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
- Basic research
- Administration
- EMBO

Hamburg
- Structural biology

Hinxton, Cambridge
- Bioinformatics

Grenoble
- Structural biology

Monterotondo, Rome
- Mouse biology

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

Gene, protein & metabolite expression
- ArrayExpress
- Expression Atlas
- Metabolights
- PRIDE

Literature & ontologies
- Europe PubMed Central
- Gene Ontology
- Experimental Factor Ontology

Protein sequences, families & motifs
- 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.

![Studies in the EGA by technology](image-url)
How can I access data stored in EGA

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

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

https://www.ebi.ac.uk/ega
How can I access data stored in EGA

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

Data Access Committees

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

https://www.ebi.ac.uk/ega
How can I access data stored in EGA

https://www.ebi.ac.uk/ega
How can I access data stored in EGA

https://www.ebi.ac.uk/ega
How can I access data stored in EGA
How can I access data stored in EGA

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)

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: EGAD00001000124
Technology: Illumina HiSeq 2000
Type: 3
Samples: Sequencing Acute Myeloid Leukaemia

Publications
How can I access data stored in EGA

DATASET: Sequencing Acute Myeloid Leukaemia

Dataset ID | Technology  | Samples
-----------|-------------|---------
EGAD000001000124 | 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.

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Study Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGAS000001000035</td>
<td>Sequencing Acute Myeloid Leukaemia</td>
</tr>
</tbody>
</table>

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:

WTSG CGP Data access committee

Access Information:
http://www.ebi.ac.uk/ega/docs/EGAC000001000000

Contact Person: Giselle Kerry
Email: gh2@cancer.ac.uk
Data Access applied directly from DAC
How can I access data stored in EGA

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

Data Access Committees

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

https://www.ebi.ac.uk/ega
How can I access data stored in EGA

Login

Login with Local Account

Username * myusername
Enter your European Genome-phenome Archive username.

Password * *******
Enter the password that accompanies your username.

Log in

Login with Federated Identity

Use a suggested selection:

Cartia Oy

Or enter your organization's name

Continue

Allow me to pick from a list
Help
How can I access data stored in EGA
How can I access data stored in EGA

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

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Study Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGAS00001000035</td>
<td>Sequencing Acute Myeloid Leukaemia</td>
</tr>
</tbody>
</table>
Secure EGA Downloader

1. Log in/select transfer protocol
2. Create key
3. Select destination directory
4. Filter and select files
5. Download!
Secure EGA Downloader

Specify key

Select directories

Select files

Decrypt!

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

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

Webin-UDT

[Chart showing transfer rates for different locations, including US (Santa Cruz CA), Sweden, India, China (Hong Kong), and Australia (Melbourne). The chart compares UDT and FTP transfer rates.]

[Link: 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

DATA TYPE
- Sequence
  - Raw-unaligned
- Variants
  - VCF
- Array based
- Phenotypes
- Structural variants

PUBLIC ACCESS
- ENA
- EVA

CONTROLLED ACCESS
- ArrayExpress
- BioSamples
- DGVarchive
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 times. 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.

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 times. 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
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.

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

How to initiate submission to EGA?

- 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
How to initiate submission to EGA?

- 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.
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>

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

File submission tools

Submission accounts

Staging area

EGA security

Meta information submission tools

Public information
EGA data submission tools

File submission tools

Meta information submission tools

processing will only start when meta data is in place

Submission accounts

Staging area

EGA security

Public information
Processing submitted file into the archive

File submission tools

Submission accounts

Staging area

EGA security

encryption using EGA public key

File validation based on md5 checksums and re-encryption using archive key

General EBI file archiving process

A

B
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


[https://www.ebi.ac.uk/ega/submission/sequence](https://www.ebi.ac.uk/ega/submission/sequence)
[https://www.ebi.ac.uk/ena/about/cram_toolkit](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/ega/submission/sequence)
[https://www.ebi.ac.uk/ena/about/cram_toolkit](https://www.ebi.ac.uk/ena/about/cram_toolkit)
Supported file formats

