12th International Biocuration Conference

APRIL 7-10, 2019

West Road Concert Hall, Cambridge, UK

Abstract Book

biocuration2019.org

biocuration2019@gmail.com

#biocuration2019

ISB International Society for Biocuration

Contents

Welcome

Dear Colleague,

We are delighted to welcome you to Cambridge for the 12th International Biocuration Society (ISB) conference which we hope will be a forum for biocurators and developers to discuss their work, promote collaboration and foster a sense of community in this very active and growing area. We have looked to encourage participation from academia, government, and industry interested in the methods and tools employed in curation of biological data and tried to give you every opportunity to share your knowledge with one another and discuss the future of biocuration.

We take this opportunity to thank our sponsors, without whose help many of the events this week could not have taken place. We thank DATABASE Journal for creating the conference virtual issue through which submitted papers can be accessed https://academic.oup.com/database/pages/biocuration_virtual_issue and also F1000 who are providing a platform for posters to be published https://f1000research.com/collections/biocuration.

The ISB is a non-profit organization for biocurators, developers, and researchers with an interest in biocuration. The society promotes the field of biocuration and provides a forum for information exchange through meeting and workshops.

We hope that by attending this meeting you too will feel welcomed into our biocuration community.

Best wishes,

The Biocuration 2019 Organising Committee

Organising Committee

International Scientific Committee

Claire O'Donovan (Chair)	EMBL-EBI, UK
Rama Balakrishnan	Genentech, US
David Lynn	InnateDB, SAHMRI, Australia
Steven Marygold	University of Cambridge, UK
Pete McQuilton	Oxford University, UK
Ilene Mizrachi	NCBI, US
Chris Mungall	Berkeley, US
Sandra Orchard	EMBL-EBI, UK
Eleanor Stanley	Eagle Genomics, UK
Kimberley Van Auken	WormBase, US
Ulrike Wittig	Heidelberg Institute for Theoretical Studies, Germany
Cathy Wu	University of Delaware, US
Val Wood	University of Cambridge, UK
Zhang Zhang	Beijing Institute of Genomics, China

Local Organising Committee

Sandra Orchard (Chair)	EMBL-EBI, UK
Claire O'Donovan	EMBL-EBI, UK
Alex Bateman	EMBL-EBI, UK
Amy Cottage	EMBL-EBI, UK
Sarah Morgan	EMBL-EBI, UK
Rebecca Foulger	Genomics England, UK
George Georghiou	EMBL-EBI, UK
Rachael Huntley	SciBite, UK
Michele Ide-Smith	EMBL-EBI, UK
Steven Marygold	University of Cambridge, UK
Pete McQuilton	Oxford University, UK
Eleanor Stanley	Eagle Genomics, UK
Val Wood	University of Cambridge, UK

General Information

Conference Badges

Please wear your name badges at all times to promote networking and to assist staff in identifying you.

Internet Access

Wifi Access – eduroam. Additional wifi access and information will be handed out at the registration desk.

Social Media Policy

To encourage the open communication of science and biocuration, we would like to support the use of social media at this year's conference. Please use the conference hashtag #biocuration2019. For poster sessions, please check with the presenter to obtain permission for sharing their work.

Poster Sessions

Poster sessions and a drinks reception will be held at 18:10 on Monday 8th and Tuesday 9th April. Odd number posters will be presented on Monday, the 8th and even numbered posters on Tuesday, the 9th. Poster abstracts are available on the conference webpage: https://www.biocuration2019.org/posters

Social Events

There are several opportunities to explore Cambridge with your fellow conference attendees. While there are numerous bars, pubs, restaurants, and museums all over the city, there are several activities we can recommend. Please visit https://www.biocuration2019.org/social-events for more details.

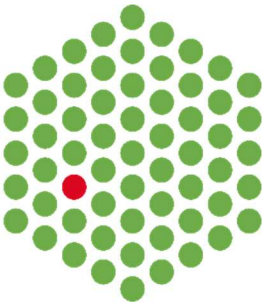
Taxis

Please find a list of local taxi numbers below:
Panther Taxis - www.panthertaxis.co.uk
+44 (0) 1223 715715
CamCab - http://camcab.co.uk/
+44 (0) 1223 704704
A1 Cabco - http://www.a1cabco.co.uk/
+44 (0) 1223 313131
Cambridge City Taxis - https://www.cambridgecitytaxis.co.uk/
+44 (0) 1223 832832

Biocuration 2019 Sponsors

We are proud to be sponsored by the following organisations:

EMBL-EBI



https://www.ebi.ac.uk/

https://www.elixir-europe.org

https://currentprotocols.onlinelibrary.wiley.com/journal/1934340x

https://www.eaglegenomics.com/

www.e-mergeglobal.com

OXFORD

UNIVERSITY PRESS

https://academic.oup.com/journals

The Company of Biologists

https://www.biologists.com/

SciBite

https://www.scibite.com/

SPRINGER NATURE

https://www.springernature.com/gp/authors/research-data?utm_source=confrence&utm_campaign=RSDT-Biocuration2019

Biocuration 2019 Programme

Day 1 - Sunday 7th April

9:00 **Pre-conference workshops –** Location TBA, please check website
Developing the Gene Regulation Knowledge Commons
Martin Kuiper
The IMEx Consortium and molecular interactions: beyond A binds to B
Pablo Porras Milan
2nd Phenotypes Traversing All the Organisms (POTATO) Workshop
David Osumi-Sutherland
Practical ontology applications, tooling and interoperability best practices for FAIRification
Danielle Welter, Randi Vita and Sirarat Sarntivijai

12:30 Lunch

13:30 **Pre-conference workshops –** Location TBA, please check website
Mapping the landscape of biocuration - where are the biocurators and what do they need?
Pete McQuilton, Patricia Palagi, Sarah Morgan, Alex Holinski and Melissa Burke

15:45 **Conference registration at West Road Concert Hall**

16:15 Tea

16:45 **Opening & Welcome**
Chairs: Claire O'Donovan and Sandra Orchard

17:00 **Keynote Lecture: Séan O'Donoghue**
Chairs: Claire O'Donovan and Sandra Orchard

18:00 Drinks reception

Day 2 - Monday 8th April

9:00 **Keynote Lecture: Paul Sternberg**
Chair: Sandra Orchard

9:55 **Global Biodata Coalition**
Rolf Apweiler

Core Data Resources
Niklas Blomberg

10:55 Tea/coffee break

Session 1: Functional Annotation
Chairs: Valerie Wood and Steven Marygold

11:20 **The SwissLipids knowledge resource for lipid biology**
Lucila Aimo

11:40 **Enhanced enzyme annotation in UniProtKB using Rhea**
Kristian Axelsen

12:00 **An Evaluation of Gene Ontology Annotation of Gene Products Associated with Immunological Processes**
Alexander Diehl

12:20 **SIGNOR and DISNOR: two sister databases for the analysis of causal relationships whose disruption underlies genetic diseases**
Luana Licata

12:40 **Capturing variation impact on molecular interactions: updates on the IMEx Consortium mutations data set**
Pablo Porras Millan

13:00 Lunch

14:00 **Parallel workshops**

How should biological resources be evaluated?
Valerie Wood, Marc Gillespie and Elena Speretta

Diverse Perspectives on Data Licensing
Andrew Su, Monica Munoz-Torres and Raja Mazumder

Curating Evidence for Gene: Disease Validity for Clinical Omics
Members of the Gene Curation Coalition (GenCC)

16:00 Tea/coffee break

Session 2 - Curation and data visualisation tools
Chairs: Kimberly Van Auken, David Lynn and Chris Mungall

16:30 **Efficient Curation of Genome Annotations through Collaboration with Apollo**
Nathan Dunn

16:50 **Submission, archival and visualisation of single-cell sequencing data**
Nancy George

17:10 **MetaboLights study editor - An open-access curation tool for metabolomics studies submission and associated meta-data annotation**
Venkata Chandrasekhar Nainala

17:30 **The BioGRID Interaction Database: Curation strategies and new developments for capturing genome-wide CRISPR/Cas9-based screens**
Rose Oughtred

17:50 **Exploring neXtProt data and beyond: A SPARQLing solution**
Monique Zhan

18:10 Poster reception (Odd numbered posters) **–** Drinks and Nibbles

19:30 Conference close

Day 3 - Tuesday 9th April

9:00 **Keynote Lecture: Ellen McDonagh**
Chair: Claire O'Donovan

9:55 **Award 1 presentation (5m) + talk (30m)**

10:55 Tea/coffee break

Session 3 - Curation for human health and nutrition
Chairs: Eleanor Stanley and Rebecca Foulger

11:20 **OncoMX: a cancer biomarker resource leveraging published literature and genomics data**
Evan Holmes

11:40 **EWAS Atlas: a curated knowledgebase of epigenome-wide association studies**
Mengwei Li

12:00 **UniProtKB and Alzheimer's Disease: Linking molecular defects to disease phenotype**
Yvonne Lussi

12:20 **LOINC2HPO: Curation of Phenotype Data from the Electronic Health Records using the Human Phenotype Ontology**
Nicole Vasilevsky

12:40 **Curating the authorship of clinical records and biomedical abstracts**
Fabio Rinaldi

13:00 Lunch

14:00 **Parallel workshops**
Biocuration in Industry
Jane Lomax
Equality, Diversity, and Inclusion
Mary Ann Tuli

16:00 Tea

Session 4 - Database journal sessions
Chairs: Mike Cherry and Sandra Orchard

16:30 **Annotation of gene product function from high-throughput studies using the Gene Ontology**
Helen Attrill

16:40 **Validation of protein-protein interactions in databases and resources: the need to identify interaction detection methods that provide binary or indirect experimental evidences**
Javier de Las Rivas

16:50 **An enhanced workflow for variant interpretation in UniProtKB/Swiss-Prot improves consistency and reuse in ClinVar**
Maria Livia Famiglietti

17:00 **Increased Interactivity and Improvements to the GigaScience Database, GigaDB**
Christopher Hunter

17:10 **Towards comprehensive annotation of Drosophila melanogaster enzymes in FlyBase**
Steven Marygold

17:20 **ccPDB 2.0: An updated version datasets of created and compiled from Protein Data Bank**
Piyush Agrawal

17:30 **Building Deep Learning Models for Evidence Classification from the Open Access Biomedical Literature**
Gully Burns

17:40 **Curating Gene Sets: Challenges and Opportunities for Integrative Analysis**
Gaurab Mukherjee

17:50 **Using Deep Learning to Identify Translational Research in Genomic Medicine Beyond Bench to Bedside**
T.B.A.

