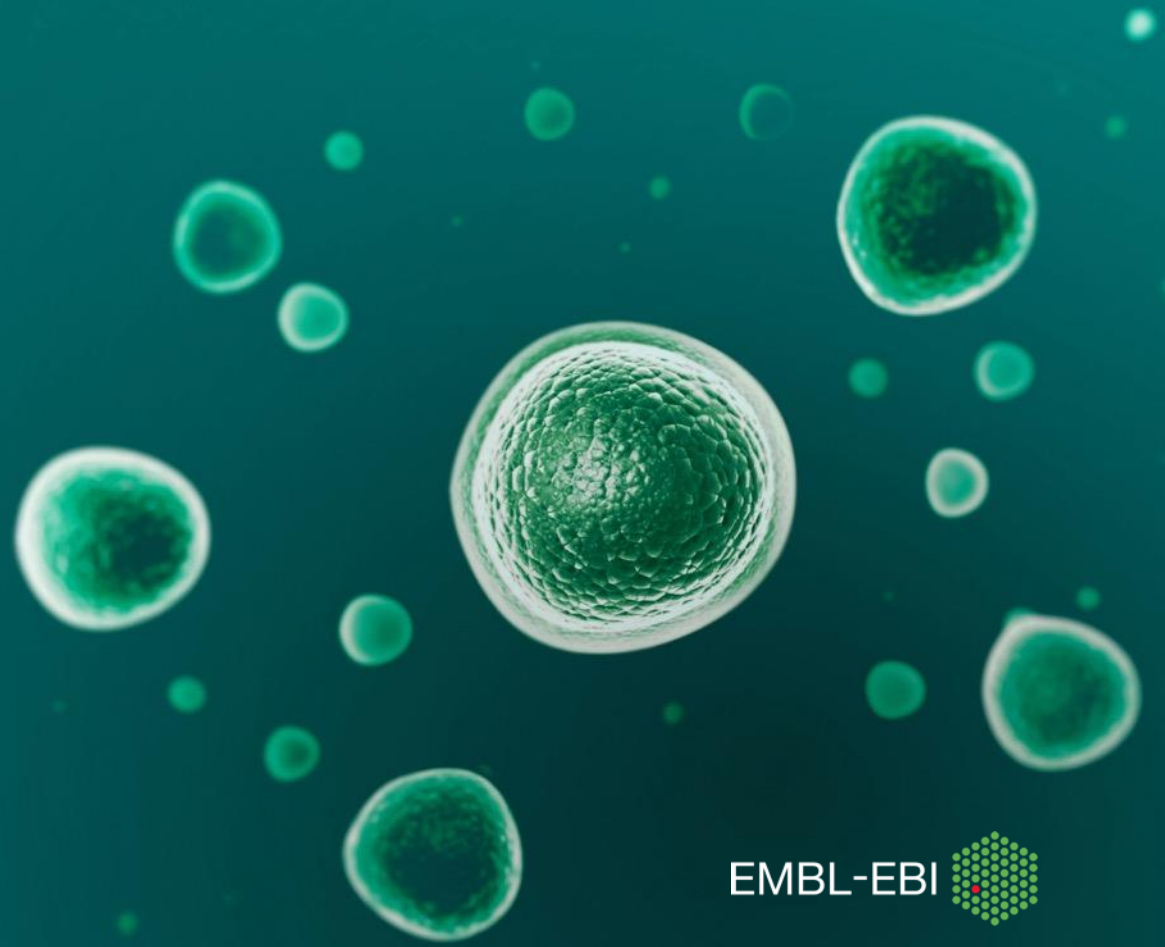
Bioimage data ecosystem

Ugis Sarkans

European Bioinformatics Institute (EMBL-EBI)



EMBL-EBI

What is EMBL-EBI?

- Europe's home for biological data services, research and training
- A trusted data provider for the life sciences
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 650 members of staff from 66 nations
- Home of the ELIXIR Technical hub.