- **SEQUENCE**
  - Raw unaligned reads
  - Aligned BAM reads
  - VCF

- **ARRAY BASED**
  - Raw files
  - Intensity files
  - Analysis files
  - All formats supported

- **PHENOTYPES**

Additional 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/
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)

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

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.
- Programmatical 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 metadata 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

- **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

ENA and EGA shared requirements

EGA specific requirements
Relationships between meta data objects

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

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

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

- Samples
- Array packets
- Dataset
- Policy
- DAC

Register using Webin
Register using Array Format (AF) template

All objects may be registered prior to file upload
Array-based submission spreadsheet – Referencing correct accessions

<table>
<thead>
<tr>
<th>Study Accession Number</th>
<th>EGAS00001000527</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submission box</td>
<td>ega-box-100</td>
</tr>
</tbody>
</table>

Data deposited outside of EGA (e.g. ArrayExpress or European Nucleotide Archive)

<table>
<thead>
<tr>
<th>Study Accession Number</th>
<th>EGAC00001000114</th>
</tr>
</thead>
<tbody>
<tr>
<td>Policy accession number/s</td>
<td>EGAP00001000115</td>
</tr>
</tbody>
</table>

**Figure 1:** Referencing accessions generated through Webin

Should your submission require multiple DAC’s or policies, use ‘;’ to separate the accession numbers.
Array-based submission spreadsheet – Describing samples

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Human_1</td>
<td>sample_1a1</td>
<td>Cell line derived from lung tissue</td>
<td>Human</td>
<td>Lung</td>
<td>EFO_0000934</td>
<td>Lung-1</td>
<td>Male</td>
<td>East England</td>
<td>Case</td>
<td>Normal</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human_1</td>
<td>sample_1a2</td>
<td>Cell line derived from lung tissue</td>
<td>Human</td>
<td>Lung</td>
<td>EFO_0000934</td>
<td>Lung-1</td>
<td>Male</td>
<td>East England</td>
<td>Control</td>
<td>Lung carcinoma</td>
<td>EFO_0001461</td>
<td></td>
</tr>
</tbody>
</table>

Attributes to associate to each sample. All samples MUST include gender information (Male/Female/Unknown) and phenotype information critical for further analysis.
Array-based submission spreadsheet – Creating dataset for the files

- **Name your dataset**
- **Add short descriptive title**
- **Platform used to generate data files**
- **Number of unique samples in dataset**
- **Accession number of Policy to be affiliated to dataset**

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Dataset accession number</th>
<th>Description</th>
<th>Data type</th>
<th>Technology</th>
<th>Case or Control</th>
<th>Number of samples</th>
<th>DMC accession number</th>
<th>Policy accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matched_normal</td>
<td></td>
<td>Matched normal controlled</td>
<td>Genotype</td>
<td>Illumina</td>
<td>Control</td>
<td>1</td>
<td>EGAC000010000114</td>
<td>EGA000001000115</td>
</tr>
<tr>
<td>Case</td>
<td></td>
<td>Lung carcinoma case</td>
<td>Genotype</td>
<td>Illumina</td>
<td>Case</td>
<td>1</td>
<td>EGAC000010000115</td>
<td>EGAP000001000116</td>
</tr>
</tbody>
</table>

If referring to an existing dataset add the accession, otherwise leave blank.