18:00 **Integration of Macromolecular Complex Data into the Saccharomyces Genome Database**
Edith Wong

18:10 Poster reception (Even numbered posters) – Drinks and nibbles

19:30 Conference Close

20:00 Conference Dinner at Downing College

Day 4 - Wednesday 10th April

9:00 **Keynote Lecture: Susanna Sansone**
Chair: Cecilia Arighi

9:55 **Award 2 presentation (5m) + talk (30m)**
Poster Prizes (15m)
Chair: Cecilia Arighi

10:55 **Tea/coffee break**

Session 5 - Data standards and ontologies: Making data FAIR
Chairs: Ilene Mizrachi and Pete McQuilton

11:20 **Research on Metadata Standards of Biomedical Data**
Jiawei Cui

11:40 **New approaches to data management: supporting FAIR data sharing at Springer Nature**
Varsha Khodiyar

12:00 **The ELIXIR Recommended Interoperability Resources (RIRs) - What tools can I use to make data FAIR?**
Sirarat Sarntivijai

12:20 **Expanding MIxS Genomic Minimal Information Standards**
Lynn Schriml

12:40 **The Ontology for Biomedical Investigations (OBI) as a Curation Tool**
Randi Vita

13:00 Lunch

14:00 **ISB meeting**
Chair: Sandra Orchard

Session 6 - Interacting with the Research Community
Chairs: Claire O'Donovan and Ulrike Wittig

14:30 **The ELIXIR Data Platform in 2019**
Rachel Drysdale

14:50 **Measuring the value of data curation as a part of the publishing process**
Varsha Khodiyar

15:10 **Involving researchers in the biocuration of plant genes and pathways**
Sushma Naithani

15:30 **Coordination and Collection of Data for a Community Global Biodiversity Initiative**
Jeena Rajan

15:50 **Research on the Collaborative Mechanisms for Journals and Data Repositories in the Integrated Publication of Papers and Scientific Data**
Jinming Wu

16:10 Conference Close

Keynote Speakers

Seán O'Donoghue

Laboratory Head at the Garvan Institute of Medical Research

Professor Seán O'Donoghue is a Laboratory Head at the Garvan Institute of Medical Research in Sydney, and also a Chief Research Scientist in Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO), and conjoint Professor in the University of New South Wales. He has created numerous award-winning bioinformatics resources such as Reflect, Minardo, and Aquaria. He chairs VIZBI, an international conference series on data visualisation methods in the life sciences. These resources are accessed by tens of thousands of scientists worldwide each year.

Professor O'Donoghue's contributions have been recognised with a C.J. Martin Fellowship from the NHMRC, an Achievement Award from Lion Bioscience AG, and by election as a Fellow of the Royal Society of Chemistry. He received his B.Sc. and PhD in biophysics from the University of Sydney. Much of his career was spent in Germany at the European Molecular Biology Laboratory (EMBL) and at Lion Bioscience AG.

Notes:

Paul Sternberg

Director of the Center for Biological Circuit Design

Paul Sternberg graduated from Hampshire College in 1978 and joined the Biology Department at M.I.T. for his Ph.D., which he received in 1984. His graduate work on the Genetic Control of Nematode Development was under the supervision of H. Robert Horvitz. He then pursued postdoctoral research on yeast mating type with Ira Herskowitz at UCSF, and returned to C. elegans when he joined the Caltech Biology Division faculty in 1987, where he is now Bren Professor of Biology.

He was an Investigator with the Howard Hughes Medical Institute, from 1989 - 2017. He became lead-PI of WormBase in 1999 and started the Caltech branch of WormBase. He served on the board of directors of the Genetics Society of America from 2000-2003.

At Caltech, he led the Biology graduate program for several years, helped found the BioEngineering graduate option, and is now director of the Center for Biological Circuit Design, part of the new Information Science and Technology initiative.

Notes:

Ellen McDonagh

Head of Curation, Genomics England

Ellen began her research career in the lab as a sandwich student with AstraZeneca as part of her undergraduate degree, before moving on to complete a PhD in Immunology research at Imperial College London. Ellen realised that she wanted to pursue opportunities outside the lab, and explored alternative career options in science. An opportunity came up for a role in curation at Stanford University, California, at PharmGKB, the Pharmacogenomics Knowledgebase, a publicly-accessible database located at Stanford University, which captures knowledge about the impact of genetic variation on drug response. This allowed Ellen to apply her science knowledge and transferable skills gained during her PhD in a desk-based role.

Now, Ellen is Head of Curation at Genomics England and has a team of curators working with her on the 100,000 Genomes Project. The project aims to create a new genomic medicine service for the NHS. The Curation Team at Genomics England are currently working with clinicians and researchers worldwide to curate evidence for gene panels to help diagnose rare diseases, using the publicly available crowdsourcing database PanelApp.

Notes:

Susanna-Assunta Sansone

Associate Director, FAIR Data Science, Oxford e-Research Centre

Susanna is an Associate Director, and Principal Investigator at the Oxford e-Research Centre, and an Associate Professor at the Department of Engineering Science of the University of Oxford. She is also the founding Honorary Academic Editor of the Springer Nature Scientific Data open access journal.

She holds a PhD in Molecular Biology from Imperial College, London, UK; after a few years working on vaccine genetics in an Imperial's spinnoff she moved to the European Bioinformatics Institute where she worked for nine years as a Project and Team Coordinator and Principal Investigator, before moving to the University of Oxford in 2010.

She is interested in data reproducibility and the evolution of scholarly publishing, which drive science and discoveries. With her team of data engineers (research software and knowledge engineers/curators) she runs research and development activities in the areas of knowledge and information management, and interoperability of applications. More specifically, they investigate and implement new ways to make digital research objects (including data, software, model and workflows) Findable, Accessible, Interoperable and Reusable, in one other word: FAIR.

Notes:

Session 1 - Functional Annotation

The SwissLipids knowledge resource for lipid biology

Lucila Aimo
Swiss-Prot, SIB Swiss Institute of Bioinformatics

SwissLipids (www.swisslipids.org) is a freely available knowledge resource for lipid biology. In SwissLipids, targeted expert biocuration of lipid metabolism using Rhea (www.rhea-db.org) and UniProt (www.uniprot.org) provides the knowledge needed to generate a library of more than 550,000 annotated lipid structures from over 300 lipid classes. This lipid library is fully mapped to ChEBI, Rhea and UniProt. It is organized in a hierarchical classification that maps lipidomics data to possible lipid structures and associated biological knowledge, like this: PC(38:4) -> PC(18:0/20:4(5Z,8Z,11Z,14Z)) -> PLA2G4A. The SwissLipids website provides a range of search and browse options as well as an identifier mapping service for resources such as LIPID MAPS and HMDB. SwissLipids is used for lipidomic data interpretation and annotation by projects such as the Innovative Medicines Joint Initiative (IMI-JU) for Diabetes (IMIDIA, http://www.imidia.org/), the EU H2020 project METASPACE (http://metaspace2020.eu/) on bioinformatics for spatial metabolomics, any by commercial lipidomics service providers such as LipoType (www.lipotype.com), as well as many individual research laboratories. The SwissLipids biocuration effort focuses on human lipids and those of major model organisms from vertebrates to yeast, experimental systems used by our collaborators in the LipidX project of the Swiss Initiative in Systems Biology SystemsX.ch. Here we provide an overview of some of our latest biocuration work, including a targeted biocuration effort that covers hundreds of classes of complex glycosphingolipids – building on the work of expert resources such as SphinGOMAP.

Notes:

Enhanced enzyme annotation in UniProtKB using Rhea

Kristian Axelsen
SIB Swiss Institute of Bioinformatics

The UniProt Knowledgebase (http://www.uniprot.org) is a large reference resource of protein sequences and functional annotation. More than 45% of UniProtKB/Swiss-Prot entries are enzymes, which were traditionally annotated using EC (Enzyme Commission) numbers, the hierarchical 4 digit enzyme classification based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Here we describe our work on the enhancement of enzyme annotation in UniProtKB using Rhea (https://www.rhea-db.org). Rhea is a comprehensive expert-curated knowledgebase of biochemical reactions that uses the ChEBI ontology to describe reaction participants, their chemical structures, and chemical transformations – a computationally tractable description of reaction chemistry. UniProt has recently adopted Rhea as the reference vocabulary for enzyme annotation in UniProtKB, and now describes all enzymatic reactions using Rhea where possible. Rhea provides improved consistency and precision of enzyme annotation in UniProtKB and allows UniProt users to search, browse, and mine enzyme data in new ways, combining approaches from the fields of cheminformatics and bioinformatics. Going forward, UniProt curators are now using Rhea in their daily work: major curation efforts are focusing on improved coverage of human and microbial metabolism in health and disease (some of which will be described here) as well as biosynthetic pathways for natural products (see the poster “Diverse taxonomies for diverse chemistries: enhanced plant and fungal metabolic pathway annotation for natural product biosynthesis in UniProtKB/Swiss-Prot”). You can learn more about Rhea in the poster “Rhea, an expert curated resource of biochemical reactions for enzyme annotation”, which also describes how we aim to improve the alignment of Rhea with the Gene Ontology (GO) and other knowledge resources such as Reactome.

Notes:

An Evaluation of Gene Ontology Annotation of Gene Products Associated with Immunological Processes

Alexander Diehl
University at Buffalo

The Gene Ontology Consortium relies on both curator-driven development and workshops of domain experts for revision of the Gene Ontology (GO) to enable better GO term representation for diverse areas of biology. One of the earliest workshops was held in 2005 and brought together expert immunologists and GO developers to improve the representation of immunological processes. The workshop and follow-up efforts resulted in the addition of over 700 new terms, improvements in existing terms, and overall rearrangement of the ontology hierarchy. While many of these terms have been used in annotation of gene products associated with the immune system, it is apparent after 12 years that a number of the terms have never been used in annotation, and many well studied immune system proteins are inadequately annotated. In order to get a better picture of the state of annotation for immunology, we have studied annotation patterns for GO terms that are subclasses of 'immune system process' for gene products from human, mouse, and rat. We find an overall trend of more annotations in all three species for GO terms that are subclasses of 'innate immune response' rather than 'adaptive immune response'. Annotation of human gene products have less of a skewing to subclasses of 'innate immune response' whereas annotation of mouse and rat gene products are strongly skewed. We believe these differences are the result of both differing types of data available in different species, as well as differing annotation practices of curators working on these species. Furthermore, analysis of annotations in all three species tends toward less granular terms in general. GO annotation for immunology can be quite complex and demands a degree of domain expertise to select the most appropriate terms. Based on our results, we are planning a focused effort in GO annotation for immunology, in order to improve the utility of the GO for term enrichment and other downstream analyses for immunological data.

Notes:

SIGNOR and DISNOR: two sister databases for the analysis of causal relationships whose disruption underlies genetic diseases.

Luana Licata
University of Rome, Tor Vergata

SIGNOR (http://signor.uniroma2.it) -- the SIGnaling Network Open Resource -- is a manually curated database that captures, organizes and displays signaling information as binary causal relationships between biological entities (proteins, chemicals, protein families, complexes, small molecules, phenotypes and stimuli). These relationships are displayed as signed directed graphs in a viewer application that places entities in specific compartments (extracellular, membrane, cytoplasm, nucleus). SIGNOR annotates about 20,000 interactions between 5,000 biological entities maintaining the link to the published experiments that support the interaction. The data in SIGNOR can be freely explored in the WEB interface or downloaded for local analysis. Users can upload a user defined list of proteins and query the database for causal relationships that link the proteins in the query list. A similar approach is implemented in DISNOR (https://disnor.uniroma2.it/), a new resource that uses a comprehensive collection of disease associated genes, as annotated in DisGeNET, to interrogate SIGNOR in order to assemble disease-specific logic networks linking disease associated genes by causal relationships. DISNOR is an open resource where more than 4000 disease-networks, linking ~ 2800 disease genes, can be explored. For each disease curated in DisGeNET, DISNOR links disease genes through manually annotated causal relationships and the inferred 'patho-pathways' can be visualised at different level of complexity.

