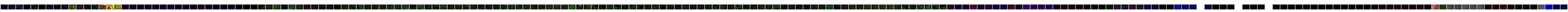


0.01% of the IDR

Image Data Repository

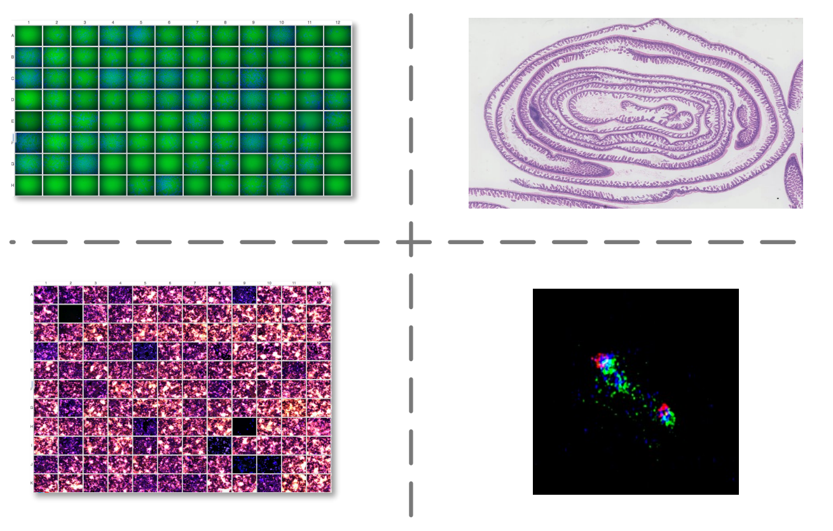
A platform for publishing, integrating and mining imaging-derived biological data at scale

University of Dundee, University of Cambridge, European Bioinformatics Institute and the OME Consortium

Abstract The Image Data Repository is a prototype platform for publishing, mining and integrating bioimaging data at scale, following the Euro-BioImaging/ELIXIR imaging strategy, using the OMERO and Bio-Formats open source software built by the Open Microscopy Environment. Deployed on an OpenStack cloud running on EMBL-EBI's Embassy resource, it includes image data linked to independent studies from genetic, RNAi, chemical, localisation and geographic high content screens, super-resolution microscopy, and digital pathology.

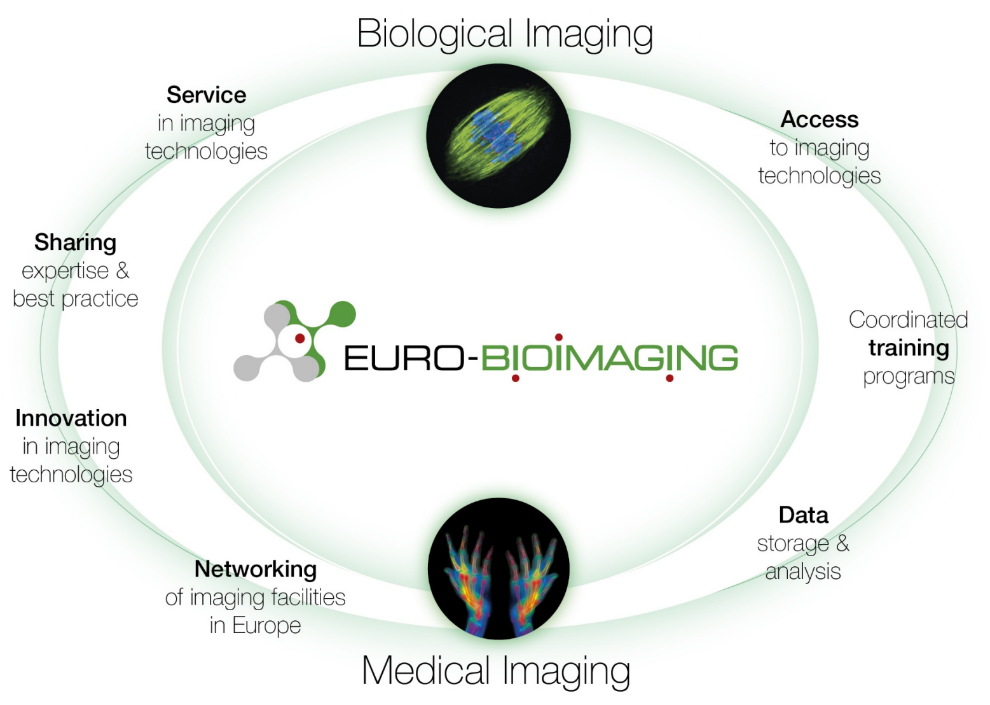
Metadata challenges

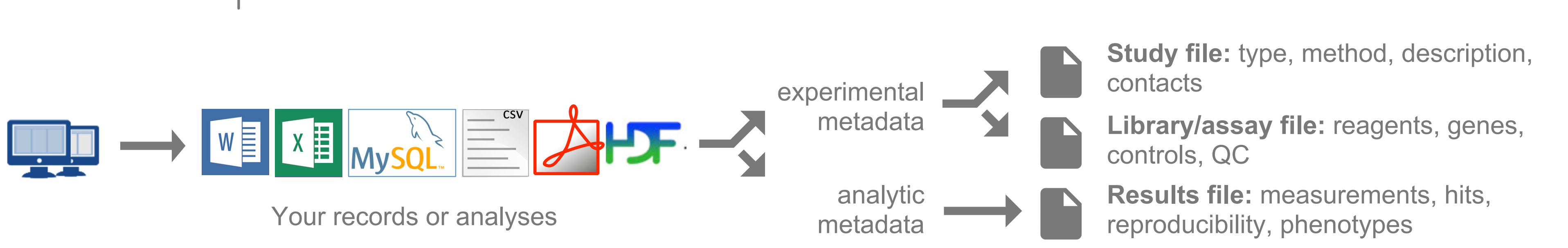
The IDR prototype aims to implement Euro-BioImaging's vision of a central resource for reference image sets, demonstrating that there is value in the curation of existing and future imaging data.



