

# A closer look at the repositories' world

---

Bruna Piereck

15th June 2023, Gent



# Do we know why?

- Why submit data?
- Why should we be aware of the repositories?



# A (somewhat) simple start



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal



# A (somewhat) simple start



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

What is written in the DMP?

# A (somewhat) simple start



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal

What is written in the DMP?

Are you cleared for depositing?

Ethical and legal issues

Data availability requirements of the funder and of the journal



# How to find a repository?



re3data.org  
REGISTRY OF RESEARCH DATA REPOSITORIES

Search... Search

<https://www.re3data.org/>



## Platform highlights

- [ELIXIR Core Data Resources](#): European data resources that are of fundamental importance to research in the life sciences and are committed to the long-term preservation of data.
- [ELIXIR Deposition Databases](#): repositories recommended for the deposition of life sciences experimental data.
- [Data resource services](#): this list is updated as Nodes finalise or review their Service Delivery Plans (see [How countries join](#)).



<https://elixir-europe.org/platforms/data>



FAIRsharing.org  
standards, databases, policies

search through all content

STANDARDS DATABASES POLICIES COLLECTIONS

<https://fairsharing.org/>



scientific data

Explore content ▾ About the journal ▾ Publish with us ▾

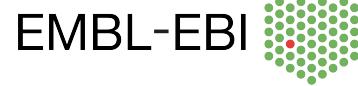
nature > scientific data > policies > data repository guidance

<https://www.nature.com/sdata/policies/repositories>

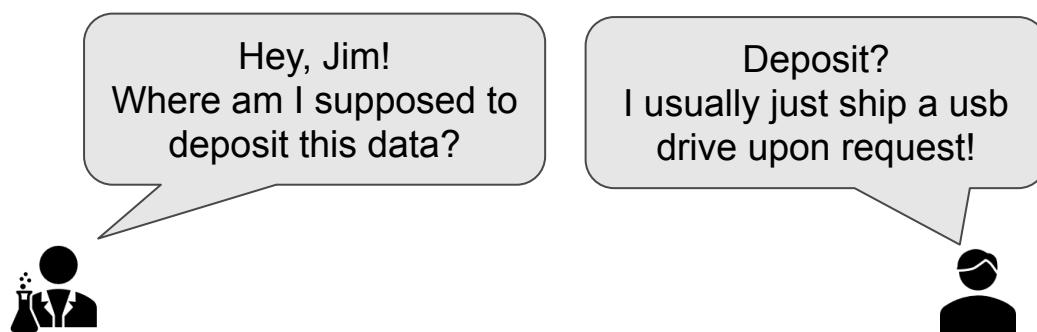
[EMBL-EBI home](#) > [Services](#) > [Data submission](#)

## Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.



<https://www.ebi.ac.uk/submission/>



# Generic vs. Domain specific

Repository Name	Information on fees/costs	Size limits	Integrated with <i>Scientific Data's</i> manuscript submission system	Re3data / FAIRsharing entry
Dryad Digital Repository	\$120 USD for first 20 GB, and \$50 USD for each additional 10 GB	<a href="#">None stated</a>	<b>Yes ✓</b>	<a href="#">view FAIRsharing entry</a>
figshare	100 GB free per <i>Scientific Data</i> manuscript.	1 TB per dataset	<b>Yes ✓</b> - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. <a href="#">Download instructions.</a>	<a href="#">view FAIRsharing entry</a>
Harvard Dataverse	<a href="#">Contact repository</a> for datasets over 1 TB	2.5 GB per file, 10 GB per dataset	No	<a href="#">view re3data entry</a>
Open Science Framework	<a href="#">Free of charge</a>	5 GB per file, multiple files can be uploaded	No	<a href="#">view FAIRsharing entry</a>
Zenodo	<a href="#">Donations towards sustainability encouraged</a>	50 GB per dataset	No	<a href="#">view re3data entry</a>
Science Data Bank	<a href="#">Free of charge</a>	8 GB per file, no limit to dataset size	No	<a href="#">view FAIRsharing entry</a>

<https://www.nature.com/sdata/policies/repositories>

- There are a few generalist repositories out there
- Your institution might have a data repository that fulfils requirements from funders and journals
- They might still not match your data type

Where do you find data?

Which repository can best connect you with the broader community?



# Let's find a repository



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal



<https://www.re3data.org/>



<https://fairsharing.org/>

Identify a couple of potential repositories using re3data and FAIRsharing

- Domain specific
- Justify the choice
- What information can you find about the repository that might help you?



# Other ways to find a repository



## Platform highlights

- ELIXIR Core Data Resources: European data resources that are of fundamental importance to research in the life sciences and are committed to the long-term preservation of data.
- ELIXIR Deposition Databases: repositories recommended for the deposition of life sciences experimental data.
- Data resource services: this list is updated as Nodes finalise or review their Service Delivery Plans (see [How countries join](#)).

<https://elixir-europe.org/platforms/data>

Focus on Life Sciences



# The ELIXIR recommended deposition databases

## ELIXIR Deposition Database list

Deposition Database	Data type	International collaboration framework <sup>1</sup>
ArrayExpress	Functional genomics data. Stores data from high-throughput functional genomics experiments.	
BioModels	Computational models of biological processes.	
BioSamples	BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry.	NCBI BioSamples database
BioStudies	Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured archives.	
EGA	Personally identifiable genetic and phenotypic data resulting from biomedical research projects.	European Bioinformatics Institute and the Centre for Genomic Regulation
EMDB	The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures.	
ENA	Nucleotide sequence information, covering raw sequencing data, contextual data, sequence assembly information and functional and taxonomic annotation.	International Nucleotide Sequence Database Collaboration

IntAct	IntAct provides a freely available, open source database system and analysis tools for molecular interaction data.	The International Molecular Exchange Consortium
MetaboLights	Metabolite structures and their reference spectra as well as their biological roles, locations and concentrations, and experimental data from metabolic experiments.	
PDB	Biological macromolecular structures.	wwPDB
PRIDE	Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence.	The ProteomeXchange Consortium

Further information: Gavin Farrell ([gavin.farrell@elixir-europe.org](mailto:gavin.farrell@elixir-europe.org))

<sup>1</sup> An International collaboration framework enables content sharing on a formal level. This is often signified by a shared Accession Number system, such that data deposited in one database becomes part of the shared data collection, and is also available through other partner portals.

<https://elixir-europe.org/platforms/data/elixir-deposition-databases>

Focus on experimental data



---

Why is it a good idea to deposit your data?



# A short side note on visibility

Why is it a good idea to deposit your data?

Core Data Resource	Data type
Europe PMC	Europe PMC is a repository, providing access to worldwide life sciences articles, books, patents and clinical guidelines.

 Europe PMC    About | Tools | Developers | Help    Europe PMC plus

Search life-sciences literature (42,499,961 articles, preprints and more)

Advanced search   

 Innovative features  
Intuitive and powerful search tools, linked resources and author services help you stay on top of the cutting edge of science. To learn more, see Why use Europe PMC.

 Comprehensive search  
Search life sciences literature from trusted sources around the globe, all in one search, accessible by anyone anywhere, for free. Learn more About Europe PMC.

 Trusted partnerships  
Europe PMC is the partner of PubMed Central (PMC), an ELIXIR core data resource, and the repository of choice for many international science Funders.

[COVID-19 full text preprints initiative →](#)

**Data citations or links search**  
Data citations are mentions of accession numbers (unique identifiers for data) in the text of articles. Data citations are available for data in a number of different databases including: European Nucleotide Archive [↗](#), UniProt, PDBe [↗](#), OMIM [↗](#), RefSNP [↗](#), RefSeq [↗](#), Pfam [↗](#), InterPro [↗](#), and Ensembl [↗](#). Data links are citations of an article by a database record. Data links from a specific data record to a Europe PMC article are provided by a number of life science databases. The databases included are listed in sections 2.7 and 2.8 of the Search syntax reference.

<https://europepmc.org/Help#searchdata>



# Cross referencing resources

Why is it a good idea to deposit your data?

The screenshot shows the ENA homepage with the search term "C10634" entered. A modal window displays detailed information about the sequence, including its organism (Caenorhabditis elegans), accession number (C10634), and various metadata fields like topology (linear) and base count (300). The modal also includes download options for EMBL and FASTA formats, navigation links, and publication and sequence version information. At the bottom, there's a "Navigation & Cross References" section listing links to EuropePMC, PMC6876284, PMC6955021, PMC6993210, PMC7380943, PMC8985095, and UNILIB.

Sequence: C10634.1

Caenorhabditis elegans cDNA clone yk150g2 : 3' end, single read.

Organism: Caenorhabditis elegans

Accession: C10634

Mol Type: mRNA

Topology: linear

Base Count: 300

Dataclass: EST

Tax Division: INV

Strain: CB1489 him-8(e1489)

Keywords: 3'-end sequence (3'-EST), EST (expressed sequence tag)

Sex: hermaphrodite, male

Show More

Navigation & Cross References

- Taxon: Taxon:6239
- EuropePMC: PMC6876284, PMC6955021, PMC6993210, PMC7380943, PMC8985095
- UNILIB: 228



# Cross referencing resources

Why is it a good idea to deposit your data?

The screenshot shows the ENA homepage with the search term 'C10634' entered. Below the search bar, the sequence identifier 'Sequence: C10634.1' is displayed. A large callout box highlights the 'Cross-references' section. The text in the callout box states: 'From this tab you can also see any links from the record out to external data resources that have used or generated these records as part of their services. These mappings are compiled as part of ENA's cross-reference service, and so only show data from resources that are registered with us. You can see more details on such registered resources [here](#)'. Below this, a note explains that the view only shows associations from the original project, while a related tab shows all associations across other projects. The 'Navigation & Cross References' section at the bottom lists links to Taxon, EuropePMC, and UNILIB databases.

EN  
European Nucleotide Archive

Home | Submit | Search | Rulespace | About | Support

Sequence: C10634.1

Enter text search terms  Search

Examples: histone, BN000065

C10634

Examples: Taxon:9606, BN000065, PRJEB402

?

View ENA FASTA

## Cross-references

From this tab you can also see any links from the record out to external data resources that have used or generated these records as part of their services. These mappings are compiled as part of ENA's cross-reference service, and so only show data from resources that are registered with us. You can see more details on such registered resources [here](#).

Note: This view only gives a view of the associated records submitted as part of the originally submitted research project and any registered cross-references. For a view showing all ENA records which are associated with this record (including any other links to this record within other ENA submission projects), you can see this in the Related ENA Records tab (available for Project, Sample and Taxon records).

Show More

Navigation & Cross References ?

- Taxon:  
Taxon:6239
- EuropePMC:  
PMC6876284, PMC6955021, PMC6993210, PMC7380943, PMC8985095
- UNILIB:  
228



# Many ways to find

How do I know which data goes where?  
When should I consider it?



# A wizard can be of help



EMBL-EBI home > Services > Data submission

## Data submission

Use this data submission wizard to find the right archive for your data in a few simple steps.

1 What **type of data** do you have?

[DNA/RNA sequence](#) [Expression data](#) [Protein data](#) [Structures](#) [Systems](#)

[Chemical biology](#) [Ontologies](#) [Images](#) [Multi-omics or other cross-domain study](#)

[Other biological research data](#)

**Why submit data to an archive?**

- Submission of primary data and derived information to public data repositories is an essential step in the scientific process.
- Through submission, the scientific community is fed the raw materials for the building and maintenance of the complete and up-to-date data sets that support searches and analysis on the latest sequences, structures and molecular profiles of living systems.
- Serving as a complement to the literature publication process and supporting early data sharing, the EMBL-EBI offers a number of submission services appropriate for different types and scales of data.

**Need help?**

If you need help with your data submission, please contact support.

