

Module 4: Public Repositories for NGS data

Presented by:

Marcela Sjöberg Herrera

Assistant professor

Laboratory of Epigenetic Regulation

Faculty of Biological Sciences

Pontificia Universidad Católica de Chile

msjoberg@bio.puc.cl

Based on slides by:

Jacqueline Keane

Next Generation Sequencing Bioinformatics Course

22-27 January 2023 - Santiago - Chile



FACULTAD DE
CIENCIAS BIOLÓGICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

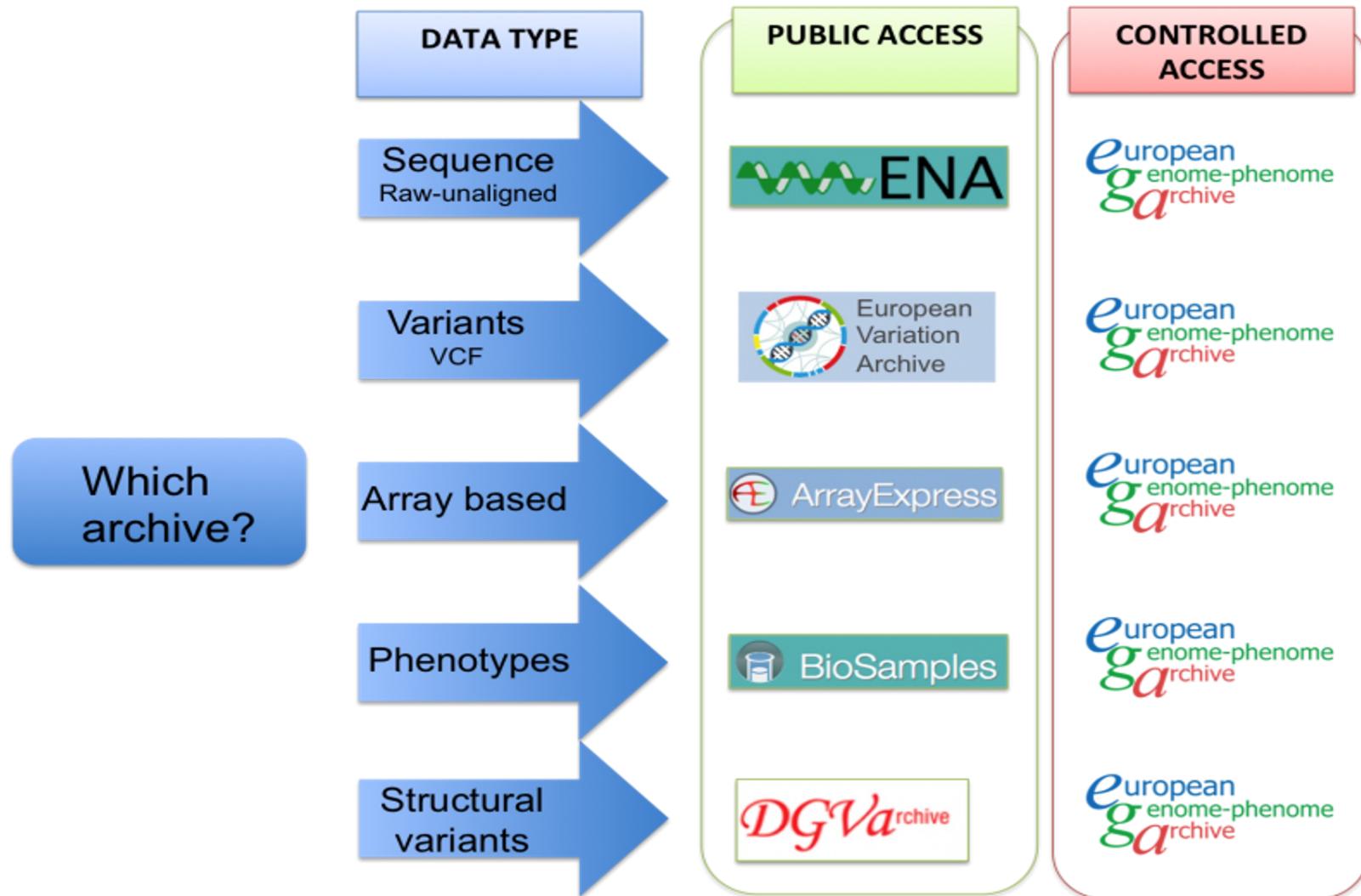
WELLCOME GENOME CAMPUS

CONNECTING SCIENCE
ADVANCED COURSES +
SCIENTIFIC CONFERENCES

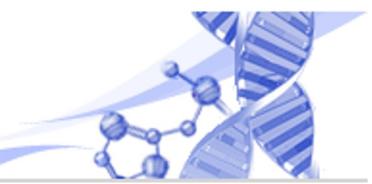
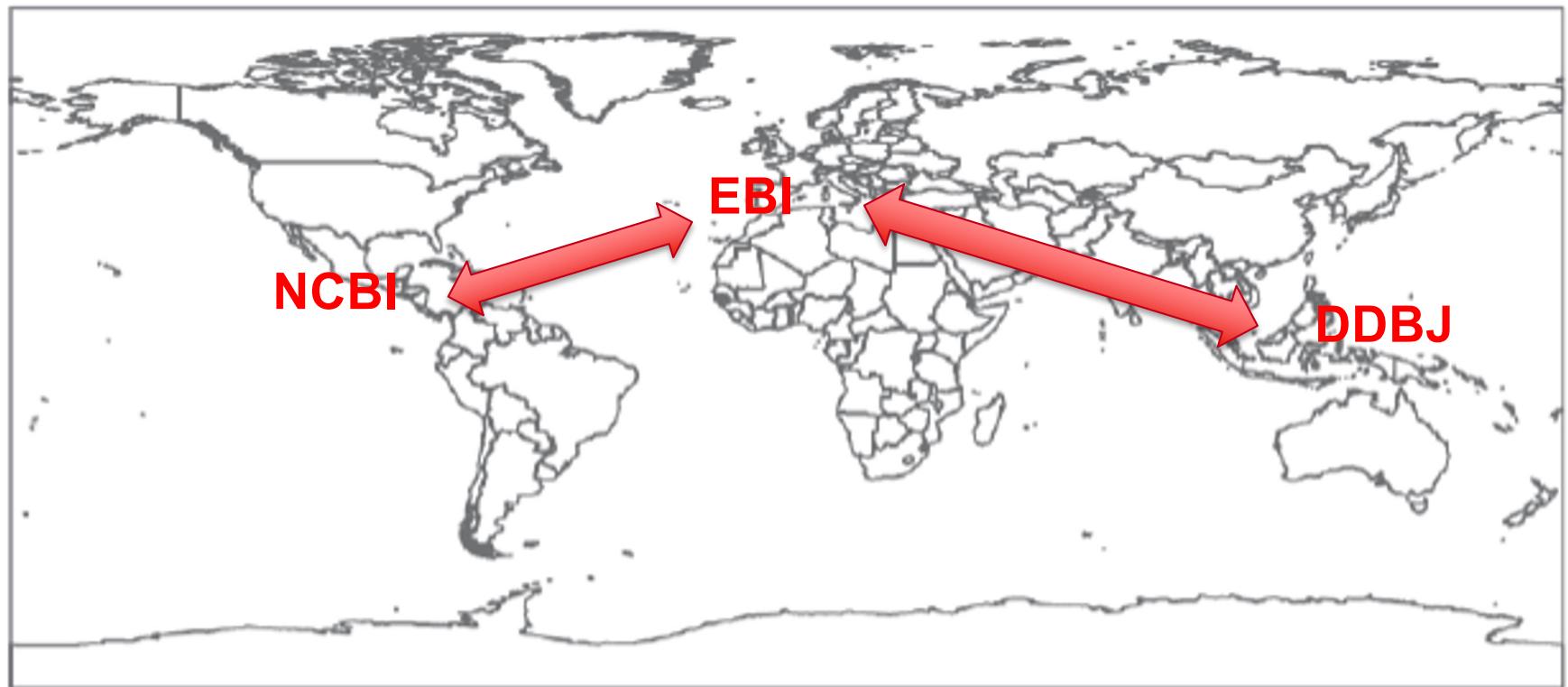
Purpose of data archives (repositories)

