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

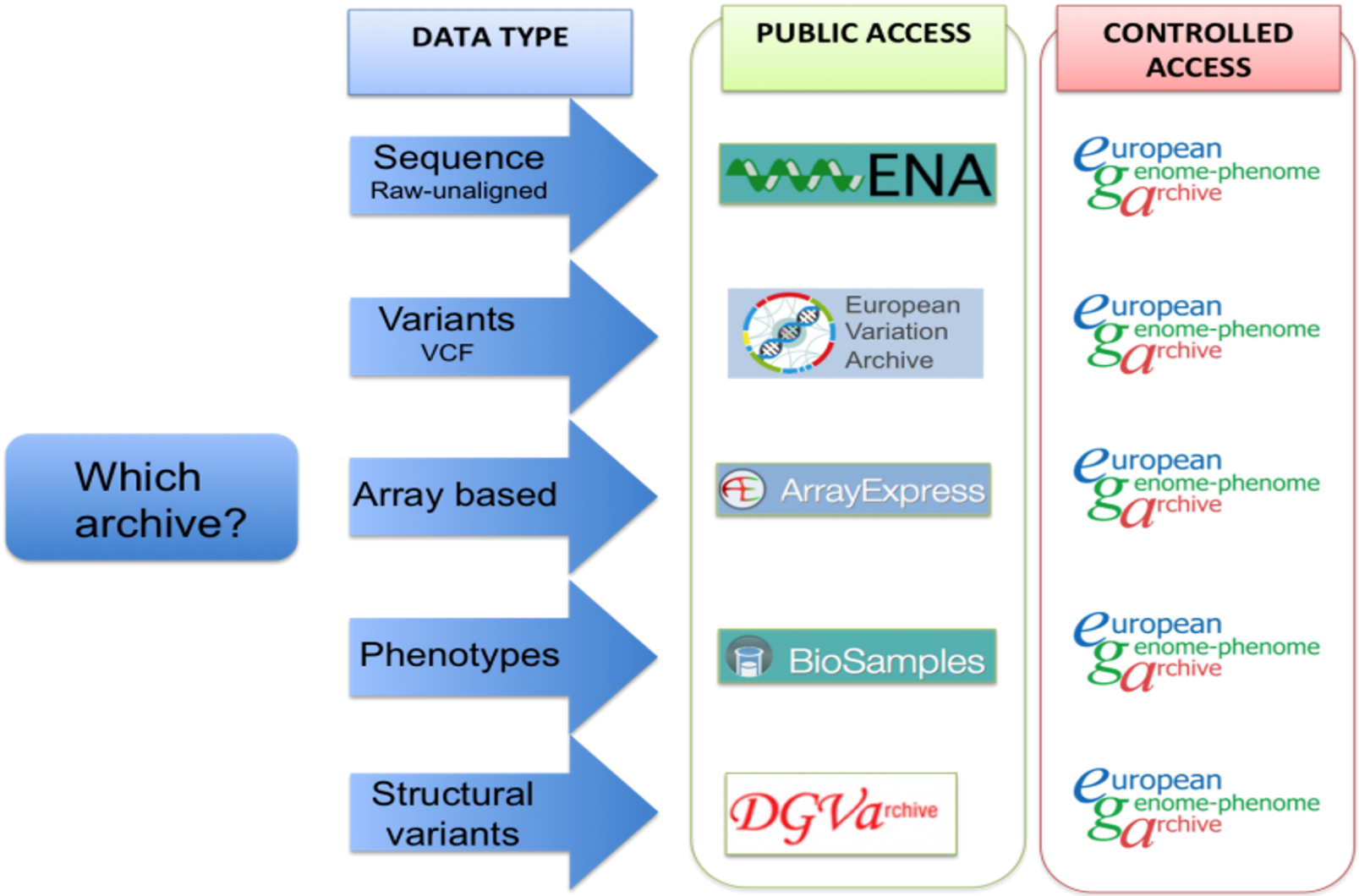
FACULTAD DE
CIENCIAS BIOLOGICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

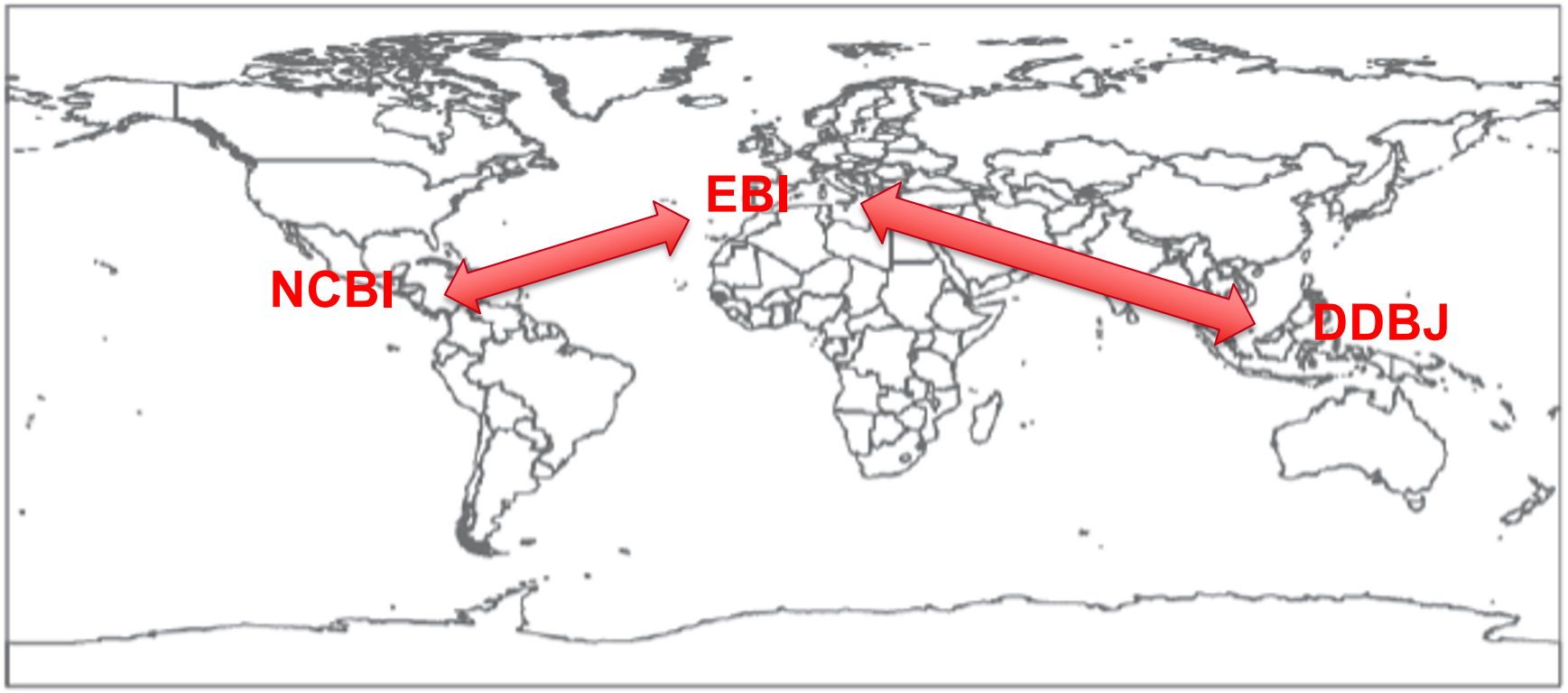
Purpose of data archives (repositories)

