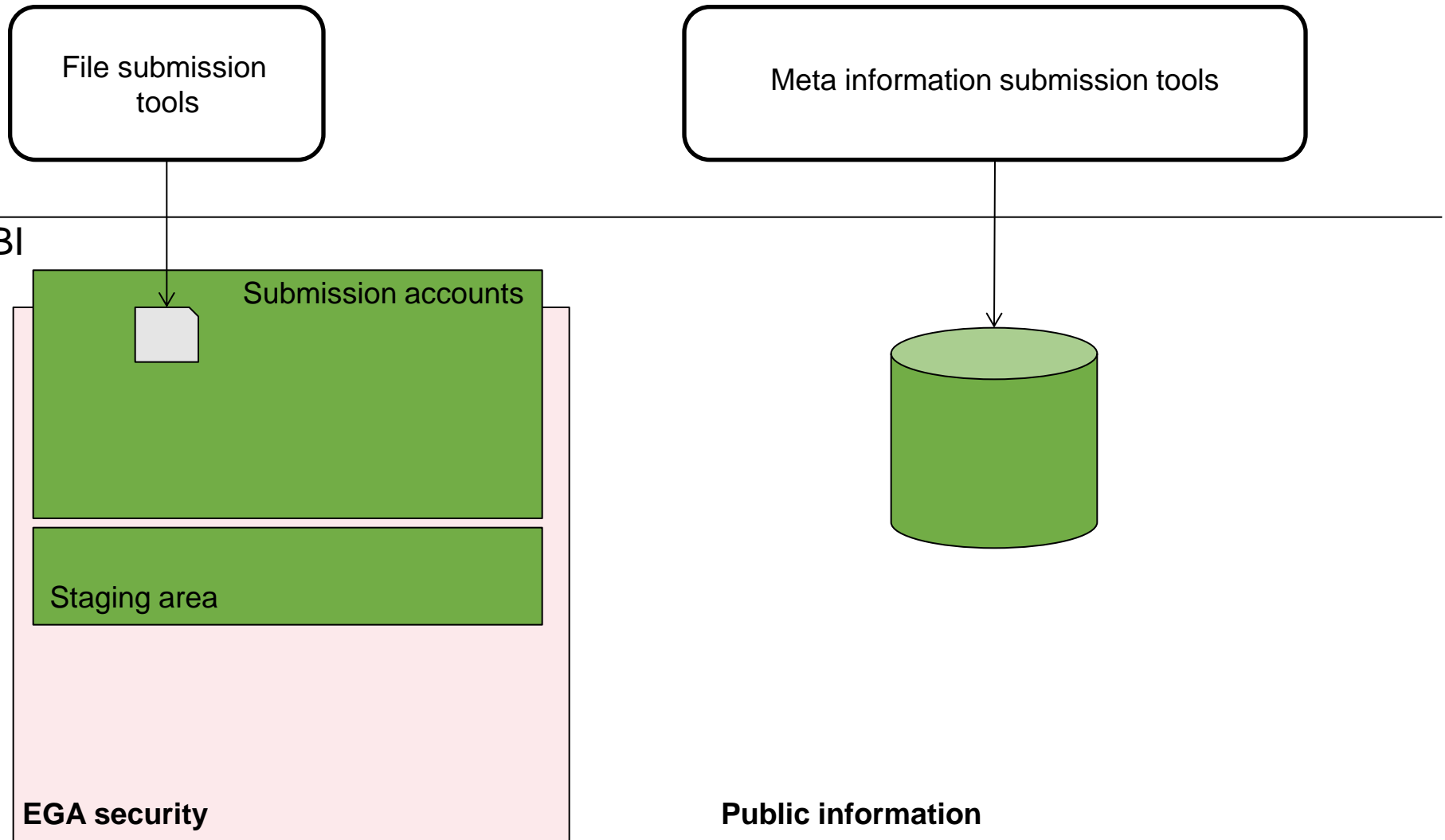
• Should I submit to EGA or use fully public data resources?

- Defined by the informed consent
- It is possible to use EGA and other archives at the EBI?
- Approval documentation for a submission
- Establishing Data Access Committee (DAC) or authorizing data access approval process for an existing DAC

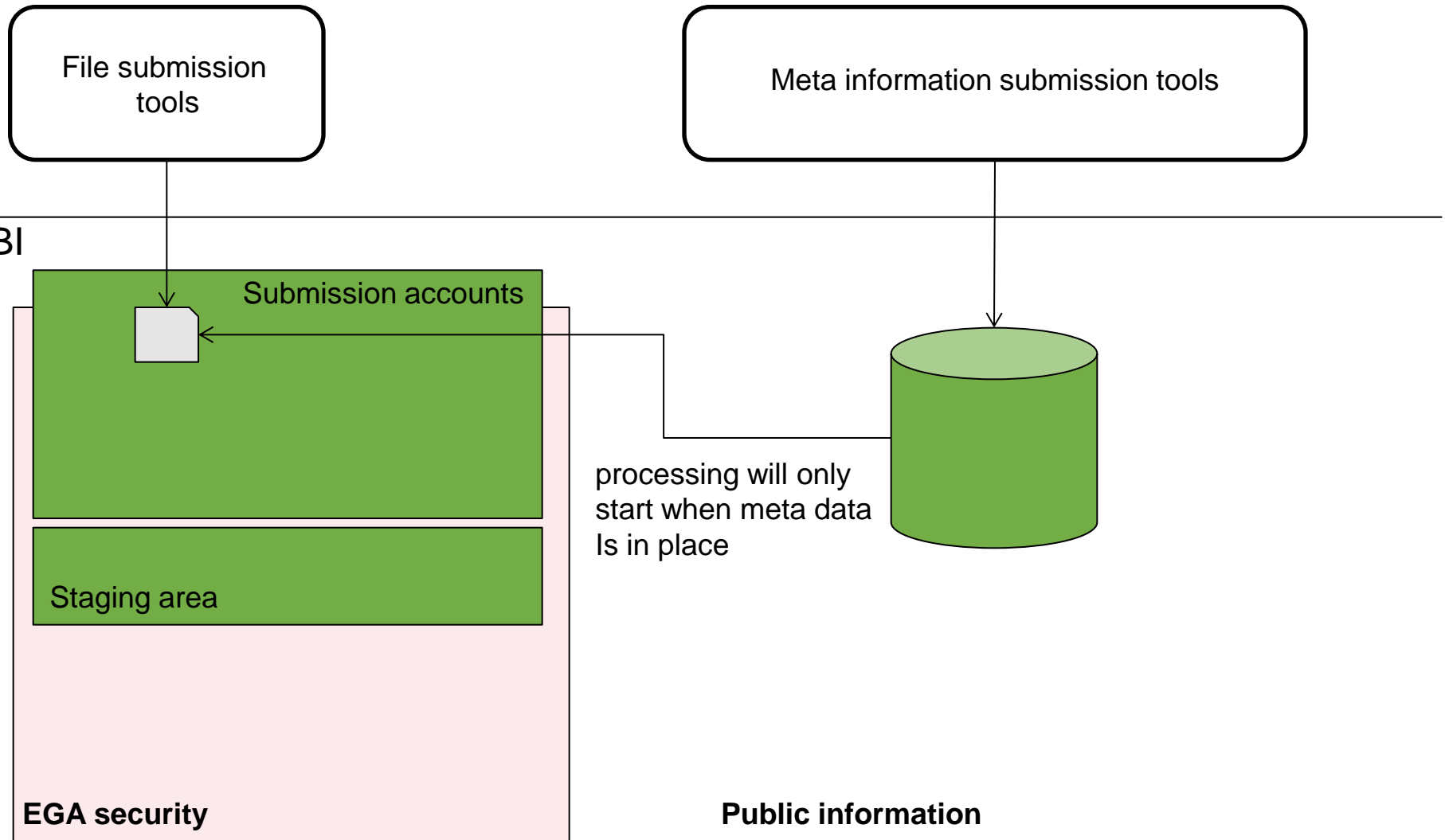
• Submission to EGA consists of two actions:

- File upload - supported file formats
- Meta data submission

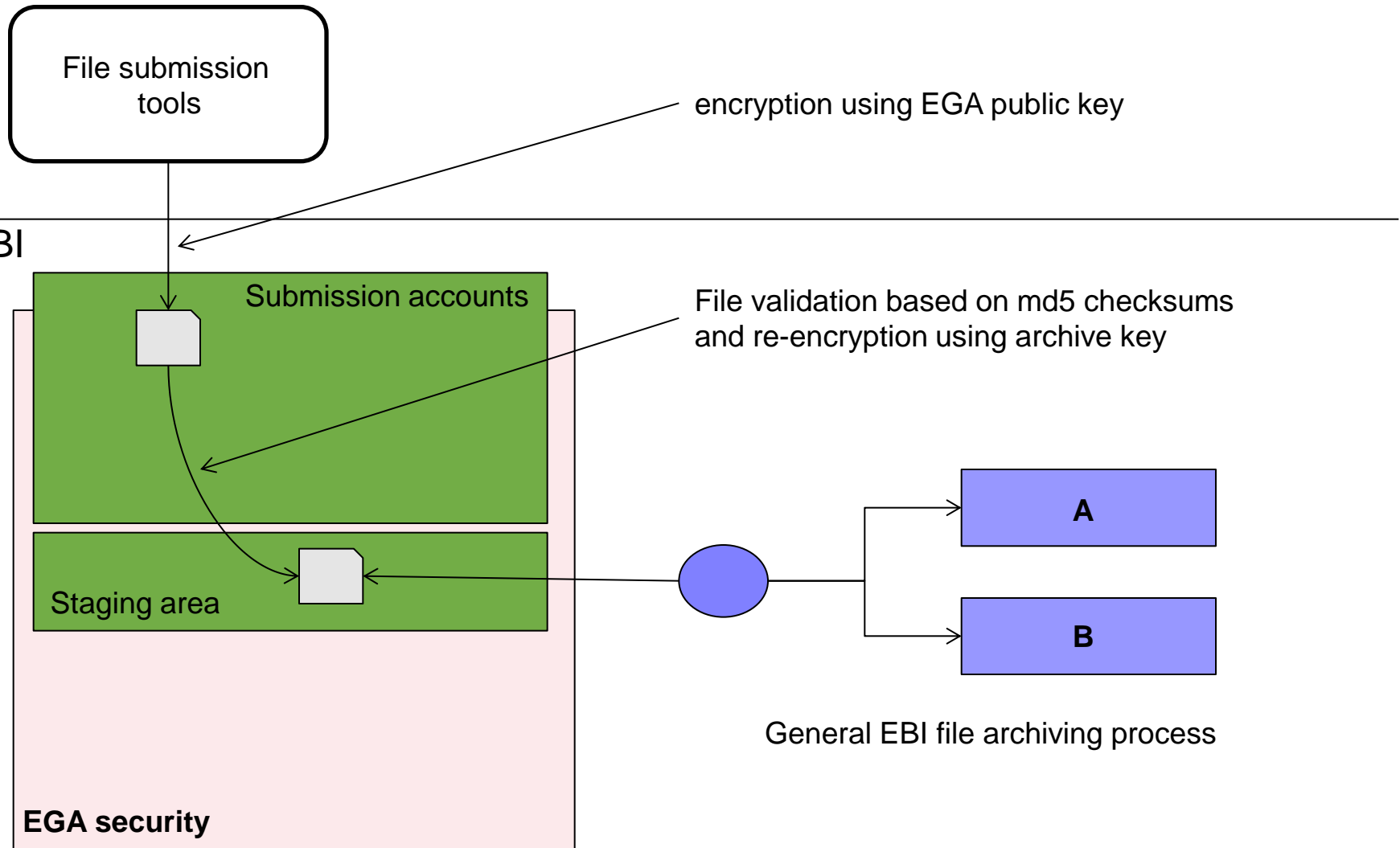
EGA data submission tools



EGA data submission tools



Processing submitted file into the archive



Supported file formats

● **All manufacturer-specific raw data formats for the major next generation sequencing platforms are accepted**

- We prefer BAM or CRAM file format for sequence data
- We prefer VCF file format for variant and genotype data

Fritz, M.H. Leinonen, R., et al. (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 21 (5), 734-40

Cochrane G., Cook C.E. and Birney E. (2012) The future of DNA sequence archiving. *GigaScience* 2012, 1:

<https://www.ebi.ac.uk/ega/submission/sequence>

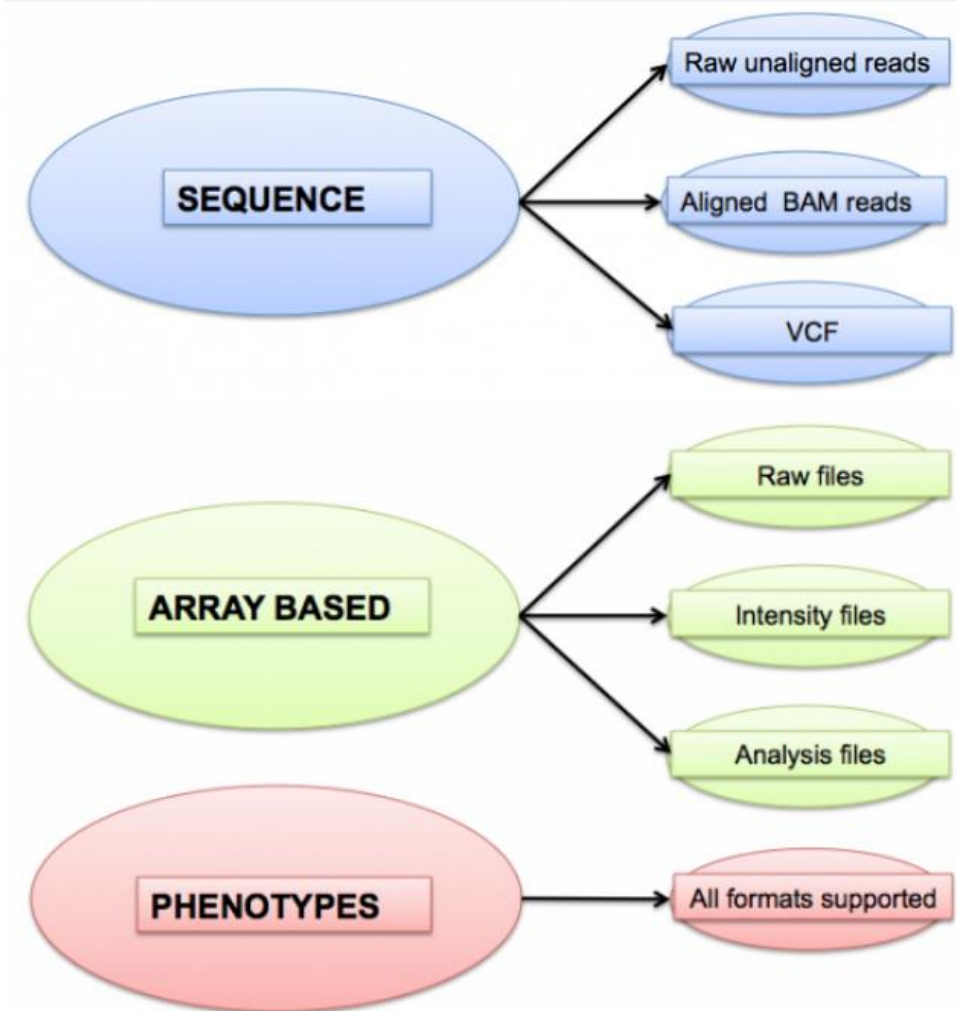
https://www.ebi.ac.uk/ena/about/cram_toolkit