- For archiving and distribution of raw and processed data generated by NGS experiments
- Submit your own data that you want to publish
- Retrieve data sets from publications
- Finding data sets that might be relevant to your own research
- Exist different data archives for different data types

Which data archive?



Data sharing across archives



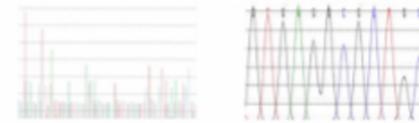
Global data archives

Data Type	DDBJ	EBI	NCBI
Primary Sequence Data	DDBJ Sequence Read Archive (DRA)	European Nucleotide Archive (ENA)	Sequence Read Archive (SRA)
Annotated Sequences	DDBJ	European Nucleotide Archive (ENA)	GenBank
Variation	-	European Variation Archive (EVA)	dbSNP
Structural Variation	-	Genomic Variants Archive (DGVa)	dbVar
Expression	DDBJ Omics Archive (DOR)	ArrayExpress (mostly sequencing)	Gene Expression Omnibus (GEO)
Restricted	Japanese Genome-phenome Archive (JGA)	European Genome-phenome Archive (EGA)	dbGAP
Samples	BioSample	BioSample	BioSample
Studies	BioProject	BioProject	BioProject

European Nucleotide Archive (ENA)

- For data from experiments based on **nucleotide sequencing**

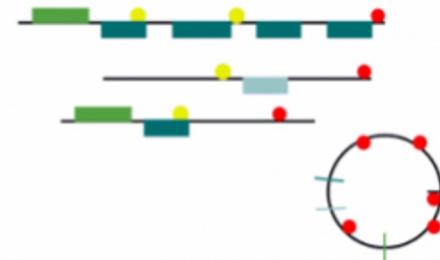
ENA-Reads:
Sequencing and
sampling information