- 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



NCBI
National Center for Biotechnology Information

EMBL-EBI

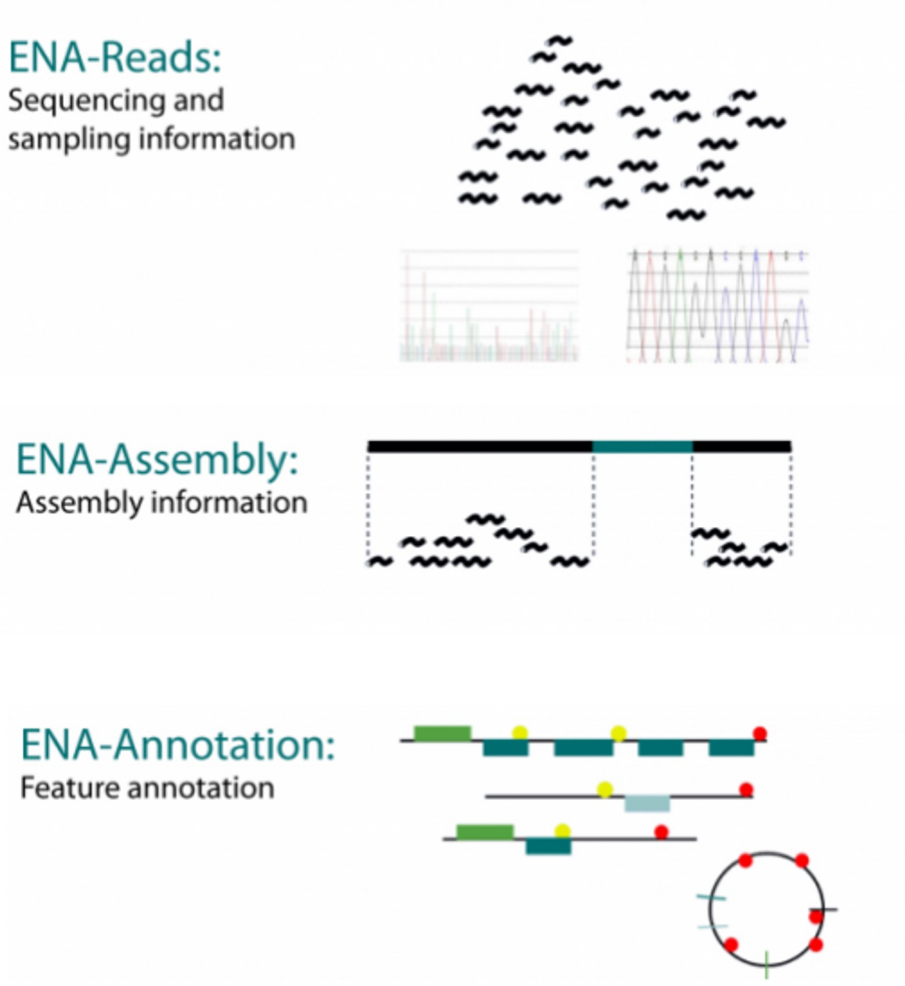
DDBJ
DNA Data Bank of Japan

Global data archives

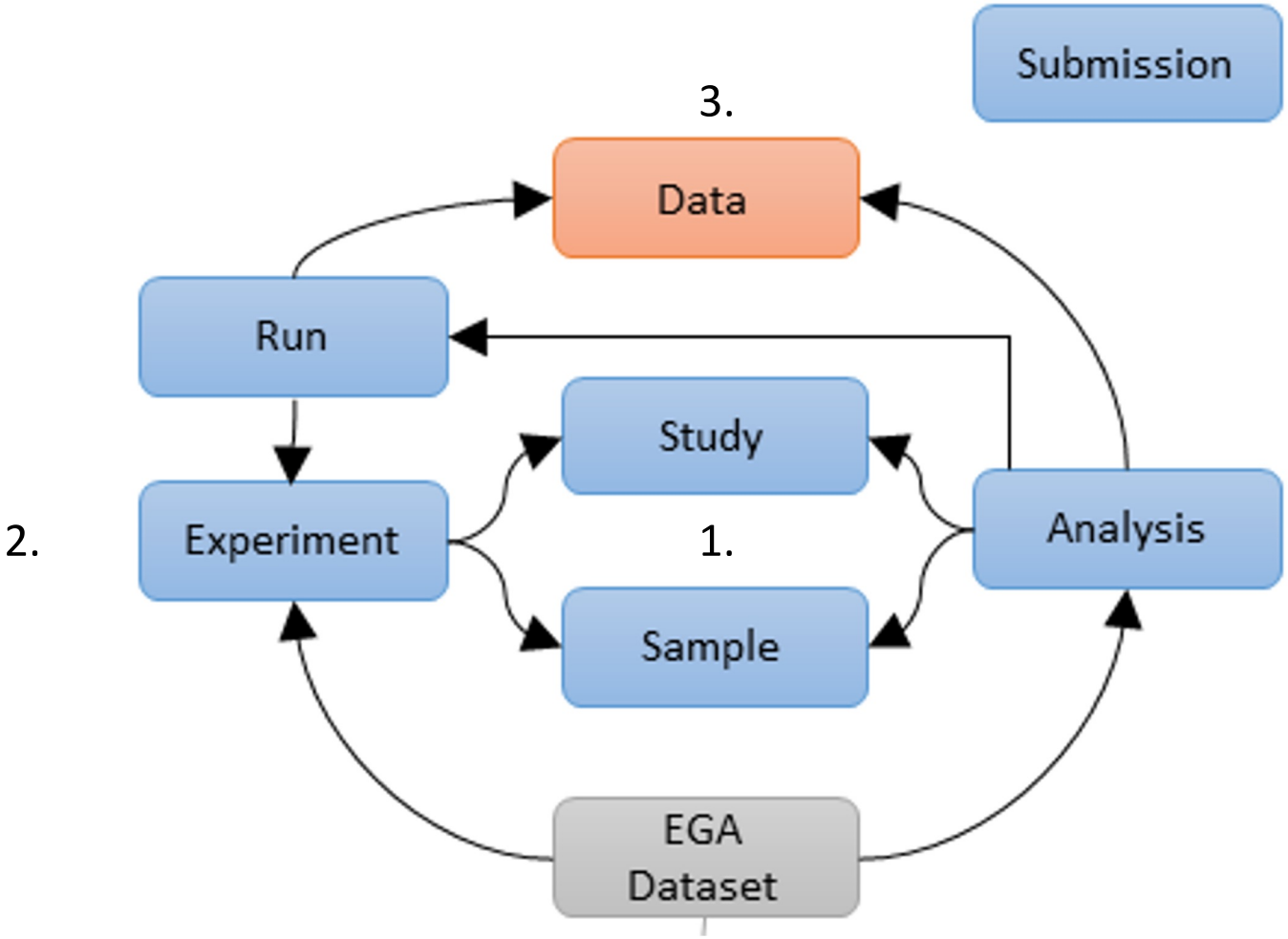
Data Type	DDBJ	EBI	NCBI
Primary Sequence Data	DDBJ Sequence Read Archive (DRA)	European Nucleotide Archive (ENA)	Sequence Read Archive (SRA)
Annotated Sequences	DDBJ		GenBank
Variation	-	European Variation Archive (EVA)	dbSNP
Structural Variation	-	Genomic Variants Archive (DGVa)	dbVar
Expression	DDBJ 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 data model