Independently collected datasets are rarely comparable due to differing terminology. Careful curation can permit comparison. The development and use of common vocabularies and/or ontologies is also key.

Common terms such as phenotypes, gene symbols, or compound treatments connect submitted studies, bridging a data divide.


Biological Imaging
Service in imaging technologies
Access to imaging technologies
Sharing expertise & best practice
EURO-BIOIMAGING
Coordinated training programs
Innovation in imaging technologies
Data storage & analysis
Networking of imaging facilities in Europe
Medical Imaging



Study file: type, method, description, contacts

Library/assay file: reagents, genes, controls, QC

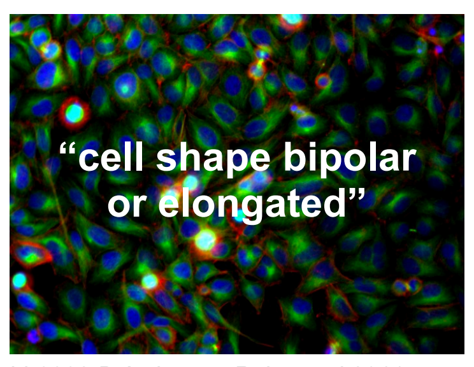
Results file: measurements, hits, reproducibility, phenotypes

Metadata and analysis results collected by authors come in a variety of formats, from spreadsheets to databases. These entries are unified in IDR files during curation at http://github.com/idr

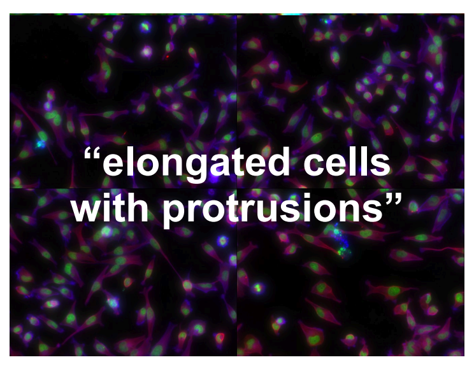
Similar in structure to MAGE-TAB and ISA-TAB, the IDR metadata formats are open and can be used by anyone to store their experimental, imaging, and analytic metadata.

"Elongated cells" Example

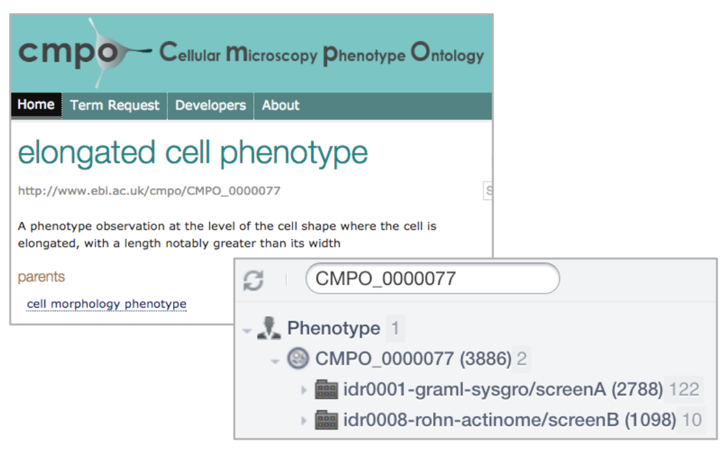
idr0001-A Sysgro, Graml et al 2014, S. pombe genome wide gene knockout screen (image-1233538)



idr0008-B Actinome, Rohn et al 2011, human HeLa targeted actinome RNAi screen (image-109296)



idr0012-A CellMorph, Fuchs et al 2010, human HeLa genome wide RNAi screen (image-1812396)



elongated cell phenotype

http://www.ebi.ac.uk/cmpo/CMPO_0000077

A phenotype observation at the level of the cell shape where the cell is elongated, with a length notably greater than its width