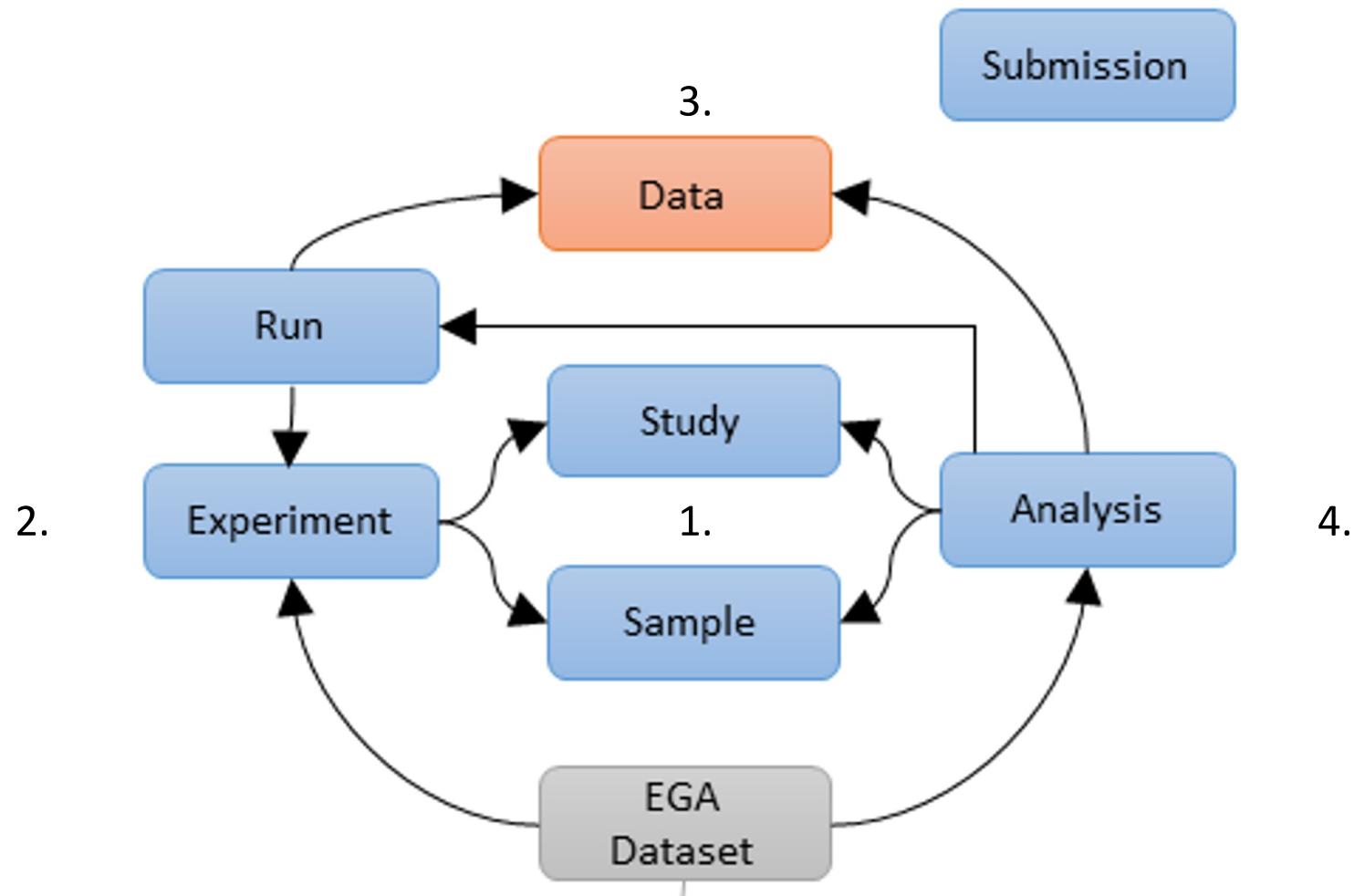
ENA-Assembly:
Assembly information



ENA-Annotation:
Feature annotation



ENA data model



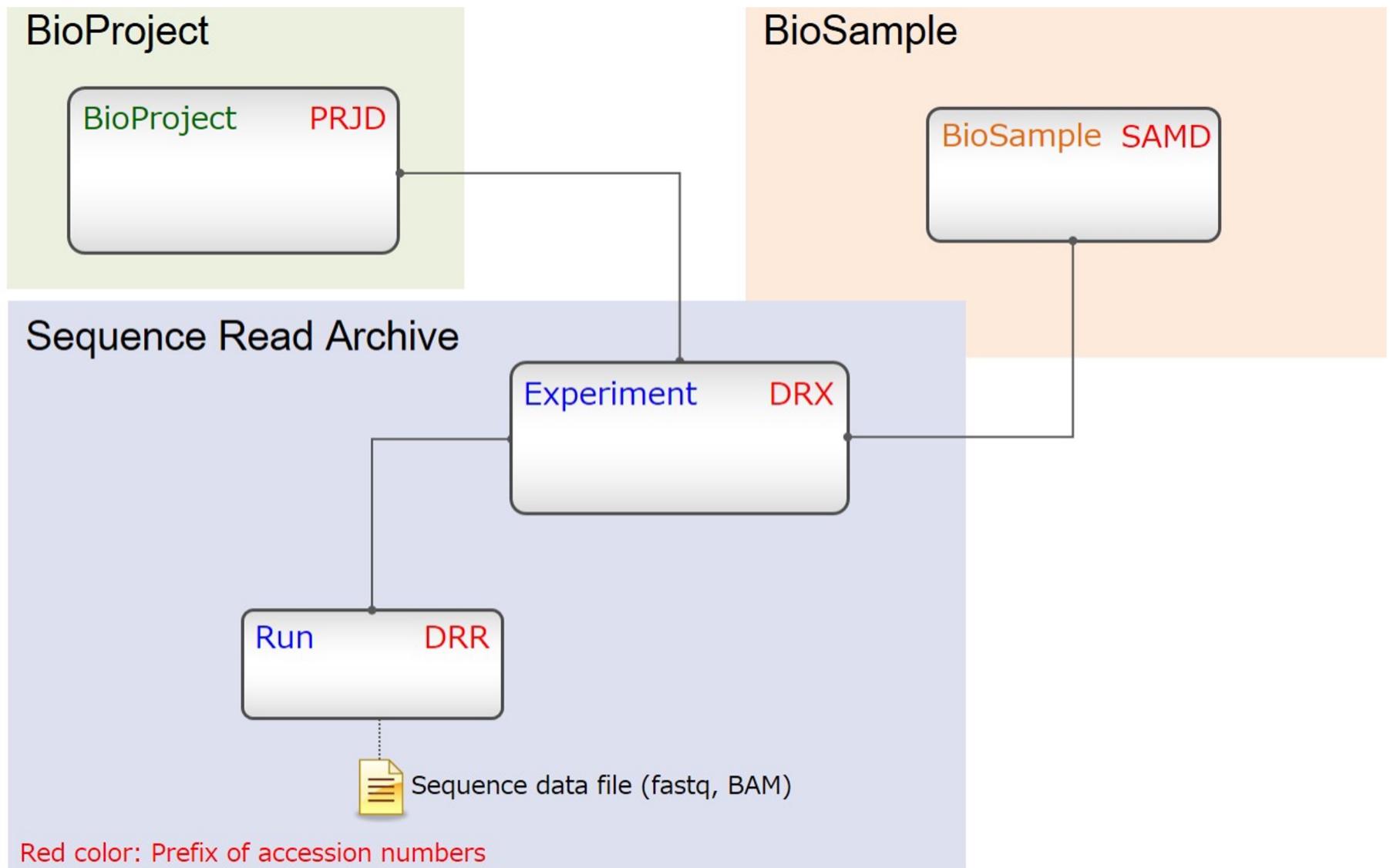
ENA accessions

Type	Accession	Description
Study	ERP/PRJE	Information about the sequencing study
Sample	ERS/SAME	Information about the samples sequenced
Experiment	ERX	Information about sequencing experiment including platform used and library information
Read	ERR	Raw data files containing sequence data (CRAM, BAM, Fastq)
Analysis	ERZ	Secondary analysis results computed from the primary sequencing reads (BAM, EMBL)
Annotated Sequence	LN CWSE	Assembled and annotated sequence, one number for each sequence e.g. CWSE01000001-CWSE01000051

ENA accession codes

Type	Accession	Description
Study	ERP/PRJE	Information about the sequencing study
Sample	ERS/SAME	Information about the samples sequenced
Experiment	ERX	Information about sequencing experiment including platform used and library information
Read	ERR	Raw data files containing sequence data (CRAM, BAM, Fastq)
Analysis	ERZ	Secondary analysis results computed from the primary sequencing reads (BAM, EMBL)
Annotated Sequence	LN CWSE	Assembled and annotated sequence, one number for each sequence e.g. CWSE01000001-CWSE01000051

DDBJ data model



ENA data submission

The screenshot shows the ENA homepage with a teal header. On the left is the ENA logo with the text "European Nucleotide Archive". On the right are two search boxes: one for text search terms ("Enter text search terms") with an example "histone, BN000065" and a search button; and another for accession numbers ("Enter accession") with an example "Taxon:9606, BN000065, PRJEB402" and a "View" button. Below these are navigation links: Home, Submit, Search, Rulespace, About, and Support. A yellow callout box contains text about subscribing to the ENA-announce mailing list and information for SARS-CoV-2 data submissions.

We recommend that you subscribe to the [ENA-announce mailing list](#) for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

Submitting and updating data

We offer a number of services through which data (including updates) can be submitted to the European Nucleotide Archive (ENA). These technologies provide options appropriate for the scale and frequency of submission, the expertise and capacity of the submitter and the nature of the data to be transferred. The choices below lead users most directly to the appropriate submission route.



General Guide on ENA Data Submission

Welcome to the general guide for the European Nucleotide Archive submission. Please take a moment to view this introduction and consider the options available to you before you begin your submission.

ENA allows submissions via three routes, each of which is appropriate for a different set of submission types. You may be required to use more than one in the process of submitting your data:

- **Interactive Submissions** are completed by either filling out web forms directly in your browser or alternatively downloading spreadsheets that can be completed off-line and uploaded to ENA. This is often the most accessible submission route.
- **Command Line Submissions** use our bespoke Webin-CLI program. This validates your submissions entirely before you complete them, allowing you maximum control of the process.
- **Programmatic Submissions** are completed by preparing your submissions as XML documents and either sending them to ENA using a program such as cURL or using the [Webin Submissions Portal](#).

