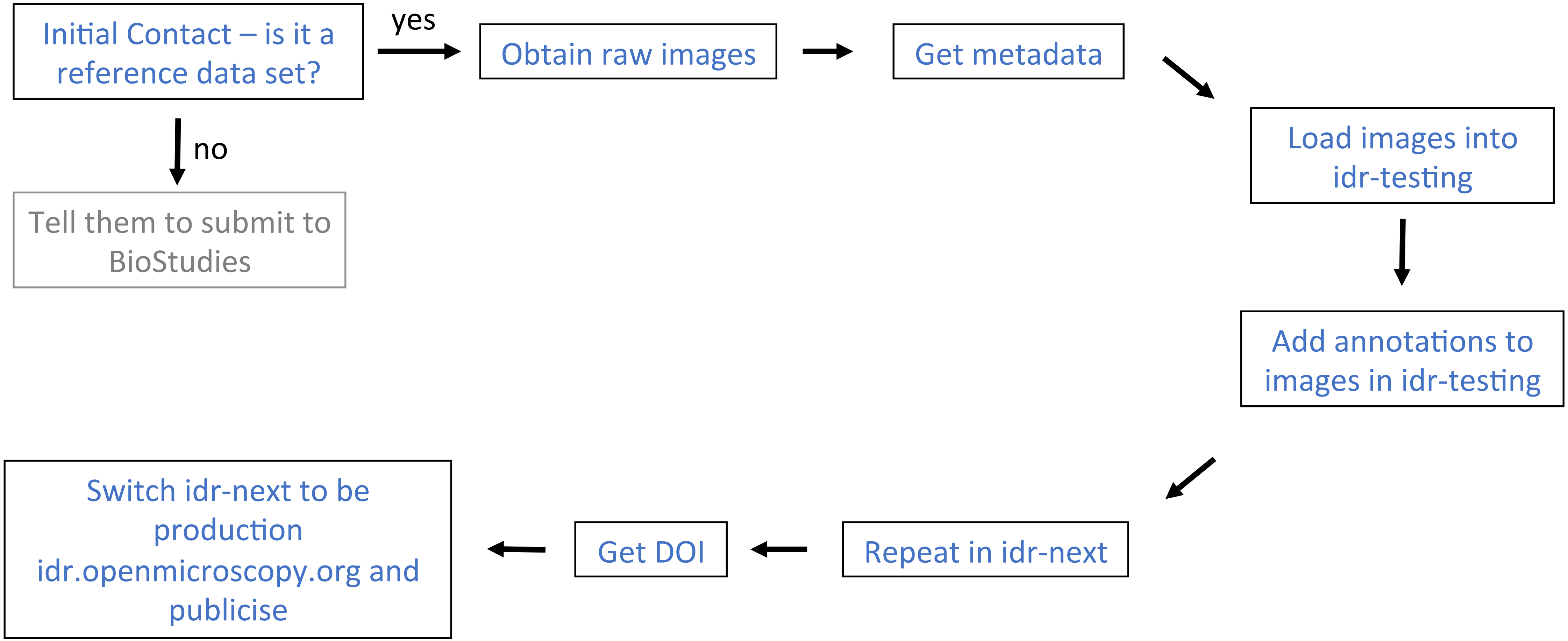
IDR submission workflow

Eleanor Williams

8th November 2017

Over view of the work flow



Initial Contact – is it a reference data set?

yes

Obtain raw images

Get metadata

Load images into idr-testing

Add annotations to images in idr-testing

Repeat in idr-next

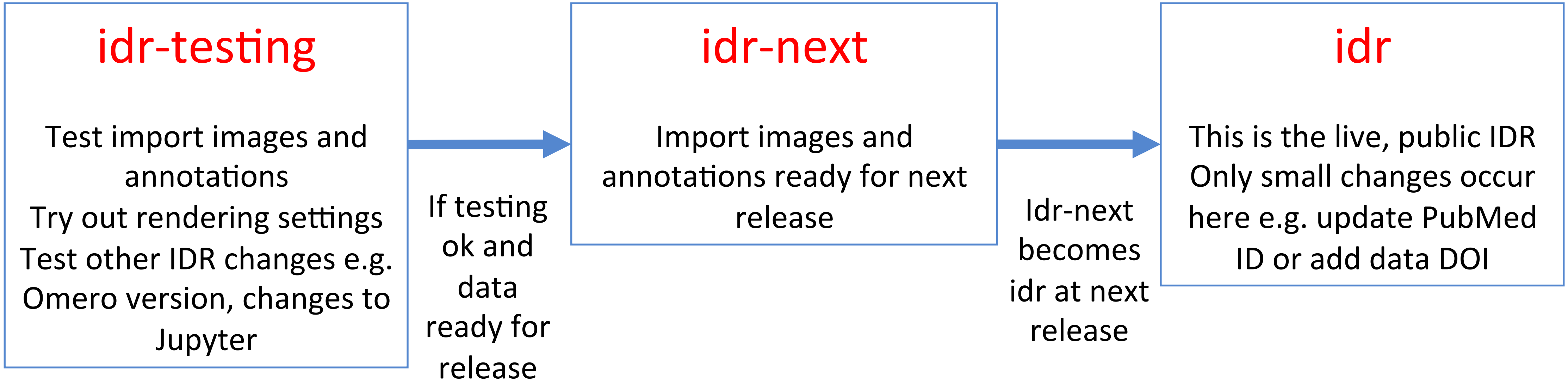
Get DOI

Switch idr-next to be production idr.openmicroscopy.org and publicise

no

Tell them to submit to BioStudies

idr-testing, idr-next and idr production



idr-testing

Test import images and annotations
Try out rendering settings
Test other IDR changes e.g. Omero version, changes to Jupyter

If testing ok and data ready for release

idr-next

Import images and annotations ready for next release

Idr-next becomes idr at next release

idr

This is the live, public IDR
Only small changes occur here e.g. update PubMed ID or add data DOI

What is a reference dataset?

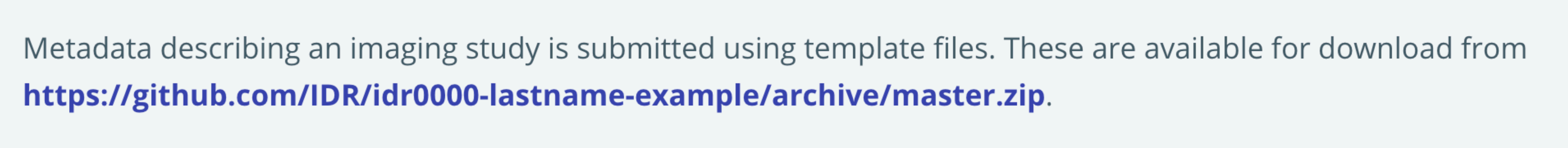
- have value beyond simply supporting an original publication
- Guidance from EuroBioimaging. See http://www.eurobioimaging.eu/sites/default/files/Euro-BioImaging_Elixir_Image_Data_Strategy_0.pdf
- Criteria we use (see our submission help page) are:
    - Datasets **associated** with an existing or upcoming publication
    - **Complete** datasets - not just images supporting one figure in the publication
    - Datasets whose metadata can be **integrated** with other datasets via identifiers from well-known biomolecular resources (Ensembl, NCBI Entrez Gene, RefSeq, PubChem, ChEBI etc)
    - Datasets generated using new imaging **methods** or new analysis methods
    - Datasets that are likely to be **re-analysed or incorporated** into other studies or integrated with other imaging datasets

Obtain raw image data and experimental metadata

- **Raw images** – send them a hard drive by post if over 500 Gb. If less than 500 Gb then we are going to set up an FTP transfer.