Supported file formats

- **All manufacturer-specific raw data formats for the major next generation sequencing platforms are accepted**
 - We prefer BAM or CRAM file format for sequence data
 - We prefer VCF file format for variant and genotype data
- **All array-based technologies investigating genotyping, gene expression or methylation are accepted.**
 - We prefer to have all supporting data and final report files

<https://www.ebi.ac.uk/ega/submission/sequence>
https://www.ebi.ac.uk/ena/about/cram_toolkit

Supported file formats



BAM, FASTQ

BAM, CRAM

Affymetrix CELs

Illumina IDAT files

VCF, Plink, WTCCC etc

Spreadsheet

Phenotype submissions to EGA

- **At this point EGA requirements are very simple** – in general large projects are currently moving to use ontologies for managing their phenotype data.
 - Gender
 - Phenotype (using any ontology is recommended as this will allow us to connect submitted terms to a future service at the EBI or to harmonize data across projects).
 - Donor identifier (anonymised subject identifier that will link samples together)
- **What phenotypic data can be made fully public?**
 - EGA will release all meta information submitted as part of sample submission.
 - Any data that falls confidential under the informed consent must be submitted as a file to EGA. File must use the EGA sample accessions or aliases and it must be linked to a dataset to be distributed under DAC approval process.

<http://www.ebi.ac.uk/efo/>


Experimental Factor Ontology (EFO)

● EGA recommends Experimental Factor Ontology (EFO)

- Originates from the ArrayExpress submission system and therefore may not cover all use cases.
- Influence EFO directly to cover the H3Africa requirements.
- Alternatively use an Ontology that already covers all your terms and submit the data to us in a way that we can understand the terms, e.g. MeSH: D003922 for Diabetes Mellitus, type 1.
- You may need a number of different ontologies to describe e.g. disease and anatomical details.

<http://www.ebi.ac.uk/efo/>

Experimental Factor Ontology (EFO)




Experimental Factor Ontology

Examples: cancer, HoLa, Li-Fraumeni syndrome

[Home](#) [Browse EFO](#) [Submit Term](#) [Semantic Web Project](#)


Representing experimental variables with EFO

The **Experimental Factor Ontology** (EFO) provides a systematic description of many experimental variables available in EBI databases, and for external projects such as the NHGRI GWAS catalogue. It combines parts of several biological ontologies, such as anatomy, disease and chemical compounds. The scope of EFO is to support the annotation, analysis and visualization of data handled by the EBI [Functional Genomics Team](#). We also add terms for external users when requested. If you are new to ontologies, there is a [short introduction](#) on the subject available and a blog post by James Malone [on what ontologies are for](#).




Browse

Browse EFO at [NCBO BioPortal \(external\)](#) or in EBI's [OLS](#). You can also search EFO using the search box, above.




Download

Download the [latest release](#) of EFO in OWL format. There is an [OBO format version](#) and an [inferred OWL view](#). Read the latest [Release Notes](#).




Tools

We provide tools to support the development and use of EFO such as [Bubastis](#), an ontology diff tool, and [Semantic Web work](#). See the complete [list of tools](#).




FAQ

Read more about [EFO](#) or see the [Frequently Asked Questions](#). You can also read about ontologies in [James Malone's blog](#).



Submit

Submit new terms or report bugs using our [JIRA ticket system](#). You can also join the [mailing list](#).



Contact

You can email [James Malone](#) directly with any questions or email the [EFO list](#).

<http://www.ebi.ac.uk/efo/>

Submission tools

● **EGA Webin Data Uploader**

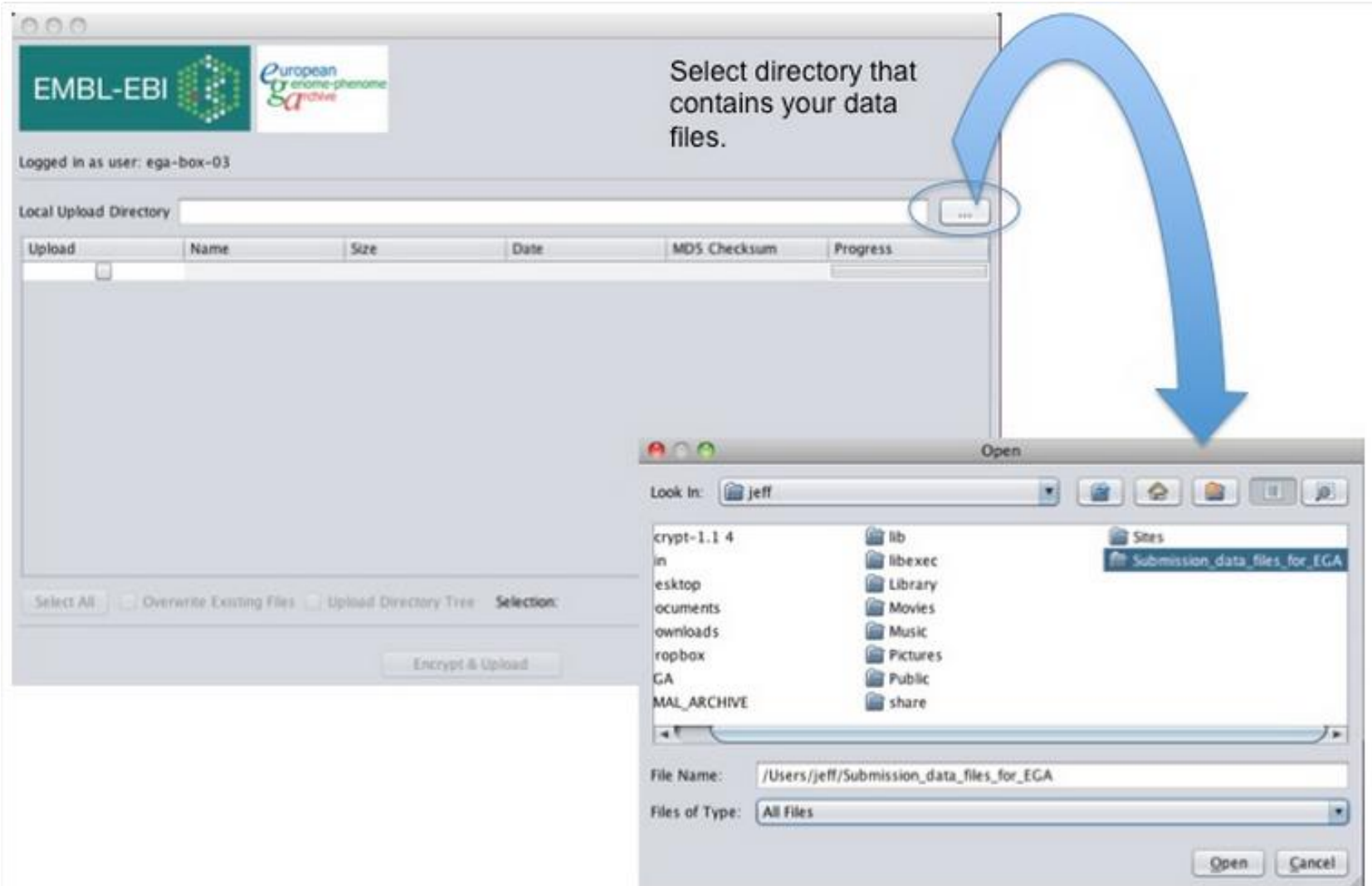
- Java Client to be installed locally
- Java command line tools
- Supports FTP and UDT

● **EGA Webin - meta data submission**

- An online tool that allows submitters describe their study and associated files available at <https://www.ebi.ac.uk/ena/submit/sra/#home>.
- Programmatical REST interface for large scale submitters
- Submitting data from platforms other than NGS – AF spreadsheet

<https://www.ebi.ac.uk/ega/submission/applications>

EGA Webin Data Uploader



https://www.ebi.ac.uk/ega/submission/tools/EGA_webin_data_uploader

EGA Webin Data Uploader

EMBL-EBI European genome-phenome archive

Logged in as user: ega-box-03

Local Upload Directory: /A/User Support Documentation/Submission_Documentation/New_new_submission_documentation/Screenshots

Upload	Name	Size	Date	MDS Checksum	Progress
<input checked="" type="checkbox"/>	Capture.PNG	6.36 KB	03-Aug-2011		0%
<input checked="" type="checkbox"/>	DAC.png	132.50 KB	02-Dec-2011		0%
<input checked="" type="checkbox"/>	dataset.png	120.40 KB	02-Dec-2011		0%
<input checked="" type="checkbox"/>	EBI_archives.PNG	110.07 KB	10-Aug-2011		0%
<input checked="" type="checkbox"/>	Genotypes_new.png	228.09 KB	02-Dec-2011		0%
<input checked="" type="checkbox"/>	Phenotypes_tab.png	50.96 KB	02-Dec-2011		0%
<input checked="" type="checkbox"/>	Policy.png	103.62 KB	02-Dec-2011		0%
<input checked="" type="checkbox"/>	Sequence_type.PNG	20.85 KB	06-Jul-2011		0%
<input checked="" type="checkbox"/>	Sequence_XML_stag...	36.32 KB	07-Jul-2011		0%
<input checked="" type="checkbox"/>	Sequence_XML_stag...	26.55 KB	07-Jul-2011		0%
<input checked="" type="checkbox"/>	Sequence_XML_stag...	10.24 KB	07-Jul-2011		0%
<input checked="" type="checkbox"/>	Sequence_XML_stag...	42.16 KB	12-Aug-2011		0%
<input checked="" type="checkbox"/>	Submission_1_recei...	26.81 KB	14-Jul-2011		0%
<input checked="" type="checkbox"/>	Submission_2_recei...	16.51 KB	14-Jul-2011		0%
<input checked="" type="checkbox"/>	Submission_types.P...	43.43 KB	05-Jul-2011		0%
<input checked="" type="checkbox"/>	XML_pipeline.PNG	17.15 KB	13-Jul-2011		0%

Select None Overwrite Existing Files Upload Directory Tree 16 files selected. Total size: 992.00 KB

Encrypt & Upload

Select your files and click on
'Encrypt & Upload'

https://www.ebi.ac.uk/ega/submission/tools/EGA_webin_data_uploader

Using Command line Data Uploader

- **Integrate Java command line application as part of the local pipeline.**
- **Prepare EGA compliant files for submission but use Aspera for data upload.**

```
java -jar webin-data-streamer-Upload-Client.jar -p -user -pass -files
```

```
java -jar ../webin-data-streamer-Upload-Client.jar -file file1 file2
```

https://www.ebi.ac.uk/ega/submission/tools/EGA_webin_data_uploader

Submission tools

● EGA Webin Data Uploader

- Java Client to be installed locally
- Java command line tools
- Supports FTP and UDT

● EGA Webin - meta data submission

- An online tool that allows submitters describe their study and associated files available at <https://www.ebi.ac.uk/ena/submit/sra/#home>.
- Programmatic REST interface for large scale submitters
- Submitting data from platforms other than NGS – AF spreadsheet

<https://www.ebi.ac.uk/ega/submission/applications>

Submission automation

- **Integrate your local LIMS system to our programmatic interface to automate submission process**
 - Contact ega-helpdesk for more information.
 - Map the mandatory and optional fields for each meta data object to the appropriate information stored within the local LIMS.
 - Test submissions using our test-server, use production server for real submissions.
 - Store EGA accessions directly into LIMS for successful submissions.
 - Submission, deprecation or update actions are also available using this interface.

