ISB International Society for Biocuration

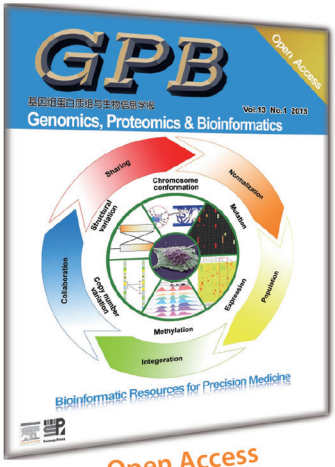
Biocuration 2015

The 8th International Biocuration Conference

From Big Data to Big Discovery

April 23–26, 2015

Beijing, China



Open Access

Submit to GPB and Enjoy

- ✓ Trending topics
- ✓ Fast and professional review
- ✓ Complimentary language service
- ✓ International readership
- ✓ Broad exposure
- ✓ APC discounts and awards
- ✓ Almetrics for article-level impact

Upcoming Special Issue Call for Papers

HOSTED BY

ELSEVIER

Metagenomics of Marine Environments (Fall, 2015)

Guest Editor: Vladimir Bajic (KAUST, Saudi Arabia)

Fangqing Zhao (BIOLS, CAS, China)

Computational Cardiology (Fall, 2015)

Guest Editor: Andreas Keller (Saarland University, Germany)

Benjamin Meder (Heidelberg University, Germany)

Big Data in Biomedicine (Spring, 2016)

Guest Editor: Xiangdong Fang (BIG, CAS, China)

Hongxing Lei (BIG, CAS, China)

creative commons

PubMed

Altmetric

GPB
Genomics, Proteomics & Bioinformatics

Genomics Frontiers Symposium

http://symposium.icgchina.org

July 16 – 19, 2015 Beijing, China

Single cell omics

Sunney Xie (Keynote, USA)

Chaoyong Yang (China)

more speakers TBA

High-dimensional structure of genome

Giacomo Cavalli (Keynote, France)

Zhijun Duan (Keynote, USA)

Guohong Li (China)

more speakers TBA

RNA modifications

Jacob Hanna (Keynote, Israel)

Akimitsu Okamoto (Keynote, Japan)

Runsheng Chen (China)

Yueqin Chen (China)

Guifang Jia (China)

Mofang Liu (China)

Yijun Qi (China)

Yun-Gui Yang (China)

Chengqi Yi ((China)

Hosted by

Beijing Institute of Genomics,
Chinese Academy of Sciences

Organized by

Genomics, Proteomics & Bioinformatics

Contact us: editor@big.ac.cn

4th International Conference of Genomics

Beyond Biological and Medical Big Data

Xi'an, China
October 23 - 25, 2015
http://www.icgchina.org

Genomics has entered a new era that bioinformatics knowledge and tools are being actively applied to a broad spectrum of research fields, including medicine, agriculture, aquaculture, environment, and energy. The amount of data being produced by sequencing, mapping, and analyzing genomes propels genomics into the realm of Big Data. Life sciences have been highly affected by the generation of large datasets, specifically by overloads of omics information (genomes, transcriptomes, epigenomes, and other omics data from cells, tissues, and organisms). Therefore, data sharing, idea exchange, interdiscipline collaboration, and direct communication are all essential for the fast developing research activities. With these in mind, Beijing Institute of Genomics, Chinese Academy of Sciences (CAS), joining Genetics Society of China and Xi'an Jiaotong University, organizes the 4th International Conference of Genomics in 2015 (ICG 2015). We would like to invite the field's leading scientists in China and from other part of the world to come together and to share their insights on the recent progress in genome science and to discuss the challenges and future directions in the relevant fields.

As the chair of the ICG 2015, I am honored to invite you to join us at this conference. We look forward to your attendance at the ICG 2015 in Xi'an!
Sincerely,

Jun Yu, Ph.D.
Chair of the 4th International Conference of Genomics
Beijing Institute of Genomics, CAS

Host

- Beijing Institute of Genomics, CAS
- Genetics Society of China
- Xi'an Jiaotong University

Conference topics

- Genomics and Big Data
- Epigenetics and Transcriptional regulation
- Forensic Genomics
- Marine Genomics and National Strategy

Welcome to Beijing and the 8th International Biocuration Conference!

Dear Colleagues,

We are delighted to host you in Beijing for the 8th International Biocuration Conference (IBC). Omics studies have accumulated a large volume of biomedical data. However, the promise to translate these big data into biomedical knowledge can be realized only if they are standardized to be interoperable, complete to ensure data integrity, and consistent in data content. Biocuration provides tools and procedures critical in realizing the promise.

IBC is the global discussion platform for biocurators, bioinformaticians, data scientists, as well as any biomedical researchers and practitioners who care about data quality and interoperability, and interested in the exploration of a great application potential by data-driven knowledge discovery. After 7 years' successful running of IBC by International Society of Biocuration, we are honored to host the 8th IBC in Beijing.

Keynote speakers for the 8th IBC include Amos Bairoch from University of Geneva, Philip Bourne from National Institutes of Health, Takashi Gojobori from King Abdullah University of Science and Technology, Michal Linial from Hebrew University of Jerusalem, Johanna McEntyre from EMBL-EBI, and Guoping Zhao from Chinese Academy of Sciences.

We thank Oxford University Press and the Editors of DATABASE for once again hosting the proceedings of the conference. As the official journal for our community, DATABASE has proven to be an excellent forum for the dissemination of biocuration-focused research. The Biocuration 2015 Virtual Issue can be accessed at: http://www.oxfordjournals.org/our_journals/databa/biocuration_virtual_issue.html.

We thank the ISB Executive Committee for valuable advice and financial support. We extend our thanks to the members of the Organizing Committee and the Scientific Committee and the Session/Workshop Chairs for their guidance and their work in helping to organize Biocuration 2015. We also thank all sponsors for their much appreciated and generous support and thank you for joining us at the 8th IBC.

We are glad you are here in Beijing and on behalf of the organizing committee, we would like to thank you for coming to Biocuration 2015 and hope you enjoy your stay in Beijing!

Sincerely yours,

Weimin Zhu
Jingchu Luo
Zhang Zhang
The 8th International Biocuration Conference Organizing Chairs

The 8th International Biocuration Conference

Beijing Friendship Hotel, Beijing, China
April 23-26, 2015

Organizing Committee

- **Weimin Zhu (Chair),** Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Jingchu Luo (Co-chair),** Peking University, China
- **Zhang Zhang (Co-chair),** Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Alex Bateman**, European Bioinformatics Institute, UK
- **Mike Cherry**, Stanford University, USA
- **Pascale Gaudet**, Swiss Institute of Bioinformatics, Switzerland
- **Melissa Haendel**, Oregon Health and Science University, USA
- **Jen Harrow**, Welcome Trust Sanger Institute, UK
- **Robin Haw**, Reactome/Ontario Institute for Cancer Research, Canada
- **Zhong Jin**, Computer Network Information Center, Chinese Academy of Sciences, Beijing
- **Jiao Li**, Institute of Medical Information, Chinese Academy of Medical Sciences, China
- **Claire O'Donovan**, European Bioinformatics Institute, UK
- **Francis Ouellette**, Ontario Institute for Cancer Research, Canada
- **Tieliu Shi**, East China Normal University, China
- **Zhigang Wang**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Jingfa Xiao**, Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Xiaolin Yang**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Meng Zhang**, Beijing Lucidus Bioinformation Co., Ltd., China

Scientific Committee

- **Yongbiao Xue (Chair)**, Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Amos Bairoch (Co-chair)**, Swiss Institute of Bioinformatics, Switzerland
- **Philip Bourne (Co-chair)**, National Institutes of Health, USA
- **Guoping Zhao (Co-chair)**, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, China
- **Cecilia Arighi**, University of Delaware, USA
- **Alex Bateman**, European Bioinformatics Institute, UK

- **Runsheng Chen**, Institute of Biophysics, CAS, China
- **Mike Cherry**, Stanford University, USA
- **Lipman David**, National Center for Biotechnology Information, USA
- **Pascale Gaudet**, Swiss Institute of Bioinformatics, Switzerland
- **Takashi Gojobori**, King Abdullah University of Science and Technology, Kingdom of Saudi Arabia
- **Bailin Hao**, Fudan University, China
- **Melissa Haendel**, Oregon Health and Science University, USA
- **Midori Harris**, European Bioinformatics Institute, Wellcome Trust Genome Campus, UK
- **Jen Harrow**, Welcome Trust Sanger Institute, UK
- **Robin Haw**, Reactome/Ontario Institute for Cancer Research, Canada
- **Fuchu He**, Beijing Proteome Research Center, China
- **Henning Hermjakob**, European Bioinformatics Institute, Wellcome Trust Genome Campus, UK
- **Li Jin**, Fudan University, China
- **Ilene Karsch-Mizrachi**, National Center for Biotechnology Information, USA
- **David Landsman**, National Center for Biotechnology Information, USA
- **Suzanna Lewis**, Lawrence Berkeley National Laboratory, USA
- **Jiao Li**, Institute of Medical Information, Chinese Academy of Medical Sciences, China
- **Yixue Li**, Shanghai Center for Bioinformation Technology, China
- **Depei Liu**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Jingchu Luo**, Peking University, China
- **Claire O'Donovan**, European Bioinformatics Institute, UK
- **Francis Ouellette**, Ontario Institute for Cancer Research, Canada
- **Kim Pruitt**, National Center for Biotechnology Information, USA
- **Boqin Qiang**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Heng Wang**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Jun Wang**, Beijing Genomics Institute (BGI), China
- **Cathy Wu**, University of Delaware, USA
- **Jun Yu,** Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Qimin Zhan**, Cancer Institute, Chinese Academy of Medical Sciences, China
- **Zhang Zhang,** Beijing Institute of Genomics, Chinese Academy of

Sciences, China

- **Weimin Zhu**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China

Secretary Team

- **Dong Zou**, Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Hao Xie**, Taicang Institute of Life Sciences Information, China
- **Lina Ma**, Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Qingshu Meng**, Beijing Institute of Genomics, Chinese Academy of Sciences, China
- **Sangang Xu**, Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, China
- **Mengjun Yang**, East China Normal University, China

Volunteers*

(*All From Beijing Institute of Genomics, Chinese Academy of Sciences, China)

- **Dandan Cao**
- **Shenghan Gao**
- **Qianqian Sun**
- **Shixiang Sun**
- **Huangkai Zhang**
- **Guangyu Wang**
- **Shuangyang Wu**
- **Lin Xia**
- **Jiayue Xu**
- **Xingjian Xu**
- **Li Yang**
- **Hongyan Yin**
- **Chunlei Yu**

Acknowledgements

- **Jennifer Harrow,** Wellcome Trust Sanger Institute, UK
- **Xiaoxin Ruan,** Beijing Friendship Hotel, China

Hosted by

- Beijing Institute of Genomics, Chinese Academy of Sciences
- Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
- Institute of Medical Information & Library, Chinese Academy of Medical Science & Peking Union Medical College
- National Scientific Data Sharing Platform for Population and Health
- Beijing Proteome Research Center
- Center for Bioinformatics, Peking University
- Computer Network Information Center, Chinese Academy of Sciences

Organized by

- Beijing Lucidus Bioinformation Co., Ltd.

Sponsors

Biocuration 2015 relies on the generosity of sponsors and exhibitors to ensure the viability of the conference. We welcome and thank these sponsors for both their financial support and their scientific contributions.

Gold Sponsors

Silver Sponsor

Bronze Sponsor

General Information

Your conference name badge will be required to gain access to all events.

Conference Venue

Beijing Friendship Hotel
No.1 Zhong Guan Cun South Road
Beijing 100873, China
Tel: +86 (10) 6849 8888
Fax: +86 (10) 6849 8866
Email: smd@BJfriendshiphotel.com

Onsite Contact Details

The Biocuration 2015 Conference Secretariat will be available every day of the conference at the conference registration/information desk in the conference venue. Information on tours around Beijing and the rest of China will be available here also.

Abstract Book

This abstract book is also available online at:
http://biocuration2015.big.ac.cn/biocuration2015book.pdf

Registration Desk Opening Times

Thursday 23 April 2015	13:00 – 18:00
Friday 24 April 2015	08:30 – 18:00
Saturday 25 April 2015	08:30 – 18:00
Sunday 26 April 2015	08:30 – 14:00

Internet Access

There is complimentary wireless internet available to all Biocuration 2015 attendees in the meeting rooms. Log onto wireless networks, then select “biocuration2015” connection and enter password “welcome2beijing”.

Catering

All catering is provided by Beijing Friendship Hotel. The program details the times of the coffee & tea and lunch breaks every day as well as the welcome reception. Alternatively, there are a number of restaurants located in and around the hotel.

Accompanying Person

If you have accompanying person who would like to attend the welcome reception (260 CNY/person) on April 23 and three lunches (120 CNY/person) on April 24~26, please book at the registration & information

desk, which is on a "first come, first serve" basis.

Pre-conference Biocuration Training

Pre-conference biocuration training courses are provided with the supports from Mark Thomas, Xiaodong Wang, Yanli Wang on April 22 at Beijing Institute of Genomics, Chinese Academy of Sciences and Ilene Mizrachi, Kim Pruitt, Claire O'Donovan and Sandra Orchard on April 23 at Beijing Friendship Hotel.

Poster Presentations

Poster sessions will be held on April 24 and 25 from 5:40pm – 7pm. Odd numbered posters will be presented on April 24 and even numbered posters on April 25.

Poster Guidelines

Posters should be printed in **A0 size** (height 118.9cm, width 84.1cm) and should be **portrait** (not landscape).

Poster Hanging & Removal

- Each poster is allocated one poster board.
- Poster boards will be numbered and a poster number will be assigned for your poster.
- Poster numbers will be indicated in the online program and on the poster board onsite.
- Posters may be put up at the assigned location on **April 24 before the first poster session**.
- Pushpins for hanging posters will be provided.
- Posters must be removed by **April 25 at 8pm**.

GPB Best Poster Awards

Six Best Poster Awards are made by a generous contribution from GPB.

- 1st Prize: 1 awardee, CNY 1000
- 2nd Prize: 2 awardees, each CNY 600
- 3rd Prize: 3 awardees, each CNY 300

Genomics, Proteomics & Bioinformatics (GPB) is the official journal of Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS) and Genetics Society of

China (GSC). As the first international journal in China that is focused on the omics and bioinformatics, GPB is indexed by PubMed/Medline, BIOSIS Previews, Chinese Science Citation Database, Chemical Abstracts, Scopus, EMBASE, etc. GPB is currently an Open Access journal produced and hosted by Elsevier and all GPB articles can be freely accessed at ScienceDirect. The Editor-in-Chief is Dr. Jun Yu. For more information, please visit http://www.journals.elsevier.com/genomics-proteomics-and-bioinformatics.

How to Vote

All conference registrants will pick up a voting ballot at the conference registration desk and vote for the 6 GPB Best Poster Awards, keeping the following issues in mind:

- Scientific value to the biocuration community
- Authors efficiency in conveying their educational message
- Innovative ideas
- Artwork and graphics

Please remember to vote by **7pm April 25, 2015**. Each participant may write a maximum of 6 Poster Numbers. The winners of the GPB Best Poster Awards will be announced on April 26 as part of the closing remarks.

Workshops

Concurrent Workshops will be held on April 24 and 25 from 16:10 - 17:40, and on April 26 from 15:50 - 17:20. Any changes to locations will be announced at the conference.

Currency and Banking

Foreign currency can be exchanged for CNY (an abbreviation for "Chinese Yuan") at the airport, banks and hotels. Major credit cards are honored at most hotels. Banks usually open at 9:00 in the morning and close at 17:00 in the afternoon. It is highly recommended you carry Chinese Yuan in cash, which will be used in small restaurants and taxi. ATM is widely available throughout the city.

Electricity

Electricity is supplied at 220V, 50Hz AC throughout China. Major hotels usually provide 115V outlet for razor.

Pre-conference Biocuration Training

Wednesday 22 April 2015

Conference Hall, Beijing Institute of Genomics, Chinese Academy of Sciences

09:00 - 10:00 **Wormbase literature curation workflow**
Xiaodong Wang
WormBase & California Institute of Technology, USA

10:00 - 11:00 **PubChem: a case study for managing big data**
Yanli Wang
National Center for Biotechnology Information, USA

11:00 - 12:00 **Overview of annotation tools and curation workflow for the GENCODE gene sets**
Mark Thomas
Wellcome Trust Sanger Institute, UK

Thursday 23 April 2015

Function Room, Building 1, Beijing Friendship Hotel

09:00 - 12:00 **Biocuration of GenBank & RefSeq**
Ilene Mizrachi and Kim Pruitt
National Center for Biotechnology Information, USA

13:30 - 16:30 **Biocuration of UniProt & Proteomics Databases**
Claire O'Donovan and Sandra Orchard
EMBL-EBI, UK

Agenda

Agenda - Thursday 23 April 2015

13:00 - 18:00 **Arrival & Registration**

17:00 - 17:50 **Plenary Session, Chaired by Alex Bateman**
Function Room, Building 1

Keynote Lecture 1
Biocuration: Driven by Functional Studies for Biology
Guoping Zhao
Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, China

18:00 - 21:00 **Welcome Reception**
Chaired by Weimin Zhu, Jingchu Luo and Zhang Zhang
Ya Shi Ting, Building 1

18:00 - 18:15 **Introduction of National Scientific Data Sharing Platform for Population and Health**
Depei Liu
Chinese Academy of Medical Sciences, China

18:15 - 21:00 **Buffet Dinner**

Agenda - Friday 24 April 2015

08:30 - 08:50	**Welcome and Opening Remarks** Chaired by Weimin Zhu JuYing Ballroom, Friendship Palace
	Tieniu Tan Deputy Secretary-General of Chinese Academy of Sciences
	Alex Bateman Chair of International Society for Biocuration
	Weimin Zhu, Jingchu Luo, and Zhang Zhang Organizing Chairs of Biocuration2015
08:50 - 10:10	**Plenary Session, Chaired by Weimin Zhu** **JuYing Ballroom, Friendship Palace**
08:50 - 09:40	Keynote Lecture 2 **Biocuration in the community** Alex Bateman, *EMBL-EBI, UK*
09:40 - 10:10	Invited Lecture 1 **Challenges and Practices of Big Data in Life Science** Yixue Li, *Shanghai Center of Bioinformation Technology, China*
10:10 - 10:40	**Group Photo and Coffee & Tea Break**
10:40 - 12:00	**Session 1: Curation and Data annotation: from molecules to communities** Chaired by Suzanna Lewis, Ilene Mizrachi JuYing Ballroom, Friendship Palace
10:40 - 11:00	**The UniRule system to enable scaling from manual curation to large data sets** Claire O'Donovan, *EMBL-EBI, UK*
11:00 - 11:20	**Functional curation of sequence data for RefSeq** Kim Pruitt, *NCBI, USA*
11:20 - 11:40	**MyVariant.info: community-aggregated variant annotations as a service** Chunlei Wu, *The Scripps Research Institute, USA*

Agenda - Friday 24 April 2015

11:40 - 12:00	**APOLLO: SCALABLE & COLLABORATIVE CURATION OF GENOMES** Monica Munoz-Torres, *Lawrence Berkeley National Laboratory, USA*
12:00 - 13:30	**Lunch** Ju He Yuan, Friendship Palace
13:30 - 14:50	**Plenary Session, Chaired by Jingchu Luo** **JuYing Ballroom, Friendship Palace**
13:30 - 14:20	Keynote Lecture 3 **neXtProt: recent developments in the context of biocuration** Amos Bairoch, *Swiss Institute of Bioinformatics, Switzerland*
14:20 - 14:50	Invited Lecture 2 **Detangling transcriptional complexity in GENCODE using cutting-edge transcriptomics and proteomics data** Jennifer Harrow, *Wellcome Trust Sanger Institute*
14:50 - 15:50	**Session 2: Curation: from genotype to phenotype** Chaired by Ming Qi JuYing Ballroom, Friendship Palace
14:50 - 15:10	**Large-scale semantic mining of disease-phenotype annotations** Robert Hoehndorf, *King Abdullah University of Science and Technology, Saudi Arabia*
15:10 - 15:30	**Annotation of functional impact of missense mutations in BRCA1** Pascale Gaudet, *Swiss Institute of Bioinformatics, Switzerland*
15:30 - 15:50	**PhenoMiner: a quantitative phenotype database for the laboratory rat, Rattus norvegicus. Application in hypertension and renal diseases** Shur-Jen Wang, *Medical College of Wisconsin, USA*
15:50 - 16:10	**Coffee & Tea Break**

Agenda - Friday 24 April 2015

16:10 - 17:40	**Concurrent Workshops**
	Workshop 1: Crowd/community curation: challenges & credit attribution Chaired by Henning Hermjakob JuYing Ballroom, Friendship Palace
	Workshop 2: Data visualization & annotation Chaired by Rama Balakrishnan and Monica Munoz-Torres Conference Room 4, Friendship Palace
17:40 - 19:00	**Poster Session 1** Conference Rooms 1, 2, and 3, Friendship Palace

Agenda - Saturday 25 April 2015

09:00 - 10:20 **Plenary Session, Chaired by Yixue Li**
JuYing Ballroom, Friendship Palace

09:00 - 09:50 Keynote Lecture 4
Big data to small data and back again: integrating biological data with the open access literature
Johanna McEntyre, *EMBL-EBI*

09:50 - 10:20 Invited Lecture 3
Interoperable metadata leads to integrative analyses
Mike Cherry, *Stanford University, USA*

10:20 - 10:40 **Coffee & Tea Break**

10:40 - 12:00 **Session 3: Biological database, tool & framework**
Chaired by Vladimir Bajic
JuYing Ballroom, Friendship Palace

10:40 - 11:00 **The human proteome in UniProtKB**
Sylvain Poux, *SIB Swiss Institute of Bioinformatics, Switzerland*