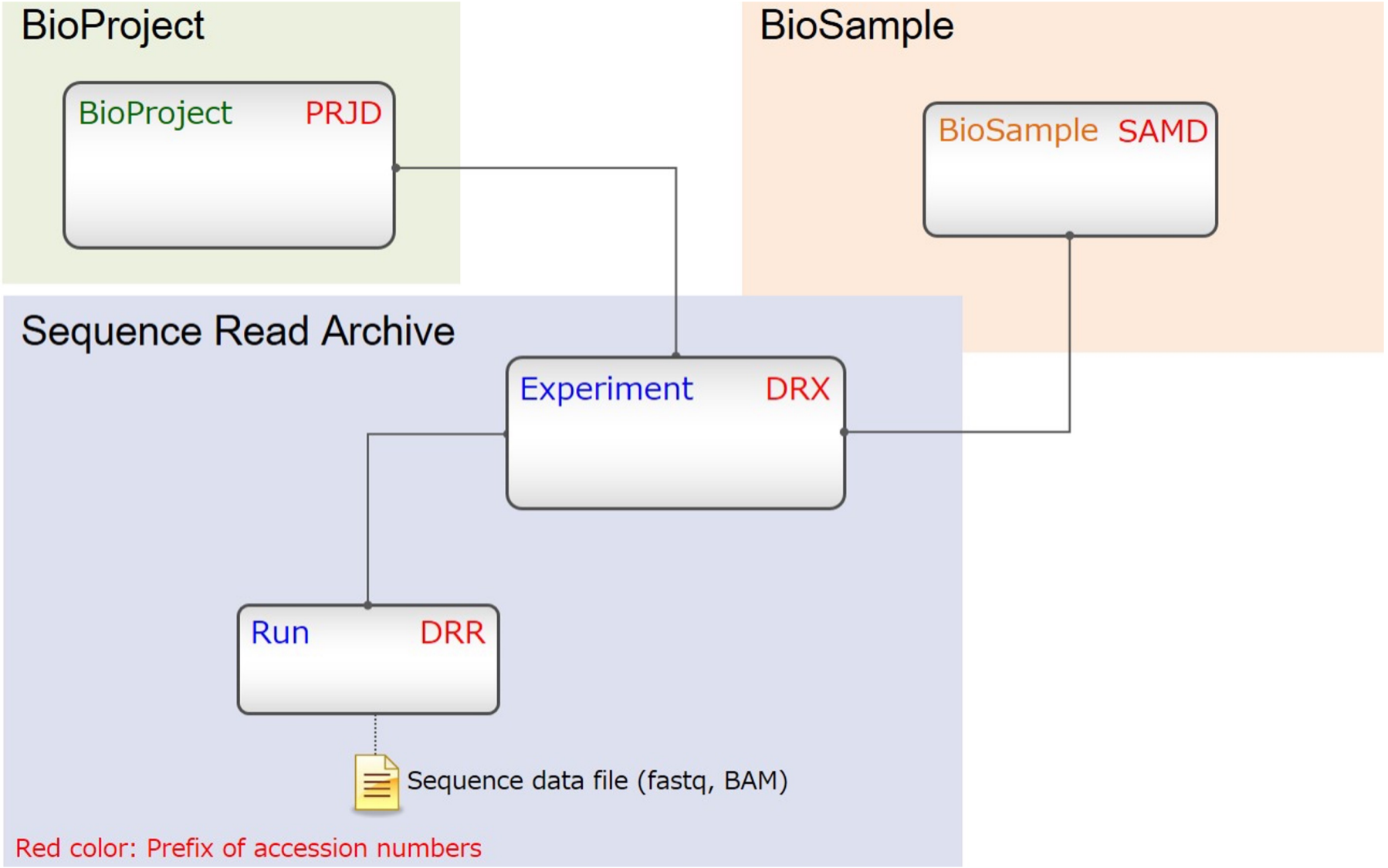
ENA accessions

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

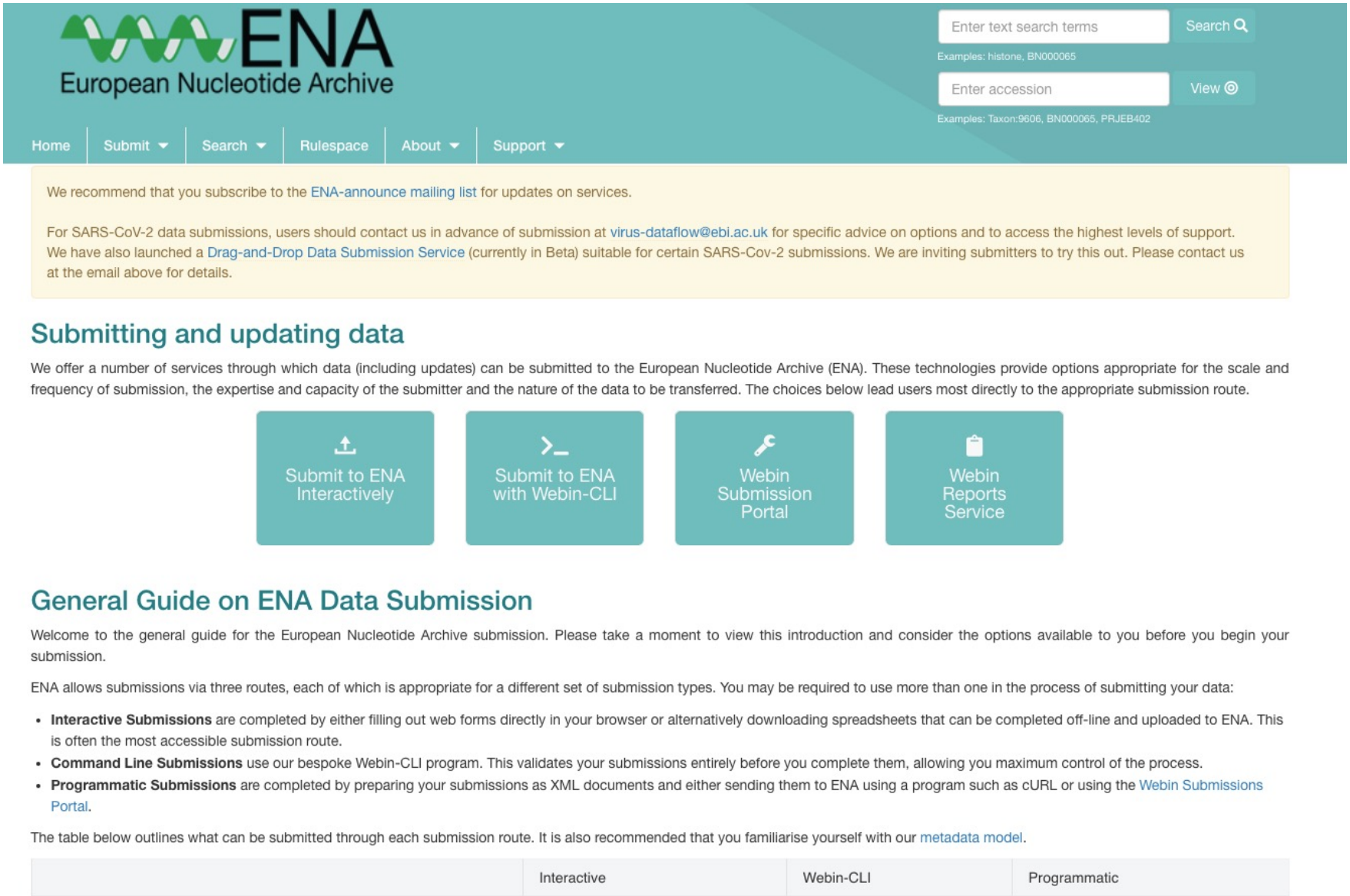
ENA accession codes

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