Image credit: Simon Li

- **Experimental metadata** – ask them to fill out metadata templates – link on https://idr.openmicroscopy.org/about/submission.html



Metadata describing an imaging study is submitted using template files. These are available for download from **https://github.com/IDR/idr0000-lastname-example/archive/master.zip**.

Experimental metadata files

High Content Screen	Non-screen study
Study file - mandatory	Study file - mandatory
Library file - mandatory	Assay file - mandatory
Processed data file - optional	Processed data file - optional
Feature data, tracking data - optional	Feature data, tracking data - optional

Lots of examples in https://github.com/IDR/idr-metadata

Study file

Title, description
Contact info
Publication info
License

Appears only once for each study

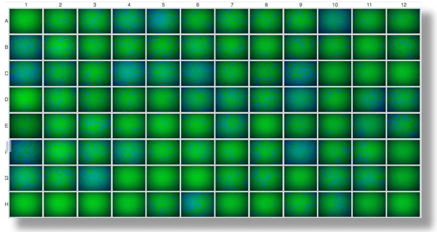
Library info for HCS
Experimental Conditions
Protocols
Phenotypes + CMPO mappings
Links to library/assay and processed files

Repeated block for each screen or experiment
e.g. screenA, experimentA

Library info for HCS
Experimental Conditions
Protocols
Phenotypes + CMPO mappings
Links to library/assay and processed files

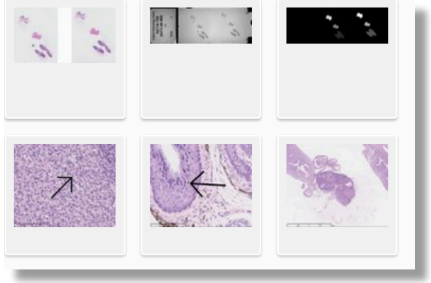
Repeated block for each screen or experiment
e.g. screenB, experimentB

Library and Assay files



Library file

- One row for each plate + well
- Describe the **sample** in the well e.g. species, cell line
- Describe **treatment** to the sample e.g. siRNA use, compound treatment, different media used to grow the cells
- Which are **control wells** – positive control – expect an effect, negative control – don't expect an effect
- **Quality control** – any wells rejected by authors e.g. out of focus, too few cells
- **Channels** – label and what is labeled e.g. DAPI:nucleus



Assays file

- One row for image file
- Describe the **sample** in the image e.g. species, cell line
- Describe **treatment** to the sample e.g. siRNA use, compound treatment, different media used to grow the cells
- List the **protocols** applied
- Group into **Datasets**
- Specify if **raw or processed** image
- **Channels** – label and what is labeled e.g. DAPI:nucleus
- Links to **processed files**