11:00 - 11:20 **VCGDB: a dynamic genome database of the Chinese population**
Jiayan Wu, *Beijing Institute of Genomics, CAS, China*

11:20 - 11:40 **Epidaurus: A Platform for Aggregation and Integration Analysis of the Epigenome**
Liguo Wang, *Mayo Clinic, USA*

11:40 - 12:00 **The Bgee database: large-scale multi-species expression data**
Frederic Bastian, *SIB Swiss Institute of Bioinformatics & University of Lausanne, Switzerland*

12:00 - 13:30 **Lunch**
Ju He Yuan, Friendship Palace

Agenda - Saturday 25 April 2015

13:30 - 14:50 **Plenary Session, Chaired by Juncai Ma**
JuYing Ballroom, Friendship Palace

13:30 - 14:20 Keynote Lecture 5
Challenges in the Future and Experiences in the Past of Genome Annotation
Takashi Gojobori, *King Abdullah University of Science and Technology, Saudi Arabia*

14:20 - 14:50 Invited Lecture 4
Disease causing repeats and novel repeats in eukaryotic proteins
Xiu-Jie Wang, *Institute of Genetics and Developmental Biology, CAS, China*

14:50 - 15:50 **Session 4: Drug & disease**
Chaired by Yanli Wang
JuYing Ballroom, Friendship Palace

14:50 - 15:10 **Standardization and global knowledge exchange in metabolomics**
Reza Salek, *EMBL-EBI*

15:10 - 15:30 **BioExpress: An integrated RNA-seq derived gene expression database for pan-cancer analysis**
Yang Pan, *The George Washington University, USA*

15:30 - 15:50 **Bridge up chemical genomics and genomic information resources at NCBI**
Yanli Wang, *NCBI, USA*

15:50 - 16:10 **Coffee & Tea Break**

16:10 - 17:40 **Concurrent Workshops**

Workshop 3: Biocuration in big data to knowledge: new strategy, process & framework
Chaired by Francis Ouellette
JuYing Ballroom, Friendship Palace

Workshop 4: International collaboration in biocuration: projects & data/expertise sharing
Chaired by Claire O'Donovan and Sandra Orchard
Conference Room 4, Friendship Palace

Agenda - Saturday 25 April 2015

17:40 - 19:00 **Poster Session 2**
Conference Rooms 1, 2, and 3, Friendship Palace

19:00 - 21:00 **Special Workshop: Biocuration in China: importance, road map & priority**
Chaired by Weimin Zhu, Jingchu Luo and Zhang Zhang
Conference Room 5, Friendship Palace

Agenda - Sunday 26 April 2015

09:00 - 10:20 **Plenary Session, Chaired by Zhang Zhang**
Ya Shi Ting, Building 1

09:00 - 09:50 Keynote Lecture 6
Genomes: The good, the bad and the ugly - the limits of automatic biocuration
Michal Linial, *The Hebrew University, Israel*

09:50 - 10:20 Invited Lecture 5
Big and better but no junk
Jun Yu, *Beijing Institute of Genomics, CAS, China*

10:20 - 10:40 **Coffee & Tea Break**

10:40 - 12:00 **Session 5: Data standards & Ontologies**
Chaired by Sandra Orchard and Pascale Gaudet
Ya Shi Ting, Building 1

10:40 - 11:00 **Ontology application and use at the ENCODE DCC**
Venkat Malladi, *Stanford University, USA*

11:00 - 11:20 **Applying OBO Foundry ontologies to model, annotate and query longitudinal field studies on malaria**
Jie Zheng, *University of Pennsylvania, USA*

11:20 - 11:40 **Phylogenetic- based gene function prediction in the Gene Ontology Consortium**
Huaiyu Mi, *University of Southern California, USA*

11:40 - 12:00 **Standards for public health genomic epidemiology to improve infectious disease outbreak detection and investigation**
Melanie Courtot, *Simon Fraser University & BC Public Health Microbiology and Reference Laboratory, Canada*

12:00 - 13:30 **Lunch**
Ju He Yuan, Friendship Palace

Agenda - Sunday 26 April 2015

13:30 - 14:30 **Session 6: Literature mining**
Chaired by Zhiyong Lu and Cecilia Arighi
Ya Shi Ting, Building 1

13:30 - 13:50 **TextpressoCentral: A universal portal to search and curate biological literature**
Yuling Li, *California Institute of Technology, USA*

13:50 - 14:10 **Construction of Phosphorylation Interaction Networks by Text Mining of Full-length Articles using the eFIP System**
Cecilia Arighi, *PIR & University of Delaware, USA*

14:10 - 14:30 **SourceData: integrating biocuration within the publishing process**
Thomas Lemberger, *EMBO, Germany*

14:30 - 15:30 **Concurrent Special Sessions: Lightning Talks**

Special Session I: Chaired by Jingchu Luo
Ya Shi Ting, Building 1

14:30 - 14:40 **Bioso! – A Search Engine & Annotation Framework for Biological Big Data**
Yin Huang, *Beijing Proteome Research Center, China*

14:40 - 14:50 **Improved Data Representation of Very Large Macromolecules at Protein Data Bank**
Jasmine Young, *RCSB Protein Data Bank, USA*

14:50 - 15:00 **An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools**
Anyuan Guo, *Huazhong University of Science and Technology, China*

15:00 - 15:10 **Gene Curation Software at the Rat Genome Database: Update 2015**
Stanley Laulederkind, *Medical College of Wisconsin, USA*

15:10 - 15:20 **dbPSP: a curated database for protein phosphorylation sites in prokaryotes**
Zexian Liu, *Huazhong University of Science and Technology*

15:20 - 15:30 **Shared Resources, Shared Costs Leveraging Biocuration Resources**
Sandra Orchard, *EMBL-EBI*

Agenda - Sunday 26 April 2015

Special Session II: Chaired by Zhang Zhang
Conference Room 4, Friendship Palace

14:40 - 14:50 **Improving the Consistency of Domain Annotation within the Conserved Domain Database (CDD)**
Myra Derbyshire, *NCBI, USA*

14:50 - 15:00 **Automated collection, curation and analysis for Transcriptional and Epigenetic Regulation Data**
Hanfei Sun, *Tongji University, China*

15:00 - 15:10 **mycoCLAP, the Database for Characterized Lignocellulose-Active Proteins of Fungal Origin: Resource and Text Mining Curation Support**
Gregory Butler, *Concordia University, Canada*

15:10 - 15:20 **dbPPT: a comprehensive database of protein phosphorylation in plants**
Han Cheng, *Huazhong University of Science and Technology, China*

15:20 - 15:30 **Generating a focused view of Disease Ontology cancer terms for pan-cancer data integration and analysis**
Raja Mazumder, *The George Washington University, USA*

15:30 - 15:50 **Coffee & Tea Break**

15:50 - 17:20 **Concurrent Workshops**

Workshop 5: Genotype-2-phenotype: Curation challenges in translational & reverse translational informatics
Chaired by Stanley Laulederkind
Ya Shi Ting, Building 1

Workshop 6: Money for biocuration: strategies, ideas & funding
Chaired by Renate Kania, Ulrike Wittig
Conference Room 4, Friendship Palace

17:20 - 17:40 **Closing Session: Discussions, Awards and Closing Remarks**
Chaired by Weimin Zhu
Ya Shi Ting, Building 1

Workshops

Workshops

Workshop 1: Crowd/community curation: challenges & credit attribution

Chair: Henning Hermjakob, EMBL-EBI, UK

Crowd sourcing and community annotation are regarded as a valuable complement to professional curators to improve the annotation and structuring of the vast and rapidly growing amount of biological and biomedical data. However, the annotation strategy based purely on voluntary contributions of many almost anonymous editors, so successful for Wikipedia, does often not work well for biomedical and biomolecular community resources. Highly specialised resources often have a limited number of potential editors, busy domain experts, who are under pressure to build their scientific careers through attributable scientific products, typically classic publications, and who see limited gain in contributing to community resources.

Professional curators are often not able to systematically reference and demonstrate the work they have done for a specific resource, because attribution of the content of large databases to specific curators is far less developed than for example author attribution on open source software development. In this workshop, we will discuss current approaches to community and professional curation, with a specific focus on incentives and attribution of contribution to specific authors.

Presentations and perspectives, panelists/presenters:

- **Henning Hermjakob**, EMBL-EBI, Cambridge, UK
- **Xiao Si Zhe and Chris Hunter**, Gigascience
- **Todd Taylor**, RIKEN Center for Integrative Medical Sciences, Japan
- **Elvira Mitraka**, University of Maryland School of Medicine, USA
- **Jennifer Polson, Anders Garlid, Tevfik Umut Dincer**, University of California at Los Angeles, USA
- **Chunlei Wu**, The Scripps Research Institute, USA
- **Thomas Lemberger**, EMBO

The first part of the workshop will consist of short presentations from submitted abstracts, followed by a panel discussion.

Workshop 2: Data visualization & annotation

Chairs: Rama Balakrishnan, Stanford University, USA and **Monica Munoz-Torres**, Lawrence Berkeley National Laboratory, USA

Explaining the most intricate biological processes often requires a degree of detail beyond the scope of equations and algorithms; in fact, most biological knowledge is represented visually as illustrations, graphs, and diagrams. Genomics data in particular require specialized forms of visualization to improve our understanding and increase our chances of extracting meaningful conclusions from our analyses. Furthermore, the heterogeneity and abundance of genomic data include widely varied sources, techniques for their obtention, and intrinsic experimental error. And even data obtained under similar conditions from two or more individuals are loaded with biological variation. So what is the best way to interpret the stories the data are telling us? Given the questions we wish to answer and the data we are generating, which tools would be most useful and effective? In this workshop we will explore the tools available for human interpretation of genomic data, specifically in the context of annotation.

The workshop will include a brief introduction to a landscape of tools available - as updated as the constantly changing field allows-, brief presentations chosen from abstract submissions and invited speakers, as well as ample discussion to capture the contributions and questions from attendants. In the end, we expect participants to walk away with a toolset in hand that may benefit the progress of their own research.

Workshop 3: Biocuration in big data to knowledge: new strategy, process & framework

Chair: Francis Ouellette, Ontario Institute for Cancer Research, Canada

While data sets that are large continue to challenge how we manage and present the data. Big Data, or Data Science has recently been recognized as a discipline. We will discuss different approaches and ways to address the analysis, capture, curation, searching and sharing of Big Data in this workshop. The chair will be prodding panelists on their opinion of this important new topic, and out it can affect biocuration activities worldwide.

Workshop Agenda:

1. Introduction and perspectives from co-chairs: Francis Ouellette, Ontario Institute for Cancer Research, ON, Canada.
2. Perspectives from panelists, and questions from the chair:
   - **Johanna McEntyre**, EMBL-EBI, UK
   - **Michael Cherry**, Stanford University, USA
   - **Takashi Gojobori**, King Abdullah University of Science and Technology, Saudi Arabia
   - **Suzi Lewis**, Berkeley Bioinformatics Open-source Projects
   - **Raja Mazumder**, The George Washington University, USA
3. Open discussion and question period (from chair and audience)
4. Closing comments.

Workshop 4: International collaboration in biocuration: projects & data/expertise sharing

Chairs: Claire O'Donovan and Sandra Orchard, EMBL-EBI, UK

The manual curation of the information of data in biomedical resources is an expensive task. Sharing curation effort is a model already being adopted by several data resources including the Gene Ontology annotation effort and the IntAct molecular interaction database. In this workshop we will present examples of such efforts and will then like to open the floor for discussion on how to further develop such models and to explore how databases can make cost savings by sharing infrastructure and tool development to ensure that the data generated by public funding can have a greater longevity and minimise redundant development of resources by multiple disparate groups.

Presentations and perspectives, panelists/presenters:

- **Sandra Orchard**, EMBL-EBI, UK
- **Rama Balakrishnan**, Stanford University, USA
- **Claire O'Donovan**, EMBL-EBI, UK
- **Yuling Li**, California Institute of Technology, USA

Workshop format: ~45 minutes short talks, ~45 minutes open floor discussion

Workshop 5: Genotype-2-phenotype: Curation challenges in translational & reverse translational informatics

Chair: Stanley Laulederkind and Shur-Jen Wang, Medical College of Wisconsin, USA

This workshop will focus on curation of diseases in model organisms and in human populations. The process of gathering the data and presenting the information in various databases will be discussed. The advantages and disadvantages of comparing animal models of disease with human disease will be debated.

Presentations and perspectives, panelists/presenters:

- **Elvira Mitraka**, University of Maryland/IGS, USA
- **Hong Sun**, Shanghai Center for Bioinformation Technology, China
- **Stacia Engel**, Stanford University, USA
- **Li Ni**, The Jackson Laboratory, USA
- **Stanley Laulederkind**, Medical College of Wisconsin, USA

Workshop format: ~45 minutes short talks, ~45 minutes open floor discussion

Workshop 6: Money for biocuration: strategies, ideas & funding

Chairs: Renate Kania and Ulrike Wittig, Heidelberg Institute for Theoretical Studies, Germany

Even though funding research infrastructures becomes more widespread, maintaining and growing existing databases is problematic, even for well-established databases. The recognition of the importance of curation still lags behind that of data generation and this is also reflected by funding policies. In this workshop we will discuss what the International Society for Biocuration (ISB), assisted by us curators, could do to improve the situation. Would it for example be appropriate to generate a letter of recommendation to increase funders' awareness of the importance of biocuration, or publish a comment article advocating explicit funding for biocuration? Or what other kind of action(s) could be helpful? Additionally alternative business models besides licensing like e.g. pay-per use or crowd funding, are planned to be discussed.

To get an overview about the current financial situation of ISB members databases, we would like you to participate in this survey. The statistics of the survey will be presented at the workshop.

Presentations and perspectives, panelists/presenters:

- **Donghui Li**, Carnegie Institution & TAIR, USA
- **Alex Bateman**, EMBL-EBI, UK
- **Ulrike Wittig**, Heidelberg Institute for Theoretical Studies, Germany
- **Renate Kania**, Heidelberg Institute for Theoretical Studies, Germany

Workshop format: short talks, discussion

Special Workshop: Biocuration in China: importance, road map & priority

Chairs: Weimin Zhu, Jingchu Luo and Zhang Zhang

Biocuration involves adding value to biomedical data by the processes of standardization, quality control and information transferring (also known as data annotation). It enhances data interoperability and consistency, and is critical in translating biomedical data into scientific discovery. Although China is becoming a leading scientific data producer, biocuration is still very new to the Chinese biomedical data community. Here we will discuss the importance, road map and priority of biocuration in China.

Workshop format: short talks, discussion

List of Abstracts

Abstract #	Presenter	Title
1	Guoping Zhao	Biocuration: Driven by Functional Studies for Biology
2	Alex Bateman	Biocuration in the community
3	Amos Bairoch	neXtProt: recent developments in the context of biocuration
4	Johanna R. McEntyre	Big data to small data and back again: integrating biological data with the open access literature
5	Takashi Gojobori	Challenges in the Future and Experiences in the Past of Genome Annotation
6	Michal Linial	Genomes: The good, the bad and the ugly - the limits of automatic biocuration
7	Yixue Li	Challenges and Practices of Big Data in Life Science
8	Jennifer Harrow	Detangling transcriptional complexity in GENCODE using cutting-edge transcriptomics and proteomics data
9	J. Michael Cherry	Interoperable metadata leads to integrative analyses
10	Xiu-Jie Wang	Disease causing repeats and novel repeats in eukaryotic proteins
11	Jun Yu	Big and better but no junk
12	Claire O'Donovan	The UniRule system to enable scaling from manual curation to large data sets
13	Kim D. Pruitt	Functional curation of sequence data for RefSeq
14	Chunlei Wu	MyVariant.info: community-aggregated variant annotations as a service
15	Monica C. Munoz-Torres	APOLLO: SCALABLE & COLLABORATIVE CURATION OF GENOMES
16	Robert Hoehndorf	Large-scale semantic mining of disease-phenotype annotations
17	Pascale Gaudet	Annotation of functional impact of missense mutations in BRCA1
18	Shur-Jen Wang	PhenoMiner: a quantitative phenotype database for the laboratory rat, Rattus norvegicus. Application in hypertension and renal diseases
19	Sylvain Poux	The human proteome in UniProtKB
20	Jiayan Wu	VCGDB: a dynamic genome database of the Chinese population
21	Liguo Wang	Epidaurus: A Platform for Aggregation and Integration Analysis of the Epigenome

Abstract #	Presenter	Title
22	Frederic B. Bastian	The Bgee database: large-scale multi-species expression data
23	Reza Salek	Standardization and global knowledge exchange in metabolomics
24	Yang Pan	BioExpress: An integrated RNA-seq derived gene expression database for pan-cancer analysis
25	Yanli Wang	Bridge up chemical genomics and genomic information resources at NCBI
26	Venkat S. Malladi	Ontology application and use at the ENCODE DCC
27	Jie Zheng	Applying OBO Foundry ontologies to model, annotate and query longitudinal field studies on malaria
28	Huaiyu Mi	Phylogenetic- based gene function prediction in the Gene Ontology Consortium
29	Melanie Courtot	Standards for public health genomic epidemiology to improve infectious disease outbreak detection and investigation
30	Yuling Li	TextpressoCentral: A universal portal to search and curate biological literature
31	Cecilia N. Arighi	Construction of Phosphorylation Interaction Networks by Text Mining of Full-length Articles using the eFIP System
32	Thomas Lemberger	SourceData: integrating biocuration within the publishing process
33	Yin Huang	Bioso! – A Search Engine & Annotation Framework for Biological Big Data
34	Jasmine Young	Improved Data Representation of Very Large Macromolecules at Protein Data Bank
35	Anyuan Guo	An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools
36	Stanley J. Laulederkind	Gene Curation Software at the Rat Genome Database: Update 2015
37	Zexian Liu	dbPSP: a curated database for protein phosphorylation sites in prokaryotes
38	Sandra Orchard	Shared Resources, Shared Costs – Leveraging Biocuration Resources
39	Myra K. Derbyshire	Improving the Consistency of Domain Annotation within the Conserved Domain Database (CDD).
40	Hanfei Sun	Automated collection, curation and analysis for Transcriptional and Epigenetic Regulation Data

Abstract #	Presenter	Title
41	Gregory Butler	mycoCLAP, the Database for Characterized Lignocellulose-Active Proteins of Fungal Origin: Resource and Text Mining Curation Support
42	Han Cheng	dbPPT: a comprehensive database of protein phosphorylation in plants
43	Raja Mazumder	Generating a focused view of Disease Ontology cancer terms for pan-cancer data integration and analysis.
44	Dong Zou	MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data
45	Lina Ma	LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs
46	Lingling Shen	Genomics external service tracking and management
47	Sari A. Ward	COSMIC - Exploring the world's knowledge of somatic mutations in cancer
48	Cheng Yan	DrugVar: An integrated Database of Germline & Somatic Non- synonymous Variations That Impact Drug Binding
49	Min He	An efficient and scalable toolset for family-based sequencing data analysis
50	Abdelkrim Rachedi	STAD the Structural Targets Annotation Database
51	Xiao SiZhe	GigaDB submission wizard to enable authors to curate their own metadata
52	Xiaodong Wang	Regulatory Sequence Feature Curation in WormBase.
53	Aziz Khan	An extensive and interactive database of super-enhancers – dbSUPER
54	Ursula Hinz	Protein 3D-structures as precious sources of information in UniProtKB
55	Bifang He	MimoDB update 2015: towards a Biopanning Data Bank
56	Elvira Mitraka	2015 Disease Ontology update: DO's expanded curation activities to connect disease-related data
57	Donghui Li	TAIR as a Model for Database Sustainability
58	Jin Gu	Annotations of cancer-related microRNAs (oncomiRs)
59	Liu Wei	lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse

Abstract #	Presenter	Title
60	Kimchi A. Strasser	mycoCLAP, a Database for Characterized Lignocellulose-Active Proteins of Fungal Origin
61	Lin Yang	Applying Curation Lifecycle Model to Manage Cross-disciplinary Data
62	Hongmei Zhang	AnimalTFDB: a comprehensive animal transcription factor database
63	Christopher I. Hunter	GigaDB schema update to accommodate the variety of data.
64	Hongxing Lei	Deep mining of big data in stem cell and Alzheimer's disease
65	Madhura R. Vipra	Controlled vocabulary for capturing cellular lower level interactions by text-mining tools.
66	Randeep Singh	Chakraview: Interactive Genome Data Exploration and Real Time Visual Analytics Software
67	Robert Hoehndorf	DERMO: an ontology for the description of dermatologic disease
68	Susan Tweedie	Human Gene Family Resources at genenames.org
69	Cecilia N. Arighi	Viral Representative Proteomes: Computational Clustering of UniProtKB Virus Proteomes
70	Lorna Richardson	eMouseAtlas and Image Informatics
71	Madhura R. Vipra	Strategic manual curation for retrieval of highly complex data from mutagenesis experiments
72	Chunlei Yu	Database Commons: a web resource for cataloguing biological databases
73	Esther T. Chan	Towards reproducible computational analyses: the ENCODE approach
74	Cricket A. Sloan	Tracking data provenance to compare, reproduce, and interpret ENCODE results
75	Stacia R. Engel	The war on disease: Homology curation at SGD to promote budding yeast as a model for eukaryotic biology
76	Chengkun Wu	Fast large-scale text mining of biomedical literature on Tianhe-2 supercomputer
77	Jean M. Davidson	The role of the ENCODE Data Coordination Center
78	Todd D. Taylor	iCLiKVAL: Interactive community resource for manual curation of all scientific literature through the power of crowdsourcing
79	Naveen Kumar	iCLiKVAL API: a RESTful Hypermedia API for literature annotation and discovery