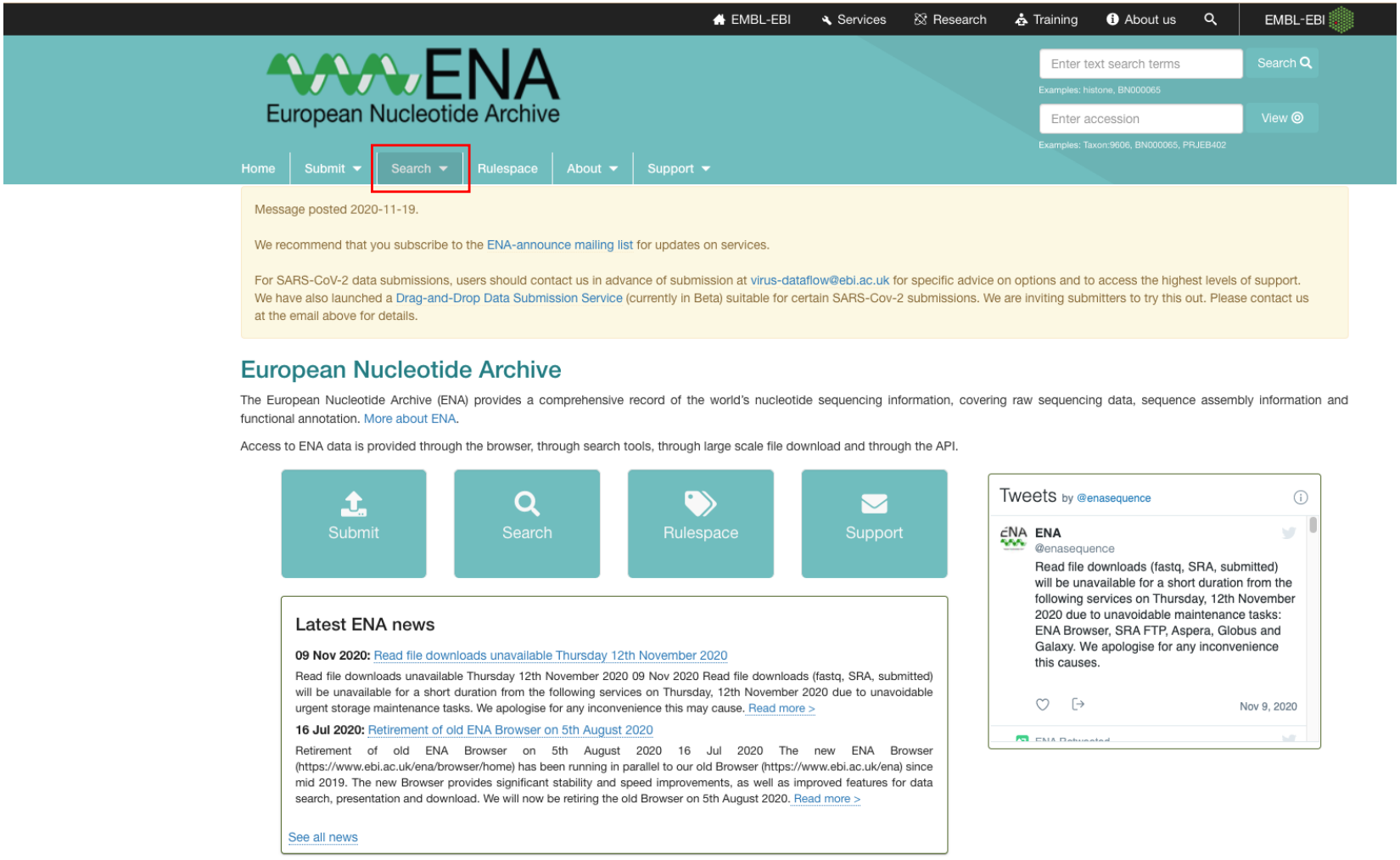


ENA data submission



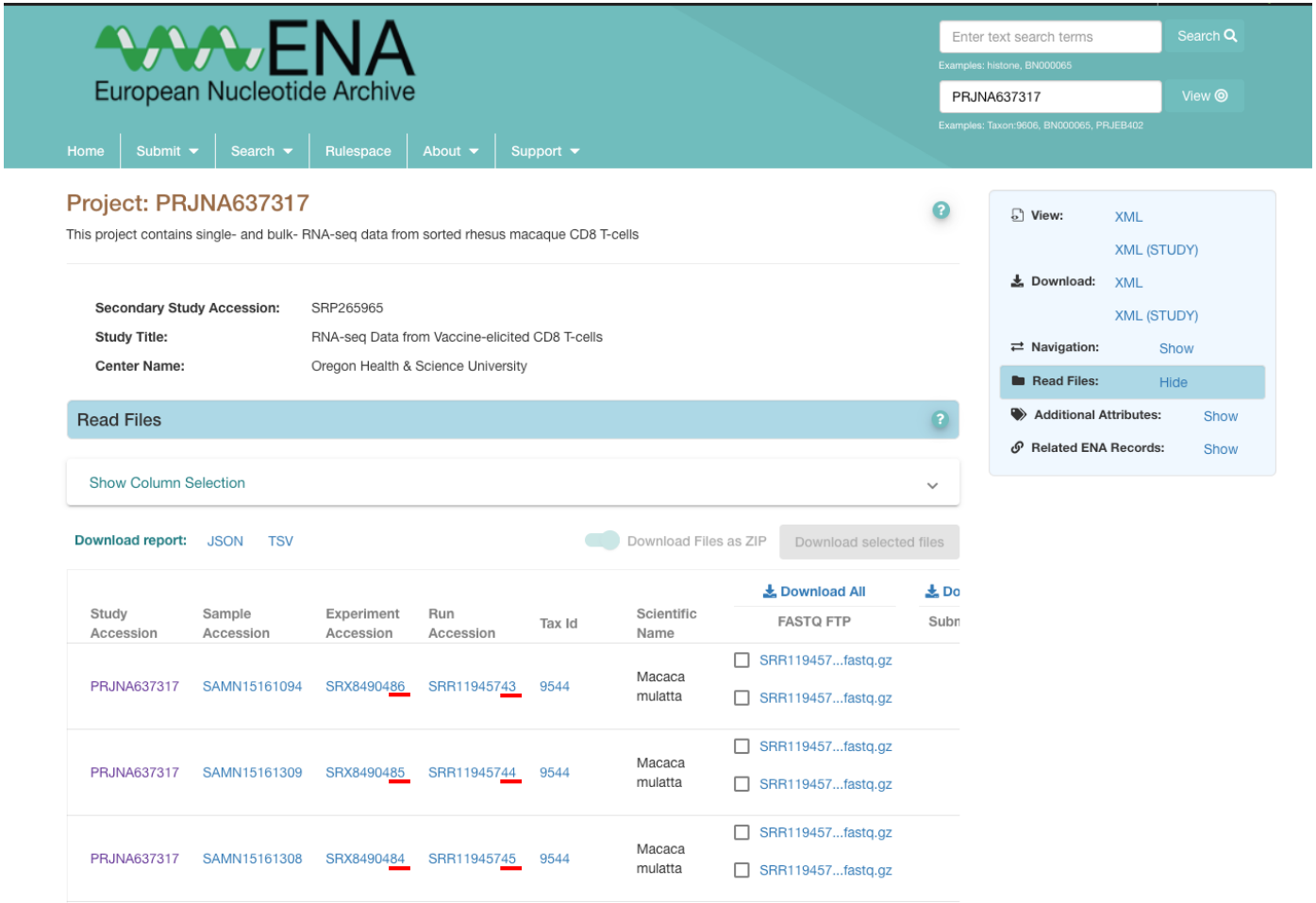
Browsing ENA

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



Browsing ENA

- Let's browse at
  - https://www.ebi.ac.uk/ena/browser/home