- Free text field for type of data submitted
- Add ‘control’ or ‘case’ value for each dataset
- Accession number of DAC to be affiliated to dataset

**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

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Platform</th>
<th>Raw data File (e.g. CEL/BDAT)</th>
<th>File format (Raw data file)</th>
<th>File format (Signal intensity file)</th>
<th>Genotype calling software</th>
<th>Genotype/Expression file</th>
<th>File format (Genotype/Expression)</th>
<th>Additional files</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample_1a1</td>
<td>Illumina_550K</td>
<td>1.cel.gpg</td>
<td>CEL</td>
<td>1_signal.txt.gpg</td>
<td>illumina</td>
<td>1_genotypes.txt.gpg</td>
<td>TXT</td>
<td>Read_me.gpg; Sample_file.gpg</td>
<td>Matched_normal</td>
</tr>
<tr>
<td>sample_1a2</td>
<td>Illumina_550K</td>
<td>2.cel.gpg</td>
<td>CEL</td>
<td>2_signal.txt.gpg</td>
<td>illumina</td>
<td>2_genotypes.txt.gpg</td>
<td>TXT</td>
<td>Read_me.gpg; Sample_file.gpg</td>
<td>Case</td>
</tr>
</tbody>
</table>

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) that SNPs in human genome
Detecting Structural Variants - Sequencing

- Depending on method, sequencing can give bp resolution
- Genotype can be determined
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)
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
Example 1: Assertion method 50% overlap
Example 2: Assertion method 100% overlap
DGVA will create regions if required
DGVA Data Flow Diagram

Direct submissions

DGVA archive

Curation from literature

dbVar

Database of Genomic Variants


www.ensembl.org/

www.ensembl.org/biomart/martview

http://projects.tcag.ca/variation/

www.ebi.ac.uk/dgva/data-download
<table>
<thead>
<tr>
<th>Study</th>
<th>Reference</th>
<th>Organism</th>
<th>Variants</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>estd209</td>
<td>Pang et al. 2013b</td>
<td>Homo sapiens</td>
<td>471817</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd208</td>
<td>Hebig et al. 2013</td>
<td>Homo sapiens</td>
<td>81</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd205</td>
<td>Zichner et al. 2012</td>
<td>Drosophila melanogaster</td>
<td>65487</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd204</td>
<td>Simon et al. 2013</td>
<td>Mus musculus</td>
<td>43</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd203</td>
<td>Vogler et al. 2010</td>
<td>Homo sapiens</td>
<td>4193</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd201</td>
<td>Wong et al. 2012b</td>
<td>Homo sapiens</td>
<td>36558</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd200</td>
<td>Wong et al. 2012</td>
<td>Homo sapiens</td>
<td>30044</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd199</td>
<td>1000 Genomes Consortium Phase 1</td>
<td>Homo sapiens</td>
<td>22531</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd198</td>
<td>Chia et al. 2012</td>
<td>Homo sapiens</td>
<td>381</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd197</td>
<td>Mokeman et al. 2009</td>
<td>Homo sapiens</td>
<td>232775</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd196</td>
<td>Simon-Sanchez et al. 2007</td>
<td>Homo sapiens</td>
<td>335</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd195</td>
<td>Atsuchi et al. 2010</td>
<td>Homo sapiens</td>
<td>856</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd194</td>
<td>Bentley et al. 2008</td>
<td>Homo sapiens</td>
<td>504912</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd193</td>
<td>Feuk et al. 2005</td>
<td>Homo sapiens</td>
<td>3</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd192</td>
<td>COSMID</td>
<td>Homo sapiens</td>
<td>15168</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd188</td>
<td>Pinto et al. 2011</td>
<td>Homo sapiens</td>
<td>60247</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd186</td>
<td>Thevenon et al. 2012</td>
<td>Homo sapiens</td>
<td>3</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd185</td>
<td>Yalcin et al. 2012</td>
<td>Mus musculus</td>
<td>1453</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd180</td>
<td>Pang et al. 2010</td>
<td>Homo sapiens</td>
<td>23887</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd176</td>
<td>Banerjee et al. 2011</td>
<td>Homo sapiens</td>
<td>734</td>
<td>Download via FTP</td>
</tr>
</tbody>
</table>
DGVA download site

### Data Download

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

<table>
<thead>
<tr>
<th>Study</th>
<th>Reference</th>
<th>Organism</th>
<th>Variants</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>estd209</td>
<td>Pang et al. 2013b</td>
<td>Homo sapiens</td>
<td>471817</td>
<td>Download via FTP</td>
</tr>
<tr>
<td>estd208</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd204</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd203</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd201</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd199</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd198</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd197</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd196</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estd195</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Index of /pub/databases/dgva/estd209_Pang_et_al_2013b/gvf/**

- [parent directory]
- estd209_Pang_et_al_2013b.2014-04-01.GRCh38.Remapped.gvf.gz: 17.5 MB 4/14 8:12:00 PM
- 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_Theverson 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 data in Ensembl
Submissions to DGVA

Contact eva-helpdesk@ebi.ac.uk.