Abstract #	Presenter	Title
80	Yinghao Cao	Improving evidence-based gene prediction using RNA-seq data
81	Jia Jia	HBV-DIAP: HBV genome sequence curation and integrated analysis platform
82	Justyna Szostak	Interpretation of Large Scale Biological Data Facilitated by Curated Causal Biological Network Models.
83	Lei Sun	lncRScan-SVM: a support vector machine based tool for long non-coding RNA prediction
84	Wenming Zhao	DoGSD: the dog and wolf genome SNP database
85	Cecilia N. Arighi	Text Mining Tools for Biocuration in the iProLINK Web Portal
86	Takatomo Fujisawa	Semantic data integration in CyanoBase and RhizoBase
87	Martin Krallinger	Text mining and curation system for enzymatic and metabolism reactions: the TeBactEn tool.
88	Elvira Mitraka	Wikidata: a central hub of linked open life science data
89	Hao-Ying Huang	WEiGEM+, a mobile SNS platform for iGEM and Synthetic Biology
90	Ramona Britto	Proteome redundancy in UniProtKB: challenges and solutions
91	Lili Hao	IC4R: Information Commons for Rice
92	Julio Collado-Vides	RegulonDB, a tool to decipher the regulation of bacterial complexity.
93	Anne Niknejad	Development of a collection of life stage ontologies
94	Cecilia N. Arighi	BioCreative V a new challenge in text mining for biocuration
95	Kambiz Karimi	Curation and classification of inherited disease variants in a high-throughput clinical-grade genetic screening laboratory environment
96	Rama Balakrishnan	Collection and curation of whole genome studies of budding yeast at the Saccharomyces Genome Database (SGD)
97	Liguo Dong	From High-throughout Analysis, Comprehensive Analysis of downstream genes of Human chromatin remodeling INO80 complex
98	Feng Gao	A large-scale identification of prokaryotic replication origins and its applications
99	Yunxia S. Zhu	Building a Unified Mouse Gene Catalog

Abstract #	Presenter	Title
100	Hong Sun	A system for pathological comparison between human and animal models
101	Zhihua Zhang	3CDB：A 3C Database
102	Jiao Yuan	NONCODE, a noncoding RNA database with emphasis on long noncoding RNAs
103	Jiao Yuan	NPInter, a database of noncoding RNA related interactions
104	Inna Kuperstein	NaviCell Web Service for network-based data visualization and analysis
105	Jie Zheng	Applying the Ontology of Biological and Clinical Statistics (OBCS) to standardize statistical analysis of literature mined vaccine investigation data
106	Dapeng Wang	Plastid-LCGbase: a collection of evolutionarily conserved plastid-associated gene pairs
107	Jinmeng Jia	Ontology-based framework for mass spectrum data standard and analysis
108	Li Ni	Challenges and Solutions to Incorporating and Using Orthology Data in Model Organism Databases
109	Clara Amid	'Next-Generation-Biocuration' at the European Nucleotide Archive (ENA)
110	Dina Vishnyakova	Functionality and Adequacy of the Biocuration Text-mining Pipeline
111	Li Ni	Mouse Genome Informatics (MGI) GO Annotations in Context: Who, What, When and Where
112	Rose Oughtred	The BioGRID Interaction Database and Yeastphenome: Featuring New Curation of Yeast Models of Human Disease, Full Coverage of Yeast Interactions, and Comprehensive Curation of Yeast Phenotypes
113	Xiaolin Yang	CNPHD：Physique and Health Database of Chinese Nationals
114	Sangang Xu	GPNS: Gene & Protein Name-Mapping Service
115	Jennifer S. Polson	Novel Software Tools for Crowdsourcing Mitochondrial Protein Knowledge in Gene Wiki
116	Jing Guo	Comprehensive transcriptome and improved genome annotation of Bacillus licheniformis WX-02
117	Ling-Ling Chen	Comprehensive resources for sweet orange genome

Abstract #	Presenter	Title
118	Jinpu Jin	PlantTFDB: A portal for the functional and evolutionary study of plant transcription factors
119	Shi-Jian Zhang	RhesusBase: Evolutionary Interrogation of Human Biology in Well-Annotated Genomic Framework of Rhesus Macaque
120	Hong Luo	SorGSD: a Sorghum Genome SNP Database
121	Yi Zhao	LSD: A leaf senescence database
122	Mei Hou	AnnoLnc: a web server for integratively annotating novel human lncRNAs

Abstracts

Biocuration: Driven by Functional Studies for Biology

Guoping Zhao[1]

[1] Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, China

Presenter: Guoping Zhao

Abstract not provided at time of printing.

Biocuration in the community

Alex Bateman[1]

[1] EMBL-EBI, UK

Presenter: Alex Bateman

In this presentation I will discuss the challenges that we have faced in developing community annotation in the area of protein and RNA families. We have used Wikipedia as the source of our annotations and as the interface for our community to use. I will present results showing the success of this approach as well as some of the more challenging aspects. Finally I will discuss the social engineering of how we have tried to motivate scientists to edit Wikipedia.

neXtProt: recent developments in the context of biocuration

Amos Bairoch[1]

[1] CALIPHO group at the SIB Swiss Institute of Bioinformatics and University of Geneva, Switzerland

Presenter: Amos Bairoch

There is a large volume of data on human proteins being generated and annotated. That data is unfortunately very heterogeneous both in quality and types. neXtProt aims to address this need by selecting high quality data and presenting users with a user-friendly way to navigate through all the information. Recently we have made progress in three areas that are pertinent to the field of biocuration: • We have developed a biocuration platform (BioEditor) that allows to perform structured annotations concerning many aspects of the proteins realm. We are currently using the Bioeditor to capture the phenotypic effect of protein amino acid variations in genetic diseases and cancers. • We recently launched (on search.nextprot.org) a new search engine that makes use of the SPARQL technology. This engine and the accompanying API allow users to make very powerful queries that take into account the richness of the annotations available in neXtProt and also permit to make “federated queries” across resources that have developed a SPARQL endpoint, such as small molecule compendia. • We are active in the creation and maintenance of standardization resources. In this context we have increased the scope and content of the Cellosaurus (www.expasy.ch/cellosaurus), a cell line thesaurus that currently describes about 35’000 cell lines from 410 different species with a special emphasis on human cancer cell lines. All these developments are currently aimed at providing high quality, detailed annotations on human proteins, with a focus on the molecular basis of human cancers.

Big data to small data and back again: integrating biological data with the open access literature

Johanna R. McEntyre[1]

[1] EMBL-EBI

Presenter: Johanna R. McEntyre

The research literature is a central resource for data curation. A combination of recent policy and technical developments around text and data are providing both new challenges and opportunities that will begin to impact all researchers using the scientific literature, and therefore curators in particular. In this presentation I will review specific developments around open access, literature-data integration and scientific credit systems. For example, ongoing initiatives on the way we cite data in research articles will lead to improvements in matters of provenance, functional literature-data integration and the assignment of credit. In addition, granular annotation of articles through text mining and by individual comments is becoming possible through emerging standards and tools. Europe PMC, in the context of EMBL-EBI databases, is actively contributing to these developments and these will be highlighted over the course of the presentation.

Challenges in the Future and Experiences in the Past of Genome Annotation

Takashi Gojobori[1]

[1] Distinguished Professor, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Presenter: Takashi Gojobori

A word of "Annotation" means to give additional information to something basic or to give interpretation to something not well-known. Around at the time when the draft data of the complete human genome was published, scientific significance of annotation was not so recognized as much as at the present. Nowadays, however, it has become so crucial to make accurate and appropriate annotation on the genome and/or genes in order to conduct further studies of research such as cancer, ethnic groups, and human biomes, particularly in the current situation that a few thousands of genomes can be easily sequenced. With acute advancement of such genome-related technologies, sequencing of ten thousands and even a hundred thousands of genomes of human and other species have been planned in a number of projects. Moreover, the single cell technologies such as the nano-scale droplet system as well as the genome editing technologies such as the CRISPR-cas9 system provide us with more complicated and advanced ways of genome/gene annotation. This is particularly because genomic-related analysis can be done at a single cell level in many cases. In such a present situation of advancement of genome-related research, I would like to present the experiences of the past genome/gene annotation activities, such as H-invitational and FANTOM, that I have been mainly involved with. Then, I would discuss the challenge of genome annotation in the coming years. In particular, I may talk on the challenges of in annotation in the studies of metagenomics.

Genomes: The good, the bad and the ugly - the limits of automatic biocuration

Michal Linial[1]

[1] Department of Biological Chemistry, Institute for Life Sciences, The Hebrew University, Israel

Presenter: Michal Linial

The staggering growth in raw protein sequences made available by next-generation sequencing technology has overwhelmed our ability to thoroughly annotate the emerging protein sequence space. With over 70 million proteins in the public databases and an unprecedented growth, rate only robust unsupervised and automated methods can realistically achieve the comprehensive functional annotations of the protein space. Importantly, in recent years, the accumulation of new sequences is mostly driven by deep sequencing efforts and the high quality assembly phase that follows. While some of the genomes are accurately assembled (the good), many genomes are poorly assembled (bad) and in specific instances the genomic characteristics (e.g. large numbers of repeats, large population heterogeneity) lead to inherently limited power for automation and annotation inference (the ugly). To address the pressing challenge of accurate annotation, we offer a classification resource that is created by an unsupervised analysis of full-length protein sequences. The charting of the protein space to families is “model free” with a quality that matches the state-of-the-art semi-manual expert systems. The system, called ProtoNet, provides analysis tools to explore the entire protein space. I will present case studies from multiple genomes. The application of the algorithm to 18 complete genomes from insects serves as an evolutionary perspective on insects’ diversity. I claim that the difficulty of curation is directly estimated from the quality of the information and the degree of the evolutionary relatedness to already studied model organisms. Challenging groups of proteins for assignment of correct annotations includes proteins having multi-domain architecture, carrying moonlighting functions, short proteins and more. I will illustrate the power and the limitation of ProtoNet to detect overlooked connections between families. The division of families to subfamilies and the requirement for a reliable inference constitute immediate benefits to the biological and biomedical community.
Selected references: Rappoport N, Stern A, Linial N, Linial M. (2014) Entropy-driven partitioning of the hierarchical protein space. Bioinformatics 30: i624-i630. Rappoport N, Linial N, Linial M. (2013) ProtoNet: charting the expanding universe of protein sequences. Nat Biotechnol. 31: 290-292.
Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, et al. (2013) A large-scale evaluation of computational protein function prediction. Nat Methods. 10: 221-227.

Challenges and Practices of Big Data in Life Science

Yixue Li[1]

[1] Shanghai Center of Bioinformation Technology, China

Presenter: Yixue Li

Overall, the big data has four hallmarks: Volume: the quantity of data; Velocity: the time in which Big Data can be processed; Variety: the type of data that Big Data can comprise; and Veracity: the degree in which a researcher trusts the used information. Biological big data, in general, has the similar properties. But, not alike the data gathered by Google, WeChat, and Ali Baba, etc, even the big biological data is highly heterogeneous, inside the data, there exist intrinsic structures determined by various biological principles and experiment designs. Because of the four v's features of big data one always says that we can only build association relationship among certain elements, such as genes, proteins, pathways, etc., across the whole big data. This is not true and not enough for biological studies and life scientists always need to know driven force or causal relationship among biological elements, such as genes, proteins, metabolize, pathways, etc., which consists of complex biological systems. And many examples can be found which shown that the existed intrinsic structures determined by various biological principles and experiment designs have already provided biological data miners or curators with possible supporting to search a right way to identify causal relationship among biological molecules in big biological data. In this case "hypothesis driven" is a key for big biological data mining, which can have the facility to reduce the time consume of data mining and the occupation of computing resources effectively.

Detangling transcriptional complexity in GENCODE using cutting-edge transcriptomics and proteomics data

Jennifer Harrow[1]

[1] Computational Genomics, Wellcome Trust Sanger Institute, UK

Presenter: Jennifer Harrow

The modern view of genome organization focuses not on genes, but on the transcripts that constitute genes. Thus, while human GENCODEv21 (the ENCODE project geneset) contains 19,881 protein-coding genes, nearly 200,000 transcripts are associated with these loci, largely as a result of alternative splicing. In fact, RNAseq experiments routinely generate thousands of transcripts not found in GENCODE. However, if processes such as spliceosomal error or stochastic polymerase binding can generate transcriptional 'nose', this suggests that the transcriptome is infinitely large. This challenges the remit of GENCODE, which is to capture the full extent of transcriptional complexity. We should therefore ask what proportion of the transcriptome is actually functional? This question is of critical importance in the study of disease. For example, spurious coding exons are falsely enriched for loss of function variants, while variation surveys that do not consider the entire transcriptome will miss informative alleles. The need to decipher the true link between complexity and functionality is urgent, and yet a small minority of transcripts have had their functionality confirmed experimentally. Here, we will discuss how GENCODE is being radically improved by the integration of next-generation technologies including RNAseq, CAGE, polyAseq, ribosome profiling and mass spectrometry. Such data allow us to identify missing transcripts of putative functionality, including single exon lncRNAs. Meanwhile, we are fine-tuning our description of transcript ends, creating functionally distinct models within existing genes that share identical splicing patterns. Furthermore, these data allow us to reappraise our functional annotation; in particular, modern proteomics can address long-standing questions about the total number of protein-coding genes and transcripts.

Interoperable metadata leads to integrative analyses

J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA

Presenter: J. Michael Cherry

Computational methods are developed to analyze the relationships between high-resolution genomics data such as epigenomic and chromatin marks. These analyses explore connections between various assay methods applied to biological samples obtained from different cell lines, tissues or developmental time points. The relationship of the biological samples and their treatment is an essential component for this integrative analysis. We have developed detailed structured metadata schema for reagents, biological samples and analysis software that adds clarity to the experimental context and promotes blending of experimental results. I will discuss the creation of reference data sets through the curation of metadata and the enhancements we provide via our web resources using these annotations.

Disease causing repeats and novel repeats in eukaryotic proteins

Ying-Ke Ma[1,2], Qing Shi[1,2], Xiu-Jie Wang[1]

[1] Key Laboratory of Genetic Network, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China

Presenter: Xiu-Jie Wang

Repeat expansion mutations in coding or non-coding regions may cause many neurological diseases. These diseases are usually inheritable and with genetic anticipation, such as Fragile X syndrome and Huntingtonâ ™s disease. Here we curated all known repeat expansion-related diseases and their causal genes. A thorough screen for repeat-containing proteins identified many unknown amino acid repeats in eukaryotic proteins. The amino acid composition, chemical features and functions of these repeats were analyzed. Evolutionary analysis revealed variations of some repeats across species, indicating that the repeat variation may contribute to the functional evolution difference of some proteins.

Big and better but no junk

Jun Yu[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Jun Yu

Genomics sequences contribute significantly to BIG DATA COMMONS. Aside from sequence quality specific to each sequencer type, the understanding of genome biology is of essence in the utilization of the genomics big data. First, the structures of plants and animals are different. Plant chromosomes are less hopeful to be assembled into single pieces due to very active biologically-defined repeats (BDRs). However, vertebrate genomes, especially the mammalian genomes are highly organized so that their gene clusters are mostly conserved, exhibiting as strong co-linearity. Second, transcriptomics data have still been poor in quality, largely due to a combination of technical draw-backs, including diverse platforms (such as different types of microarrays and sequencers) and methodologies (such as EST and RNA-seq). A compatible laboratory protocol and data acquired with single copy and single cell resolutions are of importance. Third, cellular heterogeneity (or phenotypic plasticity) calls for spatiotemporal data. How genes and their clusters are organized topologically as what are ordered on chromosomes? How methylation sites are differentially methylated as cell develops and differentiates spatiotemporally? Fourth, since big data eventually have to be ALIVE, visualization in the context of cells becomes new challenge for future genomics data curation.

The UniRule system to enable scaling from manual curation to large data sets

Claire O'Donovan[1], The UniProt Consortium[1,2,3]

[1] EMBL-European Bioinformatics Institute, UK
[2] Swiss Institute of Bioinformatics, Switzerland
[3] Protein Information Resource, USA

Presenter: Claire O'Donovan

The UniProt Knowledgebase is a central hub for the collection of functional information on proteins. It consists of 2 sections: UniProtKB/Swiss-Prot which is manually annotated and reviewed and UniProtKB/TrEMBL which is automatically annotated and not reviewed. Over the last few years, we have developed the UniRule system that leverages experimental UniProtKB/Swiss-Prot curation for the automatic annotation of UniProtKB/TrEMBL in order to address the exponential growth in uncharacterized proteins from the genome sequencing centers. UniRule consists of manually created annotation rules that specify functional annotations and the conditions which must be satisfied for them to apply (such as taxonomic scope, family membership as defined by the 11 InterPro Consortium members (GENE3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, TIGRFAMs) and the presence of specific sequence features. The UniRules are applied at each UniProt release, and their predictions are continuously evaluated against the content of matching UniProtKB/Swiss-Prot entries, guaranteeing that the predictions remain in synch with the expert curated knowledge of UniProtKB/Swiss-Prot. In Release 2015_02, UniProtKB/Swiss-Prot contains 547,599 entries while UniProtKB/TrEMBL contains 90,860,905 entries. UniRule has enabled us to annotate approximately 37% of UniProtKB/TrEMBL with functional annotation. We would very welcome the opportunity to present our work and outline our plans for the future which hopefully will interest other members of the biocuration community and create collaborations.

Functional curation of sequence data for RefSeq

Kim D. Pruitt[1], Lillian Riddick[1], Mike Murphy[1], Wendy Wu[1]

[1] NCBI, National Institutes of Health, United States

Presenter: Kim D. Pruitt

The National Center for Biotechnology Information (NCBI) reference sequence (RefSeq) project provides sequence standards for proteins, transcripts, genes, and genomes through processes that leverage computational analysis, collaboration, and targeted curation. These high quality standards provide a baseline that is used for a variety of applications ranging from individual gene, transcript, or protein analysis to high throughput next-generation sequence analysis. Manual curation of sequence and gene data for human, mouse, and other vertebrates is essential to correct errors, add representation of genes and their products that aren't included in draft assemblies, and to add functionally relevant information about both the gene and the sequence. NCBI scientific staff rely on literature review, personal communications, and sequence analysis to provide specific functionally relevant annotation in the context of RefSeq transcript and protein sequence records. This curation activity is organized by gene or protein family, community request, or other sequence-based metrics and adds information to sequence records that cannot be provided computationally. Functional information provided includes sequence motifs (e.g., histones), regulatory upstream open reading frames, antimicrobial peptides, and other relevant regions of the sequence. The presentation will provide an overview of RefSeq and examples of targeted curation activities that add functional information to RefSeq records.

MyVariant.info: community-aggregated variant annotations as a service

Chunlei Wu[1], Adam Mark[1], Sean Mooney[2], Ben Ainscough[3], Ali Torkamani[1], Andrew I. Su[1]

[1] The Scripps Research Institute, United States
[2] Buck Institute of Aging Research, United States
[3] Washington University School of Medicine, United States

Presenter: Chunlei Wu

The accumulation of genetic variant annotations has been increasing explosively with the recent technological advances. However, the fragmentation across many data silos is often frustrating and inefficient. Bioinformaticians everywhere must continuously and repetitively engage in data wrangling in an effort to comprehensively integrate knowledge from all these resources, and these uncoordinated efforts represent an enormous duplication of work. We created a platform, called MyVariant.info (http://myvariant.info), to aggregate variant-specific annotations from community resources and provide high-performance programmatic access. Annotations from each resource are first converted into JSON-based objects with their id fields as the canonical names following HGVS nomenclature (genomic DNA based). This scheme allows merging of all annotations relevant to a unique variant into a single annotation object. A high-performance and scalable query engine was built to index the merged annotation objects and provides programmatic access to the developers. As of today, MyVariant.info is serving >100M variants in total and we are actively expanding the coverage by engaging community efforts. MyVariant.info decouples two fundamental steps in management of variant annotations: the creation and maintenance of centralized web services (which requires deep software-engineering expertise), and the task of structuring biological annotations (which requires broad community effort). Annotation providers from the community can provide data parsers to convert their raw data into JSON-compatible objects. The only requirement is that a valid HGVS name is used as the id field for each object. These data can then be queryable through the query engine we built. The data provider doesn't have to worry about building their own query infrastructure. And the research community doesn't have to learn another query interface in order to access new annotations.

APOLLO: SCALABLE & COLLABORATIVE CURATION OF GENOMES

Monica C. Munoz-Torres[1], Nathan Dunn[1], Colin Diesh[2], Deepak Unni[2], Seth Carbon[3], Heiko Dietze[3], Christopher Mungall[3], Nicole Washington[3], Ian Holmes[4], Christine G. Elsik[2], Suzanna E. Lewis[3]

[1] Genomics Division, Lawrence Berkeley National Laboratory, United States of America
[2] Plant and Animal Sciences, University of Missouri, United States of America
[3] Genomics Division, Lawrence Berkeley National Laboratory, United States of America
[4] Bioengineering, University of California Berkeley, United States of America

Presenter: Monica C. Munoz-Torres

Obtaining meaningful results from genome analyses requires high quality annotations of all genomic elements. Today's sequencing projects face challenges such as lower coverage, more frequent assembly errors, and the lack of closely related species with well-annotated genomes. Apollo is a web-based application that supports and enables collaborative genome curation in real time, analogous to Google Docs, allowing curators to improve on existing automated gene models through an intuitive interface. Apollo's extensible architecture is built on top of JBrowse; its components are a web-based client, an annotation-editing engine, and a server-side data service. It allows users to visualize automated gene models, protein alignments, expression and variant data, and conduct structural and/or functional annotations. Apollo is actively used within a variety of projects, including the initiative to sequence the genomes of 5,000 Arthropod species (i5K), and will become essential to the thousands of genomes now being sequenced and analyzed. Researchers from nearly 100 institutions worldwide are currently using Apollo on distributed curation efforts for over sixty genome projects across the tree of life; from plants to echinoderms, to fungi, to species of fish and other vertebrates including human, cattle (bovine), and dog. We are training the next generation of researchers by reaching out to educators to make these tools available as part of curricula, offering workshops and webinars to the scientific community, and through widely applied systems such as iPlant and DNA Subway. We are currently integrating Apollo into an annotation environment combining gene structural and functional annotation, transcriptomic, proteomic, and phenotypic annotation. In this presentation we will describe in detail its utility to users, introduce the architecture to developers interested in expanding on this open-source project, and offer details of our future plans.