The European Molecular Biology Laboratory

80+ nationalities

>1700 personnel

6 sites in Europe

Heidelberg, Germany

Main Laboratory

Tissue Biology, Disease Modeling

Barcelona, Spain

Hinxton, Cambridge, UK

Bioinformatics

Mouse Biology

Rome, Italy

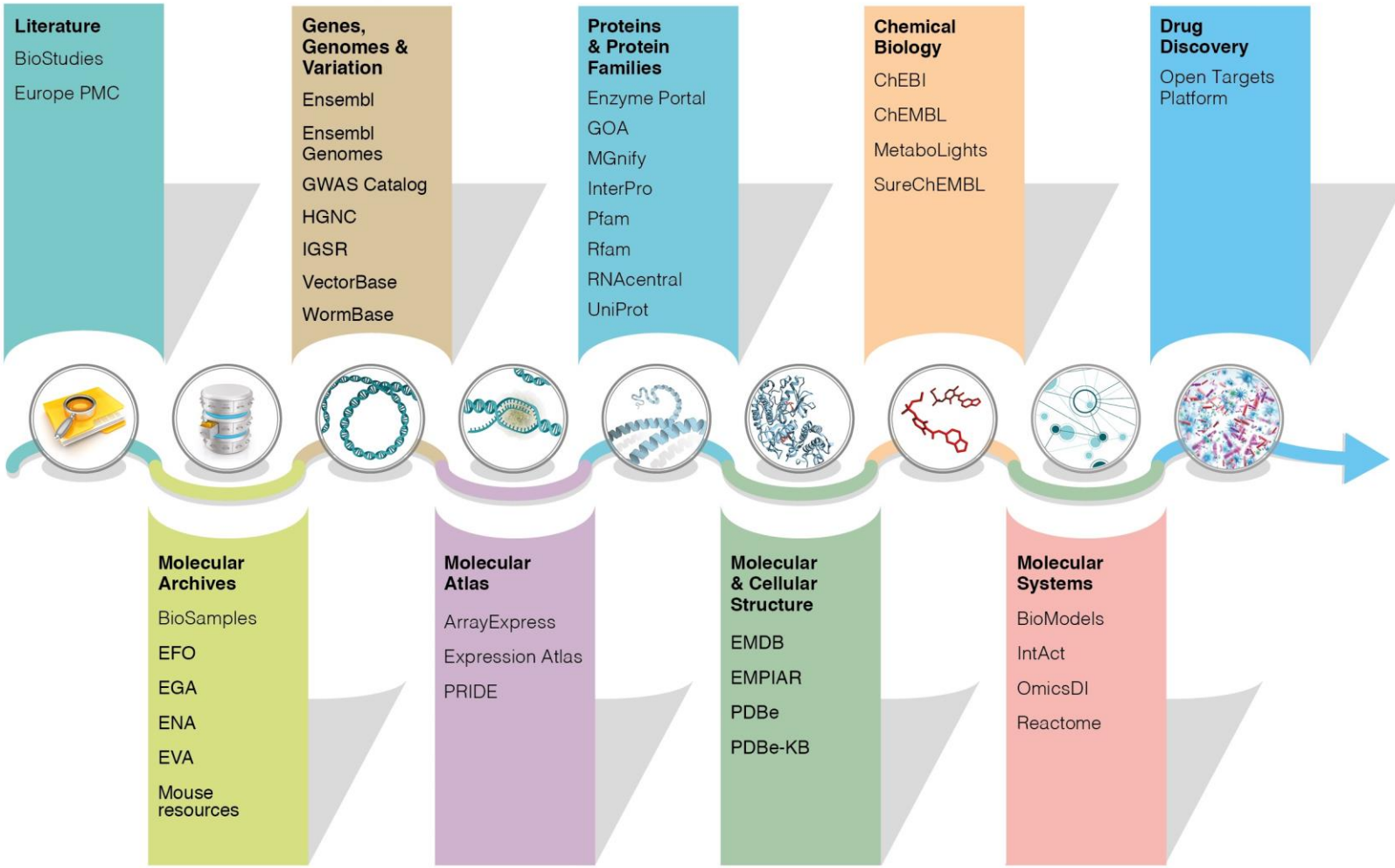
Grenoble, France

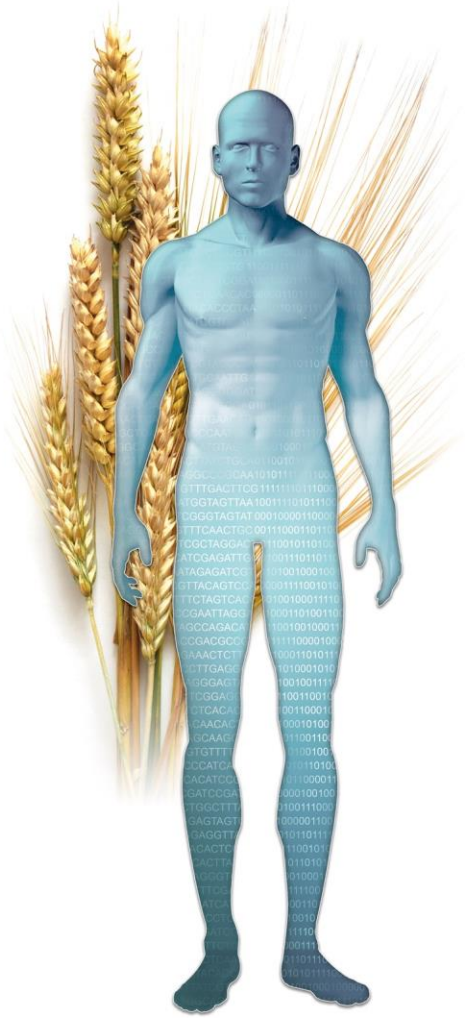
Structural Biology

Structural Biology

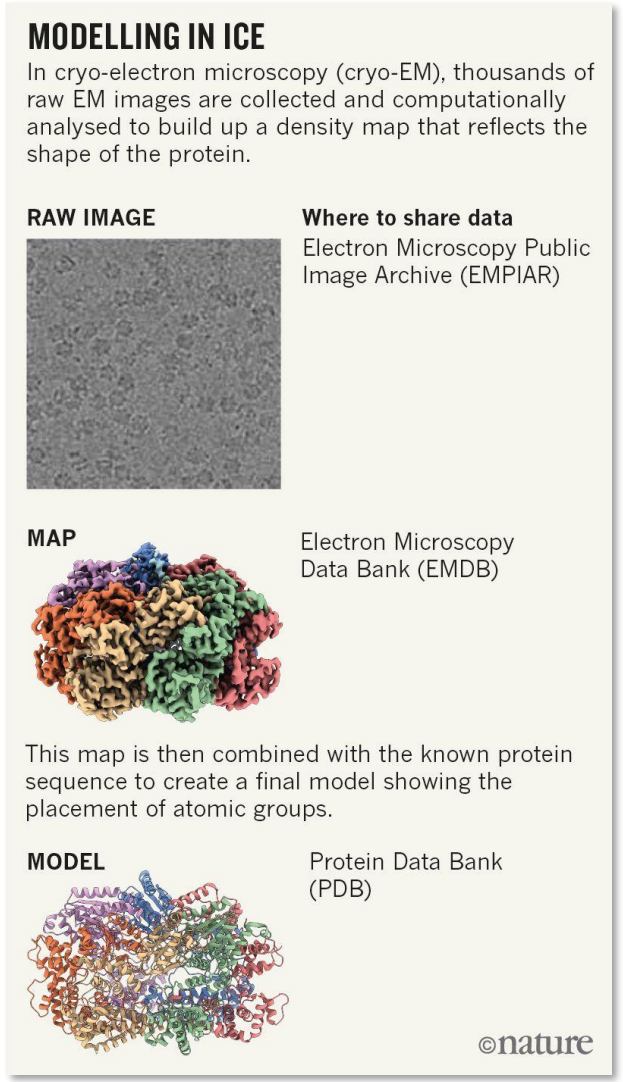
Hamburg, Germany

Data resources at EMBL-EBI





EMPIAR – EM Public Image Archive



- Established as a pilot in 2014 at EMBL-EBI
  - Community-driven initiative
- Raw 2D image data (of cryo-EM/Tomography EMDB entries) and 3D volume maps (Soft X-ray Tomography, Volume EM, …)
- 229 entries and >160 TB data
- Use: validation, teaching/training, software & methods development, community challenges, …
- Establishing new international collaborations
  - Mirror in Japan: empiar.pdbj.org
- Part of EMBL-EBI BioImage Archive

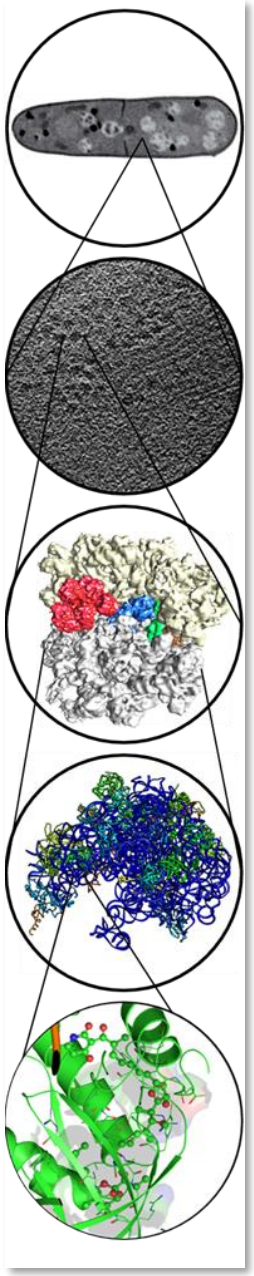
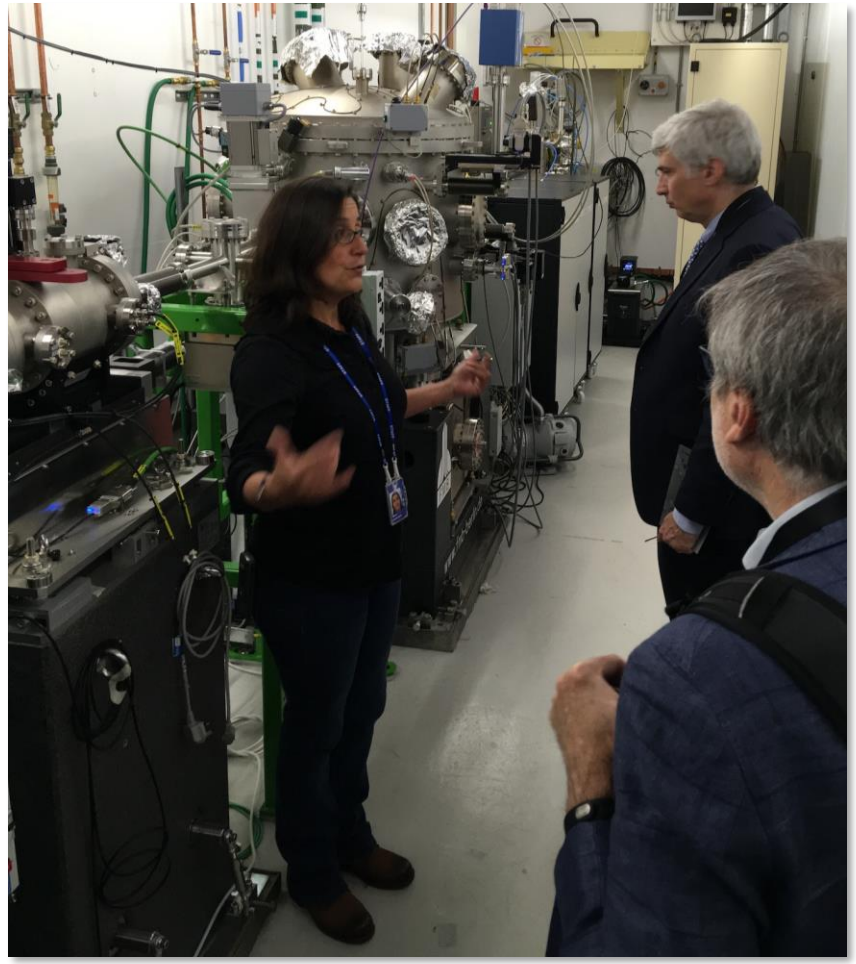