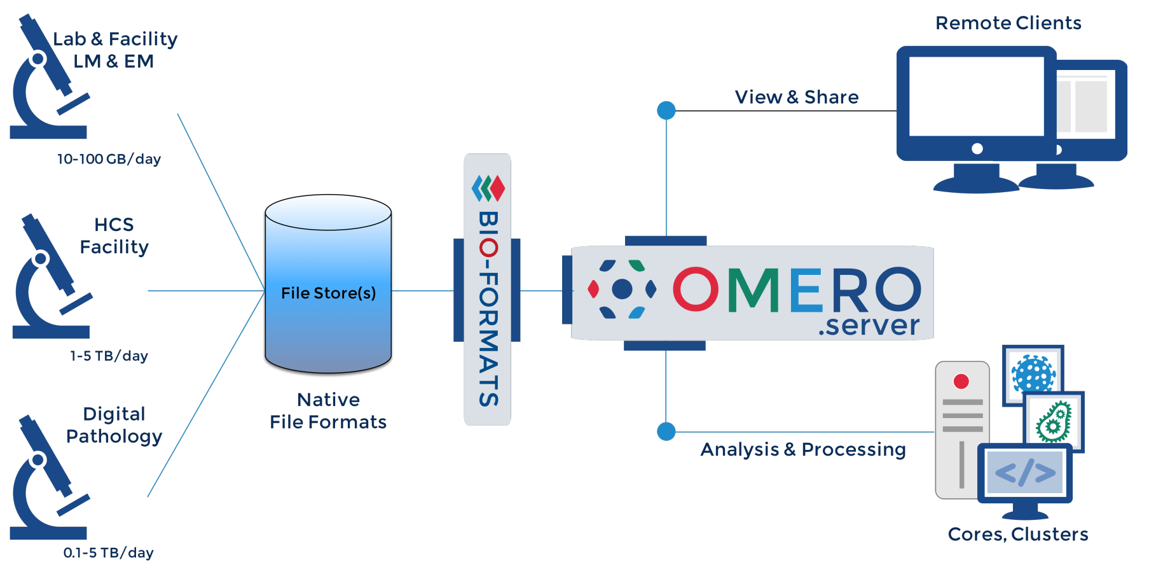
parents: cell morphology phenotype

CMPO_0000077
- Phenotype 1
  - CMPO_0000077 (3886) 2
    - idr0001-graml-sysgro/screenA (2788) 122
    - idr0008-rohn-actinome/screenB (1098) 10

Technologies used

Bio-Formats

Submitted studies come from a wide-range of acquisition systems. The IDR stores the original data without duplication and employs Bio-Formats to access the different file formats through a single API. **More than 140** proprietary formats are supported, and adapters can be written specifically for reference datasets.



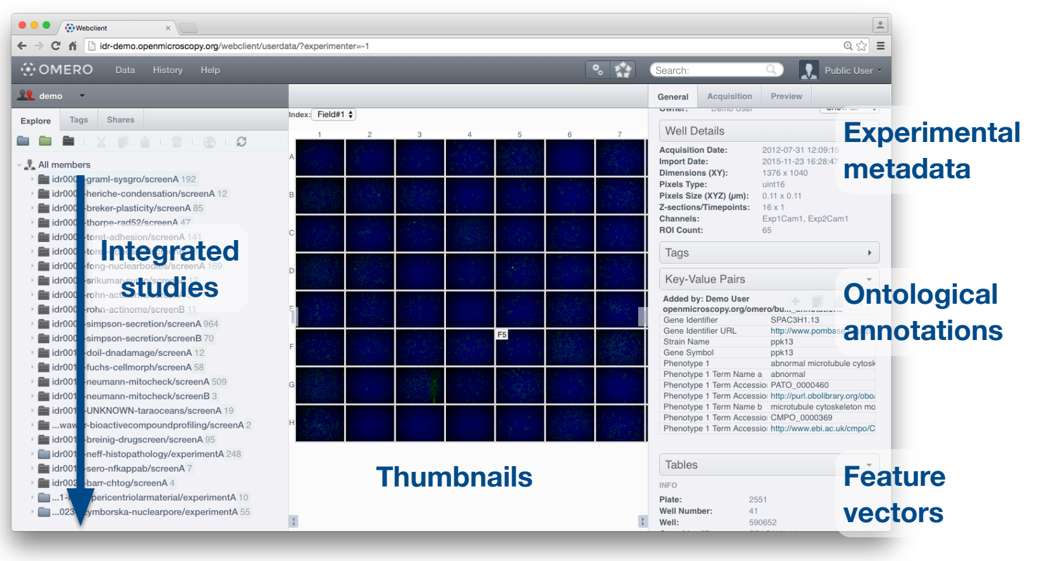
OMERO

The IDR combines submitted studies within a single, standard OMERO server. Cross-linking between studies, e.g. by phenotypes and genes, as well as full text search become possible when all the studies are brought together.

Once public, the OMERO API will enable re-analysis and comparison with existing datasets, either locally or in the cloud.

All software including source code can be found at http://downloads.openmicroscopy.org

Download local analysis **Cloud analysis** **Cross-data browsing**



Aberdeen
Dundee
Edinburgh
Glasgow
Belfast
UNITED KINGDOM
Preston
Manchester
Stoke-on-Trent
Nottingham
Birmingham
Milton Keynes
Swansea
Cardiff
Bristol
London
Southampton
openstack CLOUD SOFTWARE
ANSIBLE

Deployment infrastructure

All filesets delivered to the IDR team have been mirrored between a GPFS cluster in Dundee and a storage system at EBI, each accessible by an OpenStack cloud.

Combined they have more than **1200 VCPUs** with **almost 6 terabytes of memory.** Ansible is used to automate deployments of the system. A full clone including a copy-on-write version of the entire database can be spun up in minutes allowing for third-party investigations of interesting relationships.

Playbooks and roles for these deployments are available at:

http://github.com/openmicroscopy/infrastructure.