Large-scale semantic mining of disease-phenotype annotations

Robert Hoehndorf[1], Paul N. Schofield[2], Georgios V. Gkoutos[3]

[1] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Saudi Arabia
[2] Physiology, Development and Neuroscience, University of Cambridge, United Kingdom
[3] Computer Science, University of Aberystwyth, United Kingdom

Presenter: Robert Hoehndorf

Phenotypes are the observable characteristics of an organism arising from the interactions of its genotype with the environment. Phenotypes associated with engineered and natural genetic variation are widely recorded using phenotype ontologies in model organisms, as are signs and symptoms of human Mendelian diseases in databases such as OMIM and Orphanet. Exploiting these resources, several computational methods have been developed for integration and analysis of phenotype data to identify the genetic etiology of diseases or suggest plausible interventions. A similar resource would be highly useful not only for rare and Mendelian diseases, but also for common, complex and infectious diseases. We apply a semantic text-mining approach to identify the phenotypes (signs and symptoms) associated with over 6,000 diseases. We evaluate our text-mined phenotypes by demonstrating that they can correctly identify known disease-associated genes in mice and humans with high accuracy. The resource is freely available at http://aber-owl.net/aber-owl/diseasephenotypes/.

Annotation of functional impact of missense mutations in BRCA1

Isabelle Cusin[1], Monique Z. Zahn[1], Amos Bairoch[1], Pascale Gaudet[1]

[1] SIB Swiss Institute of Bioinformatics, Switzerland

Presenter: Pascale Gaudet

Characterization of the phenotypic effect of mutations provides evidence on which variants of unknown significance (VUS) can be evaluated. We have annotated the phenotypes caused by missense mutations in BRCA1 associated with increase susceptibility to breast and ovarian cancers. Using the information derived from 87 publications, the function, the phenotype and the binding properties of 385 unique missense mutations were captured according to impact on gene ontology function, mammalian phenotype ontology, and in-house binding relations, resulting in 1106 different annotations. Each annotation is supported by detailed experimental evidences. Well characterized and assessed functions of BRCA1 includes its ubiquitin-protein ligase activity and its role in DNA repair, as well as its transcriptional regulation activity, response to DNA damage, and UBE2D1, BARD1 and BRIP1 binding. These data provide the most comprehensive resource on phenotypes of BRCA1 variants. We are expanding this work to other disease-causing genes: BRCA2, as well as genes implicated in Lynch syndrome (MSH2, MSH6, MLH1).

PhenoMiner: a quantitative phenotype database for the laboratory rat, Rattus norvegicus. Application in hypertension and renal diseases

Shur-Jen Wang[1], Stanley J. Laulederkind[1], George T. Hyaman[1], Victoria Petri[1], Weisong Liu[1], Jennifer R. Smith[1], Rajni Nigam[1], Melinda R. Dwinell[2], Mary Shimoyama[2,3]

[1] Rat Genome Database, Medical College of Wisconsin, USA
[2] Physiology, Medical College of Wisconsin, USA
[3] Surgery, Medical College of Wisconsin, USA

Presenter: Shur-Jen Wang

Rats have been used extensively as animal models to study physiological and pathological processes involved in human diseases. Many rat strains have been selectively bred for certain biological traits related to specific medical interests. The Rat Genome Database (RGD) has initiated the PhenoMiner project to integrate quantitative phenotype data from the PhysGen Program for Genomic Applications and the National BioResource Program in Japan as well as manual annotations from the biomedical literature. PhenoMiner, the search engine for these integrated phenotype data, facilitates mining of datasets across studies by searching the database with a combination of terms from four different ontologies/vocabularies (Rat Strain Ontology, Clinical Measurement Ontology, Measurement Method Ontology, Experimental Condition Ontology). In this study, salt-induced hypertension was used as a model to retrieve blood pressure records of Brown Norway (BN), Fawn-Hooded Hypertensive (FHH), and Dawley salt sensitive (SS) rat strains. The records from these three strains served as a basis for comparing records from consomic/congenic/mutant offspring derived from them. We examined the cardiovascular and renal phenotypes of consomics derived from FHH and SS, and of SS congenics and mutants. The availability of quantitative records across laboratories in one database, such as those provided by the PhenoMiner can empower researchers to make the best use of publicly available data.

The human proteome in UniProtKB

Sylvain Poux[1], Lionel Breuza[1], Maria L. Famiglietti[1], the UniProt consortium[1,2,3]

[1] Swiss-Prot, SIB Swiss Institute of Bioinformatics, Switzerland
[2] The European Bioinformatics Institute, United Kingdom
[3] Protein Information Resource, USA

Presenter: Sylvain Poux

The Universal Protein Resource (UniProt) provides the scientific community with a comprehensive and richly curated resource of protein sequences and functional information. The centerpiece of UniProt is the knowledgebase (UniProtKB) which is composed of the expert curated UniProtKB/Swiss-Prot section and its automatically annotated complement, UniProtKB/TrEMBL. The expert curation of the human proteome constitutes the highest priority of the consortium and UniProtKB/Swiss-Prot provides representative sequences for all human protein-coding genes that are enriched with functional annotations. Expert curation combines the manually verified sequence with experimental evidence derived from biochemical and genetic analyses, 3D-structures, mutagenesis experiments, information about protein interactions and post-translational modifications. We also manually assign Gene Ontology (GO) terms from literature during the biocuration process. We continually revisit human UniProtKB/Swiss-Prot, updating functional and sequence annotation, prioritizing proteins for which knowledge evolves. Specific emphasis is given to curation of human protein variants and association with genetic diseases. Functionally characterized single amino-acid polymorphisms (SAPs) and their functional consequences are extracted from literature and described in the sequence section of UniProtKB/Swiss-Prot entries. It provides a way to map them to other important sequence features inferred, for instance, from 3D structures data. Here we describe some of our current work in standardizing the curation of variants and their functional impact using existing ontologies. Our aim is to facilitate the integration of information on protein function in normal and disease states. Release 2015_01 of UniProtKB/Swiss-Prot contained 20,199 human entries as well as 21,841 manually reviewed alternative products and more than 25’400 single amino acid polymorphisms associated with a genetic disease.

VCGDB: a dynamic genome database of the Chinese population

Yunchao Ling[1], Jun Yu[1], Jiayan Wu[1], Jingfa Xiao[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Jiayan Wu

The data released by the 1000 Genomes Project contain an increasing number of genome sequences from different nations and populations with a large number of genetic variations. As a result, the focus of human genome studies is changing from single and static to complex and dynamic. The currently available human reference genome (GRCh37) is based on sequencing data from 13 anonymous Caucasian volunteers, which might limit the scope of genomics, transcriptomics, epigenetics, and genome wide association studies.We used the massive amount of sequencing data published by the 1000 Genomes Project Consortium to construct the Virtual Chinese Genome Database (VCGDB, http://vcg.cbi.ac.cn/), a dynamic genome database of the Chinese population based on the whole genome sequencing data of 194 individuals. VCGDB provides dynamic genomic information, which contains 35 million single nucleotide variations (SNVs), 0.5 million insertions/deletions (indels), and 29 million rare variations, together with genomic annotation information. VCGDB also provides a highly interactive user-friendly virtual Chinese genome browser (VCGBrowser) with functions like seamless zooming and real-time searching. In addition, we have established three population-specific consensus Chinese reference genomes that are compatible with mainstream alignment software. VCGDB offers a feasible strategy for processing big data to keep pace with the biological data explosion by providing a robust resource for genomics studies; in particular, studies aimed at finding regions of the genome associated with diseases.

Epidaurus: A Platform for Aggregation and Integration Analysis of the Epigenome

Liguo Wang[1], Haojie Huang[2], Jean-Pierre A. Kocher[1]

[1] Division of Biomedical Informatics and Statistics, Mayo Clinic, 55905
[2] Department of Biochemistry and Molecular Biology, Mayo Clinic, 55905

Presenter: Liguo Wang

Integrative analyses of epigenetic data promise a deeper understanding of the epigenome. Epidaurus is a bioinformatics tool used to effectively reveal inter-dataset relevance and differences through data aggregation, integration and visualization. In this study, we demonstrated the utility of Epidaurus in validating hypotheses and generating novel biological insights. In particular, we described the use of Epidaurus to 1) integrate epigenetic data from prostate cancer cell lines to validate the activation function of EZH2 in castration-resistant prostate cancer and to 2) study the mechanism of androgen receptor (AR) binding deregulation induced by the knockdown of FOXA1. We found that EZH2's noncanonical activation function was reaffirmed by its association with active histone markers and the lack of association with repressive markers. More importantly, we revealed that the binding of AR was selectively reprogrammed to promoter regions, leading to the up-regulation of hundreds of cancer-associated genes including EGFR. The prebuilt epigenetic dataset from commonly used cell lines (LNCaP, VCaP, LNCaP-Abl, MCF7, GM12878, K562, HeLa-S3, A549, HePG2) makes Epidaurus a useful online resource for epigenetic research. As standalone software, Epidaurus is specifically designed to process user customized datasets with both efficiency and convenience.

The Bgee database: large-scale multi-species expression data

Frederic B. Bastian[1,2], Anne Niknejad[1,2], Marta Rosikiewicz[1,2], Sébastien Moretti[1,2], Valentine Rech De Laval[1,2], Marc Robinson-Rechavi[1,2]

[1] SIB Swiss Institute of Bioinformatics, Switzerland
[2] Department of Ecology and Evolution, University of Lausanne, Switzerland

Presenter: Frederic B. Bastian

Bgee is a database to study and compare gene expression patterns in animal species. Bgee currently provides, for 17 species, baseline expression calls, and differential expression calls over anatomy and over development, made comparable between species. This unique resource allows new insights on, e.g., the evolution of gene expression, or on functional genomics. Bgee requires the development of original annotation resources and statistical analyses, and the integration of various Model Organism Databases (MODs) and expression data repositories. For integration into Bgee, expression data are filtered based on unique QCs, analyzed using relevant statistical methods, then summarized and made comparable between species. This notably requires developing: - a repository of homology annotations between anatomical structures; - a collection of life stage ontologies, covering all species integrated in Bgee; - a new process and ontology to provide confidence information about annotations, the Confidence Information Ontology; - new QC and statistical analyses, notably to uncover duplicated content in public transcriptomics repositories, and to filter low quality Affymetrix chips or RNA-Seq libraries; - annotations of thousands of expression data to the multi-species anatomical ontology Uberon; - pipelines for integrating in situ data from various MODs. In addition, the Bgee team contributes to the development of shared resources, such as OWLTools and Uberon. Bgee currently integrates data for 17 animal species. We have the aim of integrating all animal species with sequenced genomes and expression data available. This has become realistic thanks to the curation and analyses tools that we have developed. We believe that Bgee could become in the future a major hub for multi-species expression data integration, and a reference resource for the analysis and annotation of expression data in non-model organisms. Bgee is available at http://bgee.org/.

Standardization and global knowledge exchange in metabolomics

Reza Salek[1]

[1] Cheminformatics and Metabolism, EMBL-EBI, United Kingdom

Presenter: Reza Salek

Metabolomics has become a crucial phenotyping technique in a range of research fields including medicine, the life sciences, biotechnology and the environmental sciences. This necessitates the transfer of experimental information between research groups, as well as potentially to publishers and funders. In this contribution we outline the first global-scale, open access repository for metabolomics studies - MetaboLights (1) - as well as its embedding in European and global data standards and data sharing projects such as COSMOS (COordination Of Standards In MetabOlomicS) and MetabolomeXchange.org. The rapid physiological responses that can be detected by metabolomics profiles have the potential to identify early warning markers of disease-related molecular and cellular disturbances well in advance of the appearance of medical phenotypes. However, many exciting initial findings are based on pilot studies with small sample sizes, and in order to be translated into clinical applications such studies need to be multiply replicated in larger studies across different populations. Rapid and efficient replication depends on making study protocols, data and metadata openly available -- a requirement increasingly mandated by funding agencies. However, merely sharing raw data is not as useful without addressing the additional challenges of standardized and open formats, clear and sufficient descriptions of study protocols, and the use of globally agreed standards for metadata annotation. Here we describe our concepts and efforts engagement with the metabolomics community, academics and industry, journal publishers, software and hardware vendors, as well as those interested in standardisation worldwide to addressing missing metabolomics ontologies, complex-metadata capturing and XML based open source data exchange format. References: MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. K Haug et al., Nucl. Acids Res.2012

BioExpress: An integrated RNA-seq derived gene expression database for pan-cancer analysis

Quan Wan[1], Hayley Dingerdissen[1], Yang Pan[1], Haichen Zhang[1], Tsung-Jung Wu[1], Cheng Yan[1], Naila Gulzar[1], Yu Fan[1], Raja Mazumder[1,2]

[1] Department of Biochemistry and Molecular Medicine, The George Washington University, USA
[2] McCormick Genomic and Proteomic Center, The George Washington University, USA

Presenter: Yang Pan

BioExpress is a gene expression and cancer association database in which the expression levels are mapped to genes using RNA-seq data obtained from TCGA, ICGC, Expression Atlas and publications. The BioExpress database includes expression data from 64 cancer types, 6361 patients, and 17,469 genes with 9513 of the genes displaying differential expression between tumor and normal samples. In addition to data directly retrieved from RNA-seq data repositories, manual biocuration of publications supplements the available cancer association annotations in the database. All cancer types are mapped to Disease Ontology terms to facilitate a uniform pan-cancer analysis. The BioExpress database is easily searched using HGNC gene symbol, UniProtKB accession, or protein accession, or alternatively, can be queried by cancer type with specified significance filters. This interface, along with availability of pre-computed downloadable files, enables the straightforward retrieval and display of a broad set of cancer-related genes. Access: http://hive.biochemistry.gwu.edu/tools/bioexpress

Bridge up chemical genomics and genomic information resources at NCBI

Yanli Wang[1]

[1] National Library of Medicine, National Institutes of Health, National Center for Biotechnology Information, USA

Presenter: Yanli Wang

Chemical genomics (Chemogenomics) systematically screens small molecule libraries to identify drug candidates and chemical probes to characterize protein and gene functions. Advances in RNAi screening technology have enabled genome-wide functional screens for discovering new cellular pathways and therapeutic targets. The PubChem BioAssay database (http://www.ncbi.nlm.nih.gov/pcassay/), hosted by the National Center for Biotechnology Information (NCBI) at NIH, serves as a public repository for information generated by Chemogenomics and RNAi research. The database currently contains over 1,000,000 freely accessible bioassay protocols and datasets. These include HTS data submitted by US government funded projects, as well as literature based curated information via collaborating with ChEMBL, IUPHAR, PDBBind and several other organizations, providing chemical modulators for nine thousand proteins and experimental indications of phenotypic significance for over 30,000 genes. PubChem has been utilized by a broad range of research fields including medicinal chemistry, drug discovery, chemical biology and cheminformatics. Moreover, PubChem BioAssay closes the gap between molecular and chemical biology research, and supports the study of protein, gene functions and their molecular pathways by linking the experimental data contained in PubChem to genomic resources at NCBI. The integration of PubChem with the rest resources at NCBI provides a unique annotation service for NCBI's genomic information. This presentation will describe how this effort enables the retrieval of drug and chemical modulators, as well as biological and therapeutic relevance for many GenBank records.

Ontology application and use at the ENCODE DCC

Venkat S. Malladi[1], Drew T. Erickson[1], Nikhil R. Podduturi[1], Laurence D. Rowe[1], Esther T. Chan[1], Jean M. Davidson[1], Benjamin C. Hitz[1], Marcus Ho[1], Brian T. Lee[2], Stuart Miyasato[1], Gregory R. Roe[1], Matt Simison[1], Cricket A. Sloan[1], J. Seth Strattan[1], Forrest Tanaka[1], W. James Kent[2], J. Michael Cherry[1], Eurie L. Hong[1]

[1] Department of Genetics, Stanford University, USA
[2] Center for Biomolecular Science and Engineering, University of California Santa Cruz, USA

Presenter: Venkat S. Malladi

The Encyclopedia of DNA elements (ENCODE) project is an ongoing collaborative effort to create a catalog of genomic annotations. To date, the project has generated over 4000 experiments across more than 350 cell lines and tissues using a wide array of experimental techniques to study the chromatin structure, regulatory network, and transcriptional landscape of the H. sapiens and M. musculus genomes. All ENCODE experimental data, metadata, and associated computational analyses are submitted to the ENCODE Data Coordination Center (DCC) for validation, tracking, storage, and distribution to community resources and the scientific community. As the volume of data increases, the organization of experimental details becomes increasingly complicated and demands careful curation to identify related experiments. Here we describe the ENCODE DCC's use of ontologies to standardize experimental metadata. We discuss how ontologies, when used to annotate metadata, provide improved searching capabilities and facilitate the ability to find connections within a set of experiments. Additionally, we provide examples of how ontologies are used to annotate ENCODE metadata and how the annotations can be identified via ontology-driven searches at the ENCODE portal. As genomic datasets grow larger and more interconnected, standardization of metadata becomes increasingly vital to allow for exploration and comparison of data between different scientific projects. Database URL: https://www.encodeproject.org/

Applying OBO Foundry ontologies to model, annotate and query longitudinal field studies on malaria

Jie Zheng[1,2], San Emmanuel James[2,3], Emmanuel Arinaitwe[3], Bryan Greenhouse[4], Edwin Charlebois[4], Grant Dorsey[4], Ja'Shon Cade[2,5], Brian P. Brunk[2,5], Omar S. Harb[2,5], David S. Roos[2,5], Christian J. Stoeckert[1,2]

[1] Department of Genetics, University of Pennsylvania, USA
[2] EuPathDB project, University of Pennsylvania, USA
[3] Infectious Disease Research Collaboration, Uganda
[4] Department of Medicine, University of California, San Francisco, USA
[5] Department of Biology, University of Pennsylvania, USA

Presenter: Jie Zheng

Ontologies play a crucial role in supporting consistent data representation to facilitate data integration, information retrieval and new knowledge discovery. The Ontology for Biomedical Investigations (OBI; http://obi-ontology.org/), a member of OBO (Open Biological and Biomedical Ontologies) Foundry Ontologies, provides a semantic model to cover all aspects of biomedical investigations, including terms facilitating integration of other OBO Foundry ontologies. We have employed OBI, along with extensions from other OBO Foundry ontologies, to provide a semantic framework for describing data generated by the Program for Resistance, Immunology, Surveillance and Modeling of Malaria (PRISM; http://muucsf.org/projects/prism.html), an NIH-supported International Center for Excellence in Malaria Research (ICEMR). The PRISM project aims to use comprehensive surveillance data to elucidate interactions between malaria parasites, their mosquito vectors, and human hosts. Cohort studies following >300 households in three regions of Uganda with diverse demographics and transmission intensity provide information on ~1000 study participants, the dwellings in which they live, quarterly routine visits and additional sick visits, and monthly mosquito collections for each dwelling. Integrating these data into the Plasmodium Genomics Resource (http://PlasmoDB.org), a component of the Eukaryotic Pathogen Bioinformatics Resource Center offers to make data-mining strategies accessible to PRISM project members, and ultimately the broader international research community. The challenge of understanding >280 different kinds of data and their interconnections has been addressed by exploiting OBI to model PRISM studies. This effort has yielded the EuPathDB ontology, built on the basis of OBI integrated with other OBO Foundry ontologies. Consistent interpretation and representation of PRISM data using OBO Foundry ontologies greatly facilitate data exploring and querying through PlasmoDB.

Phylogenetic- based gene function prediction in the Gene Ontology Consortium

Huaiyu Mi[1], Pascale Gaudet[2], Marc Feuermann[2], Anushya Muruganujan[1], Suzanna E. Lewis[3], Paul Thomas[1]

[1] University of Southern California, USA
[2] Swiss Institute of Bioinformatics, Switzerland
[3] Lawrence Berkeley Laboratory, USA

Presenter: Huaiyu Mi

Gene Ontology (GO) is a community resource that represents biological knowledge of gene functions through the use of structured and controlled vocabulary. Since most sequences have not been experimentally characterized, most available annotations within GO are based on predictions. Therefore it is crucial to have more accurate prediction method. Evolutionarily related genes that evolved from a common ancestor (orthologs) tend to preserve their functions. Thus inferences based on such information are more accurate. A curation tool, called Phylogenetic Annotation and INference Tool (or PAINT), has been developed to help curators to infer annotations among members within a gene family. If a gene has been previous annotated with experimental evidence, the curator can make precise assertion as when the function is evolved from based on the phylogenetic information, and propagate the function to that ancestor. All genes evolved from that ancestor would then inherit the same function. PAINT enables a biocurator to construct and record a (generally) parsimonious model of the evolution of function in the family that can be tested against, and modified by, new experimental data as it emerges. Preliminary studies show that PAINT is able to make more accurate inferences, especially to non-model organism genes. It also serves as QA process to validate the previous annotations by viewing annotations from many related genes. The new curation paradigm greatly improved the efficiency and quality of GO annotation, and will greatly help our user community to utilize the GO knowledge in their data analysis.

