Public data archives for NGS data

Jacqui Keane

@drjkeane

drjkeane@gmail.com

WELLCOME GENOME CAMPUS

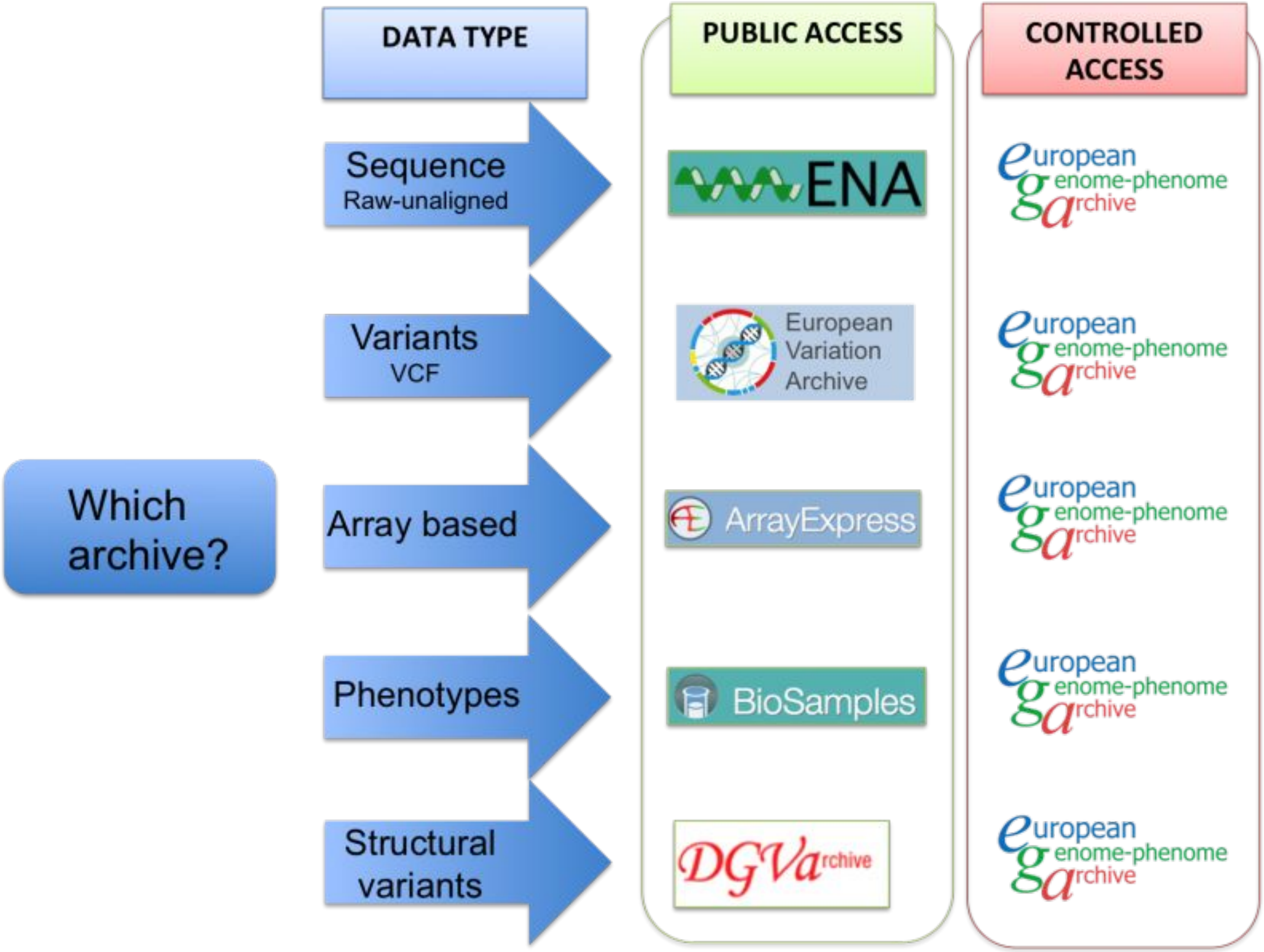
wellcome sanger institute

CONNECTING SCIENCE

Purpose of data archives

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

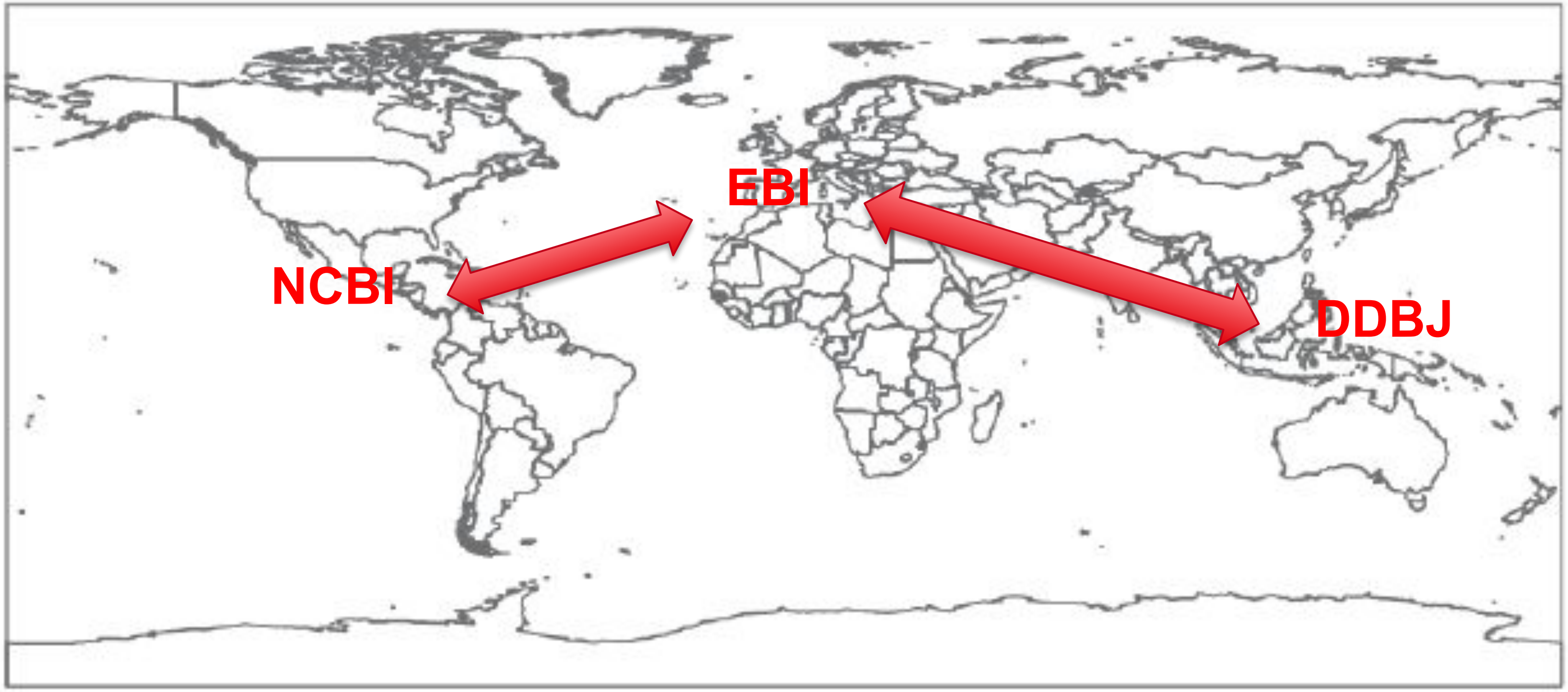
Which data archive?



Which archive?

DATA TYPE	PUBLIC ACCESS	CONTROLLED ACCESS
Sequence Raw-unaligned	ENA	european genome-phenome archive
Variants VCF	European Variation Archive	european genome-phenome archive
Array based	ArrayExpress	european genome-phenome archive
Phenotypes	BioSamples	european genome-phenome archive
Structural variants	DGVarchive	european genome-phenome archive