<https://www.ebi.ac.uk/submission/>



## All EMBL-EBI data repositories

[Array Express](#): functional genomics data

[BioImage Archive](#): bioimaging data

[BioModels](#): computational models

[BioSamples](#): reference sample data

[BioStudies](#): biological research data

[ChEBI](#): chemical entities

[DGVa](#): structural genetic variation data

[EFO](#): experimental variables

[EGA](#): human data that requires controlled access

[EMPIAR](#): raw image data

[ENA](#): nucleotide sequence data

[EVA](#): genetic variation data

[GO](#): Gene ontology annotations

[GWAS Catalog](#): Genome-wide association study data

[IntAct](#): molecular interactions

[IntEnz](#): enzyme nomenclature

[MetaboLights](#): metabolomics data

[Metagenomics](#): raw sequence data & associated meta-data

[OneDep](#): electron microscopy, X-ray crystallography & NMR data

[PRIDE](#): protein & peptide identification data

[UniProt SPIN](#): protein sequences & annotations

[UniProt](#): updates or corrections

# Other ways to find a repository

Community driven content

The screenshot shows the RDMkit homepage. At the top, there is a navigation bar with the RDMkit logo, a search bar, and links for Data management, About, Contribute, and GitHub. Below the navigation bar, the text "The Research Data Management toolkit for Life Sciences" is displayed, followed by a subtitle "Best practices and guidelines to help you make your data FAIR (Findable, Accessible, Interoperable and Reusable)". A search bar is located below this text. The main content area is titled "What can we help you find?" and features a "Browse all topics by" section with nine categories arranged in a grid:

- Data life cycle**: Start here to get an overview of research data management based on stages in the data life cycle.
- Your role**: Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.
- Your domain**: Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.
- Your tasks**: Find guidelines and solutions for tackling common data management tasks.
- Tool assembly**: Find concrete combinations of tools and resources assembled into an ecosystem for research data management.
- National resources**: Find pointers to country specific information resources and national research data management practices.
- All tools and resources**: Browse the RDMkit's catalogue of tools and resources for research data management.
- All training resources**: Browse all training resources mentioned in RDMkit pages.

<https://rdmkit.elixir-europe.org>



# Other ways to find a repository

RDMkit Data management About Contribute GitHub Search RDMkit

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks

Compliance monitoring

Costs of data management

Data analysis

Data brokering

Data management coordination

Data management plan

Data organisation

Data protection

Data provenance

Data publication

Data quality

Data storage

Data transfer

Documentation and

Your tasks

## Data publication

- Can you really deposit your data in a public repository?
- Which repository should you use to publish your data?
- How do you prepare your data for publication in data repositories?
- How do you update or delete a published entry from a data repository?
- More information
- Relevant tools and resources

### Can you really deposit your data in a public repository?

#### Description

Sometimes it is difficult to determine if publishing data you have at hand is the right thing to do. Some reasons for hesitations might be that you have not used the data in a publication yet and don't want to be scooped, that the data contains personal information about patients or that the data was collected or produced in a collaboration.

#### Considerations

- Publishing data does not necessarily mean open access nor public. Data can be published with closed or restricted access.

[https://rdmkit.elixir-europe.org/data\\_publication](https://rdmkit.elixir-europe.org/data_publication)



# Other ways to find a repository

Are there recommendations  
for the domain?

**RDMkit**

Data management    About    Contribute    GitHub    Search RDMkit

**Data management**

Data life cycle    Your role    **Your domain**

- Bioimaging data
- Biomolecular simulation data
- Epitranscriptome data
- Human data
- Intrinsically disordered proteins
- Marine metagenomics
- Microbial biotechnology
- Plant sciences
- Proteomics
- Rare disease data
- Structural bioinformatics
- Toxicology data

Your tasks    Tool assembly    National resources

All tools and resources    All training resources

**Your domain** ! ⚡

In this section, information is organised based on different domains in life sciences with different approaches on how they manage their data. You will find:

- Domain-specific best practices and guidelines for data management.
- A description of domain-specific data management challenges, considerations to be taken into account and solutions used by the community to address the challenges.
- Links to domain-specific training materials.
- Links to tool assemblies implemented by the communities to address specific data management challenges.
- Links to a Data Stewardship Wizard for your DMP and to step-by-step instructions to make your data FAIR.
- A summary table of the relevant tools and resources for the specific domain, recommended by the community.

Search Type here... ✖

<b>Bioimaging data</b> Data management solutions for bioimaging data.  Related pages <span style="color: red;">! ⚡</span>	<b>Biomolecular simulation data</b> Data management solutions for biomolecular simulation data.  Related pages <span style="color: orange;">! ⚡</span>
<b>Epitranscriptome data</b> Data management solutions for epitranscriptome data.	<b>Human data</b> Data management solutions for human data.  Related pages <span style="color: orange;">! ⚡</span>
<b>Intrinsically disordered proteins</b> Data management solutions for intrinsically disordered proteins data.	<b>Marine metagenomics</b> Data management solutions for marine metagenomics data.

<https://rdmkit.elixir-europe.org>



# Other ways to find a repository

The screenshot shows the RDMkit website interface. The top navigation bar includes links for Data management, About, Contribute, GitHub, and a search bar. The left sidebar has a 'Data management' heading with dropdown menus for Data life cycle, Your role, Your domain, and Human data (which is highlighted). Below these are links for Bioimaging data, Biomolecular simulation data, Epitranscriptome data, Intrinsically disordered proteins, Marine metagenomics, Microbial biotechnology, Plant sciences, Proteomics, Rare disease data, Structural bioinformatics, Toxicology data, Your tasks, Tool assembly, National resources, All tools and resources, and All training resources. The main content area is titled 'Human data' and contains sections for Introduction, Planning for projects with human data, Description, and Considerations. The 'Introduction' section includes a link to the URL [https://rdmkit.elixir-europe.org/human\\_data](https://rdmkit.elixir-europe.org/human_data).

## Solutions

- The European Genome-phenome Archive (EGA) is an international service for secure archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical studies and healthcare centres. All services are free of charge. The EGA platform offers secure and European law-compliant data sharing. Data treatment is FAIR-compliant, thus data is discoverable in the EGA website and shareable with other researchers through authorisation and authentication protocols. The right to allow access to any dataset belongs to the Data controllers (and not to the EGA), who are responsible to sign a Data Access Agreement (DAA) with researchers requesting access to their data. Templates of the legal documents are provided. The EGA hosts data from all around the world and distributes it where and when the data controllers permit.
- dbGAP and JGA are other international data repositories, based in the USA and Japan respectively, that adopt a controlled-access model based on their national regulations. Due to European GDPR specific requirements, it may not be possible to deposit EU subjects' data to these repositories.
- The GA4GH Beacon project is a GA4GH initiative that enables genomic and clinical data sharing across federated networks. A Beacon is defined as a web-accessible service that can be queried for information about a specific allele with no reference to a specific sample or patient, thereby reducing privacy risks.
- The GA4GH Data Use Ontology DUO is an international standard, which provides codes to represent data use restrictions for controlled access datasets.
- Crypt4gh is a Python tool to encrypt, decrypt or re-encrypt files, according to the GA4GH encryption file format.

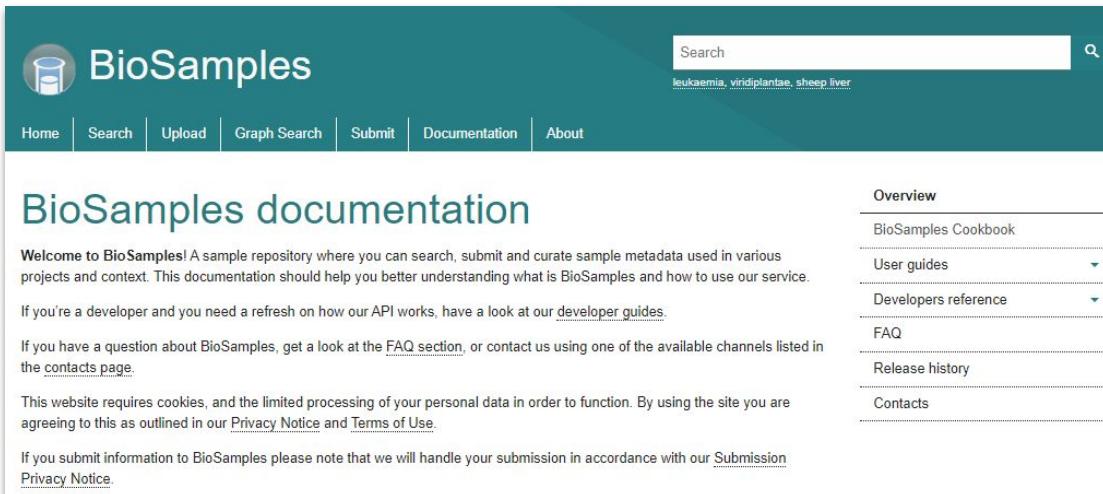
## Relevant tools and resources

Tool or resource	Description	Related pages	Registry
<b>BBMRI-ERIC's ELSI Knowledge Base</b>	The ELSI Knowledge Base is an open-access resource platform that aims at providing practical know-how for responsible research.	Data protection Data sensitivity Data Steward: policy Data Steward: research	  
<b>Beacon</b>	The Beacon protocol defines an open standard for genomics data discovery.	Researcher Data Steward: research Data Steward: infrastructure	  
<b>BIONDA</b>	BIONDA is a free and open-access biomarker database, which employs various text mining	Data storage Researcher Proteomics	

# Data submission

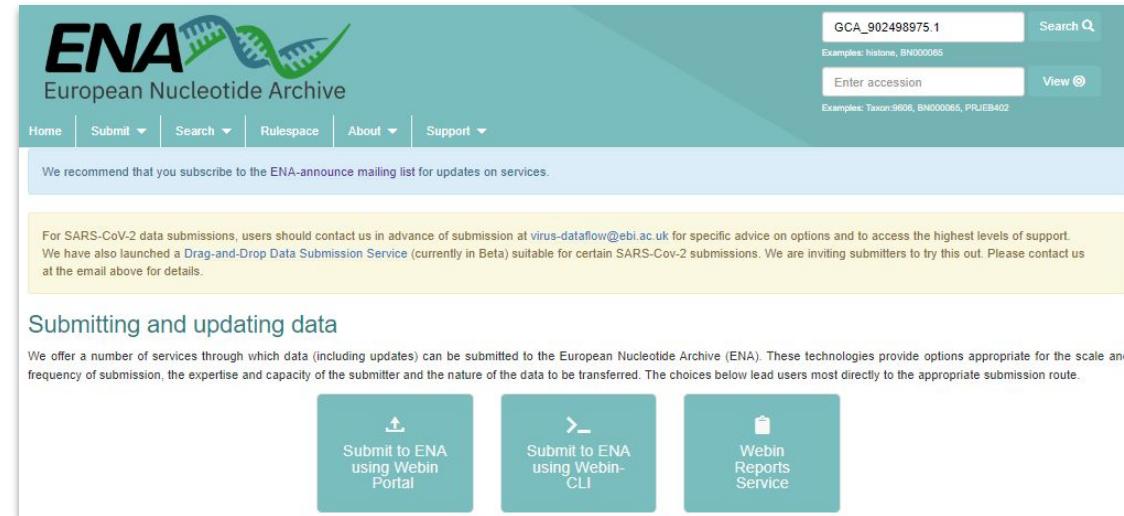
Different repositories = different routes

- What file formats are expected?
- What metadata is expected?
- Can ontologies and controlled vocabularies be used?
- Are there costs involved?
- Is there a license to be applied? Is it implicit?

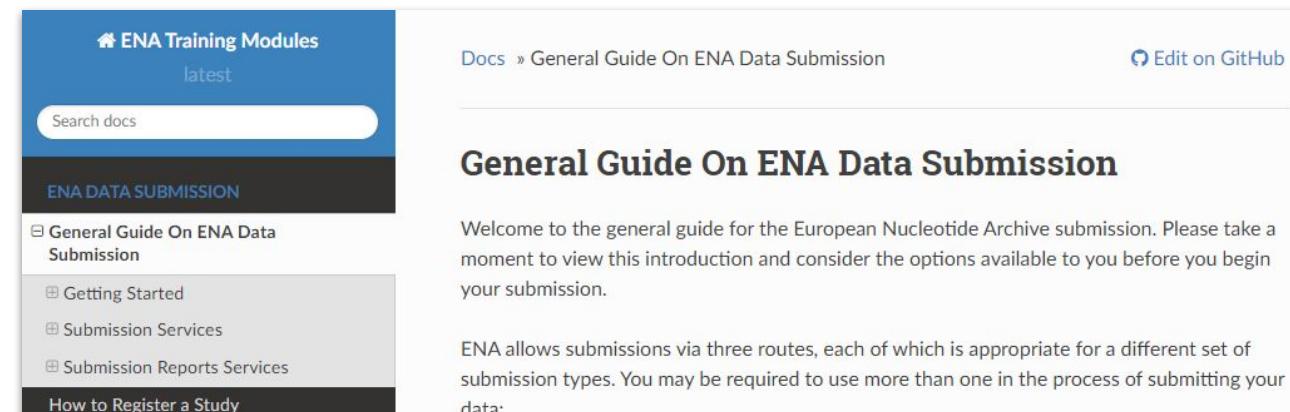


The BioSamples documentation page features a search bar at the top with the term "leukaemia, viridiplantae, sheep liver". Below the search bar, the "Home" button is highlighted. The main content area is titled "BioSamples documentation". It includes sections for "Welcome to BioSamples!", developer information, FAQ, release history, and contacts. A sidebar on the right is titled "Overview" and lists links to the BioSamples Cookbook, User guides, Developers reference, FAQ, Release history, and Contacts.