Standards for public health genomic epidemiology to improve infectious disease outbreak detection and investigation

Melanie Courtot[1,2], Emma Griffiths[1], Damion Dooley[2], Josh Adam[3], Franklin Bristow[3], Joao A. Carrico[4], Bhavjinder K. Dhillon[1], Matthew Laird[1], Raymond Lo[1], Thomas Matthews[3], Aaron Petkau[3], Geoff Winsor[1], Lynn M. Schriml[5], Morag Graham[3], Gary Van Domselaar[3], Fiona Brinkman[1], William Hsiao[2,6]

[1] Simon Fraser University, Canada
[2] BC Public Health Microbiology and Reference Laboratory, Canada
[3] National Microbiology Laboratory, Public Health Agency of Canada, Canada
[4] Faculty of Medicine, University of Lisbon, Portugal
[5] University of Maryland School of Medicine, USA
[6] University of British Columbia, Canada

Presenter: Melanie Courtot

Routine infectious disease outbreak investigations using microbial genomic data are hampered by delays incurred when manually integrating data from heterogeneous sources involved in detection and investigation of infectious disease outbreaks. The Integrated Rapid Infectious Disease Analysis (IRIDA) project, consisting of partners from national and provincial public health organizations and academic labs, is building a bioinformatics platform to support real-time infectious disease outbreak investigations and provide a suite of bioinformatics tools for researchers and public health workers. An ontology-based approach for platform development, combined with semantic web technology will enable robust data integration and more efficient analysis within IRIDA. We present how we built a new public health genomic epidemiology ontology and associated standards, based on (1) a comprehensive review of existing relevant metadata standards during which we identified missing elements to be addressed, (2) how these standards are applied to support IRIDA workflows, as well as (3) competency questions our project addresses to support epidemiology studies. By adhering to the best practices of the Open Biomedical and Biological Ontology (OBO) Consortium, our ontology allows integration of various efforts to provide a consolidated resource directly applicable to IRIDA. Our modular approach to development also ensures it can be extended to provide more comprehensive coverage, e.g., in the domain of food categories. Efficient biocuration of genomics, laboratory, clinical, and epidemiological data using the resource we are developing will enable data integration and querying over previously disparate information. Our research is a key component enabling the creation of the IRIDA pipeline, which will process data in a more automated fashion, alleviating the burden of manual analysis, and ultimately resulting in more efficient and effective infectious disease outbreak responses.

TextpressoCentral: A universal portal to search and curate biological literature

Yuling Li[1], Hans-Michael Müller[1], Paul Sternberg[1,2]

[1] Department of Biology, California Institute of Technology, USA
[2] Howard Hughes Medical Institute, USA

Presenter: Yuling Li

TextpressoCentral is a new-generation curation tool, a successor to the current Textpresso 2.5 (www.textpresso.org) system, which is a biological literature search engine developed at WormBase, and is popular among WormBase curators and some other model organism databases. The TextpressoCentral site is built from scratch with an emphasis on a one-stop search and curation experience for curators. The site currently contains approximately 880,000 full text articles from the PMC Open Archive, which was downloaded in November 2014. TextpressoCentral is a platform to perform full text literature searches, view and curate research papers, train and apply machine learning (ML) and text mining (TM) algorithm for semantic analysis and curation purposes. The user is supported in this task by giving him capabilities to select, edit and store lists of papers, sentences, terms and categories in order to perform training and mining. The system is designed to empower the user to perform as many operations on a literature corpus or a particular paper as possible. It uses state-of-the-art software packages and frameworks such as the Unstructured Information Management Architecture (UIMA), the popular search engine Lucene and Wt, a WWW framework. The corpus of papers can be built from fulltext articles that are available in PDF or NXML format. TextpressoCentral will replace the current Textpresso system in the future, expanding its capabilities from search-only to multiple tasks that are essential for curation at model organism databases. It should help improve the efficiency of curation by eliminating the overhead of switching between different tools and frameworks. Moreover, it makes collection of training sets a natural part of curation, which will accelerate the application of a variety of text-mining tools to biocuration tasks.

Construction of Phosphorylation Interaction Networks by Text Mining of Full-length Articles using the eFIP System

Catalina O. Tudor[1], Karen E. Ross[2], Gang Li[3], K Vijay-Shanker[3], Cathy H. Wu[4,5], Cecilia N. Arighi[4,6]

[1] Center for Bioinformatics and Computational Biology, Department of Computer and Information Sciences, University of Delaware, USA
[2] Center for Bioinformatics and Computational Biology, University of Delaware, USA
[3] Department of Computer and Information Sciences, University of Delaware, USA
[4] Center for Bioinformatics and Computational Biology, Department of Computer and Information Sciences, University of Delaware, USA
[5] Protein Information Resource, Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, USA
[6] Protein Information Resource

Presenter: Cecilia N. Arighi

Protein phosphorylation is a reversible post-translational modification where a protein kinase adds a phosphate group to a protein. This modification can affect protein-protein interactions (PPIs). Extracting phosphorylation information coupled with PPI information from the scientific literature will facilitate the creation of phosphorylation interaction networks of kinases, substrates and interacting partners, towards knowledge discovery of functional outcomes of protein phosphorylation. Increasingly, PPI databases are interested in capturing the phosphorylation state of interacting partners. We have previously developed the eFIP (Extracting Functional Impact of Phosphorylation) text mining system, which identifies phosphorylated proteins and phosphorylation-dependent PPIs. In this work, we present several enhancements for the eFIP system: text mining for full-length articles from the PubMed Central open-access collection; integration of the RLIMS-P 2.0 system for the extraction of phosphorylation events with kinase, substrate and site information; extension of the PPI module with new trigger words/phrases describing interactions; and addition of the iSimp tool to aid in the matching of syntactic patterns. The website supports searches based on protein roles, or using keywords; link protein entities to their corresponding UniProt identifiers if mapped; and support visual exploration of phosphorylation interaction networks using Cytoscape. To demonstrate eFIP for knowledge extraction and discovery, we constructed phosphorylation-dependent interaction networks involving 14-3-3 proteins identified from cancer-related vs. diabetes-related articles. Comparison of the phosphorylation interaction network of kinases, phosphoproteins, and interactants obtained, along with enrichment analysis of the protein set, revealed several shared interactions, highlighting common pathways discussed in the context of both diseases. URL: http://proteininformationresource.org/efip

SourceData: integrating biocuration within the publishing process

Robin Liechti[1], Lou Götz[1], Anne Niknejad[1], Ioannis Xenarios[1], Thomas Lemberger[2]

[1] Vital-IT, SIB Swiss Institute of Bioinformatics, Switzerland
[2] EMBO, Germany

Presenter: Thomas Lemberger

Science is a global endeavor that requires the ongoing exchange of ideas and research findings. Efficient access to reliable research data is of paramount importance for the advancement of science. Research data are published in scientific papers as figures, which do not allow re-analysis of the data and are inaccessible to systematic data mining or search. It is thus currently extremely difficult to verify whether an experiment has already been published before or to compare the data from related studies. As research output continues to grow massively, new solutions are urgently required to maximize the efficiency of research investments. To address these issues, the collaborative SourceData project has been initiated by several partners to develop the necessary editorial tools and workflows that enable the biocuration of figures by data editors during the production phase of the publication process. We will present the SourceData curation tool, which allows 1) to delimit coherent experimental units within a figure; 2) to efficiently tag biochemical entities in figure captions; 3) to normalize entities; and 4) specify their role in the experimental design. We will furthermore show how the resulting semantic information can be used in data-oriented search strategies that enable researchers to find, compare and combine data from different sources to generate new hypotheses and stimulate novel discoveries. SourceData aims at transforming published articles into open and enriched resources that make published figures and the underlying source data searchable and available to researchers. By improving the accessibility and discoverability of published data, SourceData will contribute to increase the transparency of the scientific output and promote data reuse.

Bioso! – A Search Engine & Annotation Framework for Biological Big Data

Yin Huang[1], Huimin Wu[2], Bin Li[1], Chi Jin[2], Tianyun Xie[2], Jifei Huang[2], Tingling Li[2], Weimin Zhu[1]

[1] Beijing Proteome Research Center, China
[2] Suzhou hadoo software Co.,Ltd, China

Presenter: Yin Huang

As the rapid increase of biological data from different domain spaces, the need to integrate disparate biological databases, not only as the valuable data warehouse for sophisticated query & retrieval, but also as the rich information & knowledge resource for omics data annotation, has never been greater. Bioso! is the effort to address this urgent need in the era of biomedical big data. As the search engine, it supports well-performed cross-database search against 20+ biomedical databases managed by Lucene & semantic rule-powered integration framework, the search results are presented in an encyclopedic view with easy access to the different facets of a biological object. As the technical framework, it is highly scalable and offering flexible interfaces for programmatic access to integrated information, making it a rich knowledgebase to annotate different omics data.

Improved Data Representation of Very Large Macromolecules at Protein Data Bank

Jasmine Young[1], ChunXiao Bi[2], Li Chen[3], Cole Christie[2], Zukang Feng[4], Vladimir Guranovic[4], Brian Hudson[4], Ezra Peisach[4], Andreas Prlic[5], Peter Rose[6], Chenghua Shao[4], John Westbrook[4], wwPDB Team[7,8,9,10]

[1] Rutgers, The State University of New Jersey, RCSB Protein Data Bank, USA; [2] UCSD, RCSB Protein Data Bank, USA; [3] Center for Integrative Proteomics Research, Protein Data Bank, USA; [4] Center for Integrative Proteomics Research, RCSB Protein Data Bank, USA; [5] SDSC, RCSB Protein Data Bank, USA; [6] SDSC, RCSB Protein Data Bank, USA; [7] Rutgers, The State University of New Jersey, RCSB PDB, United States; [8] EMBL-European Bioinformatics Institute, PDBe, United Kingdom; [9] Institute for Protein Research, Osaka University, PDBj, Japan; [10] University of Wisconsin-Madison, BMRB, United States

Presenter: Jasmine Young

The Protein Data Bank (PDB) is the single global repository for three-dimensional structures of biological macromolecules and their complexes. Over the past decade, the size and complexity of macromolecules and their complexes with small molecules deposited to the PDB have increased significantly. These structures, such as ribosomes and viruses, provide essential information for understanding biochemical processes and structure-based drug discovery. Historically, large structures in the PDB have been "split" across multiple archival entries because of file size limitations (containing >62 polymeric chains and/or 99999 atoms) related to the 80 character/line record format dating from 1970s. The PDBx/mmCIF dictionary defines deposited data items and meta data. PDB data processing software tools use this dictionary to maintain data consistency across PDB archive. To support scientific advancement and ensure the best data quality and completeness, a working group composed of community experts in software development was established to enable direct use of PDBx/mmCIF format files that is not subject to size limitation in the major macromolecular crystallographic software tools. This PDBx/mmCIF Format Working Group has also made recommendations concerning extensions essential for proper treatment of large structures deposited to the PDB. Now that macromolecular crystallographic software tools that produce PDBx/mmCIF format data files for deposition are available, the large structures in the PDB archive have been unified into single large files in PDBx/mmCIF and XML formats for improved data representation and integration. The entire newly unified PDB archive was released in December 2014. The impact of improved data representation of these large structures in the PDB archive will be presented. wwPDB members are RCSB PDB (supported by NSF, NIH, and DOE), PDBe (EMBL-EBI, Wellcome Trust, BBSRC, NIGMS, and EU), PDBj (NBDC-JST) and BMRB (NLM).

An update of miRNASNP database for better SNP selection by GWAS data, miRNA expression and online tools

Anyuan Guo[1]

[1] College of Life Science and Technology, Huazhong University of Science and Technology, China

Presenter: Anyuan Guo

MicroRNAs (miRNAs) are key regulators of gene expression involved in a broad range of biological processes. MiRNASNP aims to provide single nucleotide polymorphisms (SNPs) in miRNAs and genes that may impact miRNA biogenesis and/or miRNA target binding. Advanced miRNA research provided abundant data about miRNA expression, validated targets and related phenotypic variants. In miRNASNP v2.0, we have updated our previous database with several new data and features, including: (i) expression level and expression correlation of miRNAs and target genes in different tissues, (ii) linking SNPs to the results of genome-wide association studies (GWAS), (iii) integrating experimentally validated miRNA:mRNA interactions, (iv) adding multiple filters to prioritize functional SNPs. In addition, as a supplement of the database, we have set up three flexible online tools to analyze the influence of novel variants on miRNA:mRNA binding. A new nice web interface was designed for miRNASNP v2.0 allowing users to browse, search and download (http://bioinfo.life.hust.edu.cn/miRNASNP2/). We aim to maintain the miRNASNP as a solid resource for function, genetics and disease studies of miRNA related SNPs.

Gene Curation Software at the Rat Genome Database: Update 2015

Stanley J. Laulederkind[1], Weisong Liu[2], Marek Tutaj[1], George T. Hyaman[1], Rajni Nigam[1], Victoria Petri[1], Jennifer R. Smith[1], Shur-Jen Wang[1], Jeff De Pons[1], Melinda R. Dwinell[1], Mary Shimoyama[1]

[1] Human and Molecular Genetics Center, Medical College of Wisconsin, USA
[2] Department of Quantitative Health Sciences, University of Massachusetts Medical School, USA

Presenter: Stanley J. Laulederkind

At model organism databases data is curated for numerous biological categories using various ontologies and vocabularies. The Rat Genome Database (RGD) uses four different ontologies to standardize annotation information for genes and their associations with disease, phenotypes, pathways, and other biological data. For manual gene curation this is all done in a single user interface of a web-based annotation tool developed at RGD. The same interface is used for the curation of QTLs and strains. The development of the tool has been achieved through a collaboration of curators and software developers. Features have been tailored to the needs of the curators to allow optimum efficiency of the data entry portion of the curation process. Annotations using the Gene Ontology, Mammalian Phenotype Ontology, Pathway Ontology, and RGD Disease Ontology can be done simultaneously in the same user interface. There have been multiple upgrades and additions to the RGD curation tool in the past few years. The upgrades range from improved results organization to improved post-annotation editing. Two big additions to the curation tool are an embedded ontology browser and a literature searching interface (OntoMate). Terms from the ontology browser are hyperlinked to the curation tools ontology term “bucket”, which cuts out a middle step of searching for a term in the tool after separately finding the appropriate term in an ontology browser. From the OntoMate interface, the curation tool automatically downloads the reference and assigns an RGD ID to it. Pre-selected ontology terms can also be loaded into the curation tool from the OntoMate interface. Annotation and use data automatically gets sent back to OntoMate for feedback to the curators. Embedding the ontology browser has made selecting and loading terms about 60% faster, while the OntoMate interface has increased the speed of gene curation about 30%.

dbPSP: a curated database for protein phosphorylation sites in prokaryotes

Zhicheng Pan[1,2], Bangshan Wang[1,2], Ying Zhang[2], Yongbo Wang[2], Shahid Ullah[2], Zexian Liu[2], Yu Xue[2]

[1] School of Life Sciences, University of Science and Technology of China, China
[2] Department of Biomedical Engineering, Huazhong University of Science and Technology, China

Presenter: Zexian Liu

As one of the most important post-translational modifications (PTMs), phosphorylation is highly involved in almost all of biological processes through temporally and spatially modifying substrate proteins. Recently, phosphorylation in prokaryotes attracted much attention for its critical roles in various cellular processes such as signal transduction. Thus, an integrative data resource of the prokaryotic phosphorylation will be useful for further analysis. In this study, we presented a curated database of phosphorylation sites in prokaryotes (dbPSP, Database URL: http://dbpsp.biocuckoo.org) for 96 prokaryotic organisms, which belong to 11 phyla in 2 domains including bacteria and archaea. From the scientific literature, we manually collected experimentally identified phosphorylation sites on 7 types of residues, including serine, threonine, tyrosine, aspartic acid, histidine, cysteine, and arginine. In total, the dbPSP database contains 7,391 phosphorylation sites in 3,750 prokaryotic proteins. With the dataset, the sequence preferences of the phosphorylation sites and functional annotations of the phosphoproteins were analyzed, while the results shows that there were obvious differences among the phosphorylation in bacteria, archaea and eukaryotes. All the phosphorylation sites were annotated with original references and other descriptions in the database, which could be easily accessed through user-friendly website interface including various search and browse options. Taken together, the dbPSP database provides a comprehensive data resource for further studies of protein phosphorylation in prokaryotes.

Shared Resources, Shared Costs – Leveraging Biocuration Resources

Sandra Orchard[1], Henning Hermjakob[1]

[1] Proteomics Services Team, European Bioinformatics Institute (EMBL-EBI), UK

Presenter: Sandra Orchard

The manual curation of the information in biomedical resources is an expensive task. This paper argues the value of this approach in comparison with other apparently less costly options, such as automated annotation or text-mining, then discusses ways in which databases can make cost savings by sharing infrastructure and tool development. Sharing curation effort is a model already being adopted by several data resources. Approaches taken by two of these, the Gene Ontology annotation effort and the IntAct molecular interaction database, are reviewed in more detail. These models help to ensure long term persistence of curated data and minimises redundant development of resources by multiple disparate groups.

Improving the Consistency of Domain Annotation within the Conserved Domain Database (CDD).

Myra K. Derbyshire[1], Noreen R. Gonzales[1], Shennan Lu[1], Jane He[1], Gabriele H. Marchler[1], Zhouxi Wang[1], Aron Marchler-Bauer[1]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA

Presenter: Myra K. Derbyshire

When annotating protein sequences with the footprints of evolutionarily conserved domains, conservative score or E-value thresholds need to be applied for RPS-BLAST hits, in order to avoid many false positives. We notice that manual inspection and classification of hits gathered at a higher threshold can add a significant amount of valuable domain annotation. We report an automated algorithm which “rescues” valuable borderline-scoring domain hits that are well-supported by domain architecture (DA, the sequential order of conserved domains in a protein query), including tandem repeats of domain hit(s) reported at a more conservative threshold. This algorithm is now available as a selectable option on the public conserved domain search pages. We also report on the possibility to “suppress” domains hits close to the threshold based on a lack of well-supported DA, and implemented conservatively as an option in live conserved domain searches and for pre-computed results. Improving domain annotation consistency will in turn reduce the fraction of NR sequences with incomplete DAs. CD-Search URL: http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

Automated collection, curation and analysis for Transcriptional and Epigenetic Regulation Data

Xiaole S. Liu[1,2], Hanfei Sun[2], Qian Qin[2]

[1] Biostatistics and Computational Biology, Harvard School of Public Health, United States
[2] Tongji University, School of Life Sciences and Technology, China

Presenter: Hanfei Sun

Chromatin immunoprecipitation and DNase I hypersensitivity assays with high-throughput sequencing have greatly accelerated the understanding of transcriptional and epigenetic regulation, although meta-data and data reuse for the community has been challenging. We created a platform called CistromeDB that can collect latest information of experiments from NCBI GEO database and ENCODE project, and then curate these information automatically using our knowledge base and text-mining approaches. Apart from a curation system for meta-data, CistromeDB also has a consistent analysis and quality control pipeline, therefore, user could evaluate the overall quality of each datasets. Moreover, CistromeDB integrates similarity detection functions for analysis result which is helpful to find duplicates and mistakes in original annotation. Currently, CistromeDB focuses on curating Chromatin immunoprecipitation and DNase I hypersensitivity assays with high-throughput sequencing data in human and mouse, containing 22601 samples over 14553 datasets, 974 factors and 579 cell lines or cell populations. It also has a web interface that helps users to query, evaluate and visualize the database, which can be access at http://cistrome.org/db

mycoCLAP, the Database for Characterized Lignocellulose-Active Proteins of Fungal Origin: Resource and Text Mining Curation Support

Gregory Butler[1], Kimchi Strasser[1], Erin McDonnell[1], Carol Nyaga[1], Marie-Jean Meurs[1], Min Wu[1], Sherry Wu[1], Justin Powlowski[1], Adrian Tsang[1], Leila Kosseim[2], Hayda Almeida[2]

[1] Centre for Structural and Functional Genomics, Concordia University, Canada
[2] Computer Science and Software Engineering, Concordia University, Canada

Presenter: Gregory Butler

Enzymes active on components of lignocellulosic biomass are used for indus- trial applications ranging from food processing to biofuels production. These in- clude a diverse array of glycoside hydrolases, carbohydrate esterases, polysaccha- ride lyases and oxidoreductases. Fungi are prolific producers of these enzymes, spurring fungal genome sequencing efforts to identify and catalogue the genes that encode them. In order to facilitate the functional annotation of these genes, biochemical data on over 800 fungal lignocellulose-degrading enzymes have been collected from the literature and organized into the searchable database, mycoCLAP (http://mycoclap.fungalgenomics.ca). First implemented in 2011, and updated as described here, mycoCLAP is capable of ranking search re- sults according to closest biochemically characterized homologues: this improves the quality of the annotation, and significantly decreases the time required to anno- tate novel sequences. The database is freely available to the scientific community, as are the open source applications based on natural language processing devel- oped to support the manual curation of mycoCLAP.

dbPPT: a comprehensive database of protein phosphorylation in plants

Han Cheng[1], Wankun Deng[1]

[1] Department of Biomedical Engineering, Huazhong University of Science and Technolog, China

Presenter: Han Cheng

As one of the most important protein post-translational modifications, the reversible phosphorylation is critical for plants in regulating a variety of biological processes such as cellular metabolism, signal transduction and responses to environmental stress. Numerous efforts especially large-scale phosphoproteome profiling studies have been contributed to dissect the phosphorylation signaling in various plants, while a large number of phosphorylation events were identified. To provide an integrated data resource for further investigations, here we present a comprehensive database of dbPPT (database of Phosphorylation site in PlanTs, at http://dbppt.biocuckoo.org), which contains experimentally identified phosphorylation sites in proteins from plants. The phosphorylation sites in dbPPT were manually curated from the literatures, whereas datasets in other public databases were also integrated. In total, there were 82 175 phosphorylation sites in 31 012 proteins from 20 plant organisms in dbPPT, presenting a larger quantity of phosphorylation sites and a higher coverage of plant species in comparison with other databases. The proportions of residue types including serine, threonine and tyrosine were 77.99, 17.81 and 4.20%, respectively. All the phosphoproteins and phosphorylation sites in the database were critically annotated. Since the phosphorylation signaling in plants attracted great attention recently, such a comprehensive resource of plant protein phosphorylation can be useful for the research community.

Generating a focused view of Disease Ontology cancer terms for pan-cancer data integration and analysis.