  - PRJNA637317

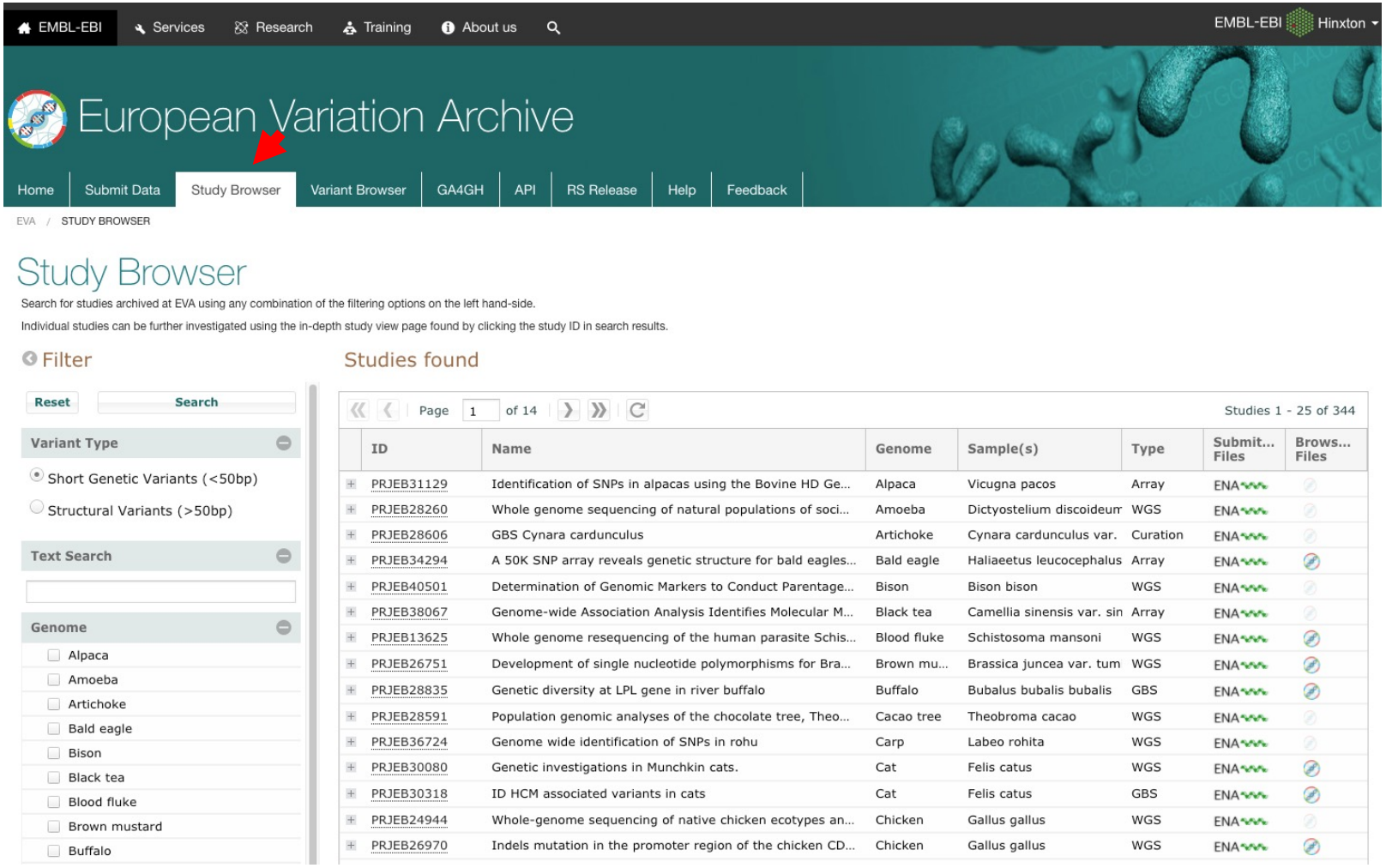


European Variation Archive (EVA)

- 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

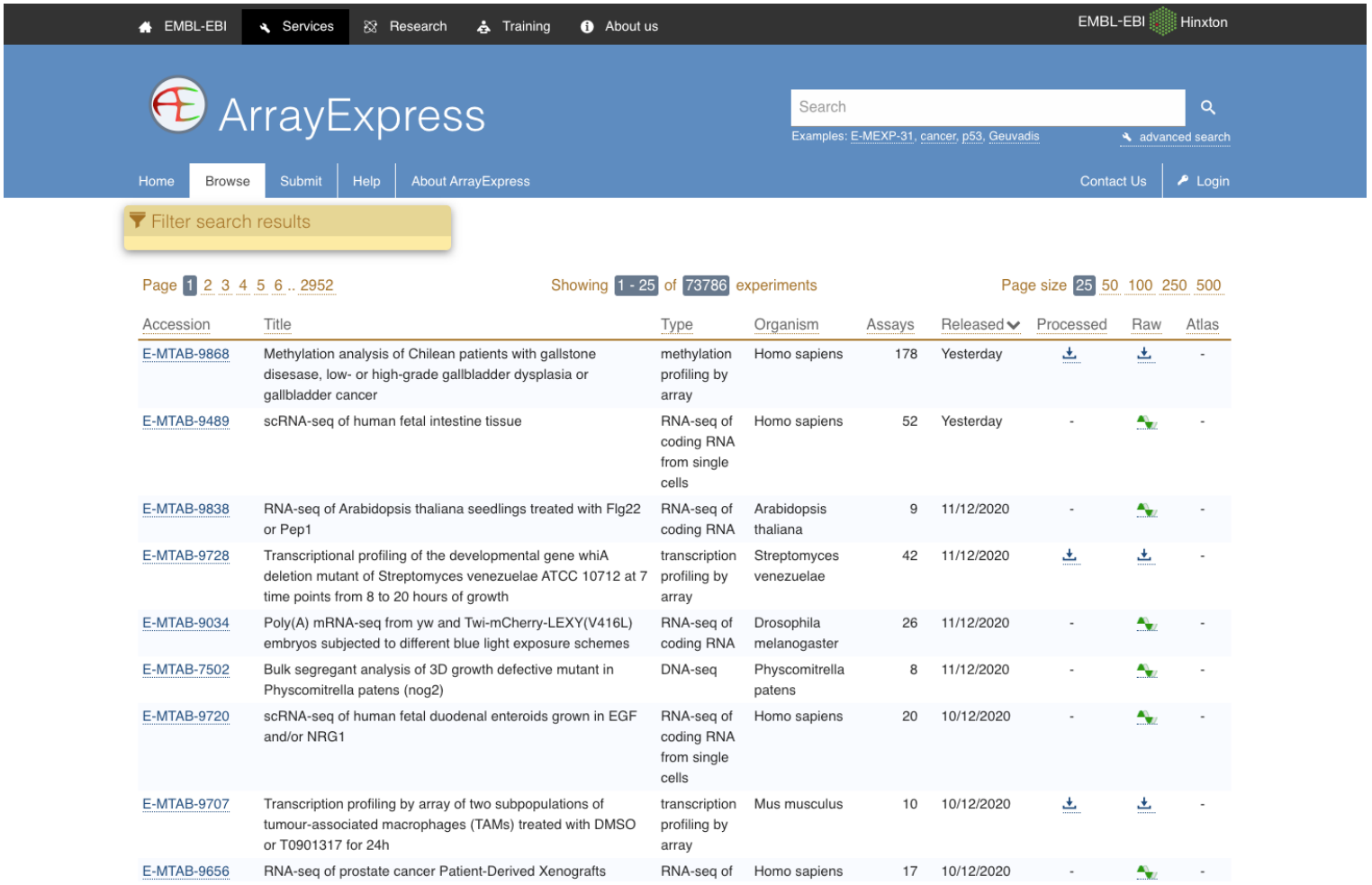


Array Express (EBI)

