

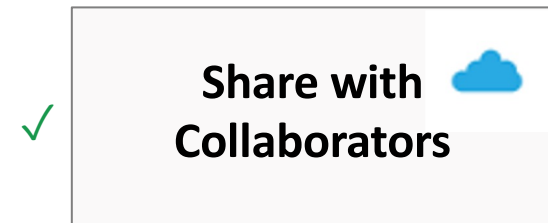
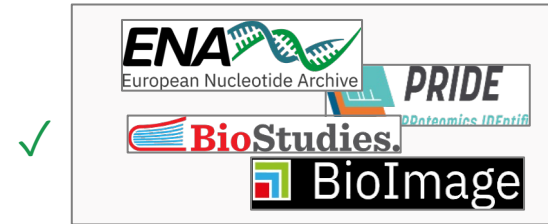
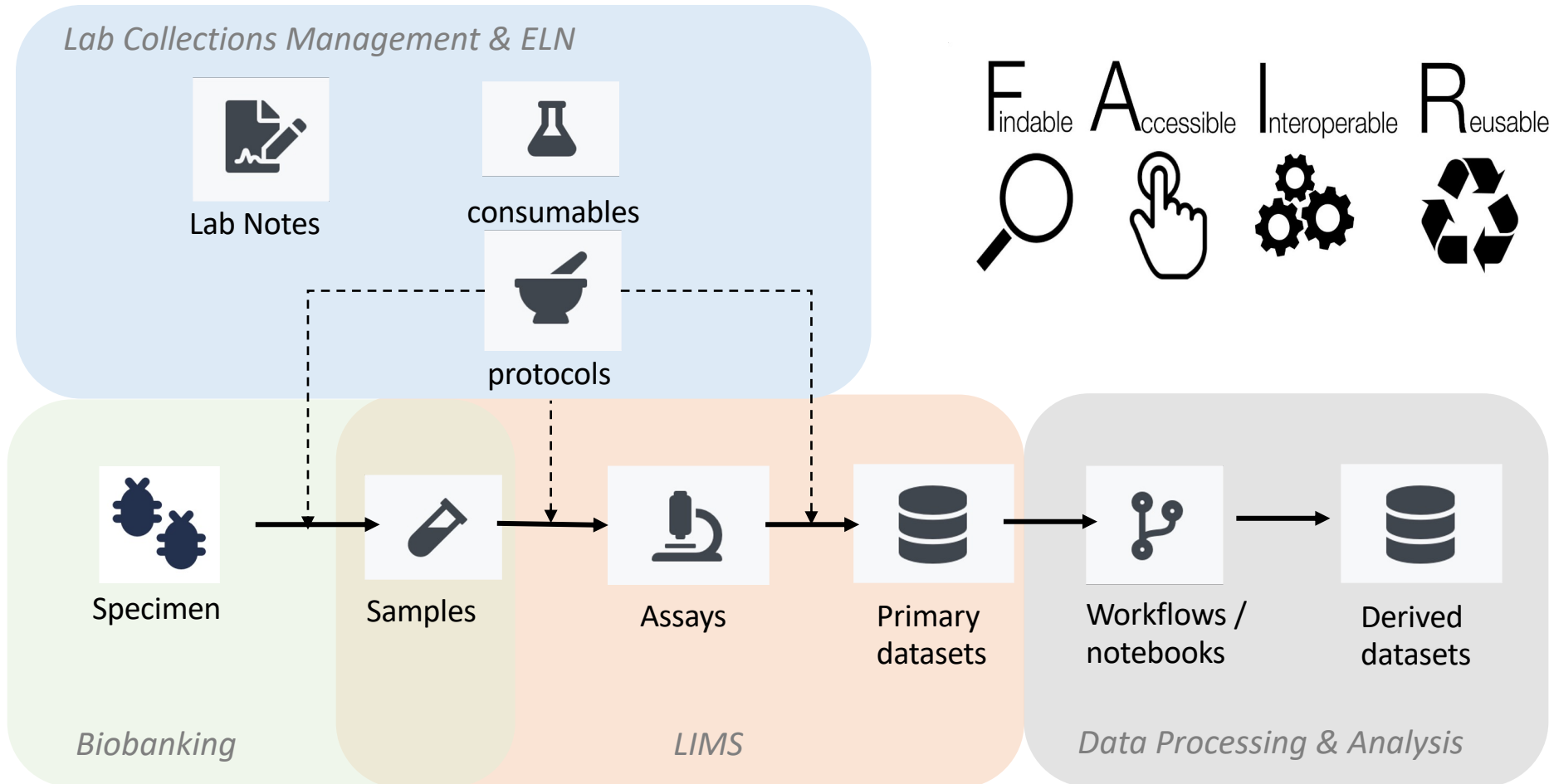
Data Export & Submission with LabID