What are the meta data requirements

● EGA requires

- *Project* – short description of the project or study
- *Sample* – description of each used sample
- *Experiment* – experiment type and platform details
- *Analysis* – results of the data processing, e.g. how the VCF file was created
- *Run* – references the raw data file
- *Dataset* – is a container for all files to be authorized to a successful applicant
- *Policy* – links DAC to a dataset
- *Data Access Committee (DAC)* – defines the data access authority

What are the meta data requirements

● EGA requires

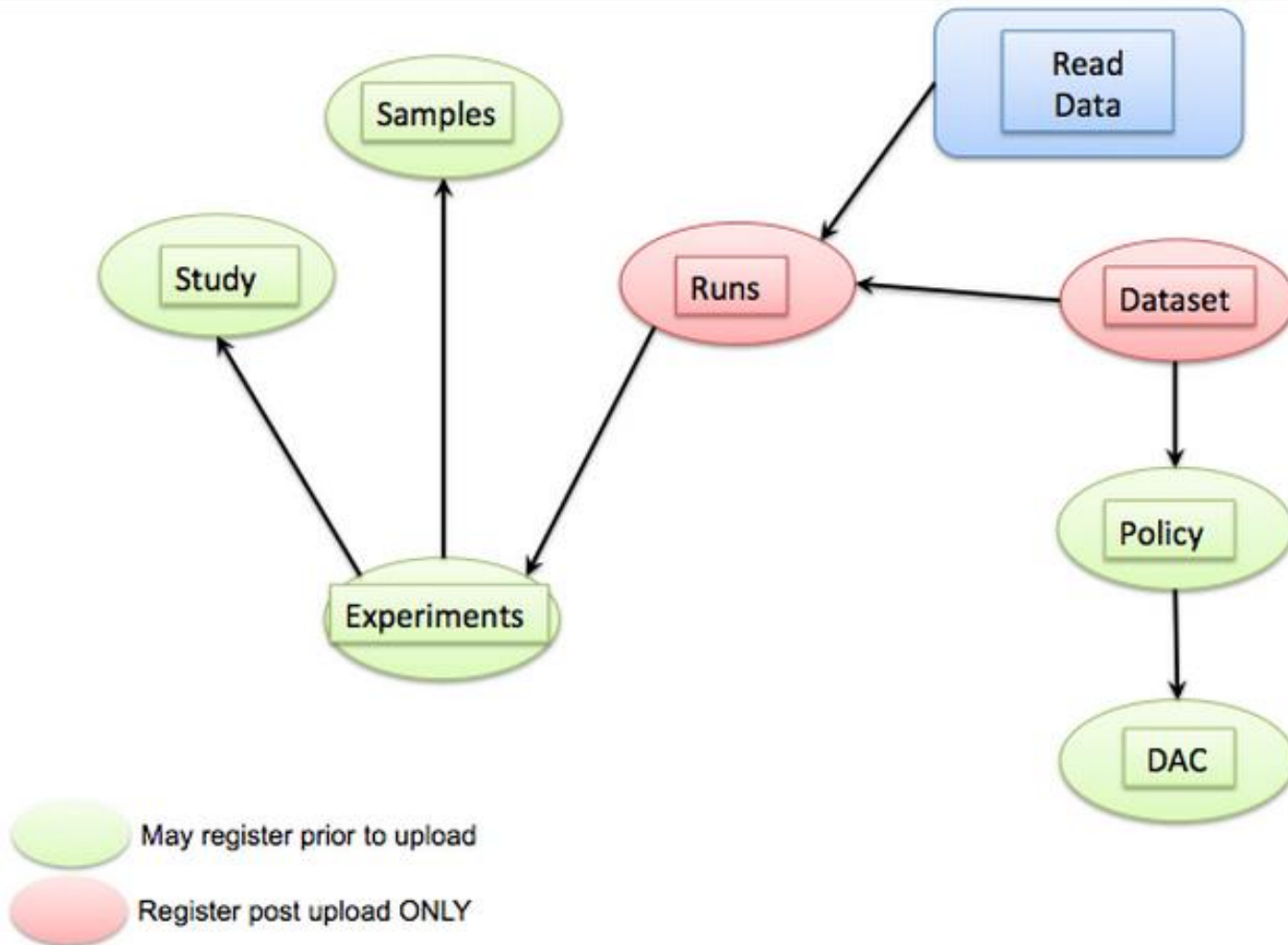
ENA and EGA shared requirements

- *Project* – short description of the project or study
- *Sample* – description of each used sample
- *Experiment* – experiment type and platform details
- *Analysis* – results of the data processing, e.g. how the VCF file was created
- *Run* – references the raw data file

- *Dataset* – is a container for all files to be authorized to a successful applicant
- *Policy* – links DAC to a dataset
- *Data Access Committee (DAC)* – defines the data access authority

EGA specific requirements

Relationships between meta data objects



ftp://ftp.sra.ebi.ac.uk/meta/xsd/sra_1_5/

EGA Webin

Home **New Submission** Studies Samples Experiments Runs Projects DACs Policies Datasets

Start >> Sample >> Finish

You have logged in as EGA submitter. You can use this service to [submit sequence reads](#) into EGA and to register [studies](#) and [samples](#). Please note that access to [data](#) submitted into EGA is controlled by submitter nominated [data access committee](#) and is available under the [data access policy](#) associated with the data.

Please select the type of submission:

- Submit sequence reads and experiments
- Register sequencing study
- Register samples
Samples can be pre-registered before submitting data into EGA.
- Submit dataset
- Register data access policy
- Register data access committee (DAC)
- Register umbrella study (project)

Please contact [ega-submission](#) to make your metadata public. Access to the data submitted to EGA is controlled by the nominated data access committee.

[Restart Submission](#)

https://www.ebi.ac.uk/ega/submission/sequence/unaligned#Webin_study

EGA Webin

The screenshot shows the EGA 'Sample' creation page. At the top, there is a navigation bar with tabs: Home, New Submission, Studies, Samples, Experiments, Runs, Projects, DACs, Policies, and Datasets. Below the navigation bar, a progress bar shows 'Start' (with a green checkmark), '>>', 'Sample', '>>', and 'Finish'. The main heading is 'Sample'. Below this, a message reads: 'Please create new samples by uploading a spreadsheet or by following the instructions below.'

The form is divided into several sections:

- Attributes Section:** A box titled 'Please select the checklist attributes you would like to include with each sample. You may also add custom attributes.' It contains a search bar 'Filter attributes...', an 'Add your own attribute' button, and an '+ Add' button. Below this is a list of attributes under a 'default' header: 'subject_id - optional' (Identifier for the subject where the sample has been derived from), 'gender - optional' (Sex), 'phenotype - optional', 'disease_site - optional' (Attached organ), and 'sample type - optional' (Attached organ). There is also a 'User Attributes' section with a checked 'i35 - user' attribute. At the bottom of this section, it says '3 of 7 attributes selected' and includes 'Expand', 'Collapse', and 'Download Template' buttons.
- Basic Details Section:** A box titled 'Please complete any fields that you would like to apply to all samples. This will act as a template for the rest of the samples.' It contains: 'Unique Name Prefix: My_samples', '* Title: Samples taken from tissue X', and 'Description: Samples affiliated to study Y'.
- Organism Details Section:** Contains: 'Search:' (with a search icon), '* Tax ID: 9606', '* Scientific Name: Homo sapiens', 'Common Name: human', and 'Anonymized Name: My_samples'.
- Additional Fields:** Below the Organism Details, there are 'phenotype:' and 'disease_site:' input fields.

At the bottom of the form, there are navigation buttons: '<< Previous', 'Next >>', and a 'Restart Submission' button with a green checkmark icon.

Create additional attributes for each sample

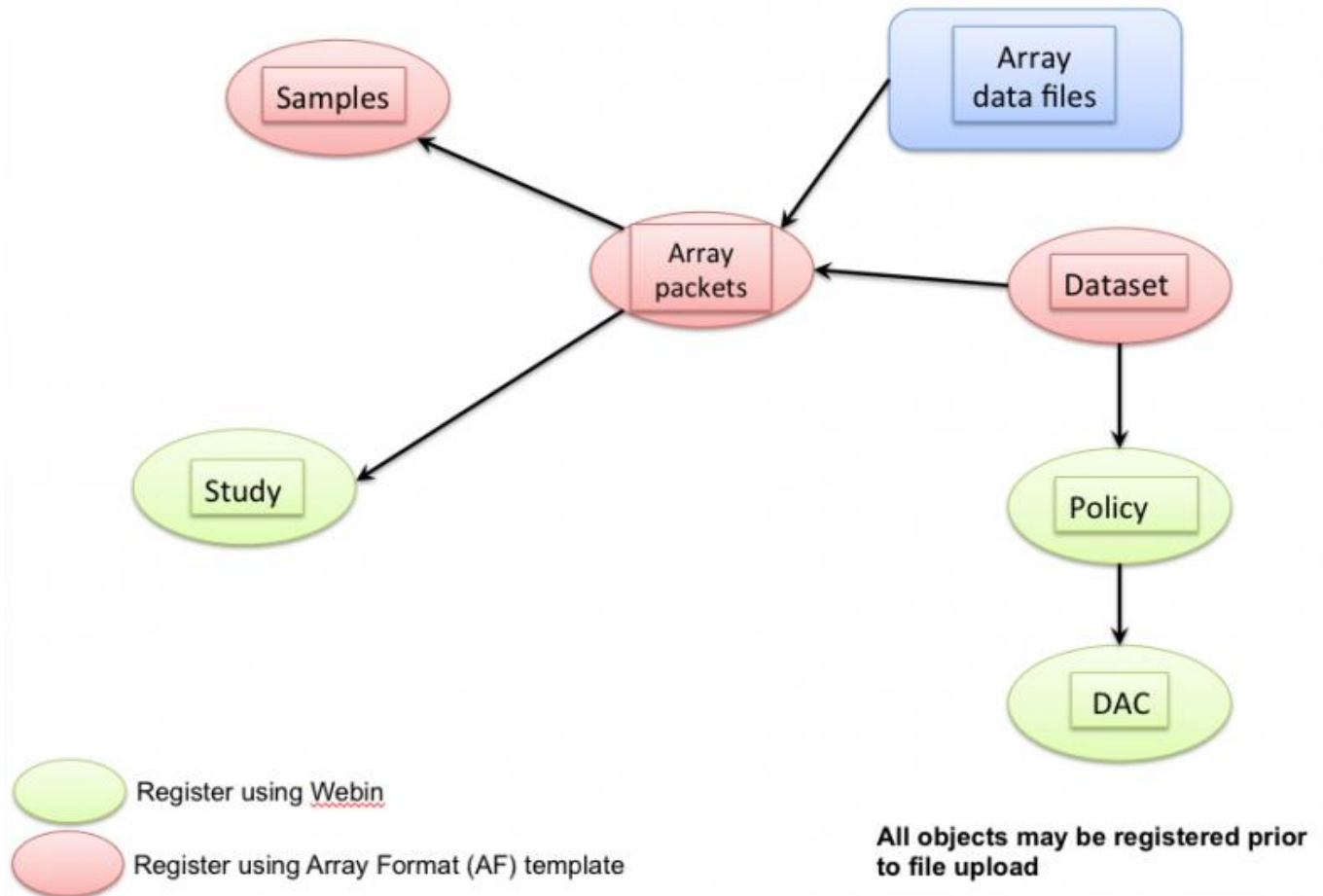
Select default attributes for each sample

Complete basic and organism details to apply to all samples

Click next or 'Download Template' to populate offline

Meta data submission for array based experiments

Required metadata objects for array-based submissions



Array-based submission spreadsheet – Referencing correct accessions

# Provide your Study accession number (EGAS00001000XXX) generated in Webin and also your data upload submission box number	
Study Accession Number	EGAS00001000527
Submission box	ega-box-100
Data deposited outside of EGA (e.g ArrayExpress or European Nucleotide Archive)	
# Provide your Data Access Committee (DAC) & Policy accession number/s generated in Webin. Multiple DAC accession numbers should be separated with a ',' (e.g. EGAC00001000XX1; EGAC00001000XX2)	
# DAC defines the body or consortium responsible for the application and approval procedures for the dataset/s submitted and Policy specifies the terms and conditions of data access	
DAC accession number/s	EGAC00001000114
Policy accession number/s	EGAP00001000115

Figure 1: Referencing accessions generated through Webin

Add study accession and submission box number

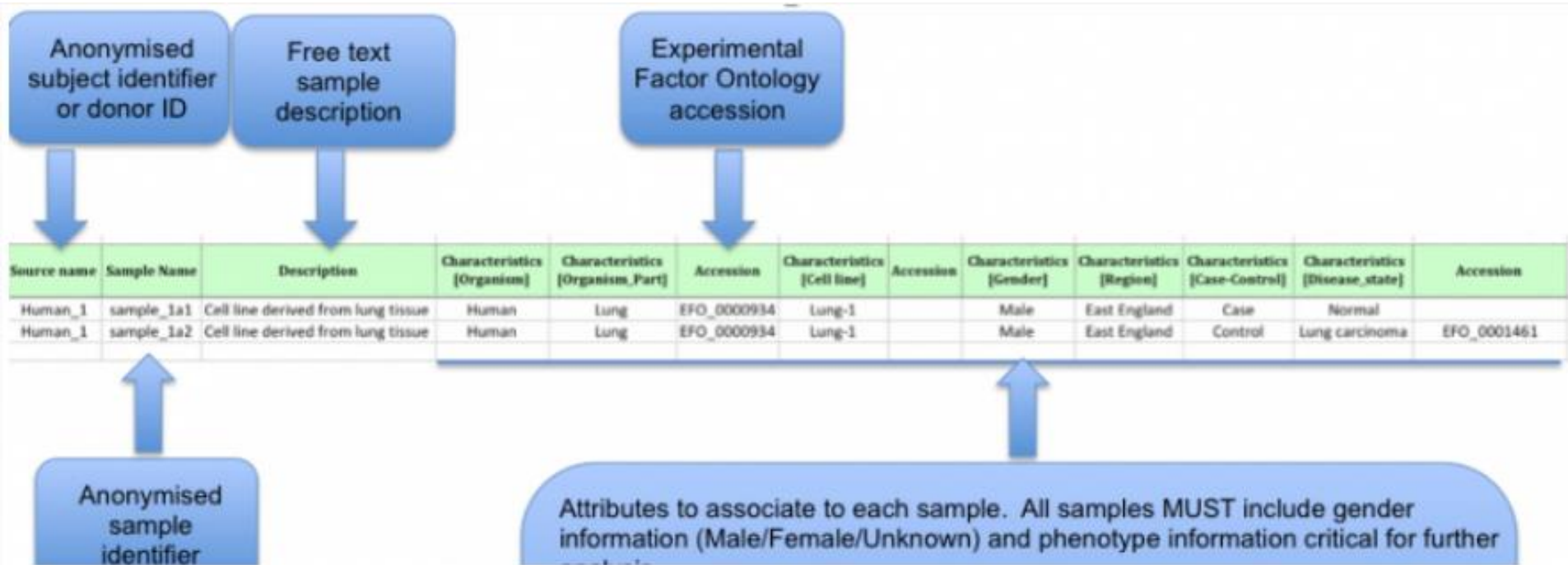
External accessions may also be referenced

Add DAC and policy accession numbers



Should your submission require multiple DAC's or policies, use ';' to separate the accession numbers.