<https://www.ebi.ac.uk/biosamples/docs>



The ENa Data Submission page has a teal header with the ENa logo and "European Nucleotide Archive". The main content area starts with a recommendation to subscribe to the ENa-announce mailing list. It then discusses SARS-CoV-2 data submissions and mentions a Drag-and-Drop Data Submission Service. Below this, a section titled "Submitting and updating data" provides options for submission: "Submit to ENa using Webin Portal" (with a file icon), "Submit to ENa using Webin-CLI" (with a terminal icon), and "Webin Reports Service" (with a document icon). A link at the bottom right leads to the submission page: <https://www.ebi.ac.uk/ena/browser/submit>.



The General Guide On ENa Data Submission page has a blue header with the ENa Training Modules logo and "latest". The main content area includes a "Search docs" bar and a sidebar with links to "General Guide On ENa Data Submission", "Getting Started", "Submission Services", "Submission Reports Services", and "How to Register a Study". The main text area welcomes users to the guide and explains that ENA allows submissions via three routes. A link at the bottom left leads to the general guide: <https://ena-docs.readthedocs.io/en/latest/submit/general-guide.html>.



# Let's find a repository



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal



<https://rdmkit.elixir-europe.org>

Identify a repository based on information from RDMkit

- Domain specific
- Justify the choice
- Bonus: can you find further resources that can help you?



# Other ways to find a repository



Your domain

## Plant sciences

- Introduction
- Plant biological materials: (meta)data collection and sharing
- Phenotyping: (meta)data collection and publication
- Genotyping: (meta)data collection and publication
- Related pages
- More information
- Relevant tools and resources

### Genotyping: (meta)data collection and publication

#### Description

Here are described the mandatory, recommended and optional metadata fields for data interoperability and re-use, as well as for data deposition in EVA (European Variation Archive), the EMBL-EBI's open-access genetic variation archive connected to BioSamples, described above.

#### Considerations

- Did you collect the metadata for the identification of your plant samples according to the recommendations provided in the above section?
- Is the reference genome assembly available in an INSDC archive and has a Genome Collections Accession number, either GCA or GCF?
- Is the analytic approach used for creating the VCF file available in a publication and has a Digital Object Identifier (DOI)?

#### Solutions

##### Checklists, ontologies and file formats

- Sharing plant genotyping data files involves the use of the Variant Call Format (VCF) standard.
- Findability and reusability of VCF files depends on the supplied metadata and in particular with MIAPPE compliant biological material description: the plant genomic and genetic variation data submission recipe helps you on that topic.

##### Data sharing and publication

- Once the VCF file is ready with all necessary metadata, it can be submitted to the European Variation Archive (EVA). You will find all necessary information on the submission steps on the [EVA submission page](#).

[https://rdmkit.elixir-europe.org/plant\\_sciences](https://rdmkit.elixir-europe.org/plant_sciences)



# Other ways to find a repository



Tool assembly

## Plant Genomics

- What is the plant genomics tool assembly?
- Who can use the plant genomics tool assembly?
- How can you access the plant genomics tool assembly?
- For what purpose can you use the plant genomics tool assembly?
- Related pages
- More information
- Relevant tools and resources

### Data sharing and publishing

All sequencing data collected in plant genotyping experiments should be submitted to ENA together with metadata compliant to the [GSC MIxS plant associated checklist](#). Final results of such studies in the form of VCF files should be submitted to EVA. Additionally, supplemental data complementing these two data types is encouraged to be submitted to [e!DAL-PGP](#) or [Recherche Data Gouv](#).

[https://rdmkit.elixir-europe.org/plant\\_genomics\\_assembly](https://rdmkit.elixir-europe.org/plant_genomics_assembly)

### More information

#### Links to other ELIXIR resources



Step-by-step process for: Improving dataset maturity - the MIAPPE use case



# Multiple repositories with a similar data type



**International Nucleotide Sequence Database Collaboration**

The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#).

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations.

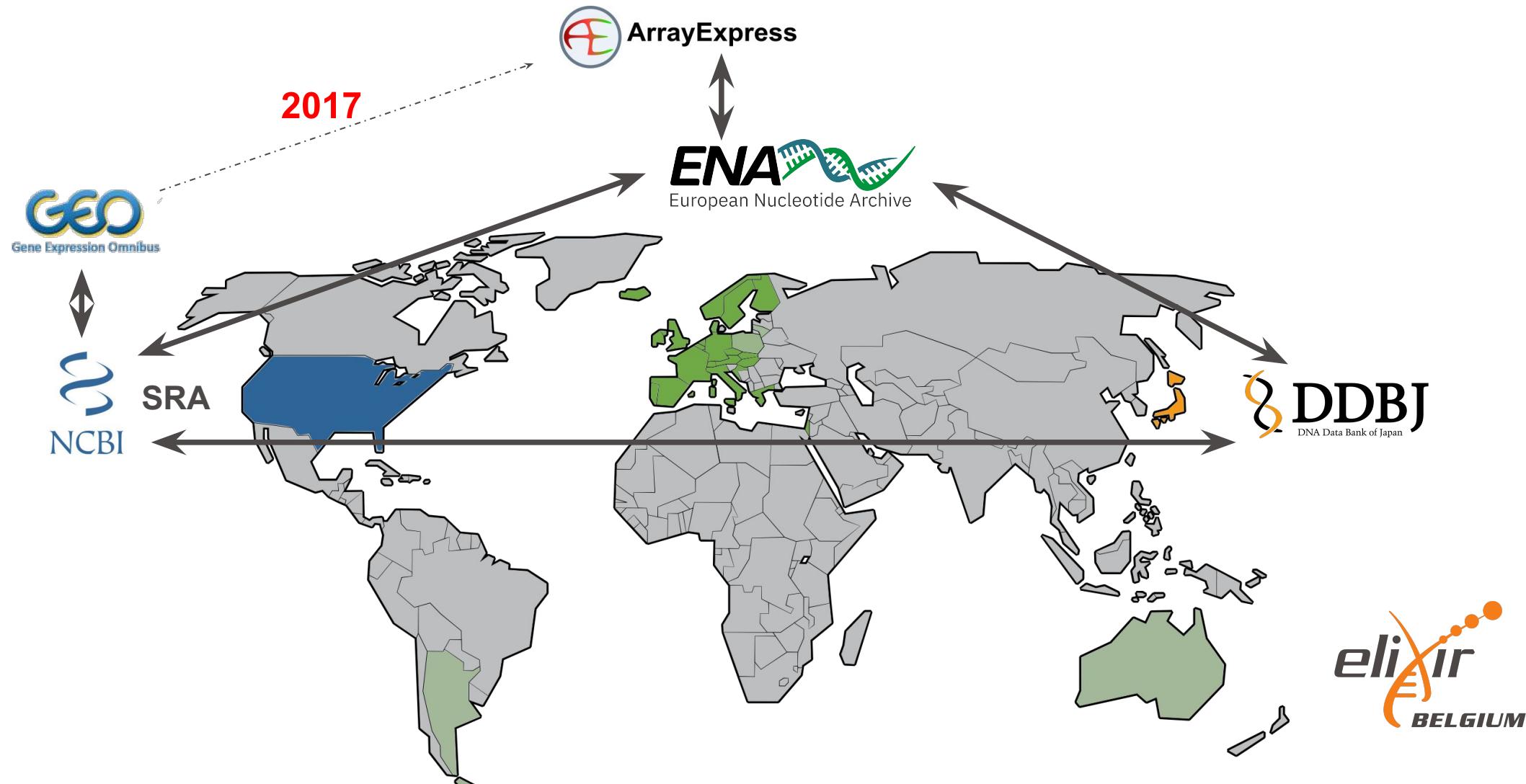
## Databases

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive		Sequence Read Archive
Assembled Sequences	DDBJ	European Nucleotide Archive	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

