

The European Bioinformatics Institute

On the way to realising the full potential of PIDs
in BioSciences:
THOR work at EMBL-EBI

Robert Petryszak

Gene Expression Team Leader

EMBL-EBI



What is EMBL-EBI?



- Europe's home for biological data services, research and training
- A trusted data provider for the life sciences – both academia and industry
- Part of the European Molecular Biology Laboratory, an intergovernmental research organisation
- International: 570 members of staff from 57 nations
- Home of the ELIXIR Technical hub

Data resources at EMBL-EBI – THOR

participants

Genes, genomes & variation

European Nucleotide Archive
European Variation Archive
European Genome-phenome Archive
Ensembl
Ensembl
Genomes
GWAS Catalog
Metagenomics portal

Gene, protein & metabolite expression

RNA Central
ArrayExpress
Expression Atlas
Metabolights
PRIDE

Protein sequences, families & motifs

InterPro
Pfam/Rfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

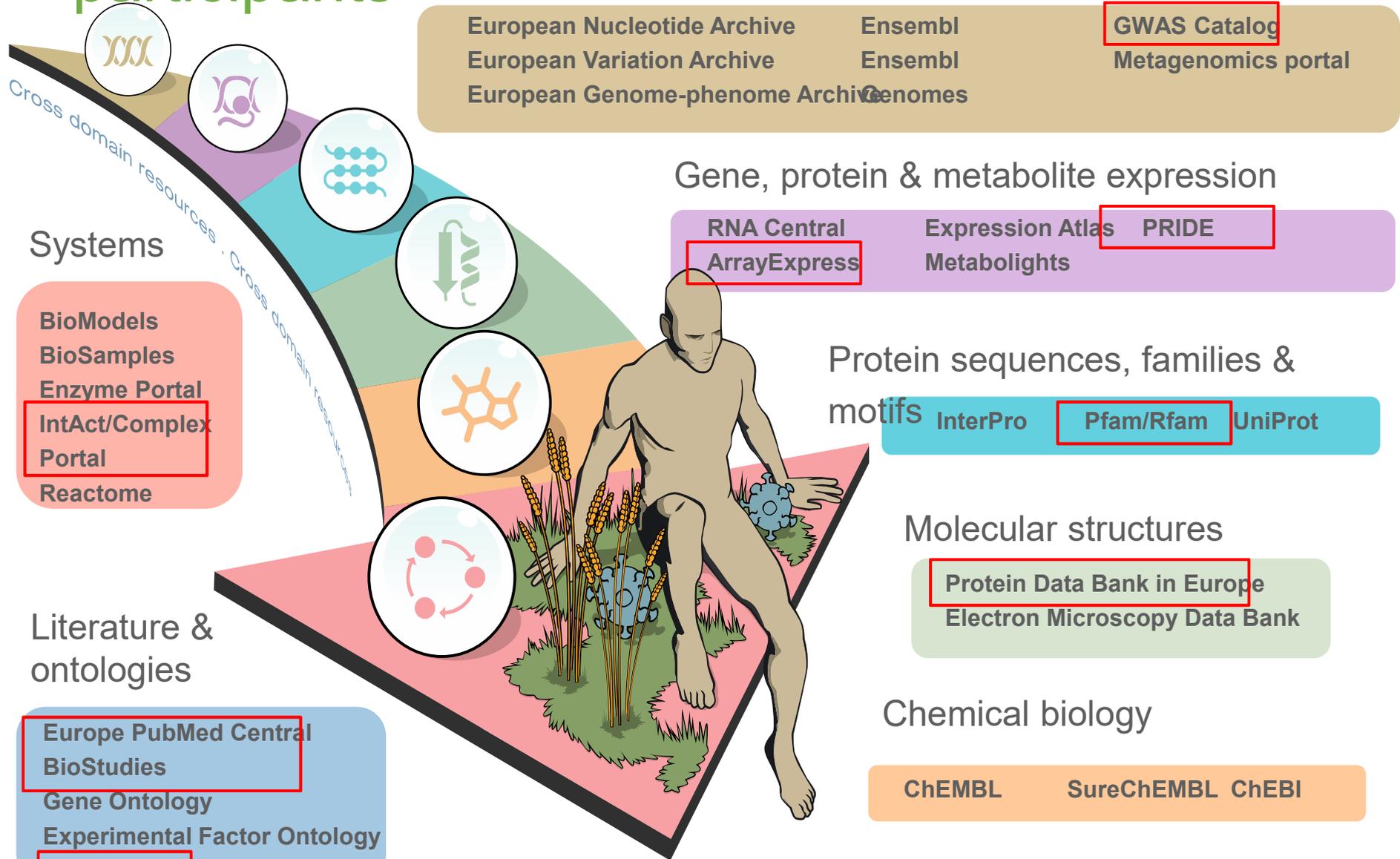
ChEMBL
SureChEMBL
ChEBI

Systems

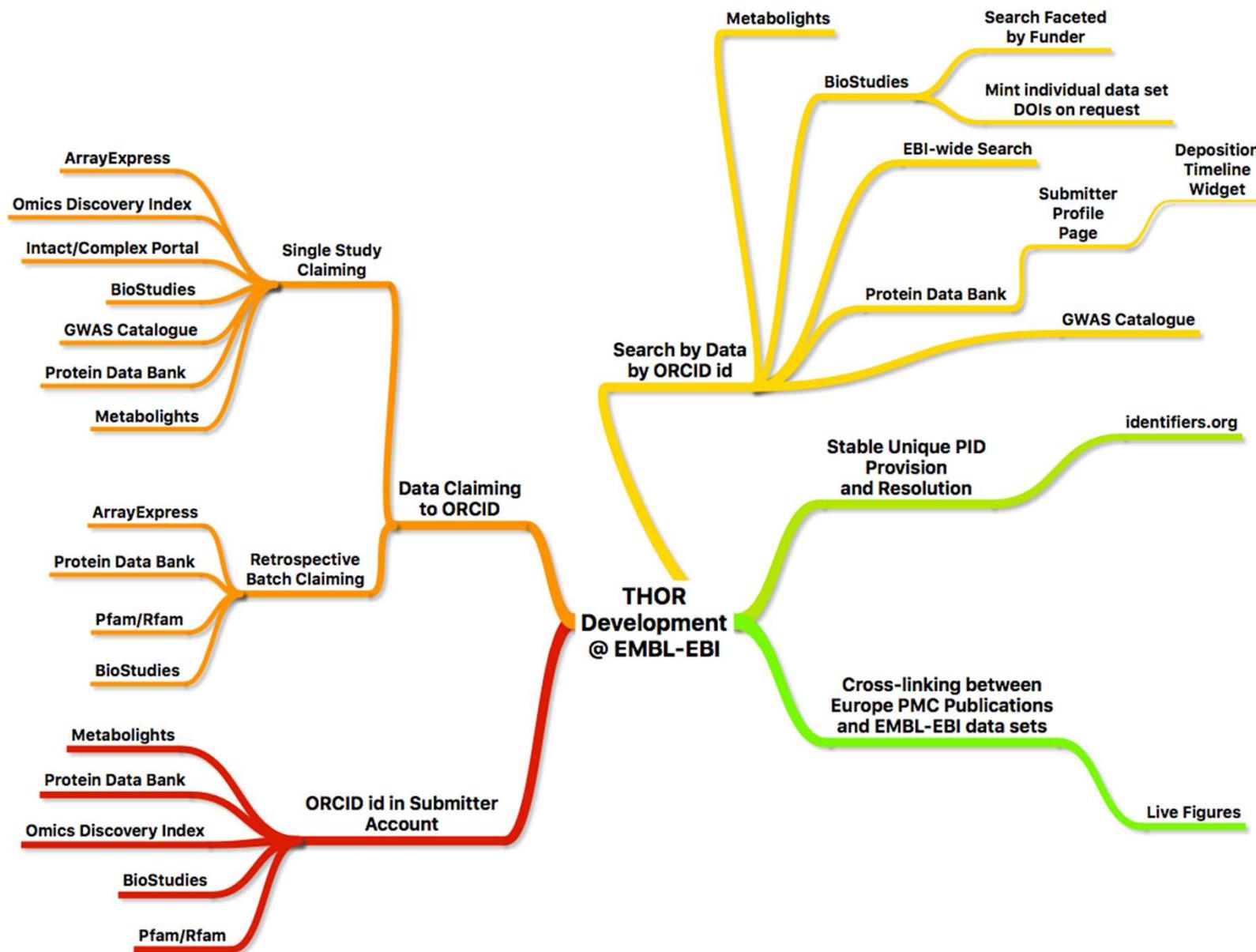
BioModels
BioSamples
Enzyme Portal
IntAct/Complex Portal
Reactome