Example library file – idr0013-screenA

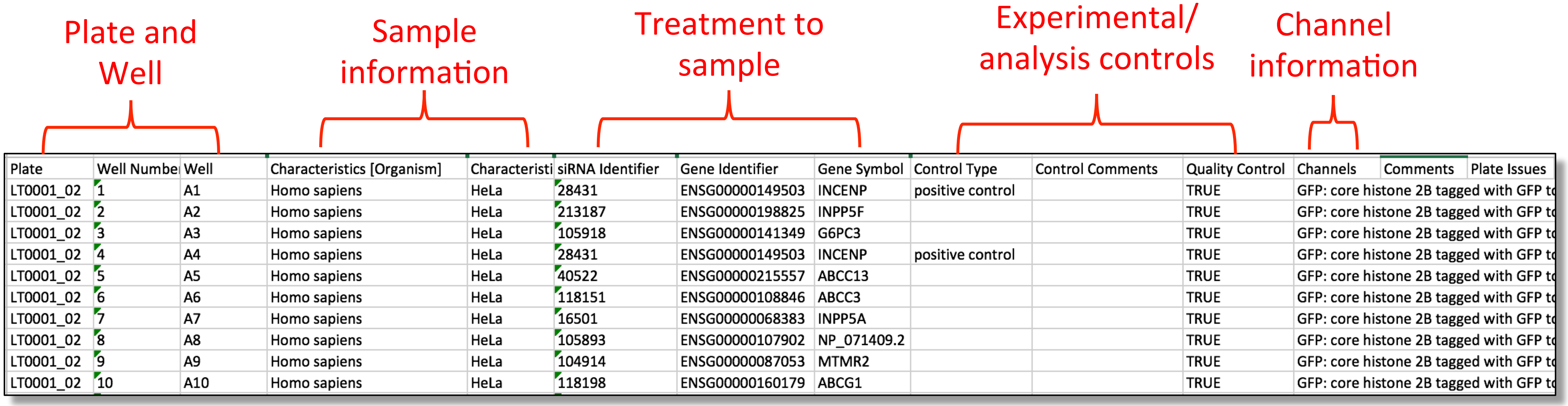


Plate and Well

Sample information

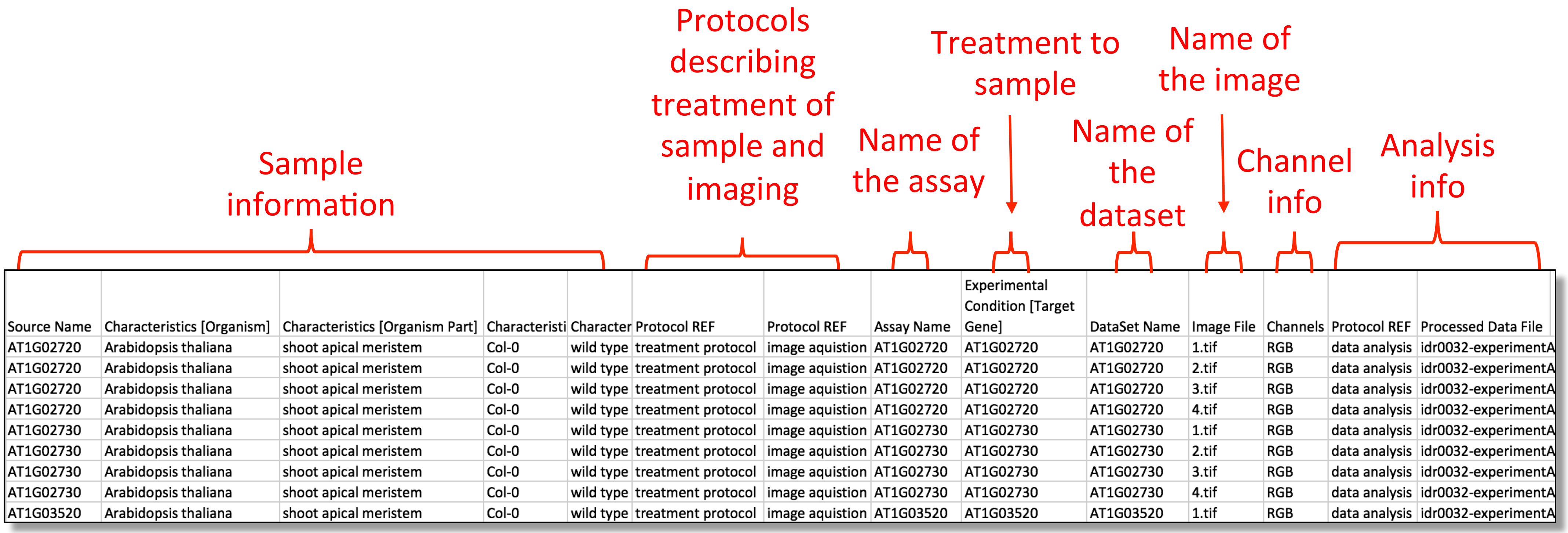
Treatment to sample

Experimental/ analysis controls

Channel information

Plate	Well Numbe	Well	Characteristics [Organism]	Characteristi	siRNA Identifier	Gene Identifier	Gene Symbol	Control Type	Control Comments	Quality Control	Channels	Comments	Plate Issues
LT0001_02	1	A1	Homo sapiens	HeLa	28431	ENSG00000149503	INCENP	positive control		TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	2	A2	Homo sapiens	HeLa	213187	ENSG00000198825	INPP5F			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	3	A3	Homo sapiens	HeLa	105918	ENSG00000141349	G6PC3			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	4	A4	Homo sapiens	HeLa	28431	ENSG00000149503	INCENP	positive control		TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	5	A5	Homo sapiens	HeLa	40522	ENSG00000215557	ABCC13			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	6	A6	Homo sapiens	HeLa	118151	ENSG00000108846	ABCC3			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	7	A7	Homo sapiens	HeLa	16501	ENSG00000068383	INPP5A			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	8	A8	Homo sapiens	HeLa	105893	ENSG00000107902	NP_071409.2			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	9	A9	Homo sapiens	HeLa	104914	ENSG00000087053	MTMR2			TRUE	GFP: core histone 2B tagged with GFP t		
LT0001_02	10	A10	Homo sapiens	HeLa	118198	ENSG00000160179	ABCG1			TRUE	GFP: core histone 2B tagged with GFP t		

Example assay file – idr0032-experimentA



Sample information

Protocols describing treatment of sample and imaging

Name of the assay

Treatment to sample

Name of the dataset

Name of the image

Channel info

Analysis info

Source Name	Characteristics [Organism]	Characteristics [Organism Part]	Characteristi	Character	Protocol REF	Protocol REF	Assay Name	Experimental Condition [Target Gene]	DataSet Name	Image File	Channels	Protocol REF	Processed Data File
AT1G02720	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02720	AT1G02720	AT1G02720	1.tif	RGB	data analysis	idr0032-experimentA
AT1G02720	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02720	AT1G02720	AT1G02720	2.tif	RGB	data analysis	idr0032-experimentA
AT1G02720	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02720	AT1G02720	AT1G02720	3.tif	RGB	data analysis	idr0032-experimentA
AT1G02720	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02720	AT1G02720	AT1G02720	4.tif	RGB	data analysis	idr0032-experimentA
AT1G02730	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02730	AT1G02730	AT1G02730	1.tif	RGB	data analysis	idr0032-experimentA
AT1G02730	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02730	AT1G02730	AT1G02730	2.tif	RGB	data analysis	idr0032-experimentA
AT1G02730	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02730	AT1G02730	AT1G02730	3.tif	RGB	data analysis	idr0032-experimentA
AT1G02730	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G02730	AT1G02730	AT1G02730	4.tif	RGB	data analysis	idr0032-experimentA
AT1G03520	Arabidopsis thaliana	shoot apical meristem	Col-0	wild type	treatment protocol	image aquistion	AT1G03520	AT1G03520	AT1G03520	1.tif	RGB	data analysis	idr0032-experimentA

Assays file – grouping into datasets

Sample	Experimental conditions	Assay	Dataset	Image File
sample1	Localization of protein X	assay1	proteinXlocalization	1.tiff
sample1	Localization of protein X	assay1	proteinXlocalization	2.tiff
sample2	Localization of protein Y	assay2	proteinYlocalization	1.tiff
sample2	Localization of protein Y	assay2	proteinYlocalization	2.tiff

idr0032

Sample	Experimental Conditions	Assay	Dataset	Image File
sample1	Embryonic kidney + localization of protein X	assay1	proteinXlocalization	1.czi
sample2	Kidney organoid + localization of protein X	assay2	proteinXlocalization	2.czi
sample3	Embryonic kidney + localization of protein Y	assay3	proteinYlocalization	3.czi
sample4	Kidney organoid + localization of protein Y	assay4	proteinYlocalization	4.czi

idr0038

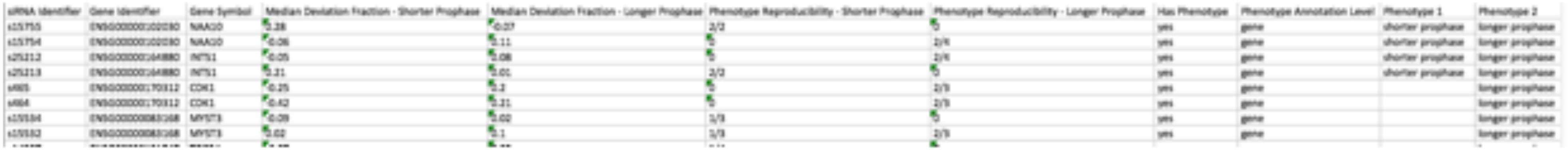
Processed data files

- Summary results and phenotypes
- Must be able to link to library file or assay file in some way – link specified in the study file

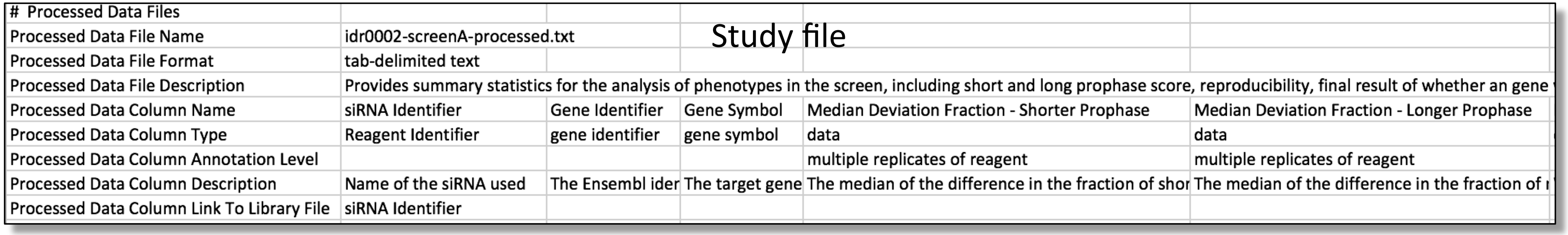
Link to library file

Results related to siRNA derived from multiple replicates of the siRNA in different wells

Result related to gene, derived from multiple siRNAs



siRNA Identifier	Gene Identifier	Gene Symbol	Median Deviation Fraction - Shorter Prophase	Median Deviation Fraction - Longer Prophase	Phenotype Reproducibility - Shorter Prophase	Phenotype Reproducibility - Longer Prophase	Has Phenotype	Phenotype Annotation Level	Phenotype 1	Phenotype 2
s15755	ENSG00000102030	NAA10	0.28	-0.07	2/2	0	yes	gene	shorter prophase	longer prophase
s15754	ENSG00000102030	NAA10	-0.06	0.11	0	2/4	yes	gene	shorter prophase	longer prophase
s25212	ENSG00000164880	INTS1	-0.05	0.08	0	2/4	yes	gene	shorter prophase	longer prophase
s25213	ENSG00000164880	INTS1	0.21	0.01	2/2	0	yes	gene	shorter prophase	longer prophase
s465	ENSG00000170312	CDK1	-0.25	0.2	0	2/3	yes	gene		longer prophase
s464	ENSG00000170312	CDK1	-0.42	0.21	0	2/3	yes	gene		longer prophase
s15504	ENSG00000083168	MYST3	-0.09	0.02	1/3	0	yes	gene		longer prophase
s15502	ENSG00000083168	MYST3	0.02	0.1	1/3	2/3	yes	gene		longer prophase