Image credit (left): *Nature* **561**, 565-567 (2018)

EMBL-EBI

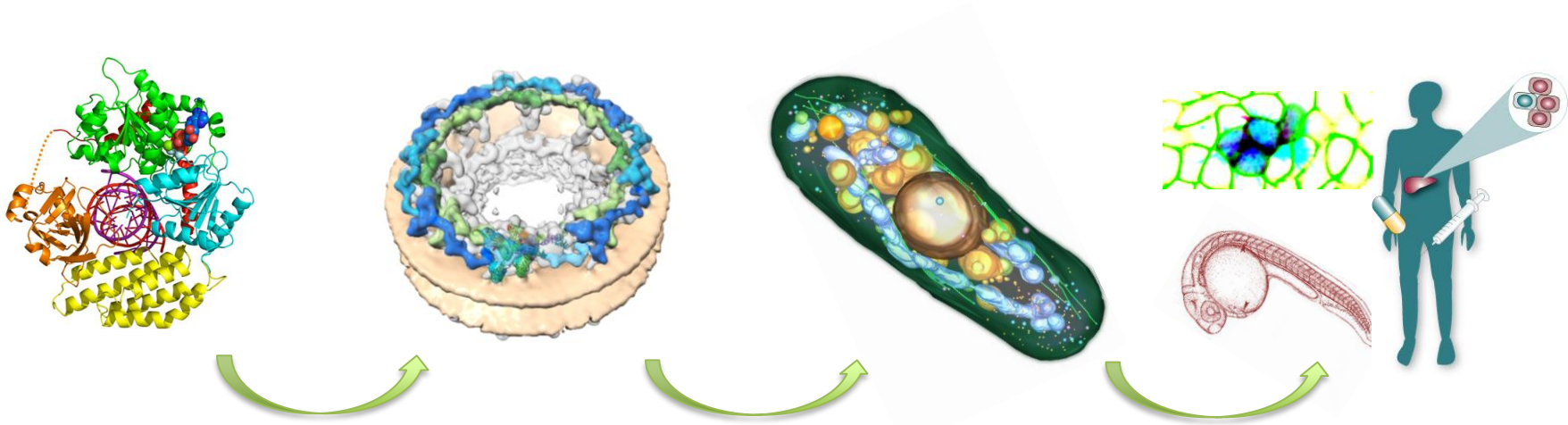
EMPIAR accepts more data types

- Raw data associated with an EMDB entry
- 2D/3D data from 3D imaging modalities not covered by EMDB (e.g., 3DSEM and SXT →)
- 2D EM data used in integrative/hybrid methods, associated with a structure deposited in the PDB or PDB-Dev archive
- Certain reference and benchmark datasets (to be decided on a case-by-case basis)
- Datasets used for certain community challenges (such as the 2015 Map Validation Challenge)
- Soon: EM/XM parts of CLEM/CLXM experiments

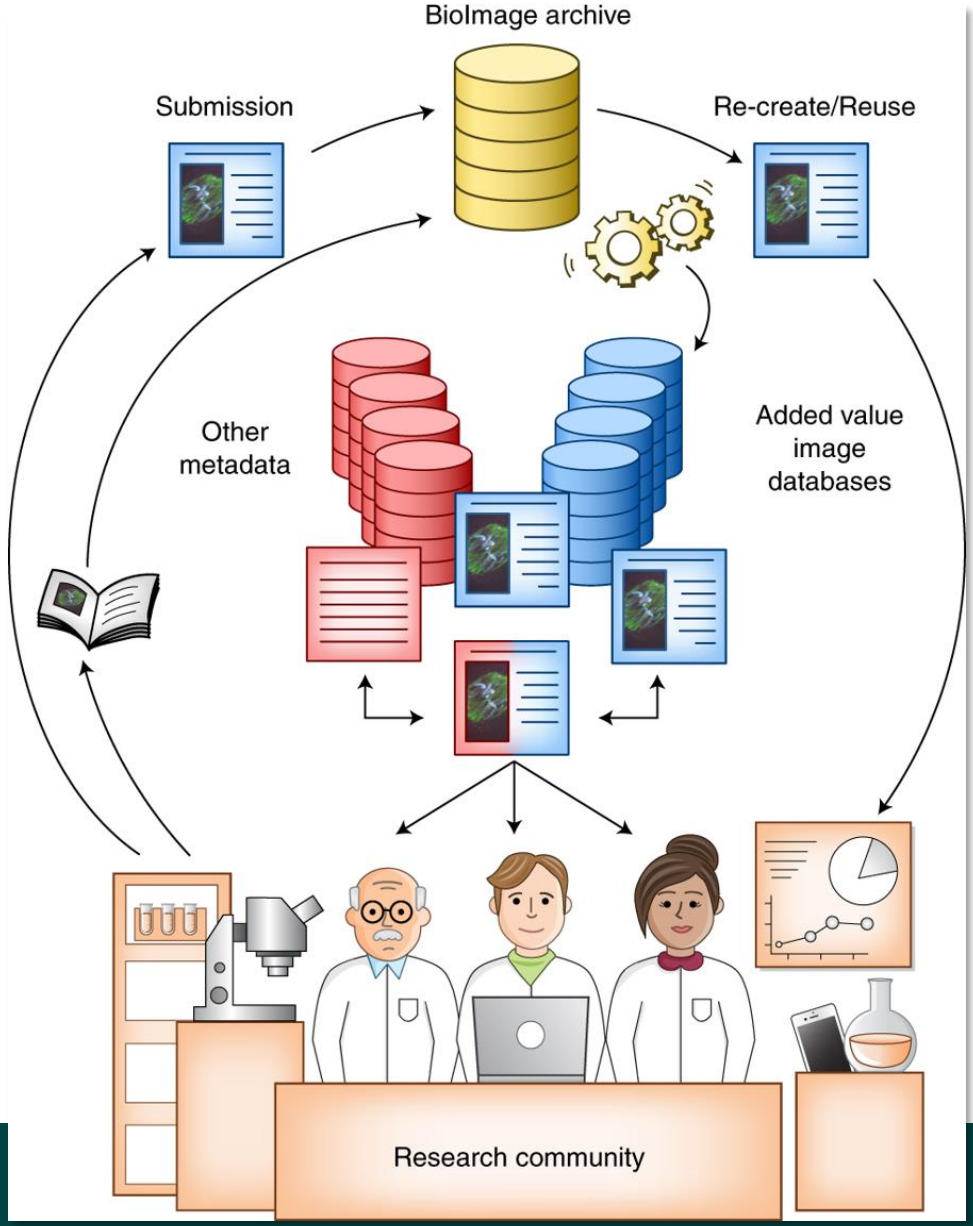


EMBL-EBI BioImage Archive

- **Advances in imaging** technologies → opportunities for research and biomedical insights
- **Open sharing** of reference image data will be important (similarities with sequencing)
- Need for a **central secure storage** for open image datasets



BioImage Archive – the big picture



BioImage Archive - scope

- data **related to publications** (in particular, figures)
  - enable re-use, support integrity and reproducibility of research
- **reference images**
  - large systematic datasets
    - e.g. systematic characterization of protein expression patterns
  - value beyond single experiment