Literature & ontologies

Europe PubMed Central
BioStudies
Gene Ontology
Experimental Factor Ontology
Identifiers.org



EMBL-EBI THOR Development – At a Glance



Claiming Individual Studies

The image shows a screenshot of the MetaboLights website. The top navigation bar includes 'EMBL-EBI', 'Services', 'Research', 'Training', and 'About us'. The main header features the 'Protein Data Bank in Europe' logo and a search bar. Below this, the 'MetaboLights' logo is prominent, with a search bar and navigation links like 'Home', 'Browse Studies', 'Browse Compounds', etc.

The main content area displays the study 'MTBLS1: A metabolomic study of urinary changes in type 2 diabetes in human compared to the control group'. The study is listed as 'Submitted' in a progress bar. The authors are 'Reza Salek, Jules Griffin'. The submission date is '14-Feb-2012', and the release date is '14-Feb-2012'. The study status is 'Public'.

A red box highlights the 'ORCID Claims' button and the resulting dropdown menu. The dropdown menu shows 'Existing ORCID' with the name 'Reza Salek' and 'Studies Claimed' with the study IDs 'MTBLS1' and 'MTBLS3'. There is also a checkbox for 'ORCID can Remember me on this computer'.

Batch Retrospective Claiming

The screenshot shows the EBI Search website. At the top, there is a search bar with the text "Ex. - hemoglobin, BRCA1_HUMAN" and a "Search" button. Below the search bar, the URL is "https://www.ebi.ac.uk/ebisearch/search.ebi?db=orcid_data_claims&query=domain_source:orcid_data_claims". The main navigation bar includes "EMBL-EBI", "Services", "Research", "Training", and "About us". The search results page displays "EBI Search" and the search query "domain_source:orcid_data_claims". Below the search bar, there are links for "Help & Documentation", "About EBI Search", and "Feedback".

Search results for **domain_source:orcid_data_claims**

Showing **15** results out of **662** in [All results](#) → [Samples & ontologies](#) → [ORCID data claims](#)

Filter your results

Source

- [All results \(662\)](#)
- [Samples & ontologies \(662\)](#)
- [ORCID data claims \(662\)](#)**

Dataset type

- Arrayexpress (641)
- Metabolights (18)
- Pride (3)

The screenshot shows a table of search results for "ORCID data claims". The table has four rows, each representing a different dataset. Each row includes a "Save result" button, a "Create RSS feed" button, and a "Related data" dropdown menu. The table columns are: Dataset ID, ORCID(s), and Source/ID.

Dataset ID	ORCID(s)	Source/ID
E-MTAB-1335	0000-0002-7030-8153	Source: ORCID data claims ID: E-MTAB-1335
E-MTAB-5631	0000-0001-8791-5729	Source: ORCID data claims ID: E-MTAB-5631
E-MTAB-3551	0000-0002-6073-4976	Source: ORCID data claims ID: E-MTAB-3551
E-MTAB-3359	0000-0003-2833-7847	Source: ORCID data claims

Linked Research Visualisation Examples – Researchers – ArrayExpress studies -

The screenshot shows a web browser window displaying the Neo4j website. The browser's address bar contains a dollar sign '\$'. The page header includes the text '\$:play start' and navigation icons for home, back, forward, and close. The main content area features the Neo4j logo on the left and three primary navigation cards:

- Learn about Neo4j**: A graph epiphany awaits you. Includes a small graph icon and a list of questions: 'What is a graph database?', 'How can I query a graph?', and 'What do people do with Neo4j?'. A blue button labeled 'Start Learning' is at the bottom.
- Jump into code**: Use Cypher, the graph query language. Includes a terminal icon and a list of links: 'Code walk-throughs' and 'RDBMS to Graph'. A blue button labeled 'Write Code' is at the bottom.
- Monitor the system**: Key system health and status metrics. Includes a heart icon and a list of metrics: 'Disk utilization', 'Cache activity', and 'Cluster health and status'. A blue button labeled 'Monitor' is at the bottom.

At the bottom of the page, the copyright notice reads: 'Copyright © Neo Technology 2002–2017'.