Simon Li @crucifixkiss

Another 20TB of #BigData for the @openmicroscopy @BBSRC @emblebi image repository

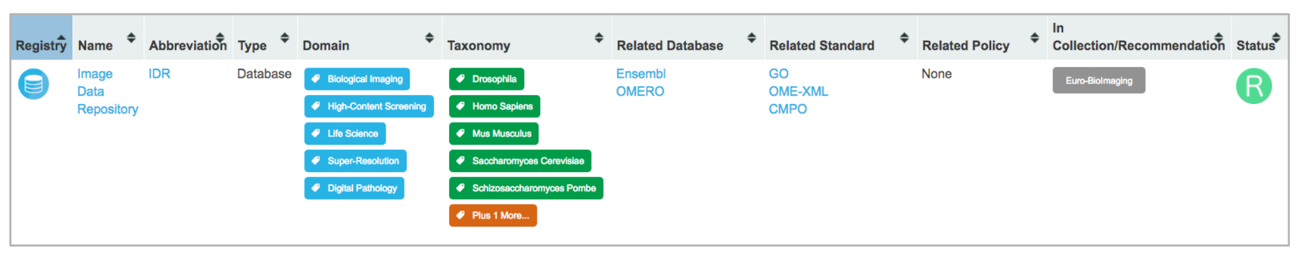
RETWEET 1 LIKES 3

1:43 PM - 15 Sep 2015

A **sneaker network** still proves to be the most convenient and reliable way to accept terabytes of data.

databases > biodbcore-000778

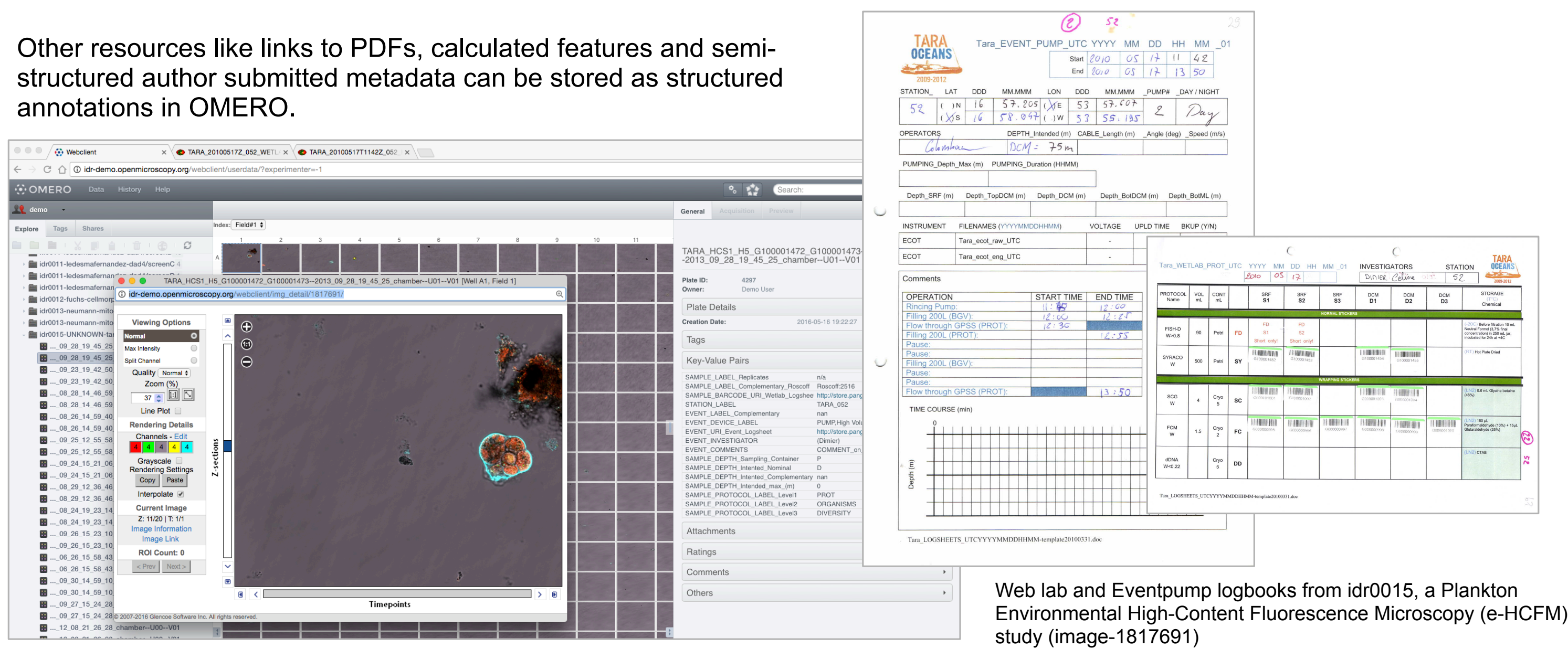
Where possible known community-accepted resources are used to simplify discovery. Potentials for re-use are tracked on the biosharing.org site.



Registry	Name	Abbreviation	Type	Domain	Taxonomy	Related Database	Related Standard	Related Policy	In Collection/Recommendation	Status
	Image Data Repository	IDR	Database	Biological Imaging High-Content Screening Life Science Super-Resolution Digital Pathology	Drosophila Homo Sapiens Mus Musculus Saccharomyces Cerevisiae Schizosaccharomyces Pombe Plus 1 More...	Ensembl OMERO	GO OME-XML CMPO	None	Euro-BioImaging	R

External resources

Other resources like links to PDFs, calculated features and semi-structured author submitted metadata can be stored as structured annotations in OMERO.