Study file

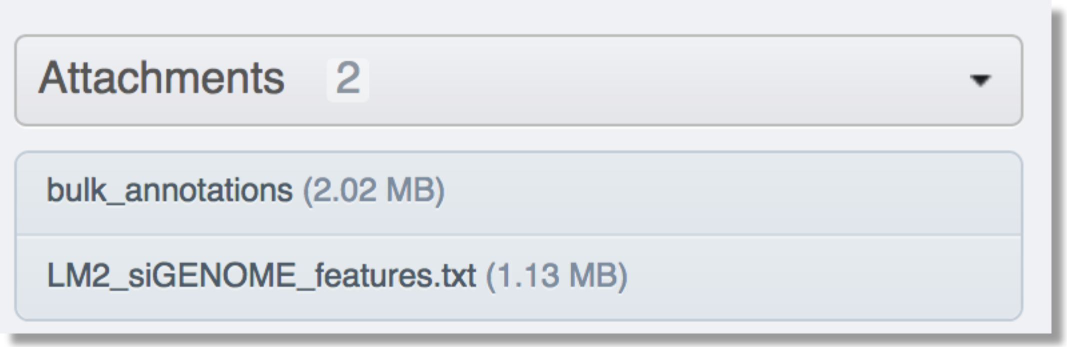
# Processed Data Files					
Processed Data File Name	idr0002-screenA-processed.txt				
Processed Data File Format	tab-delimited text				
Processed Data File Description	Provides summary statistics for the analysis of phenotypes in the screen, including short and long prophase score, reproducibility, final result of whether an gene				
Processed Data Column Name	siRNA Identifier	Gene Identifier	Gene Symbol	Median Deviation Fraction - Shorter Prophase	Median Deviation Fraction - Longer Prophase
Processed Data Column Type	Reagent Identifier	gene identifier	gene symbol	data	data
Processed Data Column Annotation Level				multiple replicates of reagent	multiple replicates of reagent
Processed Data Column Description	Name of the siRNA used	The Ensembl ider	The target gene	The median of the difference in the fraction of shor	The median of the difference in the fraction of
Processed Data Column Link To Library File	siRNA Identifier				

Other files that might be submitted

- Feature level data files + ROI/masks
- Tracking files

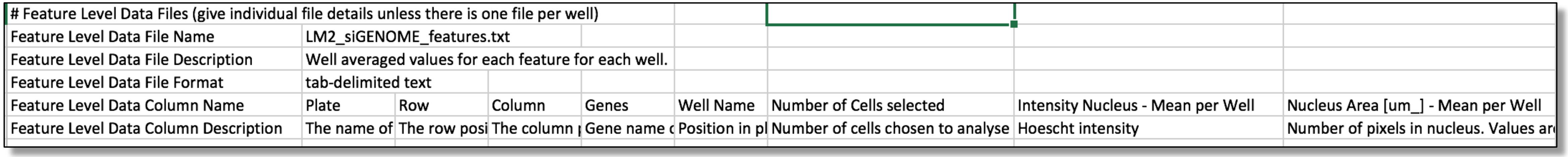
- Listed in study file but only attached to screen/assay if simple to do currently

idr0028-screenA



bulk_annotations (2.02 MB)

LM2_siGENOME_features.txt (1.13 MB)



# Feature Level Data Files (give individual file details unless there is one file per well)								
Feature Level Data File Name	LM2_siGENOME_features.txt							
Feature Level Data File Description	Well averaged values for each feature for each well.							
Feature Level Data File Format	tab-delimited text							
Feature Level Data Column Name	Plate	Row	Column	Genes	Well Name	Number of Cells selected	Intensity Nucleus - Mean per Well	Nucleus Area [um_] - Mean per Well
Feature Level Data Column Description	The name of	The row posi	The column	Gene name	Position in pl	Number of cells chosen to analyse	Hoescht intensity	Number of pixels in nucleus. Values ar

Files needed to load data into IDR

Images

- Raw images – on EBI file system but also copied to Dundee file system
- Plates.tsv or FilePaths.tsv
- Bulk.yml

Annotations

- Annotation.csv
- Bulkmap-config.yml

Plates.tsv/FilePath.tsv

	Column1	Column2
screens	Plate name	Path to directory with images or .screen files
non-screens	Dataset name	Path to image file

idr0002-screenA-plates.tsv

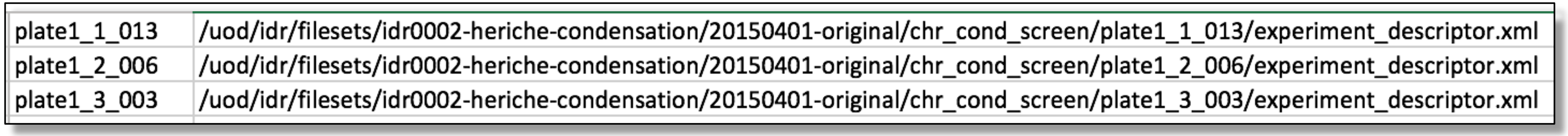
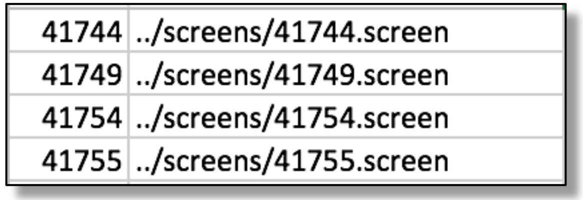


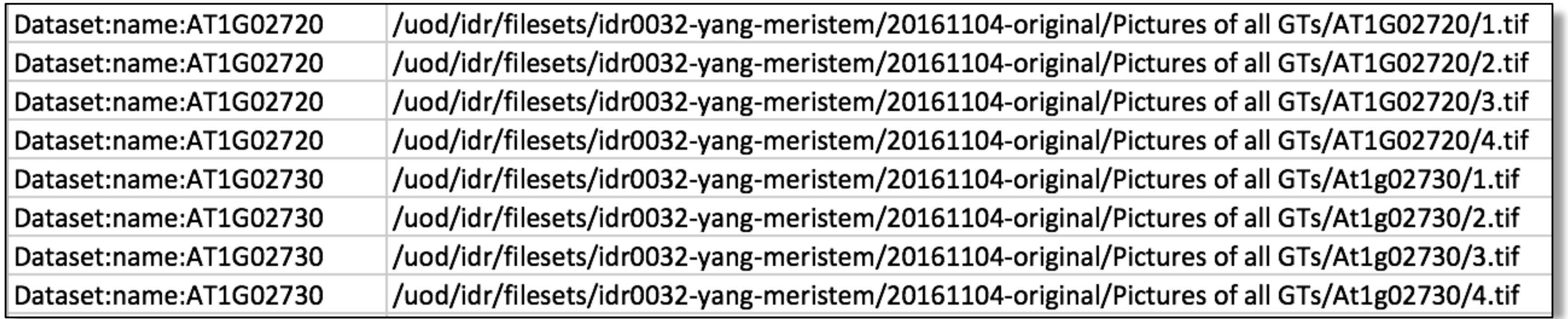
plate1_1_013	/uod/idr/filesets/idr0002-heriche-condensation/20150401-original/chr_cond_screen/plate1_1_013/experiment_descriptor.xml
plate1_2_006	/uod/idr/filesets/idr0002-heriche-condensation/20150401-original/chr_cond_screen/plate1_2_006/experiment_descriptor.xml
plate1_3_003	/uod/idr/filesets/idr0002-heriche-condensation/20150401-original/chr_cond_screen/plate1_3_003/experiment_descriptor.xml