- 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

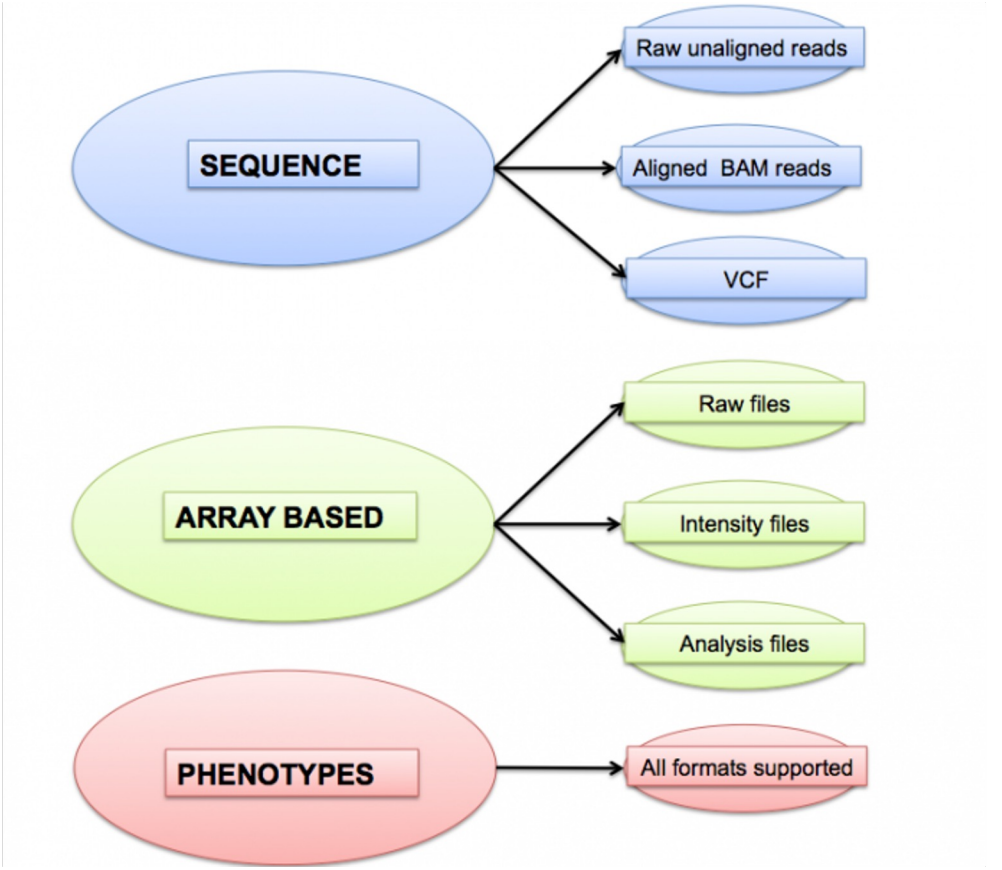
- Let's browse at
  - https://www.ebi.ac.uk/arrayexpress/browse.html

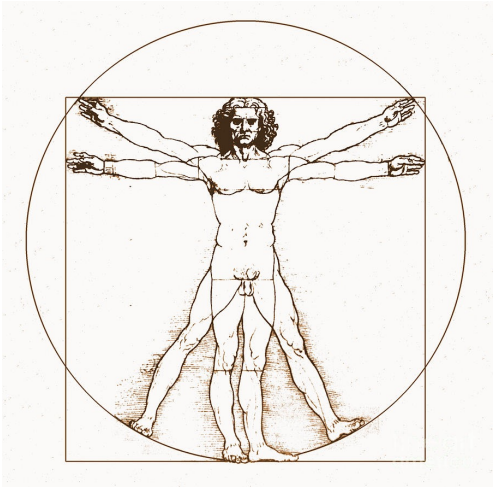


Accession	Title	Type	Organism	Assays	Released	Processed	Raw	Atlas
E-MTAB-9868	Methylation analysis of Chilean patients with gallstone disesase, low- or high-grade gallbladder dysplasia or gallbladder cancer	methylation profiling by array	Homo sapiens	178	Yesterday			-
E-MTAB-9489	scRNA-seq of human fetal 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	-		-

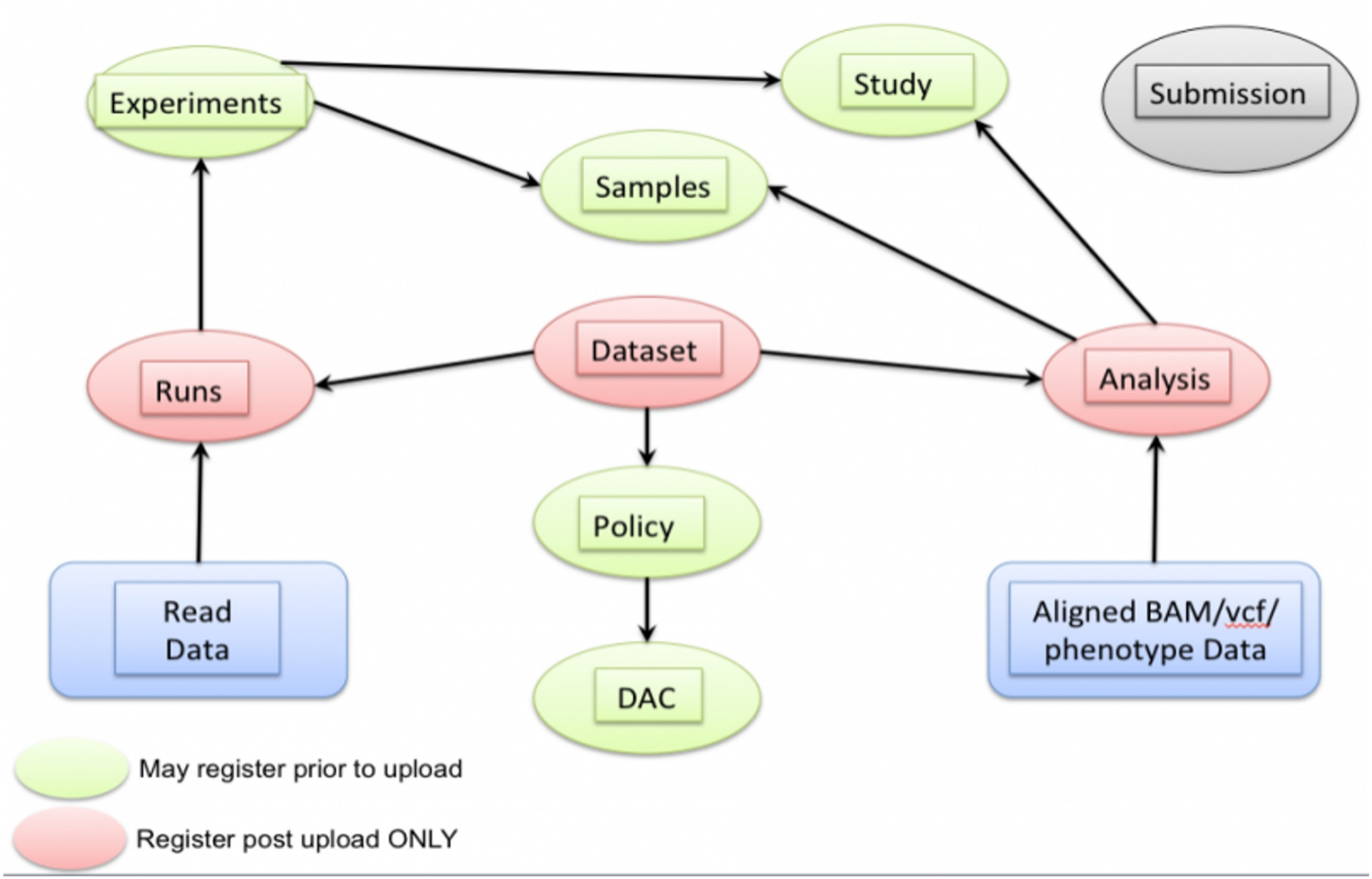
European Genome-Phenome Archive (EGA)