Data sharing across archives



NCBI National Center for Biotechnology Information

EMBL-EBI

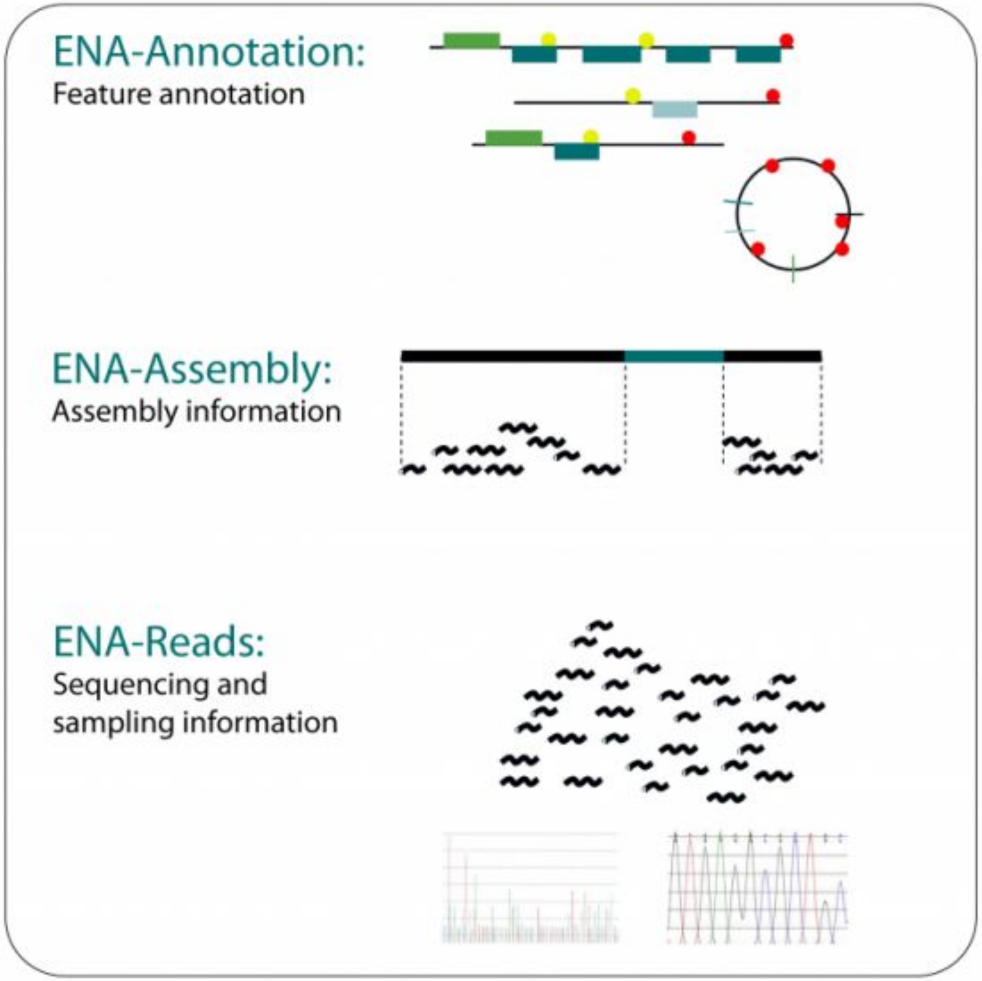
DDBJ DNA Data Bank of Japan

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	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