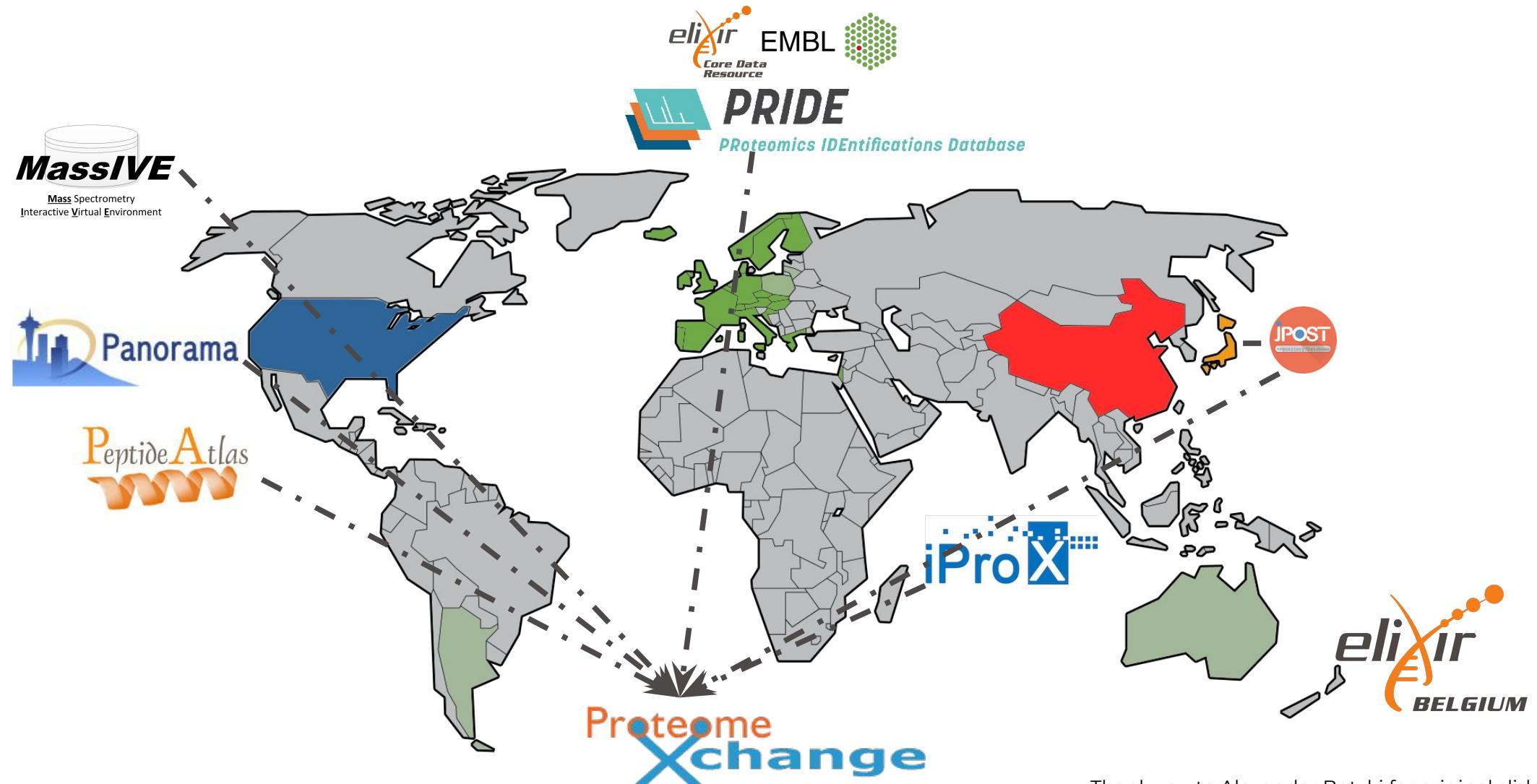
<https://www.insdc.org/>



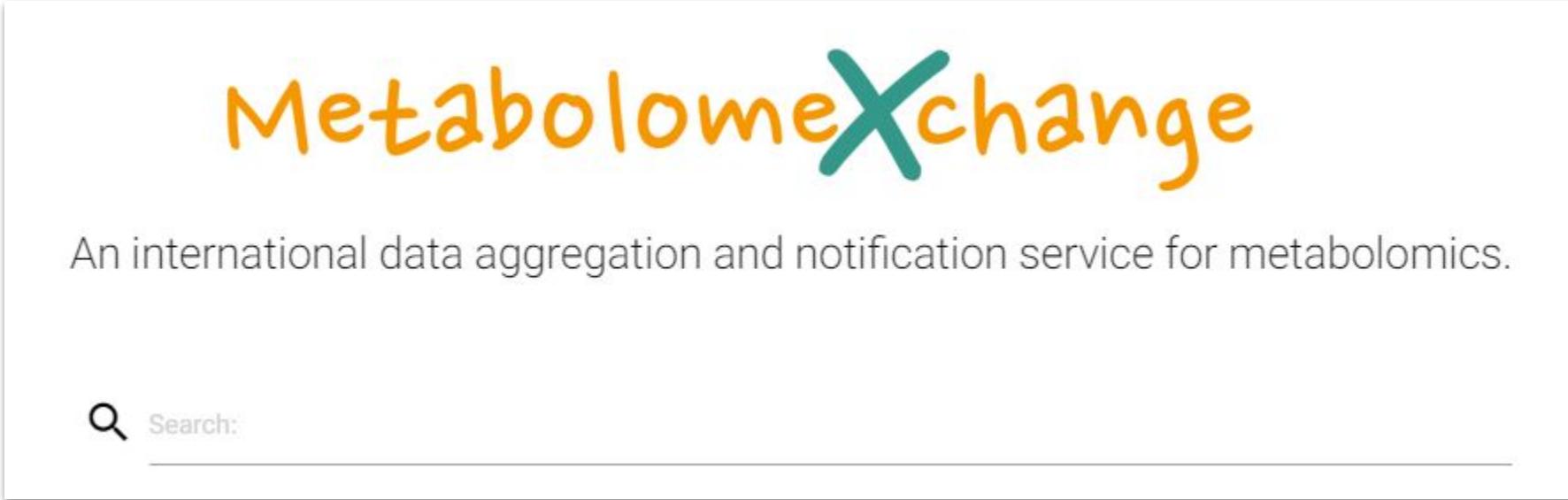
# Multiple repositories with a similar data type - Genomics



# Multiple repositories with a similar data type - Proteomics



## Multiple repositories with a similar data type - Metabolomics



The image shows the MetabolomeXchange website. At the top is the logo 'MetabolomeXchange' where 'Metabolome' is orange and 'Xchange' has a large teal 'X'. Below the logo is the text 'An international data aggregation and notification service for metabolomics.' To the left of a search bar is a magnifying glass icon. The search bar itself is labeled 'Search:'.

<http://www.metabolomexchange.org/>

Similar initiatives are being built



# Multiple repositories with a similar data type - multi-omics

## The Omics Discovery Index - OmicsDI

25 Resources  
>3M Datasets  
4 Continents

Screenshot of the OmicsDI homepage showing a grid of 25 multi-omics datasets:

Resource	Description	Number of Datasets	Last Updated
PeptideAtlas	A multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.	2365 datasets	2016-06-21
MetabolomeExpress	A public place to process, interpret and share GC/MS metabolomics datasets.	58 datasets	2015-07-29
NODE	The National Omics Data Encyclopedia (NODE) stores raw sequence data from next-generation sequencing technologies including 454, IonTorrent, Illumina, SOLID, Helicos and Complete Genomics.	523 datasets	2021-01-11
MetabolomicsWorkbench	is a scalable and extensible informatics infrastructure which will serve as a national metabolomics resource.	1735 datasets	2021-12-13
GNPS	The Global Natural Products Social Molecular Networking (GNPS) is a platform for providing an overview of the molecular features in mass spectrometry based metabolomics by comparing fragmentation patterns to identify chemical relationships.	2164 datasets	2022-03-20
PAXDB	PaxDb contains estimated abundance values for a large number of proteins in several different species. Furthermore, you can find information about inter-species variation of protein abundances.	493 datasets	2015-03-12
ENA	European Nucleotide Archive. Note that the number of datasets below is smaller than the number of datasets returned when you click on that link. This is because the former refers to the datasets provided by ENA and the latter also includes datasets in other repositories that have re-analysed ENA data.	600276 datasets	2022-07-09
LINCS	The database contains all publicly available HMS LINCS datasets and information for each dataset about experimental reagents (small molecule perturbagens, cells, antibodies, and proteins) and experimental and data analysis protocols.	447 datasets	2022-09-29
Pride	is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.	23444 datasets	2023-06-10
GEO	NCBI Gene Expression Omnibus. Note that the number of datasets below is smaller than the number of datasets returned when you click on that link. This is because the former refers to the datasets provided by GEO and the latter also includes datasets in other repositories that have re-analysed GEO data or mirrored them.	201602 datasets	2023-04-10
JPOST Repository	Japan ProteOme STandard Repository is a new data repository of sharing MS raw/processed data.	755 datasets	2022-08-18
dbGaP	The database of Genotypes and Phenotypes was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in humans.	2087 datasets	2014-04-30
BioModels	BioModels Database is a repository of computational models of biological processes. Models described from literature are manually curated and enriched with cross-references.	2521 datasets	2023-06-06
Cell Collective	Interactive Modelling of Biological Networks.	225 datasets	2021-10-25
MetaboLights	is a database for Metabolomics experiments and derived information.	1296 datasets	2023-06-08
ArrayExpress	ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.	76229 datasets	2023-01-01

<https://www.omicsdi.org/>

# Multiple repositories with a similar data type - multi-omics

## The Omics Discovery Index - OmicsDI

Indexing metadata

The richer the metadata, the easier it is to find a dataset

OmicsDI   Browse   Submit Data   Databases   API   Help   Login   Organism, repository, gene, tissue, accession   Advanced   Search

3454947 Results   Show all   Save search   Copy query

Show results for

- U Unknown (2390057)
- O Other (724873)
- T Transcriptomics (72616)
- M Multiomics (59108)
- P Proteomics (30790)
- G Genomics (2704)
- S Models (2521)
- M Metabolomics (1300)
- MG Metagenomics (69)
- RW RecuratedModel (13)
- SG Single Cell Transcriptomics (11)

Sort ▲ by: Relevance   Page size 10

**Expression data from gastric cancer and paired normal tissues**  
Gastric cancer (GC) is one of the most common cancer worldwide. Specific and reliable molecular markers are limited; it is critical to identify new biomarkers for GC to aid in early diagnosis, treatment strategy, and prognosis evaluation. Microarray technology makes it possible to measure the expression levels of thousands of genes simultaneously.  
ORGANISM(S): Homo Sapiens  
2016-04-07 | E-GEOID-79973 | ArrayExpress  
transcription profiling by array   Cite

**Histone Reader BRWD1 Targets and Restricts Recombination to the Igk Locus**  
B lymphopoiesis requires that immunoglobulin genes be accessible to RAG1-RAG2 recombinase. However, the RAG proteins bind widely to open chromatin, which suggests that additional mechanisms must restrict RAG-mediated DNA cleavage. Here we show that developmental downregulation of interleukin 7 (IL-7...  
ORGANISM(S): Mus Musculus  
2015-08-24 | E-GEOID-63302 | ArrayExpress  
RNA-seq of coding RNA   ChIP-seq   Cite

**Gene expression analysis in stigma during compatible and in-compatible pollination reactions in *Lolium perenne***  
The expression analysis had two goals: (1) look at relative transcription within mature pollen grains (2) compare expression in the stigma during pollination with either compatible or in-compatible pollen. Two pairwise comparisons, (i) unpollinated stigma vs. stigma pollinated with compatible pollen...  
ORGANISM(S): Lolium Perenne  
2015-10-26 | E-MTAB-3760 | ArrayExpress  
RNA-seq of coding RNA   Cite

**Insecticide resistance mechanisms *Myzus persicae*: Genotypes Insecticide vs. Genotypes Control**  
Transcriptional responses in three genotypes of *Myzus persicae*, each exhibiting different resistance mechanisms, in response to an anti-cholinesterase insecticide. Two-condition experiment in three different genotypes: Insecticide vs. acetone plus water. Genotypes S (exhibiting no resistance mechanism)...  
Cite

# A multi-omics project



Alice is a researcher who:

- Works with plants
- Is generating whole genome sequencing data
- Makes use of several public platforms for tools and data
- Has the current project funded by a public funding agency
- Aims at publishing work in an open and peer reviewed journal
- **Is also generating proteomics data**
- **Is also generating metabolomics data**
- **Would really like that the datasets are recognized as belonging to the same project**

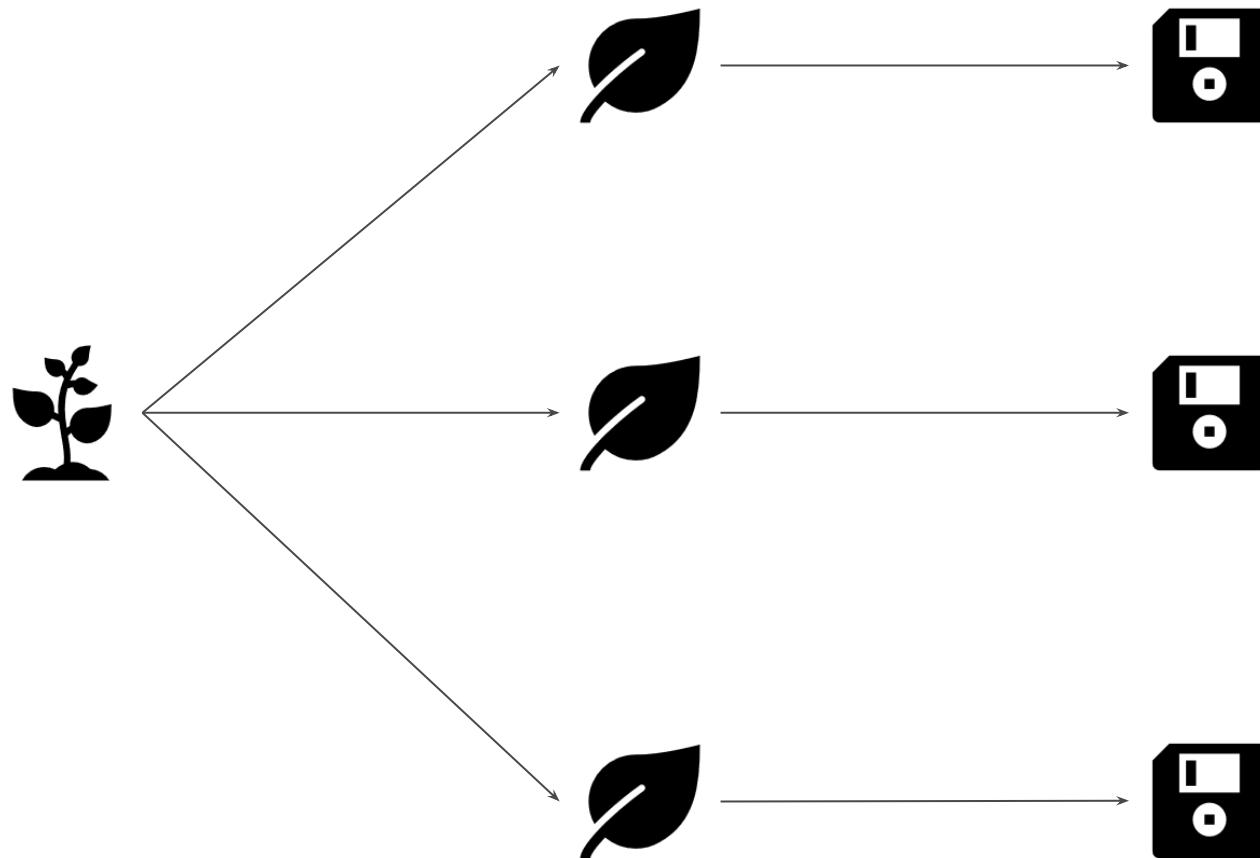
Identify

- Suitable, domain specific, repositories
- A strategy (repository) to help connect the datasets



