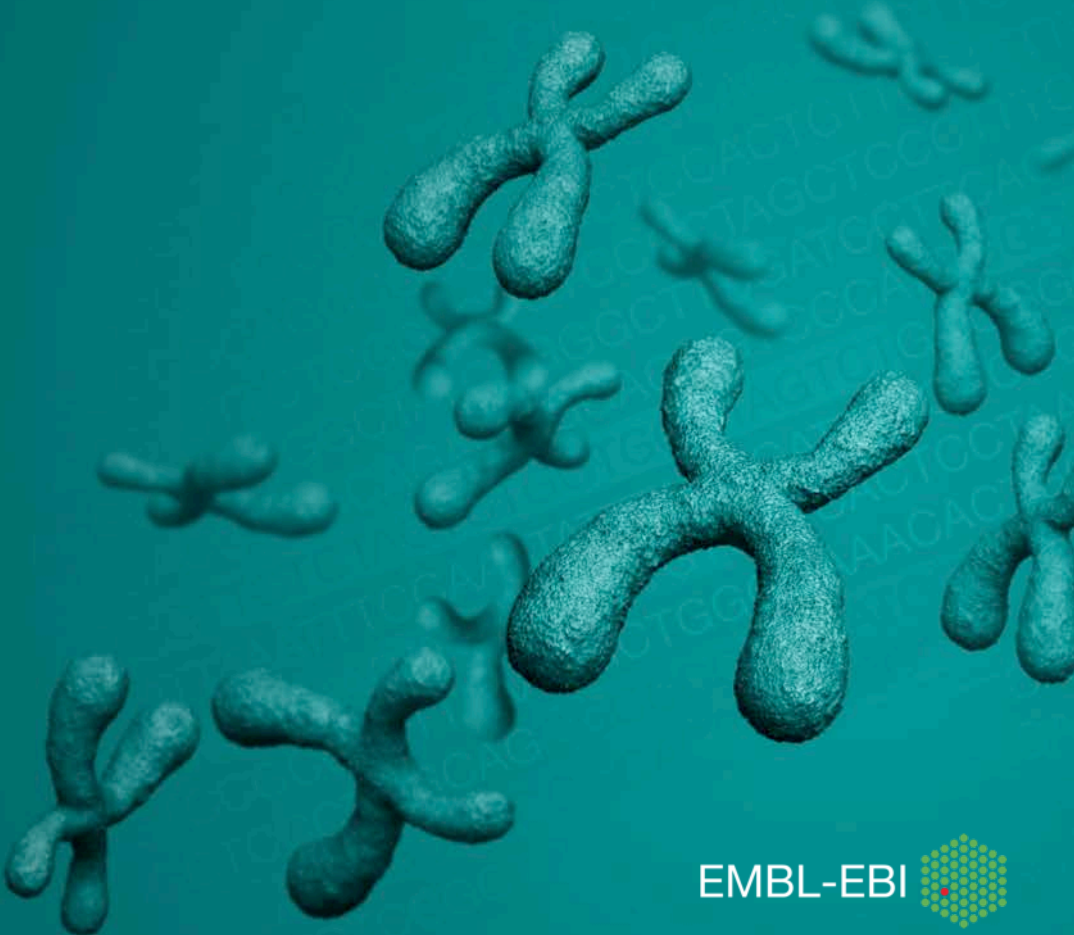
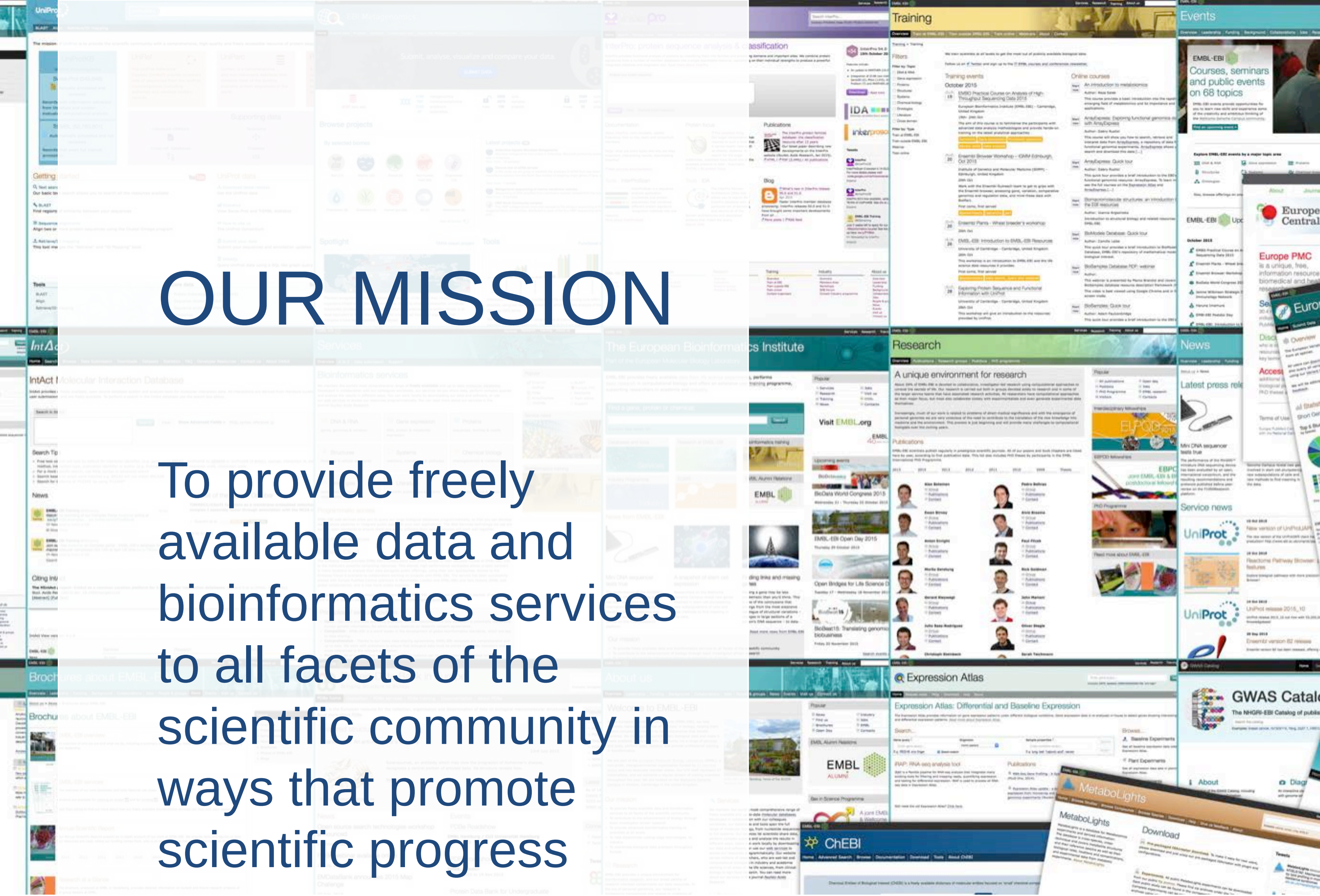
Sharing data to support Science