Synergistic Efforts at EMBL-EBI

<http://www.omicsdi.org/search>

The screenshot shows the omicsdi.org search interface. The browser address bar displays www.omicsdi.org/search. The navigation bar includes links for Home, Browse, API, Database, Help, and Login. A search bar contains the query "organism, repository, gene, tissue, accession" and is set to "Advanced" search. The results section shows 91057 results, with the first page of 10 results displayed. The results are sorted by Relevance and the page size is set to 10. The left sidebar lists various omics categories: Transcriptomics (66307), Genomics (12233), Multiomics (7481), Proteomics (8597), Metabolomics (1343), Models (1648), and UNKNOWN (129). Below this is an "Organisms" section with a search box and a list of organisms including Homo sapiens, Mus musculus, Arabidopsis thaliana, Rattus norvegicus, Drosophila melanogaster, Saccharomyces cerevisiae, Caenorhabditis elegans, and Danio rerio. The "Repository" section is partially visible at the bottom.

Q 91057 Results [show all](#) [save search](#) [copy query](#)

Show results for

- T** Transcriptomics (66307)
- G** Genomics (12233)
- M** Multiomics (7481)
- P** Proteomics (8597)
- M** Metabolomics (1343)
- M** Models (1648)
- ?** UNKNOWN (129)

Organisms

Find your Organisms

- Homo sapiens (33265)
- Mus musculus (19542)
- Arabidopsis thaliana (3773)
- Rattus norvegicus (2557)
- Drosophila melanogaster (2460)
- Saccharomyces cerevisiae (2351)
- Caenorhabditis elegans (1202)
- Danio rerio (713)

Repository

« Previous **1** 2 3 4 5 ... 9106 Next »

Sort by: Relevance Page size 10

T Irisin is a Pro-Myogenic Factor that Induces Skeletal Muscle Hypertrophy and Rescues Denervation-Induced Atrophy

C2C12 myoblasts were seeded at a density of 25,000 cells/cm² in a 10cm cell culture dish and differentiated with low serum differentiation medium at 37 °C, 5% CO₂ for 72h and further treated with eith...

ORGANISM(S): Mus musculus

2017-09-11 | E-MTAB-6024 | ArrayExpress

[transcription profiling by array](#)

Cite Claim Watch

T Microarray analysis of six brain areas from Alzheimers disease patients and normal individuals

Information about the genes that are preferentially expressed during the course of Alzheimer's disease (AD) could improve our understanding of the molecular mechanisms involved in the pathogenesis of ...

ORGANISM(S): Homo sapiens

2017-09-11 | E-GEOD-5281 | ArrayExpress

[transcription profiling by array](#)

Cite Claim Watch

T Transcription profiling by array of human neurons with and without neurofibrillary tangles from patients with Alzheimer's disease

Alzheimer's Disease (AD) is a devastating neurodegenerative disorder affecting approximately 4 million people in the U.S. alone. AD is characterized by the presence of senile plaques and neurofibrill...

ORGANISM(S): Homo sapiens

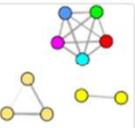
2017-09-11 | E-GEOD-4757 | ArrayExpress

[transcription profiling by array](#)