Notes:

Capturing variation impact on molecular interactions: updates on the IMEx Consortium mutations data set

Porras Millan, Pablo
EMBL-EBI - Molecular Interactions Team

Molecular interaction (MI) networks provide maps to explore cellular processes from a systems perspective. Combining them with genomic variation data can bring in-depth insight into the challenge of understanding the effector mechanisms of amino acid variation. The IMEx Consortium (www.imexconsortium.org) is an international collaboration between databases that curate MI data from the scientific literature, represent it with full experimental detail and make it freely available for the scientific community. Over the last 14 years, IMEx curators have collectively annotated 900,000 physical binary interactions, assigning details such as kinetic parameters, variable experimental conditions or construct details, including binding interfaces and mutations that affect interactions. Leveraging the IMEx detailed curation model, we have compiled a data set of over 40,000 annotations of protein mutations affecting interaction and made it freely available at www.ebi.ac.uk/intact/resources/datasets#mutationDs. The data features information about the amino acid changes, their effect over the interaction and full reference to the experimental interaction evidence from which it was extracted. Over 22,000 unique sequence changes, affecting 4500 proteins from 300 different species, are reported. Around 75% of the annotation are mapped to human proteins, providing high-quality experimental evidence of sequence change effects which directly relate to existing variation data. We present the latest updates of the data set, along with future perspectives and developments such as its integration within existing variation annotation tools, like ENSEMBL's VEP; its extension to include mutations in DNA/RNA as interacting partners; and our plans to increase accessibility. This openly available resource is an invaluable tool with immediate applications in the study of variation impact on the interactome, interaction interfaces and previously un-annotated variants, among other key questions.

Notes:

Session 2 - Curation and data visualisation tools

Efficient Curation of Genome Annotations through Collaboration with Apollo

Nathan Dunn
Lawrence Berkeley National Lab

Accurately annotated genomes are vital to understanding the biological function contributed by each genomic element. Researchers must review diverse information such as transcriptome alignments and predictive models based on sequence profiles, over potentially many iterations, and then integrate into a unified model for each genomic element. Tools that aid in the review, evaluation and integration need to be simple to install, configurable, efficient to use, and able to include additional analyses, genomes, workflows, and researchers (wherever they may be geographically located).To this end, the Apollo genome annotation editor provides a collaborative graphical platform for researchers to review and revise the predicted features on genome sequences in real-time (similar to Google Docs). Apollo can be downloaded directly to run locally (or via Docker) for individual users, and also be setup so that a single web server can concurrently support multiple research teams with hundreds of researchers and genomes. The most recent focus of Apollo has been to provide users with the ability to add and share genomes and genomic evidence (JBrowse tracks) directly through the interface using standard formats (e.g. GFF3, FASTA, BAM, CRAM, VCF), eliminating the need for an administrator to run additional scripts to load these onto the server. We are also focusing on enabling genome publishing as a browsable, graphical resource. When project researchers decide to make their genomic annotations publicly available they can generate snapshots of these in JBrowse archival hubs. Finally, we are enhancing our variant annotation capabilities, including the ability to visualize the impact a variant would have on the annotated isoforms they intersect. Apollo is used in over one hundred genome annotation projects around the world, ranging from annotation of a single species to lineage-specific efforts supporting the annotation of dozens of genomes. https://github.com/GMOD/Apollo/

Notes:

Submission, archival and visualisation of single-cell sequencing data

Nancy George
EMBL-EBI

The mission of the Gene Expression team at EMBL-EBI is the development of tools to facilitate submission, archival, reprocessing and visualisation of functional genomics data. Our tools and resources are continuously updated to incorporate data from new technologies, with our recent release of resources for single-cell RNA-sequencing (scRNA-seq) datasets which investigate the transcriptome at the single cell level. To enable capture of rich experimental metadata for scRNA-seq studies, we have created new templates for our submission tool, Annotare that represent the minimal technical information required to reproduce and reprocess single-cell experiments. A set of minimal information typically includes details describing the cell isolation protocol; cell quality measurements; library construction process and data file content. Annotations are chosen from a controlled vocabulary or mapped to Environmental Factor Ontology (EFO) terms to ensure consistency. Once submitted, datasets are reviewed by curators for accuracy before raw data is archived at the European Nucleotide Archive (ENA). Sample metadata and processed data are made available in our functional genomics archive ArrayExpress, currently hosting 120 single-cell datasets. Data sets are then reprocessed using our in-house standardised pipelines and visualised in the newest component of our added-value resource Expression Atlas, the Single Cell Expression Atlas, which contains 50 datasets across 9 species. Users can explore the expression of a specific gene of interest across different species and experiments. Results can be filtered by tissues and cell types, and point out whether the gene was identified as a "marker gene" in a particular cell population. Data points are presented in a t-SNE plot which showcases the variability of gene expression at the single cell level. Alongside the expression levels, the plots display metadata such as the cell cluster defined by the SC3 algorithm and any experimental variables.

Notes:

MetaboLights study editor - An open-access curation tool for metabolomics studies submission and associated meta-data annotation

Venkata Chandrasekhar Nainala
EMBL-EBI

MetaboLights database is an international metabolomics repository recommended by many leading journals including Nature, PLOS and Metabolomics. The service's unique manual curation maintains quality, provides helpful support for users and ensures accessibility for secondary analysis of studies. MetaboLights hosts a wealth of cross-species, cross-technique, open access experimental research. As a part of our ongoing efforts to streamline the study submission and curation process, the MetaboLights team at EMBL-EBI has developed a new tool to edit and submit studies online. The tool provides MetaboLights users and curators with an intuitive and easy to use interface to create, edit and annotate their studies online. The convenient, context-aware editor navigates curators and users through the study to define a rich description of the experimental metadata including study characteristics, protocols, technology and related factors. Metadata descriptions are enhanced by mapping this information to controlled ontologies repositories using ZOOMA. Capturing such a complete data set benefits the community by making results findable, reproducible and reusable. Going forward we have plans to incorporate text mining tools such as Named Entity Recognition (NER) to annotate metadata, enabled by the robust architecture of the online editor. Other plans include offline edit support, direct channels for curators to contact and communicate with the submitters to make the whole process of data curation more submitter-friendly.

Notes:

The BioGRID Interaction Database: Curation strategies and new developments for capturing genome-wide CRISPR/Cas9-based screens

Rose Oughtred
Lewis-Sigler Institute for Integrative Genomics, Princeton University

The Biological General Repository for Interaction Datasets (BioGRID, see thebiogrid.org) is an open-access database resource for protein, genetic and chemical interaction data, as manually curated from the literature for human and other major model organisms. As of December 2018, BioGRID contains over 1,650,600 interactions captured from more than 57,650 publications. The recent development of genome-wide knockout libraries based on CRISPR/Cas9 technology has enabled many high-throughput genetic screens in cell lines. To capture these results, a newly developed aspect of BioGRID captures gene-phenotype relationships from genome-wide CRISPR/Cas9 screens, as well as CRISPR/Cas9-derived genetic interactions from either screens or focused experiments. This new resource, called the Open Repository of CRISPR Screens (ORCS, see orcs.thebiogrid.org) currently houses over 500 curated genome-wide screens performed in 417 human or mouse cell lines. A minimal information about CRISPR screens (MIACS) record structure was developed to represent key CRISPR/Cas9 screen parameters. ORCS serves as a unified resource for CRISPR/Cas9 datasets and provides a flexible interface for searching, filtering and comparing screen datasets. To maintain consistency with the original publications, ORCS reports published screen scores according to original scoring algorithms. Results are displayed at the publication-, screen- and gene-level with original scores and significance thresholds, along with associated analytical methods and other metadata. Current screen formats in ORCS include negative and positive selection based on viability and other phenotypes in conjunction with knockout (CRISPRn), transcriptional activation (CRISPRa) or transcriptional inactivation (CRISPRi) library designs. All data are freely available for download in various standardized formats and also as the original supplementary files associated with the publication. This project is supported by NIH R01OD010929 to MT, KD.

Notes:

Exploring neXtProt data and beyond: A SPARQLing solution

Monique Zhan
SIB Swiss Institute of Bioinformatics

The neXtProt platform (www.nextprot.org) developed at SIB Swiss Institute of Bioinformatics is a one-stop-shop for human proteins proposing solutions to select, explore and reuse available genomic, transcriptomics, mass-spectrometry- and antibody-based proteomics data. The neXtProt team manually curates data from the literature (post-translational modifications, variant phenotypes, protein-protein interactions, etc.) and combines it with high quality omics data generated by systems biology projects using a single inter-operable format. neXtProt data are FAIR (Findable, Accessible, Interoperable, and Reusable), with full traceability ensured by extensive use of metadata. In the last four years, neXtProt has been promoting the use of SPARQL, a semantic query language for databases, to check, explore, and visualize its data. SPARQL queries are used to check the quality and consistency of the data loaded at each release. To date, over 450 queries have been written such that non-zero results trigger investigation. In an effort to automate these tests, all of the queries or a particular sub-set can be launched and the results written to a file. Semantic technologies can help generating innovative hypotheses where classical data mining tools have failed (protein function prediction, drug repositioning...). In order to promote the use of semantic technologies as data mining tools for life sciences, neXtProt provides over 140 pre-built queries and documentation of its data model to guide the user in his or her first steps. The use of SPARQL allows users to run federated queries across resources relevant for human biology or build customized views. All our SPARQL queries are open source and available on GitHub.

Notes:

Session 3 - Curation for human health and nutrition

OncoMX: a cancer biomarker resource leveraging published literature and genomics data

Evan Holmes
George Washington University

The massive, multiform datasets generated by cancer genomics studies serve as incredible resources for the scientific community. However, the size and heterogeneity of these datasets present challenges to interpreting biological significance across datasets. A variety of technical characteristics including file formats, attribute names, and reference data often differ between databases. While these distinctive technical characteristics are designed to facilitate highly specific research, they hamper re-use in more general studies. OncoMX is designed to broadly support the scientific community by unifying datasets to remove such challenges. The four major use-case perspectives driving the OncoMX web portal interface development are (1) exploration of cancer biomarkers, (2) evaluation of mutations and expression in an evolutionary context, (3) side-by-side exploration of published literature-mined data for mutations and expression in cancer, and (4) exploration of a specific gene or biomarker within a pathway context. To this end, OncoMX integrates and unifies sequence-based mutation data from BioMuta, cancer/normal differential expression data from BioXpress, normal expression data across organisms from Bgee, literature mining evidence for mutation and expression in cancer with DiMeX and DEXTER, biomarker data from EDRN, and pathway data from Reactome. Data integrated in OncoMX will be supplemented by functional annotations, scRNA-seq data, and additional FDA-approved biomarker data, as available. OncoMX is a collaboration between The George Washington University, NASA's Jet Propulsion Laboratory, the SIB Swiss Institute of Bioinformatics, and the University of Delaware. The multifarious research foci of this international collaboration supports diverse end-user research interests, ultimately shaping the OncoMX data model and web portal. Therefore, OncoMX is projected to widely support cancer research through relevant cancer biomarker data aggregation.

Notes:

EWAS Atlas: a curated knowledgebase of epigenome-wide association studies

Mengwei Li
Beijing Institute of Genomics

Epigenome-Wide Association Study (EWAS) has become increasingly significant in identifying the associations between epigenetic variations and different biological traits. In this study, we develop EWAS Atlas (http://bigd.big.ac.cn/ewas), a curated knowledgebase of EWAS that provides a comprehensive collection of EWAS knowledge. Unlike extant data oriented epigenetic resources, EWAS Atlas features manual curation of EWAS knowledge from extensive publications. In the current implementation, EWAS Atlas focuses on DNA methylation—one of the key epigenetic marks; it integrates a large number of 299,016 high-quality EWAS associations, involving 113 tissues/cell lines and covering 293 traits, 1,914 cohorts and 390 ontology entities, which are completely based on manual curation from 713 studies reported in 414 publications. In addition, it is equipped with a powerful trait enrichment analysis tool, which is capable of profiling trait-trait and trait-epigenome relationships. Future developments include regular curation of recent EWAS publications, incorporation of more epigenetic marks and possible integration of EWAS with GWAS. Collectively, EWAS Atlas is dedicated to the curation, integration and standardization of EWAS knowledge and has the great potential to help researchers dissect molecular mechanisms of epigenetic modifications associated with biological traits.