idr0033-screenA-plates.tsv



41744	../screens/41744.screen
41749	../screens/41749.screen
41754	../screens/41754.screen
41755	../screens/41755.screen

idr0032-experimentA-filePaths.tsv

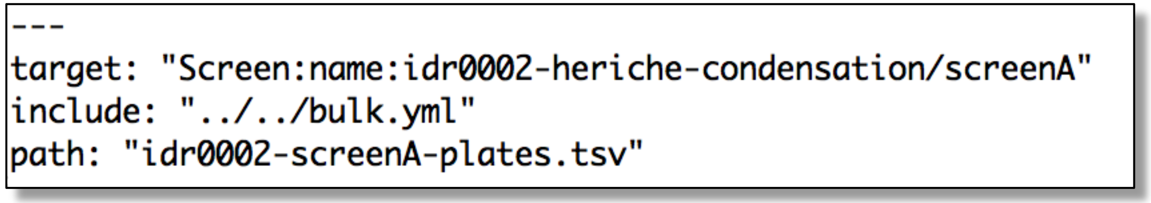


Dataset:name:AT1G02720	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/AT1G02720/1.tif
Dataset:name:AT1G02720	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/AT1G02720/2.tif
Dataset:name:AT1G02720	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/AT1G02720/3.tif
Dataset:name:AT1G02720	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/AT1G02720/4.tif
Dataset:name:AT1G02730	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/At1g02730/1.tif
Dataset:name:AT1G02730	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/At1g02730/2.tif
Dataset:name:AT1G02730	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/At1g02730/3.tif
Dataset:name:AT1G02730	/uod/idr/filesets/idr0032-yang-meristem/20161104-original/Pictures of all GTs/At1g02730/4.tif

Screen or assay bulk.yml

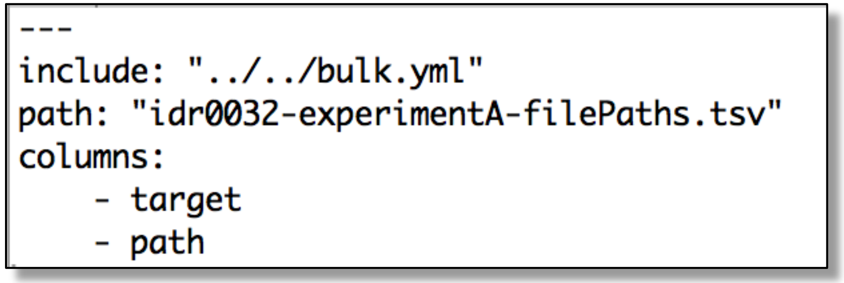
A yaml format file that allows bulk import of all the images on the command line

idr0002-screenA-bulk.yml



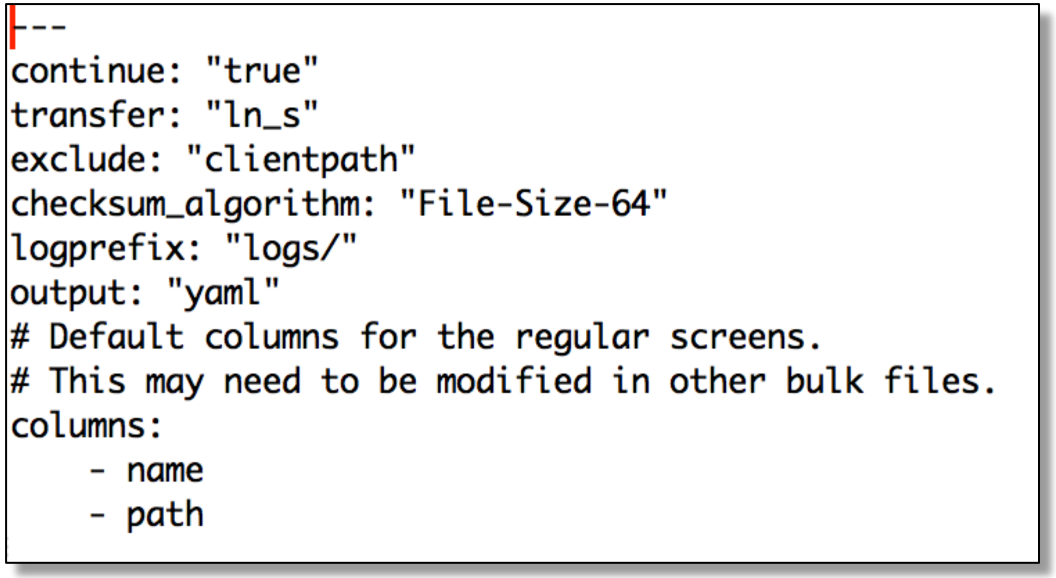
```
---
target: "Screen:name:idr0002-heriche-condensation/screenA"
include: "../../bulk.yml"
path: "idr0002-screenA-plates.tsv"
```

idr0032-experimentA-bulk.yml



```
---
include: "../../bulk.yml"
path: "idr0032-experimentA-filePaths.tsv"
columns:
    - target
    - path
```

- There is a higher level yaml file (idr-metadata/bulk.yml) file that sets some overall parameters about import



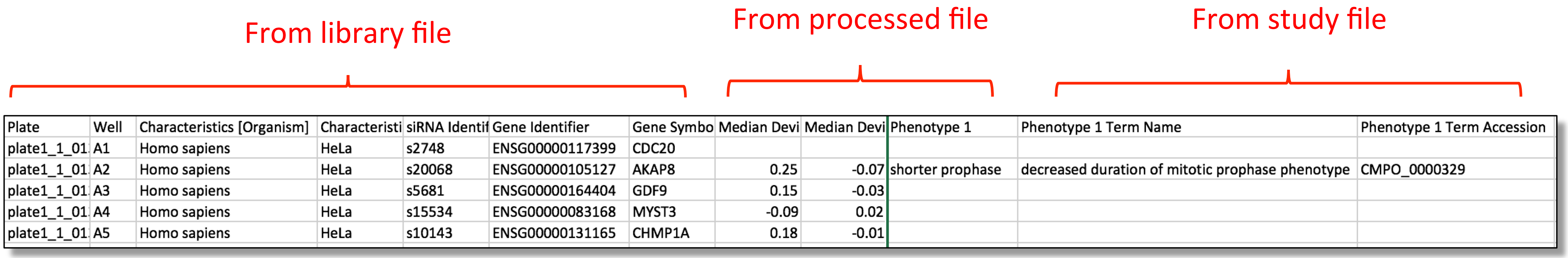
```
---
continue: "true"
transfer: "ln_s"
exclude: "clientpath"
checksum_algorithm: "File-Size-64"
logprefix: "logs/"
output: "yaml"
# Default columns for the regular screens.
# This may need to be modified in other bulk files.
columns:
    - name
    - path
```

Annotation.csv and bulkmap-config.yml

Annotation.csv

- A single annotation file is created for importing an hd5 table into IDR
- Contains metadata from the library or assay file, plus data in the processed file, plus phenotype to CMPO mappings from the study file
- All column headings must be unique
- Created from study, library and processed file for screens using perl script and manually for non-screens

```
create_bulk_annotations_file_using_studyfile.pl –s study.txt –l library.txt –p processed.txt –n screenNumber
```