Raja Mazumder[1]

[1] Department of Biochemistry and Molecular Medicine, The George Washington University, USA

Presenter: Raja Mazumder

Bio-ontologies provide terminologies for the scientific community to describe biomedical entities in a standardized manner. There are multiple initiatives that are developing biomedical terminologies for the purpose of providing better annotation, data integration and mining capabilities. Terminology resources devised for multiple purposes inherently diverge in content and structure. A major issue of biomedical data integration is the development of overlapping terms, ambiguous classifications, and inconsistencies represented across databases and publications. The Disease Ontology (DO) was developed over the past decade to address data integration, standardization and annotation issues for human disease data. We have established a DO cancer project to be a focused view of cancer terms within the Disease Ontology. The DO cancer project mapped 386 cancer terms from the Catalogue of Somatic Mutations in Cancer (COSMIC), The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), Therapeutically Applicable Research to Generate Effective Treatments (TARGET), Integrative Oncogenomics (IntOGen), and the Early Detection Research Network (EDRN) into a cohesive set of 187 DO terms represented by 63 top level DO cancer terms. For example, the COSMIC term "kidney,NS,carcinoma,clear_cell_renal_cell_carcinoma" and TCGA term "Kidney renal clear cell carcinoma [KIRC]" were both grouped to the term "DOID:4467 / renal clear cell carcinoma" which was mapped to the TopNodes_DOcancerslim term "DOID:263 / kidney cancer". Mapping of diverse cancer terms to DO and the use of top level terms (DO slims) will enable pan-cancer analysis across datasets generated from any of the cancer term sources where pan-cancer means including or relating to all or multiple types of cancer. The terms can be browsed from the DO web site (http://www.disease-ontology.org) and downloaded from the DO's SVN or GitHub repositories.

MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data

Dong Zou[1], Shixiang Sun[1,2], Rujiao Li[3], Jiang Liu[1], Jing Zhang[1], Zhang Zhang[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Dong Zou

DNA methylation plays crucial roles during embryonic development. Here we present MethBank (http://dnamethylome.org), a DNA methylome programming database that integrates the genome-wide single-base nucleotide methylomes of gametes and early embryos in different model organisms. Unlike extant relevant databases, MethBank incorporates the whole-genome single-base-resolution methylomes of gametes and early embryos at multiple different developmental stages in zebrafish and mouse. MethBank allows users to retrieve methylation levels, differentially methylated regions, CpG islands, gene expression profiles and genetic polymorphisms for a specific gene or genomic region. Moreover, it offers a methylome browser that is capable of visualizing high-resolution DNA methylation profiles as well as other related data in an interactive manner and thus is of great helpfulness for users to investigate methylation patterns and changes of gametes and early embryos at different developmental stages. Ongoing efforts are focused on incorporation of methylomes and related data from other organisms. Together, MethBank features integration and visualization of high-resolution DNA methylation data as well as other related data, enabling identification of potential DNA methylation signatures in different developmental stages and accordingly providing an important resource for the epigenetic and developmental studies.

Poster Abstract #45

LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs

Lina Ma[1], Ang Li[1], Dong Zou[1], Xingjian Xu[1,2], Lin Xia[1,2], Jun Yu[1], Vladimir Bajic[3], Zhang Zhang[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Mathematical Sciences and Engineering Division (CEMSE), King Abdullah University of Science and Technology (KAUST), Kingdom of Saudi Arabia

Presenter: Lina Ma

Long non-coding RNAs (lncRNAs) perform a diversity of functions in numerous important biological processes and are implicated in many human diseases. In this report we present lncRNAWiki (http://lncrna.big.ac.cn), a wiki-based platform that is open-content and publicly editable and aimed at community-based curation and collection of information on human lncRNAs. Current related databases are dependent primarily on curation by experts, making it laborious to annotate the exponentially accumulated information on lncRNAs, which inevitably requires collective efforts in community-based curation of lncRNAs. Unlike existing databases, lncRNAWiki features comprehensive integration of information on human lncRNAs obtained from multiple different resources and allows not only existing lncRNAs to be edited, updated and curated by different users but also the addition of newly identified lncRNAs by any user. It harnesses community collective knowledge in collecting, editing and annotating human lncRNAs and rewards community-curated efforts by providing explicit authorship based on quantified contributions. LncRNAWiki relies on the underling knowledge of scientific community for collective and collaborative curation of human lncRNAs and thus has the potential to serve as an up-to-date and comprehensive knowledgebase for human lncRNAs.

Genomics external service tracking and management

Daniel Li[1], Yang Cheng[1], Lingling Shen[1]

[1] Novartis Institutes for BioMedical Research (China), NIBR Informatics, China

Presenter: Lingling Shen

Genomics analyses, which include but not limited to both array-based platform and next generation sequencing (NGS) technologies have been proved to be powerful means to disclose the mysterious mechanisms in disease research areas. When employing these techniques, investigators have to face troublesome problems, like how to track service requests between different vendors, how to streamline the business workflow of external genomics services and how to manage the various information and data generated during the services. As we investigated, there has not been such a system or solution which can handle all the problems emerged from external genomics services. Based on the requirements collected from research practices, we have developed a new system and try to solve these problems in this single platform. The main features we implemented in our system include a single service request interface for different vendors, a real-time business workflow module to track and manage the lifecycle of a service request, an automatic fast QC pipeline and an interface for communicating with follow-up analyses like microarray analysis and data visualization. With this new system, an external genomics service request will be raised, tracked and managed with a centralized platform in a more efficient, intuitive and friendly way. All involved users including scientists, financial users and IT administrators will benefit from this system.

COSMIC - Exploring the world's knowledge of somatic mutations in cancer

Sari A. Ward[1], Sally Bamford[1], Charlotte G. Cole[1], David M. Beare[1], Nidhi Bindal[1], Charalambos Boutselakis[1], Simon A. Forbes[1], John Gamble[1], Prasad Gunasekaran[1], Mingming Jia[1], Chai Yin Kok[1], Kenric Leung[1], Ding Minjie[1], Rebecca Shepherd[1], Jon W. Teague[1], Michael R. Stratton[1], Peter J. Campbell[1]

[1] Cancer Genome Project, Wellcome Trust Sanger Institute, U.K.

Presenter: Sari A. Ward

COSMIC, the Catalogue Of Somatic Mutations In Cancer (http://cancer.sanger.ac.uk) is the world's largest and most comprehensive resource to explore the impact of somatic mutations in human cancer. Our latest release (Nov 2014) describes 2,710,449 coding point mutations in over one million tumour samples, across most human genes. To emphasise depth of knowledge on known cancer genes, mutation information is curated manually from the scientific literature, allowing very precise definitions of disease types and clinically-relevant patient details. Combination of over 20,000 published studies gives substantial resolution of how mutations and phenotypes relate in human cancer, providing insights into the stratification of populations and new diseases behind known biomarkers. Conversely, our curation of cancer genomes (over 15,000) emphasises knowledge breadth, driving discovery of new unrecognised cancer-driving hotspots and molecular targets. In addition to describing over two million coding point mutations across cancer, COSMIC also details more than six million non-coding mutations, 10,567 gene fusions, 61,232 genome rearrangements, 702,652 abnormal copy number segments, and 6,566,072 abnormal expression variants. All these types of somatic mutation are annotated to both the human genome and each affected coding gene, then correlated across disease and mutation types. Our annotations are beginning to emphasise events with a higher impact in cancer, highlighting nonsynonymous coding mutations and major amplifications and deletions. This concept of high-impact data is being extended across the entire COSMIC system, much more strongly defining genes and mutations which drive oncogenesis.

DrugVar: An integrated Database of Germline & Somatic Non-synonymous Variations That Impact Drug Binding

Cheng Yan[1]

[1] The Department of Biochemistry & Molecular Medicine, The George Washington University, USA

Presenter: Cheng Yan

Due to the advancement of next generation sequencing (NGS) technology, genome sequencing is becoming economically available to public. However, the challenge of converting NGS data to clinical interpretation remains. We have developed DrugVar, an integrated database of germline and somatic non-synonymous single nucleotide variations (nsSNV) that impact drug binding. DrugVar in conjunction with exome or whole-genome sequencing data can be used to identify patients who may not respond to a specific drug because of loss of binding sites due to nsSNVs. . The workflow used to develop DrugVar is as follows: the structure of protein-drug complex is retrieved from Protein Data Bank (PDB) followed by calculation protein-drug packing interaction and identification of the protein-drug binding sites located in the drug binding pocket. Currently, DrugVar database includes 13,011 protein-drug binding sites across 253 proteins and 235 drugs. The integration of protein-drug binding dataset and both germline and somatic nsSNVs datasets reveals 3,133 mutations affecting protein-drug binding sites. The amino acid variations in drug binding sites can potentially lead to alterations in drug efficacy. Based on protein similarity and protein-drug binding sites, 81 paralogs were identified as potential drug targeting proteins and may serve as alternative drug targets or source of side effects hence mutation profiles of these paralogs were also investigated. The complex interactive network of these key drugs and proteins provides a key knowledge source for personalized medicine.

Poster Abstract #49

An efficient and scalable toolset for family-based sequencing data analysis

Min He[1,2,3], Thomas N. Person[1], Murray H. Brilliant[1,3]

[1] Center for Human Genetics, Marshfield Clinic Research Foundation, USA
[2] Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, USA
[3] Computation and Informatics in Biology and Medicine, University of Wisconsin-Madison, USA

Presenter: Min He

Next-generation sequencing technologies are now increasingly used to find disease genes in human genomic studies, but it is difficult to infer biological insights from massive amounts of data in a short period of time. We developed a software framework called SeqHBase to quickly identify disease-contributing genes. SeqHBase is a big data-based toolset for analyzing family-based sequencing data to detect de novo, inherited homozygous or compound heterozygous mutations that may contribute to disease manifestations. It was developed based on Apache Hadoop and HBase infrastructure, which works through distributed and parallel manner over multiple data nodes. Its input includes coverage information of 3 billion sites, over 3 million variants and their associated functional annotations for each whole genome. With 20 data nodes, SeqHBase took about 5 seconds for analyzing whole-exome sequencing (WES) data for a family quartet and approximately 1 minute for analyzing whole-genome sequencing (WGS) data for a 10-member family. These results demonstrated SeqHBase's high efficiency and scalability, which is necessary as WGS and WES are rapidly becoming standard methods to study the genetics of familial disorders. In addition, it is distributed, customizable, and scalable based on the needs with available data volume. As more data become available, addition of more data nodes is possible, making the system very nimble. The newly added data nodes can be seamlessly incorporated with the existing system.

STAD the Structural Targets Annotation Database

Abdelkrim Rachedi[1]

[1] Department of Biology, University dr. Tahar Moulay, Algeria

Presenter: Abdelkrim Rachedi

Structural Targets Annotation Database (STAD) is a database that annotates and stores information about structural targets (mainly proteins selected for 3D-structure elucidation) understudy at laboratories in a number of South African, Algerian and Moroccan universities and institutes. The database is used for tracking the progress of targets as they are moved from a status to another until their 3D structures are solved and/or work on them has stopped. In addition to basic annotation of the targets, structural and functional inferred annotation is carried out for each target and provided in the search results. STAD provides a number of methods for information retrieval; the Direct search method can be used to find out information about the sequence of the targets, the stages of their production and structure determination. Patterns/Motifs search allows for finding targets with sequence motifs composed by users. Targets with overall sequence similarity to other public sequences are found by using the Sequence Alignment search method. Additional tools are also provided for further exploration of targets including a Status page, StatsBoard and a Structural Gallery. Unless targets are tagged as private based on source lab instructions, target detailed information from all contributing centers are made available for download in XML format. The project's aim is to annotate all and any target may be selected by any institute in the whole of the continent of Africa and provide data and analysis to worldwide scientific and general public communities. STAD can be accessed from the address http://www.bioinformaticstools.org/stad and is a contributing centre with target information to the Structural Biology Knowledgebase (SBKB) TargetTrack system; http://sbkb.org/tt/centers.html

GigaDB submission wizard to enable authors to curate their own metadata

Xiao SiZhe[1], Robert L. Davidson[1], Peter Li[1], Laurie Goodman[1], Scott C. Edmunds[1], Christopher I. Hunter[1]

[1] Gigascience, BGI-HK, Hong Kong SAR China

Presenter: Xiao SiZhe

GigaDB (http://www.gigadb.org) is an integrated database of 'big-data' studies from the life sciences. The initial goals of GigaDB are to assign Digital Object Identifiers (DOIs) to datasets to allow them to be tracked and cited, and to provide a user-friendly web interface for providing easy access to GigaDB datasets and files. The new GigaDB submission wizard can help authors curate their own metadata (information about the dataset), which allows searching and indexing. The submission wizard is based on the GigaDB database schema. It is divided to three main parts: Study, Authors and Sample details, with the ability to include links to Projects, genome browsers, related DOI(s) and even other external URL links. Authors simply follow the steps and fill in the information on the web-forms. After curators review the submission, authors will be given credentials to allow upload of data files to the GigaDB server. The submission wizard can then be used to add and modify any metadata about the files (description, format, links to samples etc...). GigaDB creates DataCite Metadata standard compliant XML and after our curators review your submission, this is used to assign a unique DOI to the dataset, which can be used for citation purposes. GigaDB is an open-access data repository closely linked to GigaScience journal, we aim to promote reproducibility of published research by allowing access to all the data, and maximize the reuse of published data by giving a means to credit data publication by citation. The new GigaDB submission wizard plays an important role in the GigaDB structure. It provides a new simple to use web-interface for authors to curate their own data with the most relevant and accurate information. We wish to solicit open discussions from others to help our future developments. (database@gigasciencejournal.com)

Regulatory Sequence Feature Curation in WormBase.

Xiaodong Wang[1], Daniela Raciti[1], Mary Ann Tuli[1], Gary Williams[2], Paul Sternberg[2]

[1] Division of Biology and Biological Engineering, California Institute of Technology, US
[2] EBI, UK

Presenter: Xiaodong Wang

Regulatory Sequence Feature Curation in WormBase. Xiaodong Wang, Daniela Raciti, Mary Ann Tuli, Gary Williams*, Paul Sternberg and WormBase Consortium, California Institute of Technology, Pasadena, CA, US *Hinxton, UK In WormBase, we curate cis-regulatory sequence features described in published literature in a few different ways. Firstly, transcription factor (TF) binding motifs, promoter regions and enhancers are curated as cis-regulators of cis-regulated genes in a regulatory interaction. When promoters and enhancer regions are associated with spatio-temporal gene expression patterns we curate them as expression objects using defined anatomy and developmental ontologies. Secondly, when pertinent evidence exists, for example Electrophoretic Mobility Shift Assay (EMSA), yeast one hybrid and DNaseI footprinting assays, sequence features can be curated as interactors within a physical interaction. Thirdly, we curate position weight/frequency matrices (PWM/PFM) that can be used to locate potential TF binding sites within the genome. Additionally, large-scale datasets including RNASeq & ChIP-seq data from the modEncode project, are also integrated in WormBase. All curated sequence features can be viewed by turning on tracks in GBrowse. Sequence features, as an entity involving an interaction can also be visualized in Cytoscape, which is within the Interaction widget of the Gene Summary page. The challenge now is how to organize these features into transcriptional regulatory gene networks.

An extensive and interactive database of super-enhancers – dbSUPER

Aziz Khan[1], Xuegong Zhang[1,2]

[1] MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic and Systems Biology, TNLIST/Department of Automation, Tsinghua University, China
[2] School of Life Sciences, Tsinghua University, China

Presenter: Aziz Khan

Enhancers are cis-regulatory elements of DNA that enhance the transcription of target genes and play a key role in development and diseases. Recently, a new class of enhancers named “super-enhancers” has been discovered, which refers to large clusters of transcriptional enhancers that can drive cell-type-specific gene expression and are crucial in cell identity. Many disease-associated sequence variations are enriched in these super-enhancer regions of disease-relevant cell types. Thus, super-enhancers can be used as potential biomarkers for disease diagnosis and therapeutics. Current studies have identified super-enhancers for more than 100 cell types in human and mouse. However, no centralized resource to integrate all these findings is available yet. We developed dbSUPER (http://bioinfo.au.tsinghua.edu.cn/dbsuper/), the first extensive and interactive database of super-enhancers in mouse and human genome, with the primary goal of providing a resource for further study of transcriptional control of cell identity and disease by archiving computationally produced data. We provide an interactive data transfer platform to easily send the data to Galaxy, GREAT and Cistrome web servers for further downstream analysis. dbSUPER provides a responsive and user-friendly web interface to facilitate efficient and comprehensive searching and browsing. dbSUPER provides downloadable and exportable features in a variety of data formats, and can be visualized in UCSC genome browser while custom tracks will be added automatically. Further, dbSUPER lists genes associated with super-enhancers and links to various databases, including GeneCards, UniProt and Entrez. Our database also provides an overlap analysis tool, to check the overlap of user defined regions with the current database. Currently, our database contains 66,033 super-enhancers for 96 human and 5 mouse tissue/cell types. We believe, dbSUPER is a valuable resource for the bioinformatics and genetics research community.

Protein 3D-structures as precious sources of information in UniProtKB

Ursula Hinz[1], the UniProt consortium[1,2,3]

[1] Swiss-Prot, SIB Swiss Institute of Bioinformatics, Switzerland
[2] The European Bioinformatics Institute, United Kingdom
[3] Protein Information Resource, USA

Presenter: Ursula Hinz

The UniProt knowledgebase, UniProtKB, provides access to high-quality information about proteins and their sequences, plus cross-references to multiple resources. The knowledgebase contains a manually annotated section, UniProtKB /Swiss-Prot, and another section, UniProtKB TrEMBL, annotated by automated procedures. 3D-structures are unique and precious sources of information about proteins, their functions, interactions, domains and ligand binding sites. UniProtKB entries facilitate access to these resources via cross-references to PDB, but also to the Protein Model Portal and related resources. In parallel, UniProtKB/Swiss-Prot biocurators make use of this wealth of data, combining information derived from 3D-structures and the scientific literature to identify ligand binding sites, enzyme active sites, post-translational modifications, membrane topology, and interactions between proteins, or proteins and nucleic acids. Information from well-characterized proteins is then propagated to close family members. The data are shown in a structured format, to facilitate access to specific pieces of information: protein function and subunit structure, subcellular location, the role of specific residues, domains and regions, post-translational modifications, etc., with evidence tags to indicate the sources of the information. In January 2015, UniProtKB/Swiss-Prot contained 118’000 cross-references to PDB, corresponding to over 21’600 entries, mostly from model organisms. Manual curation of information from protein 3D-structures has high priority. About 25% of the 20’000 human entries have a cross-reference to PDB, and the majority of these have at least one matching literature citation.

MimoDB update 2015: towards a Biopanning Data Bank

Bifang He[1], Huixiong Zhang[1], Zhiqiang Yan[1], Zechun Liu[2], Guoshi Chai[1], Liuyang Qiu[1], Qiang He[1], Ke Han[1], Beibei Ru[1], Peng Zhou[1], Hui Ding[1], Hao Lin[1], Xianlong Wang[1], Nini Rao[1], Feng-Biao Guo[1], Jian Huang[1]

[1] Key Laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, China
[2] School of Computer Science & Engineering, University of Electronic Science and Technology of China, China

Presenter: Bifang He

The MimoDB database (http://i.uestc.edu.cn/bdb) is a web portal providing access to information about biopanning results of combinatorial peptide library. Last updated in January 2015, this database contained 2806 sets of biopanning data collected from 1276 papers, including 24627 peptide sequences, 1652 targets, 483 known templates, 437 peptide libraries, and 263 crystal structures of target-template or target-mimotope complex. All data stored in MimoDB were revisited and new data fields on peptide affinity, measurement method and procedures were added. Qualitative or quantitative binding data and relevant measurement information of 1651 peptides from 318 sets of biopanning results were curated from 215 peer-reviewed papers. From this release, a more professional Model-View-Controller (MVC) architecture and more user-friendly web interfaces were implemented. A new on-the-fly data visualization tool and more resources for education and data analysis were hyperlinked. With the new data and services made available, we expect that the MimoDB database will continue its way to biopanning data bank and serve the biopanning and related scientific communities better.

2015 Disease Ontology update: DO's expanded curation activities to connect disease-related data

Elvira Mitraka[1], Lynn M. Schriml[1]

[1] University of Maryland School of Medicine, USA

Presenter: Elvira Mitraka

The Human Disease Ontology (DO) is a widely used biomedical resource, which standardizes and classifies common and rare human diseases. Its latest iteration makes use of the OWL language to facilitate easier curation between a variety of working groups and to take advantage of the analyses available using OWL. The DO integrates disease concepts from ICD-9, ICD-10, the National Cancer Institute Thesaurus, SNOMED-CT, MeSH, OMIM, EFO and Orphanet. The DO Team is focused on enabling mapping and curation of large disease datasets for major Biomedical Resource Centers and integration of their disease terms into DO. Constant updates and additions to the ontology allow for coverage of the vast field of human diseases. By having close collaborations with a variety of research groups, such as MGD, EBI, NCI, the Disease Ontology has established itself as the go-to tool for human disease curation. Implementing a combination of informatic tools and manual curation DO ensures that it maintains the highest standard possible. Future plans include the the use of evidence codes and the generation of various DO-Slim files to represent resource-specific views of DO for each major DO curation project.

TAIR as a Model for Database Sustainability

Donghui Li[1], Eva Huala[1], Tanya Berardini[1], Robert Muller[1]

[1] Phoenix Bioinformatics, United States

Presenter: Donghui Li

The long-term stable support of databases that serve the research community is an ongoing challenge. Traditional grant funding is an excellent mechanism for developing new resources but has drawbacks as a long-term support mechanism. However, persistence and continued availability of important datasets is an essential underpinning of scientific progress. We have pioneered a new approach to this problem that combines a nonprofit approach with subscriber support, and have successfully transitioned the TAIR database to community funding at a level that can sustain the resource for the long term. The next challenge is building an infrastructure that can extend to other databases that wish to make the same transition. Websites: www.arabidopsis.org, www.phoenixbioinformatics.org.