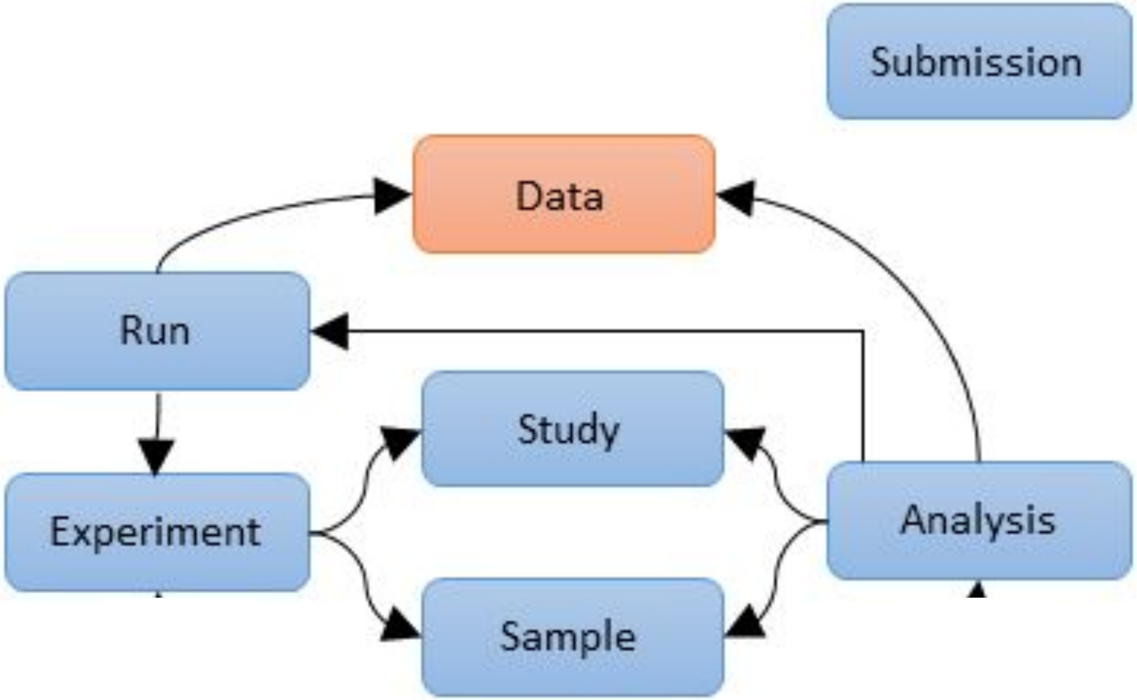


ENA-Annotation:
Feature annotation

ENA-Assembly:
Assembly information

ENA-Reads:
Sequencing and sampling information

ENA data model



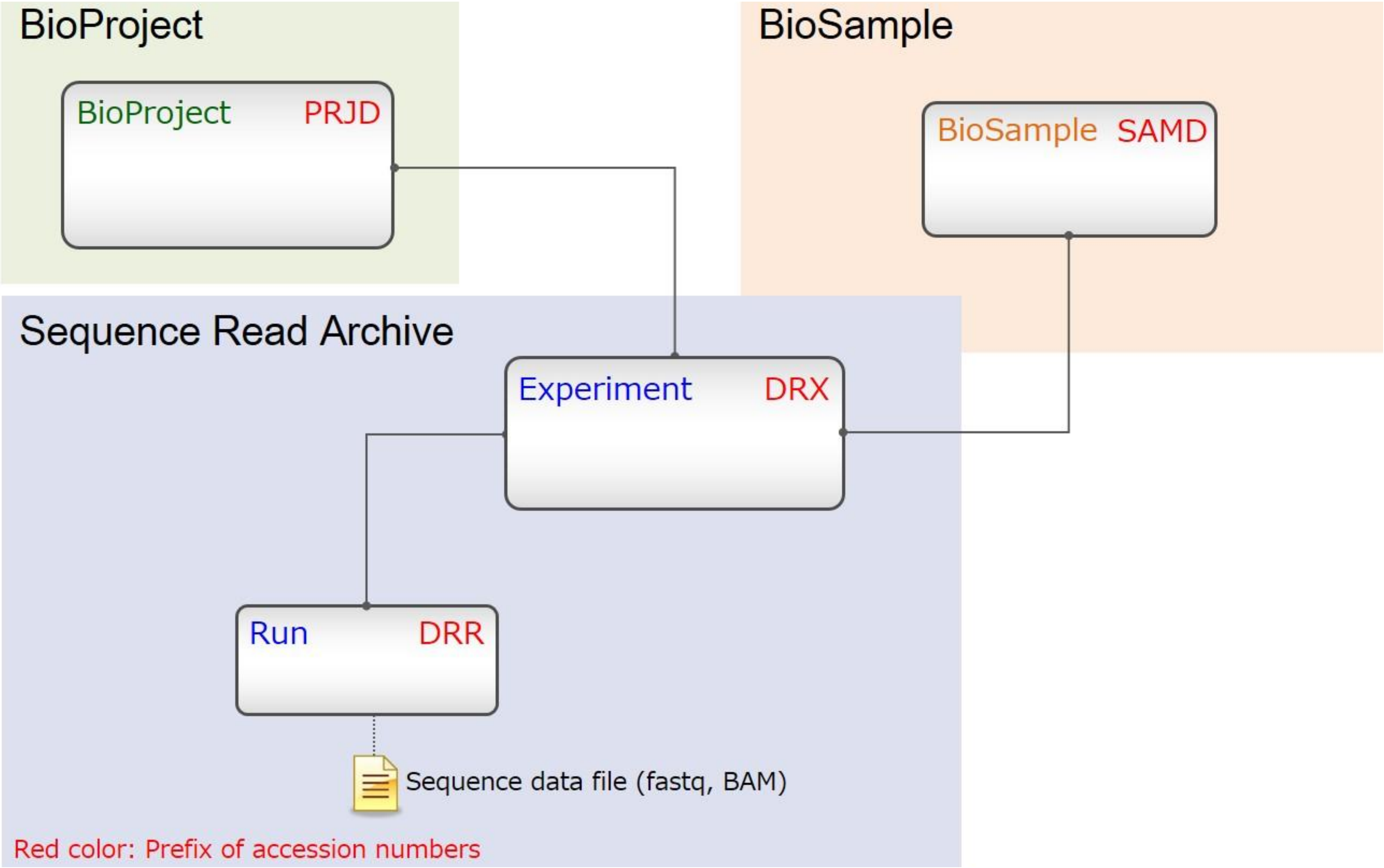
ENA accessions

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

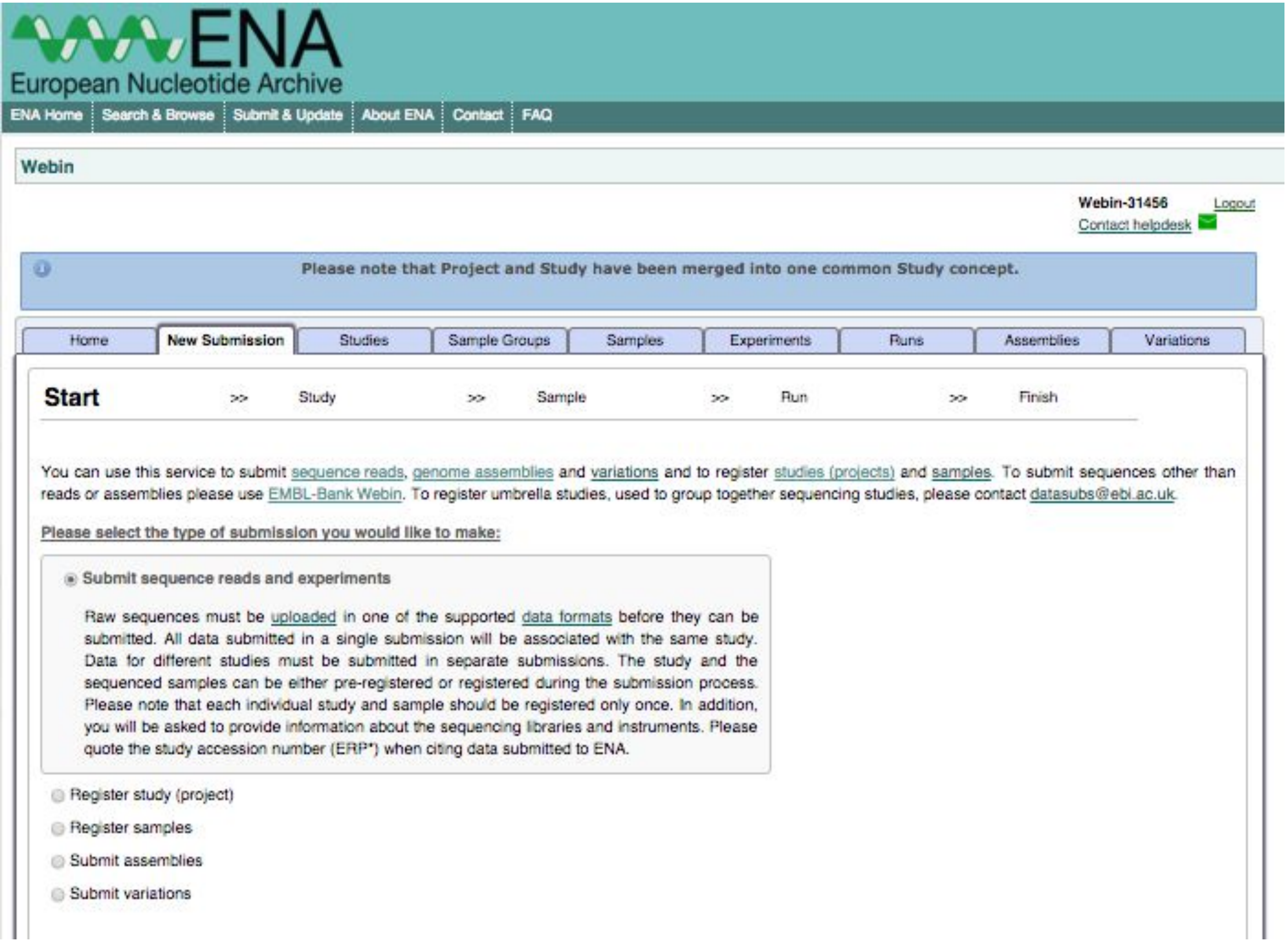
ENA accessions

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

DDBJ data model



ENA data submission



Please note that Project and Study have been merged into one common Study concept.

You can use this service to submit sequence reads, genome assemblies and variations and to register studies (projects) and samples. To submit sequences other than reads or assemblies please use EMBL-Bank