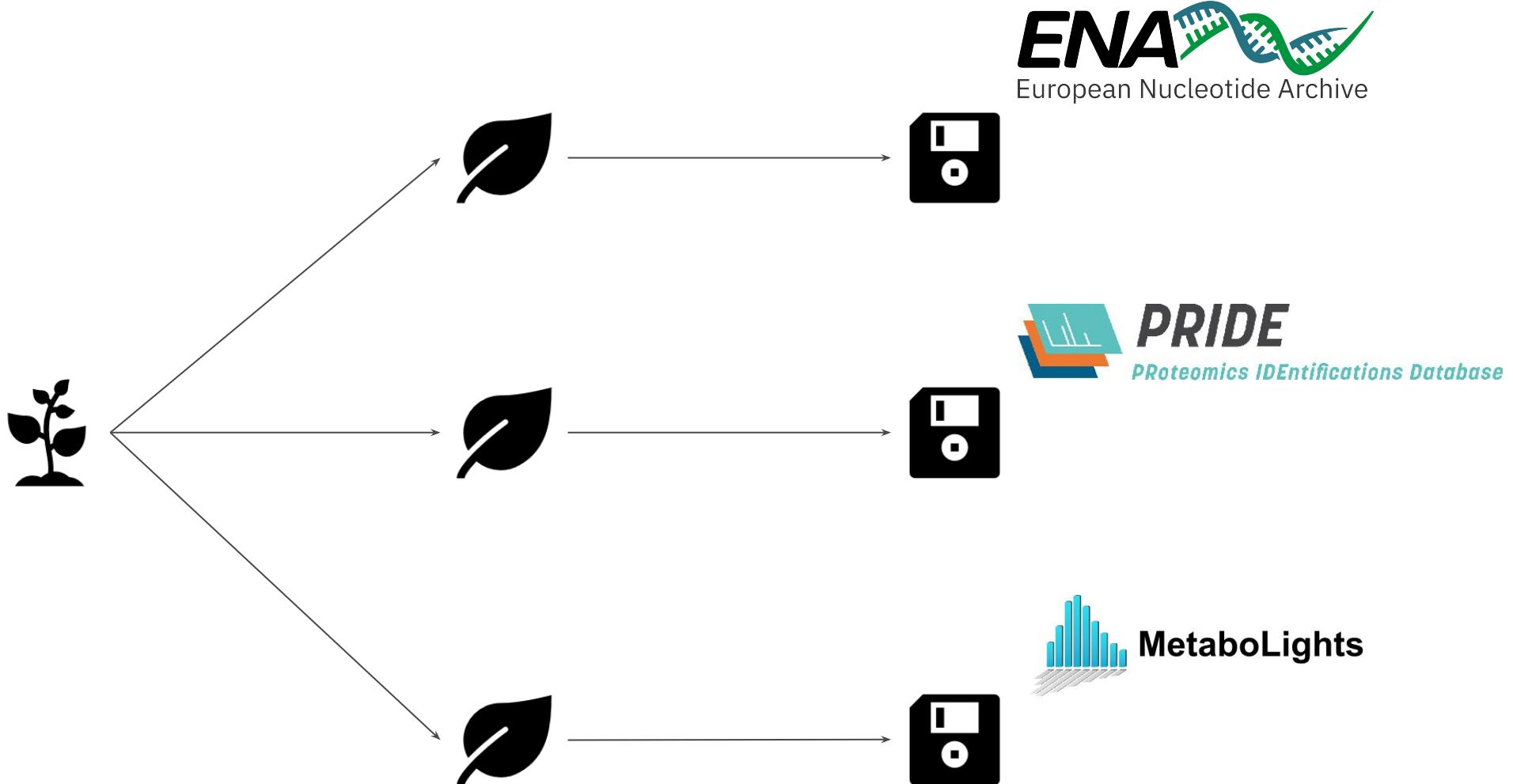
# Multi-omics

---



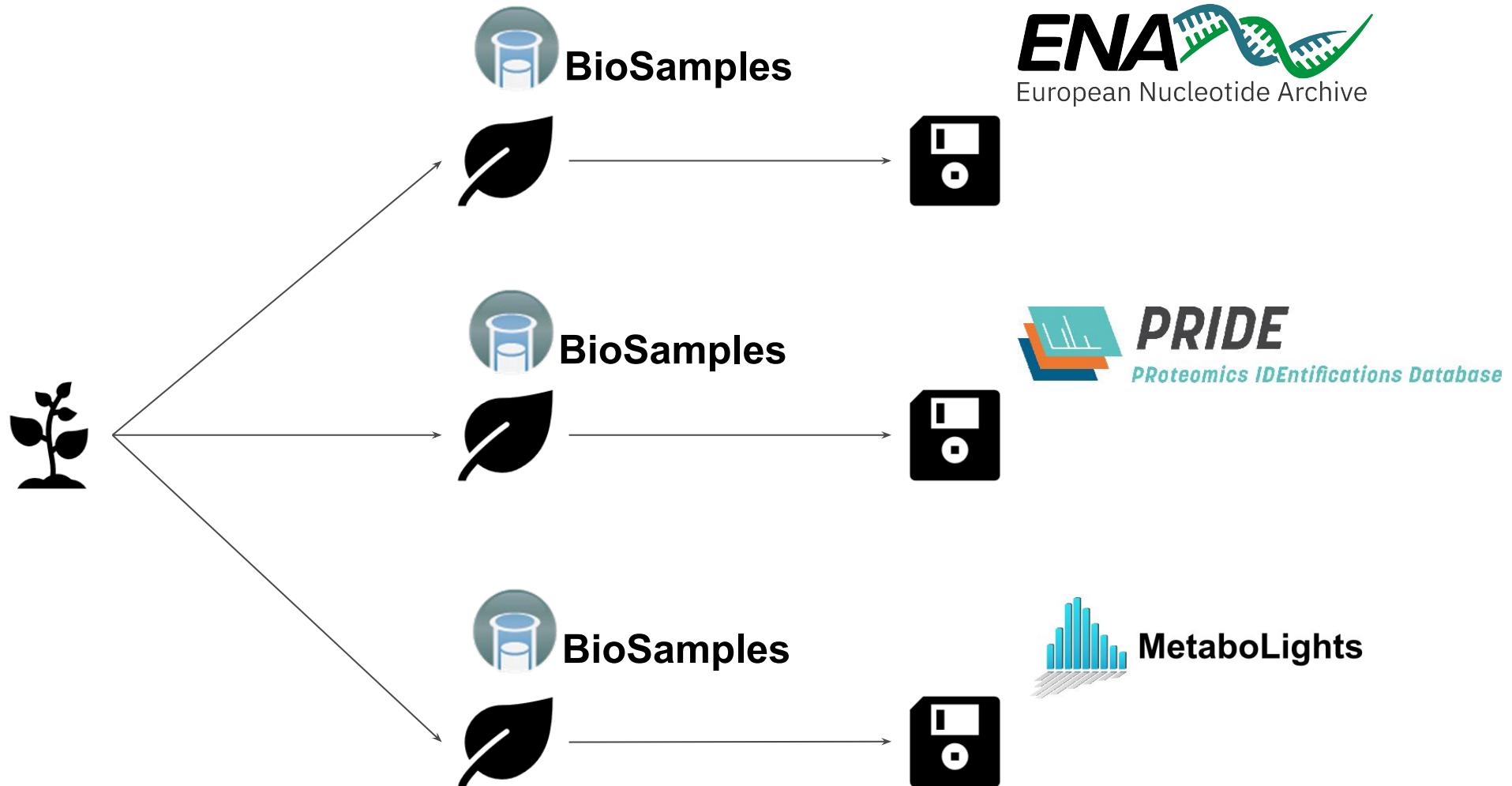
# Multi-omics

---



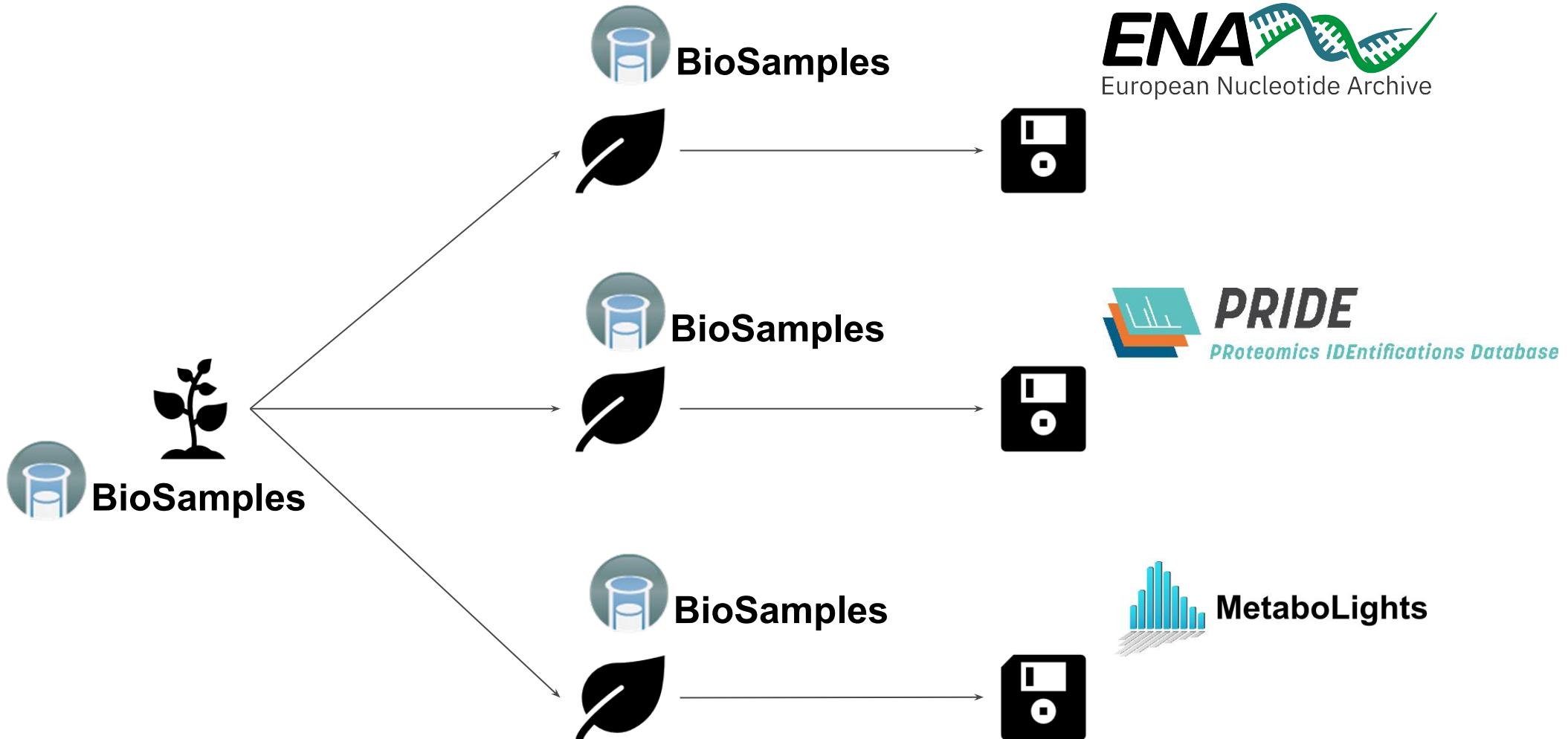
# Multi-omics

---

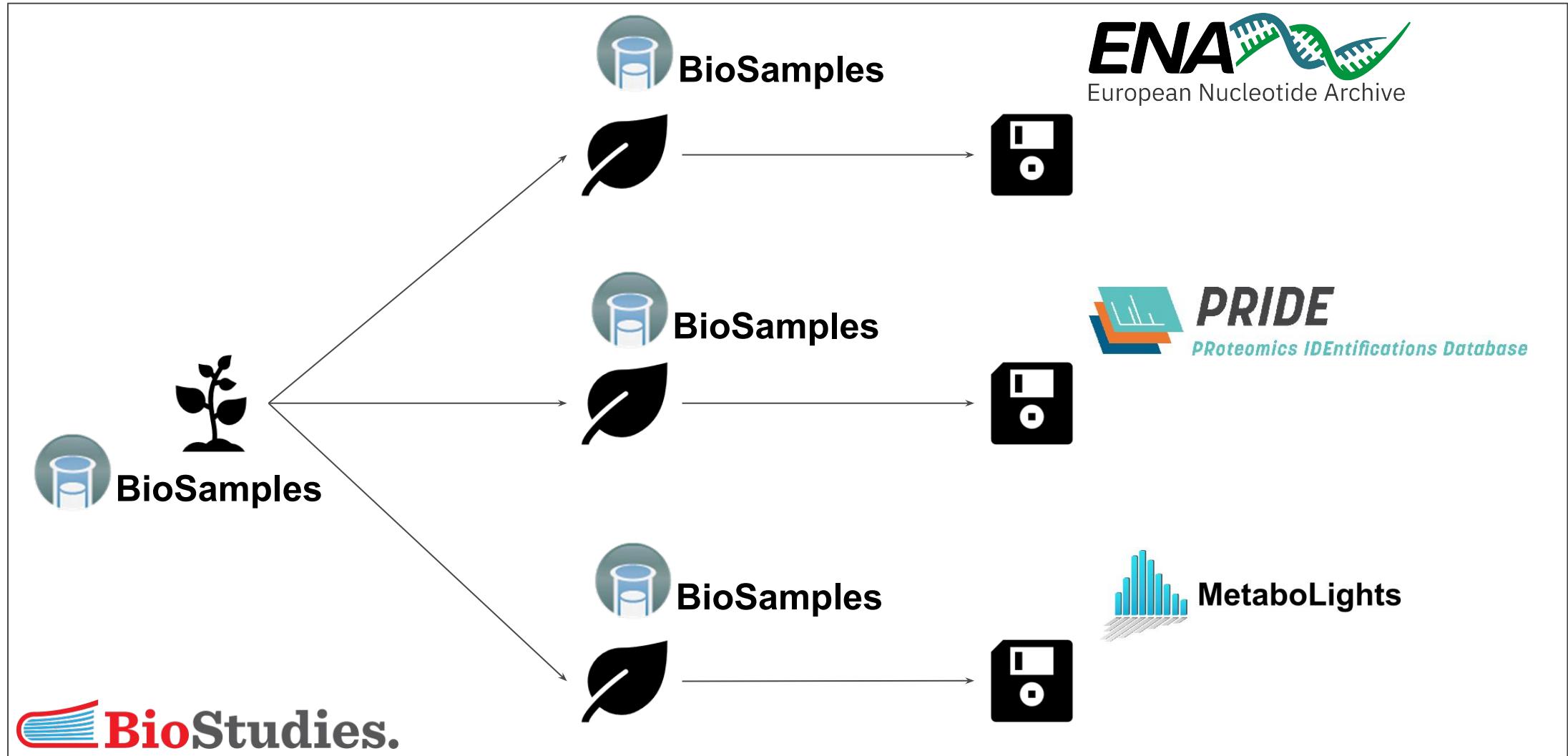


# Multi-omics

---



# Multi-omics



# What about human sensitive data?

---

“The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects or in the context of research-focused healthcare systems.”



<https://ega-archive.org/>

- Central metadata access
- Secure storage of sensitive data
- Data access defined, including Data Access Committees (DAC)
- Different datasets may have different DACs



# Access types



- **Open access:** Data is shared publicly. Open-access is a rarely used access mode for the sharing of human data. To use open-access researchers need to ensure that the data needs is anonymised, which is difficult in practice.
- **Registered access:** Data is shared with researchers, who have been vouched for by their institution and who agree to abide by data usage policies. Datasets that are shared via registered-access would typically have no restrictions besides the condition that data is to be used for research.
- **Controlled access:** Access is reviewed and approved by a data access committee (DAC). Typically researchers who were involved in the primary collection of data will form the DAC. There can be a multitude of usage conditions, including allowed research topics, geographical regions, and recipients e.g. non-profit organisations.



# Submitting data to the EGA

## Quick Guide

This is a quick guide to submit data to the EGA. Please select data type to display the right detailed instructions. Please note - Submissions to EGA can take approximately one month, so please, allow plenty of time for the submission and archiving processes.

[Array-based data](#) [Sequence data](#)



Fill the [submission form](#) and provide details of the data type(s), used platform(s) and estimated size of your submission.  
If you are affiliated to an existing consortium, such as the International Cancer Genome Project (ICGC), please add this information in the comments.