Laura Clarke

12th July 2016

EMBL-EBI

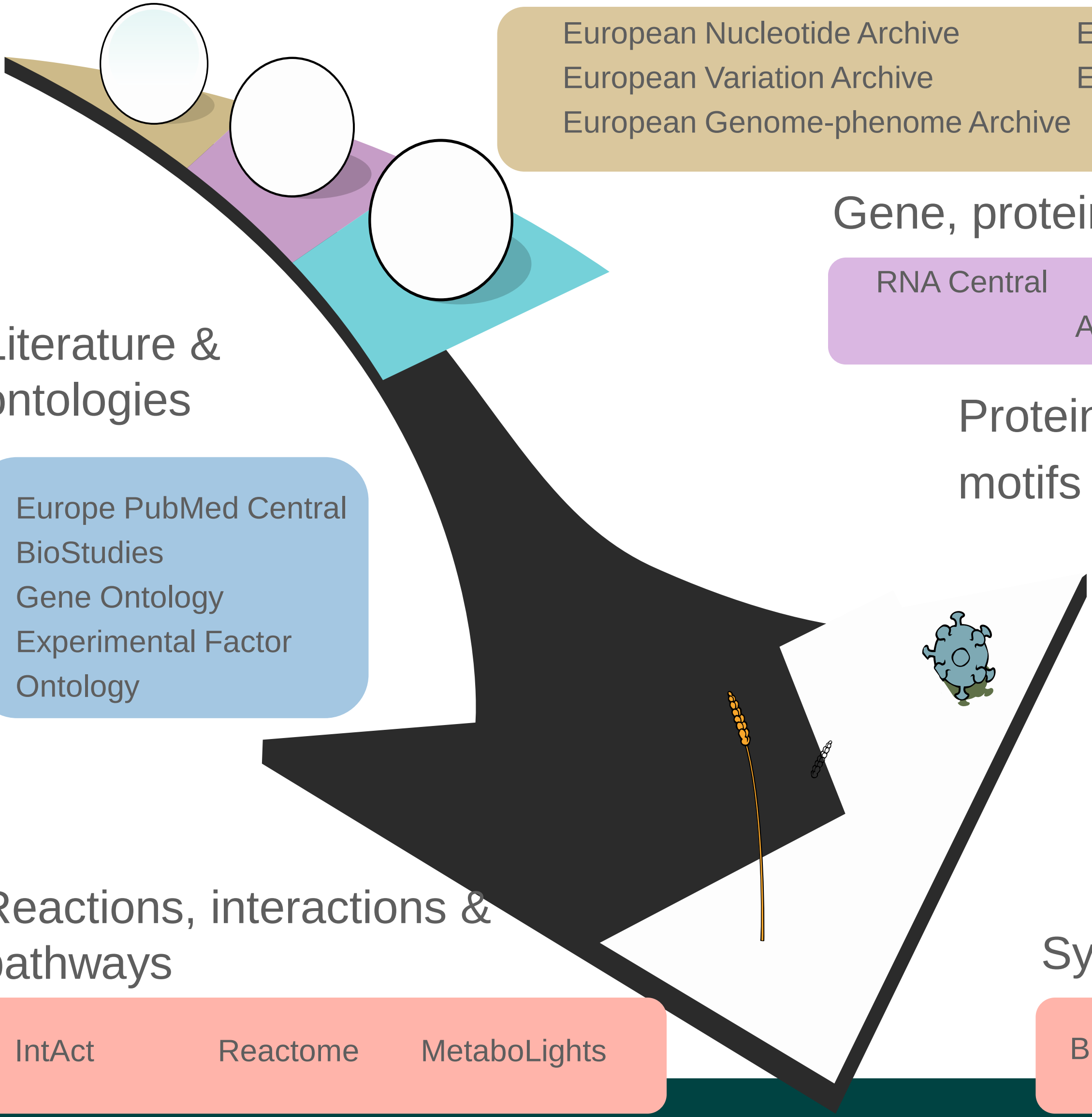


OUR MISSION

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

EMBL-EBI

Data resources at EMBL-EBI



Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
European Genome-phenome Archive
Ensembl
Ensembl Genomes
GWAS Catalog
Metagenomics portal

Gene, protein & metabolite expression

RNA Central   Express   Metabolights
Array Expression Atlas   PRIDE

Protein sequences, families & motifs

InterPro   Pfam   UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

ChEMBL   SureChEMBL   ChEBI

Systems

BioModels   Enzyme Portal   BioSamples

Literature & ontologies

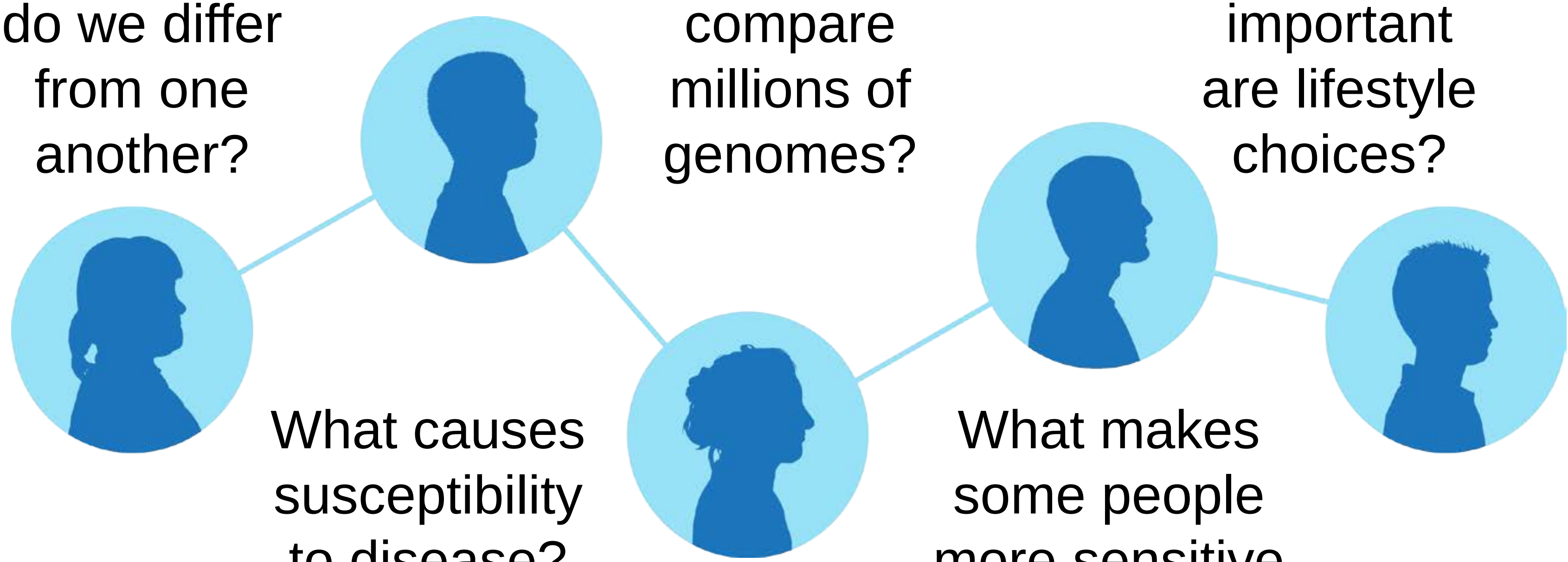
Europe PubMed Central
BioStudies
Gene Ontology
Experimental Factor Ontology