Notes:

UniProtKB and Alzheimer's Disease: Linking molecular defects to disease phenotype

Yvonne Lussi
EMBL-EBI

Alzheimer's disease (AD) is a progressive, neurodegenerative disease and the most common form of dementia in elderly people. Linkage analysis was the first milestone in unravelling the mutations in APP, PSEN1 and PSEN2 that cause early-onset AD. The development of next-generation sequencing methods over the last decade has increased the detection of genetic variants and the identification of disease-associated genes. However, establishing the relationship between the variants and disease phenotype remains a challenge. In this context, UniProtKB aims to link resources from genetic and medical information to protein sequences and associated biological knowledge. In a joint effort across the UniProt sites, over 180 proteins have been identified by text mining and by input from experts in the field to be associated with AD. By manual curation, information from peer-reviewed literature will be annotated on molecular function and involvement in disease, including disease-associated variant positions and variant characterization. By focusing our curation efforts on proteins involved in AD, we hope to shed light on the mechanisms leading to this devastating disease. We focus on a thorough review of available information on sequence variants and associated AD information, as well as normal protein function of proteins associated with the disease. The information on variants together with variant functional description, protein molecular function, structural data and protein-protein interaction should help researchers in the field of neurodegeneration, clinicians and biomedical researchers to gain a global view on the relation between variant and disease and help elucidating disease mechanism.

Notes:

LOINC2HPO: Curation of Phenotype Data from the Electronic Health Records using the Human Phenotype Ontology

Nicole Vasilevsky
Oregon Health & Science University

Electronic Health Record (EHR) data are often encoded using Logical Observation Identifier Names and Codes (LOINC) in the United States, which is a universal standard for coding medical laboratory observations in EHRs. LOINC encoded clinical tests can be inferred as phenotypic outcomes, and offer the potential for secondary reuse of EHR data for patient phenotyping. The Human Phenotype Ontology (HPO) has been widely used for deep phenotyping in research and for diagnostic purposes. It contains over 13,500 classes that represent phenotypic abnormalities encountered in human diseases. Mapping the LOINC codes to HPO terms provides an opportunity to structure this phenotypic information and enable automatic extraction of detailed deep phenotypic profiles of laboratory results for downstream studies. Multiple LOINC codes can code for lab tests that can be represented as a single phenotype; for example multiple codes exist for measurements of nitrate levels in urine, which could be interpreted as HP:0031812 Nitrituria. To harmonize lab test data from EHRs to HPO, we developed a curation tool that converts EHR observations into HPO terms. To date, over 2400 LOINC codes have been mapped to HPO terms and our mapping library is freely available online (https://w3id.org/loinc2hpo/annotations). To demonstrate the utility of these mapped codes, we performed a pilot study with de-identified data from asthma patient's health records. We were able to convert 83% of real-world laboratory tests into HPO-encoded phenotypes. Analysis of the LOINC2HPO-encoded data showed known and new phenotypes were are enriched in asthma patients such as eosinophilia and several other abnormal laboratory measurements. This preliminary evidence suggests that LOINC data converted to HPO can be used for machine learning approaches to support genomic phenotype-driven diagnostics for rare disease patients, and to perform EHR based mechanistic research.

Notes:

Curating the authorship of clinical records and biomedical abstracts

Fabio Rinaldi
F. Hoffmann-La Roche Ltd

The work of biomedical curation usually involves publication triage and annotation. We describe a type of biocuration task in which curators are required to have a more proactive role in both seeking information and using deductive reasoning. The task involves determining whether the authors of two different types of scientific biomedical documents are in fact the same person. The documents involved are MEDLINE abstracts and, for the first time analyzed in this study, ClinicalTrials.gov records. Determining the scientific authorship of a clinical finding is important to certify its validity or to gather additional information concerning it. Scientific author names, however, can be highly ambiguous and information about affiliation is often lacking in both MEDLINE and ClinicalTrials.gov. Thus, for this task we encouraged curators to seek information in the internet and make judgments based on the outcome of those searches. For their preparation we gave them methodological training to find adequate information resources and to reason over all available information. In setting up the task, we evaluated both crowdsourcing and expert curators. Crowdsourcing curators performed rather poorly, even those with a trusted track record in past curation tasks. Expert curators with appropriate training, on the other hand, were able to seek information on the internet effectively and performed with over 94% agreement. We additionally checked their judgments by emailing a set of scientific authors directly and the responses we received were in agreement with those of our curators in 98% of the cases. Thus, our experience shows that even in this apparently simplistic assignment motivated curators are necessary to independently gather appropriate information resources and produce correct annotations in a proactive fashion. Parts of this work have been approved for publication at the journal JAMIA.

Notes:

Session 4 - Database journal sessions

Annotation of gene product function from high-throughput studies using the Gene Ontology

Helen Attrill
FlyBase, University of Cambridge

High-throughput studies constitute an essential and valued source of information for researchers. However, high-throughput experimental workflows are often complex, with multiple data sets that may contain large numbers of false positives. The representation of high-throughput data in the Gene Ontology (GO) therefore presents a challenging annotation problem, when the overarching goal of GO curation is to provide the most precise view of a gene's role in biology. To address this, representatives from annotation teams within the GO Consortium reviewed high-throughput data annotation practices. We present an annotation framework for high-throughput studies that will facilitate good standards in GO curation and, through the use of new high-throughput evidence codes, increase the visibility of these annotations to the research community.

Notes:

Validation of protein-protein interactions in databases and resources: the need to identify interaction detection methods that provide binary or indirect experimental evidences

Javier de Las Rivas
Bioinformatics and Functional Genomics Group
Cancer Research Center (CiC-IBMCC, CSIC/USAL)

The collection and integration of all the known protein–protein physical interactions within a proteome framework are critical to allow proper exploration of the protein interaction networks that drive biological processes in cells at molecular level. APID Interactomes is a public resource of biological data (http://apid.dep.usal.es) that provides a comprehensive and curated collection of `protein interactomes' for more than 1100 organisms, including 30 species with more than 500 interactions, derived from the integration of experimentally detected protein-to-protein physical interactions (PPIs). We have performed an update of APID database including a redefinition of several key properties of the PPIs to provide a more precise data integration and to avoid false duplicated records. This includes the unification of all the PPIs from five primary databases of molecular interactions (BioGRID, DIP, HPRD, IntAct and MINT), plus the information from two original systematic sources of human data and from experimentally resolved 3D structures (i.e. PDBs, Protein Data Bank files, where more than two distinct proteins have been identified). Thus, APID provides PPIs reported in published research articles (with traceable PMIDs) and detected by valid experimental interaction methods that give evidences about such protein interactions (following the `ontology and controlled vocabulary': www.ebi.ac.uk/ols/ontologies/mi; developed by `HUPO PSI-MI'). Within this data mining framework, all interaction detection methods have been grouped into two main types: (i) `binary' physical direct detection methods and (ii) `indirect' methods. As a result of these redefinitions, APID provides unified protein interactomes including the specific `experimental evidences' that support each PPI, indicating whether the interactions can be considered `binary' (i.e. supported by at least one binary detection method) or not.

Notes:

An enhanced workflow for variant interpretation in UniProtKB/Swiss-Prot improves consistency and reuse in ClinVar

Maria Livia Famiglietti
SIB Swiss Institute of Bioinformatics

Personalized genomic medicine depends on integrated analyses that combine genetic and phenotypic data from individual patients with reference knowledge of the functional and clinical significance of sequence variants. Sources of this reference knowledge include the ClinVar repository of human genetic variants, a community resource that accepts submissions from external groups, and UniProtKB/Swiss-Prot, an expert curated resource of protein sequences and functional annotation. UniProtKB/Swiss-Prot provides knowledge on the functional impact and clinical significance of over 30,000 human protein coding sequence variants, curated from peer reviewed literature reports.Here we present a pilot study that lays the groundwork for the integration of curated knowledge of protein sequence variation from UniProtKB/Swiss-Prot with ClinVar. We show that existing interpretations of variant pathogenicity in UniProtKB/Swiss-Prot and ClinVar are highly concordant, with 88% of variants that are common to the two resources having interpretations of clinical significance that agree. Re-curation of a subset of UniProtKB/Swiss-Prot variants according to ACMG guidelines using ClinGen tools further increases this level of agreement, mainly due to the reclassification of supposedly pathogenic variants as benign, based on newly available population frequency data. We have now incorporated ACMG guidelines and ClinGen tools into the UniProtKB curation workflow, and routinely submit variant data from UniProtKB/Swiss-Prot to ClinVar. These efforts will increase the usability and utilization of UniProtKB variant data and will facilitate the continuing (re)evaluation of clinical variant interpretations as datasets and knowledge evolve.

Notes:

Increased Interactivity and Improvements to the GigaScience Database, GigaDB

Christopher Hunter
GigaScience, BGI

With a large increase in the volume and type of data archived in GigaDB since its launch in 2011, we have studied the metrics and user patterns to assess the important aspects needed to best suit current and future use. This has led to new front-end developments and enhanced interactivity and functionality that greatly improves user experience.In this article, we present an overview of the current practices including the Biocurational role of the GigaDB staff, the broad usage metrics of GigaDB datasets, and an update on how the GigaDB platform has been overhauled and enhanced to improve the stability and functionality of the codebase. Finally, we report on future directions for the GigaDB resource.

Database URL: http://gigadb.org/

Notes:

Towards comprehensive annotation of Drosophila melanogaster enzymes in FlyBase

Steven Marygold
FlyBase, University of Cambridge

The catalytic activities of enzymes can be described using Gene Ontology (GO) terms and Enzyme Commission (EC) numbers. These annotations are available from numerous biological databases and are routinely accessed by researchers and bioinformaticians to direct their work. However, enzyme data may not be congruent between different resources, while the origin, quality and genomic coverage of these data within any one resource is often unclear. GO/EC annotations are assigned either manually by expert curators or inferred computationally, and there is potential for errors in both types of annotation. If such errors remain unchecked, false positive annotations may be propagated across multiple resources, significantly degrading the quality and usefulness of these data. Similarly, the absence of annotations (false negatives) from any one resource can lead to incorrect inferences or conclusions. We are systematically reviewing and enhancing the functional annotation of the enzymes of Drosophila melanogaster, focusing on improvements within the FlyBase (www.flybase.org) database. We have reviewed 4 major enzyme groups to date: oxidoreductases, lyases, isomerases and ligases. Herein, we describe our review workflow, the improvement in the quality and coverage of enzyme annotations within FlyBase, and the wider impact of our work on other related databases.