Array-based submission spreadsheet – Describing samples



Array-based submission spreadsheet – Creating dataset for the files

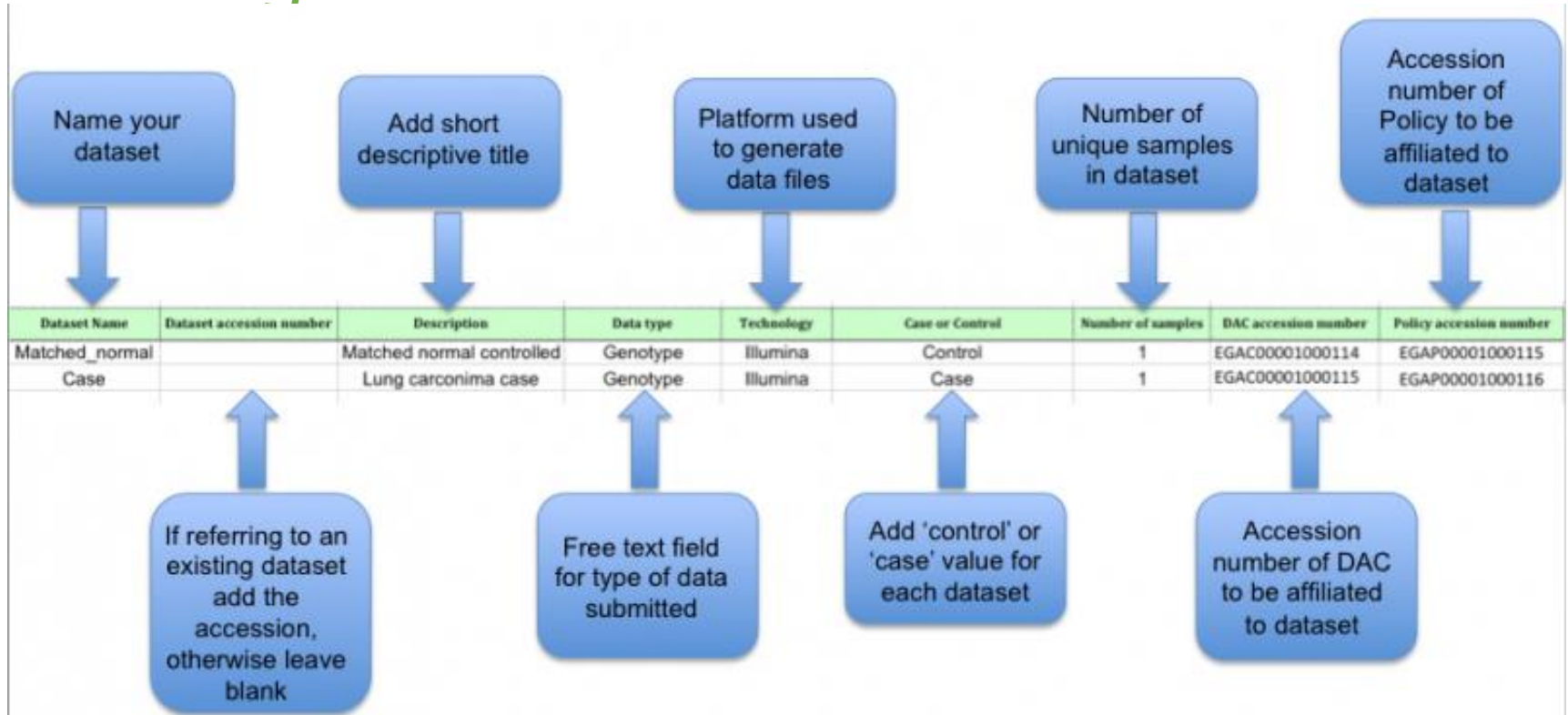


Figure 3: Creating array-based datasets



'Description', 'Data type', 'Technology', 'Case or Control' and 'Number of samples' fields are displayed on the EGA public site.

Array-based submission spreadsheet – describing files within a dataset

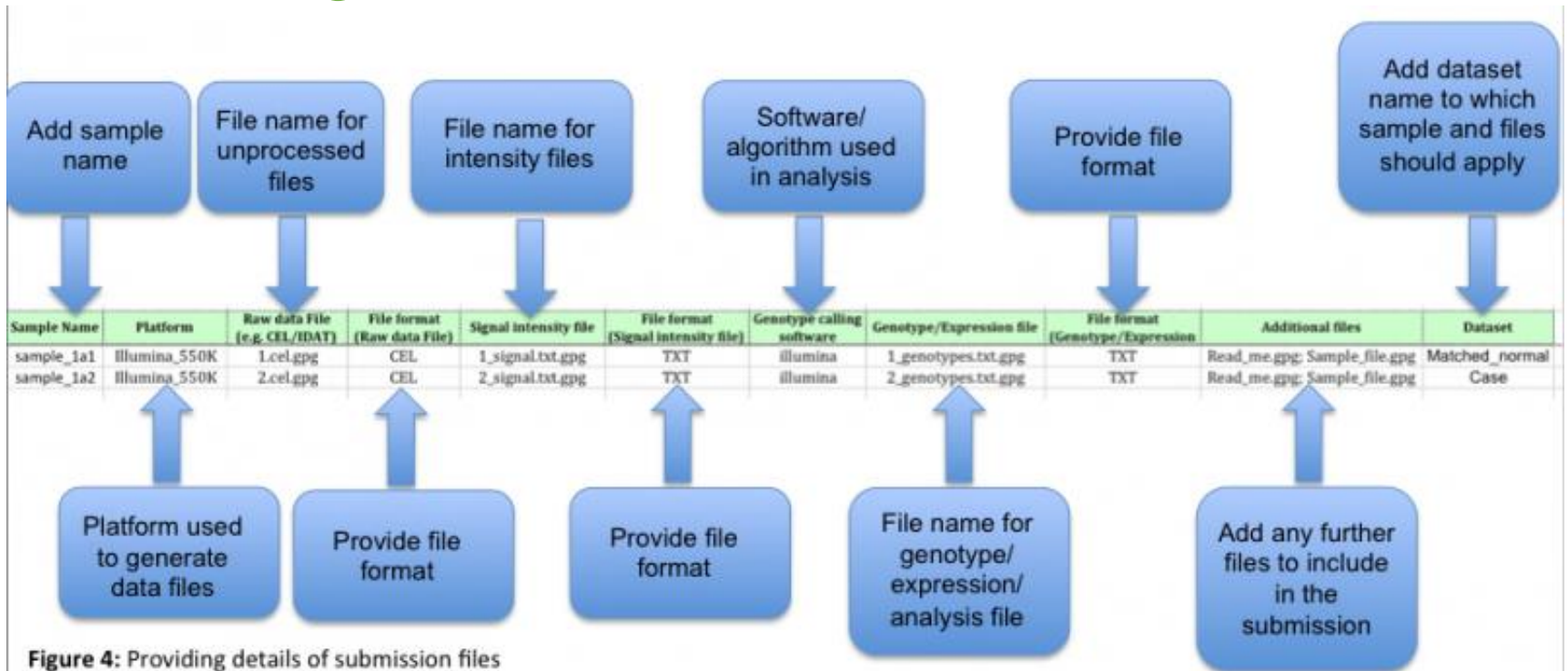
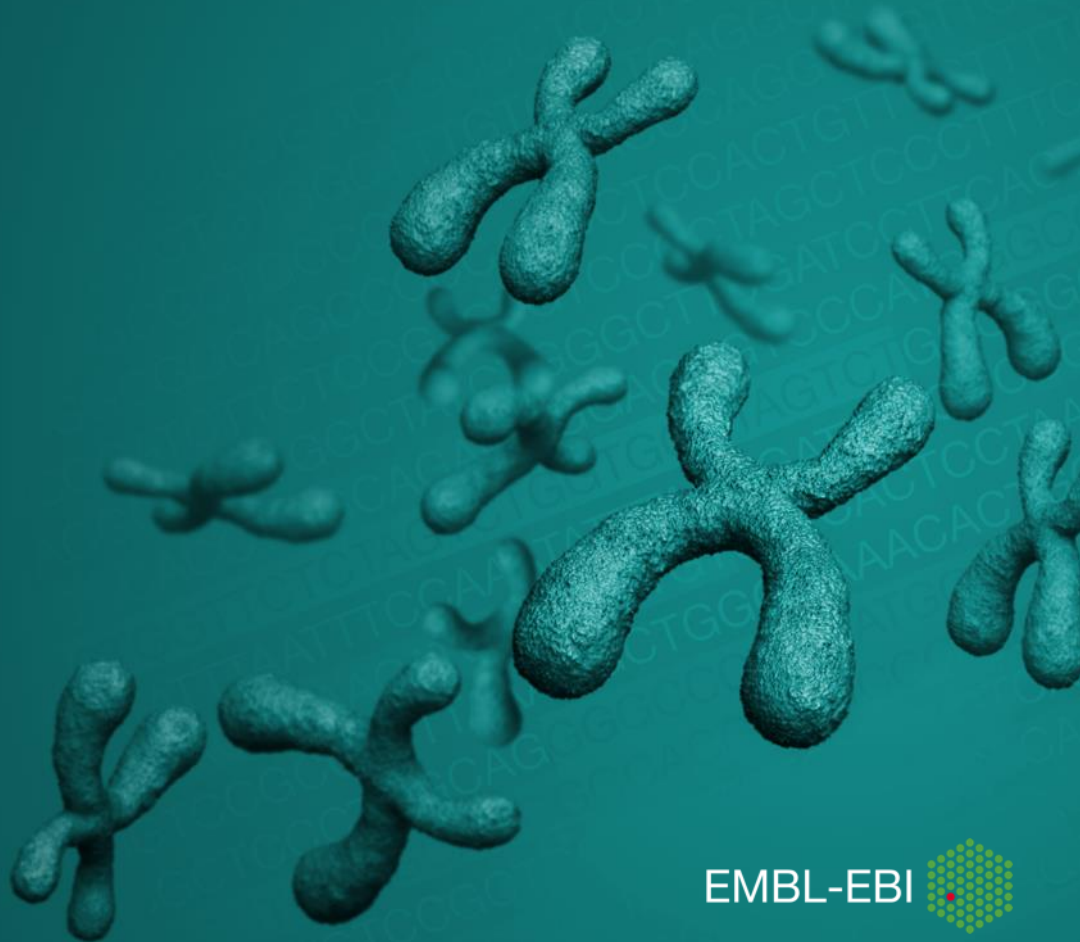


Figure 4: Providing details of submission files



'Sample Name', 'Platform' and 'Dataset' fields must be completed. The 'File format' column must be completed for all data files referenced. 'Genotype calling software' column must be completed if 'Genotype/Expression files' are referenced.

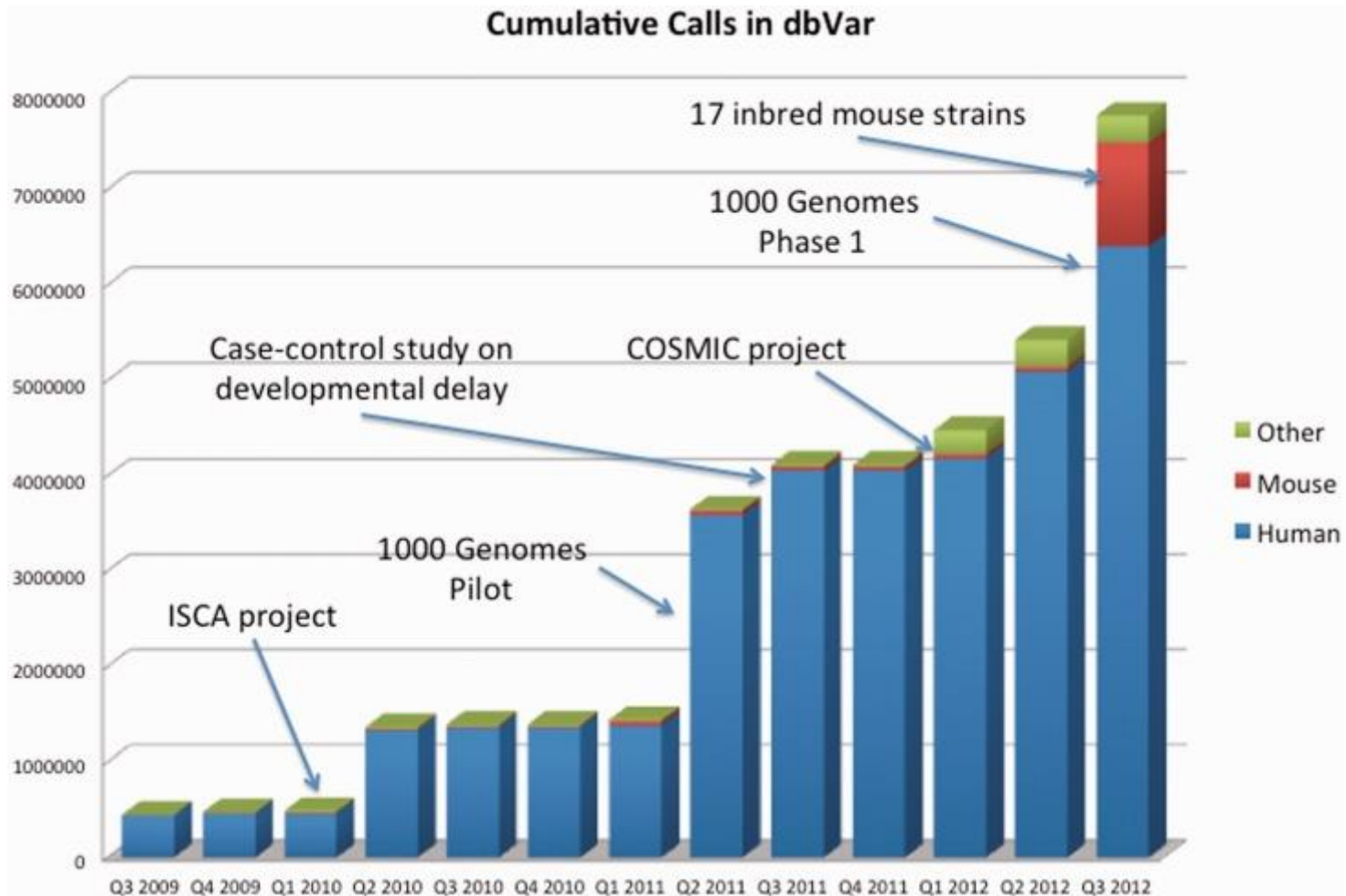
Database of Genomic Variants Archive (DGVA)



DGVa overview

- Permanent repository of all types of genomic structural variants in all species.
- Accepts direct submissions or curates data from literature.
- The data is archived on a per-study basis, often relating to an individual publication.
- Provides accession space for structural variants jointly with dbVar from NCBI, USA.
- All data are freely available from the service and integrated to other services at the EBI or outside of it.