Complete your meta data submission template and email it to the help-desk.

- 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

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.

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
PLUGIN: All EVA data available via RESTful web services

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.
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

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

Send details
Dear Sir,

I want to submit x, y and z to EVA

Gary Saunders, EVA Submissions
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.

<table>
<thead>
<tr>
<th>Worksheet</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>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.</td>
</tr>
<tr>
<td>Sample</td>
<td>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 <a href="mailto:eva-helpdesk@ebi.ac.uk">eva-helpdesk@ebi.ac.uk</a> to discuss submission.</td>
</tr>
<tr>
<td>Analysis</td>
<td>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.</td>
</tr>
<tr>
<td>Files</td>
<td>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, gff and fasta.</td>
</tr>
</tbody>
</table>

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/)
Guidelines for describing Sample(s)

Sample Info sheet

<table>
<thead>
<tr>
<th>Column Header</th>
<th>Data Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Accession</td>
<td>Accession of the sample (BioSample, EMBL or EGA)</td>
</tr>
<tr>
<td>Sampleset Accession</td>
<td>BioSample sampleset accession if appropriate</td>
</tr>
<tr>
<td>Analysis Alias</td>
<td>Alias of the analysis performed on this sample. Comma separated list allowed for multiple analyses</td>
</tr>
<tr>
<td>Description</td>
<td>Free-text description of the sample</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender of the sample. <em>Male</em> or <em>Female</em></td>
</tr>
<tr>
<td>Link(s)</td>
<td>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 &quot;http:&quot; or &quot;https:&quot;. Comma separated list allowed for multiple links</td>
</tr>
<tr>
<td>Attribute(s)</td>
<td>Comma separated list of TAG:VALUE:UNITS (units optional), e.g. Age:25:Years</td>
</tr>
<tr>
<td>Phenotype(s)</td>
<td>Phenotype(s) of the sample/sampleset, in the form DB:ID, where DB is one of &quot;OMIM&quot;, &quot;HPO&quot;, &quot;MedGen&quot;, &quot;Mesh&quot;, &quot;OMIM&quot;</td>
</tr>
</tbody>
</table>

Strain of the subject

Example of how to provide the sample information

<table>
<thead>
<tr>
<th>Sample Accession</th>
<th>Sampleset Accession</th>
<th>Analysis Alias</th>
<th>Description</th>
<th>Gender</th>
<th>Link(s)</th>
<th>Attribute(s)</th>
<th>Phenotype(s)</th>
<th>Disease Site(s)</th>
<th>Strain</th>
<th>Breed</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAMEA2417918</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417921</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417547</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417532</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417503</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417510</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417473</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417483</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417491</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417455</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417459</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417419</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMEA2417910</td>
<td>SAMEG171733</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EVA submission guidelines