Notes:

ccPDB 2.0: An updated version datasets of created and compiled from Protein Data Bank

Piyush Agrawal
CSIR-Institute of Microbial Technology

ccPDB 2.0 (http://webs.iiitd.edu.in/raghava/ccpdb) is an updated version of manually curated database ccPDB that maintains datasets required for developing methods to predict the structure and function of proteins. The number of datasets compiled from literature increased from 45 to 141 in ccPDB 2.0. Similarly, the number of protein structures used for creating datasets also increased from ~74000 to ~137000 (PDB March 2018 release). ccPDB 2.0 provides the same web services and flexible tools which were present in the previous version of the database. In the updated version, links of the number of methods developed in the past few years have also been incorporated. This updated resource is built on responsive templates which is compatible with smartphones (mobile, iPhone, iPad, tablets etc.) and large screen gadgets. In summary, ccPDB 2.0 is a user-friendly web-based platform that provides comprehensive as well as updated information about datasets.

Database URL: http://webs.iiitd.edu.in/raghava/ccpdb

Notes:

Building Deep Learning Models for Evidence Classification from the Open Access Biomedical Literature

Gully Burns
Information Sciences Institute

We investigate the application of deep learning to biocuration tasks that involve classification of text associated with biomedical evidence in primary research articles. We developed a large-scale corpus of molecular papers derived from PubMed and PMC open access records and used it to train deep learning word embeddings under the GloVe, FastText, and ELMo algorithms. We applied those models to a distant supervised method classification task based on text from figure captions or fragments surrounding references to figures in the main text using a variety or models and parameterizations. We then developed document classification (triage) methods for molecular interaction papers by using deep learning mechanisms of attention to aggregate classification-based decisions over selected paragraphs in the document. We were able to obtain triage performance with an accuracy of 0.82 using a combined convolution neural network (CNN), bi-directional Long-Short Term Memory (LSTM) architecture augmented by attention to produce a single decision for triage. In this work, we hope to encourage biocuration systems developers to apply deep learning methods to their specialized tasks by repurposing large scale word embedding to apply to their data.

Notes:

Curating Gene Sets: Challenges and Opportunities for Integrative Analysis

Gaurab Mukherjee
The Jackson Laboratory

Genomic data interpretation often requires analyses that move from a gene-by-gene focus to a focus on sets of genes that are associated with biological phenomena such as molecular processes, phenotypes, diseases, drug interactions, or environmental conditions. Unique challenges exist in the curation of gene sets beyond the challenges in curation of individual genes. Here we highlight a literature curation workflow whereby gene sets are curated from peer-reviewed published data into GeneWeaver (GW), a data repository and analysis platform. We describe the system features that allow for a flexible yet precise curation procedure. We illustrate the value of curation by gene sets through analysis of independently curated sets that relate to the integrated stress response, showing that sets curated from independent sources all share significant Jaccard similarity. A suite of reproducible analysis tools is provided in GeneWeaver as services to carry out interactive functional investigation of user-submitted gene sets within the context of over 150,000 gene-sets constructed from publicly available resources and published gene lists. A curation interface supports the ability of users to design and maintain curation workflows of gene-sets, including assigning, reviewing, and releasing gene-sets within a curation project context.

Notes:

Using Deep Learning to Identify Translational Research in Genomic Medicine Beyond Bench to Bedside

T.B.A.

Tracking scientific research publications on the evaluation, utility and implementation of genomic applications is critical for the translation of basic research to impact clinical and population health. In this work, we utilize state-of-the-art machine learning approaches to identify translational research in genomics beyond bench to bedside from the biomedical literature. We apply the convolutional neural networks (CNNs) and support vector machines (SVMs) to the bench/bedside article classification on the weekly manual annotation data of the Public Health Genomics Knowledge Base database. Both classifiers employ salient features to determine the probability of curation-eligible publications, which can effectively reduce the workload of manual triage and curation process. We applied the CNNs and SVMs to an independent test set (n = 400), and the models achieved the F-measure of 0.80 and 0.74, respectively. We further tested the CNNs, which perform better results, on the routine annotation pipeline for 2 weeks and significantly reduced the effort and retrieved more appropriate research articles. Our approaches provide direct insight into the automated curation of genomic translational research beyond bench to bedside. The machine learning classifiers are found to be helpful for annotators to enhance the efficiency of manual curation.

Notes:

Integration of Macromolecular Complex Data into the Saccharomyces Genome Database

Edith Wong
Stanford University

Proteins seldom function individually. Instead, they interact with other proteins or nucleic acids to form stable macromolecular complexes that play key roles in important cellular processes and pathways. One of the goals of Saccharomyces Genome Database (SGD; www.yeastgenome.org) is to provide a complete picture of budding yeast biological processes. To this end, we have collaborated with the Molecular Interactions team that provides the Complex Portal database at EMBL-EBI to manually curate the complete yeast complexome. These data, from a total of 589 complexes, were previously available only in SGD's YeastMine data warehouse (yeastmine.yeastgenome.org) and the Complex Portal (www.ebi.ac.uk/complexportal). We have now incorporated these macromolecular complex data into the SGD core database and designed complex-specific reports to make these data easily available to researchers. These web pages contain referenced summaries focused on the composition and function of individual complexes. In addition, detailed information about how subunits interact within the complex, their stoichiometry, and the physical structure are displayed when such information is available. Finally, we generate network diagrams displaying subunits and Gene Ontology (GO) annotations that are shared between complexes. Information on macromolecular complexes will continue to be updated in collaboration with the Complex Portal team and curated as more data become available.

Website URL: www.yeastgenome.org

Notes:

Session 5 - Data standards and ontologies: Making data FAIR

Research on Metadata Standards of Biomedical Data

Jiawei Cui
Institute of Medical Information & Library, Chinese Academy of Medical Sciences

As data-intensive scientific research has become the norm, scientific data in the biomedical field exploded, which brought enormous challenges for researchers to store, manage and share data. How to construct a biomedical data metadata standard that can meet the needs of users and function targets of data repository has become an important issue. The first step of this study is to select typical metadata standards of biomedical data, such as Datacite Metadata Schema, Data Tag Suite, metadata standard of Dryad. Then establishes a comparative analysis framework for them, the content includes basic information, content design and practical application. Finally, on the basis of the current problems of existing standards, we put forward the standard construction suggestions. Through analysis, we discovery that the format of the standard is diversified, and the content is constantly updated. In content design, the content affinity and similarity of these standards provide the basis for realize interoperability of metadata standards and integrated retrieval of biomedical data. In practice application, these standards have different scope of application, characteristics and defects, service users include data producers, data managers and data users, whose roles cover all aspects of the data life cycle. And the suggestions summarized are as follows: When developing biomedical scientific data metadata standards, it should ensure that the metadata standards are compatible with the target positioning, taking into account the simplicity and complexity of the metadata standards, and referring to the existing typical metadata standards, ensuring that the description scope covers the content and format of the data, forming basic metadata elements according to the universality principle, forming extended metadata elements according to the individualization principle, using controlled vocabularies to control standardization.

Notes:

New approaches to data management: supporting FAIR data sharing at Springer Nature

Varsha Khodiyar
Springer Nature

Since 2016, academic publishers including Springer Nature, Elsevier and Taylor & Francis have been providing standard research data policies to journal authors, reflecting key aspects of the FAIR Principles' practical applications: sharing data in repositories, using persistent identifiers and citing data appropriately. In spite of the rise of FAIR and good data management practice, recent surveys found that nearly 60% of researchers had never heard of the FAIR Principles, and 46% are not sure how to organise their data in a presentable and useful way. In this presentation we will analyse the results of a white paper which assessed the key challenges faced by researchers in sharing their data, and discuss current initiatives and approaches to support researchers to adopt good data sharing practice. These include the roll-out of research data policies since 2016, as well as the launch of a Helpdesk service which has provided support to authors and allowed the research data team to capture more granular information on the challenges they face in sharing their data. We will also discuss the development of a third-party curation service which assists authors in depositing their data into appropriate repositories, and drafting data availability statements. Finally we will assess the impacts of some of these interventions, including an analysis of data availability statements and an overview of the methods authors are currently using to share their data, and how these align with FAIR.

Notes:

The ELIXIR Recommended Interoperability Resources (RIRs) - What tools can I use to make data FAIR?

Sirarat Sarntivijai
ELIXIR

The ELIXIR Interoperability Platform (EIP) is one of the five technical platforms (Compute, Data, Interoperability, Tools, and Training) that support life science knowledge discovery by coordinating bioinformatics research infrastructures across the ELIXIR Consortium's 23 member states. The EIP aims to help people and machines to discover, access, integrate, and analyse biological data by encouraging the life science community to adopt standardised file formats, metadata, vocabularies, and identifiers. To support this goal, the Recommended Interoperability Resource (RIR) selection process has been established to identify tools that are fit-for-purpose for the myriad tasks in making data FAIR. The RIRs are promoted through ELIXIR for adoption as they have been evaluated for their practicality, and to encourage community reusability practice at large. All ELIXIR resources are publicly available. The applications for RIR consideration were reviewed by ELIXIR technical experts and evaluated by non-ELIXIR experts from international communities, based on the RIR's production maturity and quality, facilitation to scientific discovery, community support and impact, and legal framework and governance. The first round of the selection process resulted in an initial portfolio of 10 RIRs, as announced on https://www.elixir-europe.org/platforms/interoperability/rirs. These resources will be regularly evaluated for quality assurance and quality control. Additional resources will be included in future calls for RIR application, as ELIXIR evolves to accommodate emerging technologies and changing scientific needs. Recommendations for use of RIRs should also be recognised as suggestions and not a mandate as we value progressive developments from our ELIXIR members based on upcoming new use case requirements.

Notes:

Expanding MIxS Genomic Minimal Information Standards

Lynn Schriml
University of Maryland School of Medicine, Institute for Genome Sciences

The Genomics Standards Consortium's (GSC, www.gensc.org) successful development and implementation of genomic metadata MIxS standards established a community-based mechanism for sharing genomic data through a common framework. The GSC, an international open-membership working body of over 500 researchers from 15 countries, promotes community-driven efforts for the reuse and analysis of contextual metadata describing the collected sample, the environment and/or the host and sequencing methodologies and technologies. Since 2005, the GSC community has deployed MIGS genome checklists and a library of 15 MIxS environmental packages to meet evolving genomic research needs to enable standardized capture of environmental, human & host associated study data. The GSC's MIxS Compliance and Interoperability working group maintains, promotes and actively co-develops additional MIxS standards. The GSC's suite of minimal information reporting guidelines have been supported (GenBank and BioSample implementations) for over a decade by the International Nucleotide Sequence Database Collaboration (INSDC) databases, namely NCBI GenBank, EMBL-EBI ENA and DDBJ, thus allowing for an enriched environmental and epidemiological description of sequenced samples. To date, over 450,000 BioSample records have been annotated with the GSC's MIxS standards. In the past two years, the GSC community has developed novel metadata standards for capturing contextual data for single-cell genomes and genomes assembled from metagenomes for bacteria and archaea and uncultivated virus genomes. These standards contribute knowledge pertinent to these research communities, enable data reuse and integration and foster cross-study data comparisons, thus addressing the critical need for consistent data representation, data sharing and promotion of interoperability in genomic Big Data. Agriculture Microbiome, Host-Parasite Microbiome and Metabolomics MIxS standards are under development for 2019.

Notes:

The Ontology for Biomedical Investigations (OBI) as a Curation Tool

Randi Vita
La Jolla Institute for Allergy & Immunology

The Ontology for Biomedical Investigations (OBI) is a community based ontology that provides terms for all aspects of biological and medical investigations. OBI is an OBO Foundry member following its principles such as reusing existing ontology terms whenever possible. The scope of OBI includes terms covering investigations such as investigation types, subject enrollment terms, material and data transformations, and importantly, experimental assay terms. These terms can be used to curate biomedical datasets and experiments described in the scientific literature. Many projects currently use OBI and here we will describe how several example projects utilize OBI in data curation to annotate and standardize metadata, including The Immune Epitope Database (IEDB), The Eukaryotic Pathogen Genomics Database Resource (EuPathDB), and NASA GeneLab as well as any curation effort that utilizes The Evidence Ontology (ECO).