DGVa includes the most important public reference sets

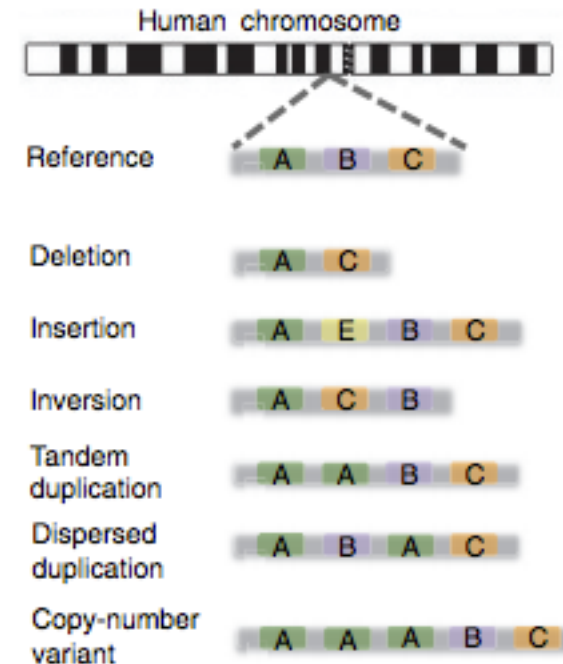


Structural Variation

• Structural variants are variations in DNA over 50bp long

• Variation types include:

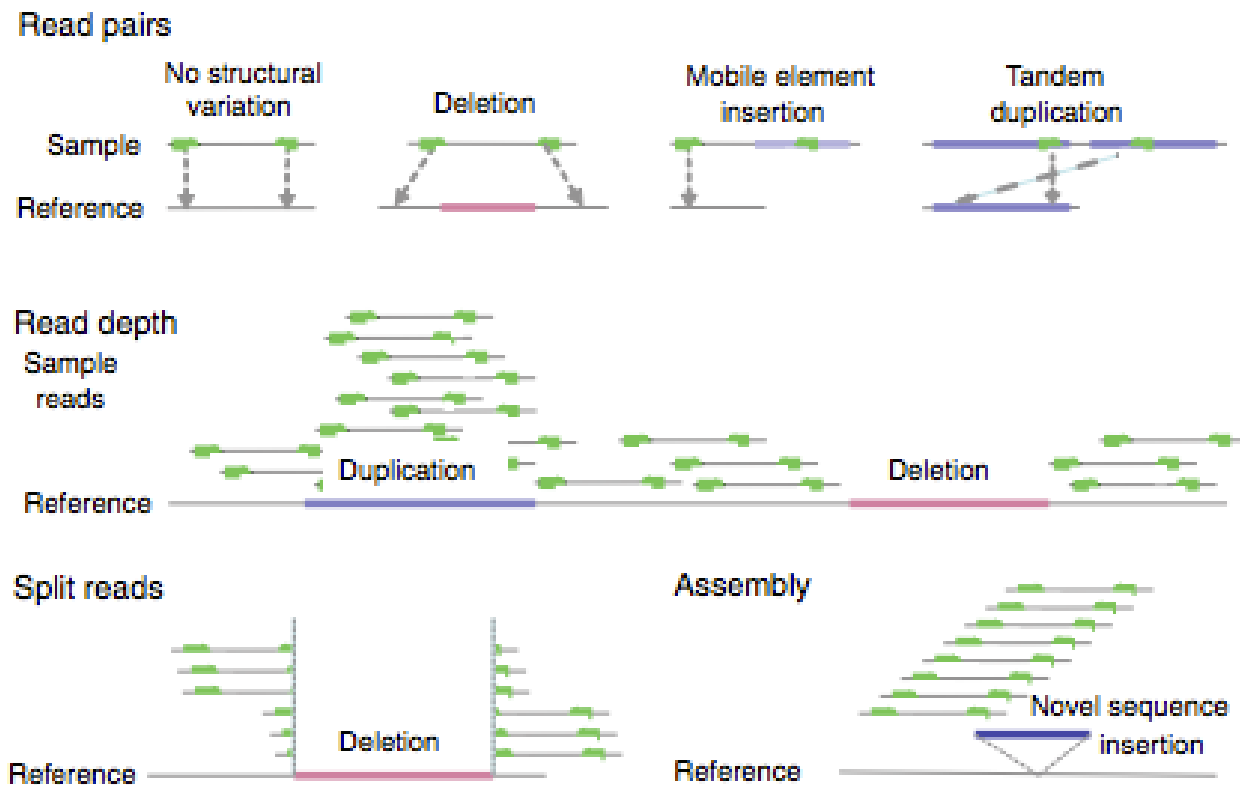
- Insertions
- Copy number gains or losses
 - Deletions, Duplications
- Inversions
- Translocations



• Accounts for more bp variation (~50Mbp) than SNPs in human genome

Detecting Structural Variants - Sequencing

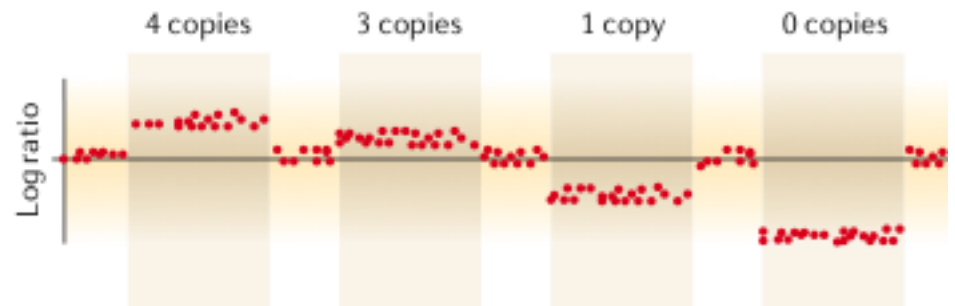
- Depending on method, sequencing can give bp resolution
- Genotype can be determined



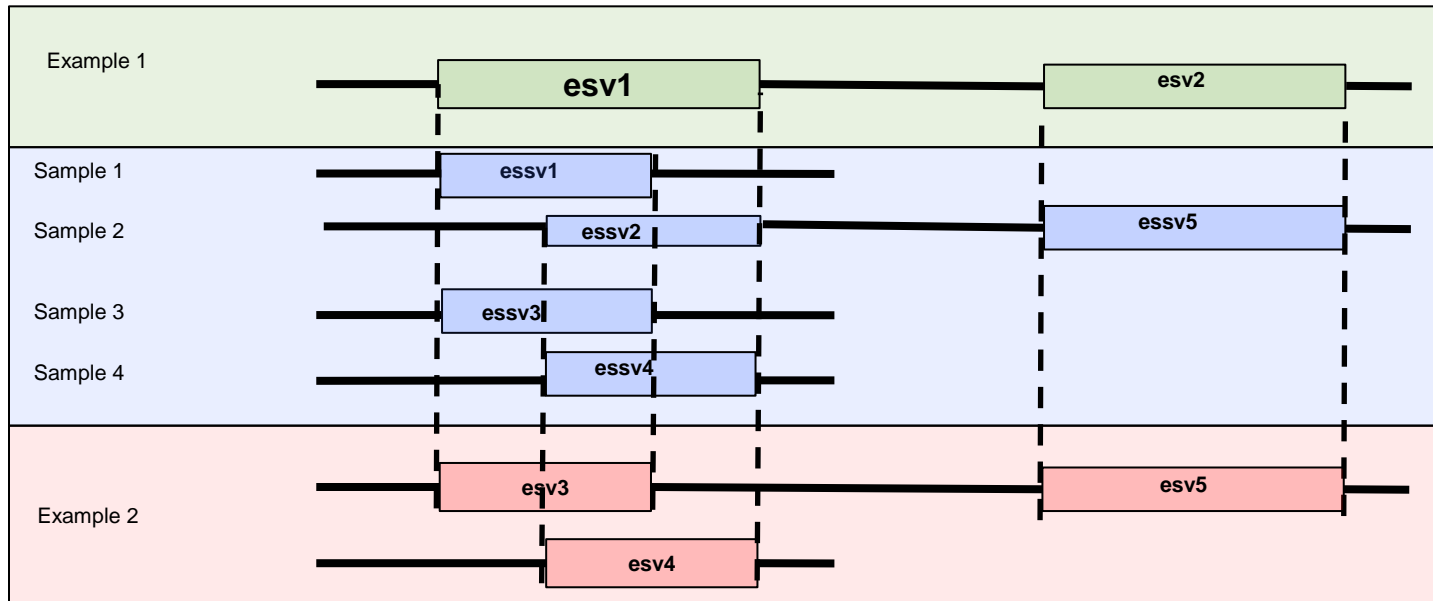
Jan Korbel, European Molecular Biology Laboratory

Detecting Structural Variants - Arrays

- Resolution depends on probe spacing
 - Inner start/stop to indicate first/last affected base
 - Outer start/stop to indicate first/last unaffected base
- No genotype information

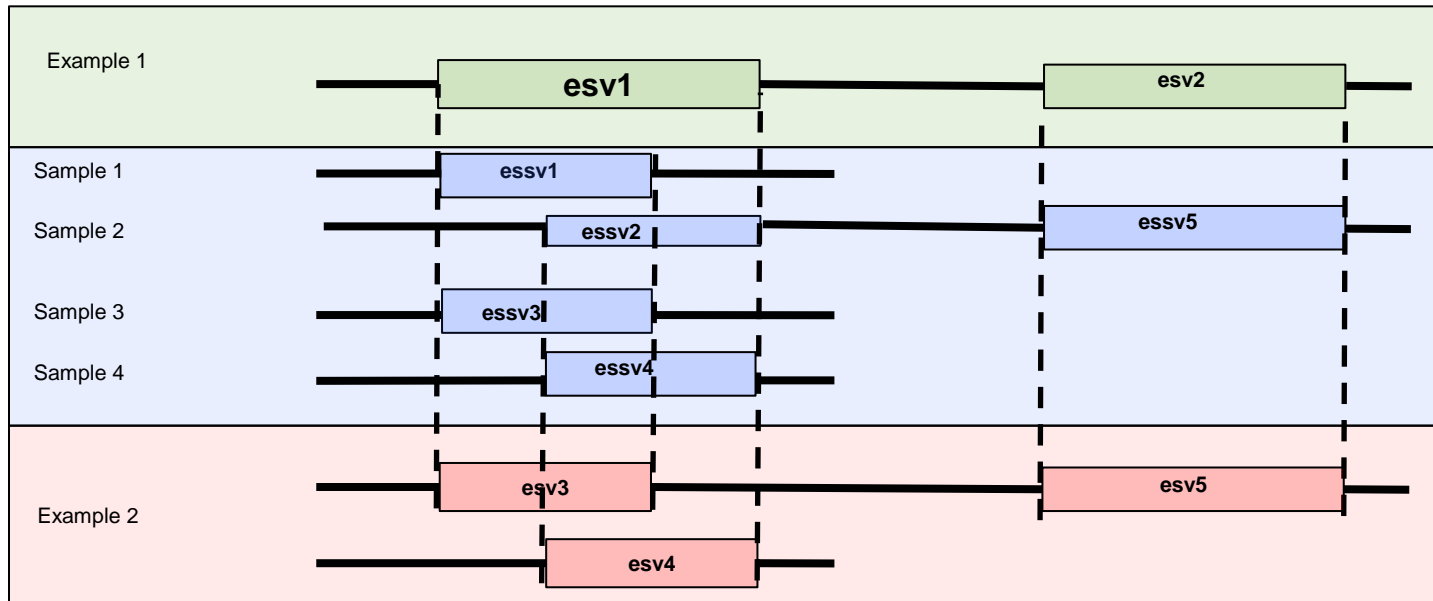


Structural Variation at DGVA



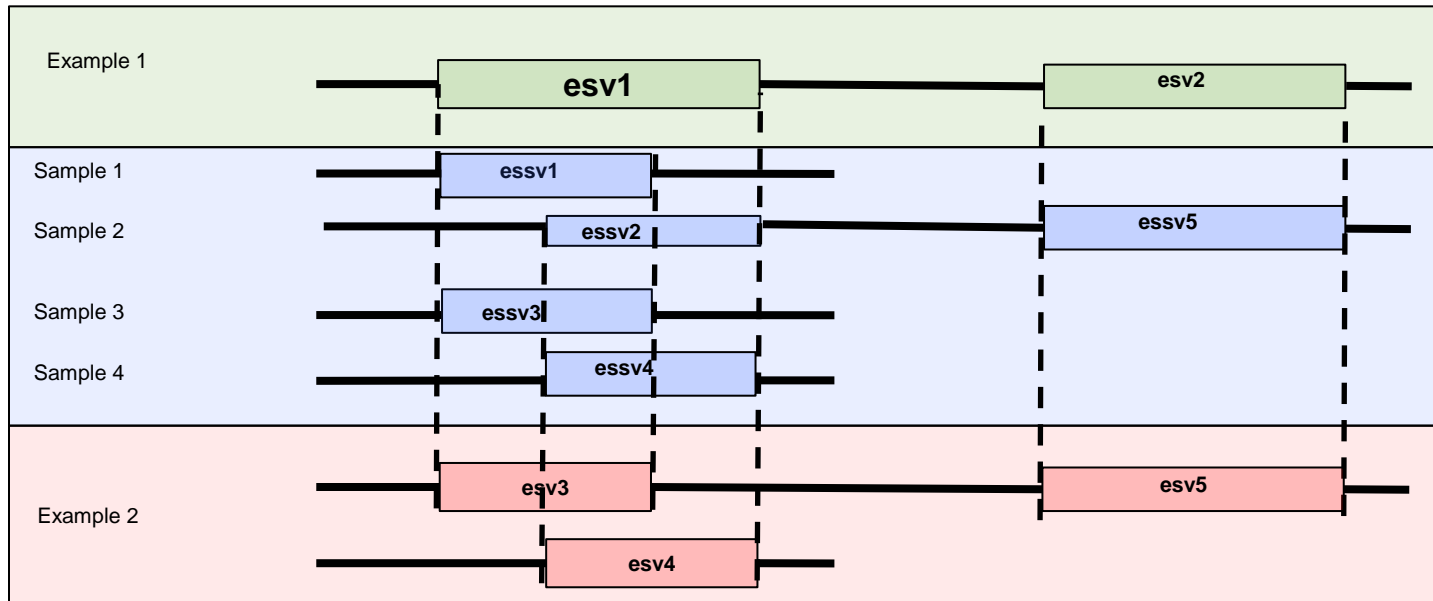
- Structural variants divided into 2 main classes:
 - Variant Call (supporting structural variant)
 - Variant Region (Structural Variant)