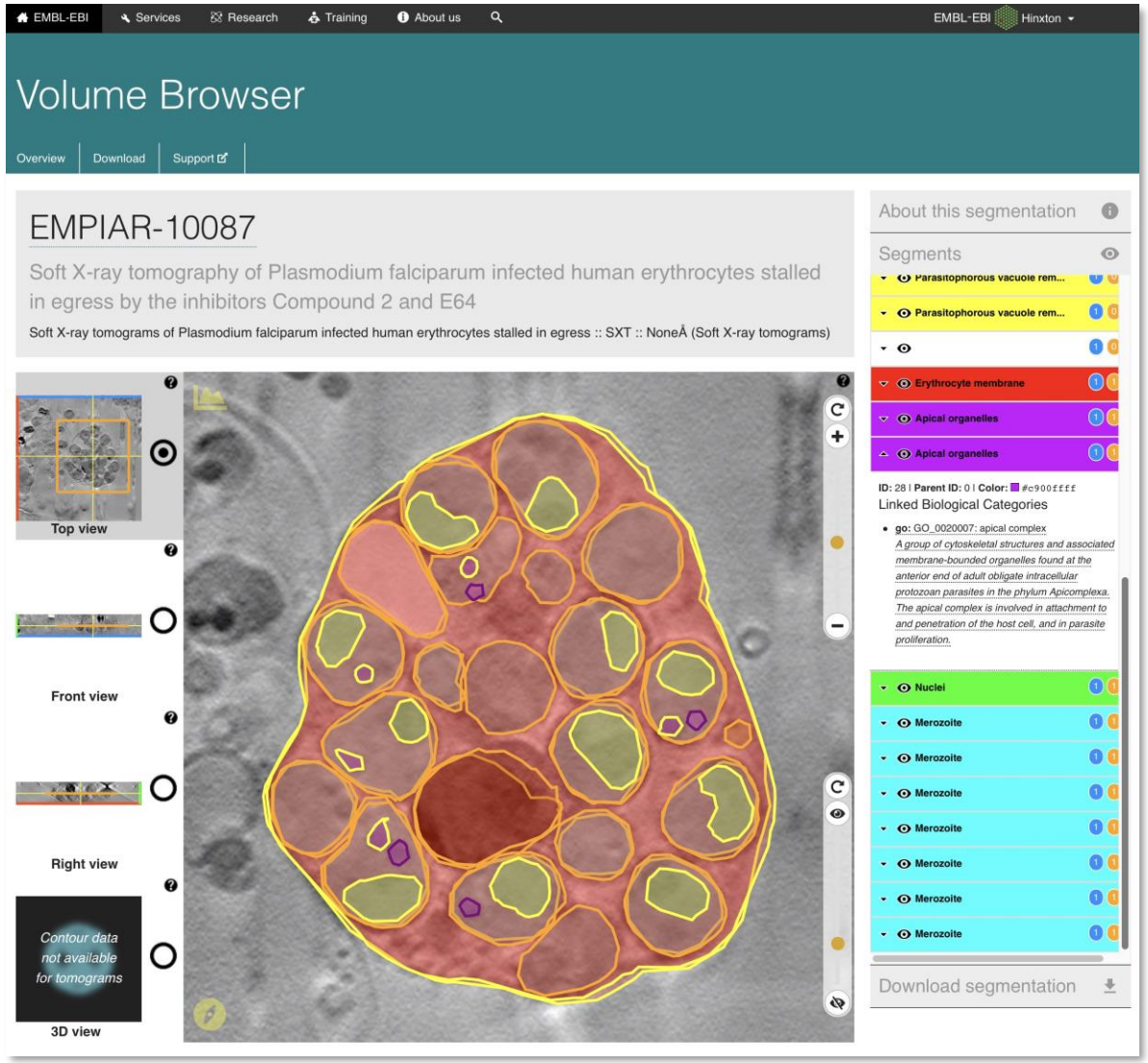
Data archive

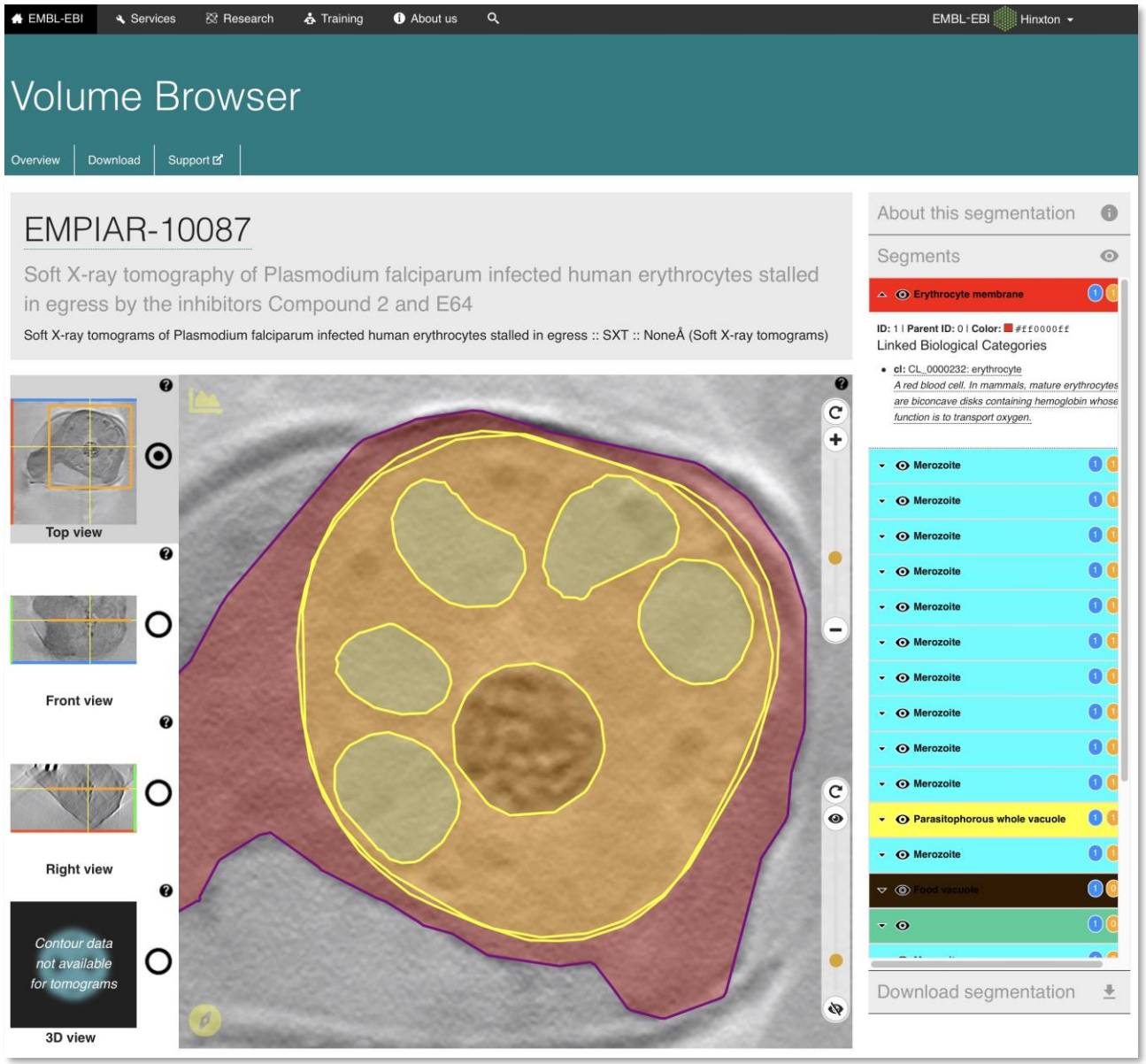
- store experimental data and supporting metadata – **scientific record**
- **straightforward** submission processes
- **searchable** and accessible
- **facilitate standards** development, and adopt existing standards
- **no data value judgements**, no data integration

Added value database

- synthetic: enrich and combine datasets
- curation
- ontologies
- additional QC
- reprocessing
- meta-analysis

Added value: Volume Browser





Try the Volume Browser prototype: https://bit.ly/2Wblo6U

EMBL-EBI

BioImage Archive

Home | Browse | About us | Our roadmap | Case studies | FAQs

The BioImage Archive stores and distributes biological images that are useful to life science researchers. It's development will provide data archiving services to the broader bioimaging database community, such as EMPIAR, Cell-IDR and Tissue-IDR, who provide added value and curation of selected datasets. Read more ›

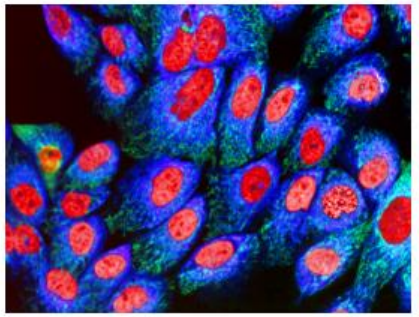
Submissions

There are multiple routes for submission of images to the archive. Depending on the type of data you have, choose one of the resources below.