Example of how to provide information about the analysis process

<table>
<thead>
<tr>
<th>Analysis Title</th>
<th>Title of the analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Alias</td>
<td>Shortened identifier for the analysis</td>
</tr>
<tr>
<td>Description</td>
<td>Description of the analysis</td>
</tr>
<tr>
<td>Project Title</td>
<td>Title of the project to which this analysis belongs</td>
</tr>
<tr>
<td>Experiment Type</td>
<td>Choose 1 of the following: &quot;whole genome sequencing&quot;, &quot;Exome sequencing&quot;, &quot;Genotyping by array&quot;, &quot;Curation&quot;</td>
</tr>
<tr>
<td>Reference</td>
<td>Reference the analysis was performed against. GRC reference name or ENA accession accepted</td>
</tr>
<tr>
<td>Platform</td>
<td>Enter the platform used in the analysis</td>
</tr>
<tr>
<td>Software</td>
<td>Enter the software used in the analysis</td>
</tr>
<tr>
<td>Imputation</td>
<td>Enter '1' if this was an imputation analysis</td>
</tr>
<tr>
<td>Centre</td>
<td>Centre which performed the analysis</td>
</tr>
<tr>
<td>Date</td>
<td>Date the analysis was performed</td>
</tr>
<tr>
<td>Link(s)</td>
<td>Link(s) to external resources related to this analysis in the form DB:ID:LABEL. Comma separated list allowed for multiple links</td>
</tr>
<tr>
<td>Run Accession(s)</td>
<td>Associated ENA run accession(s) if applicable (e.g. SRR576651, SRR576652)</td>
</tr>
</tbody>
</table>

Example of how to provide file information

<table>
<thead>
<tr>
<th>Analysis Title</th>
<th>ID of the analysis that produced the file</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Name</td>
<td>File name</td>
</tr>
<tr>
<td>File Type</td>
<td>File type from the following list: &quot;vcf&quot;, &quot;vcf_aggregate&quot;, &quot;readme_file&quot;, &quot;phenotype_file&quot;, &quot;cram&quot;, &quot;tabix&quot;, &quot;wig&quot;, &quot;bed&quot;, &quot;gff&quot;, &quot;fasta&quot;, &quot;other&quot;</td>
</tr>
<tr>
<td>MDS</td>
<td>MDS value of the file</td>
</tr>
</tbody>
</table>
How to download data from EVA

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.
How to download data from EVA

### Variant Browser

**Species:** Human  
**Project:** EVA84019 - 1000 Genomes Phase 1 Analysis  
**Location:** BRAF

Enter a location(s), e.g. 1:1000000-1200000, gene name(s) (e.g. brcal,brcar2), or id(s) (e.g. ENSG00000139818, rs77457511) to search for.

- **Download** VCF  
- **Download** TSV

#### Table

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>140423973</td>
<td>rs15528438</td>
<td>T</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424082</td>
<td>rs114228823</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424085</td>
<td>rs192388879</td>
<td>G</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424099</td>
<td>rs128474029</td>
<td>C</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424386</td>
<td>rs146207777</td>
<td>A</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424682</td>
<td>rs2603321</td>
<td>G</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424890</td>
<td>rs7350853</td>
<td>T</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424949</td>
<td>rs185077298</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424968</td>
<td>rs166273729</td>
<td>G</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>rs180985059</td>
<td>C</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
</tbody>
</table>
How to download data from EVA

Quick download of sliced data

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>14042973</td>
<td>rs15528430</td>
<td>T</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424082</td>
<td>rs114228823</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424085</td>
<td>rs19236973</td>
<td>G</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424999</td>
<td>rs128474029</td>
<td>C</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424986</td>
<td>rs14620777</td>
<td>A</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424982</td>
<td>rs260322</td>
<td>G</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424899</td>
<td>rs7556583</td>
<td>T</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424949</td>
<td>rs18507298</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424968</td>
<td>rs15627529</td>
<td>G</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>rs180985059</td>
<td>C</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERS215711</td>
</tr>
</tbody>
</table>
How to download data from EVA

Variant Browser

<table>
<thead>
<tr>
<th>CHROM</th>
<th>POS</th>
<th>ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>ANALYSIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>140423973</td>
<td>rs15028438</td>
<td>T</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424082</td>
<td>rs114228823</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424085</td>
<td>rs19238879</td>
<td>G</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424099</td>
<td>rs128474029</td>
<td>C</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424096</td>
<td>rs146202777</td>
<td>A</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424082</td>
<td>rs2630322</td>
<td>G</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424090</td>
<td>rs73550658</td>
<td>T</td>
<td>C</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424949</td>
<td>rs18667298</td>
<td>C</td>
<td>T</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424983</td>
<td>rs15623729</td>
<td>G</td>
<td>A</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>rs18098059</td>
<td>C</td>
<td>G</td>
<td>100</td>
<td>PASS</td>
<td>ERZ015711</td>
</tr>
</tbody>
</table>