- 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

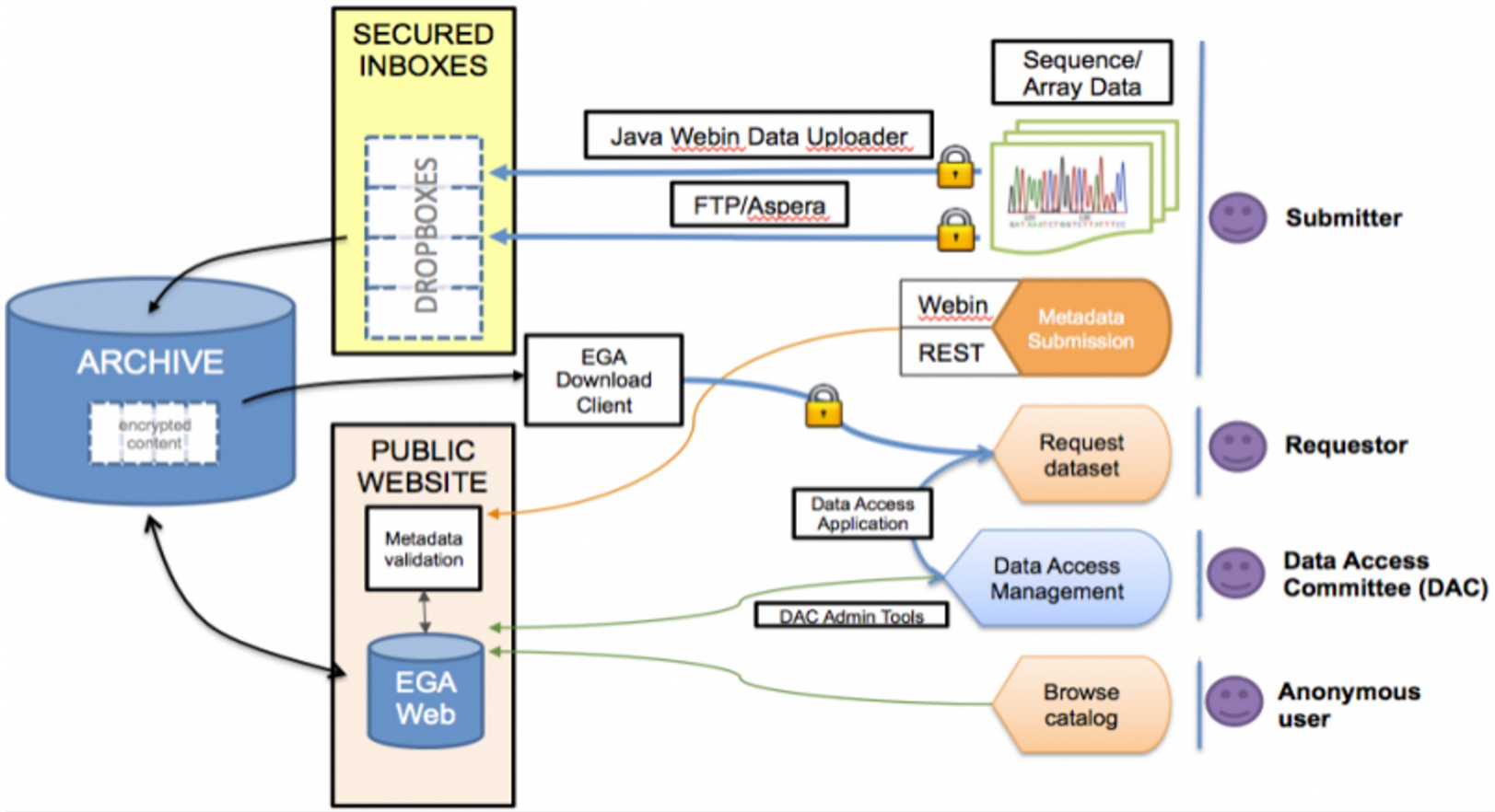
Type	Accession	Description
Study	EGAS	Information about the sequencing study
Sample	EGAN	Information about the samples sequenced
Experiment	EGAX	Information about sequencing experiment including platform used and library information
Run	EGAR	Raw data files containing sequence data (CRAM, BAM, Fastq)
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