Web lab and Eventpump logbooks from idr0015, a Plankton Environmental High-Content Fluorescence Microscopy (e-HCFM) study (image-1817691)

See http://www.embl.de/tara-oceans

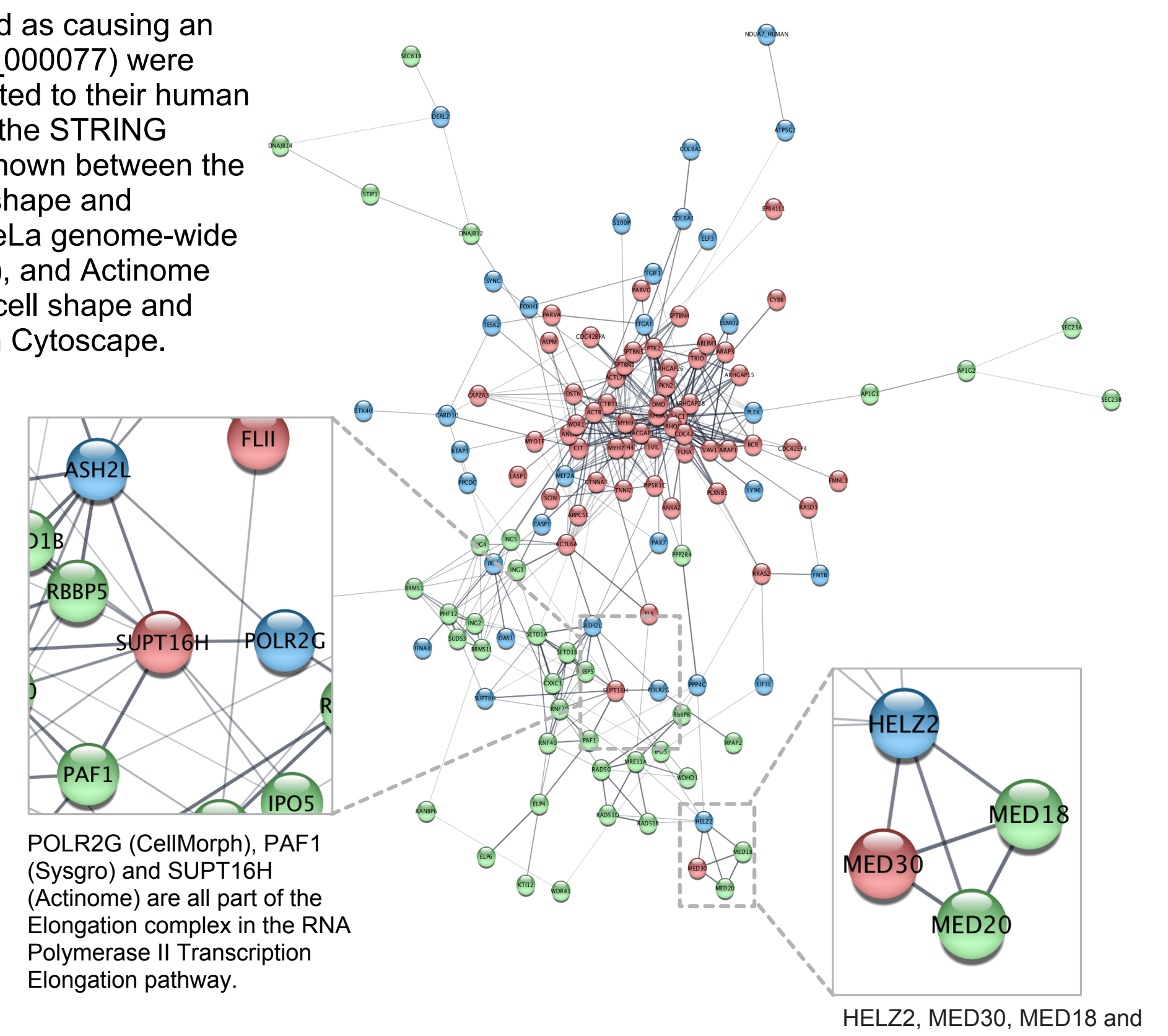
Discovery

Gene networks

Gene mutants or siRNAs that scored as causing an "elongated cell phenotype" (CMPO_000077) were compared, where necessary converted to their human orthologue, and then used to query the STRING database. Network connectivity is shown between the Sysgro (S. pombe, changes in cell shape and microtubules, green), CellMorph (HeLa genome-wide screen, changes in cell shape, blue), and Actinome (HeLa targeted screen, changes in cell shape and cytoskeleton, red) as visualised with Cytoscape.

The genes discovered in the three studies form interconnected, mostly non-overlapping, complementary networks. Since the three studies used different reporters and biological systems, they revealed different aspects of the control mechanisms of cell shape.