Annotations of cancer-related microRNAs (oncomiRs)

Jin Gu[1]

[1] Tsinghua University, China

Presenter: Jin Gu

MicroRNAs play important roles in cancer. We proposed several computational methods to identify and annotate the cancer-related microRNAs (oncomiRs) and their regulatory networks based on genomic data. Recently, we identify tumor-suppressive miR-139 regulatory networks using gene module based master regulator inference (ModMRI). We also established a manually curated oncomiR database - oncomiRDB. We manually curated 2259 entries of cancer-related miRNA regulations with direct experimental evidence from ~9000 abstracts, covering more than 300 miRNAs and 829 target genes across 25 cancer tissues. A web-based portal named oncomiRDB, which provides both graphical and text-based interfaces, was developed for easily browsing and searching all the annotations. It should be a useful resource for both the computational analysis and experimental study on miRNA regulatory networks and functions in cancer.

lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse

Liu Wei[1], Gong Jing[2], Miao Xiaoping[2], Anyuan Guo[1]

[1] College of Life Science and Technology, Huazhong University of Science & Technology, China
[2] School of Public Health, Tongji Medical College, Huazhong University of Science & Technology, China

Presenter: Liu Wei

Long non-coding RNAs (lncRNAs) play key roles in various cellular contexts and diseases by diverse mechanisms. With the rapid growth of identified lncRNAs and disease-associated single nucleotide polymorphisms (SNPs), there is a great demand to study SNPs in lncRNAs. Aiming to provide a useful resource about lncRNA SNPs, we systematically identified SNPs in lncRNAs and analyzed their potential impacts on lncRNA structure and function. In total, we identified 495,729 and 777,095 SNPs in more than 30,000 lncRNA transcripts in human and mouse, respectively. A large number of SNPs were predicted with the potential to impact on the miRNA–lncRNA interaction. The experimental evidence and conservation of miRNA–lncRNA interaction, as well as miRNA expressions from TCGA were also integrated to prioritize the miRNA–lncRNA interactions and SNPs on the binding sites. Furthermore, by mapping SNPs to GWAS results, we found that 142 human lncRNA SNPs are GWAS tagSNPs and 197,827 lncRNA SNPs are in the GWAS linkage disequilibrium regions. All these data for human and mouse lncRNAs were imported into lncRNASNP database (http://bioinfo.life.hust.edu.cn/lncRNASNP/).

mycoCLAP, a Database for Characterized Lignocellulose-Active Proteins of Fungal Origin

Kimchi Strasser[1], Erin McDonnell[1], Carol Nyaga[1], Min Wu[1], Sherry Wu[1], Hayda Almeida[2], Marie-Jean Meurs[1], Leila Kosseim[2], Justin Powlowski[3], Greg Butler[4], Adrian Tsang[5]

[1] Centre of Structural and Functional Genomics, Concordia University, Canada
[2] Department of Computer Science and Software Engineering, Concordia University, Canada
[3] Centre of Structural and Functional Genomics/Department of Chemistry and Biochemistry, Concordia University, Canada
[4] Centre of Structural and Functional Genomics/Department of Computer Science and Software Engineering, Concordia University, Canada
[5] Centre of Structural and Functional Genomics/Department of Biology, Concordia University, Canada

Presenter: Kimchi A. Strasser

Fungi possess several enzyme families capable of degrading lignocellulolytic biomass including glycoside hydrolases, carbohydrate esterases, polysaccharide lyases, and oxidoreductases. Many fungal genes belonging to these families have been identified on the basis of sequence comparisons but relatively few of the enzymes encoded by these genes have been biochemically characterized. Molecular and biochemical data on over 800 lignocellulose-degrading enzymes have been collected and organized into the database mycoCLAP (http://mycoCLAP.fungalgenomics.ca) a classification system freely available to the scientific community. This database serves as a tool for the identification and preliminary functional analysis of novel lignocellulose-active enzymes.

Poster Abstract #61

Applying Curation Lifecycle Model to Manage Cross-disciplinary Data

Lin Yang[1], Jiao Li[1], Qing Qian[1]

[1] Institute of Medical Information & Library, Chinese Academy of Medical Sciences, China

Presenter: Lin Yang

With the development of data-intensive interdisciplinary research and data open access, it is becoming an important issue to manage cross-disciplinary data effectively which supports further knowledge discovery. Curation Lifecycle Model, a standardized framework, makes data management in a stable and sustainable manner. Digital Curation Center (DCC) Curation Lifecycle Model, one of the most representative curation models, is widely used by scientific data centers for data resource development and management. In this study, we applied this model to manage cross-disciplinary data, in terms of climate data and hosptial registration data, for supporting association analysis. We selected several typical biocuration practices, formulized their key steps and presented granular functionality using the curation model. In the case study, we constructed a curation workflow for environmental health data curation. Specifically, we applied the workflow to curate climate data and hosptial regist data, where climate data was collected from China ministry of environmental protection, including hourly air pollutants monitoring data and daily weather data, and the registration data was Emergency Department visit data from several hospitals of a metropolis in China. The result shows that Curation Lifecycle Model could practically help manage cross-disciplinary data for further data mining and association exploring.

AnimalTFDB: a comprehensive animal transcription factor database

Hongmei Zhang[1]

[1] Huazhong University of Science & Technology, China

Presenter: Hongmei Zhang

Transcription factors (TFs) are proteins that bind to specific DNA sequences, thereby playing crucial roles in gene-expression regulation through controlling the transcription of genetic information from DNA to RNA. Transcription cofactors and chromatin remodeling factors are also essential in the gene transcriptional regulation. Identifying and annotating all the TFs are primary and crucial steps for illustrating their functions and understanding the transcriptional regulation. In this study, based on manual literature reviews, we collected and curated 72 TF families for animals, which is currently the most complete list of TF families in animals. Then, we systematically characterized all the TFs in 50 animal species and constructed a comprehensive animal TF database, AnimalTFDB. To better serve the community, we provided detailed annotations for each TF, including basic information, gene structure, functional domain, 3D structure hit, Gene Ontology, pathway, protein–protein interaction, paralogs, orthologs, potential TF-binding sites and targets. In addition, we collected and annotated transcription cofactors and chromatin remodeling factors. AnimalTFDB has a user-friendly web interface with multiple browse and search functions, as well as data downloading. It is freely available at http://bioinfo.life.hust.edu.cn/AnimalTFDB/.

GigaDB schema update to accommodate the variety of data.

Christopher I. Hunter[1], Xiao SiZhe[1], Robert L. Davidson[1], Peter Li[1], Laurie Goodman[1], Scott C. Edmunds[1]

[1] GigaScience, BGI-HK, Hong Kong SAR China

Presenter: Christopher I. Hunter

GigaScience (http://www.gigasciencejournal.com) is an online, open-access journal that includes, as part of its publishing activities, the database GigaDB (http://www.gigadb.org). GigaScience is co-published in collaboration between BGI and BioMed Central, to meet the needs of a new generation of biological and biomedical research as it enters the era of "big-data." The journal's scope covers studies from the entire spectrum of the life sciences that produce and use large-scale data as the center of their work. Data from these articles are hosted in GigaDB, from where they can be cited to provide a direct link between the study and the data supporting it, as well as access to relevant tools for reproducing or reusing these data. Due to the scope of GigaScience, GigaDB needs to host a wider variety of data type than most biological databases. In order to make this possible, we have created and launched a new version of GigaDB that now uses a fully extensible database schema capable of handling this variety of data types and standards. The schema is divided into 3 main areas; Dataset, Sample and Data. These are roughly analogous to those used by other common systems for submitting /curating biological data, including the SRA (http://www.ebi.ac.uk/ena/submit/metadata-model) and the ISA infrastructure (http://isatab.sourceforge.net/format.html). The dataset part includes tables to store information about the overall study design, the authors and funding bodies. It also acts as a holder to link together all the samples and data associated with it, as well as providing links to external sources. The Sample area of the schema plays host to the sample metadata and sample relationships, including their relationship to particular data files. Here we present the schema, and in an accompanying abstract we show how it is implemented for metadata capture in our Submission Wizard, see related submission by Si Zhe Xiao et al.

Deep mining of big data in stem cell and Alzheimer's disease

Hongxing Lei[1], Guangchun Han[1], Jiajia Wang[1], Zhouxian Bai[1], Fuhai Song[1], Xing Peng[1]

[1] Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Hongxing Lei

Large volume of valuable data has been accumulated in the genomics era. We are especially interested in the deep mining of big data in stem cell and Alzheimer's disease(AD). For AD, our main focus is the gene dys-regulation in the AD brains which have been extensively measured by transcriptome experiments. Other relevant information include brain transcriptome of aging and other neurological disorders, blood transcriptome of AD, GWAS studies of AD and other diseases, disease progression studies of AD, and relevant regulatory interactions. We have curated these information in a database named AlzBase (Mol Neurobiology 2015) which is available online for the examination of genes of particular interest. The enhance the functionality, we have constructed several networks including co-expression network, co-citation network and mutual information network. We have also summarized the important features of the top dys-regulated genes and top genes from the GWAS studies of AD. For big data in the field of stem cell, we have conducted analysis on data relevant to somatic cell reprogramming. We are also working on transdifferentiation or direct conversion. We plan to organize the analysis results into an online database in the coming months.

Controlled vocabulary for capturing cellular lower level interactions by text-mining tools.

Madhura R. Vipra[1], Devaki Kelkar[1]

[1] Athena Consulting, India

Presenter: Madhura R. Vipra

Biomedical research relies heavily on knowledgebase curated from literature. Massive increase in the volume of literature has made manual curation expensive and time consuming. This has resulted in increased efforts in text-mining tools development, the success of which relies on accurate and comprehensive ontologies. Controlled vocabulary development describing detailed aspects of a domain is a critical task in itself. The key challenge in translational research is establishing causal relationships between clinical outcomes, and genetic factors that alter basic cellular events and interactions. Genetic variants induced alterations in mRNA expression are known and it is assumed that it further leads to consequent changes in protein levels; however not much data is available to test if indeed this assumption is true. Increasingly, researchers are convinced that capturing cellular level effects such as altered protein expression, protein-protein interactions etc. are vital for a more holistic view of disease. Automated tools often fail to pick up this valuable information which is sparsely represented in diverse terms and cannot be tracked in absence of relevant ontology. We have used manual curation to devise ontology that can be used to semi-automate text mining tool to capture cellular level changes in the context of phenotype. Diverse strategies were employed to build this knowledgebase as typically described in literature, and cover not only mRNA effects but also those related to apoptosis, cell growth, motility, cell cycle, focal adhesion, protein activity, to quote a few to enable encoding them into a formal relational database structure. We have assessed test cases across various diseases, variants and platforms to build this repository. To quote an example, we have ontology suggestions for synaptic long-term potentiation, neuregulin signalling, sonic hedgehog signalling, cell adhesion regulation etc. for genetic variants associated with schizophrenia.

Chakraview: Interactive Genome Data Exploration and Real Time Visual Analytics Software

Dhwanit Shah[1], Hartosh Singh Bugra[1], Hitesh Mohta[1], Arvind Hulgeri[1], Shailesh Patil[1], Randeep Singh[1,2]

[1] SAP Labs India Pvt. Ltd., India
[2] ACTREC, Tata Memorial Centre, India

Presenter: Randeep Singh

A variety of data analyses problems in genomics are modeled by forming multiple sets over a collection of entities and analyzing relationships between these sets. Chakraview is an interactive browser based circular visualization software for visualizing these sets, their relationships and measures across multiple dimensions. Chakraview consists of two parts – an independent front end (client-side) JavaScript visualization library called chakraview-js for generating data driven graphs and back end (server-side) software component for parsing, storing and processing genome data available in various popular file formats like VCF, GFF, BED etc. The chakraview-js is a highly flexible and configurable API that enables end user to easily create dynamic, interactive, aesthetic and publication-quality SVG graph on the fly by adding a circular base track (referential axis) and multiple circular 2D data tracks of different types like interval track, histogram track, fusion track and text track. The key component of the Chakraview architecture is the SAP HANA in-memory column-oriented database platform based server-side data services. Highly optimized SQLScript procedures provide processing of large amount of genome data in near real-time thereby enabling Chakraview to not only serve as a visualization tool but also as an analytic tool for real time aggregation and drill down. Since Chakraview runs in a browser with predefined configuration, researchers and biologists possessing basic computer skills can easily use it for genome data analysis, common use cases which include analyzing DNA-seq data, visualizing structural variations (gene fusions, translocations, inversions and indels), SNPs, sequence similarities (alignments) and RNA-seq plots. Further analytical use cases include filtering variant data by annotation data as well as by file-format attributes, for example filtering VCF data by QUAL thresholds or specific INFO values.

DERMO: an ontology for the description of dermatologic disease

Hannah M. Fisher[1], Robert Hoehndorf[2], Soheil S. Dadras[3], Lloyd E. King[4], Georgios V. Gkoutos[5], John P. Sundberg[6], Paul N. Schofield[7]

[1] Physiology, Development and Neuroscience, University of Cambridge, UK
[2] Computational Bioscience Research Center, King Abdullah University of Science and Technology, Saudi Arabia
[3] Dermatology, UConn Health Center, USA
[4] Dermatology, Vanderbilt University Medical Center, USA
[5] Computer Science, University of Aberystwyth, United Kingdom
[6] The Jackson Laboratories, USA
[7] Physiology, Development and Neuroscience, University of Cambridge, United Kingdom

Presenter: Robert Hoehndorf

The ability to capture diagnostic information on disease using standardised terminologies has always been critical for integration and analysis of patient data . This need is now made greater by the new approaches to computing on disease information and the advent of personalised medicine. There have been many initiatives to generate structured terminologies to accurately capture skin disease diagnoses, but the available tools either suffer from inappropriate structure or significant incompleteness. We have created an ontology, DERMO, which can be used for the coding of diagnoses from patients and non-human model organisms. It is a simple hierarchy and currently contains 3425 classes. These are mapped to other major terminologies, such as DermLex, UMLS, ICD9, and provided with synonyms and textual definitions. Disease entities are categorized into 18 upper level classes which use a variety of features such as anatomical location, heritability, affected cell or tissue type, or etiology, as the features for classification. Using our recently developed semantic text-mining tool, Aber-OWL (Hoehndorf et al., BMC Bioinformatics 2015, 16:26) we have annotated all of the diseases in DERMO with their associated phenotypes mined from the fulltext index of all titles and abstracts in MEDLINE/PubMed 2014, and all fulltext articles in PubMed Central. The ontology and associated phenotypes can be accessed on http://aber-owl.net/aber-owl/dermophenotypes/ . The ontology is available in OBO format from the project web site http://dermatology.googlecode.com/. The applications of DERMO are equally to patient care, clinical training and fundamental research. It will support automated inference and reasoning and can be used to support algorithms for patient stratification, genotype/phenotype studies, and for the broader integration of skin disease information with that from other domains, such as model organism phenotypes and pharmacogenomics, for translational science.

Human Gene Family Resources at genenames.org

Susan Tweedie[1], Ruth L. Seal[1], Kristian A. Gray[1], Mathew W. Wright[2], Elspeth A. Bruford[1]

[1] HGNC, European Bioinformatics Institute (EMBL-EBI), UK
[2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, USA

Presenter: Susan Tweedie

The HUGO Gene Nomenclature Committee (HGNC) has assigned unique approved gene symbols and names to over 39,000 human loci to date. Approximately 19,000 of these are protein coding genes, but we also name pseudogenes and non-coding RNAs. Our website, genenames.org, is a searchable repository of HGNC approved nomenclature and associated resources. Approved gene symbols are based on names describing structure, function or homology, and where possible these are organised into gene groupings and families, many of which have specialist advisors who are experts in that particular area of biology. We are continually adding new gene families and currently have over 600 gene family pages. We have recently improved the display and content of these pages and sorted many of the families into hierarchies which allow genes to be browsed and downloaded at either a top level such as “G protein-coupled receptors”, or at a more specific level, such as “Melanocortin receptors”. The new gene family pages include family aliases, a family hierarchy map, a family text description, a graphical representation of protein domains for an example family member, and links to publications and relevant external resources. The gene families are fully searchable via the Search tool found at the top of each page at genenames.org and there is also a gene family index where users can browse through the full list of families. If you know of a gene family that you think we should include or update, please contact us via hgnc@genenames.org or talk to us during this meeting.

Poster Abstract #69

Viral Representative Proteomes: Computational Clustering of UniProtKB Virus Proteomes

Chuming Chen[1], Hongzhan Huang[1], Darren A. Natale[2], Raja Mazumder[3], Peter B. McGarvey[2], Jian Zhang[2], Shawn W. Polson[1], Cecilia N. Arighi[1], Cathy H. Wu[1,2]

[1] Center for Bioinformatics and Computational Biology, Department of Computer and Information Sciences, University of Delaware, USA
[2] Protein Information Resource, Department of Biochemistry and Molecular & Cellular Biology, Georgetown University, USA
[3] Department of Biochemistry and Molecular Biology, The George Washington University, USA

Presenter: Cecilia N. Arighi

The rapid increase in the number of sequenced genomes has hindered efforts to functionally annotate proteins due to sequence redundancy. Our approach was to compute Representative Proteomes (RPs) by clustering microbial and eukaryotic proteomes into groups of related organisms so that curators can focus their efforts on a smaller set of representative organisms from each cluster. However, viral proteomes poses special challenges, as the taxonomy for viruses is not uniformly well defined. Grouping viruses based on the sequence similarity of their proteins in the proteomes can help normalize against potential taxonomic nomenclature biases. We present here an initial set of Viral Representative Proteomes (RPs) and associated tools. Viral Representative Proteomes are computed from UniProtKB virus complete proteomes. For each pair of proteomes, we calculate their co-membership in UniRef50 clusters. We then hierarchically cluster the similar proteomes into a set of Representative Proteome Groups (RPGs) based on their co-memberships at five different cutoff levels. The proteomes in each RPG are ranked using a proteome priority score to facilitate the selection of a top ranked proteome as the representative from the group. We also use taxonomic group and host information to annotate the viral proteomes in each RPG. Comparison of our RPs with UniProt's curator-selected Reference Proteomes (RefPs) indicates that RPs and RefPs are consistent. Furthermore, RPs can suggest many more candidate Reference Proteomes. A Viral RPs Browser presents the RPs grouped by taxonomic group or by host. From this browser, we can see majority of the viral RPs clustered the virus proteomes consistently with the taxonomic groups and the hosts. In conclusion, Viral RPs can be used to improve proteome annotation, protein classification, and taxonomic nomenclature bias detection in the viral proteome community. Viral RPs are available from PIR web site at: http://pir.georgetown.edu/rps/viruses/.

eMouseAtlas and Image Informatics

Lorna Richardson[1], Chris Armit[1], Julie Moss[1], Liz Graham[1], Nick Burton[1], Yiya Yang[1], Bill Hill[1], Richard A. Baldock[1]

[1] IGMM

Presenter: Lorna Richardson

A significant proportion of biomedical data is presented in the form of images, and the curation of this image-based data presents different challenges to that of sequence/text-based data. The eMouseAtlas project is an online resource for mouse developmental biology that archives, annotates and curates such image data. We describe here the underlying structure of the resource, as well as some of the tools that have been developed to allow users to mine the curated image data in a similar way to mining sequence-based data. The inherent complications in the curation and analysis of image-based data are only compounded by the increased availability of full 3D image data. We describe some of our efforts to overcome issues of visualisation and analysis with respect to 3D image data. Finally we propose some future developments to the eMouseAtlas resource to facilitate increased interaction with the user community.

Strategic manual curation for retrieval of highly complex data from mutagenesis experiments

Madhura R. Vipra[1]

[1] Biocuration, Athena Consulting, India

Presenter: Madhura R. Vipra

Biocuration is integral to biomedical data analysis and biocurators have to ensure that both the experimental and inferential results are annotated correctly. Data mining tools are rapidly replacing manual curation, which is argued to be expensive and time consuming specially with large volumes of published data. Automation, however, miss out on data that is spread through the body of the paper/supplements, hidden from plain sight or described in obscure terms. For complex knowledgebase building, manual curation is still the most reliable option. Presented here is one such effort that extracted complex data for building a database of phenotype impacts of induced mutations, typically at mRNA and protein levels. Discerning the multitude of overlapping phenotypic impacts was arduous, even in a paper which described few mutations. Through strategies were devised to capture diverse array of impacts such as mRNA stability, biomarker identification, drug sensitivity, amino acid modification, protein solubility, constitutive activation, catalytic activity etc. to name a few. These were stratified further for the levels at which they were expressed (e.g. gene, protein, cell etc.), in a given model system (e.g. yeast, mice, cell lines etc.), and for their directionality, increase/decrease, fold change, statistical values etc. Use of varied adjectives such as very, slightly, absolutely, significantly, almost, rarely etc. by authors was particularly challenging to assign to controlled vocabulary. In several cases phenotypic impact was obscure and needed domain intensive interpretation of the full text and even cross references to make the data conform to the database structure. This task has yielded a unique knowledgebase which was feasible only because it was done manually. Manual curation is still the start point for complex biocuration. We feel the need of time is to scale the manual annotation methods and practices to increasing data for high quality data extraction.

Poster Abstract #72

Database Commons: a web resource for cataloguing biological databases

Dong Zou[1], Chunlei Yu[1,2], Lina Ma[1], Lili Hao[1], Zhang Zhang[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China

Presenter: Chunlei Yu

With the explosive growth of biological data, there is an increasing number of biological databases that contain a huge variety of data and cover a diverse range of organisms. As this number continues to grow, it makes difficult and time-consuming for researchers to navigate such huge volume of databases, not to mention locating a complete range of biological databases that are not only related to a specific research field but also have higher popularity in terms of both data quantity and data quality. To address this difficulty, here we develop Database Commons (http://databasecommons.org), a web resource that provides researchers with easy access to a comprehensive collection of publicly available biological databases incorporating different data types and spanning diverse organisms. It integrates meta-information for all collected databases, including URL, description, related publication, update history, etc., and catalogues each database based on its data type, organism, curation level, etc. In addition, it allows users to comment on database utility by considering data quality, quantity, web interface, etc. Therefore, Database Commons features cataloguing databases under different criteria and incorporating peer comments on database utility, thus serving as a valuable resource for efficient exploitation of all publicly available databases.