Webin. To register umbrella studies, used to group together sequencing studies, please contact datasubs@ebi.ac.uk.

Please select the type of submission you would like to make:

● Submit sequence reads and experiments

Raw sequences must be uploaded in one of the supported data formats before they can be submitted. All data submitted in a single submission will be associated with the same study. Data for different studies must be submitted in separate submissions. The study and the sequenced samples can be either pre-registered or registered during the submission process. Please note that each individual study and sample should be registered only once. In addition, you will be asked to provide information about the sequencing libraries and instruments. Please quote the study accession number (ERP*) when citing data submitted to ENA.

○ Register study (project)

○ Register samples

○ Submit assemblies

○ Submit variations

Browsing ENA

- Let’s browse at
  - http://www.ebi.ac.uk/ena
  - PRJEB6352

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
  - http://www.ebi.ac.uk/eva/?Study%20Browser&browserType=sgv

Array Express

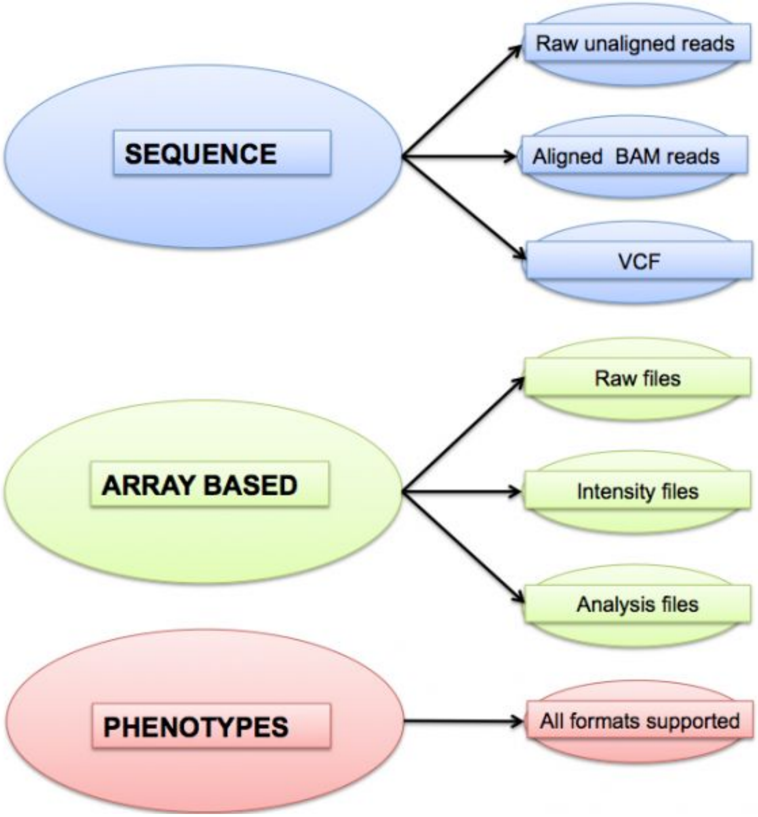
- For functional genomics data from array and sequencing based experiments (RNA-Seq, CHiP-Seq)
  - raw e.g. Affymetrix CEL files, fastq files
  - processed e.g. aligned bam, txt files of read counts
- Data submission is via ‘Annotare’ web interface
- NCBI equivalent is GEO