The table below outlines what can be submitted through each submission route. It is also recommended that you familiarise yourself with our [metadata model](#).

	Interactive	Webin-CLI	Programmatic
--	-------------	-----------	--------------

Browsing ENA

- Let's browse at
 - <https://www.ebi.ac.uk/ena/browser/home>

Message posted 2020-11-19.

We recommend that you subscribe to the [ENA-announce mailing list](#) for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a [Drag-and-Drop Data Submission Service](#) (currently in Beta) suitable for certain SARS-CoV-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. [More about ENA](#).

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

[Submit](#) [Search](#) [Rulespace](#) [Support](#)

Latest ENA news

09 Nov 2020: [Read file downloads unavailable Thursday 12th November 2020](#)
Read file downloads unavailable Thursday 12th November 2020 09 Nov 2020 Read file downloads (fastq, SRA, submitted) will be unavailable for a short duration from the following services on Thursday, 12th November 2020 due to unavoidable urgent storage maintenance tasks. We apologise for any inconvenience this may cause. [Read more >](#)

16 Jul 2020: [Retirement of old ENA Browser on 5th August 2020](#)
Retirement of old ENA Browser on 5th August 2020 16 Jul 2020 The new ENA Browser (<https://www.ebi.ac.uk/ena/browser/home>) has been running in parallel to our old Browser (<https://www.ebi.ac.uk/ena>) since mid 2019. The new Browser provides significant stability and speed improvements, as well as improved features for data search, presentation and download. We will now be retiring the old Browser on 5th August 2020. [Read more >](#)

[See all news](#)

Tweets by @enasequence

ENAS ENA @enasequence
Read file downloads (fastq, SRA, submitted) will be unavailable for a short duration from the following services on Thursday, 12th November 2020 due to unavoidable maintenance tasks: ENA Browser, SRA FTP, Aspera, Globus and Galaxy. We apologise for any inconvenience this causes.

Nov 9, 2020

Browsing ENA

- Let's browse at

- <https://www.ebi.ac.uk/ena/browser/home>
- PRJNA637317

The screenshot shows the ENA browser interface for the study PRJNA637317. At the top, there is a search bar with the text "PRJNA637317" and a "View" button. Below the search bar, there is a message: "This project contains single- and bulk- RNA-seq data from sorted rhesus macaque CD8 T-cells". On the left, there is a sidebar with options for "View" (XML, XML (STUDY), Download (XML, XML (STUDY)), Navigation (Show), Read Files (selected), Additional Attributes (Show), and Related ENA Records (Show)). The main content area displays study details: Secondary Study Accession: SRP265965; Study Title: RNA-seq Data from Vaccine-elicited CD8 T-cells; Center Name: Oregon Health & Science University. Below this, there is a "Read Files" section with a "Show Column Selection" dropdown. At the bottom, there is a "Download report" section with JSON and TSV options, and a "Download Files as ZIP" button. A table lists study details for three samples:

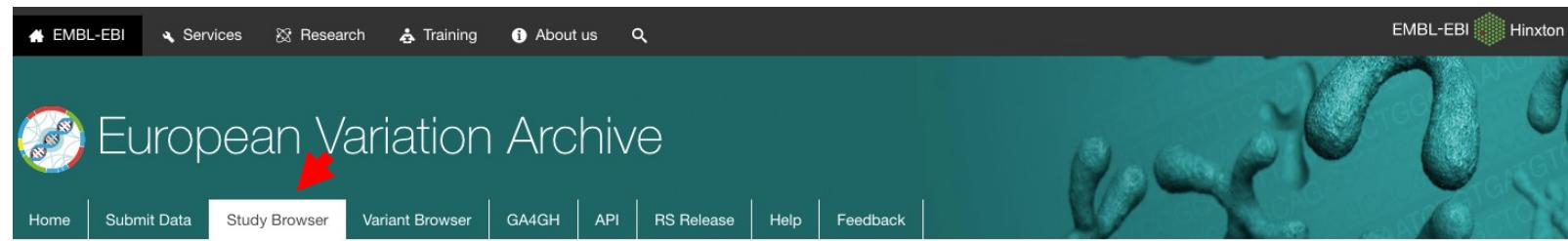
Study Accession	Sample Accession	Experiment Accession	Run Accession	Tax Id	Scientific Name	FASTQ FTP	Download All	Do Subr
PRJNA637317	SAMN15161094	SRX8490486	SRR11945743	9544	Macaca mulatta	<input type="checkbox"/> SRR119457...fastq.gz <input type="checkbox"/> SRR119457...fastq.gz		
PRJNA637317	SAMN15161309	SRX8490485	SRR11945744	9544	Macaca mulatta	<input type="checkbox"/> SRR119457...fastq.gz <input type="checkbox"/> SRR119457...fastq.gz		
PRJNA637317	SAMN15161308	SRX8490484	SRR11945745	9544	Macaca mulatta	<input type="checkbox"/> SRR119457...fastq.gz <input type="checkbox"/> SRR119457...fastq.gz		

European Variation Archive (EVA)

- ▶ For genetic variation data from all species
- ▶ Data submission
 - ▶ Same infrastructure as ENA
 - ▶ Consists of VCF file(s) and metadata that describes sample(s), experiment (s), and analysis that produced the variants
 - ▶ Accessions are ERZ
- ▶ NCBI equivalent is dbSNP

Browsing EVA

- ▶ Let's browse at
 - ▶ <https://www.ebi.ac.uk/eva/?Home>



The screenshot shows the EVA Study Browser interface. At the top, there is a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a search bar. The main title "European Variation Archive" is displayed with a logo to its left. Below the title, the "Study Browser" tab is highlighted with a red arrow pointing to it. The page content includes a heading "Study Browser", a search bar, and a "Filter" section on the left. The "Filter" section contains three expandable dropdowns: "Variant Type" (radio buttons for "Short Genetic Variants (<50bp)" and "Structural Variants (>50bp)"), "Text Search" (an empty input field), and "Genome" (checkboxes for various organisms like Alpaca, Amoeba, Artichoke, etc.). To the right, a table titled "Studies found" lists 25 of 344 studies, showing columns for ID, Name, Genome, Sample(s), Type, and ENA files. Each study row has a "View" link.