Quick download of sliced data

- external dbSNP identifiers
- Direct link to ENA analysis object
### Variant browser: single variant analysis

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>ConsequenceType</th>
<th>SNPID</th>
<th>AminoacidChange</th>
<th>GeneID</th>
<th>TranscriptID</th>
<th>FeatureID</th>
<th>FeatureName</th>
<th>FeatureType</th>
<th>FeatureBiotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001653</td>
<td>rs180985059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001659</td>
<td></td>
<td></td>
<td>ENSG00000002688</td>
<td>ENST00000476279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001659</td>
<td></td>
<td></td>
<td>ENSG00000002688</td>
<td>ENST00000476279</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001654</td>
<td></td>
<td></td>
<td>ENSG00000157754</td>
<td>ENST00000491457</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001751</td>
<td></td>
<td></td>
<td>ENSG00000009684</td>
<td>ENST0000049384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0000864</td>
<td>rs180985059</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>H3K36me3</td>
<td></td>
<td></td>
<td></td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>H3K36me3</td>
<td></td>
<td></td>
<td></td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>H3K36me3</td>
<td></td>
<td></td>
<td></td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>H3K36me3</td>
<td></td>
<td></td>
<td></td>
<td>regulatory_region</td>
<td></td>
</tr>
</tbody>
</table>
Variant browser: single variant analysis

| Chromosome | Position | ConsequenceType | SNPID          | AminoacidChange | GeneID        | TranscriptID | FeatureID | FeatureName | FeatureType         | FeatureBtype   |
|------------|----------|-----------------|----------------|-----------------|---------------|--------------|-----------|-------------|---------------------|----------------|---------------|
| 7          | 140424979| SO:0001633       | rs180985059    |                 |               |              |           |             | downstream          | protein_coding |              |
| 7          | 140424979| SO:0001633       | rs180985059    |                 |               |              |           |             | downstream          | protein_coding |              |
| 7          | 140424979| SO:0001624       | ENSG0000000157768 |                 | ENST00000476279 | ENST00000486384 | NDURF2 | downstream          | protein_coding |              |
| 7          | 140424979| SO:0001751       | ENST000004961457 |                 | ENSG000000157768 | ENSG000000157768 | NDURF2 | downstream          | protein_coding |              |
| 7          | 140424979| SO:0001751       | ENSG000000157768 |                 | ENST000004961457 | ENSG000000157768 | NDURF2 | downstream          | protein_coding |              |

Ensembl based transcript focused
### Variant browser: single variant analysis

ENCODE regulatory data information

Ensembl based transcript focused

---

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position</th>
<th>ConsequenceType</th>
<th>SNPID</th>
<th>AminoacidChange</th>
<th>GeneID</th>
<th>TranscriptID</th>
<th>FeatureID</th>
<th>FeatureName</th>
<th>FeatureType</th>
<th>FeatureBiotype</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001633</td>
<td></td>
<td></td>
<td>ENSG000000090266</td>
<td>ENST000000476279</td>
<td>ENST000000476279</td>
<td>NDUFB2</td>
<td>downstream</td>
<td>protein_coding</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001659</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>downstream</td>
<td>protein_coding</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001524</td>
<td></td>
<td></td>
<td>ENSG000000157764</td>
<td>ENST000000496384</td>
<td>ENST000000496384</td>
<td>BRAF</td>
<td>3_prime_utr</td>
<td>protein_coding</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001791</td>
<td></td>
<td></td>
<td>ENSG000000157764</td>
<td>ENST000000496384</td>
<td>ENST000000496384</td>
<td>BRAF</td>
<td>exon</td>
<td>protein_coding</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001754</td>
<td></td>
<td></td>
<td>ENST000000496384</td>
<td>ENST000000496384</td>
<td>ENST000000496384</td>
<td>BRAF</td>
<td>snp</td>
<td>protein_coding</td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>regulatory_region</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>140424979</td>
<td>SC:0001566</td>
<td></td>
<td></td>
<td>ENSG000000020859</td>
<td>ENST000000491457</td>
<td>ENST000000491457</td>
<td>NDUFB2</td>
<td>regulatory_region</td>
<td></td>
</tr>
</tbody>
</table>