Browsing ArrayExpress

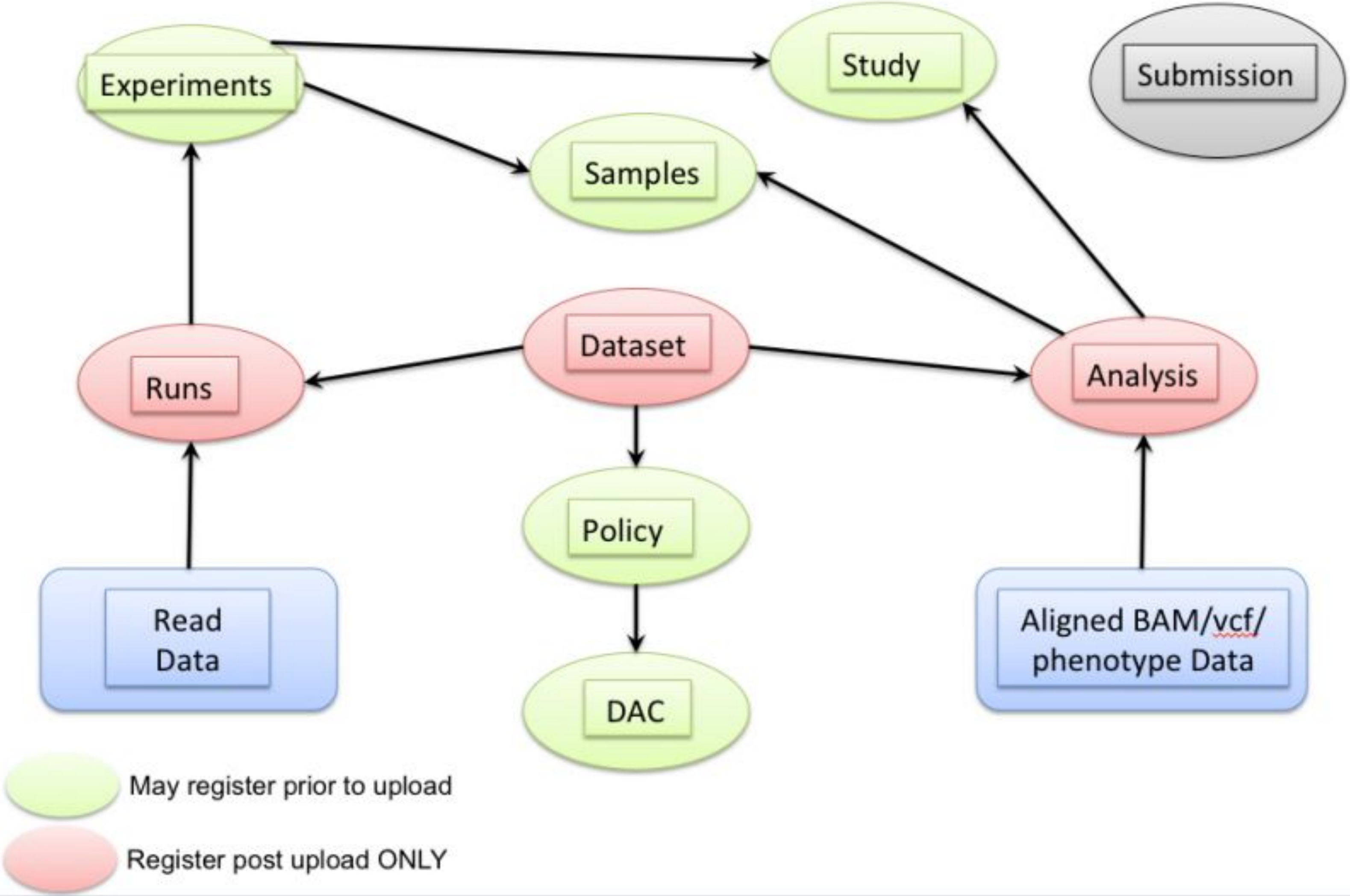
- Let’s browse at
  - https://www.ebi.ac.uk/arrayexpress/browse.html

European Genome-phenome Archive (EGA)