Cite Claim Watch

Synergistic Efforts at EMBL-EBI

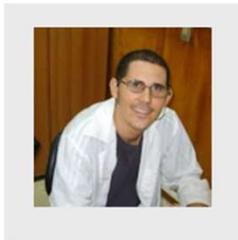
<http://www.omicsdi.org/search>

MetaboLights is a database for Metabolomics experiments and derived information.  286 datasets	MetabolomeExpress A public place to process, interpret and share GC/MS metabolomics datasets.  58 datasets	GPMDb The Global Proteome Machine Database was constructed to utilize the information obtained by GPM servers to aid in the difficult process of validating peptide MS/MS spectra as well as protein coverage patterns.  367 datasets	PAXDB PaxDb contains estimated abundance values for a large number of proteins in several different species. Furthermore, you can find information about inter-species variation of protein abundances.  493 datasets
GNPS The Global Natural Products Social Molecular Networking (GNPS) is a platform for providing an overview of the molecular features in mass spectrometry based metabolomics by comparing fragmentation patterns to identify chemical relationships.  572 datasets	LINCS The Database contains all publicly available HMS LINCS datasets and information for each dataset about experimental reagents (small molecule perturbagens, cells, antibodies, and proteins) and experimental and data analysis protocols.  350 datasets	EGA allows you to explore datasets from genomic studies, provided by a range of data providers. Access to datasets must be approved by the specified Data Access Committee (DAC).  5950 datasets	PeptideAtlas is a multi-organism, publicly accessible compendium of peptides identified in a large set of tandem mass spectrometry proteomics experiments.  2365 datasets
ArrayExpress ArrayExpress Archive of Functional Genomics Data stores data from high-throughput functional genomics experiments, and provides these data for reuse to the research community.  70411 datasets	BioModels Database BioModels Database is a repository of computational models of biological processes. Models described from literature are manually curated and enriched with cross-references.  1648 datasets	JPOST Repository jPOSTrepo (Japan ProteOme Standard Repository) is a new data repository of sharing MS raw/processed data.  57 datasets	ExpressionAtlas The Expression Atlas provides information on gene expression patterns under different biological conditions. Gene expression data is re-analysed in-house to detect genes showing interesting baseline and differential expression patterns.  2913 datasets
Pride is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.  4128 datasets	MetabolomicsWorkbench is a scalable and extensible informatics infrastructure which will serve as a national metabolomics resource.  425 datasets	Massive is a community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data.  1034 datasets	

Synergistic Efforts at EMBL-EBI

<http://www.omicsdi.org/search>

Yasset Perez-Riverol



I'm a Project Leader of Multiomics at the EMBL-European Bioinformatics Institute (Hinxton, Cambridge, UK). I earned undergraduate degrees in Software Engineer (2006) and a doctoral degree in Biochemistry (2013) from the University of Havana. After finishing my PhD in Havana he joined the PRIDE team in 2014. I have lead several development projects such as PRIDE Inspector Toolsuite, and Omics Discovery Index a major resource to find, discovery and link omics datasets.

Contact Info

EMBL-EBI

yperrez@ebi.ac.uk

<https://orcid.org/0000-0001-6579-6941>

0000-0001-6579-6941

Yasset Perez-Riverol

Datasets

PIA - Mouse Benchmark Dataset 150 0 0 0

This Dataset is no actual new study but the mouse benchmark dataset used in the PIA manuscript.
ORGANISM(S): Mus musculus

2015-05-08 | [PXD000790](#) | [Pride](#)

[mouse](#) [benchmark](#) [Reference](#) [Biomedical](#) Cite

PIA - Yeast Gold Standard Benchmark Dataset 159 0 0 0

This Dataset is no actual new study but the Yeast Gold Standard benchmark dataset used in the PIA manuscript.
ORGANISM(S): Saccharomyces cerevisiae

2015-05-08 | [PXD000792](#) | [Pride](#)

[yeast](#) [benchmark](#) [Reference](#) Cite

PIA - iPRG2008 Benchmark Dataset 198 0 0 0

This dataset is no actual new study but the iPRG2008 benchmark dataset used in the PIA manuscript.
ORGANISM(S): Mus musculus

2015-05-08 | [PXD000793](#) | [Pride](#)