---

Variant Effects (click on a row in table above to see variant effects)
Variant browser: single variant analysis

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

Ensembl based transcript focused
ENCODE regulatory data information
Any questions
Uses LOVD2 API

Coming soon:

Search EVA
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Start</th>
<th>End</th>
<th>ID</th>
<th>Type</th>
<th>REF/ALT</th>
<th>HSVS Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>30909266</td>
<td>30909266</td>
<td>rs8129935</td>
<td>SNV</td>
<td>A/T</td>
<td>21:g.30909266A&gt;T</td>
</tr>
<tr>
<td>21</td>
<td>30909492</td>
<td>30909492</td>
<td>rs44918094</td>
<td>SNV</td>
<td>C/T</td>
<td>21:g.30909492C&gt;T</td>
</tr>
<tr>
<td>21</td>
<td>30909620</td>
<td>30909620</td>
<td>rs73397568</td>
<td>SNV</td>
<td>C/T</td>
<td>21:g.30909620C&gt;T</td>
</tr>
<tr>
<td>21</td>
<td>30909640</td>
<td>30909640</td>
<td>rs116002269</td>
<td>SNV</td>
<td>T/C</td>
<td>21:g.30909640T&gt;C</td>
</tr>
<tr>
<td>21</td>
<td>30909691</td>
<td>30909691</td>
<td>rs151335244</td>
<td>SNV</td>
<td>C/A</td>
<td>21:g.30909691C&gt;A</td>
</tr>
<tr>
<td>21</td>
<td>30909751</td>
<td>30909751</td>
<td>rs201908148</td>
<td>SNV</td>
<td>G/T</td>
<td>21:g.30909751G&gt;T</td>
</tr>
<tr>
<td>21</td>
<td>30909759</td>
<td>30909759</td>
<td>rs1571581</td>
<td>SNV</td>
<td>G/T</td>
<td>21:g.30909759G&gt;T</td>
</tr>
<tr>
<td>21</td>
<td>30909814</td>
<td>30909814</td>
<td>rs75793187</td>
<td>SNV</td>
<td>A/C</td>
<td>21:g.30909814A&gt;C</td>
</tr>
<tr>
<td>21</td>
<td>30909845</td>
<td>30909845</td>
<td>rs186033596</td>
<td>SNV</td>
<td>C/G</td>
<td>21:g.30909845C&gt;G</td>
</tr>
<tr>
<td>21</td>
<td>30909850</td>
<td>30909850</td>
<td>rs192979876</td>
<td>SNV</td>
<td>C/G</td>
<td>21:g.30909850C&gt;G</td>
</tr>
<tr>
<td>FileID</td>
<td>StudyID</td>
<td>QUAL</td>
<td>FILTER</td>
<td>ERATE</td>
<td>AN</td>
<td>AA</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
<td>----</td>
</tr>
<tr>
<td>chr21</td>
<td>1000g</td>
<td>100.0</td>
<td>PASS</td>
<td>0.0003</td>
<td>2184</td>
<td>C</td>
</tr>
</tbody>
</table>

**Stats**

- MAF: 0.009157509543001652
- MGF: 0.0009157509193755686
- Allele MAF: T
- Genotype MAF: 1|1
- miss Allele: 0
- miss Genotypes: 0
- Mendel Err: 0

**Genotype Count**

- 0|0: 98.3%
- 1|0: 0.5%
- 0|1: 1.2%
- 1|1: 0.1%