Notes:

Session 6 - Interacting with the Research Community

The ELIXIR Data Platform in 2019

Rachel Drysdale
ELIXIR

ELIXIR (https://www.elixir-europe.org/) unites Europe's leading life science organisations in managing and safeguarding the increasing volume of data being generated by publicly funded research. It coordinates, integrates and sustains bioinformatics resources across its member states and enables users in academia and industry to access services that are vital for their research. There are currently 23 Nodes in ELIXIR, and we work together using a 'Hub and Nodes' model. ELIXIR's activities are coordinated across five 'Platforms': Data, Tools, Interoperability, Compute and Training. The goal of the ELIXIR Data Platform (https://www.elixir-europe.org/platforms/data) is to drive the use, re-use and value of life science data. It aims to do this by providing users with robust, long-term sustainable data resources within a coordinated, scalable and connected data ecosystem. This presentation will outline the initiatives currently underway in the ELIXIR Data Platform. Topics will include the ELIXIR Core Data Resources, selected on the basis of a set of Indicators that demonstrate their fundamental importance to the wider life-science community, and the related set of Deposition Databases for the long-term preservation of biological data. Our work on Literature-Data Integration and Scalable Curation for biocurators, which builds on our text mining work with EuropePMC, will be summarised. Our commitment to Long Term Sustainability of life science data resources, including our contribution to the Global Biodata Coalition, will also be covered. Work on all these topics will continue through the new ELIXIR Scientific Programme set for 2019-2023. Lastly, the Data Platform is currently engaged in seven Implementation Studies, involving fifteen ELIXIR Nodes working with 40 Data Resources across Europe. These studies are due to be completed mid-2019, and the tasks they are engaged in will be introduced.

Notes:

Measuring the value of data curation as a part of the publishing process

Varsha Khodiyar
Springer Nature

Journals and publishers have an important role to play in the drive to increase the reproducibility of published science. Since its launch in 2014, the Nature Research journal Scientific Data has established a reputation for publishing data papers ('Data Descriptors') that are highly reusable, as evidenced by a strong citation record. One of the ways in which Scientific Data ensures maximum reusability of published data is via the in-house data curation workflow applied to every Data Descriptor. In 2017, Springer Nature launched its Research Data Support (RDS) service to provide data curation expertise to researchers publishing at other Springer Nature journals. During curation at Scientific Data and RDS, our data editors familiarise themselves with the related manuscript and perform a thorough check of each data archive. This ensures the descriptions in the manuscript match the metadata and data at the data repositories. The curation process facilitates the identification of any discrepancies between the manuscript text and the information held at the data repository. Over the last year, the curation team have been recording the types of discrepancies rectified as a direct result of our curation process. At Scientific Data approximately 10% of the discrepancies the team find are significant enough to potentially have warranted a formal correction had the issue had not been resolved prior to publication.

Notes:

Involving researchers in the biocuration of plant genes and pathways

Sushma Naithani
Oregon State University

Curated genomic resources and databases that provide systems-level frameworks for visualization and analysis of high-throughput omics data cannot keep pace with an ongoing explosion in the generation of genomic data without the active engagement and support of the research community and other stakeholders, such as academic institutions, grant agencies, and publishers. Researchers who participate in biocuration activities can acquire Big Data literacy and additional analytical skills useful for conducting and publishing their own research. Based on a few on-site and online biocuration activities organized by Plant Reactome (http://plantreactome.gramene.org) curators, we will discuss the strategy, workflow, and outcomes of our efforts in involving researchers and database users in the curation of plant genes and pathways. The Plant Reactome database is funded by NSF award #IOS-1127112 to the Gramene project. It is produced with intellectual and infrastructure support provided by the Human Reactome award (NIH: P41 HG003751, ENFIN LSHG-CT-2005-518254), Ontario Research Fund, and European Bioinformatics Institute (EBI) Industry Programme).

Notes:

Coordination and Collection of Data for a Community Global Biodiversity Initiative

Jeena Rajan
EMBL-EBI

The UniEuk project (https://unieuk.org/project/) is an open, inclusive, community-based and expert driven international initiative to build a flexible, adaptive universal taxonomic framework for eukaryotes. It unites three complementary modules, EukBank, EukMap and EukRef, which use environmental metabarcoding surveys, phylogenetic markers and expert knowledge to inform the taxonomic framework. I will focus here on the Eukbank module, the aim of which is to standardise observations of global eukaryotic diversity across biomes (e.g., saturation, relative frequencies, phylogeny), and allow identification and preliminary naming of novel eukaryotic lineages of ecological and phylogenetic relevance. This module involves analysis of high-throughput metabarcoding datasets using V4 regions of 18S rRNA from various ecosystems. The European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) has been directly involved in the collection and coordination of data for EukBank. In late 2017 a letter was circulated among the protist community calling for 18S V4 rRNA metabarcoding datasets. Data submitters were encouraged to contact ENA and associate their public and confidential data to a "datahub" at ENA accessible by the EukBank team. A dedicated sample checklist was created for this project to ensure the capture of high-quality metadata standards. There was an excellent response from the community with 190 contributors submitting approximately 160 datasets. The data have geographic coverage across various planetary biomes from oceans to freshwater environments, soils, and forests. Preliminary analysis has begun on the data and final results are expected to be available by early 2019. In this presentation, I will cover the mechanisms put into place to mobilise a community and their data around a particular scientific challenge; describe how these leverage an existing core data resource and community standards and deliver FAIR practice; and preview the key outputs so far from the analysis.

Notes:

Research on the Collaborative Mechanisms for Journals and Data Repositories in the Integrated Publication of Papers and Scientific Data

Jinming Wu
Institute of Medical Information & Library, Chinese Academy of Medical Sciences

OBJECTIVE: To clarify collaborative mechanisms for journals and data repositories in the integrated publication of papers and scientific data by investigating well-known journals and data repositories in the field of biomedicine, and to provide reference for the optimization of this mechanism, which thereby promote the integration, management and sharing of scientific data.

METHODS: Ten well-known journals were selected to survey, including BioMed Central, Nature, Scientific Data, etc. Ten typical data repositories were selected as well, including GenBank, UniProt, Dryad, Genome Sequence Archive and so on. The research compares and analyzes how different journals and data repositories carry out the integrated submission, associated review, linked release, synchronized update, joint reference of papers and data by the mutual authentication and collaborative cooperation.

RESULTS: Three different collaborative mechanisms were summarized, and the integrated publication workflow of papers and data in three different situations will be detailed and comparatively analyzed in the text.

CONCLUSION: Finally, We have provided suggestions for further cooperation between journals and repositories, such as establishing an integrated submission platform that allows authors to submit data and papers from one entry; conducting a dual data review in a semi-manual or semi-automatic manner; collecting as much detailed metadata as possible, and automatically building bidirectional linking between papers and data; improving joint reference specifications for papers and data, and so on.

Notes:

Posters

Abstracts are available online at https://www.biocuration2019.org/posters

Functional Annotation

#	Poster Title	Presenter	Affiliation
1	FlyBase: A Valuable Source of Molecular Interaction Data	Agapite, Julie	FlyBase, Harvard University
2	The SwissLipids knowledge resource for lipid biology	Aimo, Lucila	Swiss-Prot, SIB
3	Adding knowledge to the UniProt resource by proteomics and genomics integration	Alpi, Emanuele	EMBL-EBI
4	An evidence-based model for representing signaling pathways in FlyBase.	Antonazzo, Giulia	FlyBase, University of Cambridge
5	Collaborative curation of antigen presentation and recognition in UniProtKB/Swiss-Prot with IMGT®	Argoud-Puy, Ghislaine	SIB Swiss Institute of Bioinformatics
6	Automated generation of modular and standardized gene descriptions using structured data at the Alliance of Genome Resources	Arnaboldi, Valerio	WormBase
7	Enhanced enzyme annotation in UniProtKB using Rhea	Axelsen, Kristian	SIB Swiss Institute of Bioinformatics
8	Using Wikidata for community engagement and display of curated Plasmodium genomes	Böhme, Ulrike	Wellcome Sanger Institute
9	A portable annotation pipeline for genomes and proteomes – HAMAP as SPARQL rules	Bolleman, Jerven	SIB Swiss Institute of Bioinformatics
10	InterPro2COGs: A mapping between InterPro and Clusters of Orthologous Genes (COGs)	Chang, Hsin-Yu	EMBL-EBI

11	MetaboLights Open Access Metabolomics Resource	Cochrane, Keeva	EBI
12	Capturing phenotypes for inclusion in a multi-species interaction database	Cuzick, Alayne	Rothamsted Research
13	Annotation of tRNA modifications genes in model organisms	de Crecy-Lagard, Valerie	University of Florida
14	An Evaluation of Gene Ontology Annotation of Gene Products Associated with Immunological Processes	Diehl, Alexander	University at Buffalo
15	Functional annotation of the population-specific variations from whole genome analyses of a Chinese population	Du, Zhenglin	Beijing Institute of Genomics, Chinese Academy of Sciences
16	The Chinese Genomic Variation Database: an integrated genomic variation repository for Chinese populations	Du, Zhenglin	Beijing Institute of Genomics, Chinese Academy of Sciences
17	Pfam and MGnify: using metagenomics to improve the Pfam coverage of microbial sequence space	El-Gebali, Sara	EMBL-EBI
18	Capra hircus and Ovis aries IGK loci: simultaneous annotation in IMGT®	Folch, Géraldine	IMGT - IGH - CNRS
19	The challenges of annotation and integration of scRNA-Seq into Bgee	Fonseca Costa, Sara	University of Lausanne
20	AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors	Guo, An-Yuan	Huazhong Univ. of Sci. and Tech
21	Integrating expression intersections into FlyBase and Virtual Fly Brain	Holmes, Alex	FlyBase/Virtual Fly Brain
22	CROssBAR: Comprehensive Resource of Biomedical Relations with Deep Learning and Network Representations	Joshi, Vishal	EMBL-EBI

23	ViralZone: recent updates to the virus knowledge resource.	Le Mercier, Philippe	SIB Swiss Institute of Bioinformatics
24	SIGNOR and DISNOR: two sister databases for the analysis of causal relationships whose disruption underlies genetic diseases.	Licata, Luana	University of Rome Tor Vergata
25	Diverse taxonomies for diverse chemistries: enhanced plant and fungal metabolic pathway annotation for natural product biosynthesis in UniProtKB/Swiss-Prot	Lieberherr, Damien	Swiss-Prot Group, SIB Swiss Institute of Bioinformatics
26	Functional annotation of dementia-related miRNAs using the Gene Ontology	Lovering, Ruth	University College London
27	LncBook: a curated knowledgebase of human long non-coding RNAs	Ma, Lina	Beijing Institute of Genomics, CAS
28	NLM's Conserved Domain Database (CDD): current curation efforts	Marchler-Bauer, Aron	National Institutes of Health
29	Yeast Complexome - The Complex Portal rising to the challenge	Meldal, Birgit	EMBL-EBI
30	Database Resources of the BIG Data Center	Memebers, BIG Data Center	BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences
31	IMGT® genomic annotation of the dog (Canis lupus familiaris) seven immunoglobulin (IG) or antibody and T cell receptor (TR) loci	Michaloud, Joumana	IMGT, the international ImMunoGeneTics information systemInstitut de Génétique Humaine IGH, UMR 2009 UM-CNRS
32	Experimental tools: a new way to categorise transgenic alleles in FlyBase.	Millburn, Gillian	