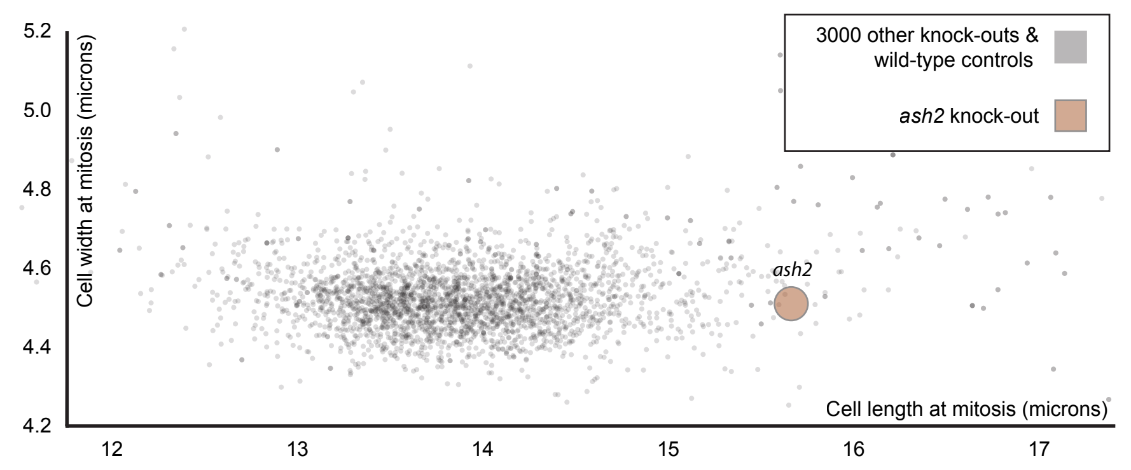
Images from http://string-db.org/ using http://www.cytoscape.org/



POLR2G (CellMorph), PAF1 (Sysgro) and SUPT16H (Actinome) are all part of the Elongation complex in the RNA Polymerase II Transcription Elongation pathway.

HELZ2, MED30, MED18 and MED20 are all part of the PPARG:FattyAcid:RXRA: Mediator:Coactivator Complex. HELZ2 was an elongated cell shape hit in the human HeLa CellMorph screen, MED30 was an hit in the human HeLa Actinome screen and MED18 and MED20 were hits in the *S. pombe* screen Sysgro.

ash2 knock-out leads to abnormal cell elongation in *S. pombe*



Mineotaur.org - Combining data from different screens can also drive the re-assessment of existing published studies. For example, ASH2L is in the elongated cell network based on data from CellMorph, yet the *S. pombe* ortholog of this gene, *ash2*, was not annotated as related to cell elongation and indeed *ash2* was not identified in that screen as a "hit" in terms of cell shape regulation. However, by inspecting the original cell length feature data extracted from *S. pombe* cell populations knocked out for the *ash2* gene, we found that these cells are actually elongated compared with wild-type cells.

Next-Gen analysis

Custom analyses can be performed against the IDR as well. Here, an IPython Notebook has been used to score the similarity between the siRNA treated wells using recomputed features of the raw images.

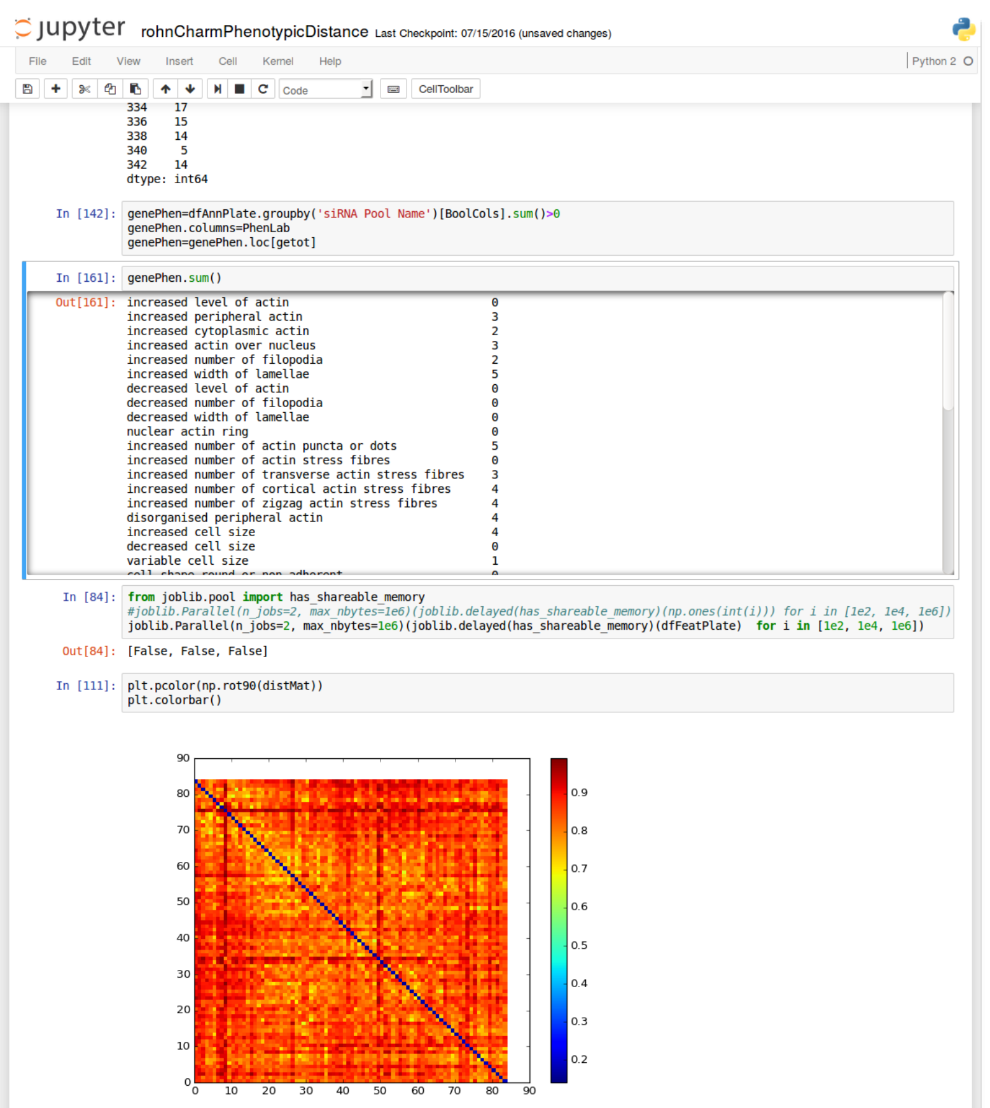
A Jupyter instance is run in the OpenStack cloud with read-only access to the IDR. The notebook has access to the image data, thumbnails, metadata annotations as well as pre-calculated features stored in HDF5.