Reactions, interactions & pathways

IntAct   Reactome   MetaboLights

EMBL-EBI

Interpreting human biology



How and why do we differ from one another?

How do you compare millions of genomes?

How important are lifestyle choices?

What causes susceptibility to disease?

What makes some people more sensitive to drugs?

Data Archives

EVA

European Variation Archive

BioStudies

PRIDE

DGVarchive

Array Express

Metabo lights

EUROPEAN GENOME-PHENOME ARCHIVE

ENA

European Nucleotide Archive

BioSD

Added Value Services

e!Ensembl

Gene Expression atlas

Ve!P

GWAS Catalog

REACTOME
A CURATED PATHWAY DATABASE

Open Targets

Locus·Reference·Genomic

Europe PMC

UniProt

Sharing Good Data

Good data is well described data

- Needs
  - Well structured
  - Consistent naming
  - Specific descriptions
- Enables
  - Aggregation
  - Integration
  - Tracking

Motivations

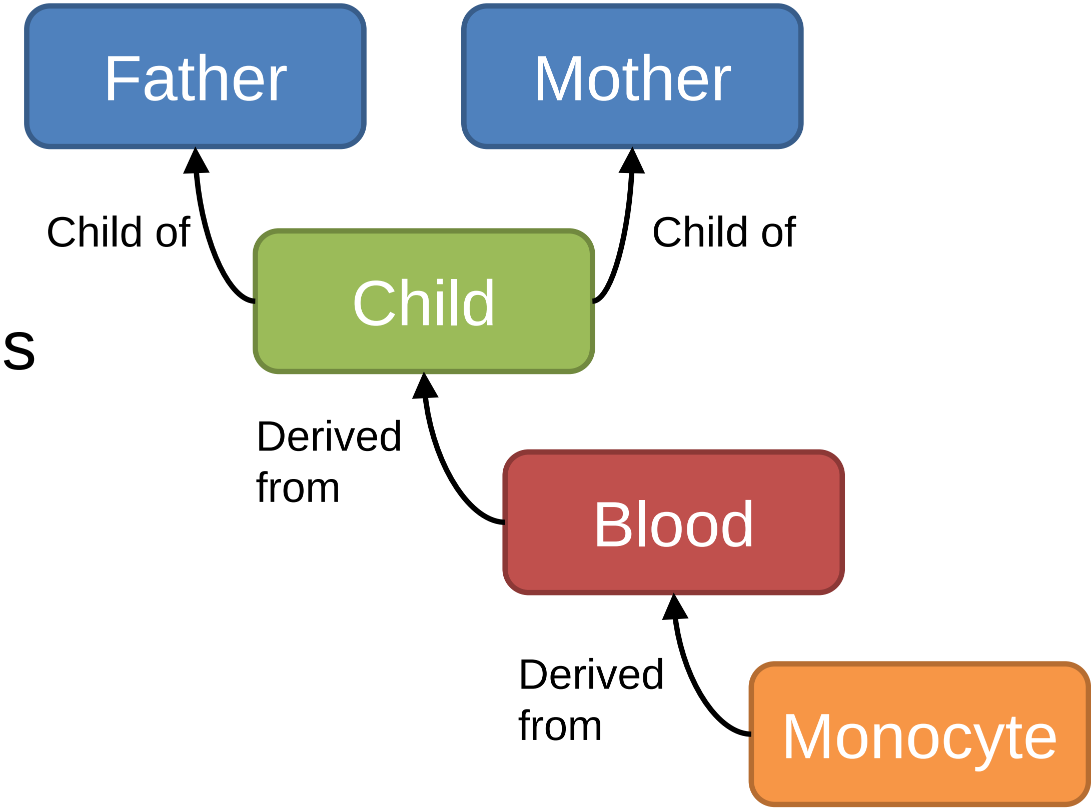
- Make your data usable
  - Reduce ambiguity
  - Facilitate reproduction of results
  - Improve integration across labs, projects and data modalities
- Make your data discoverable
  - Other researchers
  - Informatics services (Ensembl, Gene Expression Atlas)
- Improve your analysis
  - Easier to find batch effects and confounding factors

BioSamples

BioSamples

BioSD

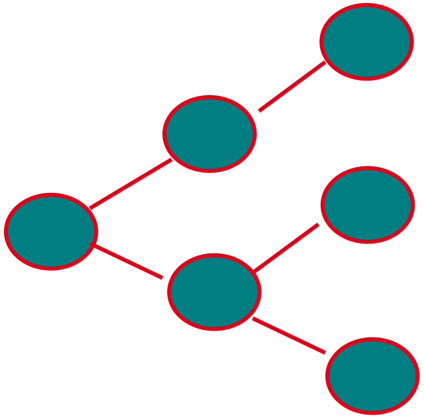
- Provides
  - Unique Identifier
    - SAMEA3105765
  - User defined attributes
  - Ontology Support
  - Relationship support
    - Child of
    - Derived from
    - Same as