From library file

From processed file

From study file

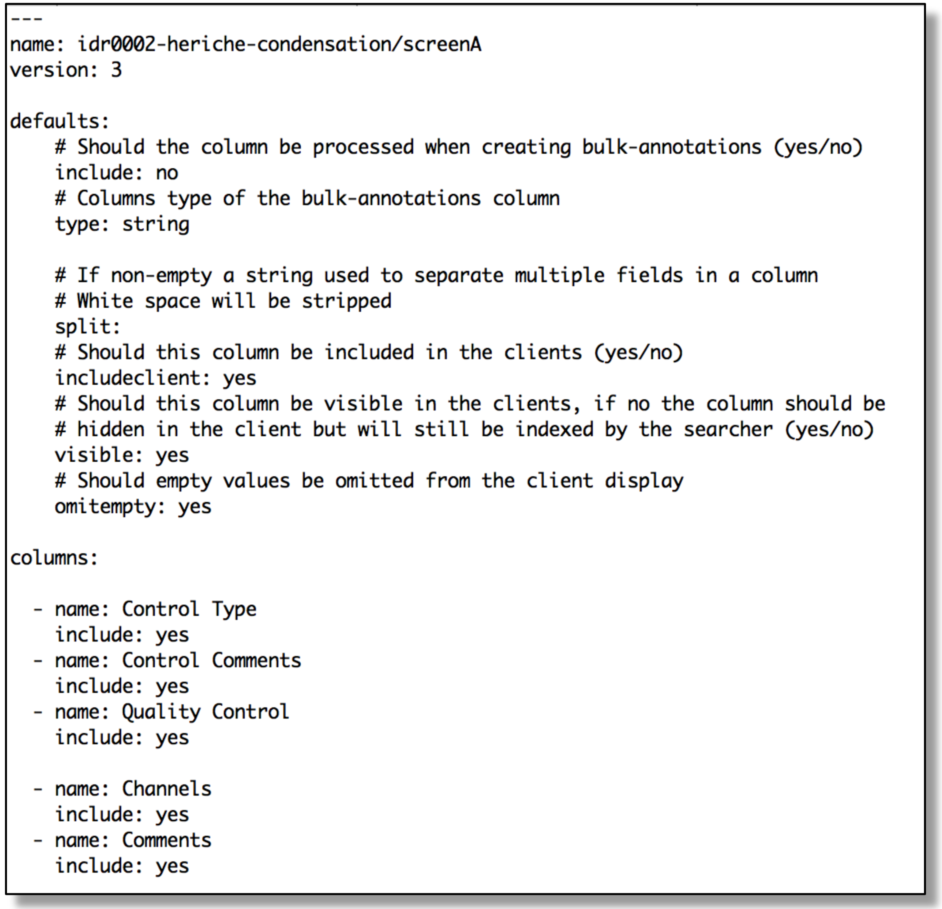
Plate	Well	Characteristics [Organism]	Characteristi	siRNA Identif	Gene Identifier	Gene Symbo	Median Devi	Median Devi	Phenotype 1	Phenotype 1 Term Name	Phenotype 1 Term Accession
plate1_1_01	A1	Homo sapiens	HeLa	s2748	ENSG00000117399	CDC20					
plate1_1_01	A2	Homo sapiens	HeLa	s20068	ENSG00000105127	AKAP8	0.25	-0.07	shorter prophase	decreased duration of mitotic prophase phenotype	CMPO_0000329
plate1_1_01	A3	Homo sapiens	HeLa	s5681	ENSG00000164404	GDF9	0.15	-0.03			
plate1_1_01	A4	Homo sapiens	HeLa	s15534	ENSG00000083168	MYST3	-0.09	0.02			
plate1_1_01	A5	Homo sapiens	HeLa	s10143	ENSG00000131165	CHMP1A	0.18	-0.01			

idr0002-screenA-annotation.csv

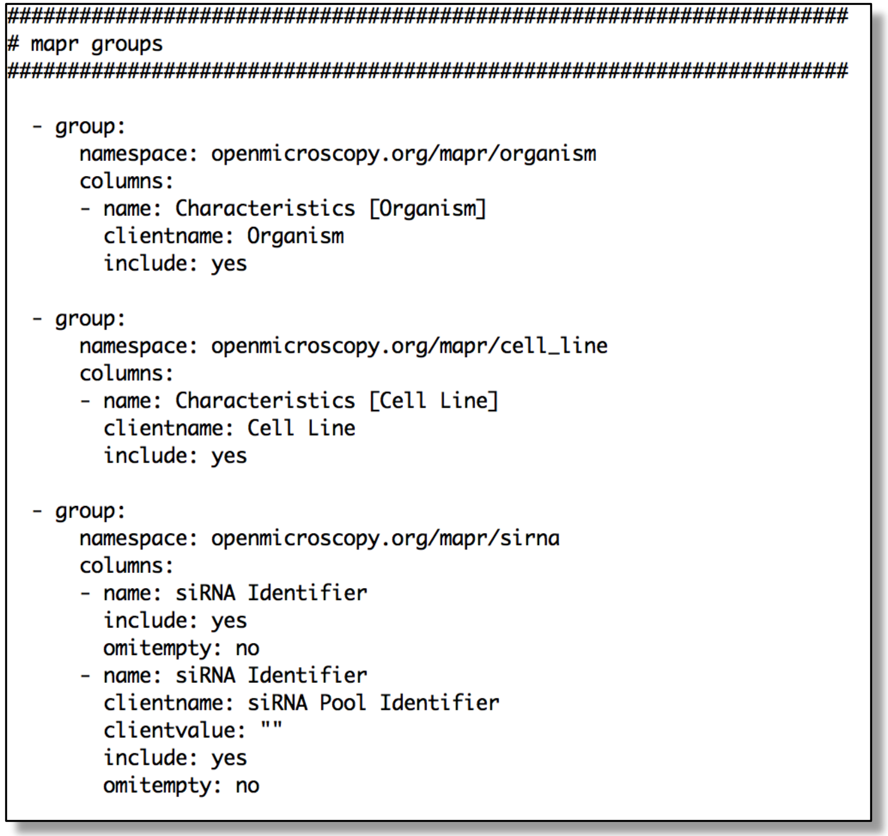
Annotation.csv and bulkmap-config.yml

Bulkmap-config.yml

- A yaml file that says what columns from annotation.csv to create map annotations from and how to display them



```
---
name: idr0002-heriche-condensation/screenA
version: 3

defaults:
    # Should the column be processed when creating bulk-annotations (yes/no)
    include: no
    # Columns type of the bulk-annotations column
    type: string

    # If non-empty a string used to separate multiple fields in a column
    # White space will be stripped
    split:
    # Should this column be included in the clients (yes/no)
    includeclient: yes
    # Should this column be visible in the clients, if no the column should be
    # hidden in the client but will still be indexed by the searcher (yes/no)
    visible: yes
    # Should empty values be omitted from the client display
    omitempty: yes

columns:

  - name: Control Type
    include: yes
  - name: Control Comments
    include: yes
  - name: Quality Control
    include: yes

  - name: Channels
    include: yes
  - name: Comments
    include: yes
```