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



EGA data model



Experiments

Study

Submission

Samples

Runs

Dataset

Analysis

Policy

Read Data

DAC

Aligned BAM/vcf/ phenotype Data

🟩 May register prior to upload

🟥 Register post upload ONLY

EGA accessions

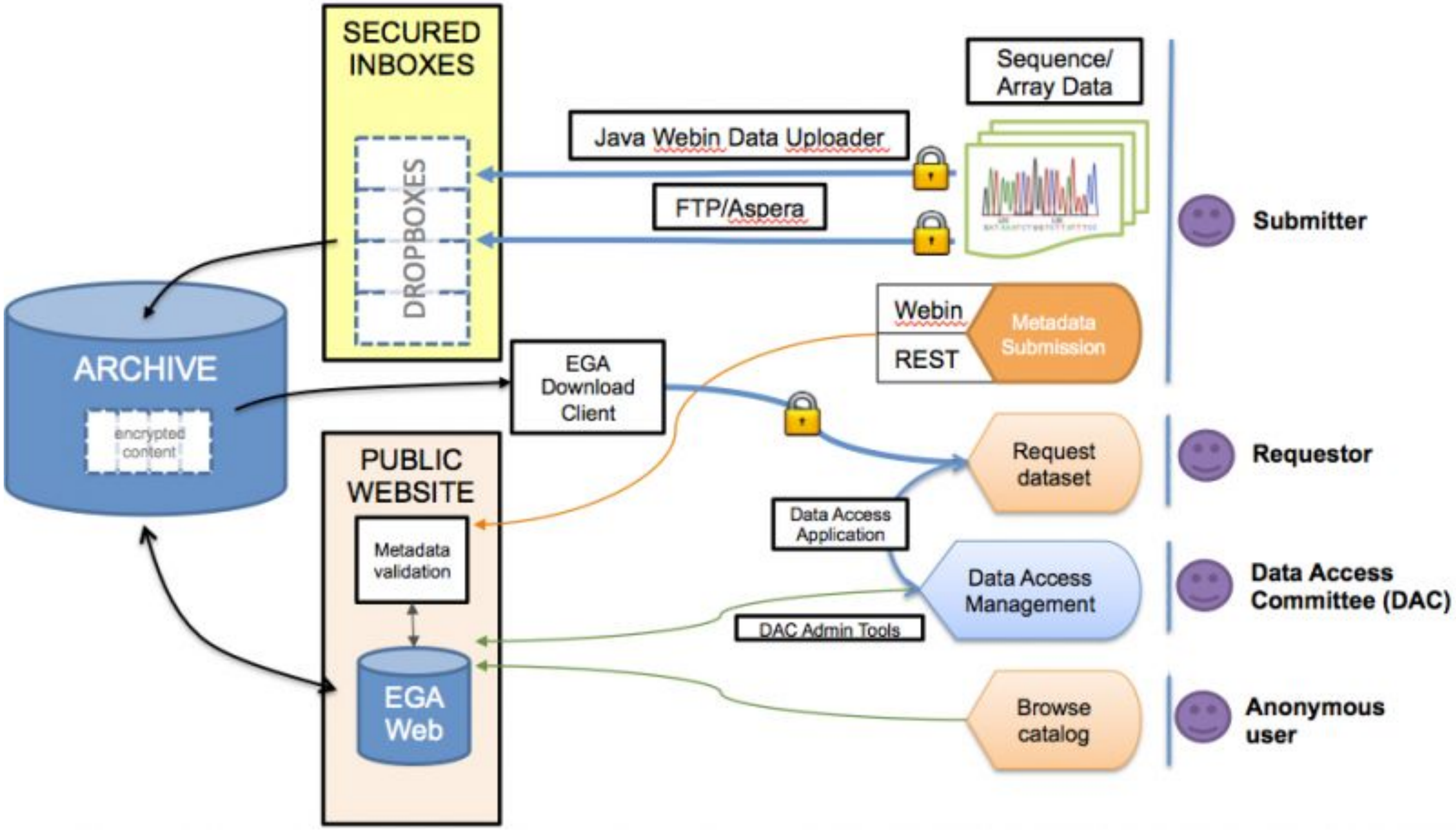
Type	Accession	Description
Study	EGAS	Information about the sequencing study
Sample	EGAN	Information about the samples sequenced
Experiment	EGAX	Information about sequencing experiment including platform used and library information
Run	EGAR	Raw data files containing sequence data (CRAM, BAM, Fastq)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

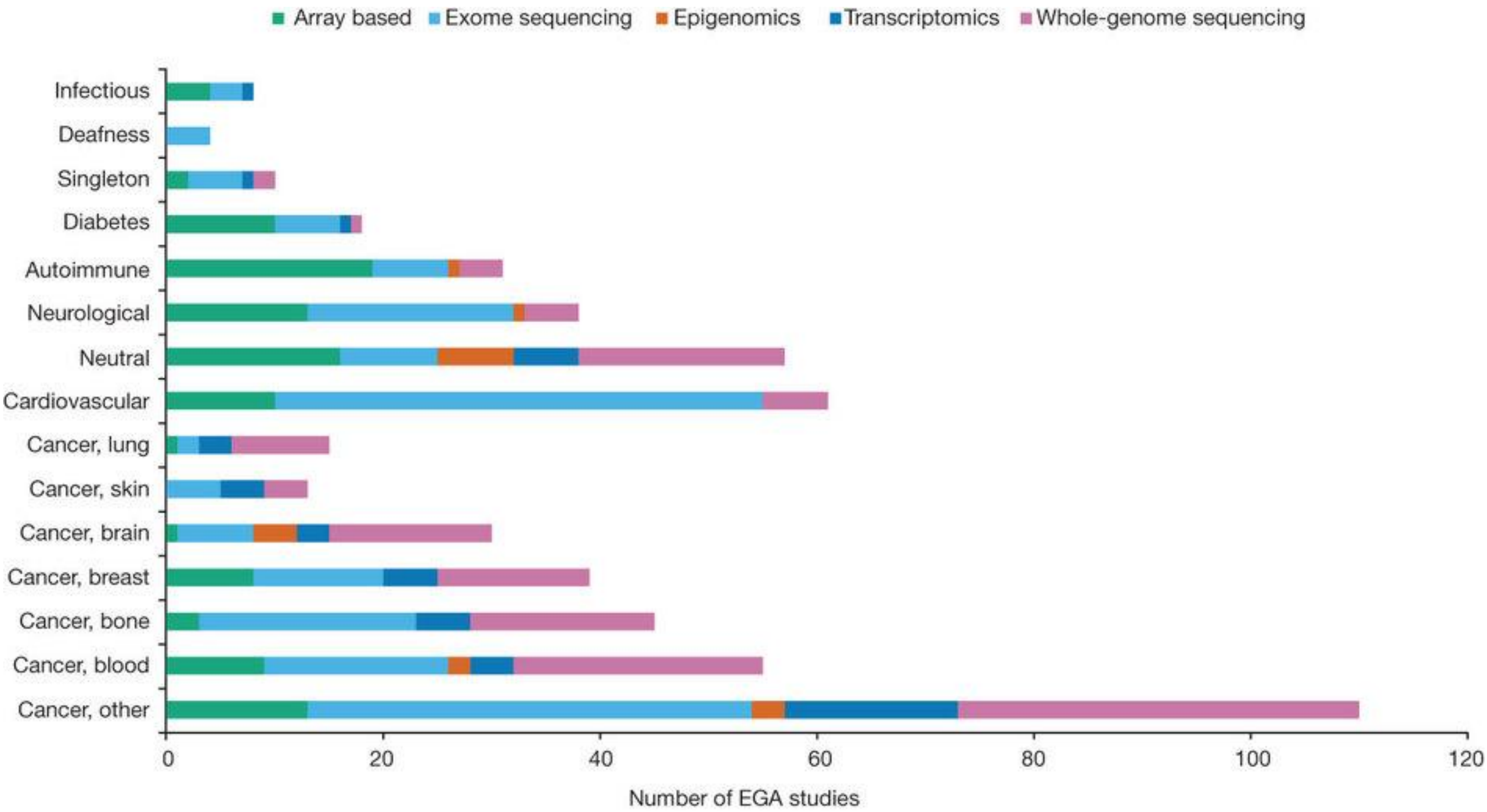
Type	Accession	Description
Study	EGAS	Information about the sequencing study
Sample	**EGAN**	Information about the samples sequenced
Experiment	EGAX	Information about sequencing experiment including platform used and library information
Run	**EGAR**	Raw data files containing sequence data (CRAM, BAM,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