33	Rhea, an expert curated resource of biochemical reactions for enzyme annotation	Morgat, Anne	SIB Swiss Institute of Bioinformatics
34	DDBJ (DNA Data Bank of Japan) Activity	Okido, Toshihisa	Bioinformation and DDBJ center, National Institute of Genetics
35	Building non coding RNA networks in IntAct: from yeast to human	Panni, Simona	DiBEST University of Calabria
36	Capturing variation impact on molecular interactions: updates on the IMEx Consortium mutations data set	Porras Millan, Pablo	EMBL-EBI - Molecular Interactions Team
37	Annotating the complex cellular orchestration of protein functions using Genome Properties	Richardson, Lorna	EMBL-EBI
38	FungiDB: Integrating genomic data for pathogens and model organisms and providing advanced search capabilities and large-scale data analysis	Shanmugasundram, Achchuthan	University of Liverpool
39	Structured Design Patterns in the Human Disease Ontology for Enhanced Genetic Disease Classification	Sinclair, Michael S.	Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA
40	The Drosophila nuclear pore complex - a one-stop shop for all we know about it, UniProtKB	Speretta, Elena	EMBL-EBI
41	ChlamBase: a Wikidata-backed genome database for the Chlamydia research community	Su, Andrew	Scripps Research
42	Protein Structures and their features in UniProt	Tyagi, Nidhi	EMBL-EBI
43	Developing a novel approach to characterize genes	Vallat, Bastien	UNIL

	essential to the function of a tissue		
44	Functional annotations in the PDBe Knowledge Base (PDBe-KB)	Varadi, Mihaly	EMBL-EBI
45	Autophagy Targeted Curation	Varusai, Thawfeek	European Bioinformatics Institute (EMBL-EBI)
46	Experiment-based computational method for proper annotation of the molecular function of enzymes	Veronique, de Berardinis	CEA/Genoscope/UMR8030
47	Making curated genome annotations available for expression calls of RNA-Seq	Wollbrett, Julien	SIB, UNIL
48	NucMap: a database of genome-wide nucleosome positioning map across species	Xiao, Jingfa	Beijing Institute of Genomics, Chinese Academy of Sciences
49	Challenges in the annotation and identification of pseudoenzymes in UniProt Knowledgebase	Zaru, Rossana	EMBL-EBI

Curation and data visualisation tools

#	Poster Title	Presenter	Affiliation
50	Validation in the Protein Data Bank	Berrisford, John	EBI
51	Maintaining Balance in a Data Ecosystem	Bolton, Elizabeth	Rat Genome Database
52	Optimizing Collaboration and Workflow with BioAssay Express	Bunin, Barry	Collaborative Drug Discovery
53	The Genome Reference Consortium: Curation of the human, mouse and zebrafish reference genome assemblies	Collins, Joanna	Wellcome Sanger Institute

54	Enabling federated queries over heterogeneous bioinformatics databases through data integration: the case of Bgee, OMA and UniProt	de Farias, Tarcisio Mendes	University of Lausanne
55	Efficient Curation of Genome Annotations through Collaboration with Apollo	Dunn, Nathan	Lawrence Berkeley National Lab
56	IMGT/mAb-DB and IMGT/2Dstructure-DB for IMGT standard definition of an antibody: from receptor to amino acid changes	Duroux, Patrice	IMGT, IGH, CNRS
57	GEO data on Xenbase: A pipeline to curate, process and visualize genomic data for Xenopus.	Fisher, Malcolm	Cincinnati Children's Hospital Medical Center
58	Accelerating annotation of articles via automated approaches: evaluation of the neXtA5 curation-support tool by neXtProt	Gaudet, Pascale	SIB Swiss Institute fo Bioinformatics
59	Submission, archival and visualisation of single-cell sequencing data	George, Nancy	European Bioinformatics Institute, EMBL-EBI
60	The Bio-Entity Recognition for SABIO-RK Database	Ghosh, Sucheta	HITS
61	IMGT/HighV-QUEST for NGS analysis of IG and TR: statistical analysis of IMGT clonotypes (AA), novel interface and functionalities	Giudicelli, Véronique	IMGT - IGH - CNRS
62	Manually Curated Database of Rice Proteins: Semantically digitizing experimental data	Gour, Pratibha	University of Delhi, South Campus
63	UniProt Automated Annotation in 2019: Combined Expert Curation and Computation	Hatton-Ellis, Emma	EMBL-EBI

64	Linking chemical mentions to Medical Subject Headings in Full Text	Islamaj, Rezarta	National Center for Biotechnology Information
65	Mouse Models of Human Cancer Database (MMHCdb) – New Visualization Tools	Krupke, Debra M.	The Jackson Laboratory
66	Finding the data in research publications	Levchenko, Maria	EMBL-EBI
67	A deterministic algorithm to lay out reactions with nested compartments	Lorente, Pascual	European Bioinformatics Institute
68	Automatic extraction of transcriptional regulatory interactions from literature	Méndez-Cruz, Carlos-Francisco	Universidad Nacional Autónoma de México
69	MetaboLights study editor - An open-access curation tool for metabolomics studies submission and associated meta-data annotation	Nainala, Venkata Chandrasekhar	EMBL-EBI
70	Leveraging curation efforts about discarded data: proposal of a new resource to report discarded data, stemming from the case of Bgee transcriptomics annotations	Niknejad, Anne	SIB Swiss Institute of Bioinformatics - Department of Ecology and Evolution, University of Lausanne
71	The BioGRID Interaction Database: Curation strategies and new developments for capturing genome-wide CRISPR/Cas9-based screens	Oughtred, Rose	Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ
72	Visualizing protein residue conservation in InterPro and PDBe	Paysan-Lafosse, Typhaine	EBI
73	The Vertebrate Genomes Project: Curating reference genome assemblies of all 66,000 extant vertebrate species	Pelan, Sarah	Wellcome Sanger Institute
74	Exploring the non-curated literature in search of	Perfetto, Livia	EBI

	molecular interactions: IMEx's Dark Space Project		
75	MANE Select: a set of matched representative transcripts from NCBI-RefSeq and EMBL-EBI GENCODE gene sets for every human protein-coding gene.	Pujar, Shasikant	National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
76	Implementing assembly quality tools in a core data resource: the BlobToolKit – ENA case	Richards, Edward	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)
77	Bgee 14.1: providing full access to curated expression calls and their source data for 29 species	Robinson-Rechavi, Marc	University of Lausanne and SIB
78	Web-Based Interactive Tools for Semantic Annotation of Electron Microscopy Volume Segments in EMDB and EMPIAR	Salih, Osman	EMBL-EBI
79	ImexCentral: A platform for coordinated curation of interaction data within The IMEx Consortium.	Salwinski, Lukasz	UCLA
80	Feature-Viewer, a visualization tool for positional annotations on sequences	Schaeffer, Mathieu	CALIPHO group - SIB
81	Reactome Icon Library	Sevilla, Cristoffer	EMBL-EBI
82	Curation and integration of single-cell RNA-Seq data for cross-study analysis and interpretation	Sponarova, Jana	NEBION AG
83	CBSAS: a collaborative text-annotation tool for disease-centered relation extraction from biomedical text in Chinese	Sun, Yueping	Institute of Medical Information & Library, Chinese Academy of Medical Sciences/Peking Union

			Medical College, Beijing, China
84	Gene Ontology Causal Activity Modeling (GO-CAM): a semantic framework to improve the expressivity and searchability of GO annotations	Thomas, Paul D.	University of Southern California
85	A visualization system for navigating neurotoxic effects of chemicals in ChemDIS: Fipronil as a case study	Tung, Chun-Wei	Kaohsiung Medical University
86	LIPID MAPS: Lipidomics Gateway	Valdivia-Garcia, maria	Cardiff University
87	Laying the foundation for an infrastructure to support biocuration	Venkatesan, Aravind	EMBL-EBI
88	VSM-box: the multi-purpose curation interface as an open-source web component	Vercruysse, Steven	NTNU
89	Triaging PubMed literature to discover novel mutations for the Catalogue of Somatic Mutations in Cancer (COSMIC) with PubTator	Ward, Sari	Wellcome Trust Sanger Institute
90	Nightingale: a library of re-usable data visualisation components	Watkins, Xavier	European Bioinformatics Institute
91	Expanding the reach of data with new visualisation tools in GigaDB	Xiao, SiZhe	GigaScience
92	Exploring neXtProt data and beyond: A SPARQLing solution	Zahn, Monique	SIB Swiss Institute of Bioinformatics
93	CNHPP Data Portal – a database application framework for proteome-centric multi-omics projects	Zhu, Weimin	Beijing Proteomics Research Center

Curation for human health and nutrition

#	Poster Title	Presenter	Affiliation

94	COSMIC: integrating and interpreting the world's knowledge of somatic mutations in cancer	Bamford, Sally	Wellcome Trust Sanger Institute
95	The BioGRID Interaction Database: Curation of Genetic, Protein and Chemical Interactions and Post-Translational Modifications	Boucher, Lorrie	Lunenfeld-Tanenbaum Research Institute
96	The Utilization of Public Health Services and Its Influence Factors among Migrants in China's Cities	Cui, Jiawei	Institute of Medical Information & Library, Chinese Academy of Medical Sciences
97	Curating with the Clinical Community: Gene Panel Annotation in PanelApp	Foulger, Rebecca	Genomics England
98	OncoMX: a cancer biomarker resource leveraging published literature and genomics data	Holmes, Evan	George Washington University
99	Adapting Disease Vocabularies for Curation at the Rat Genome Database	Laulederkind, Stanley	Medical College of Wisconsin
100	EWAS Atlas: a curated knowledgebase of epigenome-wide association studies	Li, Mengwei	Beijing Institute of Genomics
101	UniProtKB and Alzheimer's Disease: Linking molecular defects to disease phenotype	Lussi, Yvonne	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)
102	Atlas of Cancer Signaling Network: a resource of multi-scale biological maps to study disease mechanisms	Monraz Gomez, Luis Cristobal	Institute Curie
103	Large scale variant annotation in UniProt and tools for interpreting the molecular mechanisms of disease	Nightingale, Andrew	EMBL-EBI
104	PathoPhenoDB: linking human pathogens to their disease phenotypes in support of infectious disease research	Schofield, Paul	University of Cambridge

105	BioModels, a repository of curated mathematical models	Sheriff, Rahuman	EMBL-EBI
106	CCDR: a Corpus for Chemical Disease Semantic Relations in Chinese	Sun, Yueping	Institute of Medical Information & Library, Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing, China
107	Systemic analysis and targeted Curation of Blood coagulation cascade models	Tiwari, Krishna	Babraham Institute
108	LOINC2HPO: Curation of Phenotype Data from the Electronic Health Records using the Human Phenotype Ontology	Vasilevsky, Nicole	Oregon Health & Science University
109	Curating the authorship of clinical records and biomedical abstracts	Vishnyakova, Dina	F. Hoffmann-La Roche Ltd
110	Study on the Construction of Knowledge Graph of Tumor based on Chinese Electronic Medical Records	Xiu, Xiaolei	IMICAMS