Structural Variation at DGVA



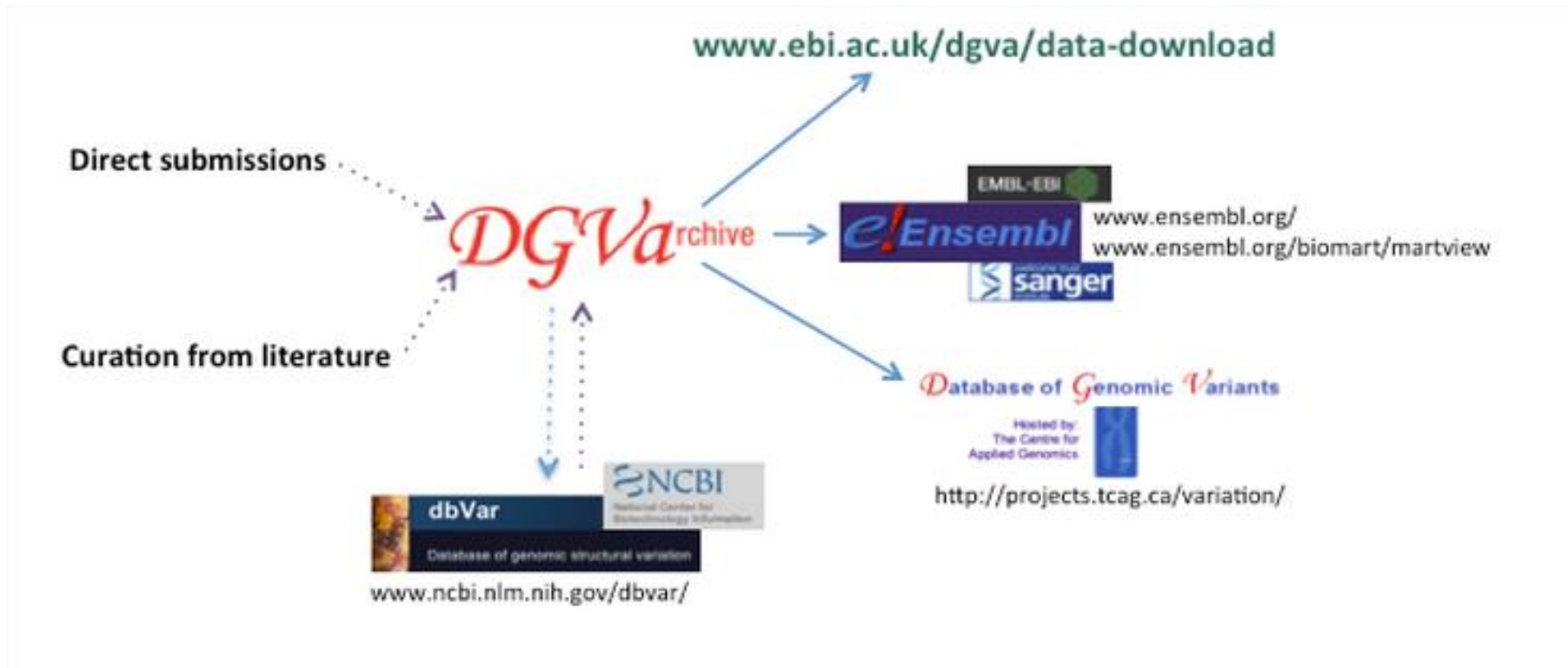
- A Variant Region is supported by 1 or more Variant Calls
- A Variant Call is the actual variant seen in a particular sample or set of samples
- Assertion method records how the calls support the region

Structural Variation at DGVA



- Example 1: Assertion method 50% overlap
- Example 2: Assertion method 100% overlap
- DGVA will create regions if required

DGVA Data Flow Diagram



DGVA download site

Data Download

Genomic structural variant study data can be downloaded via ftp by following the appropriate link.

Studies				
Study	Reference	Organism	Variants	Link
estd209	Pang et al 2013b	Homo sapiens	471817	Download via FTP
estd208	Helbig et al 2013	Homo sapiens	81	Download via FTP
estd205	Zichner et al 2012	Drosophila melanogaster	65487	Download via FTP
estd204	Simon et al 2013	Mus musculus	43	Download via FTP
estd203	Vogler et al 2010	Homo sapiens	4193	Download via FTP
estd201	Wong et al 2012b	Homo sapiens	36558	Download via FTP
estd200	Wong et al 2012	Mus musculus	30044	Download via FTP
estd199	1000 Genomes Consortium Phase 1	Homo sapiens	22531	Download via FTP
estd198	Chia et al 2012	Homo sapiens	381	Download via FTP
estd197	McKernan et al 2009	Homo sapiens	232775	Download via FTP
estd196	Simon-Sanchez et al 2007	Homo sapiens	335	Download via FTP
estd195	Altshuler et al 2010	Homo sapiens	856	Download via FTP
estd194	Bentley et al 2008	Homo sapiens	504912	Download via FTP
estd193	Feuk et al 2005	Homo sapiens	3	Download via FTP
estd192	COSMIC	Homo sapiens	15168	Download via FTP
estd188	Pinto et al 2011	Homo sapiens	60247	Download via FTP
estd186	Thevenon et al 2012	Homo sapiens	3	Download via FTP
estd185	Yalcin et al 2012	Mus musculus	1453	Download via FTP
estd180	Pang et al 2010	Homo sapiens	23887	Download via FTP
estd176	Banerjee et al 2011	Homo sapiens	734	Download via FTP

DGVA download site

Data Download

Genomic structural variant study data can be downloaded via ftp by following the appropriate link.

Studies				
Study	Reference	Organism	Variants	Link
estd209	Pang et al 2013b	Homo sapiens	471817	Download via FTP
estd208	<h2 style="text-align: center;">Index of /pub/databases/dgva/estd209_Pang_et_al_2013b/gvf/</h2>			
estd205				
estd204				
estd203				
estd201				
	Name	Size	Date Modified	
estd200	[parent directory]			
estd200	estd209_Pang_et_al_2013b.2014-04-01.GRCh37.Submitted.gvf	206 MB	4/1/14 8:12:00 PM	
estd199	estd209_Pang_et_al_2013b.2014-04-01.GRCh37.Submitted.gvf.gz	17.4 MB	4/1/14 8:12:00 PM	
estd198	estd209_Pang_et_al_2013b.2014-04-01.GRCh38.Remapped.gvf	219 MB	4/1/14 8:12:00 PM	
estd197	estd209_Pang_et_al_2013b.2014-04-01.GRCh38.Remapped.gvf.gz	17.5 MB	4/1/14 8:12:00 PM	
estd196	previous/		4/1/14 8:12:00 PM	
estd195				
estd194	Bentley et al 2008	Homo sapiens	504912	Download via FTP
estd193	Feuk et al 2005	Homo sapiens	3	Download via FTP
estd192	COSMIC	Homo sapiens	15168	Download via FTP
estd188	Pinto et al 2011	Homo sapiens	60247	Download via FTP
estd186	Thevenon et al 2012	Homo sapiens	3	Download via FTP
estd185	Yalcin et al 2012	Mus musculus	1453	Download via FTP
estd180	Pang et al 2010	Homo sapiens	23887	Download via FTP
estd176	Banerjee et al 2011	Homo sapiens	734	Download via FTP

Submissions to DGVA

• **Contact** eva-helpdesk@ebi.ac.uk.

• **Complete your meta data submission template and email it to the helpdesk.**

- We will validate your submission template
- Exchange data with dbVar
- Coordinate data release with your publication

<http://www.ebi.ac.uk/dgva/data-submission>

What are the meta data requirements

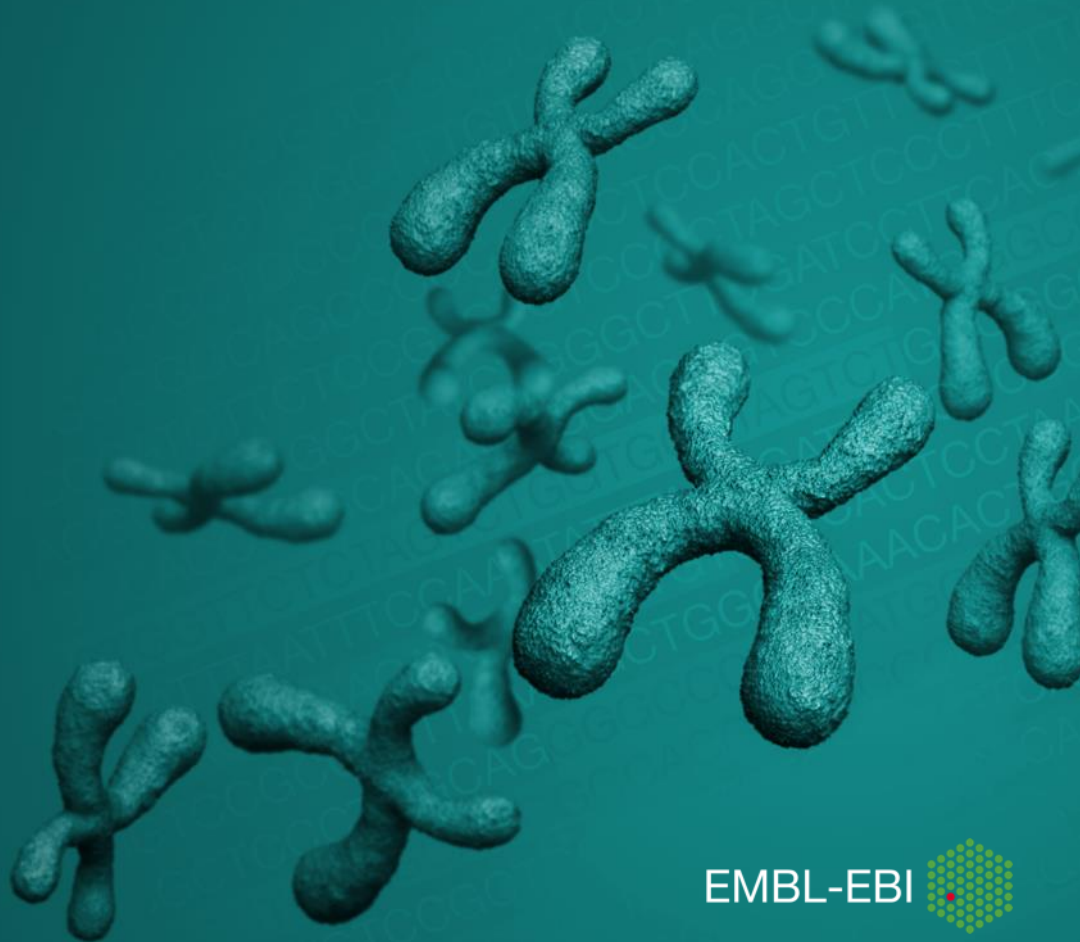
● **DGVa requires**

- *Project* – short description of the project or study
- *Sample* – description of each used sample
- *SampleSets* – description of sample grouping
- *Experiment* – experiment type and platform details
- *Variant Calls* – supporting structural variants observed in individual samples.
- *Variant regions* – submitter asserted regions, one or more variants calls are supporting as evidence.

● **The logic requires submitter to define each experiment and then to describe the Variant Calls using these experiments.**

http://www.ebi.ac.uk/dgva/sites/ebi.ac.uk.dgva/files/documents/dgvasubmissionnotes_v2.7.pdf

European Variation Archive (EVA)



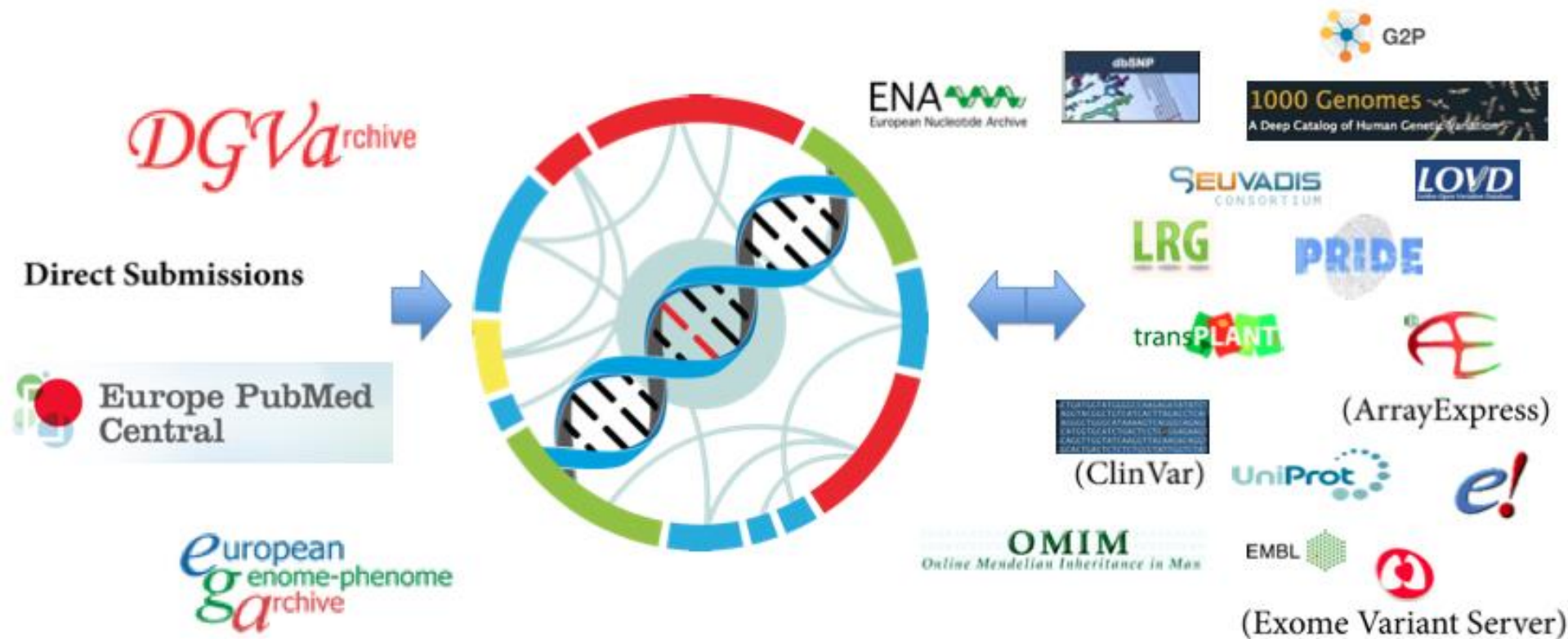
European Variation Archive (EVA)