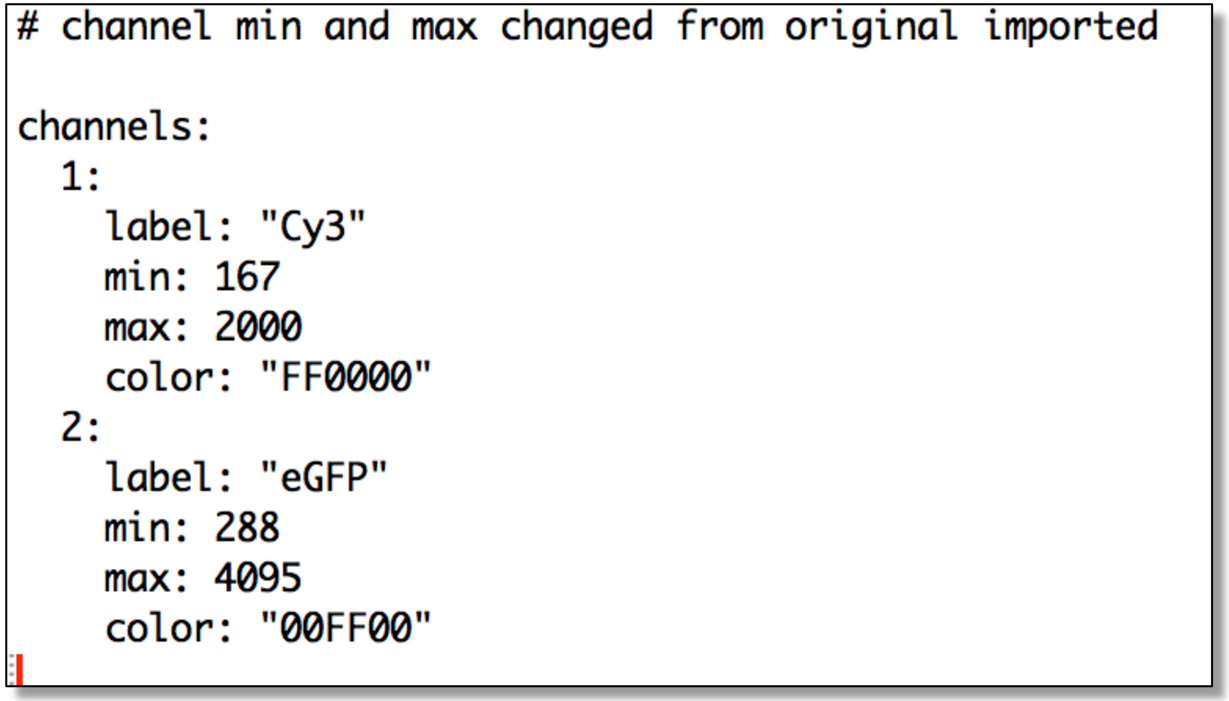


```
######################################################################
# mapr groups
######################################################################

  - group:
      namespace: openmicroscopy.org/mapr/organism
      columns:
      - name: Characteristics [Organism]
        clientname: Organism
        include: yes

  - group:
      namespace: openmicroscopy.org/mapr/cell_line
      columns:
      - name: Characteristics [Cell Line]
        clientname: Cell Line
        include: yes

  - group:
      namespace: openmicroscopy.org/mapr/sirna
      columns:
      - name: siRNA Identifier
        include: yes
        omitempty: no
      - name: siRNA Identifier
        clientname: siRNA Pool Identifier
        clientvalue: ""
        include: yes
        omitempty: no
```

idr0002-screenA-bulkmap-config.yml

Renderdef.yml

- A yaml file that allows you to specify the channel labels, colour, min and max intensity
- Can be applied to an image, plate, screen (don’t think dataset)



```
# channel min and max changed from original imported

channels:
  1:
    label: "Cy3"
    min: 167
    max: 2000
    color: "FF0000"
  2:
    label: "eGFP"
    min: 288
    max: 4095
    color: "00FF00"
```

Idr0002-screenA-renderdef.yml

Git workflow

- Create all input files for a study
- Commit to a branch on your own forked version of https://github.com/IDR/idr-metadata/
- Create a PR against https://github.com/IDR/idr-metadata/
- Create a merge build using “MASTER-push” in Jenkins
- On idr-testing server clone the merge build
- Test files
- If ok, then PR can be merged and https://github.com/IDR/idr-metadata/ can be used on idr-next.

Import of images

Screens

```
omero import --bulk idr0030-screenA-bulk.yml
```

- Will create a screen with the name specified in the bulk.yml file

Non-screen projects

```
omero import --bulk idr0032-experimentA-bulk.yml
```

- Creates datasets that are specified in the filePaths.tsv file but you have to create the project manually via the Web UI

Adding annotations

- Can be done directly or via shell scripts
- In both, first add the annotation.csv file then create map annotations from the value in the bulk annotations table

Directly

```
omero  metadata populate --file idr0002-screenA-annotation.csv Screen:102
```

```
omero  metadata populate --context bulkmap --cfg idr0002-screenA-bulkmap-config.yml Screen:102
```

Via shell scripts in https:/github.com/IDR/idr-metadata/scripts

```
./bulk.sh prod37_input_bulk.txt
```

```
./annotate.sh prod37_input.txt
```

Applying rendering settings

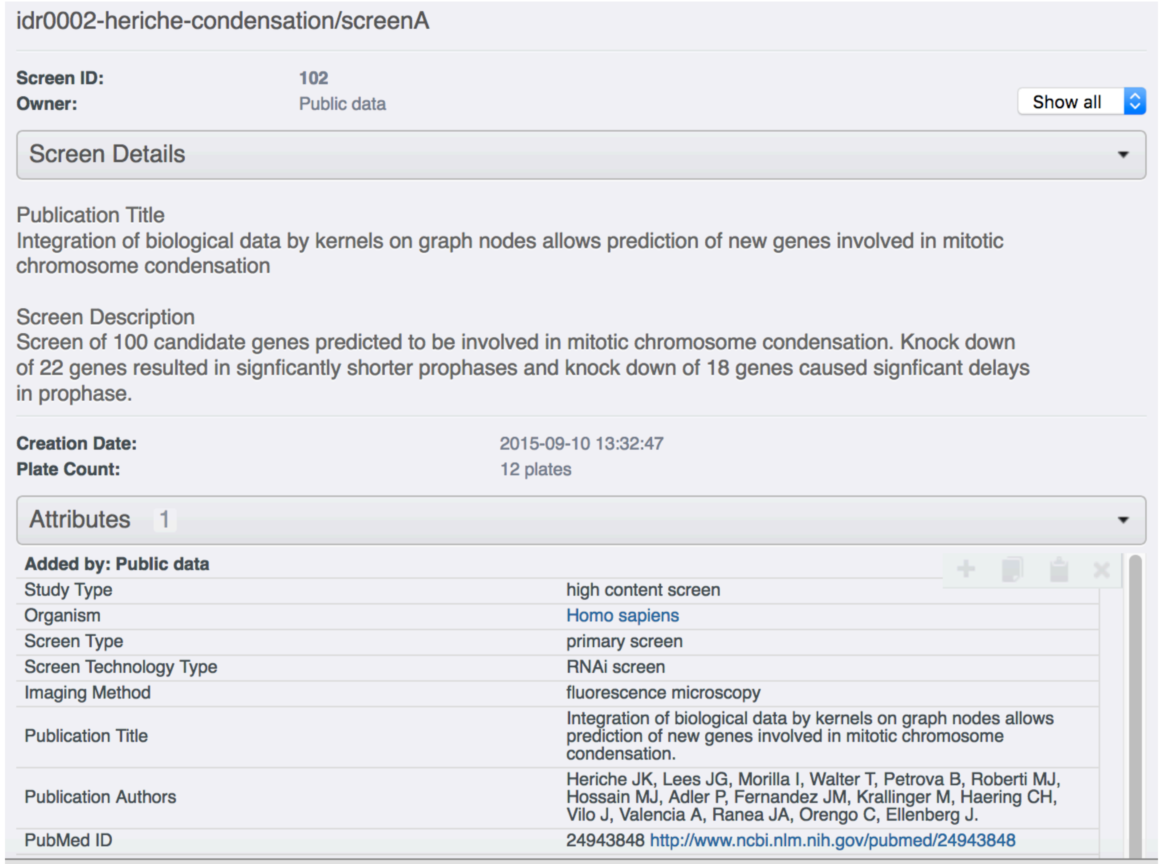
```
omero render edit Plate:1203 idr0019-screenA-renderdef.yml
```

```
omero render edit Screen:1203 idr0019-screenA-renderdef.yml
```

```
omero render edit Image:3427370 idr0038-experimentA-wtFK-cleared-Wt1-
Pax2-renderdef.yml
```

Note: `omero render edit --copy Screen:1203 idr0019-screenA-renderdef.yml` will copy the min and max from the first well to all images in the screen even if we are just specifying channel names in the renderdef.yml file

Adding study/screen/experiment level information



idr0002-heriche-condensation/screenA

Screen ID: 102
Owner: Public data

Screen Details

Publication Title
Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation

Screen Description
Screen of 100 candidate genes predicted to be involved in mitotic chromosome condensation. Knock down of 22 genes resulted in signficantly shorter prophases and knock down of 18 genes caused signficant delays in prophase.