ID	Name	Genome	Sample(s)	Type	Submit... Files	Brows... Files
PRJEB31129	Identification of SNPs in alpacas using the Bovine HD Ge...	Alpaca	Vicugna pacos	Array	ENA 	
PRJEB28260	Whole genome sequencing of natural populations of soci...	Amoeba	Dictyostelium discoideum	WGS	ENA 	
PRJEB28606	GBS Cynara cardunculus	Artichoke	Cynara cardunculus var. Curation		ENA 	
PRJEB34294	A 50K SNP array reveals genetic structure for bald eagles...	Bald eagle	Haliaeetus leucocephalus	Array	ENA 	
PRJEB40501	Determination of Genomic Markers to Conduct Parentage...	Bison	Bison bison	WGS	ENA 	
PRJEB38067	Genome-wide Association Analysis Identifies Molecular M...	Black tea	Camellia sinensis var. sin	Array	ENA 	
PRJEB13625	Whole genome resequencing of the human parasite Schis...	Blood fluke	Schistosoma mansoni	WGS	ENA 	
PRJEB26751	Development of single nucleotide polymorphisms for Bra...	Brown mu...	Brassica juncea var. tum	WGS	ENA 	
PRJEB28835	Genetic diversity at LPL gene in river buffalo	Buffalo	Bubalus bubalis bubalis	GBS	ENA 	
PRJEB28591	Population genomic analyses of the chocolate tree, Theo...	Cacao tree	Theobroma cacao	WGS	ENA 	
PRJEB36724	Genome wide identification of SNPs in rohu	Carp	Labeo rohita	WGS	ENA 	
PRJEB30080	Genetic investigations in Munchkin cats.	Cat	Felis catus	WGS	ENA 	
PRJEB30318	ID HCM associated variants in cats	Cat	Felis catus	GBS	ENA 	
PRJEB24944	Whole-genome sequencing of native chicken ecotypes an...	Chicken	Gallus gallus	WGS	ENA 	
PRJEB26970	Indels mutation in the promoter region of the chicken CD...	Chicken	Gallus gallus	WGS	ENA 	

Array Express (EBI)

- ▶ For functional genomics data from array and sequencing based experiments (**RNA-seq, ChIP-seq**)
 - ▶ raw e.g. Affymetrix CEL files, fastq files
 - ▶ processed e.g. aligned bam, txt files of read counts
- ▶ Data submission is via '**Annotare**' a web interface submission tool (<https://www.ebi.ac.uk/fg/annotare/login/>)
- ▶ NCBI equivalent is: **GEO**
- ▶ **1 October 2020 - ArrayExpress is moving to BioStudies**

Browsing ArrayExpress

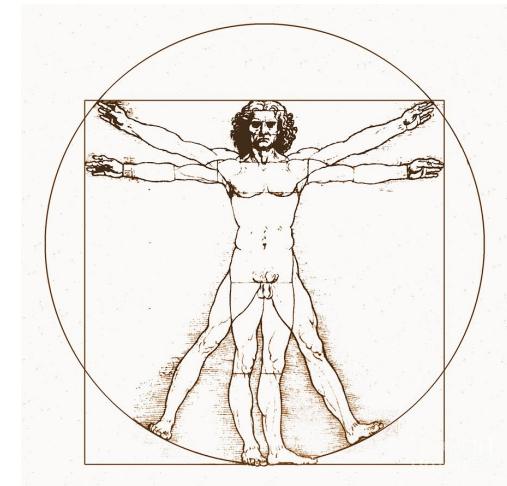
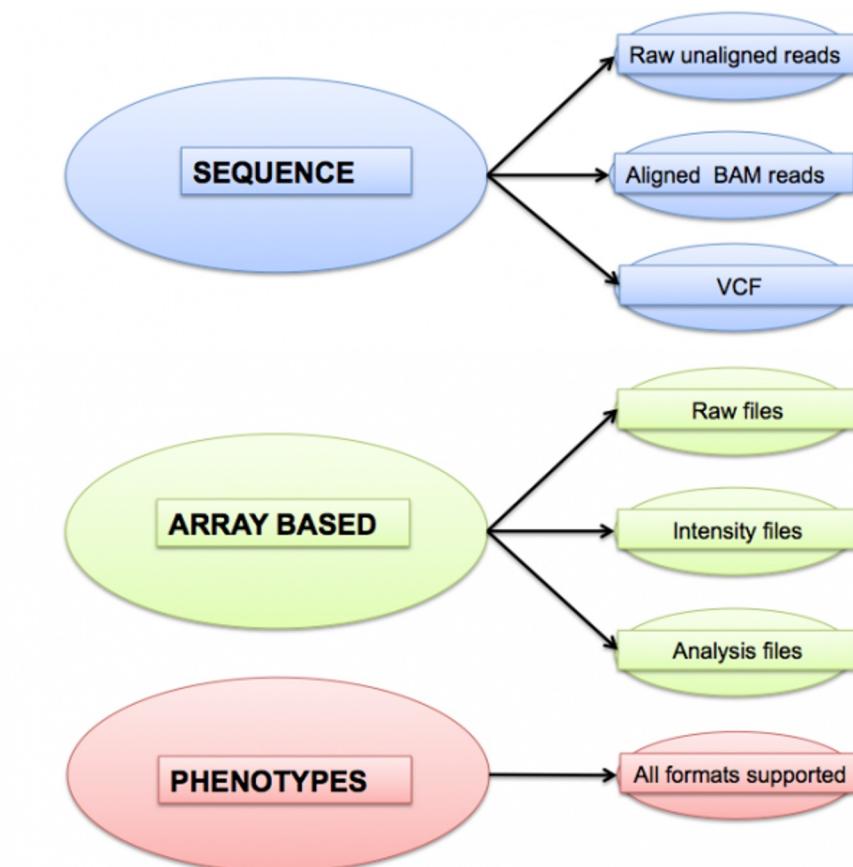
- ▶ Let's browse at
 - ▶ <https://www.ebi.ac.uk/arrayexpress/browse.html>

The screenshot shows the ArrayExpress browse interface. At the top, there is a navigation bar with links for EMBL-EBI, Services, Research, Training, and About us. The EMBL-EBI Hinxton logo is also present. Below the navigation bar, the ArrayExpress logo is displayed next to a search bar containing the placeholder "Search" and examples like "E-MEXP-31, cancer, p53, Geuvadis". There is also a link for "advanced search". The main content area shows a table of experiments. The table has columns for Accession, Title, Type, Organism, Assays, Released, Processed, Raw, and Atlas. The first experiment listed is E-MTAB-9868, which is a methylation analysis of Chilean patients with gallstone disease, low- or high-grade gallbladder dysplasia or gallbladder cancer. The table shows 73786 experiments in total, with the current view showing pages 1 to 25.