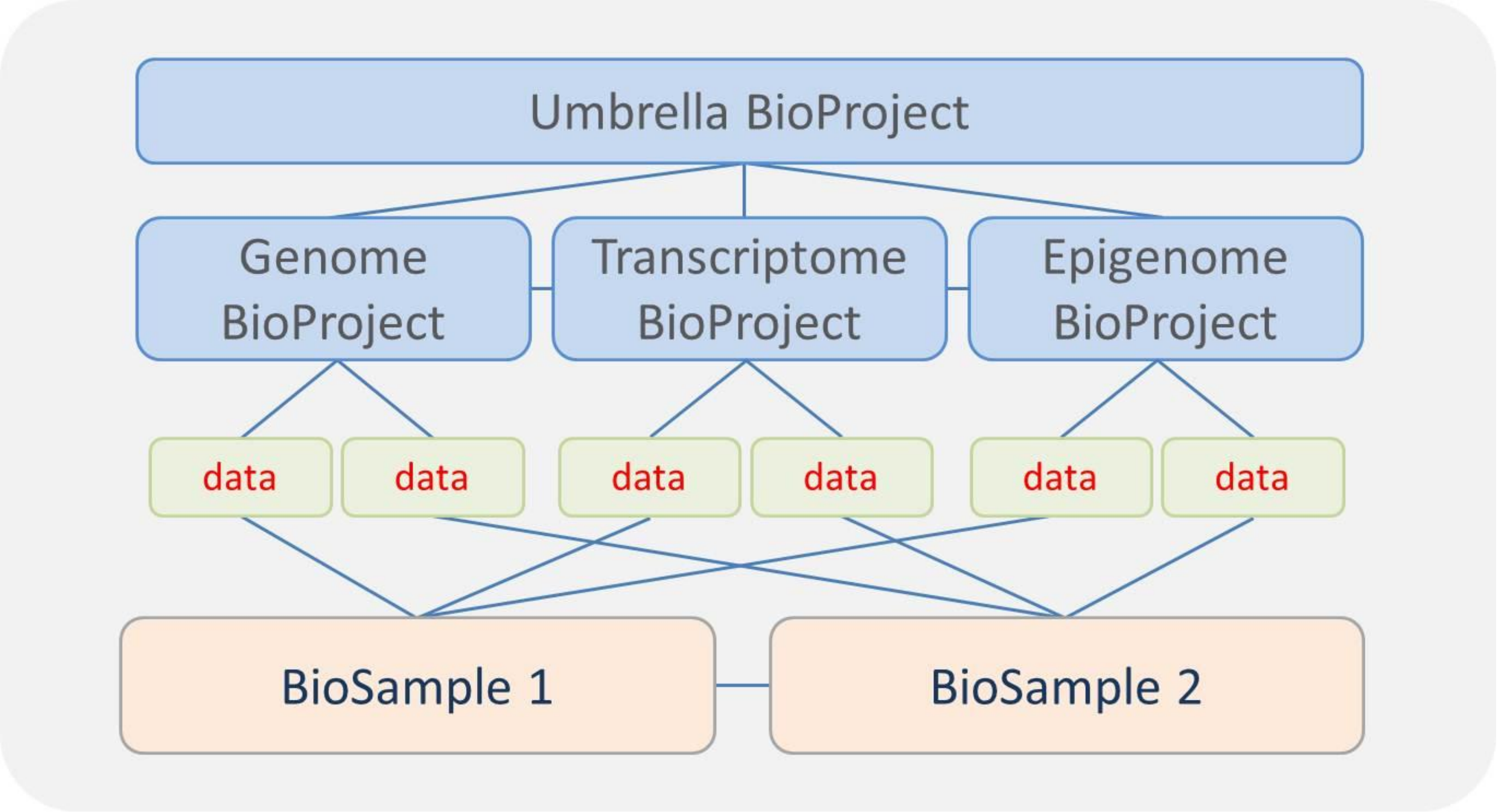
- Strict protocols govern how information is managed, stored and distributed



Breakdown of EGA studies (2014)



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 individual

- Can link up data from different archives for 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

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

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
  - http://www.ncbi.nlm.nih.gov/books/NBK143764/
- DDBJ Training
  - http://trace.ddbj.nig.ac.jp/index_e.html
- NAR Journal
  - http://nar.oxfordjournals.org/