Scope	Technologies	Biological Samples and Models	Data resource
Molecules to cells	Electron cryo-microscopy (EM)*, Electron cryo-tomography*, Scanning EM (FIB SEM, SBF SEM, array tomography), Soft X-ray tomography, X-ray microscopy	Macromolecular structure by single-particle methods, high-resolution subcellular organelles and structure	EMPIAR
Cells	Wide field, TIRF Confocal, Quantitative Phase Microscopy, High content screening	Single molecule (LM), Subcellular and cellular structure architecture, dynamics; Single cell macromolecule localization; Cellular response to small molecule- or genome-wide perturbation	Cell-IDR
Tissues	Light-Sheet Microscopy, Multiphoton Microscopy, Digital Pathology, Optical Projection Tomography, Micro-CT Mass cytometry	Tissue structure, molecular localization and dynamics, Pathology and histology sections.	Tissue-IDR
All biological images which do not fit into any of the resources above	Magnetic resonance imaging, Ultrasound imaging, Super-resolution microscopy, photoacoustic imaging, light-field microscopy, Brillouin microscopy, Atomic Force Microscopy, other	Non-human data (human medical data is out of scope)	BioImage Archive

Further information

A paper detailing the BioImage Archive was featured in Nature Methods in November 2018. Read more ›



Examples of bioimages shared and reused. Read more ›

EMBL-EBI

Submissions

There are multiple routes for submission of images to the archive. Depending on the type of data you have, choose one of the resources below.

Scope	Technologies	Biological Samples and Models	Data resource
Molecules to cells	Electron cryo-microscopy (EM)*, Electron cryo-tomography*, Scanning EM (FIB SEM, SBF SEM, array tomography), Soft X-ray tomography, X-ray microscopy	Macromolecular structure by single-particle methods, high-resolution subcellular organelles and structure	EMPIAR
Cells	Wide field, TIRF Confocal, Quantitative Phase Microscopy, High content screening	Single molecule (LM), Subcellular and cellular structure architecture, dynamics; Single cell macromolecule localization; Cellular response to small molecule- or genome-wide perturbation	Cell-IDR
Tissues	Light-Sheet Microscopy, Multiphoton Microscopy, Digital Pathology, Optical Projection Tomography, Micro-CT Mass cytometry	Tissue structure, molecular localization and dynamics, Pathology and histology sections.	Tissue-IDR
All biological images which do not fit into any of the resources above	Magnetic resonance imaging, Ultrasound imaging, Super-resolution microscopy, photoacoustic imaging, light-field microscopy, Brillouin microscopy, Atomic Force Microscopy, other	Non-human data (human medical data is out of scope)	BioImage Archive

Image Data Resource Vision

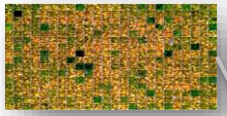
- Public access
- Reference datasets - complete datasets, associated with an existing or upcoming publication.
- Study integration - integrating studies or datasets with other datasets, e.g., through genes or phenotypes.
- Curated metadata
- Cloud re-analysis

https://idr.openmicroscopy.org

The IDR @ EBI Embassy

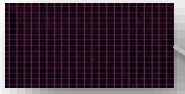
https://idr.openmicroscopy.org

Gene Product Targeting HCS

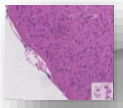


Genetic HCS

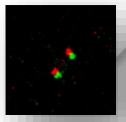
Geographic HCS



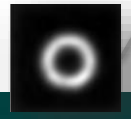
Chemical HCS



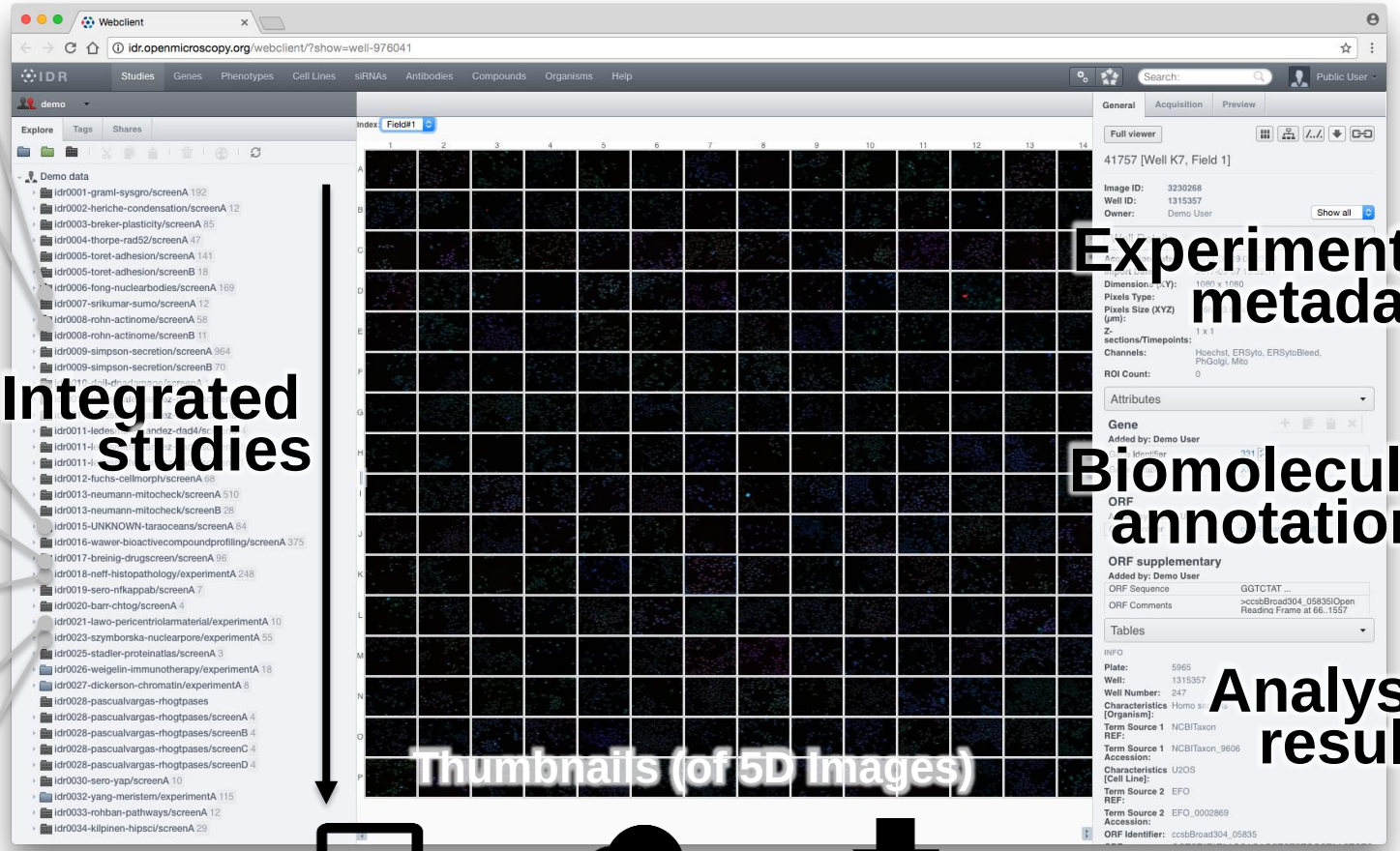
Histopathology



3D-Sim



Super-resolution



Integrated studies

Experimental metadata

Biomolecular annotations

Analysis results

Cross-data browsing

Cloud analysis

Download (local analysis)