Database journal

#	Poster Title	Presenter	Affiliation
111	Annotation of gene product function from high-throughput studies using the Gene Ontology	Attrill, Helen	FlyBase, University of Cambridge
112	ccPDB 2.0: An updated version datasets of created and compiled from Protein Data Bank	Agrawal, Piyush	CSIR-Institute of Microbial Technology
113	Building Deep Learning Models for Evidence Classification from the Open Access Biomedical Literature	Burns, Gully	Information Sciences Institute
114	Validation of protein-protein interactions in databases and	de Las Rivas, Javier	Bioinformatics and Functional Genomics

	resources: the need to identify interaction detection methods that provide binary or indirect experimental evidences		Group Cancer Research Center (CiC-IBMCC, CSIC/USAL)
115	An enhanced workflow for variant interpretation in UniProtKB/Swiss-Prot improves consistency and reuse in ClinVar	Famiglietti, Maria Livia	SIB Swiss Institute of Bioinformatics
116	Increased Interactivity and Improvements to the GigaScience Database, GigaDB.	Hunter, Christopher	GigaScience, BGI
117	Towards comprehensive annotation of Drosophila melanogaster enzymes in FlyBase	Marygold, Steven	FlyBase, University of Cambridge
118	Curating Gene Sets: Challenges and Opportunities for Integrative Analysis	Mukherjee, Gaurab	The Jackson Laboratory
119	ImmunoSPdb: An Archive of Immunosuppressive Peptides	Usmani, Salman Sadullah	CSIR-Institute of Microbial Technology
120	Integrated curation and data mining for disease and phenotype models at the Rat Genome Database	Wang, Shur-Jen	Marquette University and Medical College of Wisconsin
121	Integration of Macromolecular Complex Data into the Saccharomyces Genome Database	Wong, Edith	Stanford University
122	Using Deep Learning to Identify Translational Research in Genomic Medicine Beyond Bench to Bedside	T.B.A.	

Data standards and ontologies: Making data FAIR

#	Poster Title	Presenter	Affiliation
123	Building a Better "Trap" to Capture Information on Mouse Study Populations	Bello, Susan M	Jackson Laboratory

124	The EMBL-EBI Genome Editing Catalogue	Corbett, Sybilla	EMBL EBI
125	Research on Metadata Standards of Biomedical Data	Cui, Jiawei	Institute of Medical Information & Library, Chinese Academy of Medical Sciences
126	Challenges of Capturing Large-Scale Longitudinal Mouse Phenotyping Data	Delbarre, Daniel	MRC Harwell Institute
127	Making reproducible science	Gabdank, Idan	Stanford
128	New approaches to data management: supporting FAIR data sharing at Springer Nature	Grant, Rebecca	Springer Nature
129	Building a Pharmacogenomics Knowledge Representation Model: the case of melanoma caused by BRAF gene mutation	Kang, Hongyu	Institute of Medical Information &Library, Chinese Academy of Medical Sciences
130	Improvement of mouse strain ontology in GENEVESTIGATOR	Kinsky, Slavomir	NEBION AG, odstepny zavod, Rimska 526/20, Prague, 12000, Czech Republic
131	How TBs of proteomics data can be efficiently handled and curated in the PRIDE database	Kundu, Deepti J	EMBL-EBI
132	SPARQL-Powered Search Engine and RESTful APIs for Protein Ontology Database	Li, Xiang	Center for Bioinformatics and Computational Biology, University of Delaware
133	Maximising community participation in the FAIR-sharing of data from small-scale publications	Lock, Antonia	University of Cambridge
134	ENCODE data standards maximize quality and use of high-throughput genomic data	Luo, Yunhai	Stanford University

135	InterMine: towards supporting the FAIR principles and widening integrative data analysis	Lyne, Rachel	University of Cambridge
136	International Protein Nomenclature Guidelines: helping to standardise protein naming	Magrane, Michele	EMBL-EBI
137	uPheno 2.0: Framework for standardised representation of phenotypes across species	Matentzoglu, Nicolas	The European Bioinformatics Institute (EMBL-EBI)
138	Providing semantically-rich subject and knowledge domain annotation of FAIRsharing standards, databases and policies	McQuilton, Peter	University of Oxford
139	FAIRsharing: Mapping the landscape of biocuration - what standards should you use, what kind of data is in each database, which resources are FAIR	McQuilton, Peter	University of Oxford
140	Creation and development of a marine-community data coordination service: the EMBRIC Configurator	Milano, Annalisa	EMBL-EBI
141	Virtual Fly Brain - semantic solutions for building query-able web resources for neurobiology and 3D image data	Osumi-Sutherland, David	EMBL/EBI
142	CausalTab: PSI-MITAB updated for signaling data representation and dissemination	Panneerselvam, Kalpana	EMBL-EBI
143	BioCompute: Establishing standards for communication of HTS workflows and knowledgebase curation	Patel, Janisha	George Washington University
144	The Drosophila Anatomy Ontology	Pilgrim, Clare	

145	GlyGen - Computational and Informatics Resources for Glycoscience	Ranzinger, Rene	Complex Carbohydrate Research Center
146	Standardization of glycosaminoglycan sequences binding to proteins and creation of a pipeline for the curation of protein-glycosaminoglycan interactions	Ricard-Blum, Sylvie	University Lyon 1, ICBMS, UMR 5246 CNRS
147	Gene Ontology Annotation (GOA) Database	Rodriguez Lopez, Milagros	EBI/EMBL
148	The ELIXIR Recommended Interoperability Resources (RIRs) - What tools can I use to make data FAIR?	Sarntivijai, Sirarat	ELIXIR
149	DO: The FAIR human disease ontology domain standard	Schriml, Lynn	University of Maryland School of Medicine, Institute for Genome Sciences
150	Expanding MIxS Genomic Minimal Information Standards	Schriml, Lynn	University of Maryland School of Medicine, Institute for Genome Sciences
151	Comparative genetics and genomics of mouse strains and species at Mouse Genome Informatics (MGI)	Smith, Cynthia	The Jackson Laboratory
152	iDog: an integrated resource for domestic dogs and wild canids	Tang, Bixia	BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, 100101, China
153	Evidence and Conclusion Ontology: 2019 Update	Tauber, Rebecca	University of Maryland School of Medicine
154	HGNC: promoting standardized gene names for 40 years	Tweedie, Susan	HUGO Gene Nomenclature Committee (HGNC)
155	The Ontology for Biomedical Investigations (OBI) as a Curation Tool	Vita, Randi	La Jolla Institute for Allergy & Immunology

156	The application of ontologies in analyzing the similarities of thousands of drug pairs	Wang, Zhigang	Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
157	Metadata in the Human Cell Atlas	Welter, Danielle	EMBL-EBI
158	Identifiers.org Compact Identifiers for robust data citation	Wimalaratne, Sarala	EBI
159	FAIRDOM: supporting FAIR data and model management	Wittig, Ulrike	Heidelberg Institute for Theoretical Studies
160	SABIO-RK: extraction of enzyme function data from STRENDA DB	Wittig, Ulrike	Heidelberg Institute for Theoretical Studies
161	Translation of OBO biomedical ontologies to Chinese and their visualization	Yang, Xiaolin	Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences

Interacting with the Research Community

#	Poster Title	Presenter	Affiliation
162	In the Know About GO: A Newly Redesigned Website for the Gene Ontology	Aleksander, Suzi	Stanford University
163	Because my favorite protein paper should be in UniProt	Arighi, Cecilia	Center for Bioinformatics and Computational Biology, University of Delaware
164	Combining text mining and author participation to improve curation at WormBase	Arnaboldi, Valerio	California Institute of Technology
165	Towards comprehensive quality assessment of proteome space	Britto, Ramona	European Bioinformatics Institute (EMBL-EBI)
166	How Structural Biologists and the Protein Data Bank Contributed to Recent US FDA New Drug Approvals	Burley, Stephen	RCSB Protein Data Bank

167	Gephebase, a Database of Genotype-Phenotype Relationships for natural and domesticated variation in Eukaryotes	Courtier, Virginie	CNRS - Institut Jacques Monod
168	PHI-Canto: introducing the concept of the meta-genotype to curate information on multi-species interactions	Cuzick, Alayne	Rothamsted Research
169	The ELIXIR Data Platform in 2019	Drysdale, Rachel	ELIXIR
170	Introducing Project FREYA: opportunities for biocurators	Ferguson, Christine	EMBL-EBI
171	Integration and Presentation of Glycobiology Resources in GlyGen	Fochtman, Brian	George Washington University
172	Downloading Data from SGD	Gondwe, Felix	Stanford University
173	Protein Data Bank in Europe Knowledge Base (PDBe-KB) - a new community-driven resource for functional annotations of macromolecular structures	Gutmanas, Aleksandras	EBI-EMBL
174	User-driven PomBase website redesign improves knowledge search, display, analysis, and reuse	Harris, Midori	University of Cambridge
175	ELIXIR 5 years on : Providing a coordinated European Infrastructure for Life Science Data and Services	Harrow, Jennifer	ELIXIR
176	International Mouse Phenotyping Consortium: Capturing Multidimensional Large-Scale Phenotyping Data	Keskivali-Bond, Piia	MRC Harwell Institute
177	STRENDA DB – monitoring the completeness of enzyme function data	Kettner, Carsten	Beilstein-Institut

178	Facilitating community-based curation of transcription factor binding profiles in JASPAR	Khan, Aziz	Centre for Molecular Medicine Norway (NCMM), University of Oslo, Norway
179	Measuring the value of data curation as a part of the publishing process	Khodiyar, Varsha	Springer Nature
180	The Gene Regulation Consortium (GRECO) and the COST Action GREEKC	Kuiper, Martin	NTNU
181	PomBase at a Glance	Lock, Antonia	University College London
182	Facilitating researcher engagement with curation during the data paper publication	Matthews, Tristan	Springer Nature
183	Involving researchers in the biocuration of plant genes and pathways	Naithani, Sushma	Oregon State University
184	microPublication – incentivizing authors to publish research findings in a machine readable format	Raciti, Daniela	California Institute of Technology
185	Coordination and Collection of Data for a Community Global Biodiversity Initiative	Rajan, Jeena	European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)
186	Data Integration and Visualization at the 4D Nucleome Data Portal	Reiff, Sarah	Harvard Medical School
187	BioStudies – database of biological studies	Snow, Catherine	EBI
188	User-focused development of the NHGRI-EBI Genome-Wide Association Studies Catalog	Sollis, Elliot	European Molecular Biology Laboratory, European Bioinformatics Institute
189	TriTrypDB: A web-based resource offering the improvement of structural gene models of pathogenic kinetoplastids through	Starns, David	University of Liverpool

	community annotation using Apollo.		
190	FlyBase community curation and outreach activities	Urbano, Jose M.	FlyBase-University of Cambridge
191	CNGBdb: China National GeneBank DataBase	Wei, Xiaofeng	CNGB
192	Maximising community participation in the FAIR-sharing of data from small-scale publications	Wood, Valerie	University of Cambridge
193	Research on the Collaborative Mechanisms for Journals and Data Repositories in the Integrated Publication of Papers and Scientific Data	Wu, Jinming	Institute of Medical Information & Library, Chinese Academy of Medical Sciences
194	Enabling Findability, Accessibility, Interoperability, and Reusability with Improved Data Representation of Carbohydrates in the Protein Data Bank	Young, Jasmine	RCSB Protein Data Bank

Notes:

Notes:

Notes: