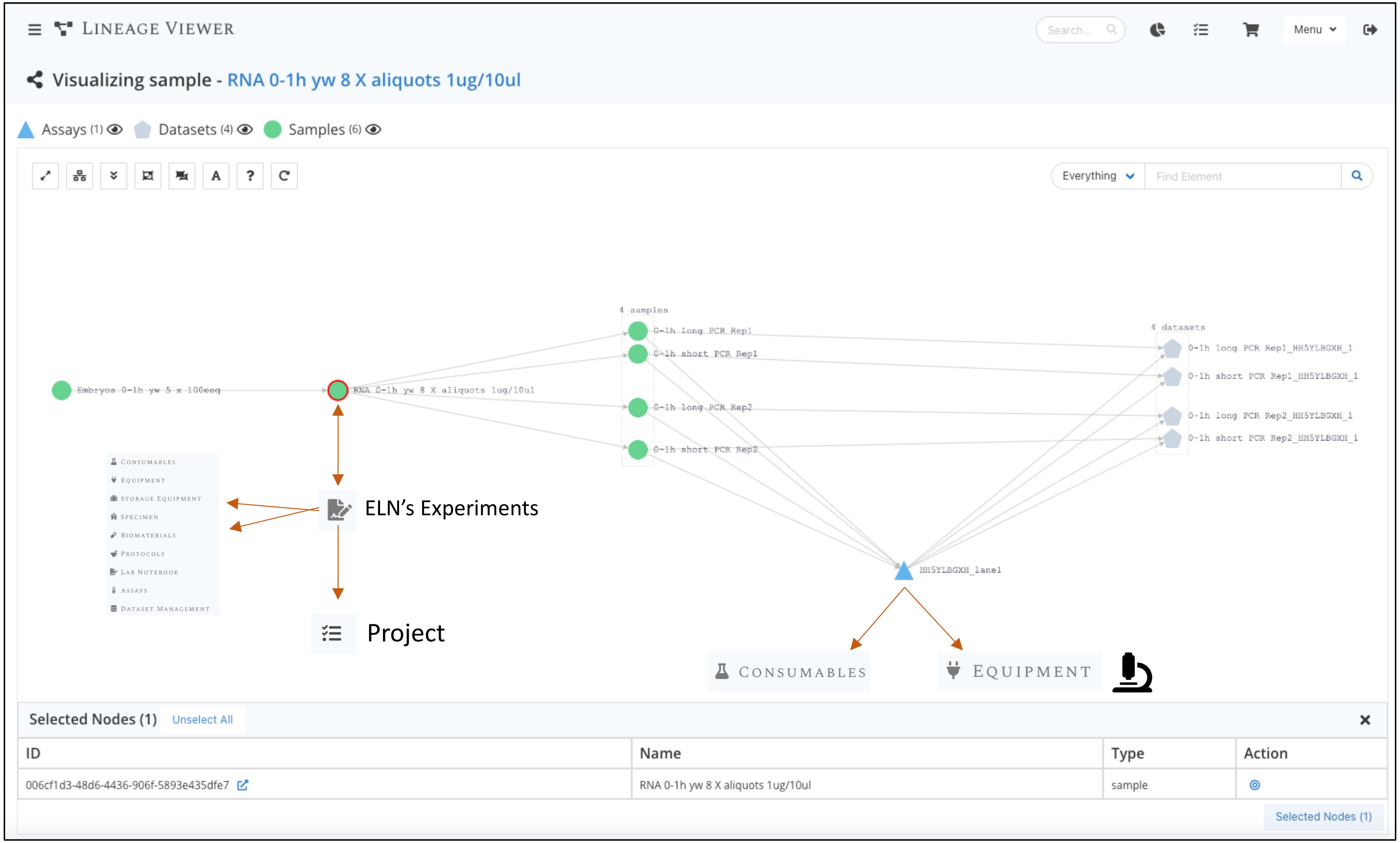
STOCKS Training: Assay & Raw Data Management

Charles Girardot

Genome Biology Computational Support (GBCS)

From Specimen to Datasets: Complete Sample Lineage



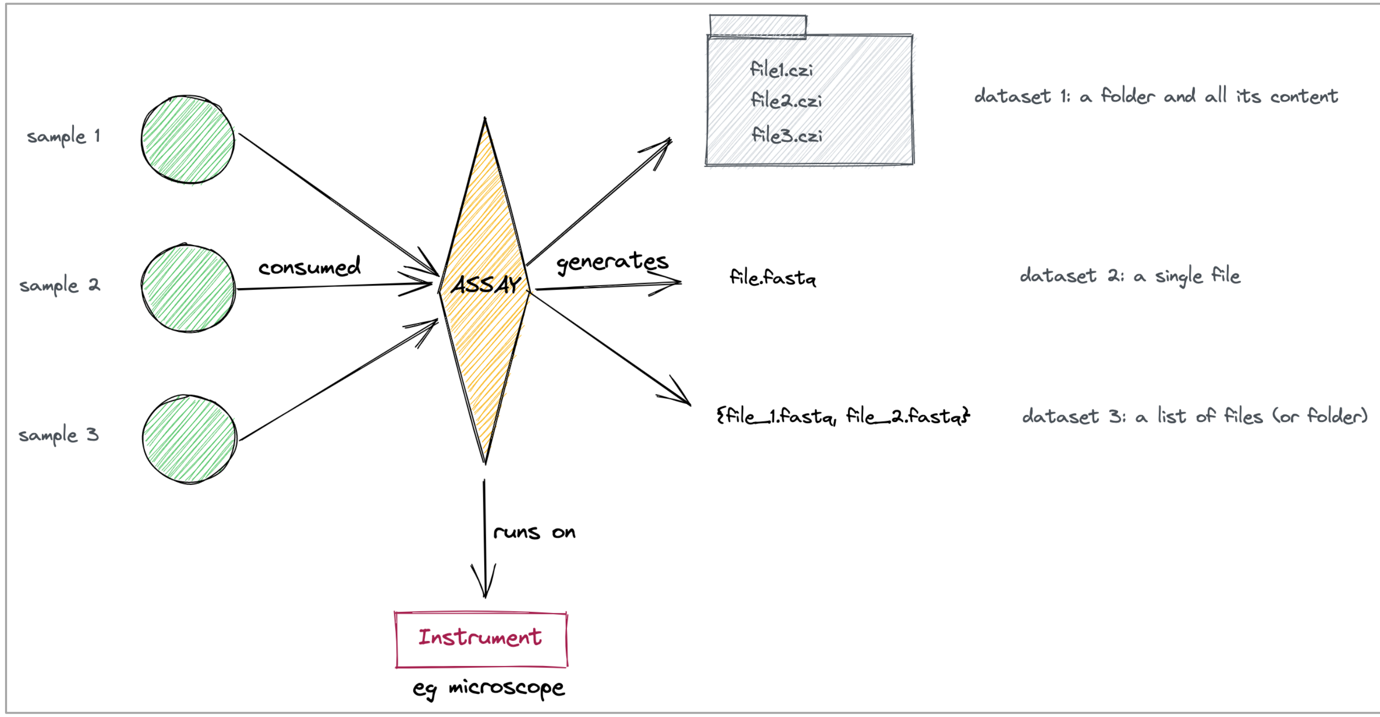
Visualizing sample - RNA 0-1h yw 8 X aliquots 1ug/10ul

Selected Nodes (1)

ID	Name	Type	Action
006cf1d3-48d6-4436-906f-5893e435dfe7	RNA 0-1h yw 8 X aliquots 1ug/10ul	sample	◎

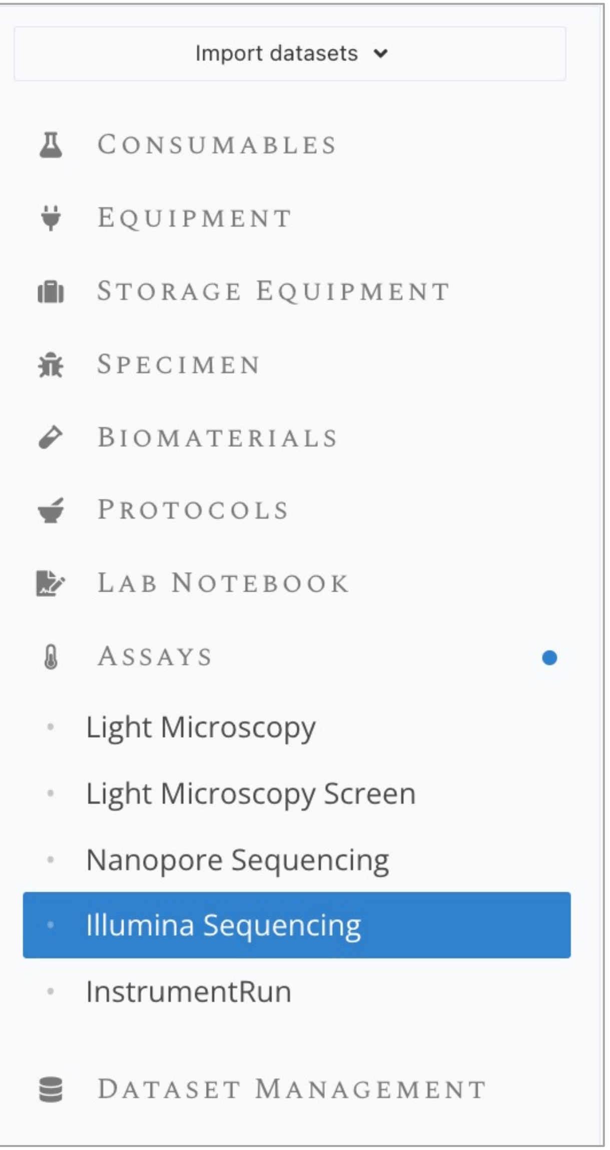
Assays consume samples and generate raw datasets

- *Assay* is short for **raw data generating assay**
- Raw data is generated:
  - **from samples** prepared in a specific way e.g. a sequencing library
  - by ***running a specific instrument*** e.g. a sequencer or a microscope



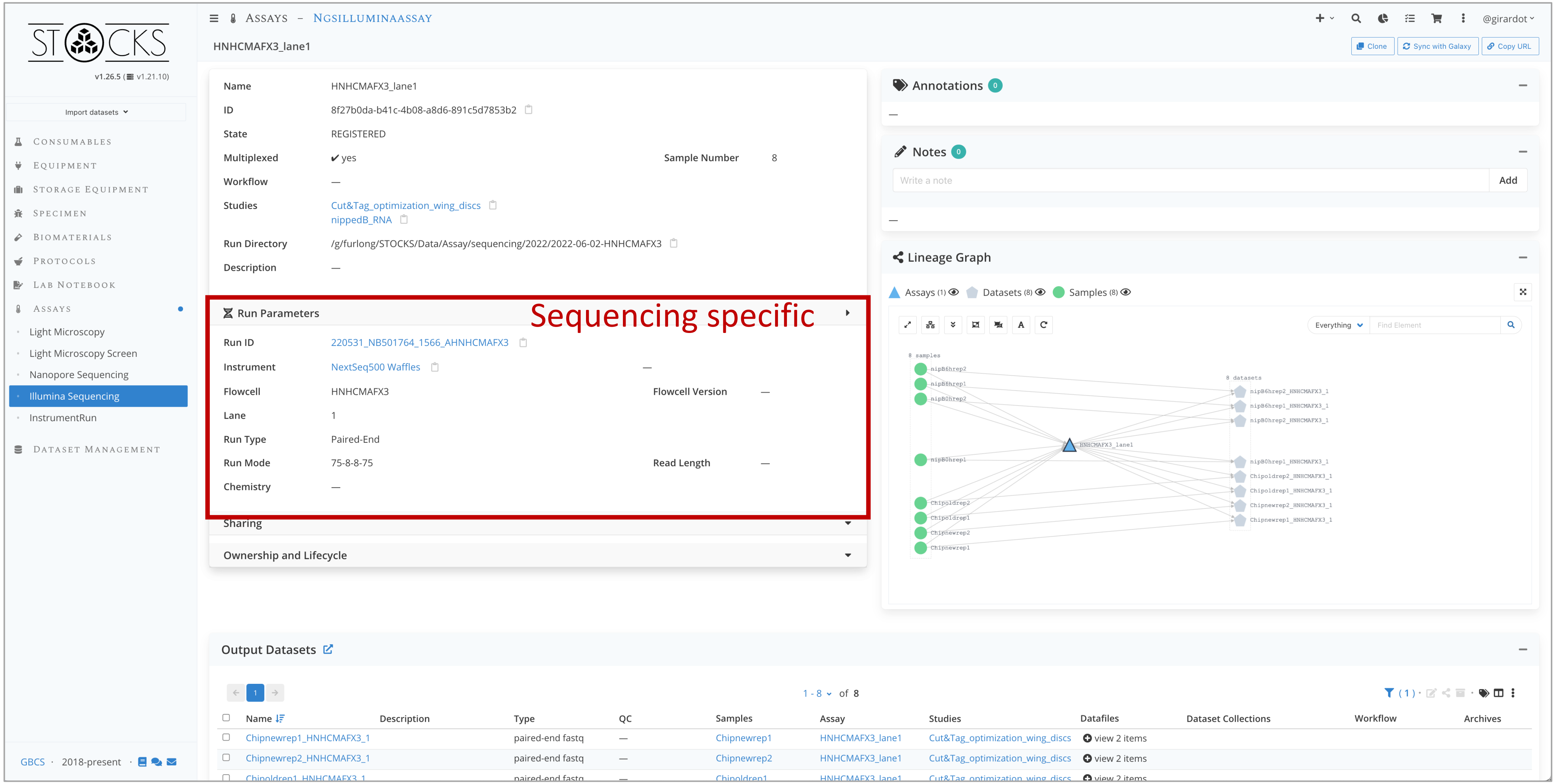
- Raw data can be :
  - One file
  - Multiple files
  - One folder
  - Multiple folders
- A file (folder) is a *datafile*
- *Registering raw data == assay creation*

Assay types reflect data standard and user needs



- Assay types are specific to a particular experiment
- They belong to one of the technology:
  - Sequencing : Illumina Sequencing, Nanopore Sequencing
  - Light microscopy: LM or LM screen
  - [electron microscopy, proteomics, metabolomics, flow cytometry]
- Together with sample sub-types, they catch the FAIR aspects of a technology
  - allow easy data deposition
- Assay (and sample) sub-types also accommodate the user and institute needs
  - Custom properties available

The Illumina & Nanopore Sequencing Assay



HNHCMAFX3_lane1

Name	HNHCMAFX3_lane1		
ID	8f27b0da-b41c-4b08-a8d6-891c5d7853b2		
State	REGISTERED		
Multiplexed	✔ yes	Sample Number	8
Workflow	—		
Studies	Cut&Tag_optimization_wing_discs nippedB_RNA		
Run Directory	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022-06-02-HNHCMAFX3		
Description	—		

Run Parameters

Sequencing specific

Run ID	220531_NB501764_1566_AHNHCMAFX3		
Instrument	NextSeq500 Waffles		—
Flowcell	HNHCMAFX3	Flowcell Version	—
Lane	1		
Run Type	Paired-End		
Run Mode	75-8-8-75	Read Length	—
Chemistry	—		

Annotations 0

—

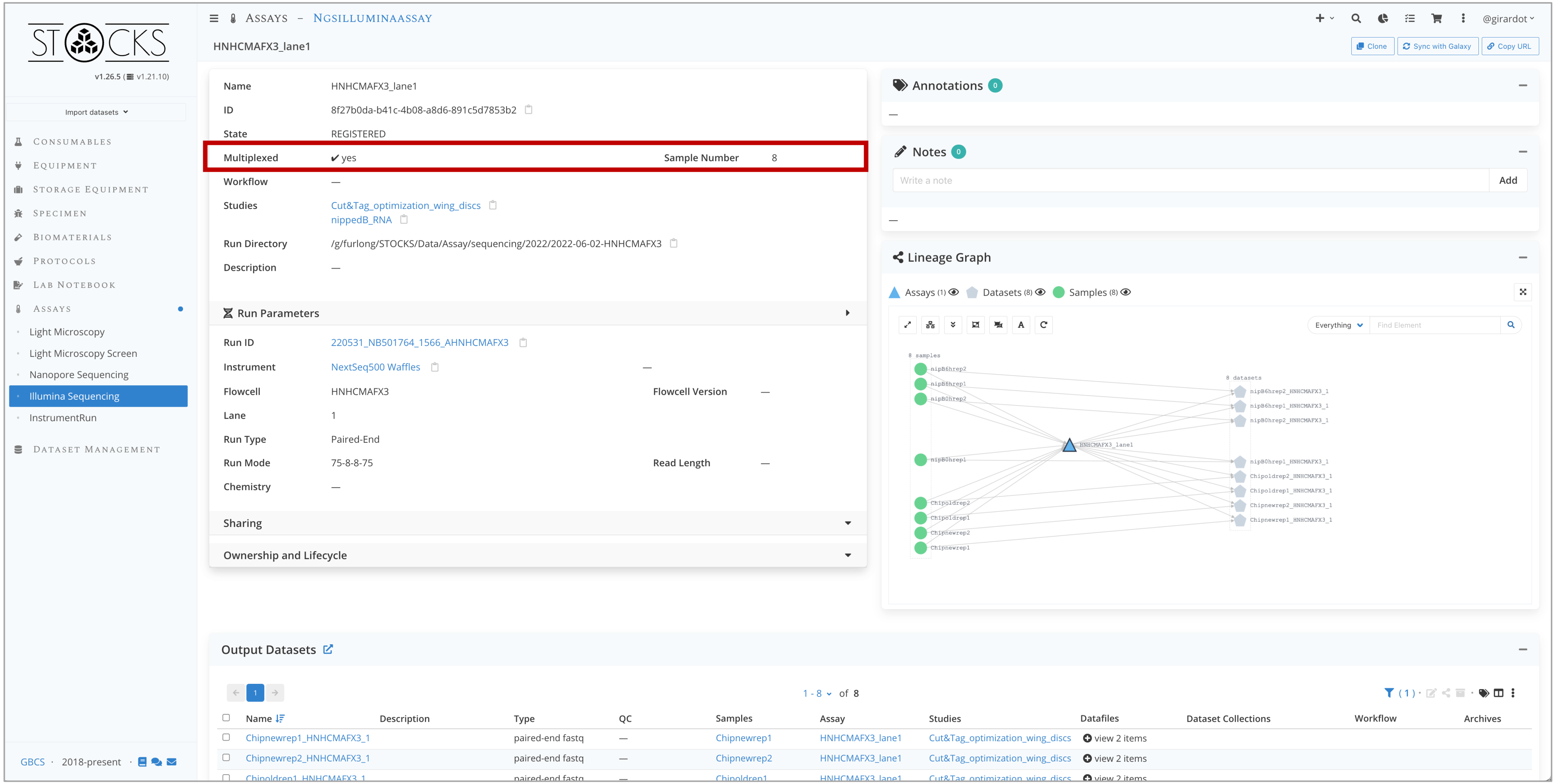
Notes 0

—

Output Datasets

☐	Name	Description	Type	QC	Samples	Assay	Studies	Datafiles	Dataset Collections	Workflow	Archives
☐	Chipnewrep1_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipnewrep2_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep2	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipoldrep1_HNHCMAFX3_1		paired-end fastq	—	Chipoldrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			

Assays have a multiplexing status (and sample number)



HNHCMAFX3_lane1

Name	HNHCMAFX3_lane1		
ID	8f27b0da-b41c-4b08-a8d6-891c5d7853b2		
State	REGISTERED		
Multiplexed	✔ yes	Sample Number	8
Workflow	—		
Studies	Cut&Tag_optimization_wing_discs nippedB_RNA		
Run Directory	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022-06-02-HNHCMAFX3		
Description	—		

Run Parameters

Run ID	220531_NB501764_1566_AHNHCMAFX3		
Instrument	NextSeq500 Waffles		—
Flowcell	HNHCMAFX3	Flowcell Version	—
Lane	1		
Run Type	Paired-End		
Run Mode	75-8-8-75	Read Length	—
Chemistry	—		

Sharing

Ownership and Lifecycle

Annotations 0

—

Notes 0

Write a note Add

—

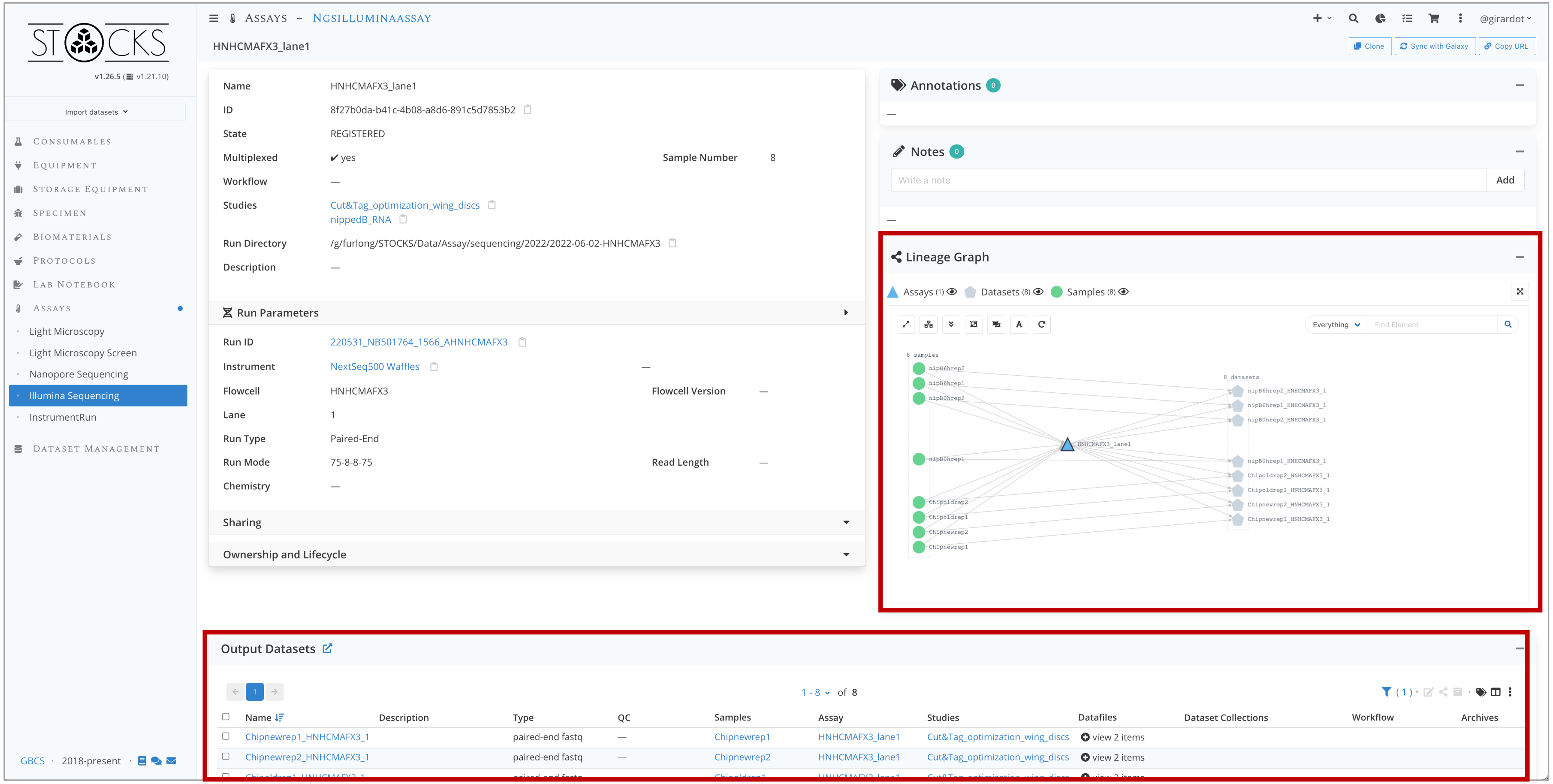
Lineage Graph

Output Datasets

☐	Name	Description	Type	QC	Samples	Assay	Studies	Datafiles	Dataset Collections	Workflow	Archives
☐	Chipnewrep1_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipnewrep2_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep2	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipoldrep1_HNHCMAFX3_1		paired-end fastq	—	Chipoldrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			

Also : Workflow can be linked ; Annotations and Notes as often

Easily navigate to consumed samples and produced datasets



HNHCMAFX3_lane1

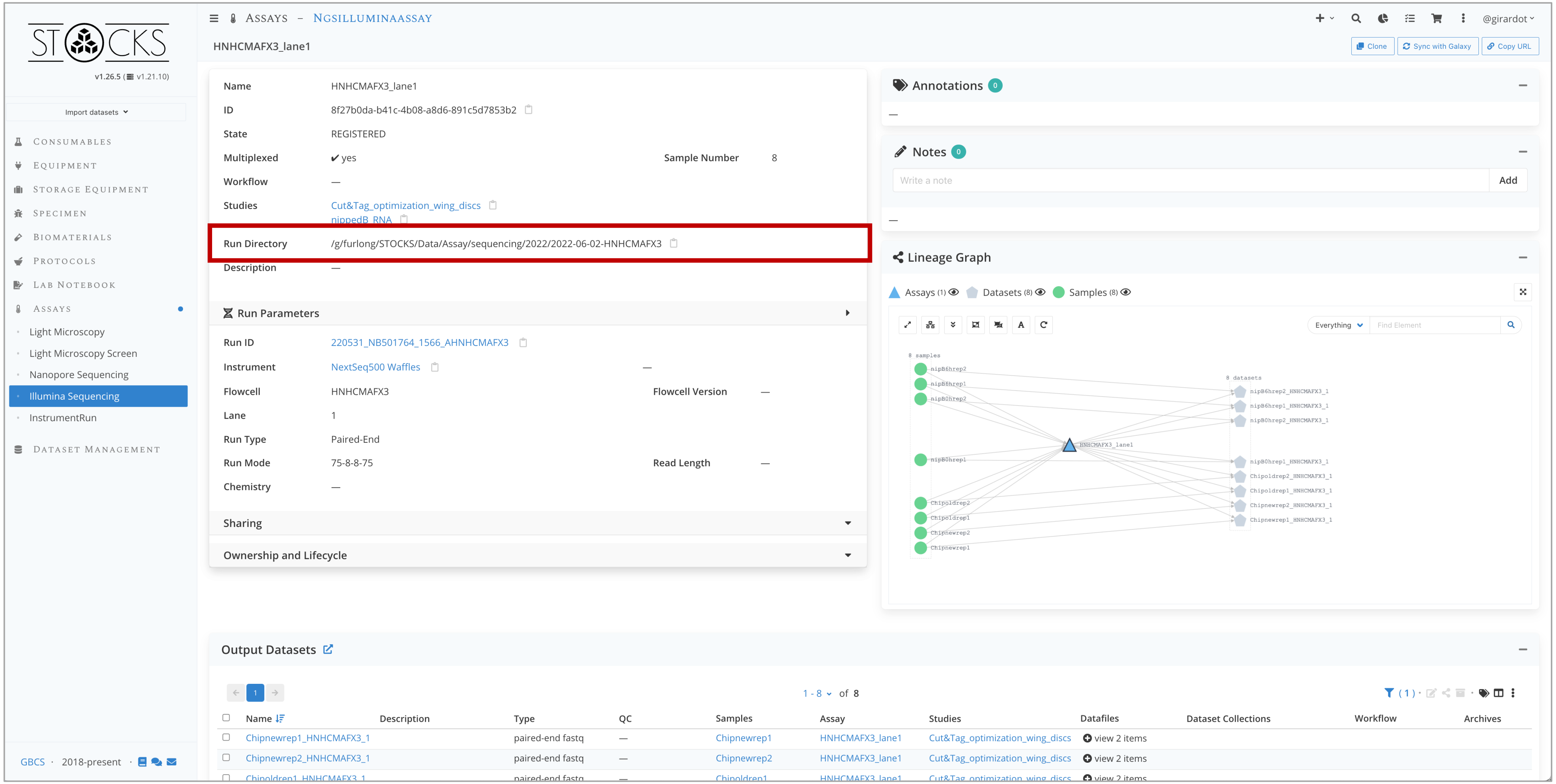
Name: HNHCMAFX3_lane1
ID: 8f27b0da-b41c-4b08-a8d6-891c5d7853b2
State: REGISTERED
Multiplexed: ✔ yes
Sample Number: 8
Workflow: —
Studies: Cut&Tag_optimization_wing_discs
nippedB_RNA
Run Directory: /g/furlong/STOCKS/Data/Assay/sequencing/2022/2022-06-02-HNHCMAFX3
Description: —

Run Parameters

Run ID: 220531_NB501764_1566_AHNHCMAFX3
Instrument: NextSeq500 Waffles
Flowcell: HNHCMAFX3
Flowcell Version: —
Lane: 1
Run Type: Paired-End
Run Mode: 75-8-8-75
Read Length: —
Chemistry: —

☐	Name	Description	Type	QC	Samples	Assay	Studies	Datafiles	Dataset Collections	Workflow	Archives
☐	Chipnewrep1_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipnewrep2_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep2	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			

The assay datafiles are located in the *run directory*



HNHCMAFX3_lane1

Name: HNHCMAFX3_lane1
ID: 8f27b0da-b41c-4b08-a8d6-891c5d7853b2
State: REGISTERED
Multiplexed: ✔ yes
Sample Number: 8
Workflow: —
Studies: Cut&Tag_optimization_wing_discs, nippedB_RNA
Run Directory: /g/furlong/STOCKS/Data/Assay/sequencing/2022/2022-06-02-HNHCMAFX3
Description: —

Run Parameters

Run ID: 220531_NB501764_1566_AHNHCMAFX3
Instrument: NextSeq500 Waffles
Flowcell: HNHCMAFX3
Flowcell Version: —
Lane: 1
Run Type: Paired-End
Run Mode: 75-8-8-75
Read Length: —
Chemistry: —

Sharing

Ownership and Lifecycle

Output Datasets

1 - 8 of 8

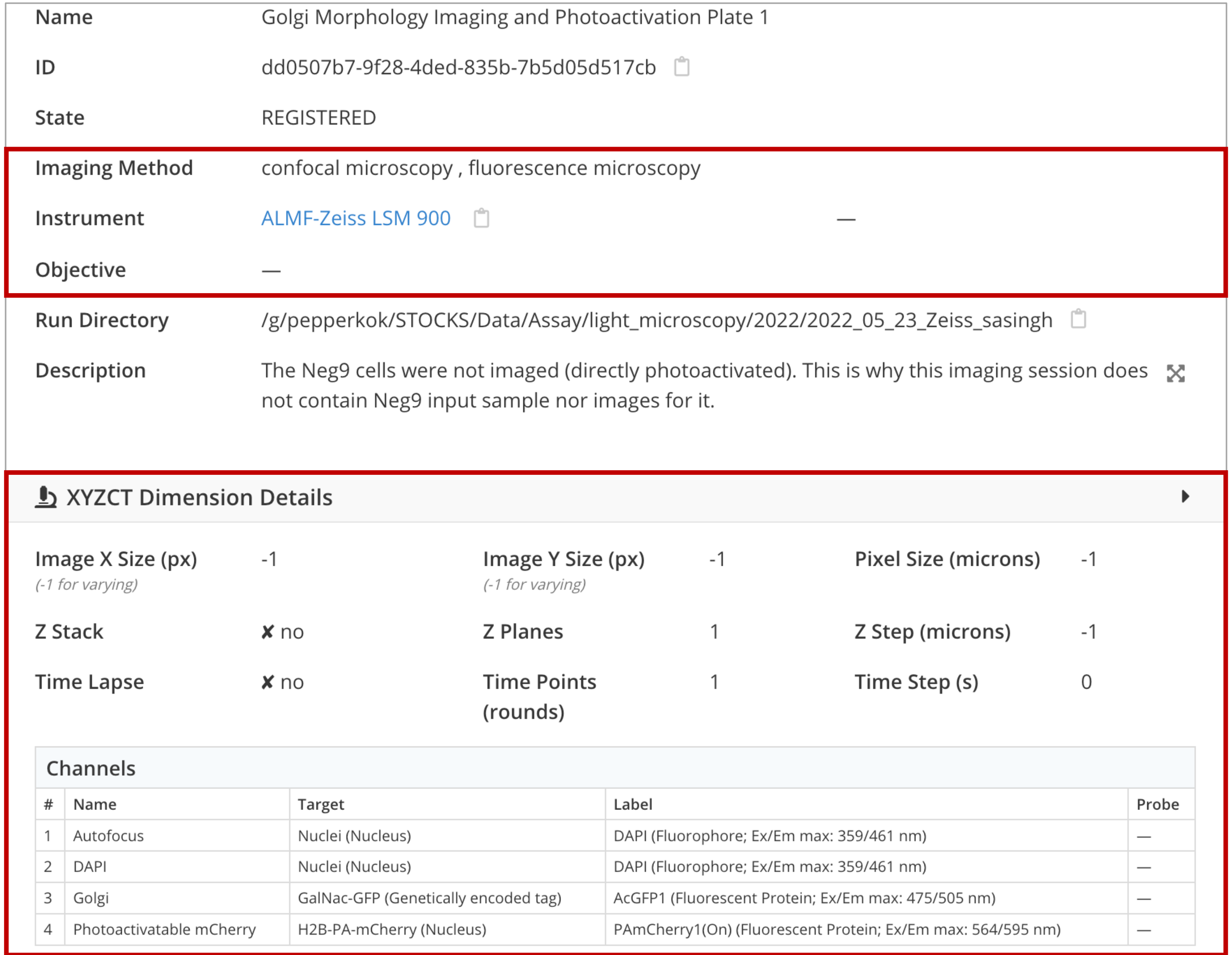
☐	Name	Description	Type	QC	Samples	Assay	Studies	Datafiles	Dataset Collections	Workflow	Archives
☐	Chipnewrep1_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipnewrep2_HNHCMAFX3_1		paired-end fastq	—	Chipnewrep2	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			
☐	Chipoldrep1_HNHCMAFX3_1		paired-end fastq	—	Chipoldrep1	HNHCMAFX3_lane1	Cut&Tag_optimization_wing_discs	view 2 items			

The Light Microscopy Assay Model

Pick a method, a microscope and an optional objective

Key values for the X, Y, Z and T dimensions

Channels description (1+)



Name	Golgi Morphology Imaging and Photoactivation Plate 1
ID	dd0507b7-9f28-4ded-835b-7b5d05d517cb
State	REGISTERED
Imaging Method	confocal microscopy , fluorescence microscopy
Instrument	ALMF-Zeiss LSM 900 —
Objective	—
Run Directory	/g/pepperkok/STOCKS/Data/Assay/light_microscopy/2022/2022_05_23_Zeiss_sasingh
Description	The Neg9 cells were not imaged (directly photoactivated). This is why this imaging session does not contain Neg9 input sample nor images for it.

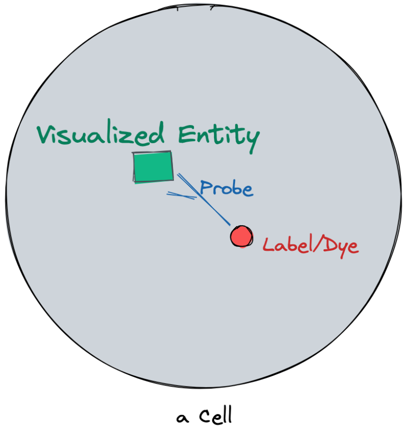
XYZCT Dimension Details

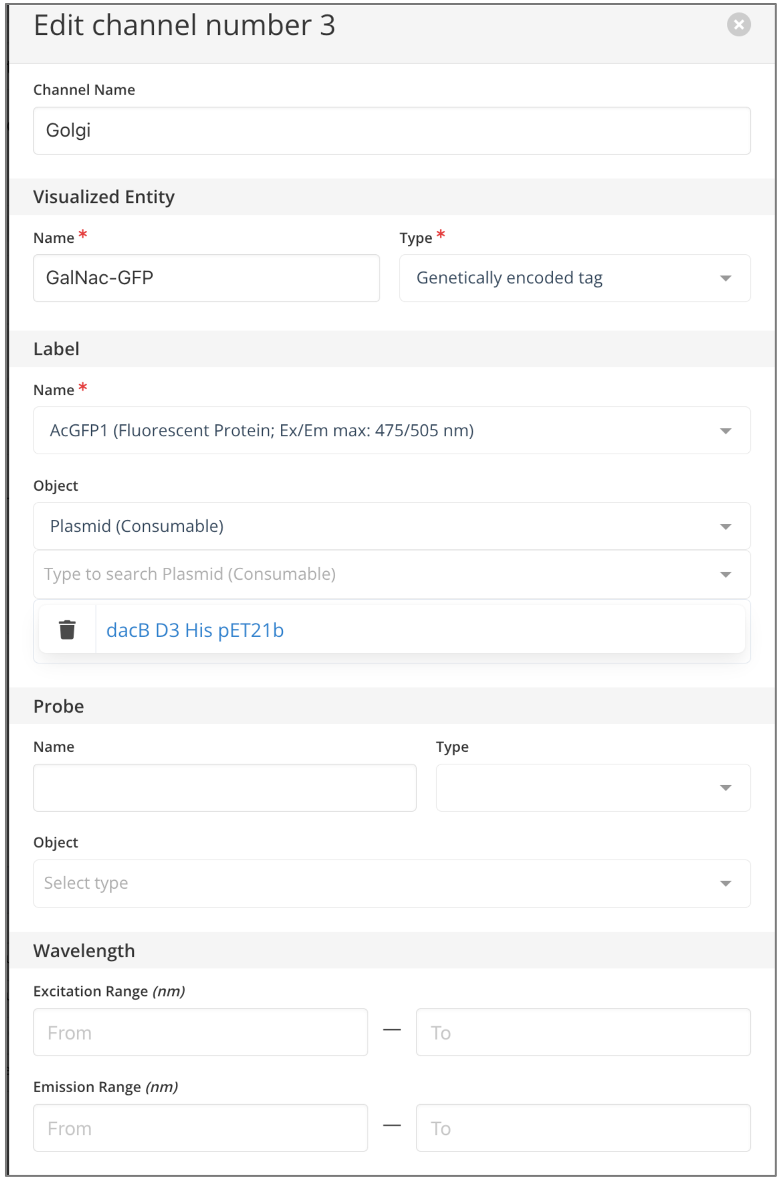
Image X Size (px) *(-1 for varying)*	-1	Image Y Size (px) *(-1 for varying)*	-1	Pixel Size (microns)	-1
Z Stack	✘ no	Z Planes	1	Z Step (microns)	-1
Time Lapse	✘ no	Time Points (rounds)	1	Time Step (s)	0

Channels

#	Name	Target	Label	Probe
1	Autofocus	Nuclei (Nucleus)	DAPI (Fluorophore; Ex/Em max: 359/461 nm)	—
2	DAPI	Nuclei (Nucleus)	DAPI (Fluorophore; Ex/Em max: 359/461 nm)	—
3	Golgi	GalNac-GFP (Genetically encoded tag)	AcGFP1 (Fluorescent Protein; Ex/Em max: 475/505 nm)	—
4	Photoactivatable mCherry	H2B-PA-mCherry (Nucleus)	PAmCherry1(On) (Fluorescent Protein; Ex/Em max: 564/595 nm)	—

The Light Microscopy Assay Model – Channel description





Edit channel number 3

Channel Name: Golgi

Visualized Entity

Name *: GalNac-GFP

Type *: Genetically encoded tag

Label

Name *: AcGFP1 (Fluorescent Protein; Ex/Em max: 475/505 nm)

Object: Plasmid (Consumable)

Type to search Plasmid (Consumable)

dacB D3 His pET21b

Probe

Name:

Type:

Object: Select type

Wavelength

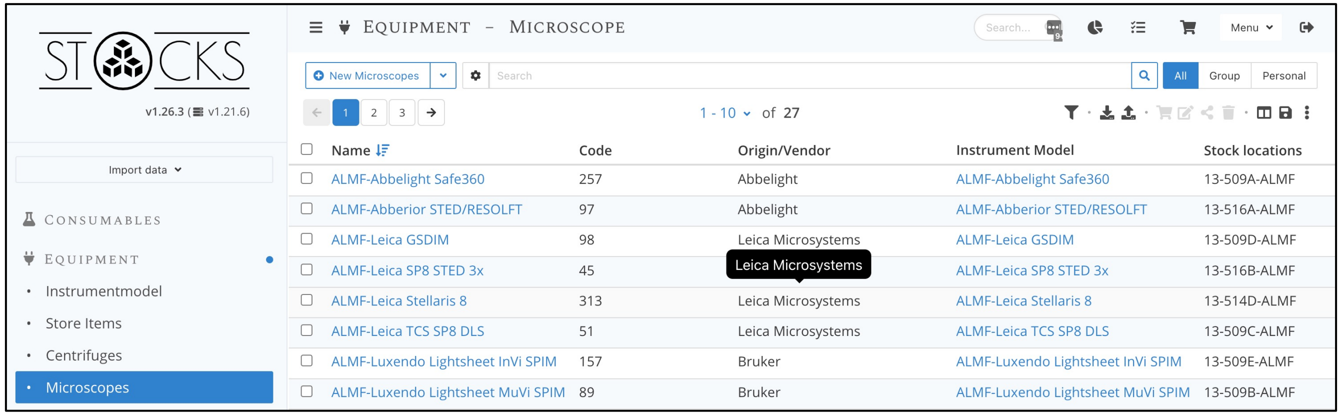
Excitation Range *(nm)*: From — To

Emission Range *(nm)*: From — To

- Hundreds of Labels/Dyes imported from
  - Fluorescent proteins https://www.fpbase.org/
  - Fluorophores https://www.semrock.com/
- Also Designed with Flow Cytometry in mind
- Exact item used for the label & probe can be linked from the consumable collections

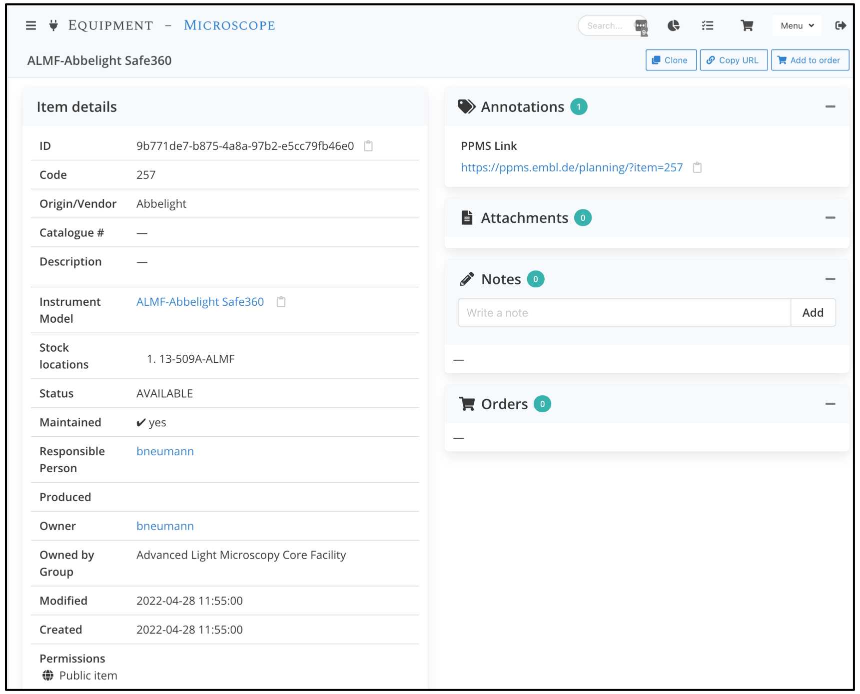
The Light Microscopy Assay Model – Microscopes

- All ALMF Microscopes available
- Groups can add their own microscopes



☐	Name	Code	Origin/Vendor	Instrument Model	Stock locations
☐	ALMF-Abbelight Safe360	257	Abbelight	ALMF-Abbelight Safe360	13-509A-ALMF
☐	ALMF-Abberior STED/RESOLFT	97	Abbelight	ALMF-Abberior STED/RESOLFT	13-516A-ALMF
☐	ALMF-Leica GSDIM	98	Leica Microsystems	ALMF-Leica GSDIM	13-509D-ALMF
☐	ALMF-Leica SP8 STED 3x	45	Leica Microsystems	ALMF-Leica SP8 STED 3x	13-516B-ALMF
☐	ALMF-Leica Stellaris 8	313	Leica Microsystems	ALMF-Leica Stellaris 8	13-514D-ALMF
☐	ALMF-Leica TCS SP8 DLS	51	Leica Microsystems	ALMF-Leica TCS SP8 DLS	13-509C-ALMF
☐	ALMF-Luxendo Lightsheet InVi SPIM	157	Bruker	ALMF-Luxendo Lightsheet InVi SPIM	13-509E-ALMF
☐	ALMF-Luxendo Lightsheet MuVi SPIM	89	Bruker	ALMF-Luxendo Lightsheet MuVi SPIM	13-509B-ALMF

- Code == PPMS ID
- Link to PPMS booking
- Responsible Person
- Room Location



ALMF-Abbelight Safe360

Item details

ID	9b771de7-b875-4a8a-97b2-e5cc79fb46e0
Code	257
Origin/Vendor	Abbelight
Catalogue #	—
Description	—
Instrument Model	ALMF-Abbelight Safe360
Stock locations	1. 13-509A-ALMF
Status	AVAILABLE
Maintained	✔ yes
Responsible Person	bneumann
Produced	
Owner	bneumann
Owned by Group	Advanced Light Microscopy Core Facility
Modified	2022-04-28 11:55:00
Created	2022-04-28 11:55:00

Permissions
Public item

Annotations 1

PPMS Link
https://ppms.embl.de/planning/?item=257

Attachments 0

Notes 0

Write a note Add

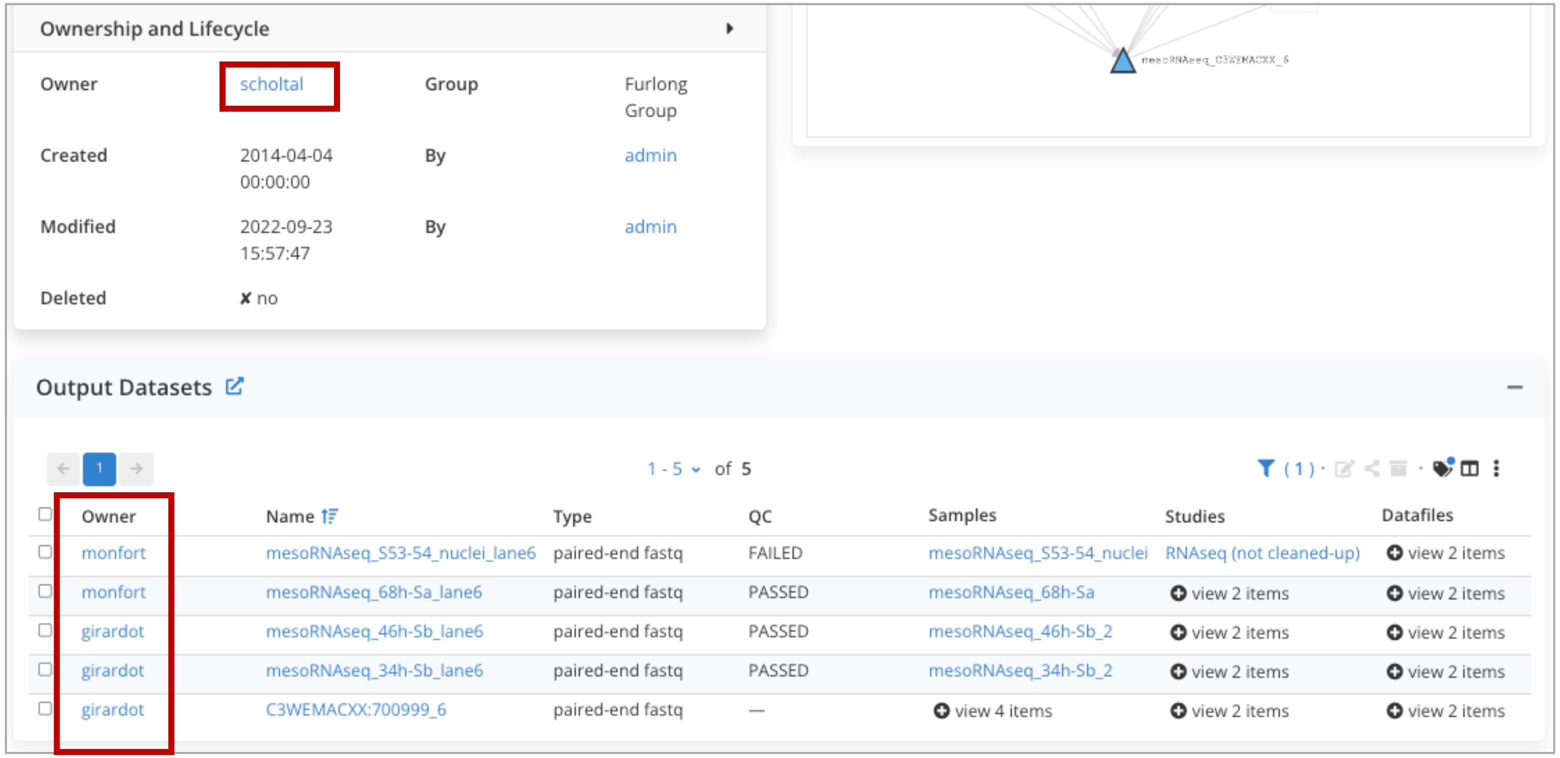
—

Orders 0

—

Multiplexed assays may contain datasets that belong to different users and projects

- An assay does not necessarily contain related datasets
- The assay, the produced and consumed samples can have different owners ; e.g.
  - The lab technicians performed the assay => they own the assay
  - Datasets (and samples) are owned by different researchers
- Related datasets, produced by one or more assays, are grouped into a **Study**



Ownership and Lifecycle

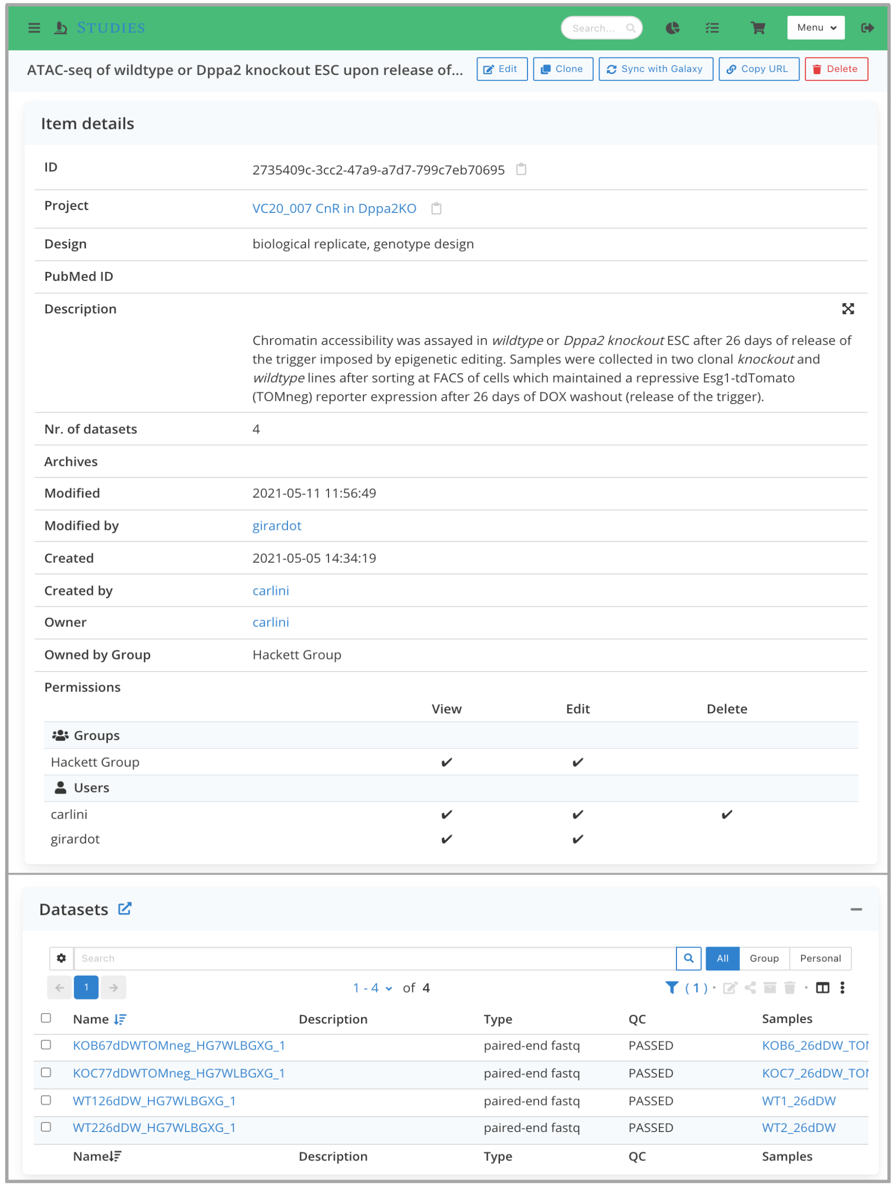
Owner	scholtal	Group	Furlong Group
Created	2014-04-04 00:00:00	By	admin
Modified	2022-09-23 15:57:47	By	admin
Deleted	✗ no		

Output Datasets

1 - 5 of 5

☐	Owner	Name	Type	QC	Samples	Studies	Datafiles
☐	monfort	mesoRNAseq_S53-54_nuclei_lane6	paired-end fastq	FAILED	mesoRNAseq_S53-54_nuclei	RNAseq (not cleaned-up)	view 2 items
☐	monfort	mesoRNAseq_68h-Sa_lane6	paired-end fastq	PASSED	mesoRNAseq_68h-Sa	view 2 items	view 2 items
☐	girardot	mesoRNAseq_46h-Sb_lane6	paired-end fastq	PASSED	mesoRNAseq_46h-Sb_2	view 2 items	view 2 items
☐	girardot	mesoRNAseq_34h-Sb_lane6	paired-end fastq	PASSED	mesoRNAseq_34h-Sb_2	view 2 items	view 2 items
☐	girardot	C3WEMACXX:700999_6	paired-end fastq	—	view 4 items	view 2 items	view 2 items

Studies are at the basis of data publication to public repositories



ATAC-seq of wildtype or Dppa2 knockout ESC upon release of...

Item details

ID	2735409c-3cc2-47a9-a7d7-799c7eb70695
Project	VC20_007 CnR in Dppa2KO
Design	biological replicate, genotype design
PubMed ID	
Description	Chromatin accessibility was assayed in *wildtype* or *Dppa2 knockout* ESC after 26 days of release of the trigger imposed by epigenetic editing. Samples were collected in two clonal *knockout* and *wildtype* lines after sorting at FACS of cells which maintained a repressive Esg1-tdTomato (TOMneg) reporter expression after 26 days of DOX washout (release of the trigger).
Nr. of datasets	4
Archives	
Modified	2021-05-11 11:56:49
Modified by	girardot
Created	2021-05-05 14:34:19
Created by	carlini
Owner	carlini
Owned by Group	Hackett Group

Permissions

	View	Edit	Delete
Groups			
Hackett Group	✔	✔	
Users			
carlini	✔	✔	✔
girardot	✔	✔	

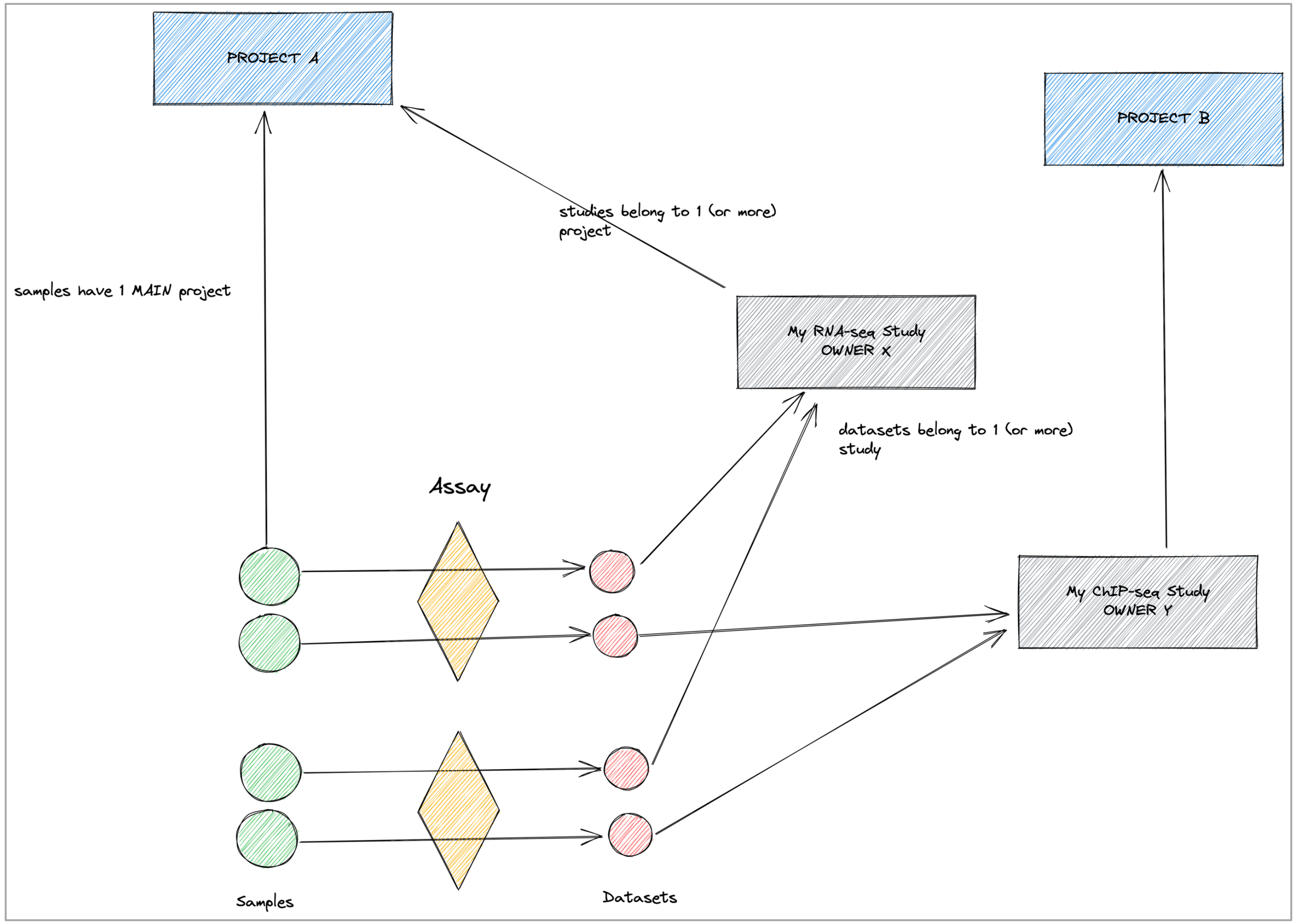
Datasets

☐	Name	Description	Type	QC	Samples
☐	KOB67dDWTOMneg_HG7WLBGXG_1		paired-end fastq	PASSED	KOB6_26dDW_TO
☐	KOC77dDWTOMneg_HG7WLBGXG_1		paired-end fastq	PASSED	KOC7_26dDW_TO
☐	WT126dDW_HG7WLBGXG_1		paired-end fastq	PASSED	WT1_26dDW
☐	WT226dDW_HG7WLBGXG_1		paired-end fastq	PASSED	WT2_26dDW
	Name	Description	Type	QC	Samples

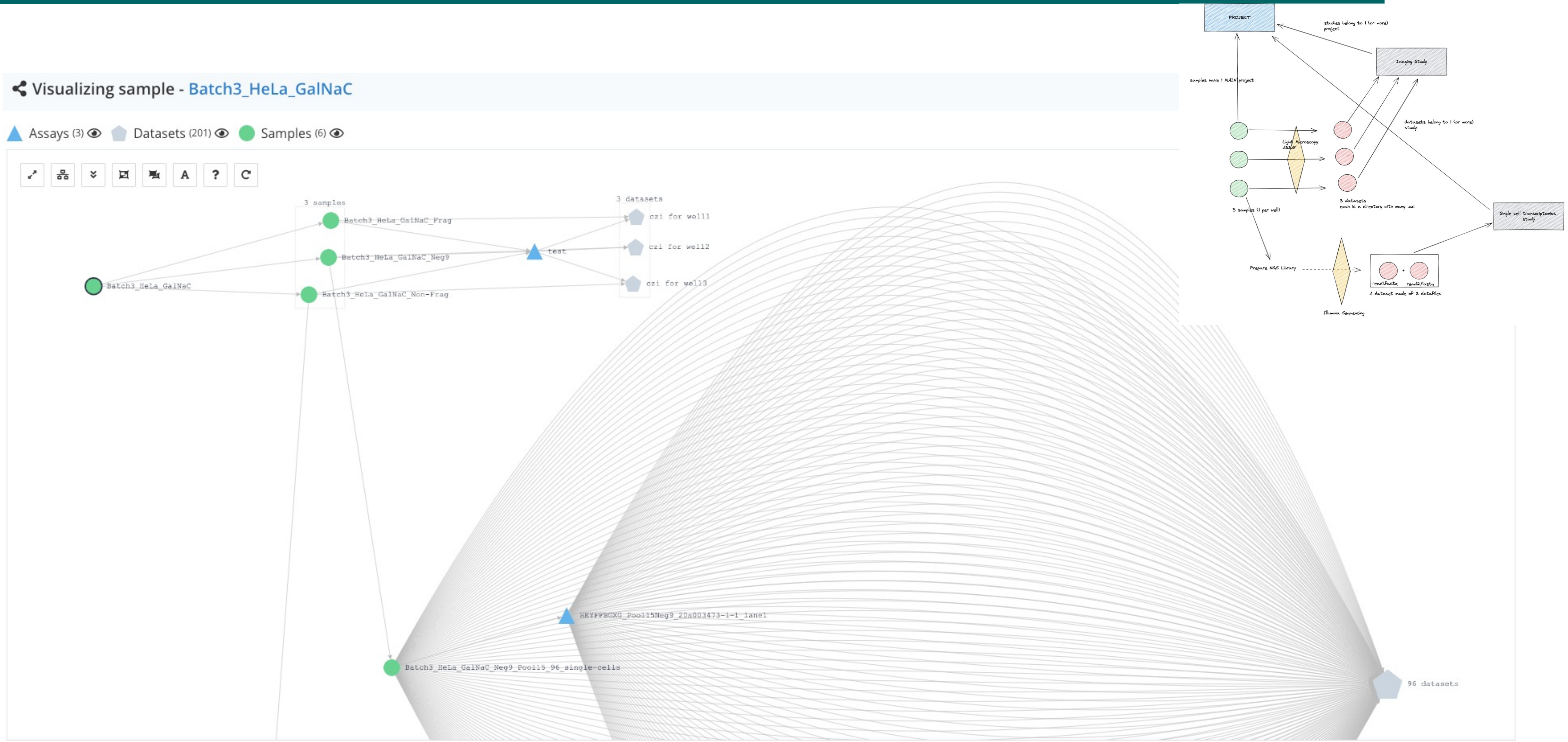
- Related raw data must accumulate in a *study*
  - Produced at different times, in different runs
  - e.g. Chip-seq study
- Belongs to one project
- Datasets can belong to <u>multiple</u> studies
  - E.g. when re-used
- The Study is what gets submitted to data repository
- Your paper may have multiple studies of different data types
  - Different sequencing flavors e.g. rna-seq, chip-seq
  - Different technologies e.g. imaging, sequencing

Assay vs Study vs Project

- The **project** regroups different items i.e. samples & datasets (more in the future)
- The **study** regroups related datasets of the same experiment type, those are analyzed together
- The **assay** is the run of one instrument and regroup the produced datasets (==same run metadata)

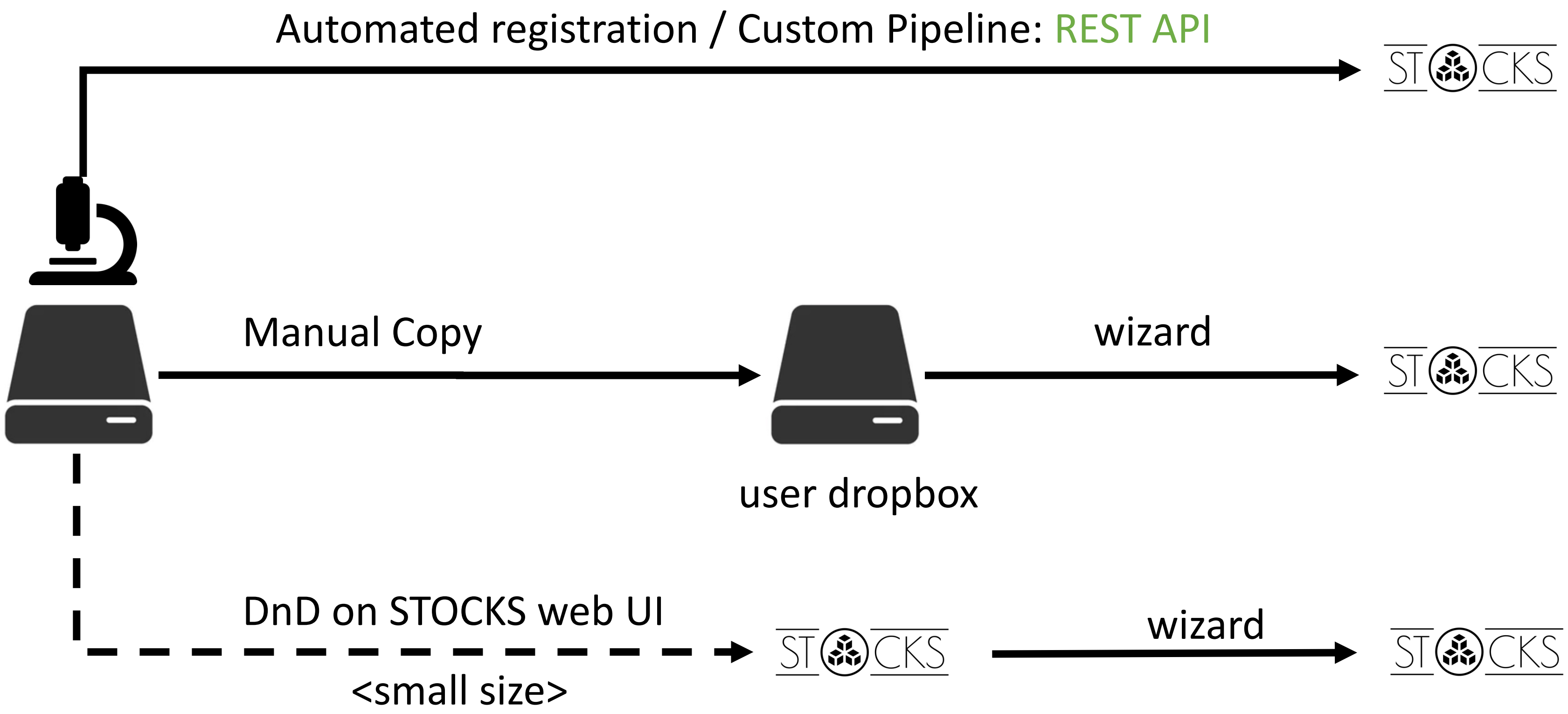


A cross-technology sequencing/imaging “multi-omics” project



Visualizing sample - Batch3_HeLa_GalNaC

Data Ingestion Scenarios



Automated registration / Custom Pipeline: REST API

Manual Copy

wizard

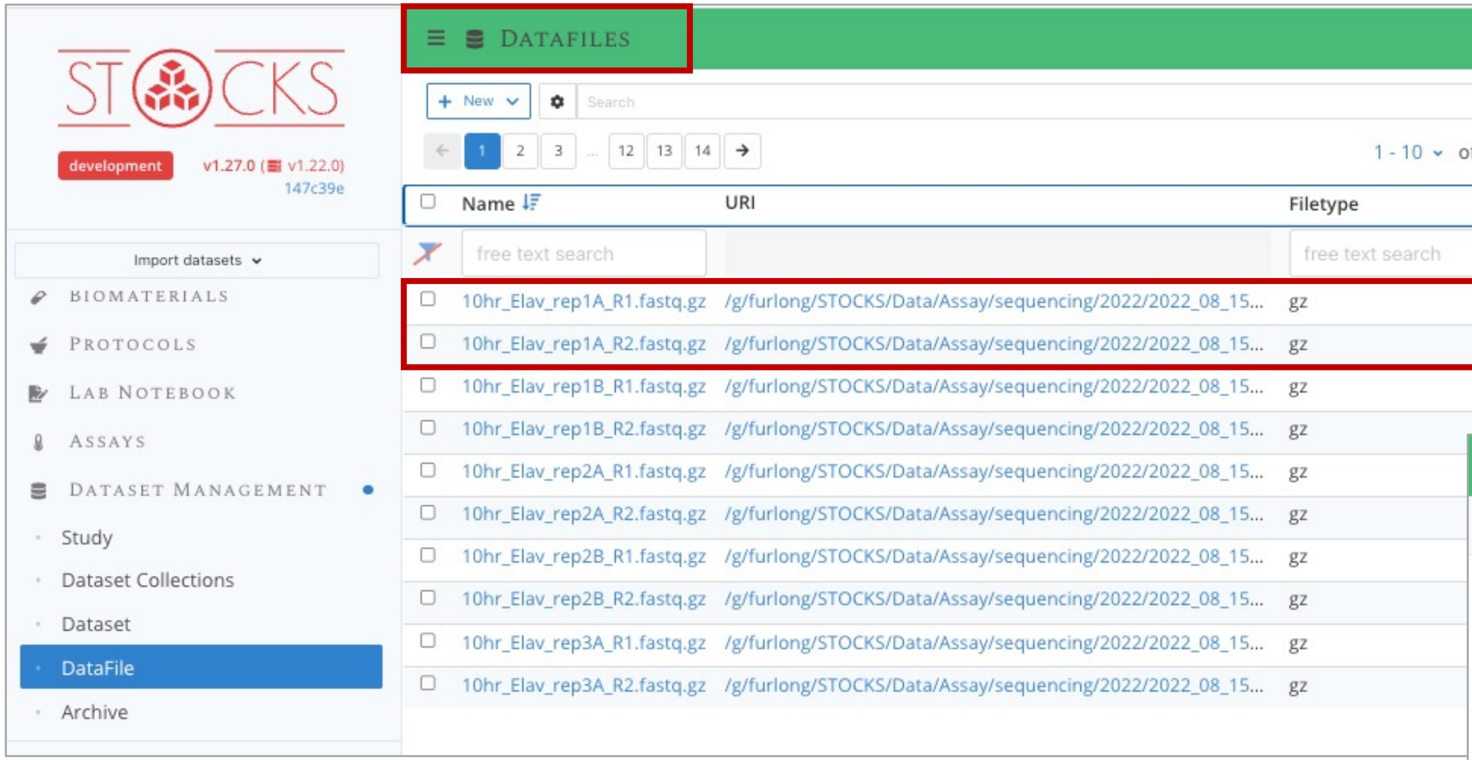
user dropbox

DnD on STOCKS web UI

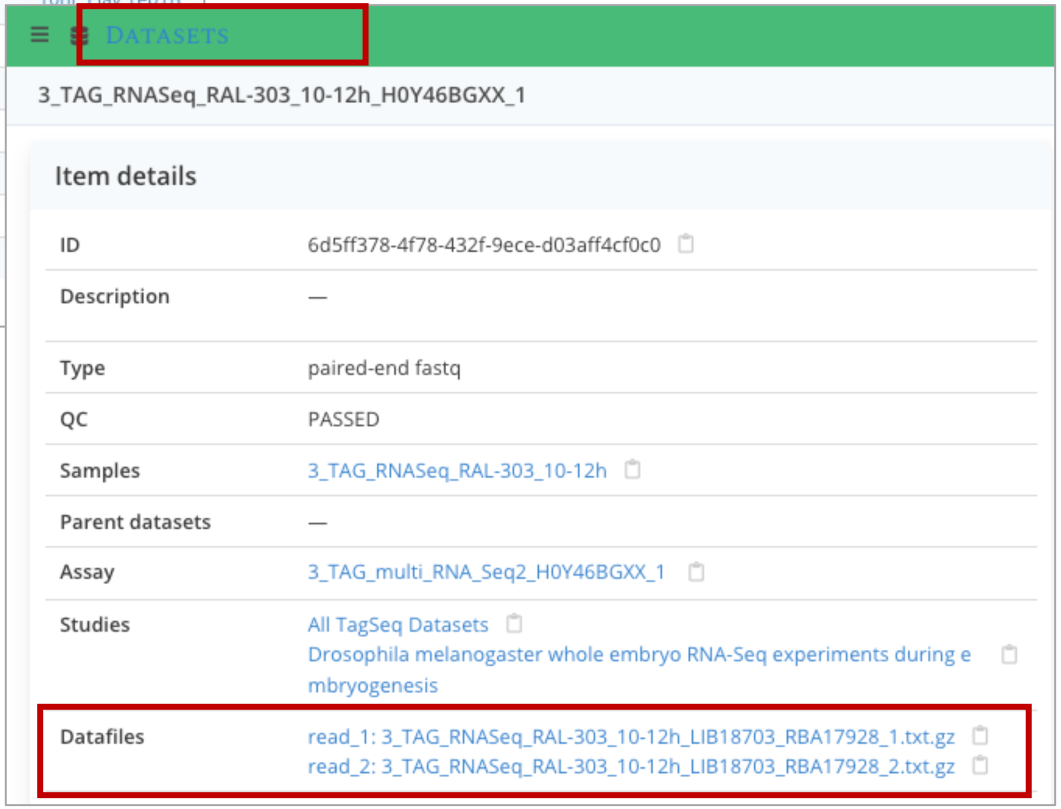
<small size>

wizard

Datasets are made of one or more Datafiles



☐	Name	URI	Filetype	Datasets
☐	10hr_Elav_rep1A_R1.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	10hr_Elav_rep1A
☐	10hr_Elav_rep1A_R2.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	10hr_Elav_rep1A
☐	10hr_Elav_rep1B_R1.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	10hr_Elav_rep1B
☐	10hr_Elav_rep1B_R2.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	10hr_Elav_rep1B
☐	10hr_Elav_rep2A_R1.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	
☐	10hr_Elav_rep2A_R2.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	
☐	10hr_Elav_rep2B_R1.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	
☐	10hr_Elav_rep2B_R2.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	
☐	10hr_Elav_rep3A_R1.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	
☐	10hr_Elav_rep3A_R2.fastq.gz	/g/furlong/STOCKS/Data/Assay/sequencing/2022/2022_08_15...	gz	



3_TAG_RNASeq_RAL-303_10-12h_H0Y46BGXX_1

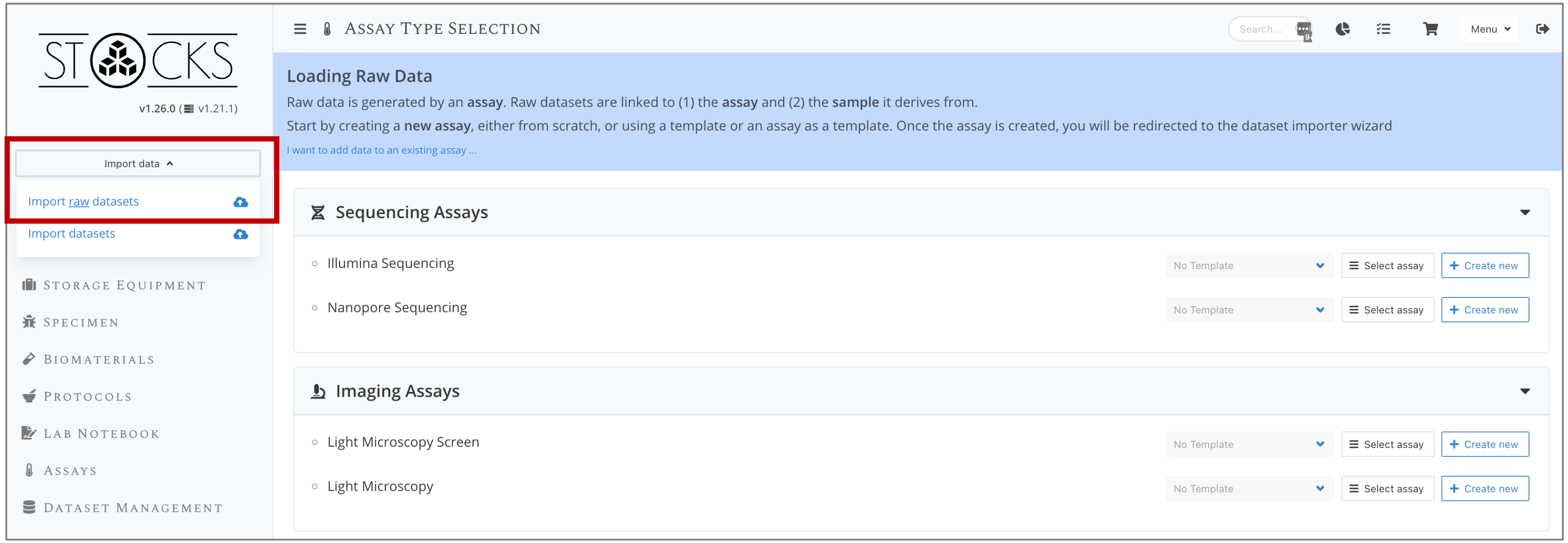
Item details

ID	6d5ff378-4f78-432f-9ece-d03aff4cf0c0
Description	—
Type	paired-end fastq
QC	PASSED
Samples	3_TAG_RNASeq_RAL-303_10-12h
Parent datasets	—
Assay	3_TAG_multi_RNA_Seq2_H0Y46BGXX_1
Studies	All TagSeq Datasets Drosophila melanogaster whole embryo RNA-Seq experiments during embryogenesis
Datafiles	read_1: 3_TAG_RNASeq_RAL-303_10-12h_LIB18703_RBA17928_1.txt.gz read_2: 3_TAG_RNASeq_RAL-303_10-12h_LIB18703_RBA17928_2.txt.gz

- One Paired-End FASTQ dataset is composed of 2 FASTQ *datafiles*
- Datasets can have as many datafiles as needed
- A dataset should be regarded as an atomic entity
  - i.e. considering a unique datafile would not make sense
- You always manipulate datasets in STOCKS

Flexible assay data import with the (raw) Dataset Loader

> Uploading **Raw data** *implies* to described **the Assay (e.g. Light Microscopy)** and **the parent sample(s)**



Loading Raw Data

Raw data is generated by an **assay**. Raw datasets are linked to (1) the **assay** and (2) the **sample** it derives from.

Start by creating a **new assay**, either from scratch, or using a template or an assay as a template. Once the assay is created, you will be redirected to the dataset importer wizard

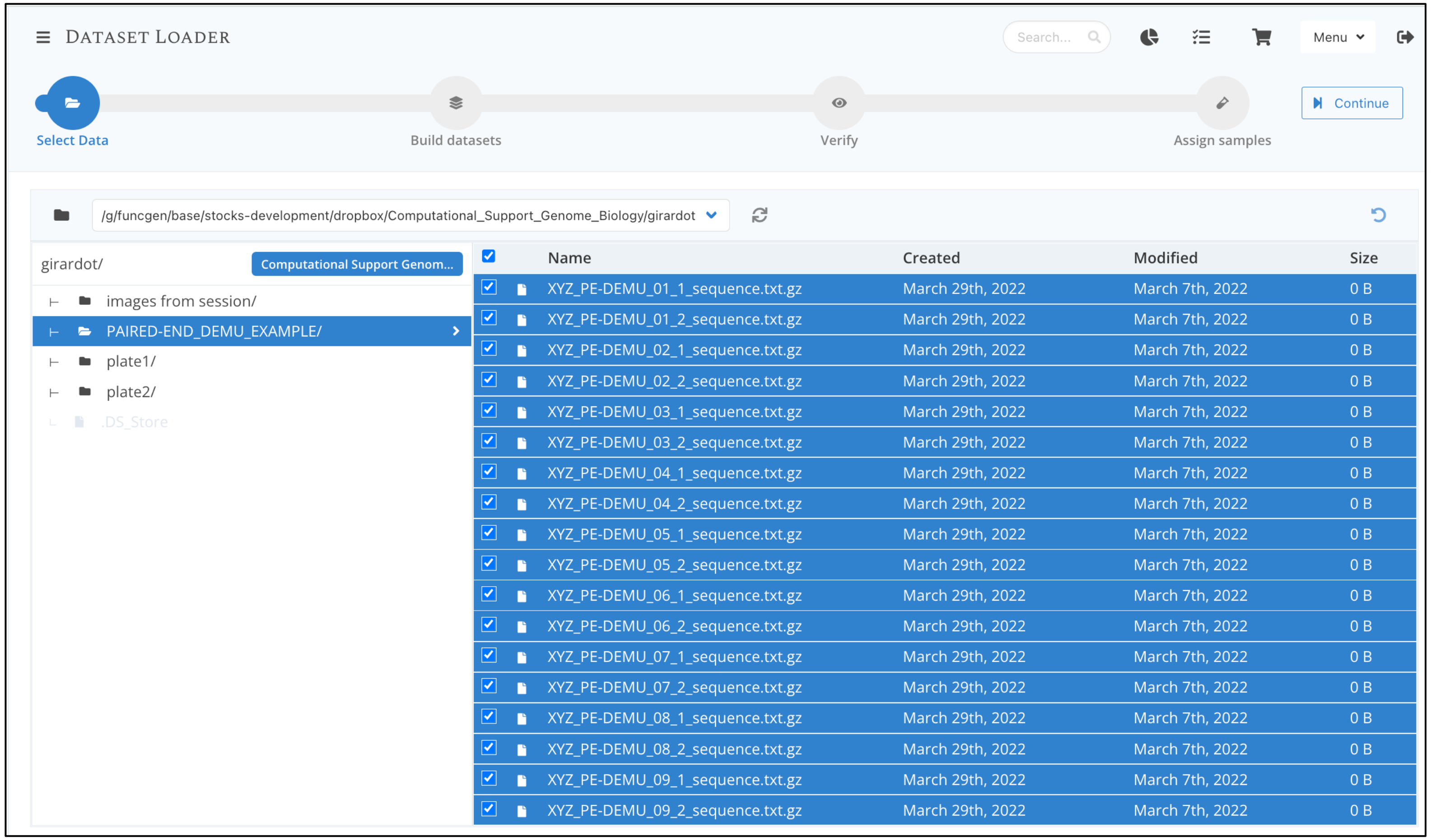
I want to add data to an existing assay ...

- Templates for frequent setups
- Re-use **your own assay as template**

=> The **Dataset Loader Wizard** drives the user through this process

Data must be first organized into a “run folder”

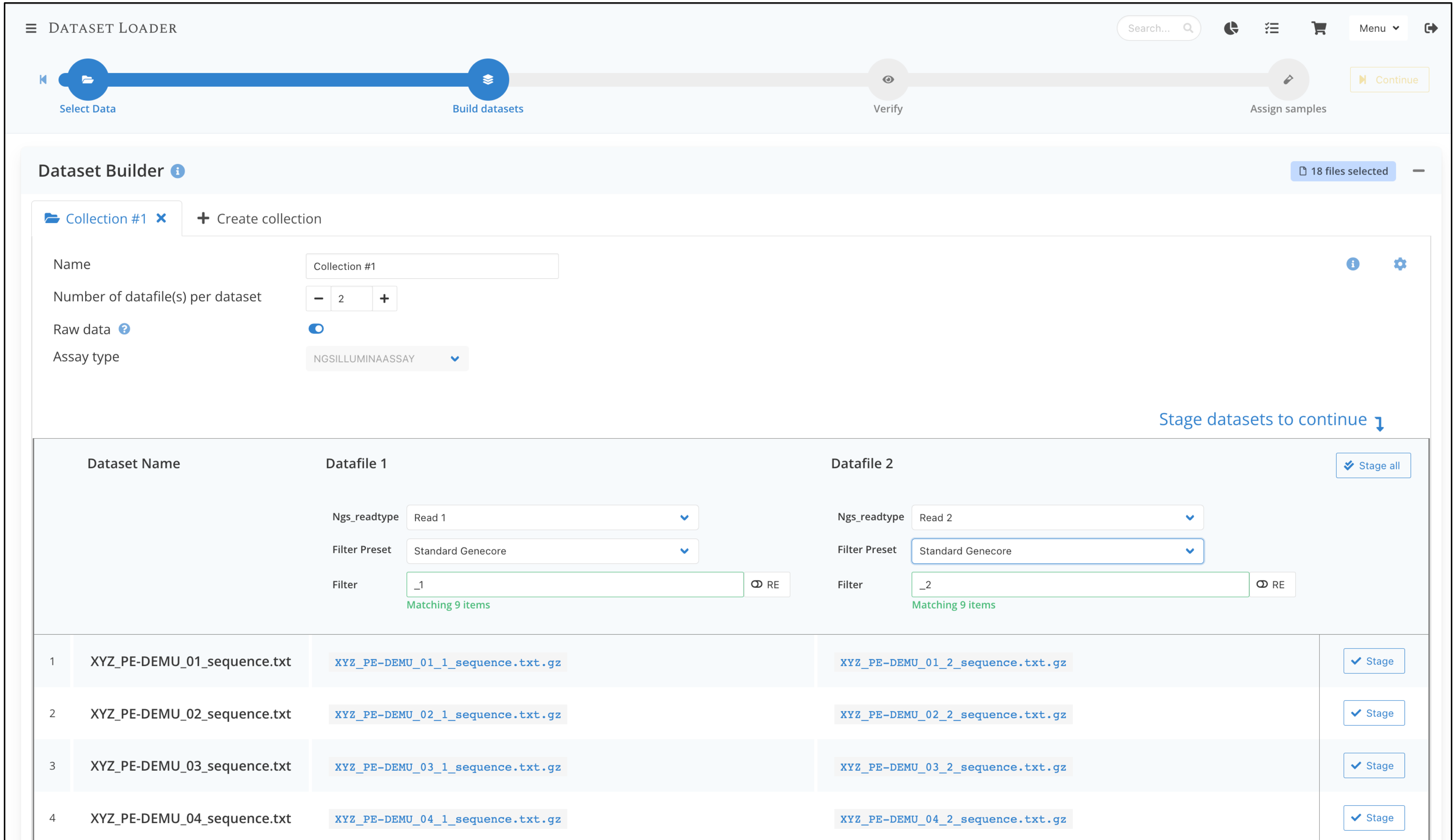
- One folder == One assay
- Data (files, folders) must be in your user “dropbox” organized in a logical way
- Group data per assay as data for a unique assay can be loaded at once
- Individual files or complete folders can be selected as “data files”



☑	Name	Created	Modified	Size
☑	XYZ_PE-DEMU_01_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_01_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_02_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_02_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_03_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_03_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_04_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_04_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_05_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_05_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_06_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_06_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_07_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_07_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_08_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_08_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_09_1_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B
☑	XYZ_PE-DEMU_09_2_sequence.txt.gz	March 29th, 2022	March 7th, 2022	0 B

Multi-file datasets can be assembled in the Dataset Builder

- Build multi-file datasets with the Dataset Builder
- Datasets can be organized in different dataset collections



Dataset Builder

Name: Collection #1

Number of datafile(s) per dataset: 2

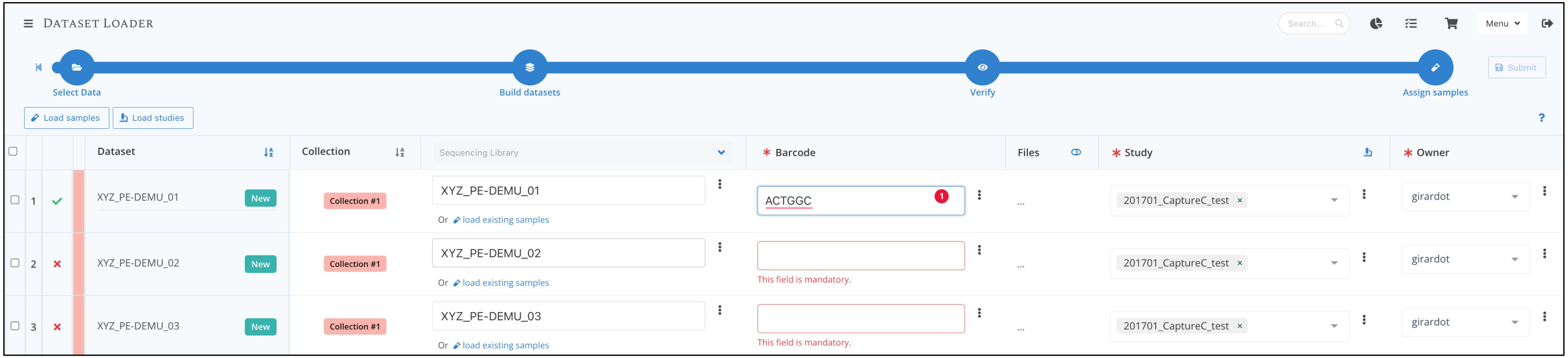
Raw data

Assay type: NGSILLUMINAASSAY

Dataset Name	Datafile 1	Datafile 2	
	Ngs_readtype: Read 1 Filter Preset: Standard Genecore Filter: _1 RE Matching 9 items	Ngs_readtype: Read 2 Filter Preset: Standard Genecore Filter: _2 RE Matching 9 items	
1 XYZ_PE-DEMU_01_sequence.txt	`XYZ_PE-DEMU_01_1_sequence.txt.gz`	`XYZ_PE-DEMU_01_2_sequence.txt.gz`	✓ Stage
2 XYZ_PE-DEMU_02_sequence.txt	`XYZ_PE-DEMU_02_1_sequence.txt.gz`	`XYZ_PE-DEMU_02_2_sequence.txt.gz`	✓ Stage
3 XYZ_PE-DEMU_03_sequence.txt	`XYZ_PE-DEMU_03_1_sequence.txt.gz`	`XYZ_PE-DEMU_03_2_sequence.txt.gz`	✓ Stage
4 XYZ_PE-DEMU_04_sequence.txt	`XYZ_PE-DEMU_04_1_sequence.txt.gz`	`XYZ_PE-DEMU_04_2_sequence.txt.gz`	✓ Stage

Raw Datasets must be linked to Samples

- Link Datasets to new or existing samples, associate to study & owner

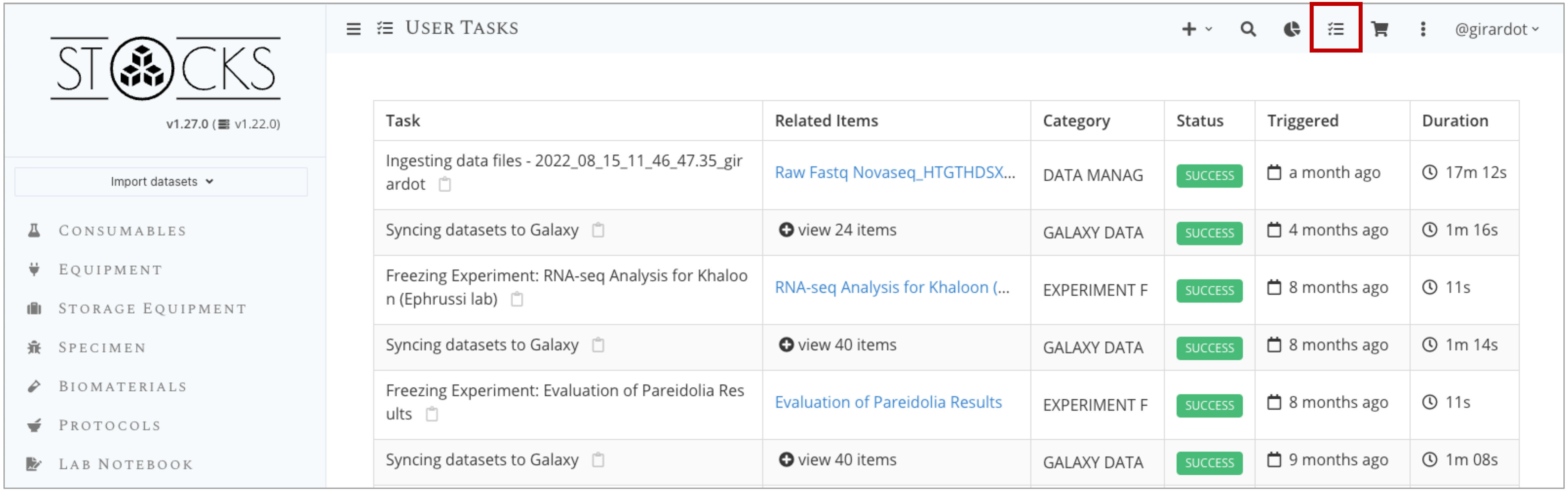


☐			Dataset	Collection	Sequencing Library	* Barcode	Files	* Study	* Owner
☐	1	✓	XYZ_PE-DEMU_01 New	Collection #1	XYZ_PE-DEMU_01 Or load existing samples	ACTGGC 1	...	201701_CaptureC_test ×	girardot
☐	2	×	XYZ_PE-DEMU_02 New	Collection #1	XYZ_PE-DEMU_02 Or load existing samples	This field is mandatory.	...	201701_CaptureC_test ×	girardot
☐	3	×	XYZ_PE-DEMU_03 New	Collection #1	XYZ_PE-DEMU_03 Or load existing samples	This field is mandatory.	...	201701_CaptureC_test ×	girardot

- Once completed, the files and folders must be imported into STOCKS
  - Async process (can take quite some time)
  - A job is created
  - You get a notification once ready

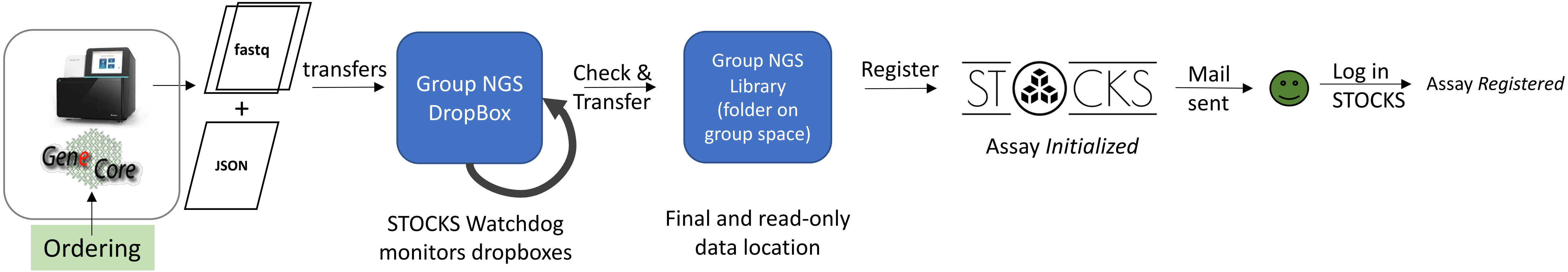
Check the status of your async tasks

- Different tasks occur in an asynchronous fashion



Task	Related Items	Category	Status	Triggered	Duration
Ingesting data files - 2022_08_15_11_46_47.35_girardot	Raw Fastq Novaseq_HTGTHDSX...	DATA MANAG	SUCCESS	a month ago	17m 12s
Syncing datasets to Galaxy	view 24 items	GALAXY DATA	SUCCESS	4 months ago	1m 16s
Freezing Experiment: RNA-seq Analysis for Khaloon (Ephrussi lab)	RNA-seq Analysis for Khaloon (...	EXPERIMENT F	SUCCESS	8 months ago	11s
Syncing datasets to Galaxy	view 40 items	GALAXY DATA	SUCCESS	8 months ago	1m 14s
Freezing Experiment: Evaluation of Pareidolia Results	Evaluation of Pareidolia Results	EXPERIMENT F	SUCCESS	8 months ago	11s
Syncing datasets to Galaxy	view 40 items	GALAXY DATA	SUCCESS	9 months ago	1m 08s

Automated data ingestion using the STOCKS API



Ordering

transfers

Group NGS DropBox

STOCKS Watchdog monitors dropboxes

Check & Transfer

Group NGS Library (folder on group space)

Final and read-only data location

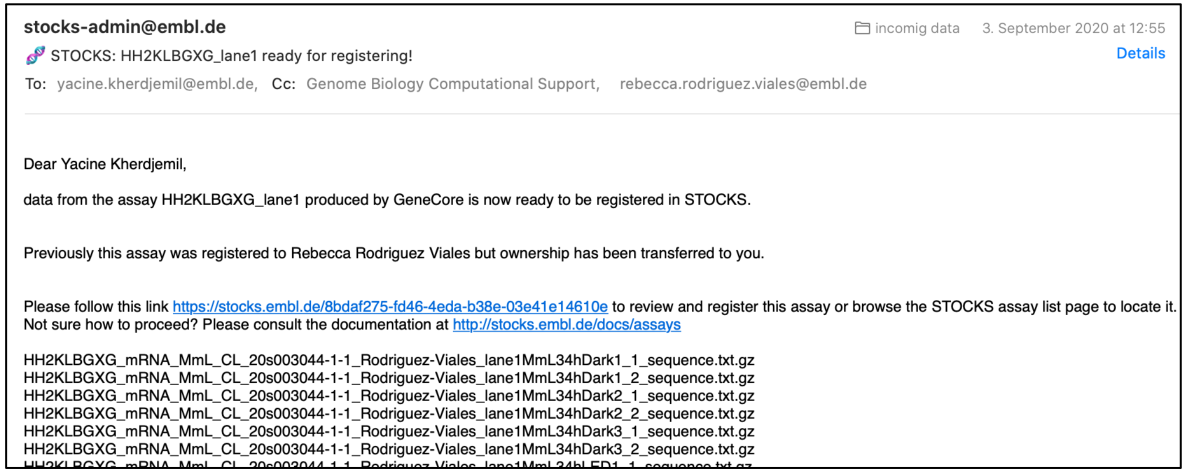
Register

Assay *Initialized*

Mail sent

Log in STOCKS

Assay *Registered*



stocks-admin@embl.de
incomig data 3. September 2020 at 12:55
STOCKS: HH2KLBGXG_lane1 ready for registering!
Details
To: yacine.kherdjemil@embl.de, Cc: Genome Biology Computational Support, rebecca.rodriguez.viales@embl.de

Dear Yacine Kherdjemil,

data from the assay HH2KLBGXG_lane1 produced by GeneCore is now ready to be registered in STOCKS.

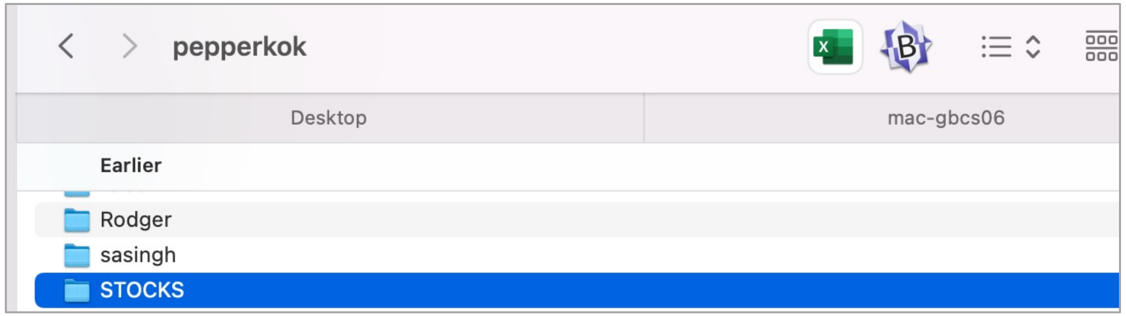
Previously this assay was registered to Rebecca Rodriguez Viales but ownership has been transferred to you.

Please follow this link https://stocks.embl.de/8bdaf275-fd46-4eda-b38e-03e41e14610e to review and register this assay or browse the STOCKS assay list page to locate it.
Not sure how to proceed? Please consult the documentation at http://stocks.embl.de/docs/assays

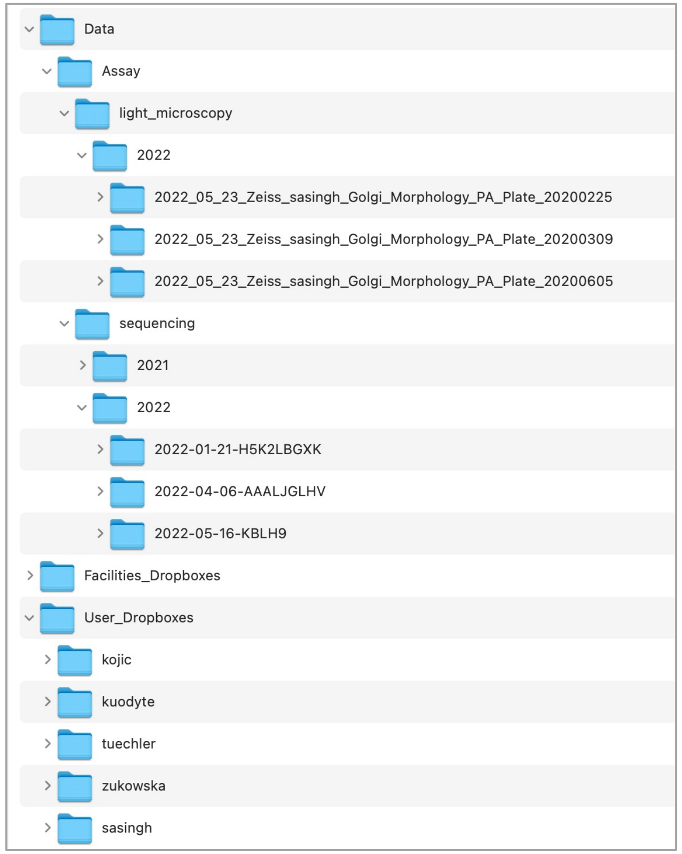
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark1_1_sequence.txt.gz
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark1_2_sequence.txt.gz
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark2_1_sequence.txt.gz
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark2_2_sequence.txt.gz
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark3_1_sequence.txt.gz
HH2KLBGXG_mRNA_MmL_CL_20s003044-1-1_Rodriguez-Viales_lane1MmL34hDark3_2_sequence.txt.gz

- No more data loss
- Information available at GeneCore is passed on
- Enforce to connect datasets with samples
- Enforce to collect minimum information
- User never manipulate the files
- Files remain accessible (read-only) for analysis
- Model to be extended to other Core Facilities

Files remain accessible but are managed by STOCKS



- Rodger
- sasingh
- STOCKS



- Data
  - Assay
    - light_microscopy
      - 2022
        - 2022_05_23_Zeiss_sasingh_Golgi_Morphology_PA_Plate_20200225
        - 2022_05_23_Zeiss_sasingh_Golgi_Morphology_PA_Plate_20200309
        - 2022_05_23_Zeiss_sasingh_Golgi_Morphology_PA_Plate_20200605
    - sequencing
      - 2021
      - 2022
        - 2022-01-21-H5K2LBGXK
        - 2022-04-06-AAALJGLHV
        - 2022-05-16-KBLH9
- Facilities_Dropboxes
- User_Dropboxes
  - kojic
  - kuodyte
  - tuechler
  - zukowska
  - sasingh

One STOCKS repo per group

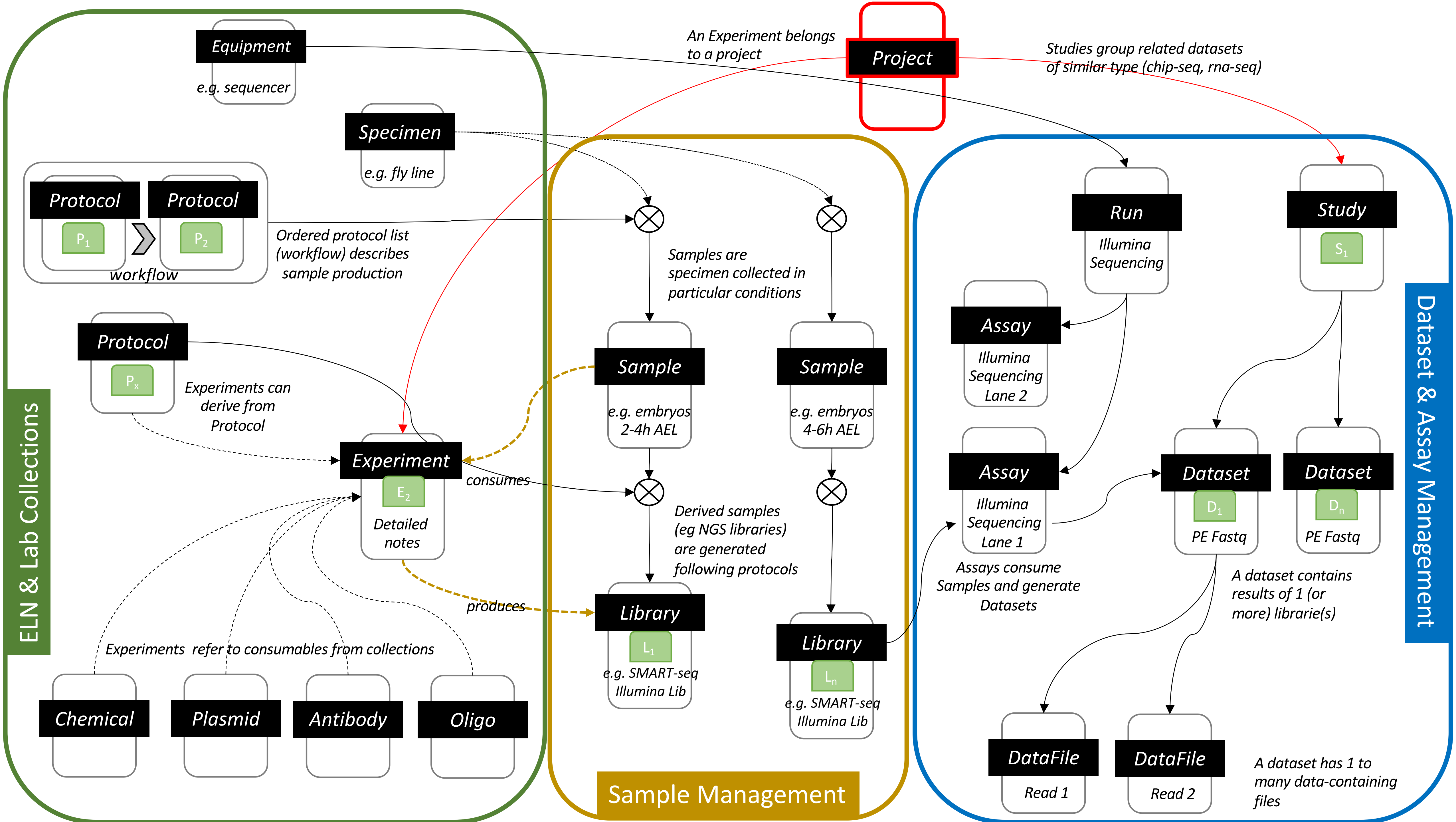
- Read only to ensure traceability
- Each group pays for its storage
- Unix group sticky bit: group can read by default

Intuitive organization

User can import data from their “dropbox”

Dataset deletion rules differ between managed and unmanaged assays

- Automatically loaded assays are managed
  - They are automatically loaded from a trusted source
  - They must be complete (no post raw dataset addition) at load time
    - Uncomplete assays must not be loaded => get to your admin
  - They cannot be deleted by users
    - Admin can unload wrong assays
  - Raw datasets cannot be deleted unless a backup copy exist
- Manually loaded datasets and associated assays are unmanaged
  - User can delete datasets and assays as needed
  - You will soon be able to convert **unmanaged** assays/datasets to **managed** ones
    - Prevent mistakes to happen



Project

An Experiment belongs to a project

Studies group related datasets of similar type (chip-seq, rna-seq)

ELN & Lab Collections

Equipment

e.g. sequencer

Specimen

e.g. fly line

Protocol

P_1

Protocol

P_2

workflow

Ordered protocol list (workflow) describes sample production

Protocol

P_x

Experiments can derive from Protocol

Experiment

E_2

Detailed notes

consumes

produces

Experiments refer to consumables from collections

Chemical

Plasmid

Antibody

Oligo

Sample Management

Samples are specimen collected in particular conditions

Sample

e.g. embryos 2-4h AEL

Sample

e.g. embryos 4-6h AEL

Derived samples (eg NGS libraries) are generated following protocols

Library

L_1

e.g. SMART-seq Illumina Lib

Library

L_n

e.g. SMART-seq Illumina Lib

Dataset & Assay Management

Run

Illumina Sequencing

Study

S_1

Assay

Illumina Sequencing Lane 2

Assay

Illumina Sequencing Lane 1

Assays consume Samples and generate Datasets

Dataset

D_1

PE Fastq

Dataset

D_n

PE Fastq

A dataset contains results of 1 (or more) librarie(s)

DataFile

Read 1

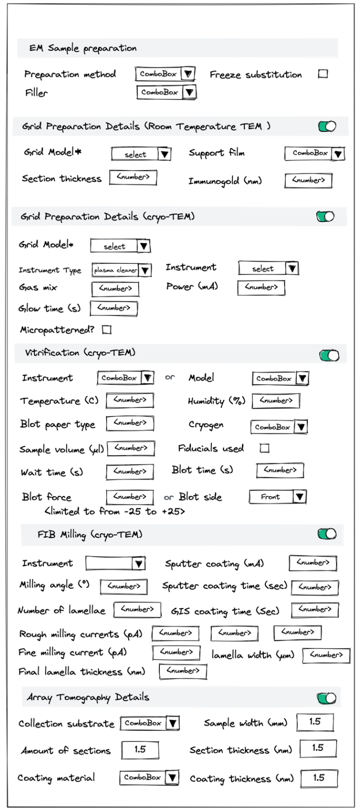
DataFile

Read 2

A dataset has 1 to many data-containing files

EM Data Support coming soon

EM Sample Type



EM Sample preparation

Preparation method	ComboBox ▼	Freeze substitution	☐
Filler	ComboBox ▼		

Grid Preparation Details (Room Temperature TEM)

Grid Model*	select ▼	Support film	ComboBox ▼
Section thickness	<number>	Immunogold (nm)	<number>

Grid Preparation Details (cryo-TEM)

Grid Model*	select ▼		
Instrument Type	plasma cleaner ▼	Instrument	select ▼
Gas mix	<number>	Power (mA)	<number>
Glow time (s)	<number>		
Micropatterned?	☐		

Vitrification (cryo-TEM)

Instrument	ComboBox ▼	or Model	ComboBox ▼
Temperature (C)	<number>	Humidity (%)	<number>
Blot paper type	<number>	Cryogen	ComboBox ▼
Sample volume (µl)	<number>	Fiducials used	☐
Wait time (s)	<number>	Blot time (s)	<number>
Blot force	<number>	or Blot side	Front ▼

<limited to from -25 to +25>

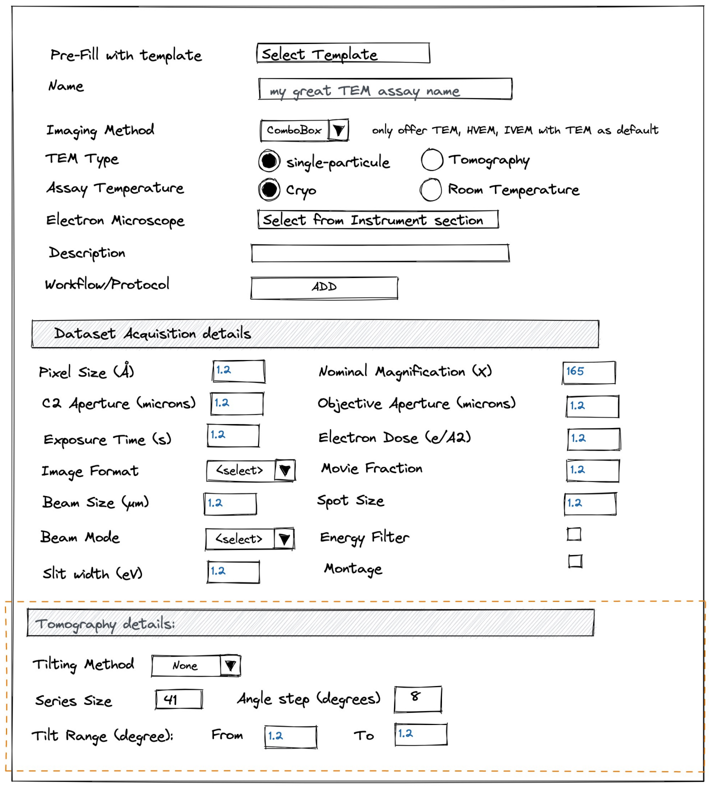
FIB Milling (cryo-TEM)

Instrument	▼	Sputter coating (mA)	<number>
Milling angle (°)	<number>	Sputter coating time (sec)	<number>
Number of lamellae	<number>	GIS coating time (Sec)	<number>
Rough milling currents (pA)	<number> <number> <number>		
Fine milling current (pA)	<number>	lamella width (µm)	<number>
Final lamella thickness (nm)	<number>		

Array Tomography Details

Collection substrate	ComboBox ▼	Sample width (mm)	1.5
Amount of sections	1.5	Section thickness (nm)	1.5
Coating material	ComboBox ▼	Coating thickness (nm)	1.5

Transmission EM Assay



Pre-Fill with template	Select Template
Name	my great TEM assay name
Imaging Method	ComboBox ▼ only offer TEM, HVEM, IVEM with TEM as default
TEM Type	● single-particule ○ Tomography
Assay Temperature	● Cryo ○ Room Temperature
Electron Microscope	Select from Instrument section
Description	
Workflow/Protocol	ADD

Dataset Acquisition details

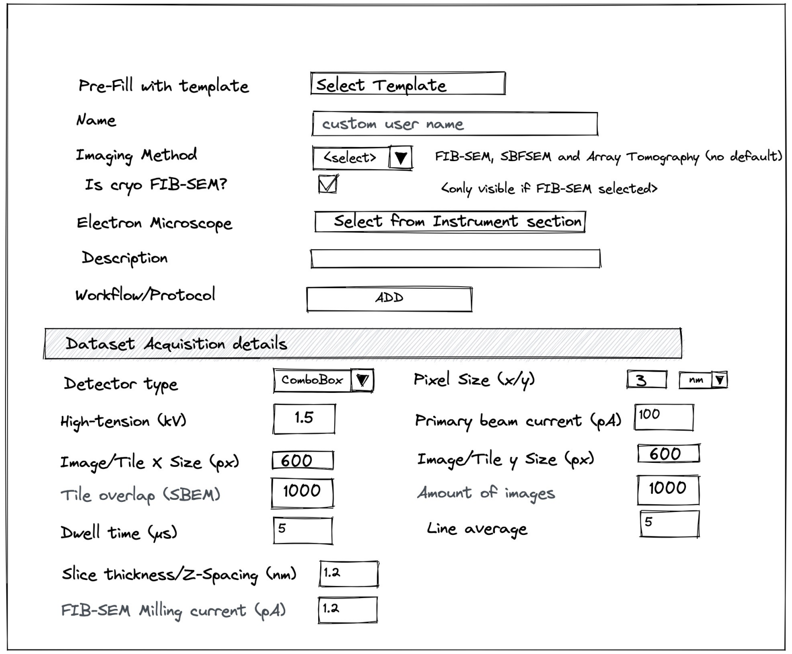
Pixel Size (Å)	1.2	Nominal Magnification (X)	165
C2 Aperture (microns)	1.2	Objective Aperture (microns)	1.2
Exposure Time (s)	1.2	Electron Dose (e/A2)	1.2
Image Format	<select> ▼	Movie Fraction	1.2
Beam Size (µm)	1.2	Spot Size	1.2
Beam Mode	<select> ▼	Energy Filter	☐
Slit width (eV)	1.2	Montage	☐

Tomography details:

Tilting Method	None ▼		
Series Size	41	Angle step (degrees)	8
Tilt Range (degree): From	1.2	To	1.2

or

Volume EM (Assay)



Pre-Fill with template	Select Template
Name	custom user name
Imaging Method	<select> ▼ FIB-SEM, SBFSEM and Array Tomography (no default)
Is cryo FIB-SEM?	☑ <only visible if FIB-SEM selected>
Electron Microscope	Select from Instrument section
Description	
Workflow/Protocol	ADD

Dataset Acquisition details

Detector type	ComboBox ▼	Pixel Size (x/y)	3 nm ▼
High-tension (kV)	1.5	Primary beam current (pA)	100
Image/Tile X Size (px)	600	Image/Tile y Size (px)	600
Tile overlap (SBEM)	1000	Amount of images	1000
Dwell time (µs)	5	Line average	5
Slice thickness/Z-Spacing (nm)	1.2		
FIB-SEM Milling current (pA)	1.2		

=> Flow Cytometry and other omics assays to follow !

Thank you for your attention

GBCS

Jelle Scholtalbers

Matthias Monfort

Nayeem Reza

Questions ?