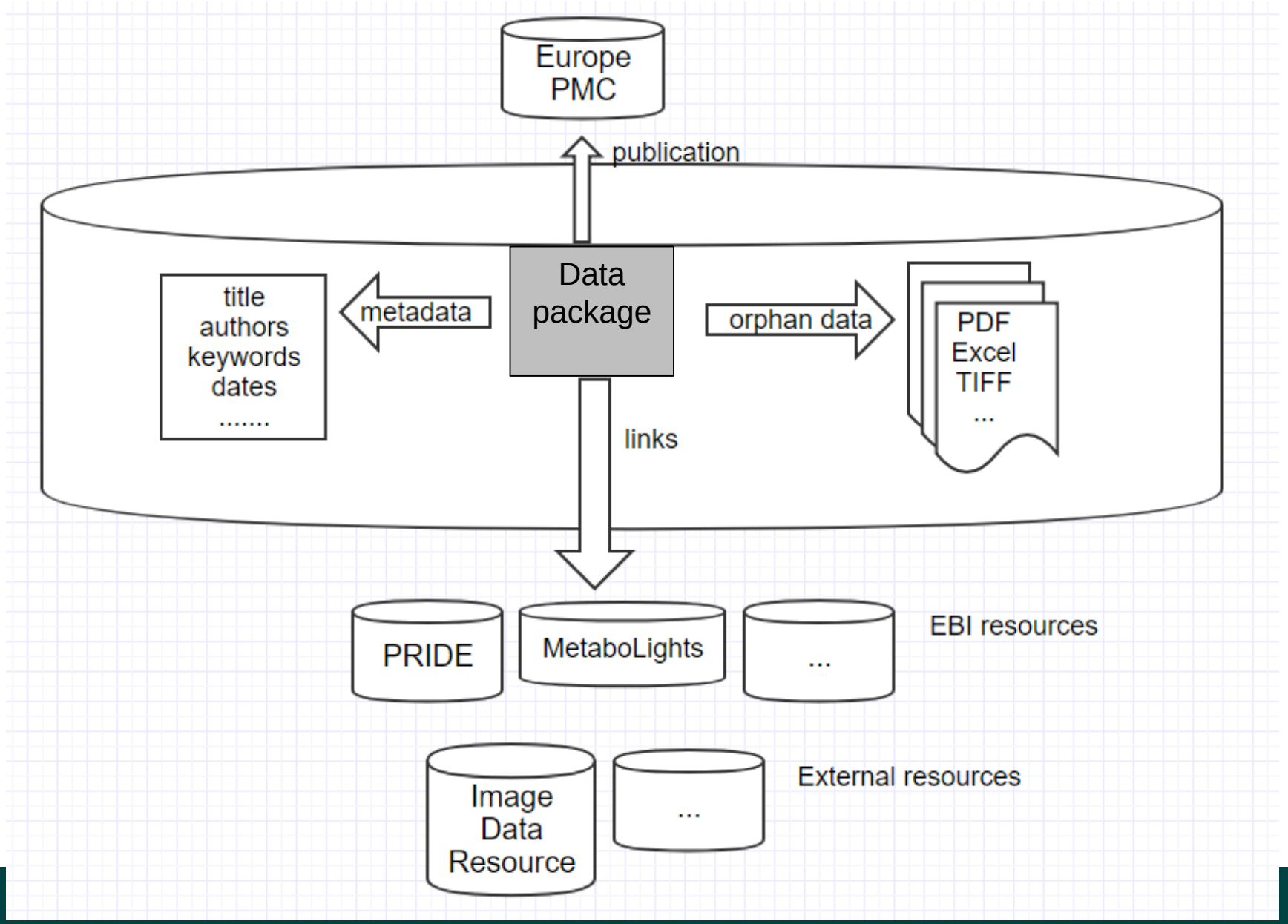
Slide: Francis Wong, IDR team

BioStudies database: aim

Facilitate **transparency and reproducibility of research**, by aggregating all the outputs of a study, the **data package**, in a single place

Latest publication:
Sarkans et al. The BioStudies database—one stop shop for all data supporting a life sciences study. *Nucleic Acids Research*, 2018, gkx965, https://doi.org/10.1093/nar/gkx965

A data package



BioImage Archive – challenges

- Data volumes
- Streamlined data and metadata acquisition
- Useful data delivery
- Useful, sufficient metadata

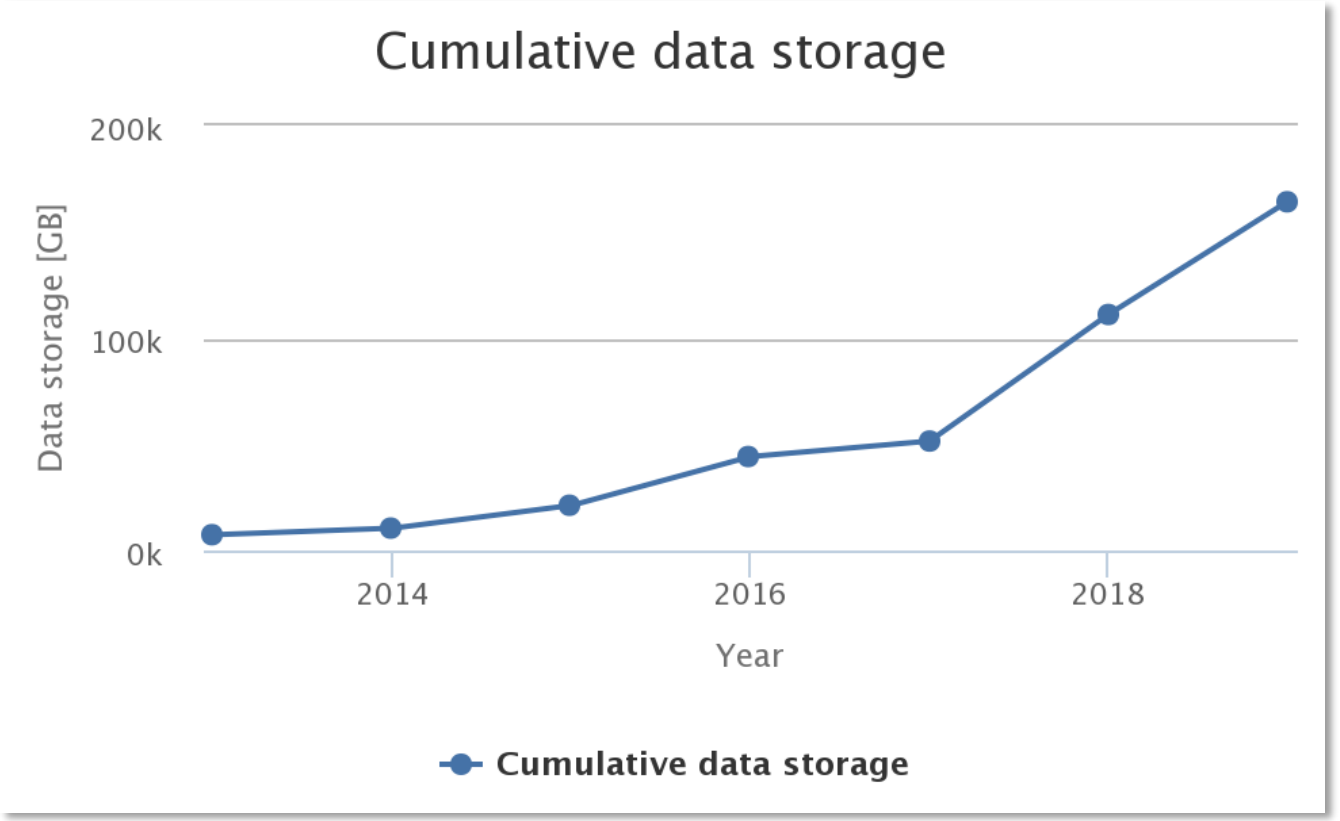
BioImage Archive – ongoing work

- Technical work – scalability
  - use object storage system instead of NFS
- User-facing work
  - develop scalable data deposition infrastructure
    - jointly work with IDR team
  - data access – exploration and bulk downloads/cloud access