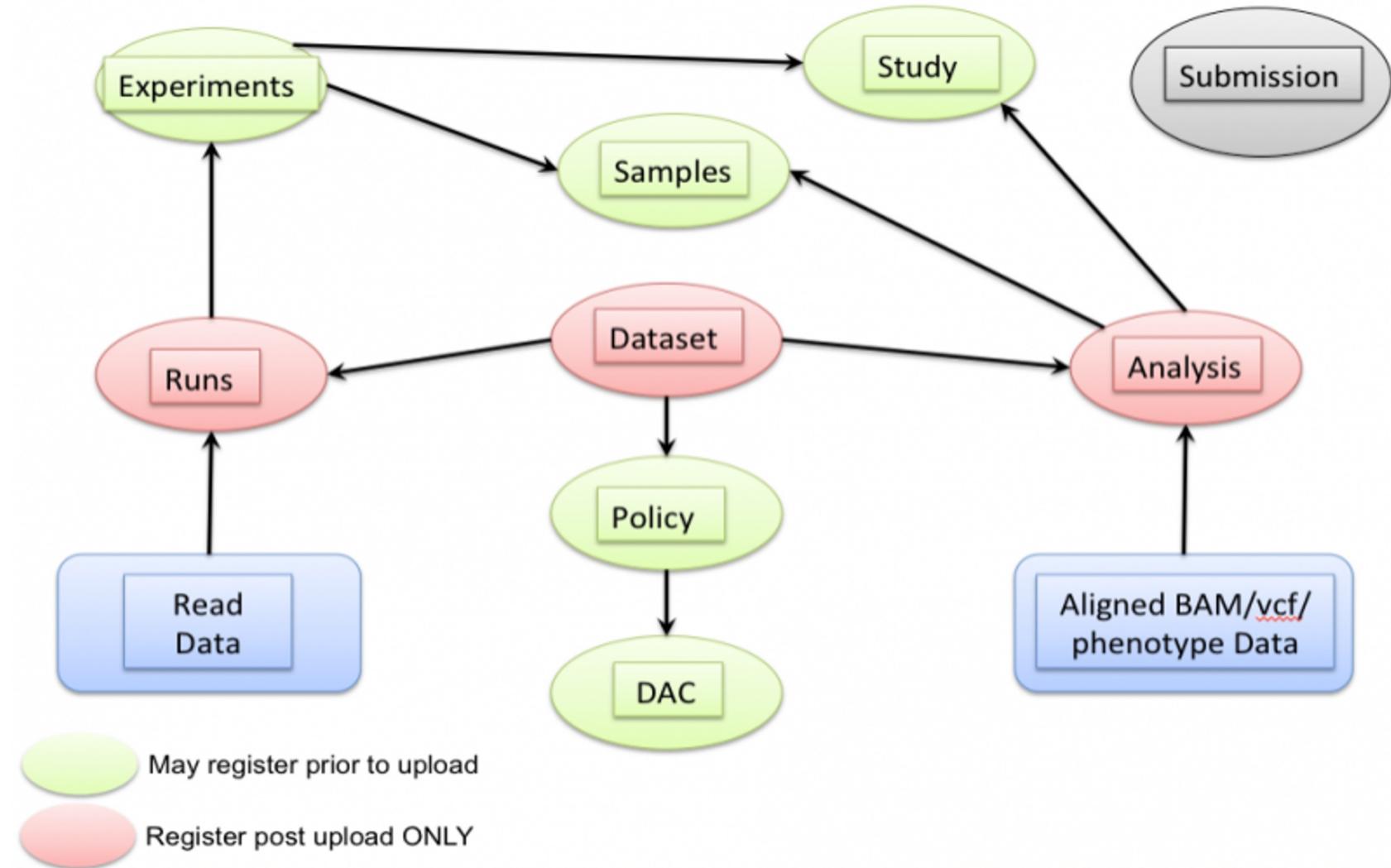
Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-9868	Methylation analysis of Chilean patients with gallstone disease, low- or high-grade gallbladder dysplasia or gallbladder cancer	methylation profiling by array	Homo sapiens	178	Yesterday			-
E-MTAB-9489	scRNA-seq of human fetal intestine tissue	RNA-seq of coding RNA from single cells	Homo sapiens	52	Yesterday	-		-
E-MTAB-9838	RNA-seq of Arabidopsis thaliana seedlings treated with Flg22 or Pep1	RNA-seq of coding RNA	Arabidopsis thaliana	9	11/12/2020	-		-
E-MTAB-9728	Transcriptional profiling of the developmental gene whiA deletion mutant of Streptomyces venezuelae ATCC 10712 at 7 time points from 8 to 20 hours of growth	transcription profiling by array	Streptomyces venezuelae	42	11/12/2020			-
E-MTAB-9034	Poly(A) mRNA-seq from yw and Twi-mCherry-LEXY(V416L) embryos subjected to different blue light exposure schemes	RNA-seq of coding RNA	Drosophila melanogaster	26	11/12/2020	-		-
E-MTAB-7502	Bulk segregant analysis of 3D growth defective mutant in Physcomitrella patens (nog2)	DNA-seq	Physcomitrella patens	8	11/12/2020	-		-
E-MTAB-9720	scRNA-seq of human fetal duodenal enteroids grown in EGF and/or NRG1	RNA-seq of coding RNA from single cells	Homo sapiens	20	10/12/2020	-		-
E-MTAB-9707	Transcription profiling by array of two subpopulations of tumour-associated macrophages (TAMs) treated with DMSO or T0901317 for 24h	transcription profiling by array	Mus musculus	10	10/12/2020			-
E-MTAB-9656	RNA-seq of prostate cancer Patient-Derived Xenografts	RNA-seq of	Homo sapiens	17	10/12/2020	-		-

European Genome-Phenome Archive (EGA)

- ▶ For personally identifiable genetic and phenotypic data
- ▶ Individuals whose consent agreements authorize that data is release for specific research use only