• **A new EMBL-EBI service for all types of fully public genetic variation data from all species.**

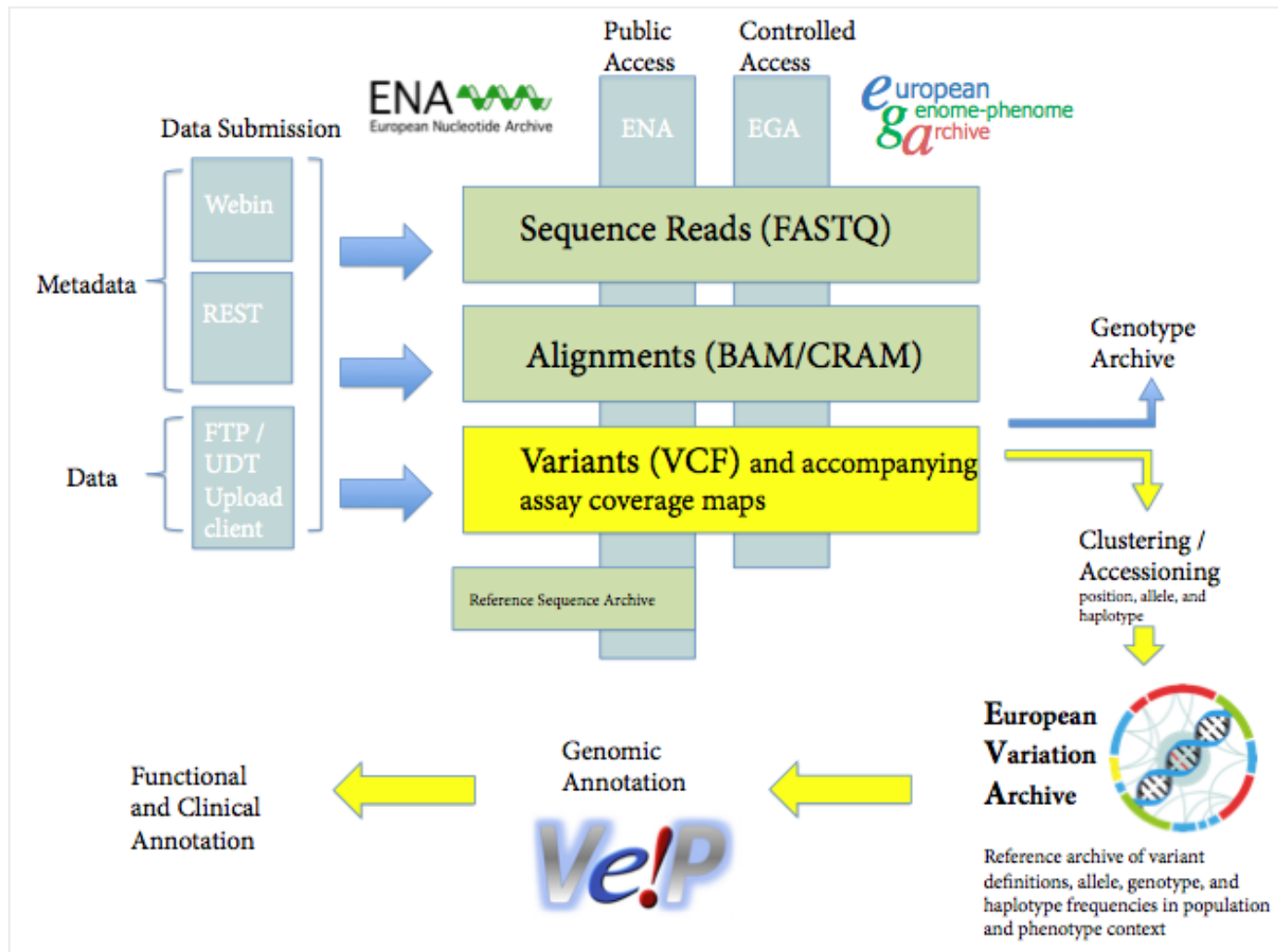
- Beta release April 2014 – first official release scheduled to June.
- Accepting submissions in VCF format. Associated files welcomed (e.g. bed, ped)
- Data dissemination in VCF or TSV format.
- Browse data – variant and study specific views supported by our variant browser
- SNPs are accessioned jointly with dbSNP (NCBI, USA)
- Structural variants are accessioned by DGVA at the EBI.



EVA collaborations



EVA Data Flow Diagram



EVA website



European Variation Archive

Search

Examples: BRAF, 3:1000000-1200000

- Home
- Submit Data
- Variant Browser
- File Browser
- About EVA
- Support & Feedback

BETA Version

This website is still in development. Please send all feedback to eva-helpdesk@ebi.ac.uk, thank you.

EVA - genetic variation at all scales

The European Variation Archive is a database that accepts submission of, and provides access to, all types of genetic variation data from all species. All users are able to download any dataset, or query our study catalogue via our variation table. Access to EVA data is also provided by RESTful web services for a variety of applications, such as annotation pipelines.

Statistics

Species	<input data-bbox="1773 495 1808 521" type="button" value="+"/>
Variants	<input data-bbox="1773 529 1808 555" type="button" value="-"/>
Homo Sapiens (153850312)	
Total (153850312)	
Projects	<input data-bbox="1773 618 1808 644" type="button" value="-"/>
Homo Sapiens (6)	
Analyses	<input data-bbox="1773 678 1808 704" type="button" value="-"/>
Homo Sapiens (80)	

Submit Data

SUBMIT: EVA welcomes direct submission of all types of genetic variation from all species

Access Data

DOWNLOAD: All of our data is open-access and can be downloaded

BROWSE: Our variant catalogue is searchable via our variation table

PLUG-IN: All EVA data available via RESTful web services

News

Tweets



Gary Saunders
@EBIvariation

24 Feb

EMBL-EBI plans to launch a new variation database that shall archive all variants from all species: European Variant Archive (EVA)

Tweet to @EBIvariation


Related Projects

All data submitted to EVA shall be available at dbSNP and vice versa.

Please contact eva-helpdesk@ebi.ac.uk for more details on this collaboration.

Additionally, EVA data is shared with Ensembl Variation, COSMIC, 1000Genomes, LOVD and [many others](#)

EVA Submission page



European Variation Archive

Examples: BRAF, 3:1000000-1200000

[Home](#) [Submit Data](#) [Variant Browser](#) [File Browser](#) [About EVA](#) [Support & Feedback](#)

European Variation Archive submissions

EVA follows the infrastructure of fellow EMBL-EBI resources European Nucleotide Archive ([ENA](#)) and European Genome-phenome Archive ([EGA](#)) to accept, archive, and accession [VCF](#) files. Submissions consist of VCF file(s) and metadata that describe sample(s), experiment(s), and analysis that produced the variant and/or genotype call(s).

Key stages of EVA submissions

Contact



Contact the EVA Helpdesk via [this webform](#) in order to provide details of your submission.

Receive



Receive your submission pack, which will include:

- i) Details for your submission uploads
- ii) [Templates](#) to capture your associated metadata
- iii) Key stages for your submission

Submit




Upload your data files to your private submission upload account or directly to the [EVA helpdesk](#).

Document



Provide details of your study, samples, experiments, runs/analysis, policy and datasets

Start submission by filling a form



European Variation Archive

Examples: BRAF, 3:1000000-1200000

[Home](#) | [Submit Data](#) | [Variant Browser](#) | [File Browser](#) | [About EVA](#) | [Support & Feedback](#)

Submission Form

User Details:

Full name *

E-mail *

Institution/Company Name *

Preferred Centre Acronym (subject to availability) *

Webpage

Country of Origin *

Type of Submission

Genomic DNA

Exonic DNA

Transcribed RNA

Unknown

Other

Comments

Submit form to EVA help-desk

The screenshot shows the EVA Submission Form interface. At the top, there is a navigation bar with the EVA logo and the text "European Variation Archive". To the right of the logo is a search bar with a "Search" button and the text "Examples: BRAF, 3:1000000-1200000". Below the navigation bar are links for "Home", "Submit Data", "Variant Browser", "File Browser", "About EVA", and "Support & Feedback".

The main content area is titled "Submission Form" and contains the following fields:

- User Details:**
 - Full name *
 - E-mail * (example: yourname@yourinsitution.edu)
 - Institution/Company Name *
 - Preferred Centre Acronym (subject to availability) *
 - Webpage
 - Country of Origin * (- Select -)
- Type of Submission**
 - Genomic DNA
 - Exonic DNA
 - Transcribed RNA
 - Unknown
 - Other
- Comments**
 - Send details (button circled in red)

A green arrow labeled "EMAIL" points from the "Send details" button to a grey box containing the text:

Dear Sir,
I want to submit x, y and z to EVA

To the right of this box is a portrait of Gary Saunders, EVA Submissions, with the text "Gary Saunders, EVA Submissions" below it.

EVA Submission Template - Cover

PLEASE READ FIRST

The aim of this sheet is to facilitate effective completion of this template.

The minimum information required to be completed in this template in order for data to be submitted to EVA is: submitter, sample, method and file names.

However, we encourage our users to submit as much meta-data as possible; such information allows for effective use of the data in future applications and permits efficient archiving of the files and enables dynamic querying of all data in the archive via the search tools at our website (www.ebi.ac.uk/eva).

Please email all questions and feedback to eva-helpdesk@ebi.ac.uk

This template is grouped into four sections, split into worksheets. Each worksheet is preceded by an "INFO" sheet which provides more information and instructions for each column.

Worksheet	Explanation
Project	The objective of this sheet is to gather general information about the Project including submitter, submitting centre, collaborators and publications. Importantly, one project can have more than one analysis.
Sample	Projects consist of analyses that are run on samples. We accept sample information in the form of BioSample, ENA or EGA accession(s). As an alternative to providing individual sample information, we also accept BioSamples sampleset accessions. If you do not have a BioSamples sampleset accession, please provide individual sample accessions. If your samples are not yet accessioned, and are therefore anonymous, please contact eva-helpdesk@ebi.ac.uk to discuss submission.
Analysis	For EVA, each analysis is one vcf file. This sheet allows EVA to link vcf files to a project and to other EVA analyses. Additionally, this worksheet contains experimental meta-data detailing the methodology of each analysis.
Files	Filenames and associated checking data associated with this EVA submission should be entered into this worksheet. Each file should be linked to one, or more, analysis. We accept all common types of file associated with variation data including vcf, cram, tabix, wig, bed, gff, ped and fasta.

Each worksheet contains a number of fields -

Completion of the remaining highlighted in **BOLD** is **REQUIRED**. **GREEN** indicates **EITHER/OR** requirement.

Completion of the remaining fields is optional, however please provide as much information as you can and avoid the use of non-ASCII characters in any fields.

An example of a completed template suitable for EVA submission is available at our website (www.ebi.ac.uk/eva/)

65

66

PLEASE READ FIRST

INFO Project

Project

INFO Sample(s)

Sample(s)

INFO Analysis

Analysis

INFO Files

Files

+



Guidelines for describing Sample(s)

Sample Info sheet

1	Column Header	Data Expected																		
2	Sample Accession	Accession of the sample (BioSamples, ENA or EGA)																		
3	Sampleset Accession	BioSamples sampleset accession if appropriate																		
4	Analysis Alias	Alias of the analysis performed on this sample. Comma separated list allowable for multiple analyses																		
5	Description	Free-text description of the sample																		
6	Gender	Gender of the sample: "Male" or "Female"																		
7	Link(s)	Links to resources related to this sample/sampleset (publication(s), dataset(s), online database(s)). Format DB:ID:LABEL (label optional, a text label to display for the link), or URL:LABEL (URL must start with "ftp:" or "http:" or "file:"). Comma separated list allowed for multiple links																		
8	Attribute(s)	Comma separated list of TAG:VALUE:UNITS (Units optional), e.g. AGE:25:Years																		
9	Phenotype(s)	Phenotype(s) of the sample/sampleset, in the form DB:ID, where DB is one of "ClinVar", "HPO", "MedGen", "MeSH", "OMIM"																		
10	Disease Site(s)	Site(s) of the disease in the subject																		
11	Strain	Strain of the subject																		
12	Breed	Breed of the subject																		
13																				

Example of how to provide the sample information

1	Sample Accession	Sampleset Accession	Analysis Alias	Description	Gender	Link(s)	Attribute(s)	Phenotype(s)	Disease Site(s)	Strain	Breed
3	SAMEA2417918	SAMEG171733	1								
4	SAMEA2417921	SAMEG171733	1								
5	SAMEA2417547	SAMEG171733	1								
6	SAMEA2417532	SAMEG171733	1								
7	SAMEA2417503	SAMEG171733	1								
8	SAMEA2417510	SAMEG171733	1								
9	SAMEA2417473	SAMEG171733	1								
10	SAMEA2417483	SAMEG171733	1								
11	SAMEA2417491	SAMEG171733	1								
12	SAMEA2417455	SAMEG171733	1								
13	SAMEA2417459	SAMEG171733	1								
14	SAMEA2417419	SAMEG171733	1								
15	SAMEA2417910	SAMEG171733	1								

EVA submission guidelines


Example of how to provide information about the analysis process

2	Analysis Title	Title of the analysis
3	Analysis Alias	Shortened identifier for the analysis
4	Description	Description of the analysis
5	Project Title	Title of the project to which this analysis belongs
6	Experiment Type	Choose 1 of the following "whole genome sequencing", "Exome sequencing", "Genotyping by array", "Curation"
7	Reference	Reference the analysis was performed against. GRC reference name or ENA accession accepted
8	Platform	Enter the platform used in the analysis
9	Software	Enter the software used in the analysis
10	Imputation	Enter '1' if this was an imputation analysis
11	Centre	Centre which performed the analysis
12	Date	Date the analysis was performed
13	Link(s)	Link(s) to external resources related to this analysis in the form DB:ID:LABEL. Comma separated list allowed for multiple links
14	Run Accession(s)	Associated ENA run accession(s) if applicable (e.g. SRR576651, SRR576652)

Example of how to provide file information

Analysis Title	ID of the analysis that produced the file
File Name	File name
File Type	File type from the following list "vcf", "vcf_aggregate", "readme_file", "phenotype_file", "cram", "tabix", "wig", "bed", "gff", "fasta", "other"
MD5	MD5 value of the file

How to download data from EVA



European Variation Archive

Examples: BRAF, 3:1000000-1200000

[Home](#) | [Submit Data](#) | [Variant Browser](#) | [File Browser](#) | [About EVA](#) | [Support & Feedback](#)

BETA Version

This website is still in development. Please send all feedback to eva-helpdesk@ebi.ac.uk, thank you.