Moving EMPIAR to the petabyte scale

- Average EMPIAR entry ~0.7 TB
- If all EMDB depositions would be accompanied by raw data in EMPIAR → 1-2 PB per year
- NFS storage systems do not scale

- Move to object storage
- Infrastructure provided through EMBL-EBI BioImage Archive



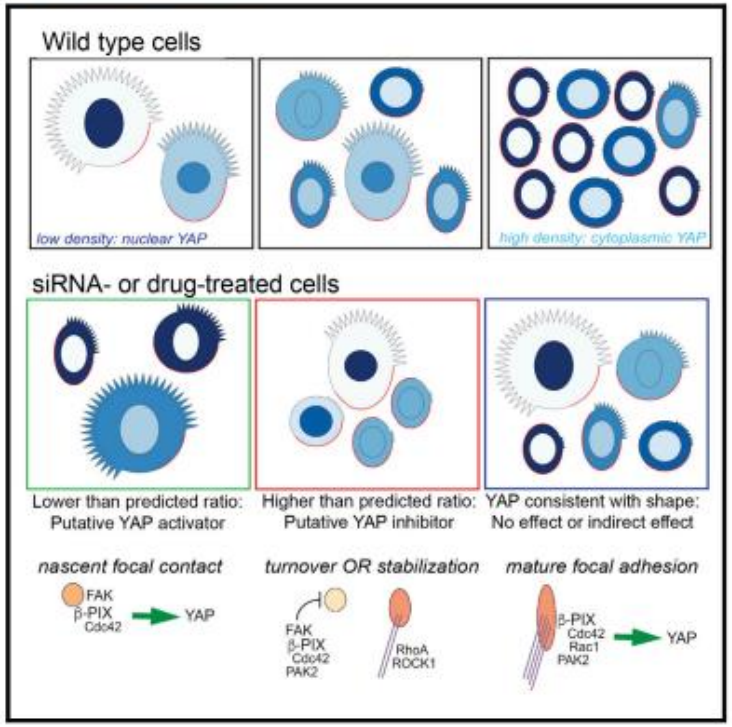
https://en.wikipedia.org/wiki/Object_storage

Article

Cell Systems

Multiparametric Analysis of Cell Shape Demonstrates that β-PIX Directly Couples YAP Activation to Extracellular Matrix Adhesion

Graphical Abstract



Authors

Julia E. Sero, Chris Bakal

Correspondence

juliasero@post.harvard.edu

In Brief

A statistical model is used to control for the effects of cell shape on protein localization in an RNAi screen, revealing regulators of the transcriptional coactivator Yes-associated protein (YAP). This work illustrates how the naturally occurring heterogeneity of cellular populations can be harnessed to distinguish indirect from direct effects.

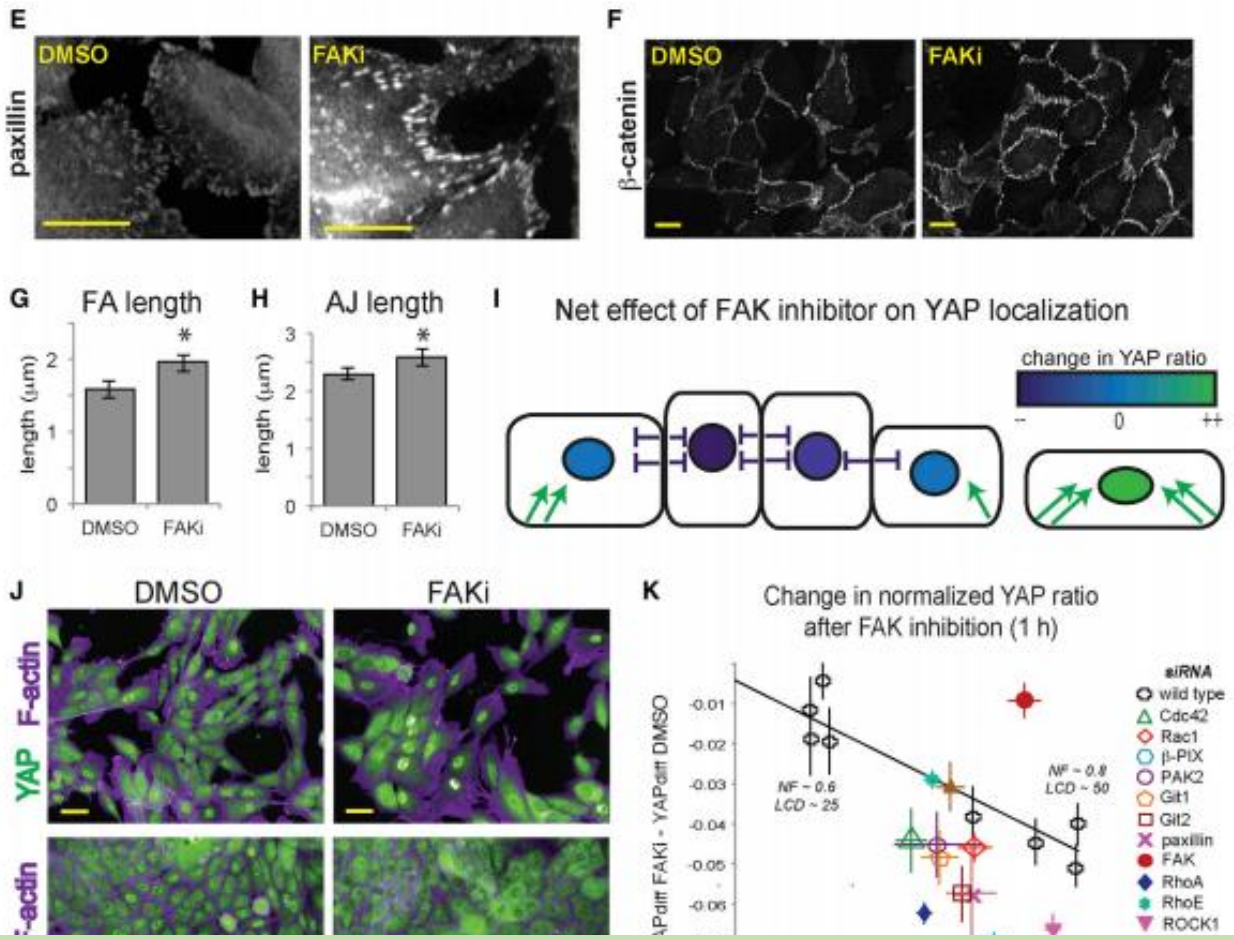
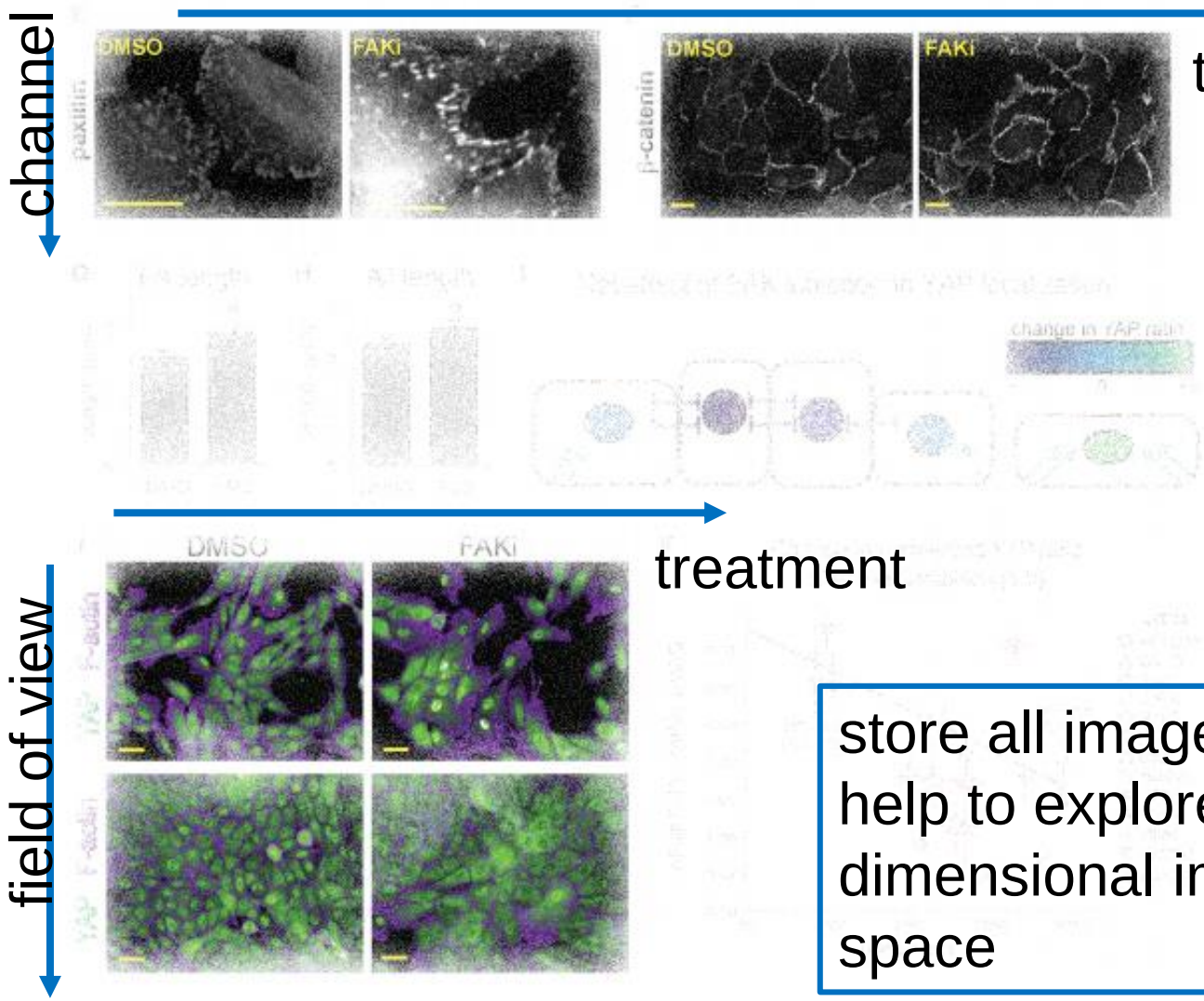