EGA data model



EGA accessions

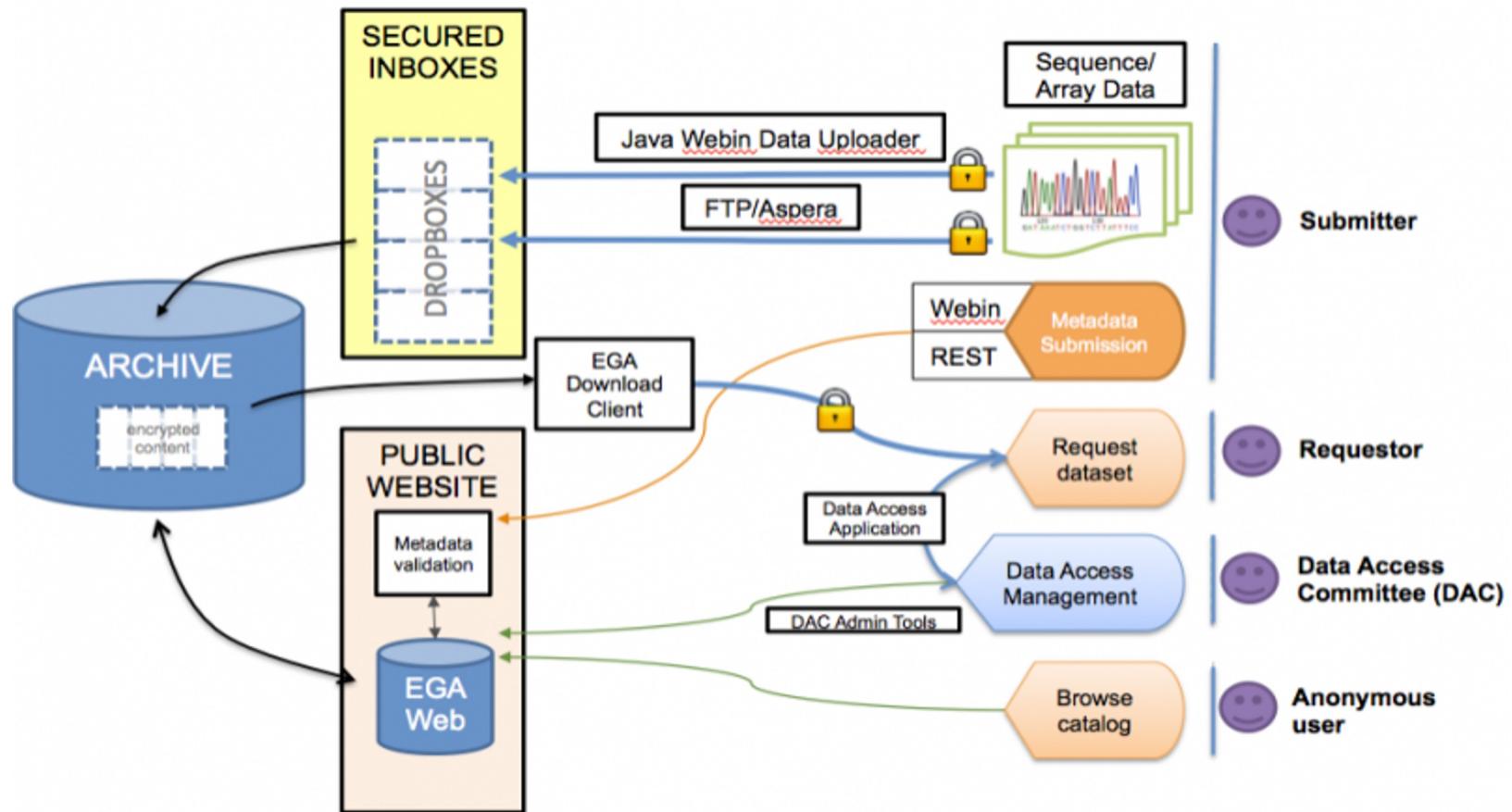
Type	Accession	Description
Study	EGAS	Information about the sequencing study
Sample	EGAN	Information about the samples sequenced
Experiment	EGAX	Information about sequencing experiment including platform used and library information
Run	EGAR	Raw data files containing sequence data (CRAM, BAM, Fastq)
Analysis	EGAZ	Analysis data files associated with study and sample : BAM, VCF, array and phenotype data
Dataset	EGAD	Collection of runs/analysis data files to be subject to controlled access
Policy	EGAP	Contains the data access agreement (DAA)
DAC	EGAC	Information about the data access committee

EGA accessions

Type	Accession	Description
Study	EGAS	Information about the sequencing study
Sample	EGAN	Information about the samples sequenced
Experiment	EGAX	Information about sequencing experiment including platform used and library information
Run	EGAR	Raw data files containing sequence data (CRAM, BAM, Fastq)
Analysis	EGAZ	Analysis data files associated with study and sample : BAM, VCF, array and phenotype data
Dataset	EGAD	Collection of runs/analysis data files to be subject to controlled access
Policy	EGAP	Contains the data access agreement (DAA)
DAC	EGAC	Information about the data access committee

EGA overview

- ▶ Strict protocols govern how information is managed, stored and distributed (data access policy & DACS)



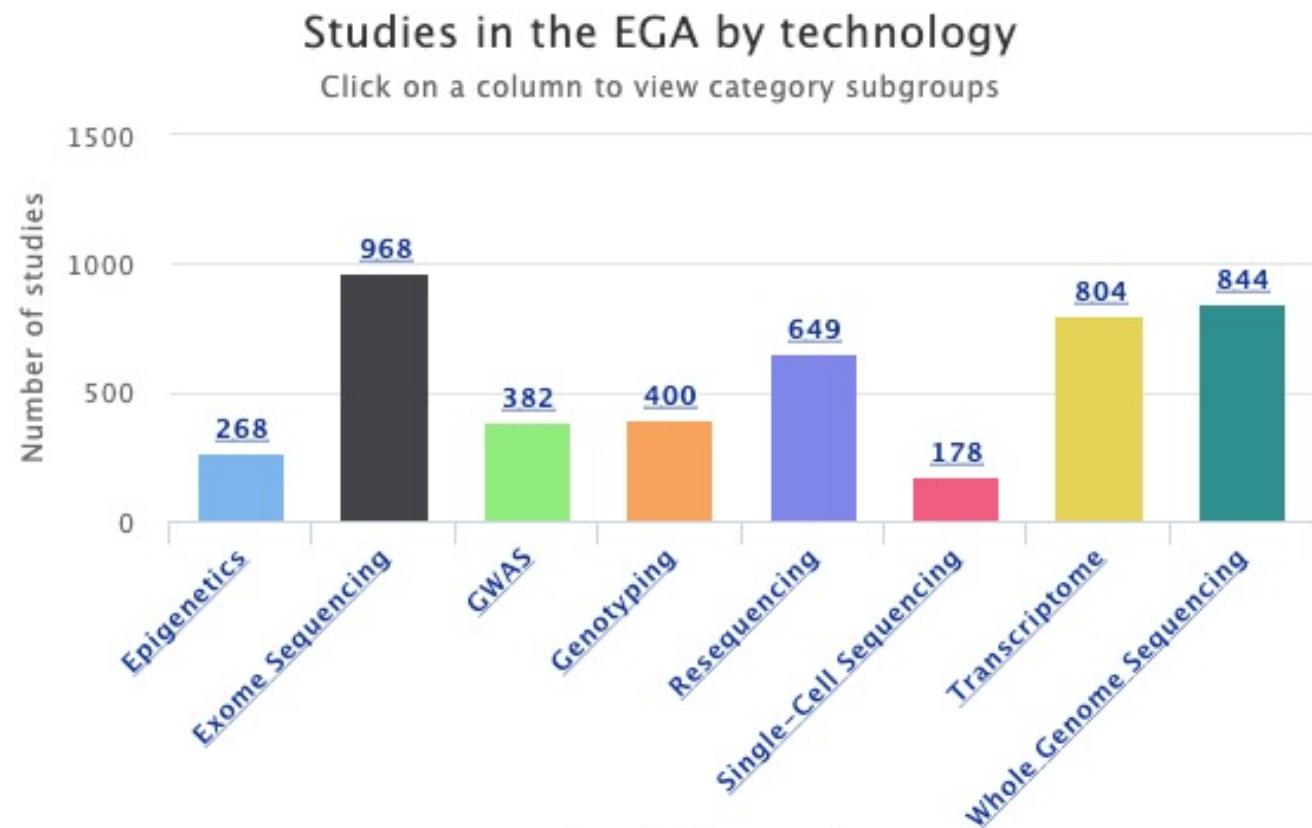
Breakdown of EGA studies (2020)