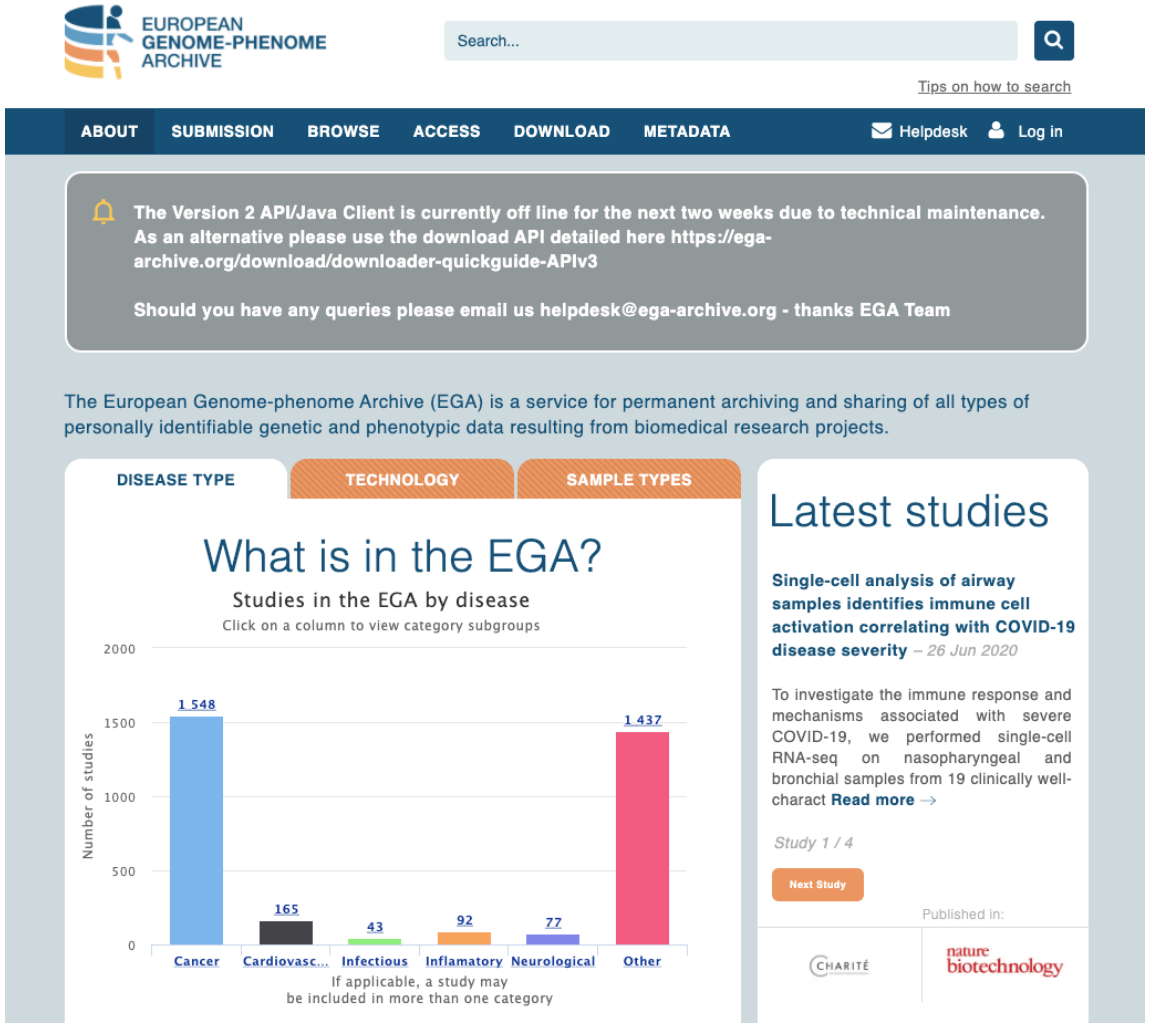
EGA overview

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



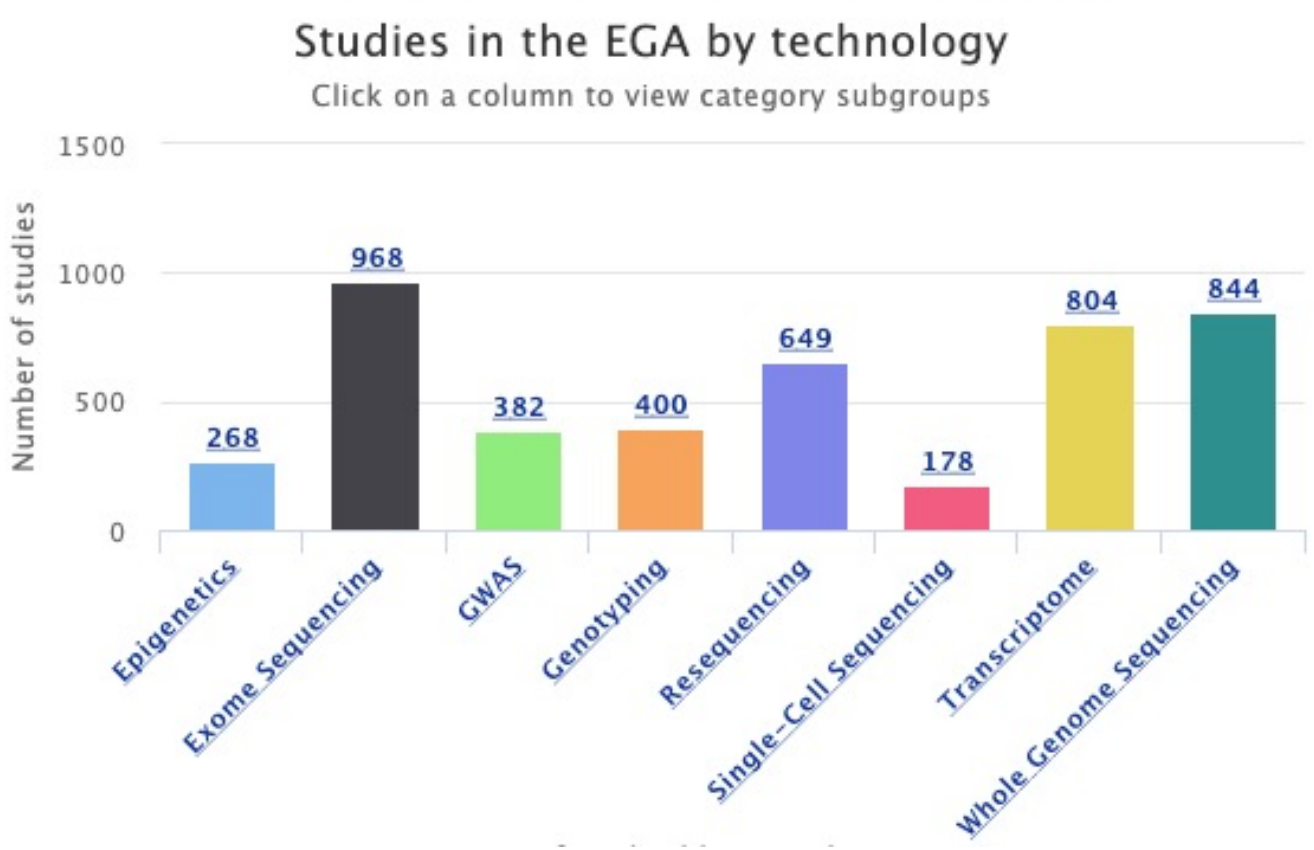
Breakdown of EGA studies (2020)

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

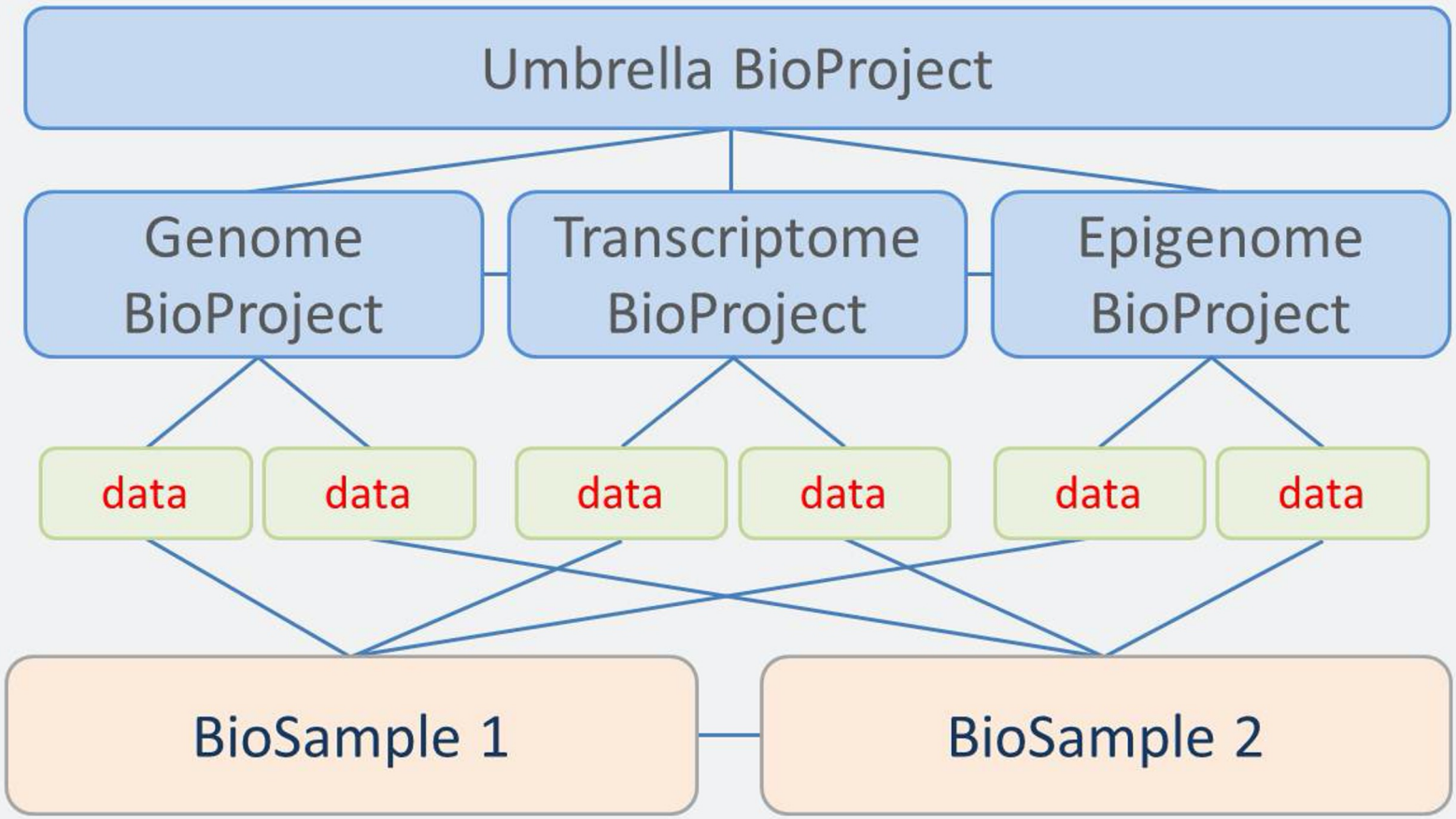


Breakdown of EGA studies (2020)

- 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
  - Fore 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