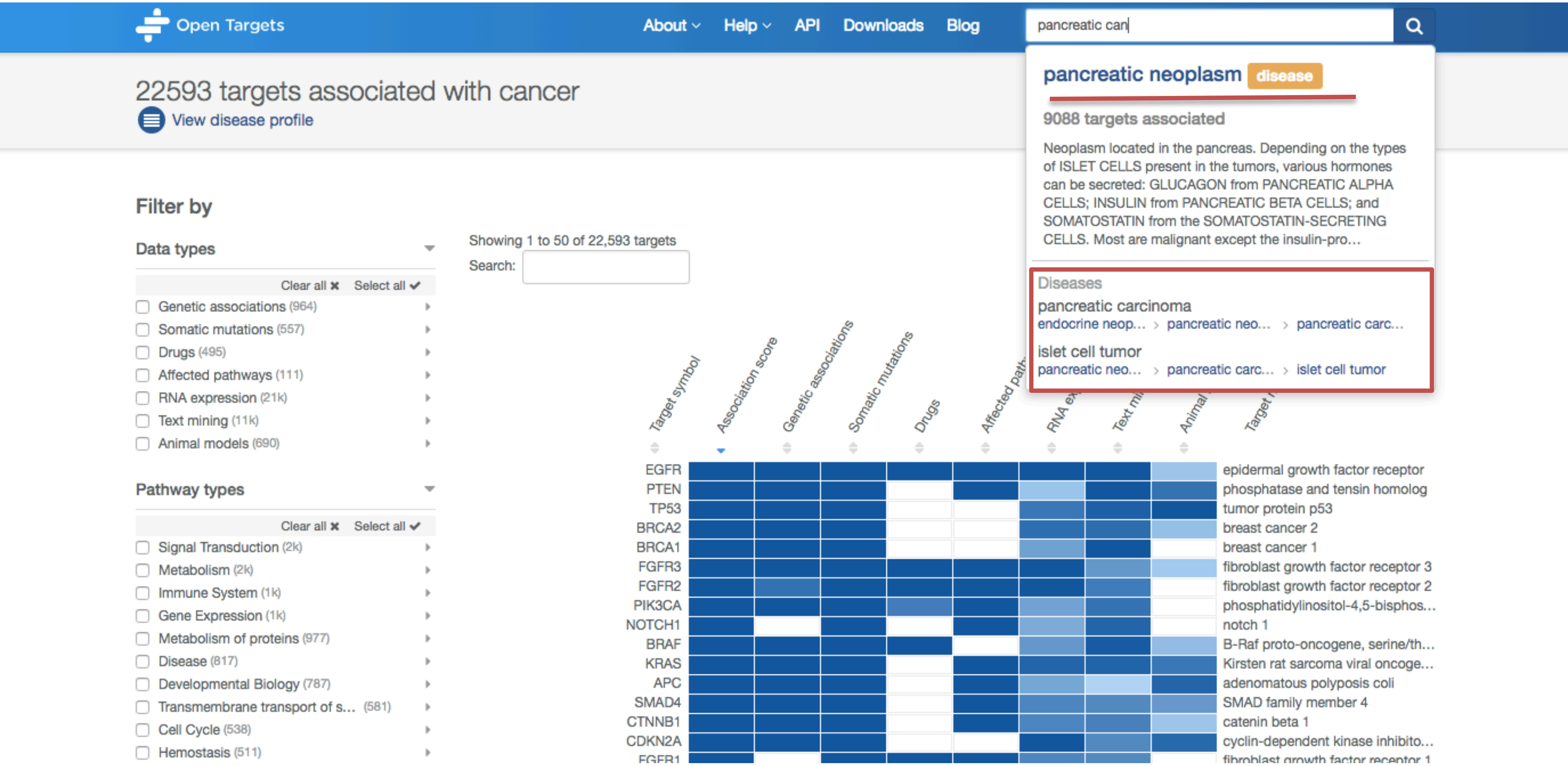
Ontology Support

BioSD

- An ontology is
  - A classification of the kinds of entities that exist.
  - A specification of the meanings of terms in a conceptual vocabulary.
- Ontologies make it easier to
  - Understand what the description means
  - Enable more intuitive searching of the data



Ontology driven search



pancreatic can

pancreatic neoplasm disease

9088 targets associated

Neoplasm located in the pancreas. Depending on the types of ISLET CELLS present in the tumors, various hormones can be secreted: GLUCAGON from PANCREATIC ALPHA CELLS; INSULIN from PANCREATIC BETA CELLS; and SOMATOSTATIN from the SOMATOSTATIN-SECRETING CELLS. Most are malignant except the insulin-pro...

Diseases

pancreatic carcinoma
endocrine neop... › pancreatic neo... › pancreatic carc...

islet cell tumor
pancreatic neo... › pancreatic carc... › islet cell tumor

22593 targets associated with cancer

View disease profile

Filter by

Data types

Clear all ✖ Select all ✔

- □ Genetic associations (964)
- □ Somatic mutations (557)
- □ Drugs (495)
- □ Affected pathways (111)
- □ RNA expression (21k)
- □ Text mining (11k)
- □ Animal models (690)

Pathway types

Clear all ✖ Select all ✔

- □ Signal Transduction (2k)
- □ Metabolism (2k)
- □ Immune System (1k)
- □ Gene Expression (1k)
- □ Metabolism of proteins (977)
- □ Disease (817)
- □ Developmental Biology (787)
- □ Transmembrane transport of s... (581)
- □ Cell Cycle (538)
- □ Hemostasis (511)

Showing 1 to 50 of 22,593 targets

Search:

EGFR	epidermal growth factor receptor
PTEN	phosphatase and tensin homolog
TP53	tumor protein p53
BRCA2	breast cancer 2
BRCA1	breast cancer 1
FGFR3	fibroblast growth factor receptor 3
FGFR2	fibroblast growth factor receptor 2
PIK3CA	phosphatidylinositol-4,5-bisphos...
NOTCH1	notch 1
BRAF	B-Raf proto-oncogene, serine/th...
KRAS	Kirsten rat sarcoma viral oncoge...
APC	adenomatous polyposis coli
SMAD4	SMAD family member 4
CTNNB1	catenin beta 1
CDKN2A	cyclin-dependent kinase inhibito...
FGFR1	fibroblast growth factor receptor 1

https://www.targetvalidation.org

Ontology Annotation

- Finding the right term can be challenging
- EMBL-EBI has tools to help
- Zooma
  - Using past knowledge to inform new annotation
  - http://www.ebi.ac.uk/spot/zooma/
- Ontology Lookup service (OLS)
  - Indexes 150 biomedical ontologies (4.5 million terms, 11 million relations)
  - Web and programmatic interfaces
  - http://www.ebi.ac.uk/ols

Sharing Data

Data Archives

BioStudies

EVA
European Variation Archive

PRIDE

Array Express

Metabo lights

DGVarchive

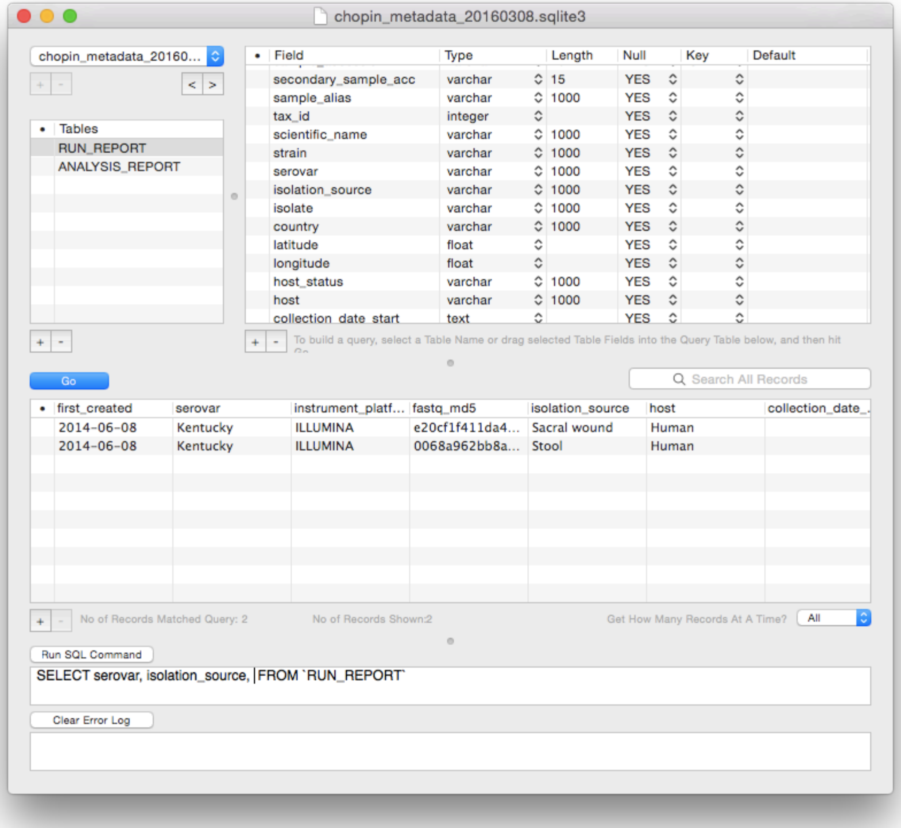
EUROPEAN GENOME-PHENOME ARCHIVE

ENA
European Nucleotide Archive

BioSD

Data Hubs

- ENA tool available to collaborators
- Allows consortium wide access to targeted prepublication and public data
- Data and metadata
- Supported data types
  - Sequence
  - Alignment
  - Variants
  - Assemblies



chopin_metadata_20160308.sqlite3

Tables
RUN_REPORT
ANALYSIS_REPORT

Field	Type	Length	Null	Key	Default
secondary_sample_acc	varchar	15	YES		
sample_alias	varchar	1000	YES		
tax_id	integer		YES		
scientific_name	varchar	1000	YES		
strain	varchar	1000	YES		
serovar	varchar	1000	YES		
isolation_source	varchar	1000	YES		
isolate	varchar	1000	YES		
country	varchar	1000	YES		
latitude	float		YES		
longitude	float		YES		
host_status	varchar	1000	YES		
host	varchar	1000	YES		
collection_date_start	text		YES		

To build a query, select a Table Name or drag selected Table Fields into the Query Table below, and then hit Go

first_created	serovar	instrument_platf...	fastq_md5	isolation_source	host	collection_date_.
2014-06-08	Kentucky	ILLUMINA	e20cf1f411da4...	Sacral wound	Human	
2014-06-08	Kentucky	ILLUMINA	0068a962bb8a...	Stool	Human	

No of Records Matched Query: 2 No of Records Shown:2 Get How Many Records At A Time? All

```
SELECT serovar, isolation_source, FROM `RUN_REPORT`
```

Sharing controlled access data

What is controlled access data

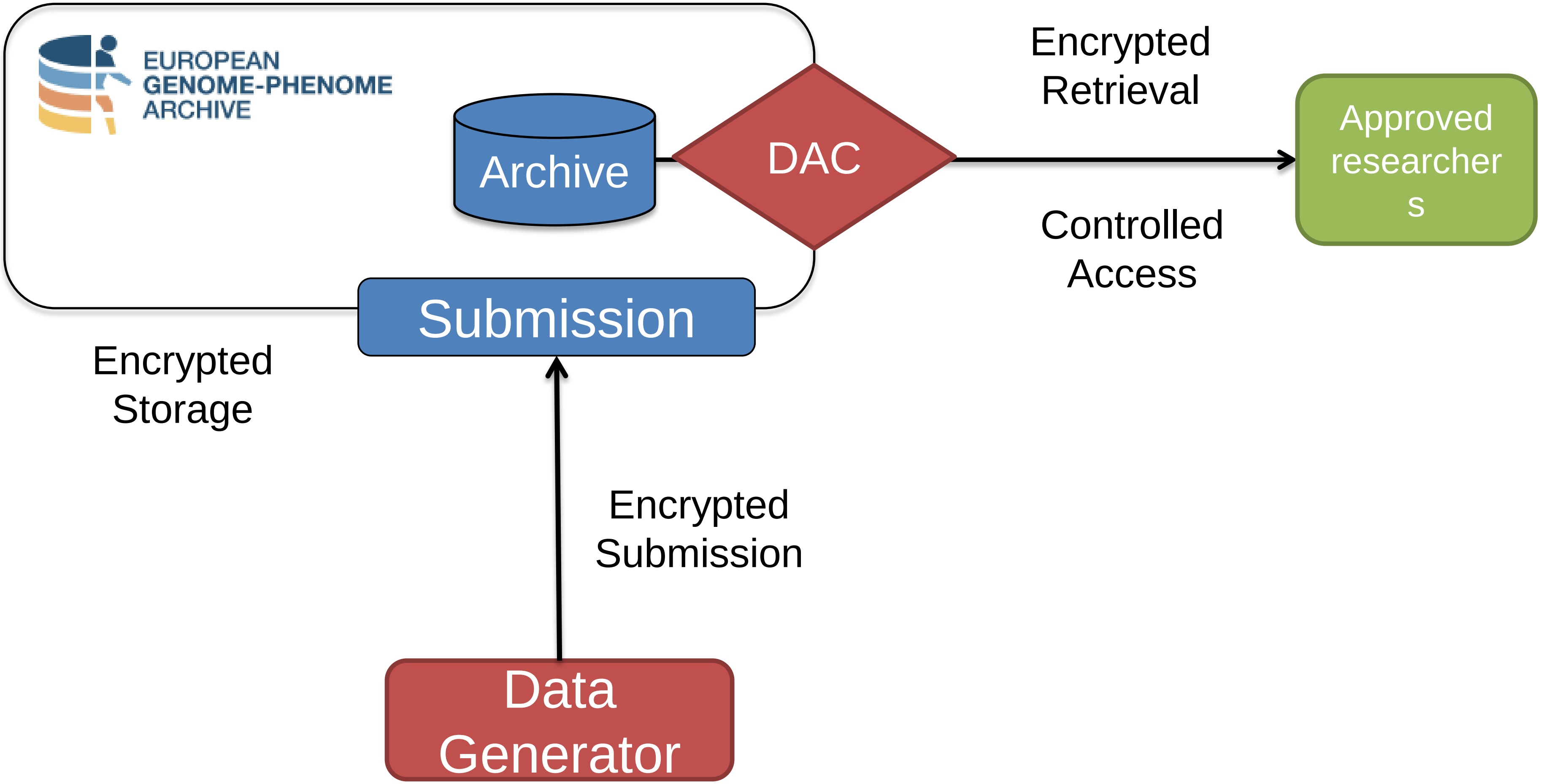
- Driven by participant consent
- Some consents specify data is released
  - In a controlled manner
  - To bone fide researchers
- This data tends to be
  - Human, personally identifiable data types (‘raw’, processed, phenotypic)
  - Affiliated to bio-medical research or consortium projects

What is the EGA?

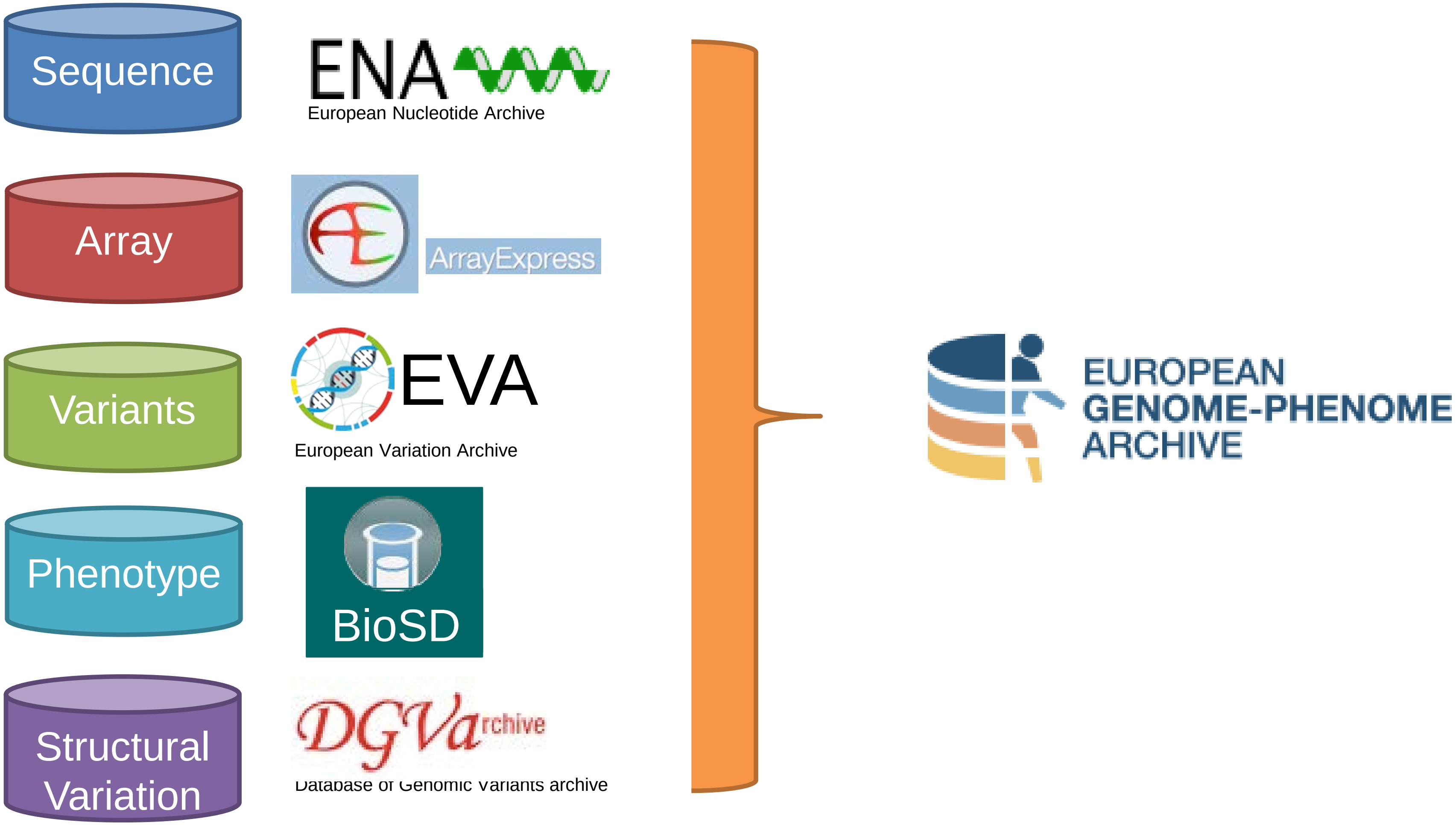
EUROPEAN GENOME-PHENOME ARCHIVE

- Launched 14th July 2008
- Role: **secure** archive for **controlled** distribution of **consented** genetic & phenotypic data
- **8000+** data access accounts; **395** submission accounts
- **~2.5PB** available for download; **~1900** datasets; **1.6PB** distributed last year
- **+200** contacts to Helpdesk/month

EGA - Architecture



Where do I submit data?



Sequence

ENA

European Nucleotide Archive

Array

ArrayExpress

Variants

EVA

European Variation Archive

Phenotype

BioSD

Structural Variation

DGVarchive

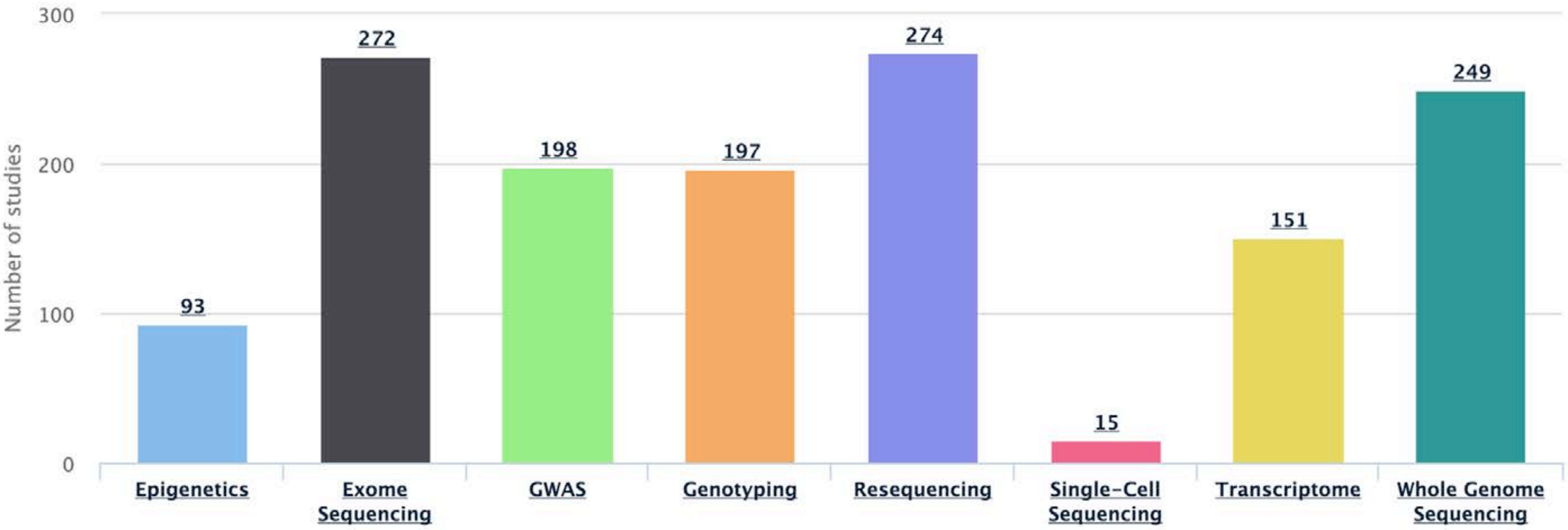
Database of Genomic Variants archive

EUROPEAN GENOME-PHENOME ARCHIVE

Open & public archives

Controlled archives

EGA Studies by Technology



EUROPEAN GENOME-PHENOME ARCHIVE

Added Value Services

e!Ensembl

Gene Expression atlas

Ve!P

GWAS Catalog

REACTOME
A CURATED PATHWAY DATABASE

Open Targets

Locus·Reference·Genomic

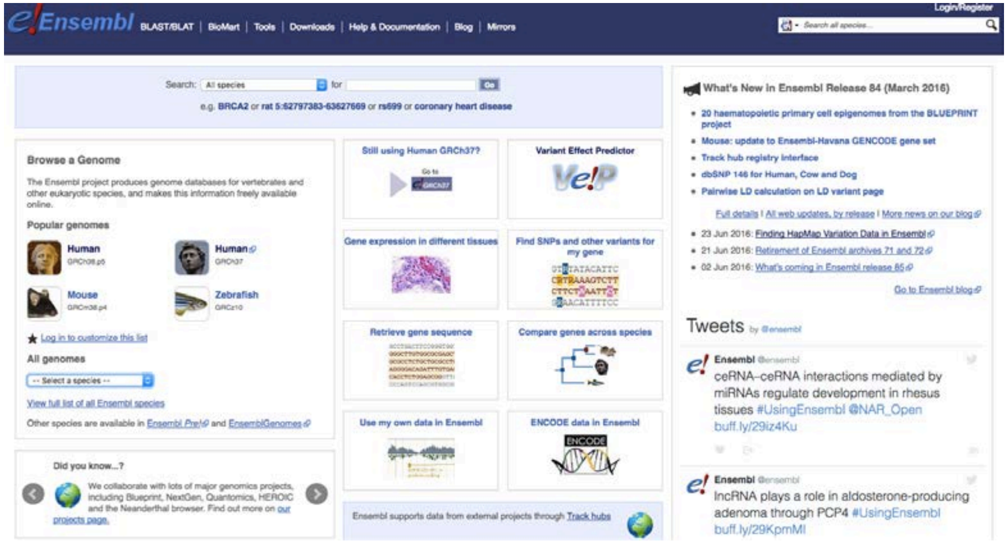
Europe PMC

UniProt

Putting you data in context

Ensembl

- Genome Browser
  - 65 vertebrate species
- Open Data
- Open Code
- High quality annotation
  - Genes
  - Regulatory regions
- Comparative Analysis
- Variation Data
- Data mining tools



Ensembl
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species...
Search: All species for Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human GRCh38.p5 · Human GRCh37 · Mouse GRCm38.p4 · Zebrafish GRCz10

★ Log in to customize this list

All genomes

-- Select a species --

View full list of all Ensembl species

Other species are available in Ensembl Pre! and EnsemblGenomes

Did you know...?

We collaborate with lots of major genomics projects, including Blueprint, NextGen, Quantomics, HEROIC and the Neanderthal browser. Find out more on our projects page.

Still using Human GRCh37?
Go to GRCh37
Variant Effect Predictor
VEP
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
ENCODE data in Ensembl
ENCODE

Ensembl supports data from external projects through Track hubs

What's New in Ensembl Release 84 (March 2016)

- 20 haematopoietic primary cell epigenomes from the BLUEPRINT project
- Mouse: update to Ensembl-Havana GENCODE gene set
- Track hub registry interface
- dbSNP 146 for Human, Cow and Dog
- Pairwise LD calculation on LD variant page

Full details | All web updates, by release | More news on our blog

- 23 Jun 2016: Finding HapMap Variation Data in Ensembl
- 21 Jun 2016: Retirement of Ensembl archives 71 and 72
- 02 Jun 2016: What's coming in Ensembl release 85

Go to Ensembl blog

Tweets by @ensembl

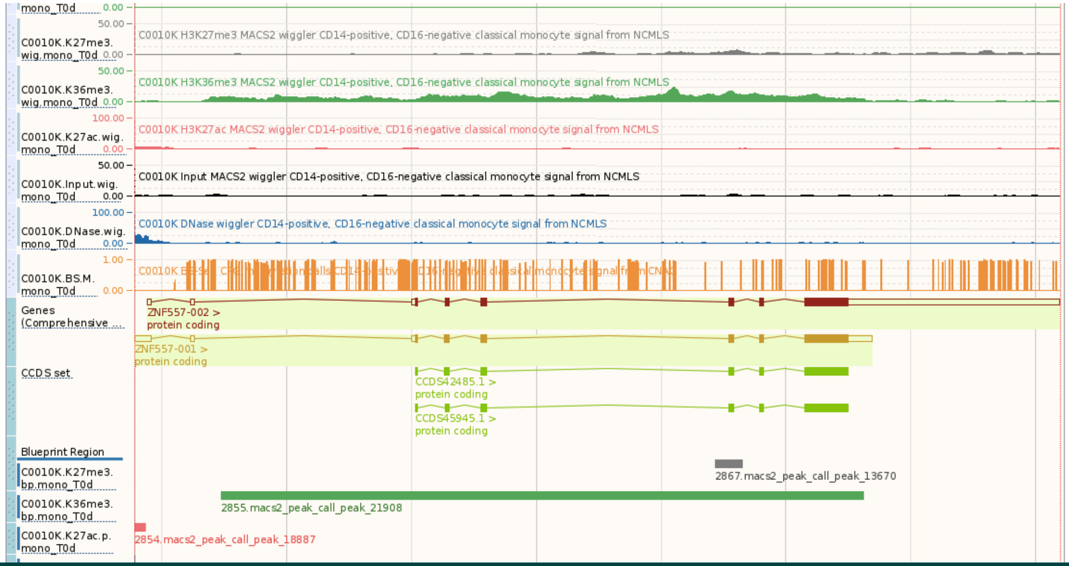
Ensembl @ensembl
ceRNA–ceRNA interactions mediated by miRNAs regulate development in rhesus tissues #UsingEnsembl @NAR_Open buff.ly/29iz4Ku

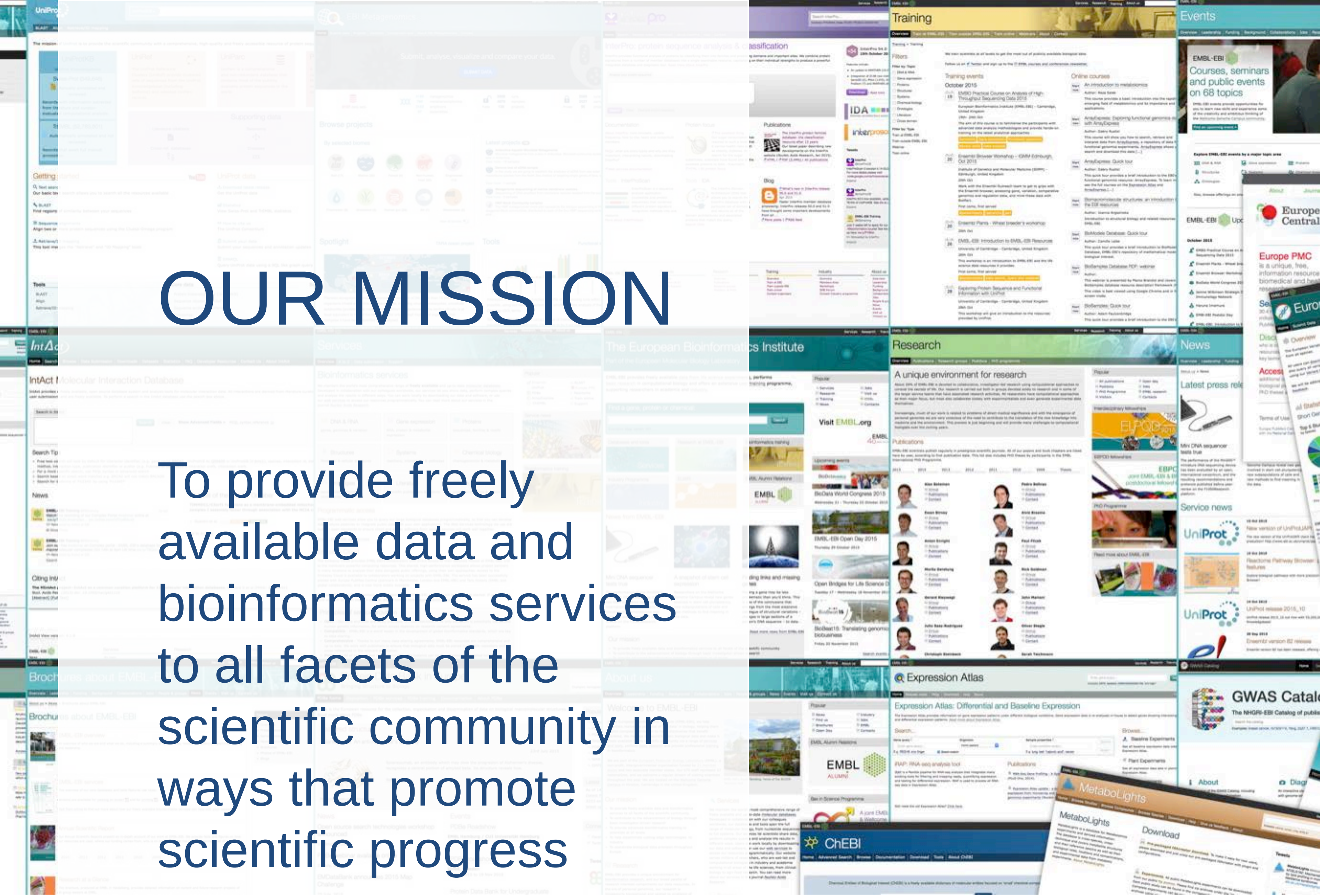
Ensembl @ensembl
lncRNA plays a role in aldosterone-producing adenoma through PCP4 #UsingEnsembl buff.ly/29KpmMI

EMBL-EBI

Trackhubs

- Publicly hosted text file
- Supports indexed “Big” formats
- See your data in context of broad genomic annotation
- Attach many files to Ensembl (or UCSC) at once
- Provide community with clearly defined data collections


mono_T0d 0.00
50.00
C0010K.K27me3.wig.mono_T0d 0.00
C0010K H3K27me3 MACS2 wiggler CD14-positive, CD16-negative classical monocyte signal from NCMLS
50.00
C0010K.K36me3.wig.mono_T0d 0.00
C0010K H3K36me3 MACS2 wiggler CD14-positive, CD16-negative classical monocyte signal from NCMLS
100.00
C0010K.K27ac.wig.mono_T0d 0.00
C0010K H3K27ac MACS2 wiggler CD14-positive, CD16-negative classical monocyte signal from NCMLS
50.00
C0010K.Input.wig.mono_T0d 0.00
C0010K Input MACS2 wiggler CD14-positive, CD16-negative classical monocyte signal from NCMLS
100.00
C0010K.DNase.wig.mono_T0d 0.00
C0010K DNase wiggler CD14-positive, CD16-negative classical monocyte signal from NCMLS
1.00
C0010K.BS.M.mono_T0d 0.00
C0010K BS-Seq CpG methylation calls CD14-positive, CD16-negative classical monocyte signal from CNAG
Genes (Comprehensive ...
ZNF557-002 > protein coding
ZNF557-001 > protein coding
CCDS set
CCDS42485.1 > protein coding
CCDS45945.1 > protein coding
Blueprint Region
C0010K.K27me3.bp.mono_T0d
2867.macs2_peak_call_peak_13670
C0010K.K36me3.bp.mono_T0d
2855.macs2_peak_call_peak_21908
C0010K.K27ac.p.mono_T0d
2854.macs2_peak_call_peak_18887


Training
Events
Courses, seminars and public events on 68 topics
Research
A unique environment for research
News
Service news
UniProt
Expression Atlas
Expression Atlas: Differential and Baseline Expression
GWAS Catalog
The NHGRI-EBI Catalog of published
MetaboLights
Download
ChEBI
IntAct Molecular Interaction Database
Europe PMC
The European Bioinformatics Institute
Brochures about EMBL-EBI
InterPro
InterPro: protein sequence analysis & classification
Services
Bioinformatics services
About us
Welcome to EMBL-EBI

OUR MISSION

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

Collaboration opportunities

- Data Coordination Services
- Specialized support for archive use
- Increased data integration with value added services
- New technology development

TRANSFORMING GENETIC
MEDICINE INITIATIVE

BLUEPRINT
epigenome

Global Alliance
for Genomics & Health

IGSR
International Genome Sample Resource

HipSci

EBiSC
European Bank for induced pluripotent Stem Cells

Thanks

Questions?