An abstract graphic featuring a teal background. On the right side, there is a glowing, yellowish-white network of lines and nodes, resembling a data visualization or a molecular structure. Several colorful circles (pink, teal, and white) are scattered across the image, some overlapping the network. The overall aesthetic is modern and technological.

Charles Girardot
Multimodal Open Data Integration Support (MODIS)
Data Science Center

Sharing Data following the FAIR principles

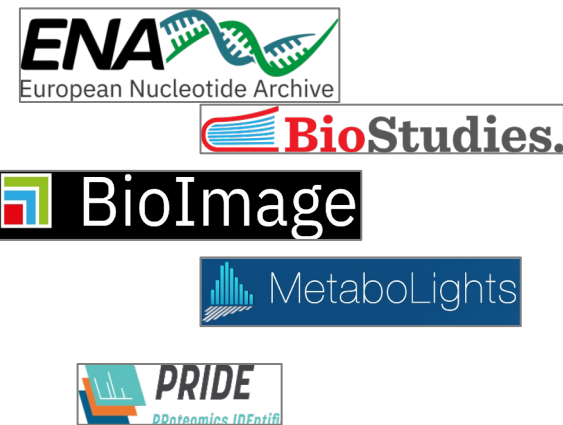
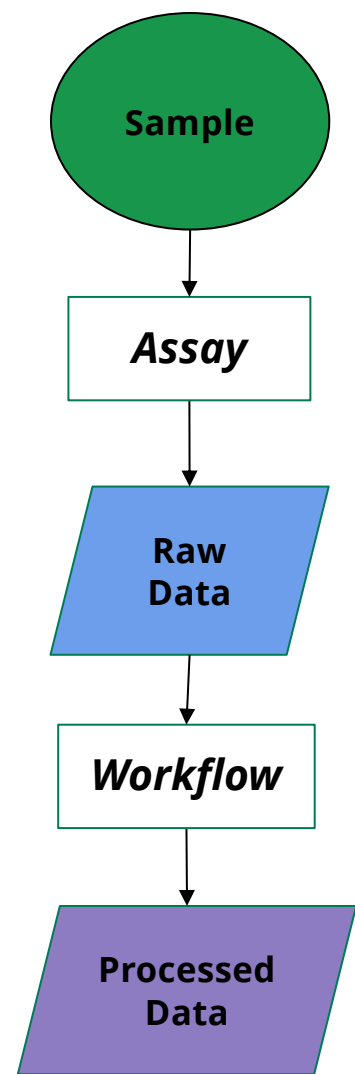
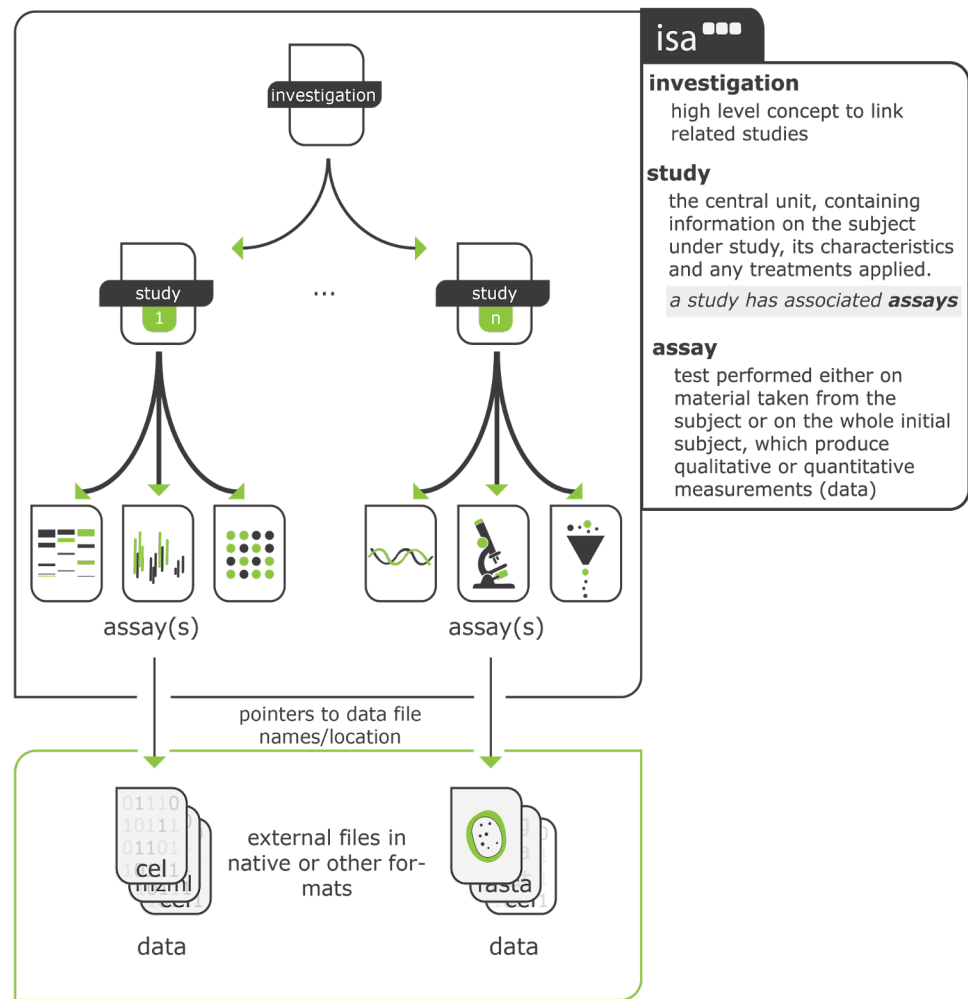
- Deposition of your data in a public repository is required by many journals
- Is the basis of Open Science (data reuse)
- Curated data is key to develop AI models (e.g. AlphaFold)



The different EBI Data Repositories

Repository	Data type
 BioSamples	Biological sample metadata
 ENA	DNA and RNA sequences; raw sequencing reads
 EVA	Genetic variants
 EGA	Human genetic and phenotypic data requiring controlled access
 ArrayExpress in BioStudies	Gene expression and other functional genomics data
 PRIDE	Proteomics data
 MetaboLights	Metabolomics data
 PDBe	3D macromolecular structures
 BioImage Archive	Biological images
 EMPIAR	Raw 3D electron microscopy imaging data
 BioStudies	Multimodal studies and data without a specialised archive

Common data representation in many EBI data repositories



Data submission consists in providing this (meta)data with their relationship

What's different between (EBI) data repositories?

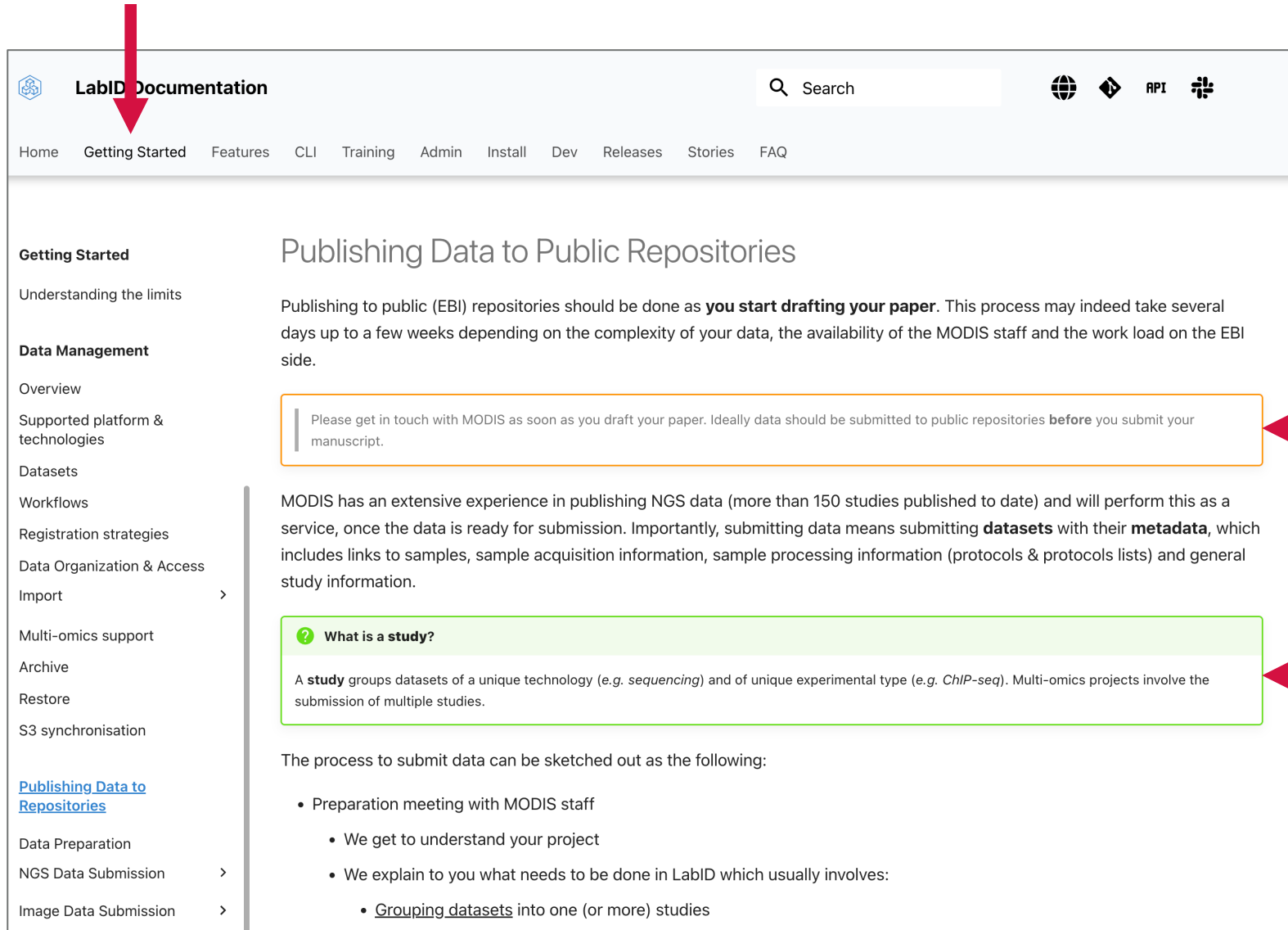
- Type of data they accept
- Submission routes
 - API, CLI, UI
- Programmatic submission formats:
 - Not all repositories support programmatic submission
 - No unified data submission format !
 - XML (e.g. ENA)
 - JSON (e.g. EGA)
 - Spreadsheet and variation around MAGE-TAB *like* (discontinued?), PAGE-TAB (BioStudies)...

Which submission is supported by LabID

- LabID supports ENA/EGA submissions
 - Annotare dropped MAGE-TAB submission type
 - Easy to copy-paste info from MAGE-TAB export
- Image Bio Archive in progress
 - Waiting them to release JSON support
- Supporting other repos first requires they provide programmatic support

=> We'll demo a submission to ENA using the data we registered during this training

Publishing guidelines available in the Doc



The screenshot shows the LabID Documentation website. A red arrow points to the 'Getting Started' link in the top navigation bar. Another red arrow points to the 'Publishing Data to Public Repositories' section in the left sidebar. A third red arrow points to an orange-bordered box containing text about contacting MODIS. A fourth red arrow points to a green-bordered box titled 'What is a study?'. A fifth red arrow points to the 'Publishing Data to Repositories' link in the sidebar.

LabID Documentation

Home Getting Started Features CLI Training Admin Install Dev Releases Stories FAQ

Getting Started

Understanding the limits

Data Management

Overview

Supported platform & technologies

Datasets

Workflows

Registration strategies

Data Organization & Access

Import >

Multi-omics support

Archive

Restore

S3 synchronisation

[Publishing Data to Repositories](#)

Data Preparation

NGS Data Submission >

Image Data Submission >

Publishing Data to Public Repositories

Publishing to public (EBI) repositories should be done as **you start drafting your paper**. This process may indeed take several days up to a few weeks depending on the complexity of your data, the availability of the MODIS staff and the work load on the EBI side.

Please get in touch with MODIS as soon as you draft your paper. Ideally data should be submitted to public repositories **before** you submit your manuscript.

MODIS has an extensive experience in publishing NGS data (more than 150 studies published to date) and will perform this as a service, once the data is ready for submission. Importantly, submitting data means submitting **datasets** with their **metadata**, which includes links to samples, sample acquisition information, sample processing information (protocols & protocols lists) and general study information.

? What is a study?

A **study** groups datasets of a unique technology (e.g. *sequencing*) and of unique experimental type (e.g. *ChIP-seq*). Multi-omics projects involve the submission of multiple studies.

The process to submit data can be sketched out as the following:

- Preparation meeting with MODIS staff
 - We get to understand your project
 - We explain to you what needs to be done in LabID which usually involves:
 - Grouping datasets into one (or more) studies

Restricted to EMBL

Data Preparation Check List

Getting Started

Understanding the limits

Data Management

Overview

Supported platform & technologies

Datasets

Workflows

Registration strategies

Data Organization & Access

Import >

Multi-omics support

Archive

Restore

S3 synchronisation

Publishing Data to Repositories

[Data Preparation](#)

NGS Data Submission >

Image Data Submission >

Data Preparation

Common data preparation steps to perform in LabID.

Step 1: Prepare the Study's dataset content

The data to submit comprises sample descriptions, protocols and files (fastq or image files) associated with **one Study** (in LabID a Study is very similar to a study in [ISA](#) standards terms). One study should contain data of one *experimental type* only. For example, in the NGS case we don't mix data from *ChIP-seq*, with *RNA-seq*, *HiC* or *DNA-seq* so if you have to submit different experimental types, you will need to organize them into different studies. One exception to this rule is single cell multiome studies where e.g. the RNA-seq and ATAC-seq data is linked to the same study.

1. Make sure to have a study with only the datasets to be submitted

Your study should only contain the datasets to be submitted. If your study contains more datasets than the ones to be submitted, the easiest is to (1) create a new study and (2) associate all relevant datasets with it using the batch edit option.

From the study page containing more datasets than the ones to be submitted, select the datasets to be submitted and use the **Batch Edit** option to associate them with a new study.

Datasets 

On this page

Step 1: Prepare the Study's dataset content

1. Make sure to have a study with only the datasets to be submitted
2. Consolidate the study name and description
3. Position the study design terms, experimental factors and suggested annotations

4. Study contributors

Step 2: Connect protocols to samples

Expected protocols for NGS data

Expected protocols for image data

Step 3: Annotate the Samples

Additional expectations for sequencing libraries (NGS)

Additional information for NGS single-cell studies

Additional expectations for EM Samples (Electron Microscopy)

Help! My samples are duplicated



Standard-compliant metadata for (EBI) submission

ENA Library Information including single-cell support

Sequencing Library Preparation			
Library Kit Type	10x 3' v3.1	Orientation	5'-3'-3'-5' (5'-3'-3'-5')
Barcode	GTCCTTCGGC CTGTGCATGA		
Source	TRANSCRIPTOMIC SINGLE CELL (RNA-seq of coding RNA from single cells or RNA-seq of non coding RNA from single cells)	Strategy	ssRNA-seq (ssRNA-seq)
Selection	cDNA_oligo_dT (cDNA_oligo_dT)	RT Primer Type	oligo-dT
Library Strand	NA	End Bias	3 prime end bias

Single-Cell Sample			
Single Cell Type	normal cell	Cell Number	12927
Single Cell Library Construction	10x 3' v3	Single Cell Isolation	10X

Study Design, Exp Factor & Contributors

Name

Bait-capture based single molecule footprinting of Sox2 degran mESC line

Project

Genetic_variation

Design

biological replicate , genetic modification design , strain or line design

PubMed ID

—

Description

Bait-capture based Single Molecule Footprinting (SMF) data. SMF data is obtained by treating extracted nuclei with a GpC and a CpG methyltransferase, where binding of proteins on DNA, e.g. nucleosomes and transcription factors (TFs), leave behind unmethylated cytosines as footprints. Data in this experiment comprises SMF data obtained from a previously reported Sox2 degran mESC line (see PMID: 33318687). Sequencing libraries were prepared using Agilent Sure-Select Mouse Methyl-Seq kit, enriching the sample for cis-regulatory regions of the mouse genome prior to library preparation. Thus, these data contain high coverage accessibility information at regulatory loci in different cell types.

Nr. of datasets

7

Archives

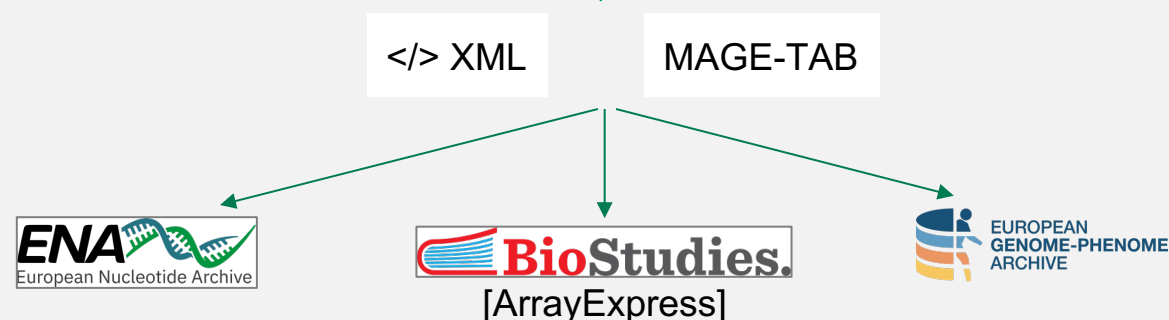
—

Experimental Factors & Required Annotations

7

Contributors

3



✓ 180+ studies submitted to date

We have almost done most of the preparation already during the workshop

Hands-on - Publish a Sequencing Study

 60 min ·  Moderate

This hands-on follows a complete RNA-seq experimental story from biological setup to submission-ready exports.

Overview

The study compares **Brpf1 knockout (KO)** versus **wild-type (WT)** neurons derived from BALB/c mouse brain tissue, with **two biological replicates per condition** (4 libraries in total).

The experimental work can be summarized as follows:

- brain tissue is dissected from four BALB/c mice
- neurons are isolated and cultured
- Brpf1 knockout is introduced in the KO condition
- RNA is extracted
- mRNA libraries are prepared
- paired-end sequencing is performed on Illumina HiSeq 4000.

In this training, you will represent this experimental flow in LabID so that the final metadata, annotations, protocols and exports are coherent, reusable, and compatible with repository requirements.

Learning Objectives

On this page

Step 1. Register the datasets

Step 2. Edit the study metadata

2.1 Name, description & design terms

2.2 Experimental factors and annotations

2.3 Contributors

Step 3. Annotate sequencing libraries

3.1 Set study context and download the Excel template for batch edition

3.2 Excel-based batch annotation

Step 4. Link protocols

4.1 Mandatory vs recommended protocols

4.2 Protocol summary vs description

4.3 Protocols used in this tutorial

Step 5. Export MAGE-TAB to validate your study

Step 6. Submit to ENA using XML format

6.1 Export tables suitable for the ENA upload CLI tool

6.2 Submit XML to ENA

Walkthrough

We Made It This Far Without Mentioning AI!

This week end I wrote a skill reference in the LabID CLI...

Prompt to Claude: I want to submit study 738d48b8-e07d-489e-8622-249fa72ab50c to ENA, is it ready to go?

Context

The user wants to submit the study "Paired-end sequencing - mouse [Trainee99]" (training server, trainee99) to ENA. I checked it read-only against references/submission-readiness.md. The samples are *Mus musculus*, so ENA is the right archive. It has 4 bulk mRNA-seq PE libraries (Brpf1 KO ×2, WT ×2, NovaSeq 6000) with 4 FASTQ datasets and 4 BAM datasets.

Verdict: NOT READY. It has 6 blockers.

Findings

Sev	Doc step	Issue	Affected	Fix
B	§1.2	Study description is empty	study	Write a technical description. I'll draft it only once the other blockers are fixed.
B	§1.3	No design term	study	Add genotype design + genetic modification design (UI, or labid manage set-field -m study -f design)
B	§2	No Culture & Growth protocol linked	all 4 libraries	Create or link a growth/husbandry protocol
B	§2	"Mouse brain extraction" summary is a placeholder ("Describe the extraction...")	protocol fc6dad6e	Rewrite the summary
B	§3 NGS	Library Strategy / Source / Selection / Orientation are empty	T99_brain_{WT,KO}_rep{1,2}	labid manage set-field -m sample -i <lib> -f strategy -v RNA-Seq; source=TRANSCRIPTOMIC; selection=PolyA; orientation (PE)
B	ENA run	No MD5 checksum on any file	all 12 files	Compute and register checksums
W	§1.1	4 processed BAMs (is_raw=false) sit next to the raw FASTQ	T99_{WT,KO}{1,2}.bam	Move them to another study
W	§1.1	QC flag not set	all 8 datasets	Set QC
W	§1.2	Name has an internal tag [Trainee99]	study	Rename to a public title
W	§1.3	Varying annotations aren't declared as factors	IndividualGeneticCharacteristics, GeneticModification	Add required_annotations (mandatory + factor, SEQUENCINGLIBRARY)
W	§1.4	No contributors stored	study	Add a submitter or PI with email and affiliation
W	§2	Two summaries only repeat the protocol name	RNeasy extraction, NEBNext library prep	Say what was done, e.g. "per manufacturer's instructions, kit ref ..."
W	§3	DevelopmentalStage annotation missing	4 libraries	Annotate (e.g. adult)
N	§3	Collection Date / Country not set	-	The export fills defaults, which is acceptable
N	§3	Material Type comes out as # TODO in MAGE-TAB	-	Pass --source-type-name at export

Everything else passes: organism is a species taxon, there's only one experiment type, instrument model is set, factor annotations are filled on every sample, and no library is duplicated.

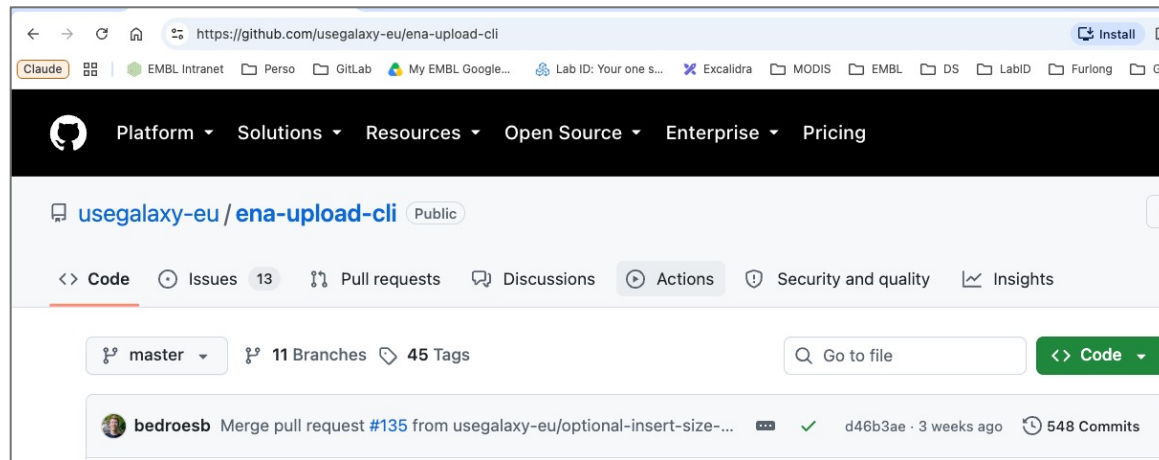
Suggested design terms

- genotype design (varying): IndividualGeneticCharacteristics = Brpf1- / wild-type

- genetic modification design (varying): GeneticModification = gene knock out / none

DEMO TIME

still the old manual way :-)



- The submission leverages the open-source **ena-upload-cli** tool

```
S=738d48b8-e07d-489e-8622-249fa72ab50c
```

```
OUT=~/.labid_ena_export/$Smkdir -p $OUT
```

```
labid export study -s $S -o $OUT. # export MAGE-TAB
```

```
labid export study -s $S --format enatables -o $OUT --study-type rnaseq # export tables
```

```
labid export ena -f /Users/girardot/.labid/ena.yaml -c EMBL -s $S -d $OUT
```

Thank you

Multimodal Open Data Integration Support (MODIS) part of EMBL Data-Science Center



Charles Girardot

Head of Multimodal Open Data Integration Support

ORCID: [0000-0003-4301-3920](https://orcid.org/0000-0003-4301-3920)



Laurent Thomas

Postdoctoral Fellow

ORCID: [0000-0001-7686-3249](https://orcid.org/0000-0001-7686-3249)



Md Nayeem Reza

Senior Software Engineer

ORCID: [0000-0003-2068-5812](https://orcid.org/0000-0003-2068-5812)



Matthias Monfort

Web Developer / Workflow Management

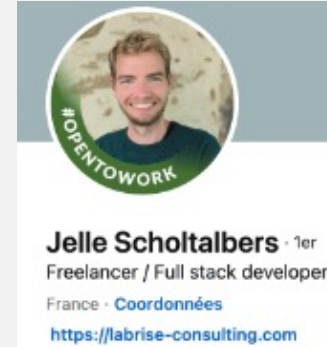


Marco Wetter

Full Stack Developer

Contact : modis@embl.de

Chat: <https://chat.embl.org/embl/channels/labid>



Join LabID on Slack