Currently, wnd-charm features are being pre-generated for a number of studies.

For more information, see:

 * https://github.com/IDR/jupyter-docker
 * https://github.com/wnd-charm/wnd-charm

Public access to this computational facility is being planned and interested parties should feel free to contact us.



rohnCharmPhenotypicDistance Last Checkpoint: 07/15/2016 (unsaved changes)

```
334    17
336    15
338    14
340     5
342    14
dtype: int64

In [142]: genePhen=dfAnnPlate.groupby('siRNA Pool Name')[BoolCols].sum()>0
          genePhen.columns=PhenLab
          genePhen=genePhen.loc[getot]

In [161]: genePhen.sum()

Out[161]: increased level of actin                              0
          increased peripheral actin                            3
          increased cytoplasmic actin                           2
          increased actin over nucleus                          3
          increased number of filopodia                         2
          increased width of lamellae                           5
          decreased level of actin                              0
          decreased number of filopodia                         0
          decreased width of lamellae                           0
          nuclear actin ring                                    0
          increased number of actin puncta or dots              5
          increased number of actin stress fibres               0
          increased number of transverse actin stress fibres    3
          increased number of cortical actin stress fibres      4
          increased number of zigzag actin stress fibres        4
          disorganised peripheral actin                         4
          increased cell size                                   4
          decreased cell size                                   0
          variable cell size                                    1

In [84]: from joblib.pool import has_shareable_memory
         #joblib.Parallel(n_jobs=2, max_nbytes=1e6)(joblib.delayed(has_shareable_memory)(np.ones(int(i))) for i in [1e2, 1e4, 1e6])
         joblib.Parallel(n_jobs=2, max_nbytes=1e6)(joblib.delayed(has_shareable_memory)(dfFeatPlate)  for i in [1e2, 1e4, 1e6])

Out[84]: [False, False, False]

In [111]: plt.pcolor(np.rot90(distMat))
          plt.colorbar()
```

The IDR was made possible by funding from:

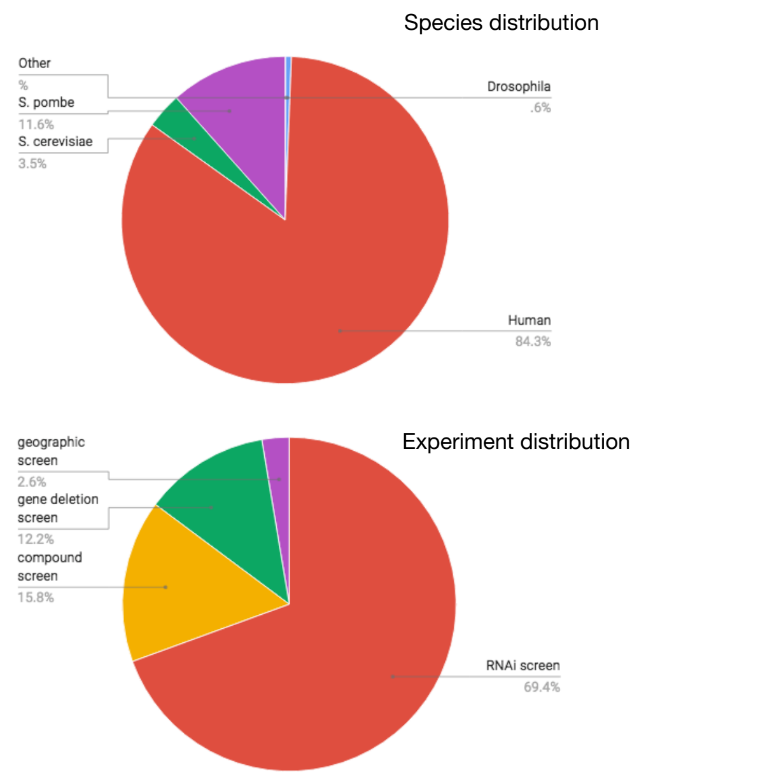
bbsrc biotechnology and biological sciences research council