EVA - genetic variation at all scales

The European Variation Archive is a database that accepts submission of, and provides access to, all types of genetic variation data from all species. All users are able to download any dataset, or query our study catalogue via our variation table. Access to EVA data is also provided by RESTful web services for a variety of applications, such as annotation pipelines.

Statistics

Species	<input type="button" value="+"/>
Variants	<input type="button" value="-"/>
Homo Sapiens (153850312)	
Total (153850312)	
Projects	<input type="button" value="-"/>
Homo Sapiens (6)	
Analyses	<input type="button" value="-"/>
Homo Sapiens (80)	

Submit Data

SUBMIT: EVA welcomes direct submission of all types of genetic variation from all species

Access Data

DOWNLOAD: All of our data is open-access and can be downloaded

BROWSE: Our variant catalogue is searchable via our variation table

PLUG-IN: All EVA data available via RESTful web services

News

Tweets



Gary Saunders @EBIvariation 24 Feb

EMBL-EBI plans to launch a new variation database that shall archive all variants from all species: European Variant Archive (EVA)

Tweet to @EBIvariation

Related Projects

All data submitted to EVA shall be available at dbSNP and vice versa.

Please contact eva-helpdesk@ebi.ac.uk for more details on this collaboration.

Additionally, EVA data is shared with Ensembl Variation, COSMIC, 1000Genomes, LOVD and [many others](#)

How to download data from EVA

Variant Browser

Species * Project Info columns to Display

Location *

Enter a location(s), e.g. 1:1000000-1200000, gene name(s) (e.g. brca1, brca2), or id(s) (e.g. ENSG00000139618, rs77475411) to search for.

Download

Show entries

CHROM	POS	ID	REF	ALT	QUAL	FILTER	ANALYSIS
7	140423973	rs150528438	T	C	100	PASS	ERZ015711
7	140424082	rs114228823	C	T	100	PASS	ERZ015711
7	140424085	rs192388879	G	C	100	PASS	ERZ015711
7	140424099	rs138474029	C	A	100	PASS	ERZ015711
7	140424386	rs149292777	A	G	100	PASS	ERZ015711
7	140424582	rs2930322	G	C	100	PASS	ERZ015711
7	140424890	rs79550658	T	C	100	PASS	ERZ015711
7	140424949	rs185077298	C	T	100	PASS	ERZ015711
7	140424968	rs188275729	G	A	100	PASS	ERZ015711
7	140424979	rs180985059	C	G	100	PASS	ERZ015711

How to download data from EVA

Variant Browser

Species * Project Info columns to Display

Location *

Enter a location(s), e.g. 1:1000000-1200000, gene name(s) (e.g. brca1, brca2), or id(s) (e.g. ENSG00000139618, rs77475411) to search for.

[Download](#)

Show entries

Quick download of sliced data

CHROM	POS	ID	REF	ALT	QUAL	FILTER	ANALYSIS
7	140423973	rs150528438	T	C	100	PASS	ERZ015711
7	140424082	rs114228823	C	T	100	PASS	ERZ015711
7	140424085	rs192388879	G	C	100	PASS	ERZ015711
7	140424099	rs138474029	C	A	100	PASS	ERZ015711
7	140424386	rs149292777	A	G	100	PASS	ERZ015711
7	140424582	rs2930322	G	C	100	PASS	ERZ015711
7	140424890	rs79550658	T	C	100	PASS	ERZ015711
7	140424949	rs185077298	C	T	100	PASS	ERZ015711
7	140424968	rs188275729	G	A	100	PASS	ERZ015711
7	140424979	rs180985059	C	G	100	PASS	ERZ015711

How to download data from EVA

Variant Browser

Species * Human Project PRJEB4019 - 1000 Genomes Phase 1 Analysis Info columns to Display Select option(s)

Location * BRAF

Enter a location(s), e.g. 1:1000000-1200000, gene name(s) (e.g. brca1, brca2), or id(s) (e.g. ENSG00000139618, rs77475411) to search for.

Download VCF TSV

Add another filter Search

Show 10 entries

CHROM	POS	ID	REF	ALT	QUAL	FILTER	ANALYSIS
7	140423973	rs150528438	T	C	100	PASS	ERZ015711
7	140424082	rs114228823	C	T	100	PASS	ERZ015711
7	140424085	rs192388879	G	C	100	PASS	ERZ015711
7	140424099	rs138474029	C	A	100	PASS	ERZ015711
7	140424386	rs149292777	A	G	100	PASS	ERZ015711
7	140424582	rs2930322	G	C	100	PASS	ERZ015711
7	140424890	rs79550658	T	C	100	PASS	ERZ015711
7	140424949	rs185077298	C	T	100	PASS	ERZ015711
7	140424968	rs188275729	G	A	100	PASS	ERZ015711
7	140424979	rs180985059	C	G	100	PASS	ERZ015711

external dbSNP identifiers

Direct link to ENA analysis object

Variant browser: single variant analysis

7	140424979	rs180985059	C	G	100	PASS	ERZ015711
---	-----------	-------------	---	---	-----	------	-----------

Showing 1 to 10 of 2,000 entries

First Previous **1** 2 3 4 5 Next Last

Variant Effects (click on a row in table above to see variant effects)

Show entries

Chromosome	Position	ConsequenceType	SNPID	AminoacidChange	GeneID	TranscriptID	FeatureID	FeatureName	FeatureType	FeatureBiotype
7	140424979	SO:0001633			ENSG00000090266	ENST00000476279	ENST00000476279	NDUFB2	downstream	protein_coding
7	140424979	SO:0001633			ENSG00000090266	ENST00000461457	ENST00000461457	NDUFB2	downstream	protein_coding
7	140424979	SO:0001624			ENSG00000157764	ENST00000496384	ENST00000496384	BRAF	3_prime_utr	protein_coding
7	140424979	SO:0001791			ENSG00000157764	ENST00000496384	ENSE00001847804	BRAF	exon	protein_coding
7	140424979	SO:0000694	rs180985059				rs180985059		snp	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	

Variant browser: single variant analysis

7	140424979	rs180985059	C	G	100	PASS	ERZ015711
---	-----------	-------------	---	---	-----	------	-----------

Showing 1 to 10 of 2,000 entries

First Previous **1** 2 3 4 5 Next Last

Variant Effects (click on a row in table above to see variant effects)

Show entries

Chromosome	Position	ConsequenceType	SNPID	AminoacidChange	GeneID	TranscriptID	FeatureID	FeatureName	FeatureType	FeatureBiotype
7	140424979	SO:0001633			ENSG00000090266	ENST00000476279	ENST00000476279	NDUFB2	downstream	protein_coding
7	140424979	SO:0001633			ENSG00000090266	ENST00000461457	ENST00000461457	NDUFB2	downstream	protein_coding
7	140424979	SO:0001624			ENSG00000157764	ENST00000496384	ENST00000496384	BRAF	3_prime_utr	protein_coding
7	140424979	SO:0001791			ENSG00000157764	ENST00000496384	ENSE00001847804	BRAF	exon	protein_coding
7	140424979	SO:0000694	rs180985059				rs180985059		snp	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	



Ensembl based transcript focused

Variant browser: single variant analysis

7	140424979	rs180985059	C	G	100	PASS	ERZ015711
---	-----------	-------------	---	---	-----	------	-----------

Showing 1 to 10 of 2,000 entries

First Previous **1** 2 3 4 5 Next Last

Variant Effects (click on a row in table above to see variant effects)

Show 10 entries

Chromosome	Position	ConsequenceType	SNPID	AminoacidChange	GeneID	TranscriptID	FeatureID	FeatureName	FeatureType	FeatureBiotype
7	140424979	SO:0001633			ENSG00000090266	ENST00000476279	ENST00000476279	NDUFB2	downstream	protein_coding
7	140424979	SO:0001633			ENSG00000090266	ENST00000461457	ENST00000461457	NDUFB2	downstream	protein_coding
7	140424979	SO:0001624			ENSG00000157764	ENST00000496384	ENST00000496384	BRAF	3_prime_utr	protein_coding
7	140424979	SO:0001791			ENSG00000157764	ENST00000496384	ENSE00001847804	BRAF	exon	protein_coding
7	140424979	SO:0000694	rs180985059				rs180985059		snp	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	



Ensembl based transcript focused



ENCODE regulatory data information

Variant browser: single variant analysis

7	140424979	rs180985059	C	G	100	PASS	ERZ015711
---	-----------	-------------	---	---	-----	------	-----------

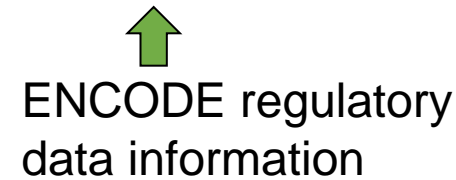
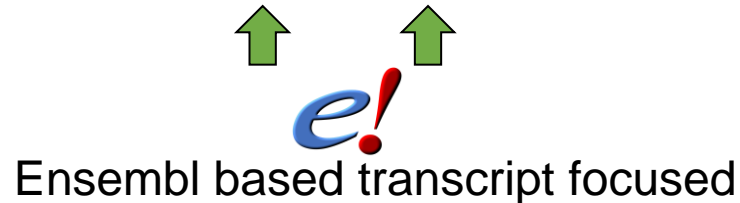
Showing 1 to 10 of 2,000 entries

First Previous **1** 2 3 4 5 Next Last




Variant Effects (click on a row in table above to see variant effects)

Show 10 entries

Chromosome	Position	ConsequenceType	SNPID	AminoacidChange	GeneID	TranscriptID	FeatureID	FeatureName	FeatureType	FeatureBiotype
7	140424979	SO:0001633			ENSG00000090266	ENST00000476279	ENST00000476279	NDUFB2	downstream	protein_coding
7	140424979	SO:0001633			ENSG00000090266	ENST00000461457	ENST00000461457	NDUFB2	downstream	protein_coding
7	140424979	SO:0001624			ENSG00000157764	ENST00000496384	ENST00000496384	BRAF	3_prime_utr	protein_coding
7	140424979	SO:0001791			ENSG00000157764	ENST00000496384	ENSE00001847804	BRAF	exon	protein_coding
7	140424979	SO:0000694	rs180985059				rs180985059		snp	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	
7	140424979	SO:0001566					H3K36me3		regulatory_region	



Any questions

-  View all genes
-  View the GRIK1 gene homepage
-  View graphs about the GRIK1 gene database

Chromosome	GRIK1
Chromosomal band	glutamate receptor, ionotropic, kainate 1
Imprinted	21
Genomic reference	q22
Transcript reference	Unknown
Associated with diseases	NC_000021.8
Citation reference(s)	NM_000830.3
Curators (0)	-
Total number of public variants reported	-
Unique public DNA variants reported	178
Individuals with public variants	178

[Edit variant entry](#) | [Add variant description to additional transcript](#) | [Delete variant entry](#) | [Search public LOVDs](#)

Graphical displays and utilities	
Graphs	Graphs displaying summary information of all variants in the database »
UCSC Genome Browser	Show variants in the UCSC Genome Browser (full view , compact view)
Ensembl Genome Browser	Show variants in the Ensembl Genome Browser (full view , compact view)
NCBI Sequence Viewer	Show distribution histogram of variants in the NCBI Sequence Viewer

Links to other resources	
HGNC	4579
Entrez Gene	2897
OMIM - Gene	138245

Active transcripts

ID	Chr	Name	NCBI ID	NCBI Protein ID	Variants
07123	21	transcript variant 1	NM_000830.3	NP_000821.1	178

Uses LOVD2 API

Coming soon:

Search EVA

Filter By

SNP id:

Region:

21:30909254-31312351

Gene:

GRIK1

Study

Variant Data

Chromosome	Start	End	ID	Type	REF/ALT	HGVS Name
21	30909266	30909266	rs8129935	SNV	A/T	21:g.30909266A>T
21	30909492	30909492	rs144918094	SNV	C/T	21:g.30909492C>T
21	30909620	30909620	rs73897668	SNV	C/T	21:g.30909620C>T
21	30909640	30909640	rs116002269	SNV	T/C	21:g.30909640T>C
21	30909691	30909691	rs151335244	SNV	C/A	21:g.30909691C>A
21	30909751	30909751	rs201908048	SNV	G/T	21:g.30909751G>T
21	30909759	30909759	rs1571681	SNV	C/T	21:g.30909759C>T
21	30909814	30909814	rs75793187	SNV	A/C	21:g.30909814A>C
21	30909845	30909845	rs189033596	SNV	C/G	21:g.30909845C>G
21	30909850	30909850	rs192979876	SNV	C/G	21:g.30909850C>G

FileID	StudyID	Attributes							
		QUAL	FILTER	ERATE	AN	AA	AC	SNPSOURCE	AF
chr21	1000g	100.0	PASS	0.0003	2184	C	20	LOWCOV,EXOME	0.01

Stats	
MAF	0.009157509543001652
MGF	0.0009157509193755686
Allele MAF	T
Genotype MAF	1 1
miss Allele	0
miss Genotypes	0
Mendel Err	0
Cases Percent Dominant	
Controls Percent Dominant	
Cases Percent Recessive	
Controls Percent Recessive	

Genotype Count