Use [Submitter Portal](#) to register your study, samples, Data Access Committee (DAC) and Policy.



Encrypt your data files using the [EgaCryptor](#) or [encrypt your files locally](#) and upload it using default FTP clients or Aspera.



Associate each data file to a registered sample and study by [Linking files to samples](#). Details of the experimental procedure you followed must be provided.

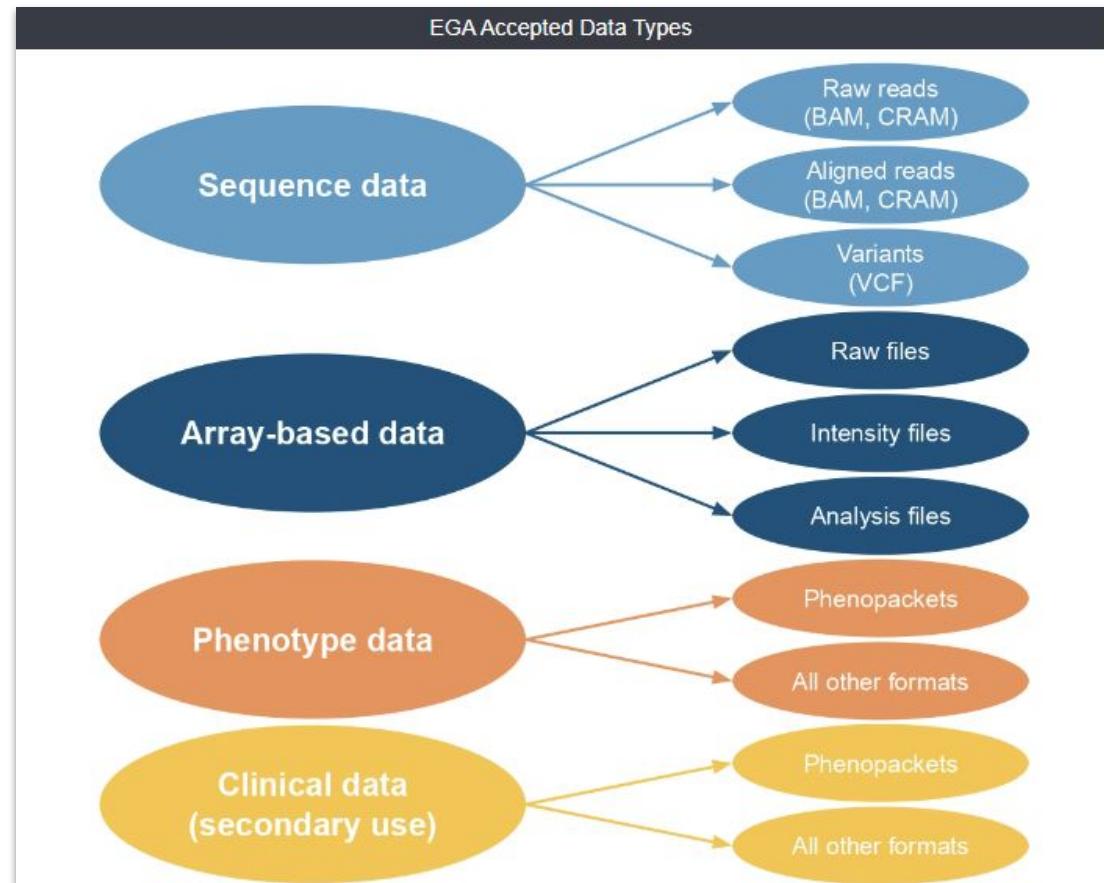


Group your runs/analysis into datasets and link them to your new or existing DAC and policy using [Submitter Portal](#) or an [XML based programmatic submission](#). Data requests are done at a dataset level, thus files within a datasets must share release conditions.



Instruct our Helpdesk to release your study. All registered studies are automatically placed on hold until the named submission or DAC contact instructs our Helpdesk for the study to be released.

<https://ega-archive.org/submission/quickguide>



What if the data cannot leave my country?

# Federated EGA

---



<https://ega-archive.org/federated>

“Federated EGA strives to support the discovery of and secure access to human data globally, while respecting national data protection regulations, with the goal of accelerating disease research and understanding and improving human health.”

<https://doi.org/10.7490/f1000research.1118988.1>

- Launched in September 2022
- Data does not leave the Local EGA
- Dataset discovery through the Central EGA



# Federated EGA



Central EGA



Signed FEGA Collaboration  
Agreement



Joined FEGA Committees



Engaged in work to establish a  
FEGA node



Expressed interest in joining FEGA  
Network



Search docs

Establishing a Federated EGA Node

TOPICS

Maturity Model

Data and Metadata Management

Outreach and Training

Technical and Operational

Governance and Legal

Docs » Establishing a Federated EGA Node

## Establishing a Federated EGA Node



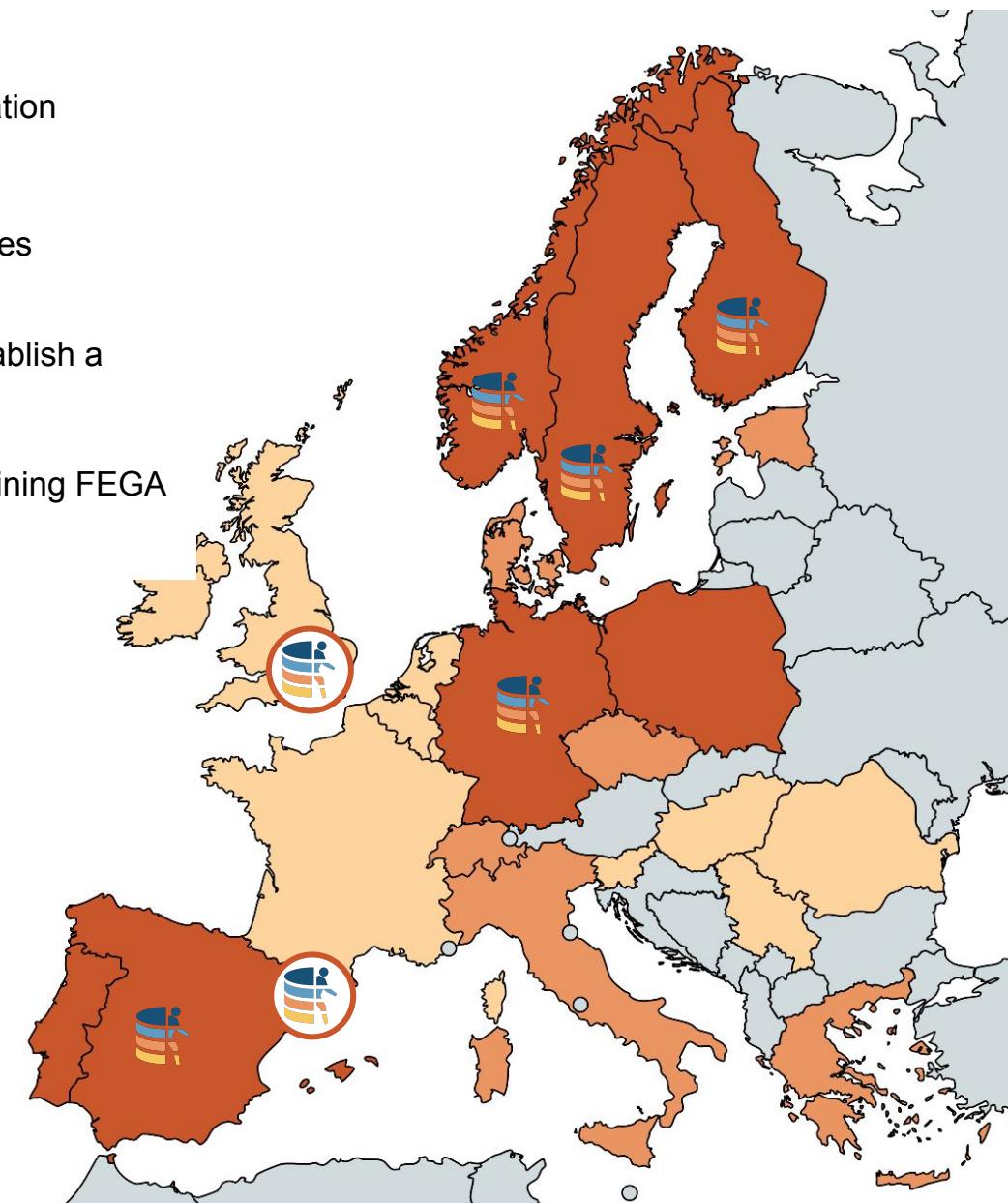
Welcome to this collection of onboarding materials for establishing a Federated EGA Node!

### What am I doing here?

If you are reading this, you are probably looking for information on how to join the Federated EGA. Great! There is a lot of information here for you.

These materials provide guidance for establishing a node within the Federated EGA. The materials are based on the knowledge and experiences of current nodes and their use cases. Your node's development might differ depending on your use cases and mandates from stakeholders. Please view these materials as suggestions and best practices - not hard requirements!

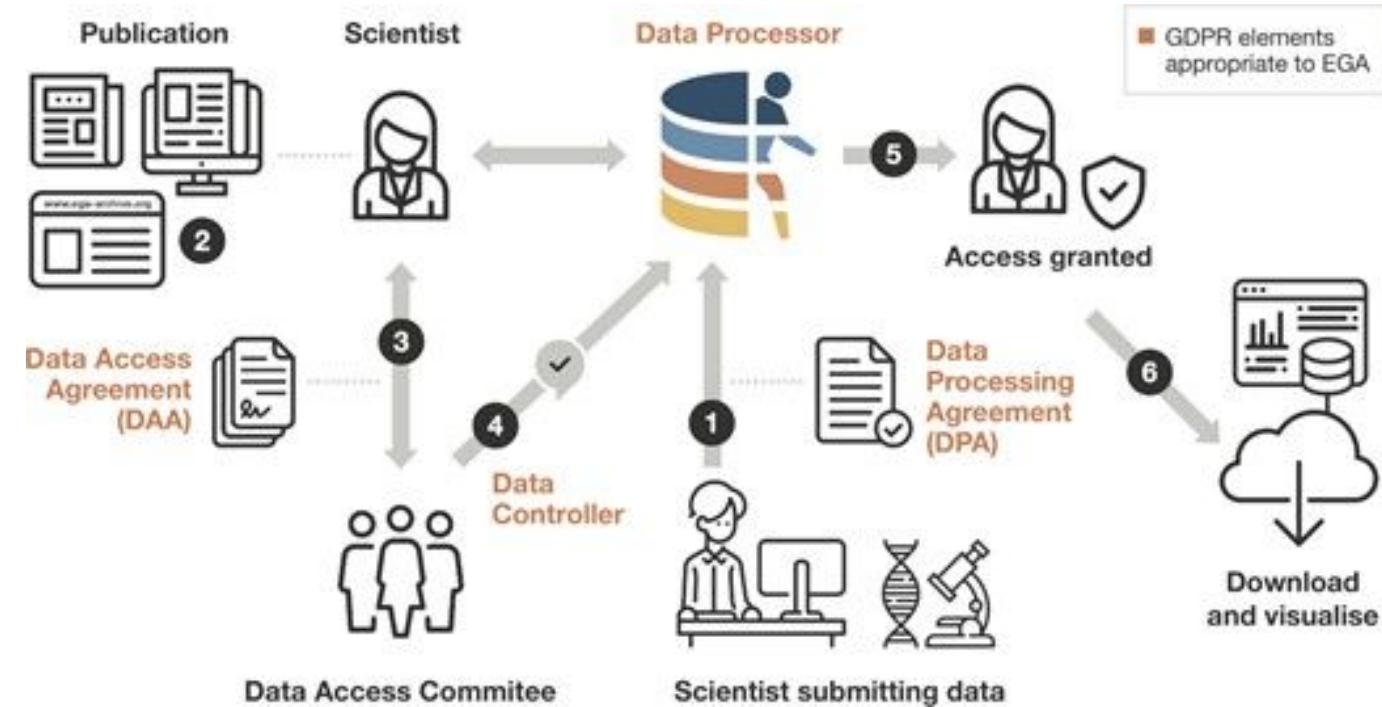
<https://ega-archive.github.io/FEGA-onboarding/>



Modified from <https://doi.org/10.7490/f1000research.1119478.1>

# Re-using data from the EGA

1. Submission after a Data Processing Agreement
2. A scientist discovers the dataset (e.g. through a publication or by navigating the EGA)
3. The DAC is contacted and a DAA is defined
4. EGA is informed of access permission
5. Scientist can access the data
6. Data is downloaded and re-used



<https://doi.org/10.1093/nar/gkab1059>



# Re-using data from repositories

1. Look into Licenses and Policies
  - a. For Reuse
  - b. Before depositing



# Re-using data from repositories

1. Look into Licenses and Policies
  - a. For Reuse
  - b. Before depositing

The screenshot shows a web browser displaying a dataset page from [data.gov.be/nl/search/datasets/water](https://data.gov.be/nl/search/datasets/water). The page title is "Waste Water Treatment Plant discharge points". A red circle highlights the "License" section, which lists various open data licenses. Below the main content, there is a large black box containing the same "License" section.

**License**

- Open Data License Flanders (159)
- CC Attribution (CC BY) (94)
- CC Zero (CC 0) (29)
- Open Data License Brussels (18)
- Etalab Open License (7)
- CC Attribution NonCommercial (CC BY NC) (3)
- Statbel Open License (3)
- CC Attribution NonCommercial NoDerivs (CC BY NC ND) (1)

Brussels-Capital Region: localization of the discharge points of the two treatment stations of waste water and rain water of the Brussels agglomeration.

Environment WMS Brussels region

OGC API Features Flemish Delivery Areas Drink water ...

Direct download service for the Flemish Delivery Areas Drinking water as compiled for reporting under the drinking water directive to the European Commission.

**elixir**  
BELGIUM

# Re-using data from repositories

The INSDC is an outstanding example of success in building an immensely valuable, widely used public resource through voluntary cooperation across the international scientific community. This success has been achieved by following the guidelines and principles outlined above.

## Data availability policy

While the INSDC databases hold public data, there are several levels of data availability which control access to these data. See the [INSDC Data Availability Policy](#) for full details of INSDC data access and control.

The two main levels to data availability are when data are confidential pre-publication and then after public release.

Confidential Data	Public Data
A data owner can indicate during study/project registration that confidentiality is required until an owner-managed release date or publication in the literature, whichever comes earlier.	A project is subsequently and automatically released as Public on reaching the specified release date or when the relevant INSDC accession cited online or in a publication prior to this date.
During the confidential phase, data are not available publicly through any means.	In the event that a release date must be extended, data owners can extend the release of their data before it becomes public.

**ENA**  
European Nucleotide Archive

Enter text search term  
Examples: histone, BN000065

Search

Enter accession  
Examples: Taxon:9606, BN000065, PRJEB402

View

Home | Submit ▾ | Search ▾ | Rulespace

**ENA and INSDC Policies**

The International Nucleotide Sequence Database Collaboration between DDBJ, EMBL, and GenBank for over 20 years. The International Advisory Committee (IAC), is made up of European, Japanese and American members. It overlaps that of the ENA Scientific Advisory Board (SAB) which oversees the data-sharing policy of the three databases that make up the ENA. Individuals submitting data to the international sequence databases DDBJ, EMBL and GenBank should be aware of the following:

1. The INSDC has a uniform policy of free and unrestricted access to all data contained in the databases. Scientists worldwide can access these records and use them for analysis and publication. Appropriate credit is given by citing the original source and utilising published scientific literature.
2. The INSDC will not attach statements to records that restrict their use or publication.

This website requires cookies, and the limited processing of personal data in order to function. By using the site you accept this as outlined in our [Privacy Notice](#) and [Terms of Use](#).

About ENA  
ENA Content  
Using ENA  
Data Standards  
Statistics  
**Policies**  
Data Coordination  
Funding  
Events  
News

I agree, dismiss this banner

# Re-using data from repositories

EUROPEAN GENOME-PHENOME ARCHIVE

Search... Tips on how to search

Helpdesk Log in

ABOUT SUBMISSION BROWSE ACCESS DOWNLOAD METADATA

Introduction Catalog Statistics EGA Statistics About the EGA Projects and Partnerships Quality Control Reports Privacy Notice Federated EGA GA4GH Security The Team Citing EGA

TECHNOLOGY SAMPLE TYPES

## What is in the EGA?

Studies in the EGA by disease

Number of studies

Disease Category	Number of Studies
Cancer	2500
Cardiovascular	201
Infectious	60
Inflammatory	237
Neurological	101
Other	2250

If applicable, a study may be included in more than one category

## Latest studies

Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. — 2021-09-30

Bottleneck Sequencing Of Human Tissue Wgs [Read more →](#)

Study 2 / 5 [Next Study](#)

Published in:

welcome trust sanger institute nature genetics

### 3.1 Genetic and phenotypic data

Within GDPR, there are two main actors: data controllers and data processors. Data controllers are persons or entities which determine the purposes and means that the personal data may be processed, e.g. companies, researchers, or universities. For EGA, the data controller is ultimately the data producer and the submitter(s) who submit the data to EGA. The data controller also creates a Data Access Committee (DAC) who will decide on data access permissions at EGA. Data processors are the persons or entities which process the data on behalf of a data controller. With regard to GDPR, EGA is a data processor as it processes data as instructed by the data controller. GDPR applies to any organization which accesses personal data from an individual within the EU. Under GDPR, personal

# Re-using data from repositories

1. Look into Licenses and Policies
  - a. For Reuse
  - b. Before depositing

<https://creativecommons.org/>



No conditions stated! Most open.



Credit for the original creators



New creations must be shared under  
the same terms



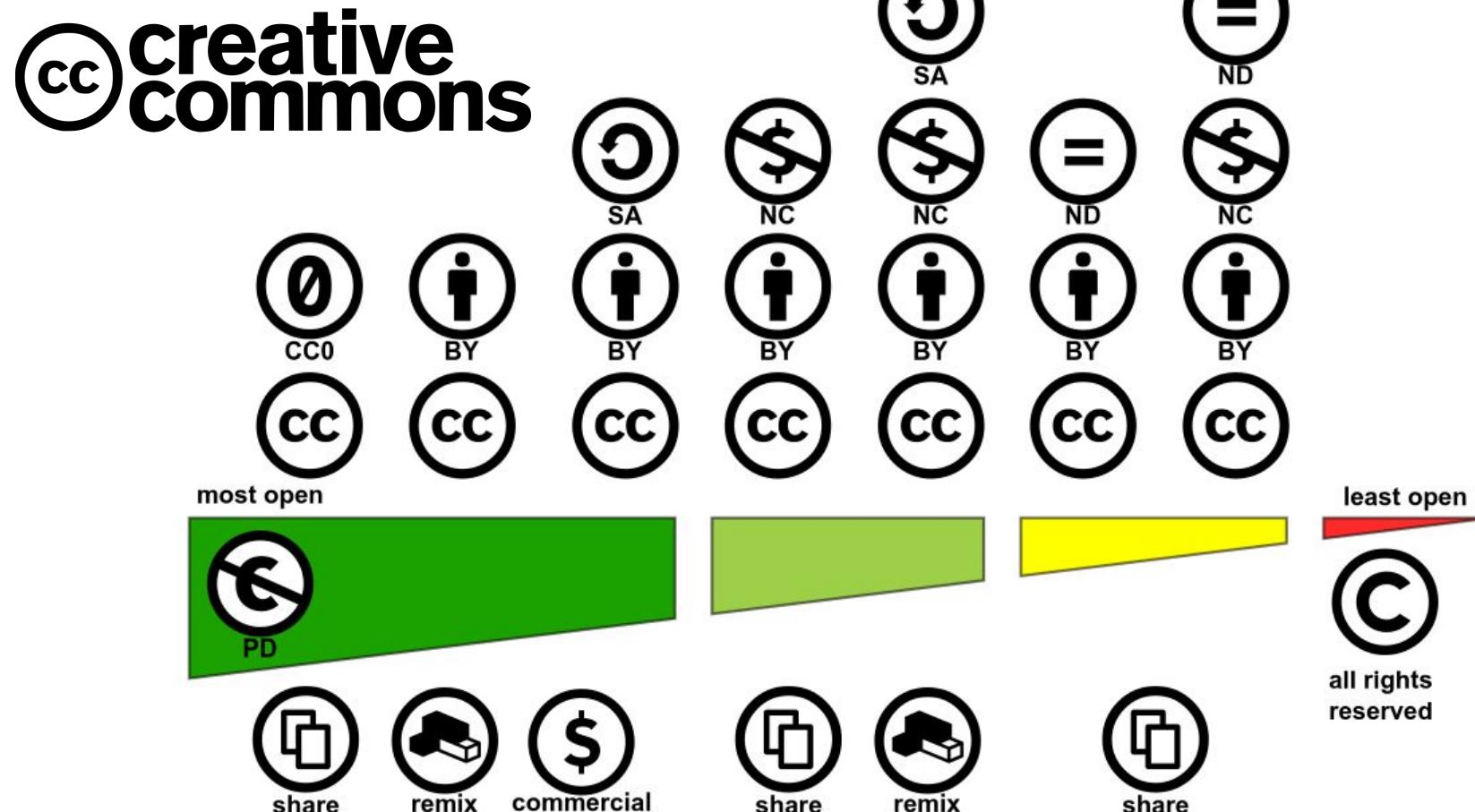
Only non-commercial use



Non derivatives or adaptations are  
allowed, only re-use in original form

# Re-using data from repositories

<https://creativecommons.org/share-your-work/cclicenses/>



CC-BY 4.0 Shaddim; original CC license symbols by <https://creativecommons.org/>

<https://creativecommons.org/choose>



# Re-using data from repositories

- **GNU GPLv3:**
- **GNU AGPLv3:** Complete source code modifications need to be available
- **GNU LGPLv3:** Large modifications can be distributed under different license.
- **MIT license:** No attribution license
- **Apache 2.0 ~ GNU PLv3**

Permissions	Conditions	Limitations
● Commercial use	● Disclose source	● Liability
● Distribution	● License and copyright notice	● Warranty
● Modification	● Same license	
● Patent use	● State changes	
● Private use		



<https://opensource.org/licenses>  
<https://choosealicense.com/>

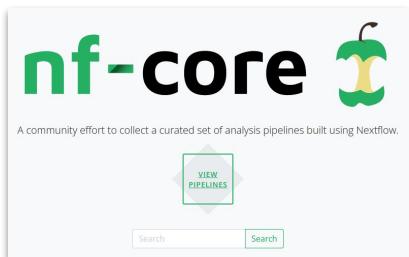
# A short side note on computational workflows

Where can I find workflows?

Platform/System  
specific repositories



IWC - Intergalactic Workflow Commission



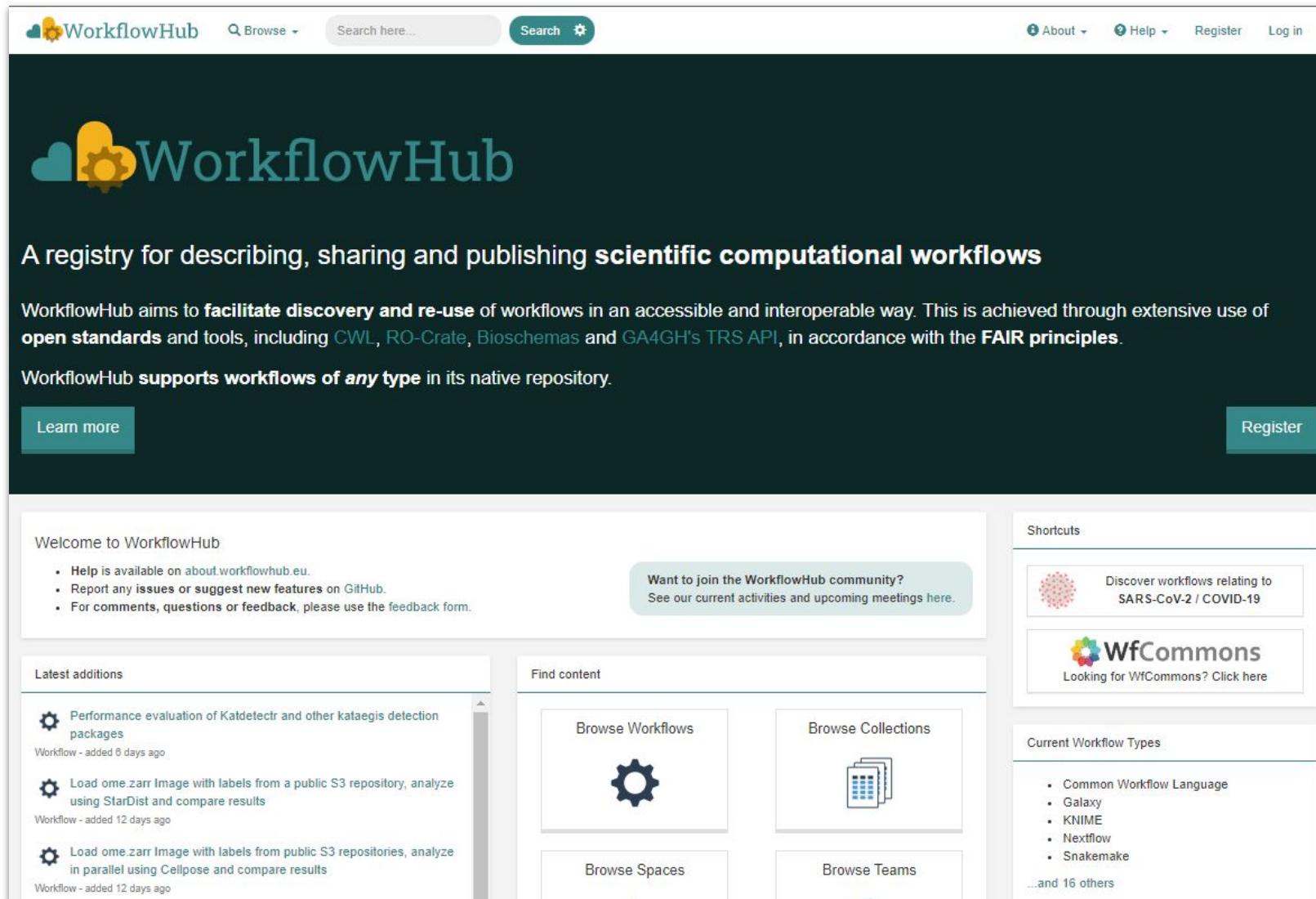
The Git based  
repositories



Generalist data  
repositories



# Using a registry to overcome fragmentation



The screenshot shows the WorkflowHub homepage. At the top, there's a navigation bar with links for 'About', 'Help', 'Register', and 'Log in'. Below the header, the WorkflowHub logo (a blue cloud with a yellow gear and heart icon) is displayed, followed by the text 'WorkflowHub' and 'A registry for describing, sharing and publishing **scientific computational workflows**'. A descriptive paragraph explains the platform's goal of facilitating discovery and re-use through open standards like CWL, RO-Crate, Bioschemas, and GA4GH's TRS API, in accordance with FAIR principles. It also mentions supporting workflows of any type. Two buttons, 'Learn more' and 'Register', are visible. The main content area includes sections for 'Welcome to WorkflowHub', 'Latest additions', 'Find content', and 'Shortcuts' (which links to WfCommons). The 'Latest additions' section lists three recent workflow submissions. The 'Find content' section provides links to browse workflows, collections, spaces, and teams. The 'Shortcuts' section links to SARS-CoV-2 / COVID-19 workflows and WfCommons.

Start: April 2020  
Launch: Sept 2022

<https://workflowhub.eu/>



# WorkflowHub

## Key features

- System agnostic
- Native repository support
- Git integration
- Versioning
- Links to external files (docs, test and reference data)
- Metadata extraction and standards
- Integration with Bio.tools
- Author credit
- Citable
- DOI minting

A sandbox is available

<https://dev.workflowhub.eu/>

The screenshot shows a detailed view of a workflow entry on the WorkflowHub platform. The main title is 'MC\_COVID19like\_Assembly\_ Reads'. Key sections visible include:

- access**: Includes links to 'Visit source', 'Download RO-Crate', and 'Run on usegalaxy.eu'.
- versions & status**: Details the deposition process, mentioning SARS-CoV-2 assembly and re-assembly steps.
- licensing**: Apache Software License 2.0.
- analytics**: Shows views (398), downloads (24), and activity logs.
- other workflows**: A section listing related workflows.
- Emphasis on metadata**: A prominent callout box highlighting the focus on metadata.

The interface also shows a Galaxy tool panel and a sidebar with creator information, citation details, and attribution tags like 'covid-19'.

# Integration with different platforms

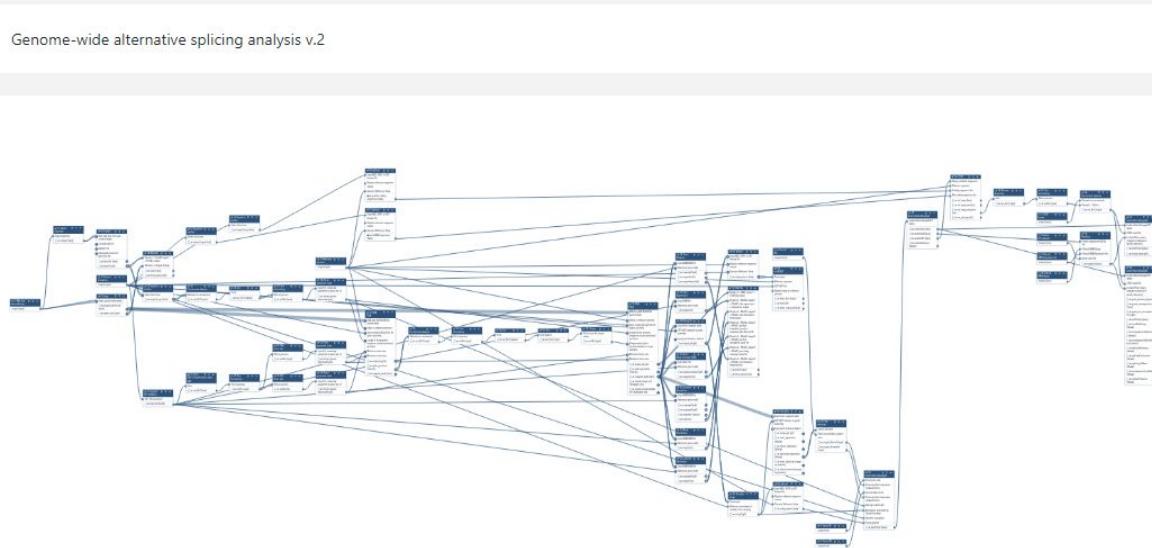
WorkflowHub Search here... Search

Galaxy **Genome-wide alternative splicing analysis v.2** Version 7 (latest)

Overview Files Related items

Workflow Type: Galaxy Stable

Genome-wide alternative splicing analysis v.2



SEEK ID: <https://workflowhub.eu/workflows/482?version=7>

### Inputs

ID	Name	Description	Type
Active sites dataset	n/a	Active sites dataset.	File

DOI: 10.48546/workflowhub.workflow.482.3

Creators and Submitter

Creator Cristóbal Gallardo  
Submitter Cristóbal Gallardo

License Creative Commons Attribution-NonCommercial 4.0

Activity Views: 97  
Created: 25th May 2023 at 23:01  
Last updated: 7th Jun 2023 at 17:00

Annotated Properties

Topic annotations Biomedical science, Transcriptomics  
Operation annotations Transcriptome assembly

Tags



<https://usegalaxy.eu/>

Connecting workflow discovery to analysis





# Get in touch with us.



[info@elixir-belgium.org](mailto:info@elixir-belgium.org)



<https://www.elixir-belgium.org>



This work is licensed under the Creative Commons Attribution 4.0 International License.  
To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>