Horizon 2020 European Union funding for Research & Innovation

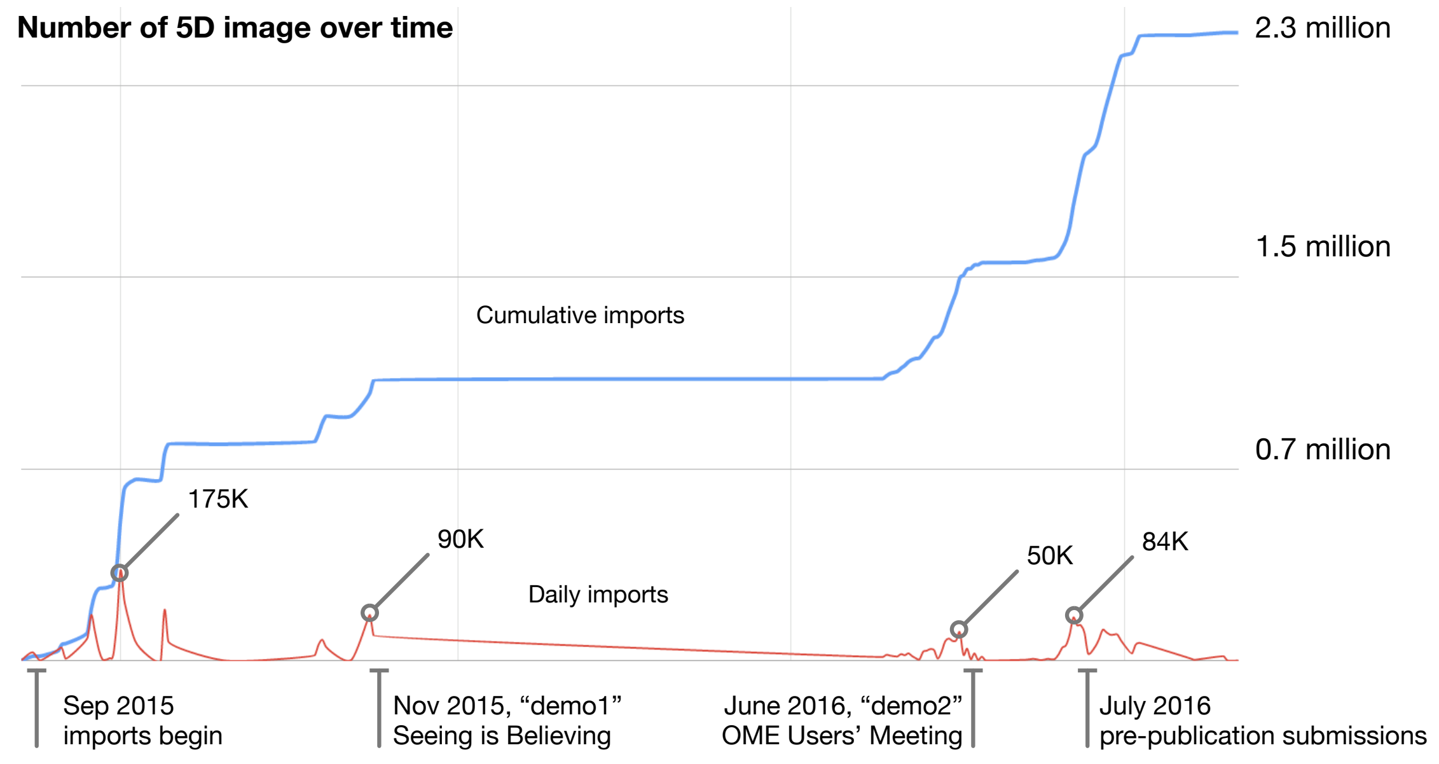
wellcome trust

Current status

Species	Type	5D Images	2D Planes	Size (TB)	Phenotype count (avg)	Term count (avg)	Target count (e.g. genes)	Experiment count
Drosophila	RNAi screen	90,330	184,782	0.22	9.33	9.33	26054	37250
Human	RNAi screen	683,200	20,275,782	19.75	10.67	11.44	75053	697249
Human	compound screen	1,017,276	4,644,012	5.66	1.00	1.00	30823	180864
Human	high content image analysis	25,872	77,616	0.03	0.00	0.00	198	2156
Human	protein localization screen	240,848	481,696	1.40	8.00	8.00	12744	15547
Human	protein localization using 3D-SIM	414	935	0.00	1.00	1.00	9	414
Human	protein localization using dSTORM	524	106,085	0.00	1.00	1.00	7	362
Mus musculus	histopathology of gene knockouts	899	2,237	0.27	48.00	48.00	9	230
S. cerevisiae	3D-tracking of tagged chromatin loci	229	697,100	0.00	0.00	0.00	8	112
S. cerevisiae	gene deletion screen	3,765	75,308	0.17	1.00	1.00	4195	4272
S. cerevisiae	protein localization screen	3,456	6,912	0.02	23.00	7.00	377	1131
S. cerevisiae	protein screen	97,920	293,760	0.20	14.00	11.00	6234	31170
S. pombe	gene deletion screen	109,728	3,511,296	10.06	19.00	21.00	3006	17270
multi-species	geographic screen	7,362	777,725	0.61	0.00	0.00	84	84
Total		**2,281,823**	**31,135,246**	**38.39**	**9.71**	**8.56**	**11342.92857**	**988111**



Number of 5D image over time



Sourcing of datasets began for the IDR with the project inception, early 2015. In the roughly 12 months since data submissions began, more than **2 million** 5D images have been imported. These images, largely from high-content screens, represent over **30 million** individual 2D planes.

Most studies were previously published but the data was not available online. Capacity exists for growing the **40 terabytes** of imaging data ten-fold, with the intent of increasing that further if interest exists. The primary goal is to enable further discoveries among the **1 million** experiments represented by the data.

If interested in submitting data or performing re-analysis, please contact idr-submission@lists.openmicroscopy.org.uk

http://j.mp/idr0008
image-123004

http://idr-demo.openmicroscopy.org