- Let's browse at
 - <https://ega-archive.org/>

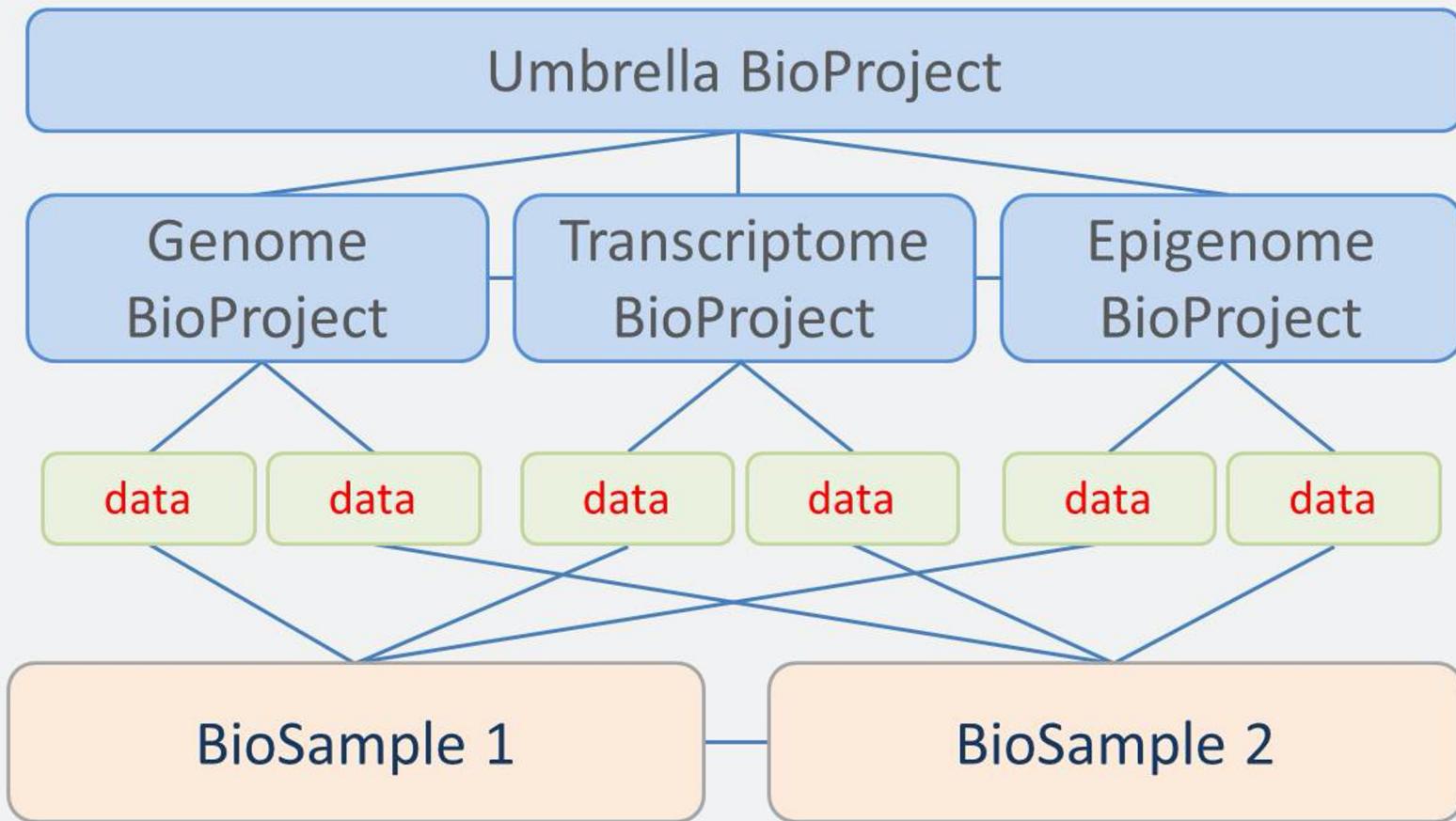


Breakdown of EGA studies (2020)

- Let's browse at
 - <https://ega-archive.org/>



BioProjects and BioSamples



BioSample database

- ▶ Stores descriptive information about biological samples used to generate experimental data
 - ▶ e.g., cell line, blood sample, environmental isolate
 - ▶ species, phenotypic information e.g., disease state, clinical info on individuals
- ▶ Can link up data from different archives for the same sample
- ▶ Accessions always begin with **SAM**
 - ▶ Next is E, N or D, for EBI, NCBI or DDBJ respectively
 - ▶ Next is A or a G, for a sample or a group of samples
 - ▶ Finally, is a numeric component
 - ▶ For more information:
<https://www.ebi.ac.uk/biosamples/submit>

BioProject database

- ▶ Organises samples & data produced by projects
 - ▶ Deposited by several research groups
 - ▶ Deposited into several archival databases
- ▶ Can be created for
 - ▶ Genome sequencing and assembly
 - ▶ Transcriptome sequencing and expression
 - ▶ Targeted locus sequencing
 - ▶ Variation detection
- ▶ Accessions always begin with **PRJ**
 - ▶ Next is E, N or D for EBI, NCBI and DDBJ respectively
 - ▶ Finally, is a numeric component

BioProject database

Find BioProjects by:	Search text example(s)
Project data type	"metagenome"[Project Data Type]
Publication information	19643200[PMID]
Material used	"material transcriptome"[Properties]
Sample scope	"scope environment"[Properties]
Species name	Escherichia coli[organism]
Submitter organization, consortium, or center	JGI[Submitter Organization]
Taxonomic Class	Insecta[organism]
BioProject database identifier	PRJNA33823[bioproject] or 33823[uid] or 33823[bioproject]

From: <https://guides.lib.berkeley.edu/ncbi/bioproject>

WSI data sharing policy

- ▶ Aim to provide **rapid and open access** to data produced
- ▶ Immediate release
 - ▶ Register sequencing studies in BioProject database
 - ▶ Register samples in BioSample database
- ▶ Within 90 days
 - ▶ Primary sequence data (CRAM) in ENA/EGA
- ▶ At publication
 - ▶ Secondary analysis in other archives
 - ▶ VCF, expression data, annotated sequences

Useful resources

- EBI Training
 - <https://www.ebi.ac.uk/training/online/course-list>
- NCBI Handbook
 - <https://www.ncbi.nlm.nih.gov/books/NBK143764/>
- DDBJ Training
 - <https://www.ddbj.nig.ac.jp/training-e.html>