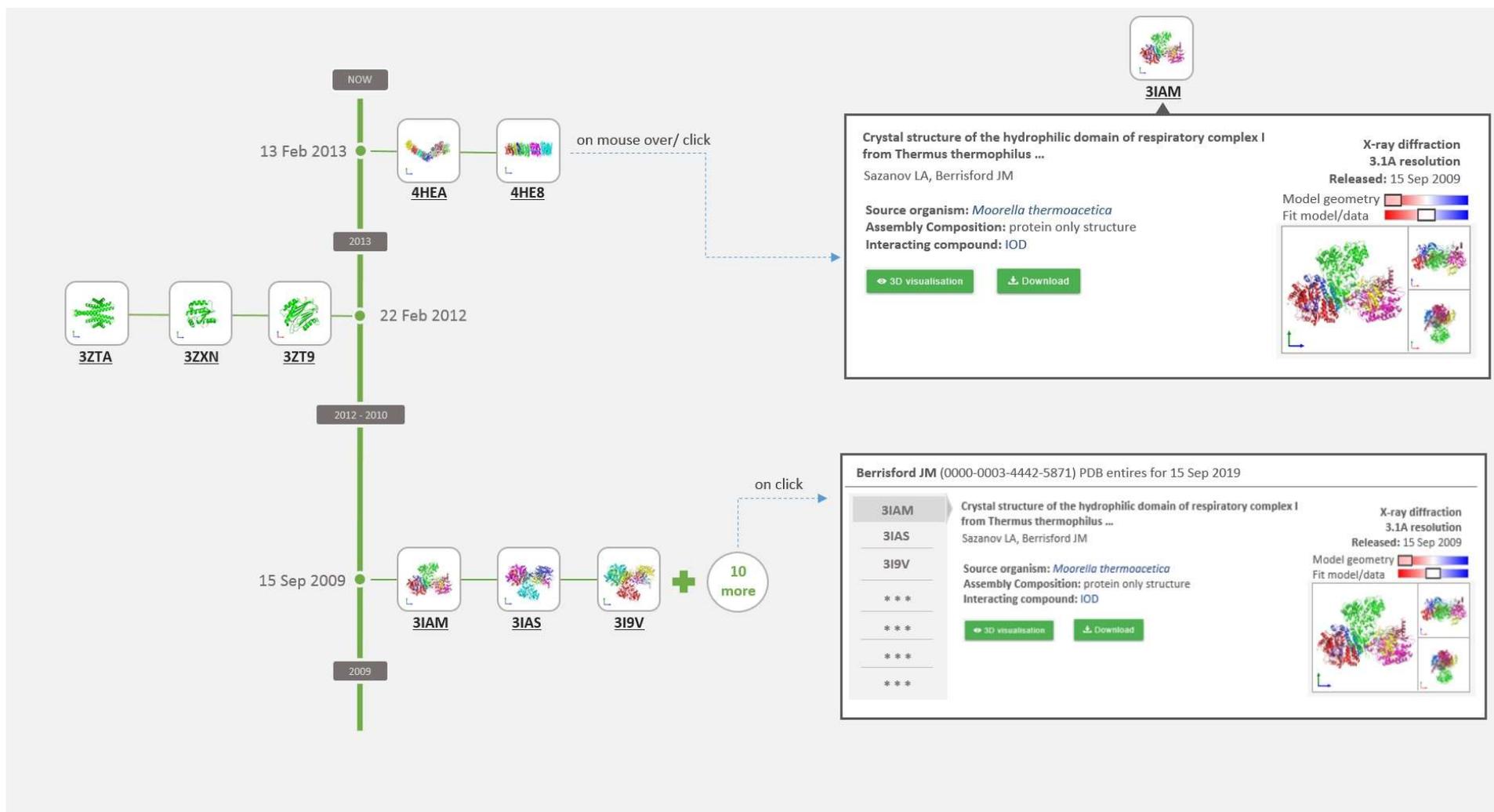
[mouse](#) [iPRG2008](#) [benchmark](#) [Technical](#) [Reference](#) Cite