Creation Date: 2015-09-10 13:32:47
Plate Count: 12 plates

Attributes 1

Added by: Public data

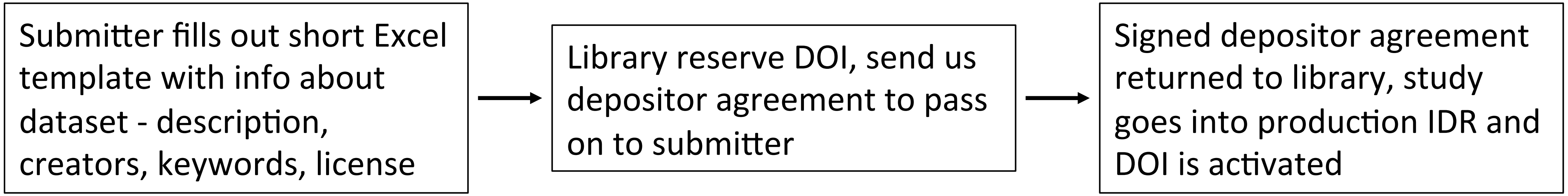
Study Type	high content screen
Organism	Homo sapiens
Screen Type	primary screen
Screen Technology Type	RNAi screen
Imaging Method	fluorescence microscopy
Publication Title	Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation.
Publication Authors	Heriche JK, Lees JG, Morilla I, Walter T, Petrova B, Roberti MJ, Hossain MJ, Adler P, Fernandez JM, Krallinger M, Haering CH, Vilo J, Valencia A, Ranea JA, Orengo C, Ellenberg J.
PubMed ID	24943848 http://www.ncbi.nlm.nih.gov/pubmed/24943848

```
ssh idr-next.openmicroscopy.org -L 12345:test44-omeroreadwrite:80
```

- Login via private window in browser and edit the right hand panel

Get DOI for screen/project

- Arranged through discovery@dundee.ac.uk and Philippa Sterlini in the library
- Minted through DataCite
- Can reserve DOI once complete in idr-next but not activated until study in idr.openmicroscopy.org



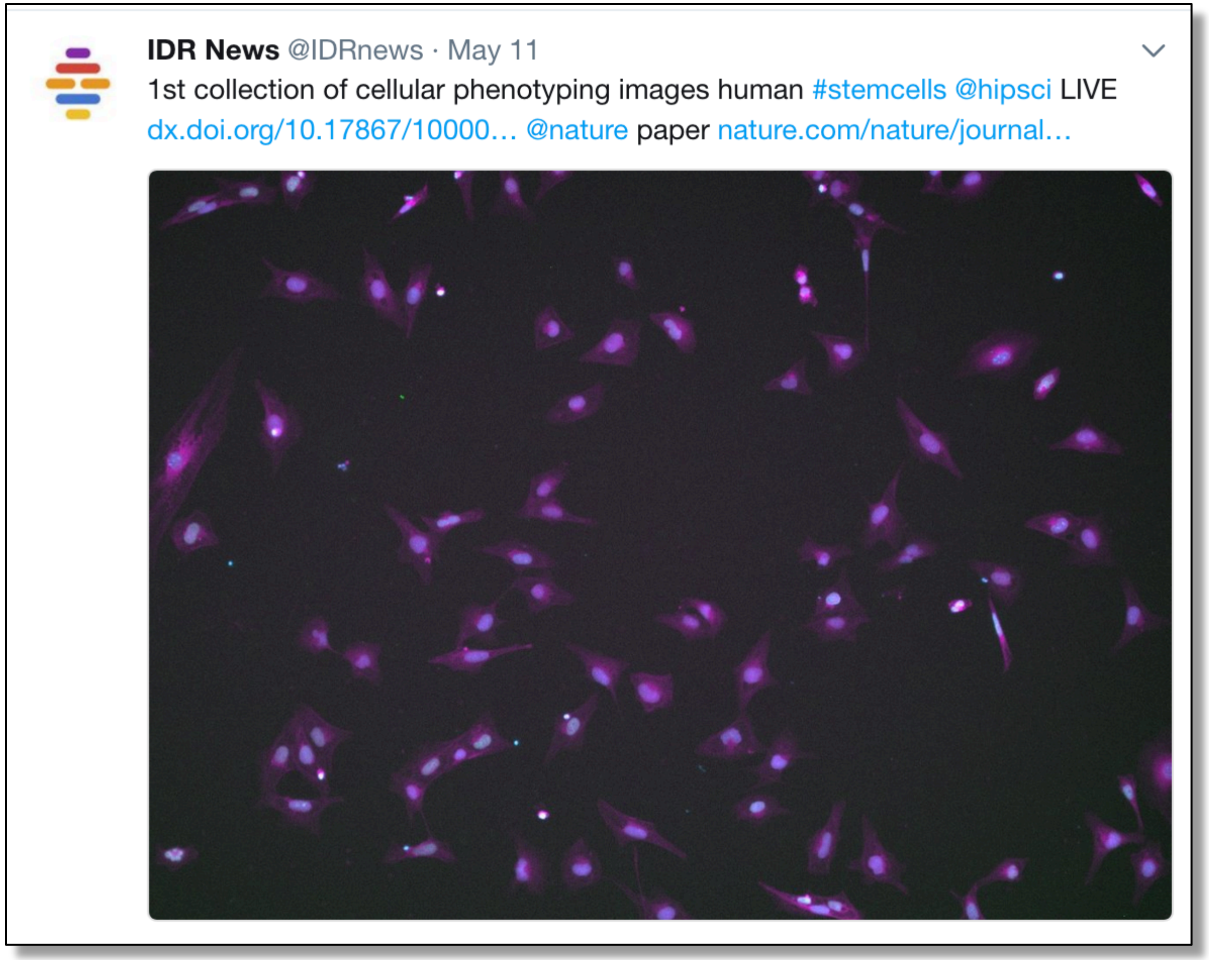
Submitter fills out short Excel template with info about dataset - description, creators, keywords, license

Library reserve DOI, send us depositor agreement to pass on to submitter

Signed depositor agreement returned to library, study goes into production IDR and DOI is activated

- Can create single DOI for a screen/project or parent and child DOIs e.g. idr0028 with a ‘study level’ parent DOI and 4 child DOIs for the 4 screens

Publicize using @IDRnews on Twitter



IDR News @IDRnews · May 11

1st collection of cellular phenotyping images human #stemcells @hipsci LIVE dx.doi.org/10.17867/10000… @nature paper nature.com/nature/journal…

Detailed notes

Detailed notes about every step are being written at https://docs.google.com/a/openmicroscopy.org/document/d/1TmBZ43_yhiO3AOua8oMk4mPWKWJtpeYNc2KLP17h-1I/edit?usp=sharing