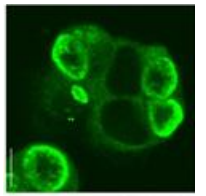
Figure from: Sero JE, Bakal C. Multiparametric Analysis of Cell Shape Demonstrates
that β-PIX Directly Couples YAP Activation to Extracellular Matrix Adhesion

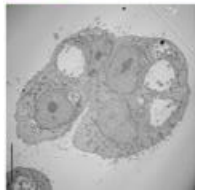


store all image data;
help to explore the multi-dimensional image space

Archiving CLEM/CLXM data in the BioImage Archive

- BioStudies
  - Original 2D or 3D LM data
  - Description of image-registration process
  - ec-CLEM metadata XML file (or list of landmark pairs)
  - If warping was used, warped image(s) as well

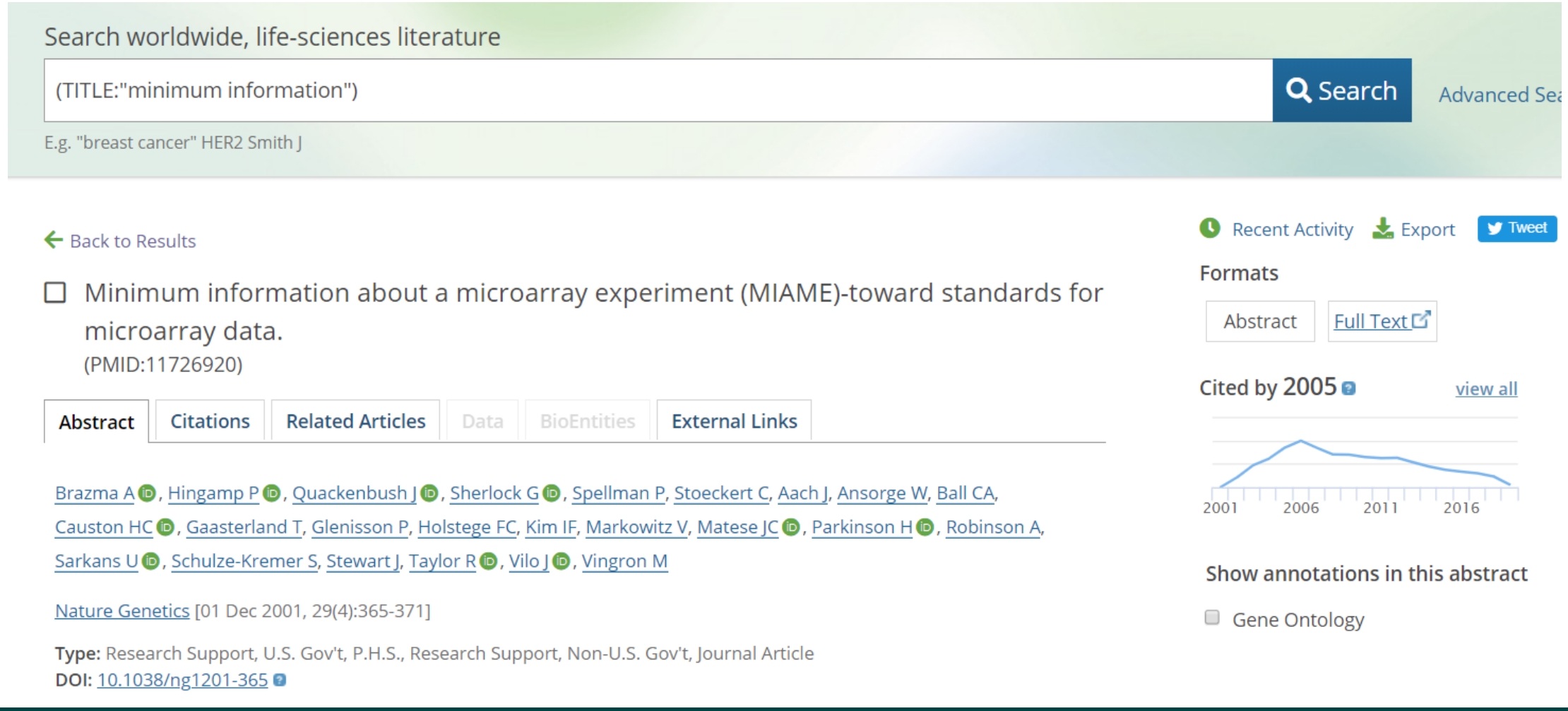




- EMPIAR
  - 2D or 3D EM or SXT data
  - Raw data optional (e.g., tilt series)
  - BioStudies accession ID (for corresponding LM and registration data)

We are looking for (patient) guinea pigs to help us test this…

“Minimum information about imaging dataset” draft development



Thanks!

- Cellular Structure and 3D Bioimaging
  - Ardan Patwardhan, Gerard Kleywegt, Andrii Iudin
- Technical Services Cluster
  - Marc Riera, Sarah Butcher
- Molecular Archives Cluster
  - Helen Parkinson, Terry Meehan
- Strategic Project Management Office
  - Jessica Vamathevan, Mary Barlow
- Molecular Atlas Cluster
  - Alvis Brazma
- Literature Services
  - Jo McEntyre
- BioStudies team

- Swedlow's team (U Dundee)

Funding and support

UK Research and Innovation

EMBL

wellcome

EBSRC 20 Years of Pioneering Great British Bioscience

MRC Medical Research Council

EURO-BIOIMAGING

GLOBAL BIOIMAGING growing collaboration