Search Europe PMC
By ORCID

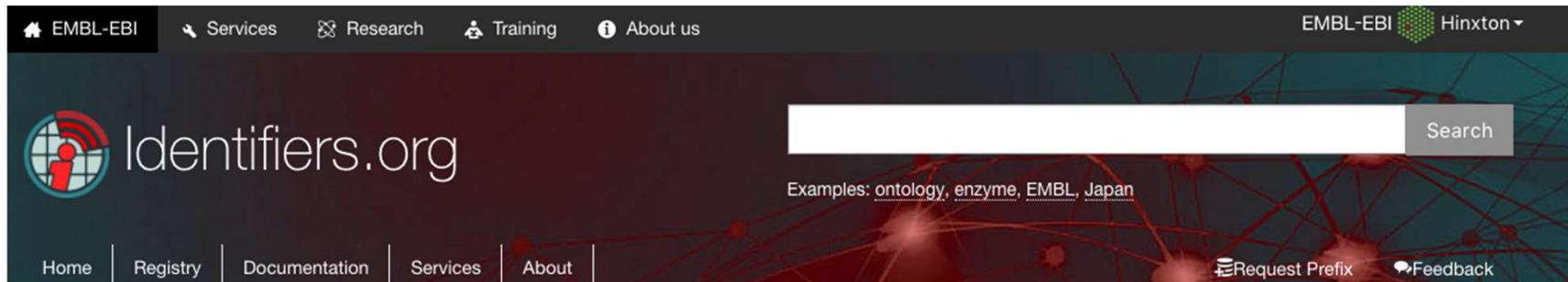
EBI ORCID HUB

ORCID

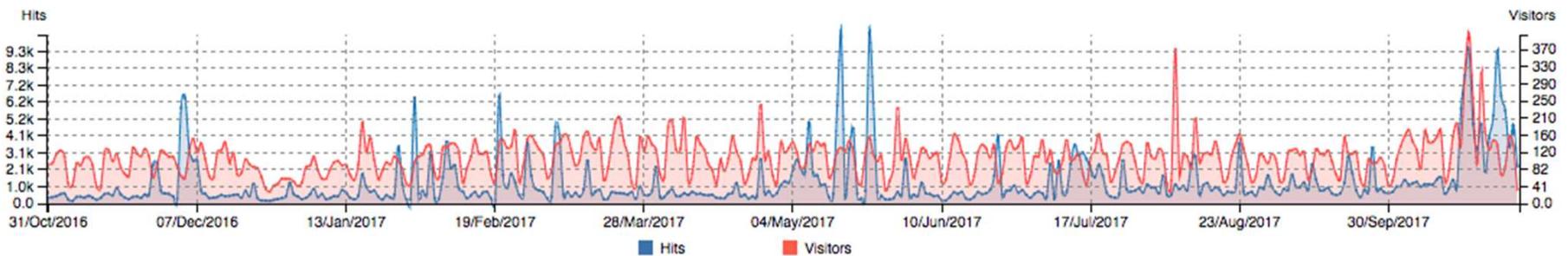
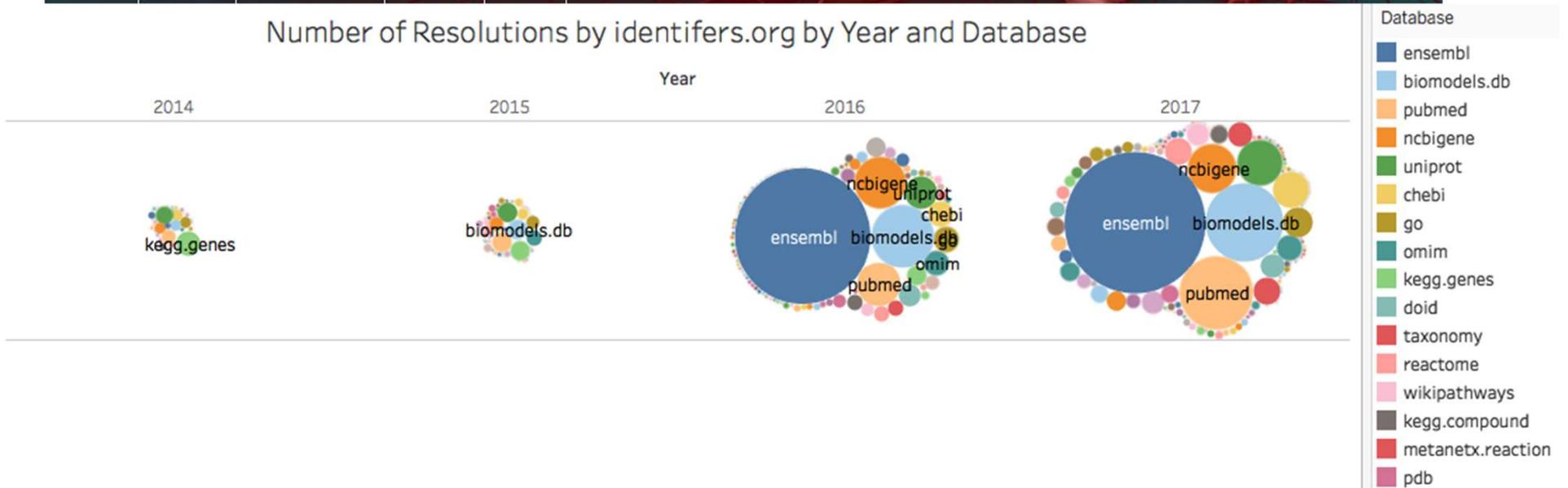
Protein Data Bank - Submitter Profile: Submission Timeline Widget



identifiers.org



Number of Resolutions by identifiers.org by Year and Database



Acknowledgements

- Jo McEntyre, Guilherme Mello, Florian Graef – Europe PMC
- Sarala Wilamaratne, Henning Hermjakob – identifiers.org
- EMBL-EBI teams implementing THOR functionality:
 - ArrayExpress, PDB,
 - PRIDE, OmicsDI,
 - Intact, BioStudies,
 - GWAS Catalogue, Pfam/Rfam,
 - EBI Search
- All my THOR collaborators