Towards reproducible computational analyses: the ENCODE approach

Esther T. Chan[1], Venkat S. Malladi[1], Jean M. Davidson[1], Benjamin C. Hitz[1], Marcus Ho[1], Brian T. Lee[2], Nikhil R. Podduturi[1], Laurence D. Rowe[1], Cricket A. Sloan[1], J. Seth Strattan[1], Forrest Tanaka[1], Eurie L. Hong[1], J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA
[2] Center for Biomolecular Science and Engineering, University of California Santa Cru, USA

Presenter: Esther T. Chan

Use of high-throughput sequencing technologies is becoming routine in the lab and with it, a reliance on computational approaches for the management, analysis and interpretation of the generated “data deluge”. Scientists are now faced with a challenge to ensure that their analysis methods are communicated clearly and fully despite their complexity and that findings can be replicable. The ENCODE Consortium has produced >80Tb of data, measuring transcription, RNA- and DNA-protein interactions, DNA methylation and chromatin structure to interrogate genome function. This corpus serves as a reference for the scientific community, so that new results can be readily compared and analyzed against it if processed uniformly to minimize technical variation. To this end, the ENCODE Data Coordination Center (DCC) and Data Analysis Center (DAC), have taken a novel approach to defining standard analysis methodologies, expressing them in a machine and human readable format to allow for automation, reproducibility and transparency. First, the DAC specified uniform analysis pipelines for four major data types: ChIP-seq, RNA-seq, DNase-seq and whole-genome bisulfite sequencing to be run on ENCODE data. Second, the DCC abstracted the pipeline workflows into modular software components and identified key metadata about the analysis process and production of results, capturing details such as software versions, input and output files and file formats. Last, the DCC implemented the pipelines to automatically determine which data need to be processed and re-run, in the event of failure, according to the specifications provided by the captured metadata, such as the assay type, reference assembly for mapping, appropriate controls for normalization and QC requirements defined by the ENCODE standards. All released ENCODE raw and processed experimental data and metadata are publicly accessible through a web portal, https://www.encodeproject.org and programmatically, via a REST API.

Tracking data provenance to compare, reproduce, and interpret ENCODE results

Cricket A. Sloan[1], Esther T. Chan[1], Venkat S. Malladi[1], Jean M. Davidson[1], Benjamin C. Hitz[1], Nikhil R. Podduturi[1], Laurence D. Rowe[1], Forrest Tanaka[1], J. Seth Strattan[1], Marcus Ho[1], Brian T. Lee[2], Eurie L. Hong[1], W. James Kent[2], J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA
[2] Center for Biomolecular Science and Engineering, University of California Santa Cruz, USA

Presenter: Cricket A. Sloan

Data provenance is crucial to compare, reproduce and interpret experimental data. The task of tracking this information can be especially challenging in large-scale high-throughput sequencing projects like ENCODE (ENcyclopedia Of DNA Elements) which produces a variety of assays to investigate chromatin structure, RNA and DNA protein interactions, DNA methylation, transcription and histone modifications. The complex path that leads from the collection of liver tissue through to transcript quantification or to a filtered list of transcription factor binding site predictions is long and has many people and techniques involved along the way. The decisions made at each step can make results more or less comparable. For example, experiments performed on liver from similar donors are comparable, but liver tissue from the same exact dissection would provide even more powerful evidence. The tracking of the specific details of both the wet lab and data processing techniques can make the results more or less reproducible. Details about the software versions and parameters as well as the library preparation are required to repeat analysis. Documenting and displaying these experimental variations clearly can make the data more or less interpretable. It needs to be immediately obvious to the user how data files and biosamples are related to each other in order to make a clear assessment of the power of the results. At the ENCODE DCC (Data Coordination Center) we have created a rich data model to capture the provenance of experimental methods and computational results in both structured and unstructured ways. We are developing graphical displays to make these details immediately understandable. These metadata can be viewed at the ENCODE Portal (https://www.encodeproject.org/).

The war on disease: Homology curation at SGD to promote budding yeast as a model for eukaryotic biology

Stacia R. Engel[1], Maria C. Costanzo[1], Rob Nash[1], Edith Wong[1], Janos Demeter[1], J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA

Presenter: Stacia R. Engel

The foundation for much of our understanding of basic cellular biology has been learned from the budding yeast Saccharomyces cerevisiae. Studies with yeast have also provided powerful insights into human genetic diseases and the cellular pathways in which they are involved. We will present an update on new developments at the Saccharomyces Genome Database (SGD; http://www.yeastgenome.org/), the premier community resource for budding yeast. We are expanding the scope of SGD to include high quality manually curated information regarding functional complementation between yeast and human homologs. This new information is provided in meaningful ways allowing data mining and discovery by integrating these data into this encyclopedic online resource. In addition to introducing our presentation of these newly curated data we will highlight other new developments, such as written summaries about yeast genes and their mutant phenotypes, their human homolog disease associations, and presentation of the yeast/human ortholog set. We also associate sequence changes with variations in cellular phenotypes and protein function. SGD maintains these different datatypes, and distributes them to the scientific community via the web and file transfer. These expanded efforts are part of our continuing mission to educate students, enable bench researchers and facilitate scientific discovery. This work is supported by a grant from the NHGRI (U41 HG001315).

Fast large-scale text mining of biomedical literature on Tianhe-2 supercomputer

Chengkun Wu[1], Shao-Liang Peng[1]

[1] School of Computer Science, National University of Defense Technology, China

Presenter: Chengkun Wu

Information about genes and pathways involved in a disease is usually 'buried' in scientific literature, making it difficult to perform systematic studies for a comprehensive understanding. Text mining has provided opportunities to retrieve and extract most relevant information from literature, and thus might enable collecting and exploring relevant data to a certain disease systematically. However, biomedical literature is growing explosively, with over 20 million abstracts and 1 million full-texts available at the moment. It is therefore a great challenge to utilize the massive amount of information and knowledge embedded in those unstructured texts. Tianhe-2, the world's fastest supercomputer built by the National University of Defense Technology, has seen a number of successful applications in mid/long term weather forecasting, oil prospecting, seismology, astrophysics and so on. Now boosting biomedical studies is one of the emerging focuses. We deployed a large scale text mining pipeline for extracting molecular interactions from all MEDLINE abstracts and the PMC open access full-texts. Previous projects like the construction of BioContext and EVEX databases can take a long time (3 months for BioContext on a cluster on campus). Utilizing the enormous compute power provided by Tianhe-2, we managed to finish the processing within a considerable short time. We aim to build upon this and set up a biological big data platform that can either provide pre-processed information or tailored bio-computing solutions.

The role of the ENCODE Data Coordination Center

Jean M. Davidson[1], Esther T. Chan[1], Venkat S. Malladi[1], Benjamin C. Hitz[1], Marcus Ho[1], Brian T. Lee[2], Nikhil R. Podduturi[1], Laurence D. Rowe[1], Cricket A. Sloan[1], J. Seth Strattan[1], Forrest Tanaka[1], Eurie L. Hong[1], W. James Kent[2], J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA
[2] Center for Biomolecular Science and Engineering, University of California Santa Cruz, USA

Presenter: Jean M. Davidson

In recent years there has been an influx in the quantity of publicly available, large genomic datasets from individual labs and large consortia. Here we describe the efforts of the Data Coordination Center (DCC) in improving the accessibility of data available for the Encyclopedia of DNA Elements (ENCODE) project. ENCODE is a collaborative effort to generate a comprehensive catalog of functional elements in human and mouse genomes. The ENCODE database currently includes more than 40 experimental techniques in over 400 tissue types and cell lines to analyze DNA and RNA-binding proteins, transcription and chromatin structure. All experimental data and computational analyses of these data are submitted to the DCC for validation, tracking, storage, and distribution to the scientific community. To ensure that the data generated by the production labs and the analysis performed on these data are accurately represented, the ENCODE DCC works closely with members of the Consortium groups to capture structured metadata related to experimental conditions, data quality metrics, and analysis methods. These experiments can be accessed via the ENCODE portal (http://www.encodeproject.org). Portal users query the database by searching for specific metadata terms, such as “p53”, or “K562” or by utilizing faceted searching of the structured metadata. The portal also supports the visualization of certain data files by launching a Genome Browser track hub. Data files can be downloaded either directly from the experiment pages at the portal or via bulk download by programmatically accessing the ENCODE REST API. By providing direct data downloads based on flexible and powerful search capabilities that rely on highly organized metadata, the DCC strives to expand the access of ENCODE data to the scientific community.

iCLiKVAL: Interactive community resource for manual curation of all scientific literature through the power of crowdsourcing

Todd D. Taylor[1], Naveen Kumar[1]

[1] Laboratory for Integrated Bioinformatics, RIKEN Center for Integrative Medical Sciences, Japan

Presenter: Todd D. Taylor

There are over 24 million scientific literature citations in PubMed. Searching this vast resource does not always give satisfactory results due to many factors: missing abstracts, unavailability of full-article text, non-English articles, no keywords, etc. Ideally, every citation should include a complete set of terms that describe the original article so that searches, using natural language, return more relevant results; however, this requires countless hours of manual curation. Our objective is to make manual curation 'fun', social and self-correcting, thus enriching resources such as PubMed so that users can precisely find the desired information. We developed a web-based open-access crowdsourcing tool for manual curation of all types of scientific media. We encourage the use of ontologies and support them as auto-suggest terms, but we do not restrict users to these terms and allow them to add their own. Non-English annotation is also supported. Research communities, regardless of location or language, can use this tool to work together on the manual curation of any type of literature or media related to their projects. We constructed a cross-browser platform-independent application using the latest web technologies and a NoSQL database. Users perform searches to find articles of interest, mark them for review, review them immediately or add them to a review-request queue, load PDFs into the viewer, select annotations within the text, and add appropriate keywords and optional relationship terms. Users can make article-specific comments, edit and rate annotations, chat with others reviewing the same article, add annotations via Twitter, etc. The more annotations made, the richer the database will become and the more relevant the search results. We plan to work with the curation community to capture valuable information already generated; and, via REST API, make iCLiKVAL (http://iclikval.riken.jp/) annotations easily accessible to the entire research community.

iCLiKVAL API: a RESTful Hypermedia API for literature annotation and discovery

Naveen Kumar[1], Todd D. Taylor[1]

[1] Laboratory for Integrated Bioinformatics, RIKEN Center for Integrative Medical Sciences, Japan

Presenter: Naveen Kumar

The field of literature mining has matured as a valuable companion for scientists from hypothesis construction to discovery. Usually for biologists, literature mining means a keyword search in PubMed or similar literature databases. However, not all search results provide relevant information due to inadequate semantic value linked with the literature. There are a myriad of information embedded in published literatures, though it is often not discoverable, and over the course of time new keywords or biological terms emerge, which will not be found in older content. There are lots of efforts being made to automatically annotate experimental data that need biological knowledge through the integration of high-throughput data and scientific literature. The iCLiKVAL API is a machine-friendly hypermedia-driven RESTful API, which can be seamlessly integrated with these kinds of applications. The API accepts requests in JSON format and provides responses in hypermedia-compliant JSON format, a self-descriptive machine-readable and machine-understandable format. The API allows for creation of free-form annotation as “key-value” pairs with optional “relationship” between them and assigns semantic value to various types of literature. These key-value pairs are crowd-sourced by the scientific community using a cross-platform and cross-browser web interface. The core application stores data in a secure NoSQL database to take advantage of schema-less and evolving data structures. The documentation for the API can be found at http://iclikval.riken.jp/documentation. We would like to present the iCLiKVAL API to more users and to get their valuable feedback so that we can continue to improve this service to meet the demands of the research community.

Poster Abstract #80

Improving evidence-based gene prediction using RNA-seq data

Yinghao Cao[1], Yan Li[1], ChengZhi Liang[1]

[1] Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, China

Presenter: Yinghao Cao

RNA-seq not only accelerates the completion of genome sequencing projects but also provides insight into gene annotation projects. We have integrated RNA-seq data to improve gene prediction using Gramene pipeline, which was proved to be one of the most accurate evidence-based pipelines. The original Gramene pipeline uses expression evidence such as protein, full-length cDNA, EST from closely related species. However, the accuracy of Gramene pipeline highly depends on the availability of evidences. RNA-seq data, due to its high abundance for many species, provides a new source of expression evidence for gene annotation, especially for species specific genes. For testing purpose, we did a gene build using rice Nipponbare genome. The traditional evidence gives 51976 genes. After adding RNA-seq data, the gene number reached to more than 120k. While many of those genes are non-coding or transposon-related, domain analysis revealed that 74546 genes could be assigned with functions, from which 1888 go-terms and 112 KEGG pathways were classified. Compared with MSUv7.0 and IRGSPv1.0 gene sets, 43515 and 30512 genes have been found to be overlapped with MSUv7.0 and IRGSPv1.0, respectively. We will extend the analysis to several other plant genomes such as Arabidopsis, Soybean, Wheat and Maize.

HBV-DIAP: HBV genome sequence curation and integrated analysis platform

Guoqing Zhang[1], Yangfan Guo[1], Jia Jia[1]

[1] Shanghai Center for Bioinformation Technology, China

Presenter: Jia Jia

Abstract The high degree of sequence divergence of hepatitis B virus (HBV) has important impacts on disease prevention, detection and treatment. It is useful to develop an integrated in-silico platform to complement laboratory experiments for discovering novel mechanisms of HBV infection and drug resistance, improving clinical diagnosis and prognosis accuracy, and optimizing therapeutic protocols to combat the disease. A HBV data integrated analysis platform (HBV-DIAP) was developed by encompassing various type of HBV sequences and clinical information with a sequence curation and analysis pipeline. The latest version of HBV-DIAP comprises 100262 annotated sequences and 10191 clinical data. For all collected sequences, we employed a standardization numbering system and annotation process in the curation process. In total, we identified 9 genotypes and 23 serotypes from 3978 curated whole-genome sequences. All clinical information on HBV patients were collected based on a standardized protocol. The sequence analysis pipeline includes the most useful tools for the analyses of HBV genotyping/sub-genotypes, drug-resistance-associated mutations, and protein-functional-associated mutations. We also developed a dynamic surveillance to tracing the treatment outcome of anti-hbv drug including lamivudine, telbivudine, adefovir, entecavir, and tenofovir. As a comprehensive data management and analysis platform, HBV-DIAP may greatly facilitate disease diagnosis and drug resistant prediction, and to benefit the standardization of HBV clinical data collection and analysis.

Interpretation of Large Scale Biological Data Facilitated by Curated Causal Biological Network Models.

Justyna Szostak[1], Sam Ansari[1], Stephanie Boue[1], Marja Talikka[1], Juliane Fluck[2], Manuel Peitsch[1], Julia Hoeng[1]

[1] Biological System Research, PMI R&D, Switzerland
[2] Fraunhofer Institute for Algorithms and Scientific Computing, Germany

Presenter: Justyna Szostak

We have previously shown that a semi-automated knowledge extraction workflow, featuring a text mining pipeline as well as a curation interface, provides an efficient workflow for knowledge extraction and the building of causal biological network models. The network models describe biological processes from upstream events to downstream measureable and quantifiable nodes, i.e., the expression of genes (Big Data) measured with microarray experiments. We now demonstrate the relevance of one of these curated network models for the interpretation of complex molecular data. Our example includes an atherosclerotic plaque destabilization network model that was created and curated with the knowledge extraction workflow as well as multiple transcriptomics datasets in a cardio vascular disease context of aortic tissue. Once the network is overlaid with experimental data, reverse causal reasoning and network perturbation amplitude algorithms allow the quantification of the impact an exposure/treatment/disease has on the experimental system. We also compare and discuss the results against a more conventional approach for transcriptomics data interpretation based on gene sets enrichment analysis. Our results show that by using curated causal biological network models with large datasets and perturbation quantification algorithms, the interpretation becomes more efficient, objective, and interpretable, as biological processes are described more holistically and are based on causal relationships and relevant context information. In summary, the semi-automated knowledge extraction workflow facilitates the construction of causal biological network models describing disease specific and complementary biological processes that can be fully contextualized and used for the interpretation of large scale biological data. In the context of Big Data, this toolset allows the organization of data in a more meaningful and accessible way as well as the simplification to incorporate new data.

lncRScan-SVM: a support vector machine based tool for long non-coding RNA prediction

Lei Sun[1]

[1] Yangzhou University, China

Presenter: Lei Sun

Functional long non-coding RNAs (lncRNAs) have been bringing novel insight into current biological study. However, it is still not trivial to accurately distinguish the lncRNA transcripts (LNCTs) from the protein coding ones (PCTs) due to the similarities between lncRNAs and message RNAs (mRNAs). Although several computational methods have been developed to deal with the problem, we present a novel approach for predicting lncRNAs more accurately. Our method named lncRScan-SVM aims at classifying PCTs and LNCTs using support vector machine (SVM). The gold-standard datasets for lncRScan-SVM model training, lncRNA prediction and method comparison were constructed according to GENCODE gene annotations. By feeding features derived from gene structure, transcript sequence, potential codon sequence and conservation into the SVM framework, lncRScan-SVM outperforms other methods in classifying PCTs and LNCTs. In addition, the prediction model can be re-trained by users according to the species they focus on. The lncRScan-SVM package is freely available on http://sourceforge.net/projects/lncrscansvm/?source=directory.
LncRScan-SVM is an efficient tool for predicting lncRNAs, and it is quite useful for current lncRNA study.

DoGSD: the dog and wolf genome SNP database

Bing Bai[1,2], Wenming Zhao[3], Bixia Tang[3], Yanqing Wang[3], Junwei Zhu[3], Guodong Wang[2], Yaping Zhang[2]

[1] Laboratory for Conservation and Utilization of Bioresource & Key Laboratory for Microbial Resources of the Ministry of Education, Yunnan University, China
[2] State Key Laboratory of Genetic Resources and Evolution, and Yunnan Laboratory of Molecular Biology of Domestic Animals, Kunming Institute of Zoology, Chinese Academy of Sciences, China
[3] Core Genomic Facility, Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Wenming Zhao

The rapid advancement of next-generation sequencing technology has generated a deluge of genomic data from domesticated dogs and their wild ancestor, grey wolves, which have simultaneously broadened our understanding of domestication and diseases that are shared by humans and dogs. To address the scarcity of single nucleotide polymorphism (SNP) data provided by authorized databases and to make SNP data more easily/friendly usable and available, we propose DoGSD (http://dogsd.big.ac.cn), the first canidae-specific database which focuses on whole genome SNP data from domesticated dogs and grey wolves. The DoGSD is a web-based, open-access resource comprising ~19 million high-quality whole-genome SNPs. In addition to the dbSNP data set (build 139), DoGSD incorporates a comprehensive collection of SNPs from two newly sequenced samples (1 wolf and 1 dog) and collected SNPs from three latest dog/wolf genetic studies (7 wolves and 68 dogs), which were taken together for analysis with the population genetic statistics, Fst. In addition, DoGSD integrates some closely related information including SNP annotation, summary lists of SNPs located in genes, synonymous and non-synonymous SNPs, sampling location and breed information. All these features make DoGSD a useful resource for in-depth analysis in dog-/wolf-related studies.

Text Mining Tools for Biocuration in the iProLINK Web Portal

Cecilia N. Arighi[1,2,3], Ruoyao Ding[3], Samir Gupta[3], Gang Li[3], A.S.M. Ashique Mahmood[3], Yifan Peng[3], Jia Ren[1], Karen E. Ross[1], Catalina O. Tudor[1], Jian Zhang[2], Hongzhan Huang[1,2,3], Carl Schmidt[4], Cathy H. Wu[1,2,3], K Vijay-Shanker[3]

[1] Center for Bioinformatics and Computational Biology, University of Delaware, USA
[2] Protein Information Resource
[3] Department of Computer and Information Sciences, University of Delaware, USA
[4] Department of Food and Animal Sciences, University of Delaware, USA

Presenter: Cecilia N. Arighi

Text mining tools can reduce the labor of manual literature curation by efficiently extracting information from the rapidly increasing biomedical literature. Literature-based biocuration workflows include common steps such as: document triage (finding relevant documents), entity recognition and normalization (finding protein names and mapping them to database IDs), and event extraction (finding relations among various entities). We have developed a suite of tools addressing such biocuration tasks. eGIFT identifies informative terms (iTerms) and documents that are relevant to a gene/protein. The iTerms are ranked, linked to evidence sentences, and placed into term categories for quick review. pGenN is a gene normalization tool tailored for plants, linking gene/protein names to Entrez Gene and UniProt identifiers. RLIMS-P identifies articles relevant to protein phosphorylation, automatically extracting information on protein kinases, substrates and phosphorylation sites, with protein links to UniProt. eFIP identifies phosphorylated proteins and phosphorylation-dependent protein-protein interactions. It supports searches based on protein roles (kinase, substrate, or interacting partner) or keywords, links protein entities to UniProt identifiers and supports visual exploration of phosphorylation interaction networks using Cytoscape. Finally, miRTex extracts miRNA-target relations as well as miRNA-gene and gene-miRNA regulation relations. These tools have been evaluated with annotated literature corpora, achieving similar or superior performance to comparable state-of-the-art systems. Each tool is hosted from a website for searching, linking to external databases, and downloading of results (in CSV or BioC formats); some also offer a web service. RLIMS-P and eFIP further provide results from full-text articles in the PMC Open Access subset. All tools are available from the PIR text mining portal iProLINK at: http://www.proteininformationresource.org/iprolink/.

Semantic data integration in CyanoBase and RhizoBase

Takatomo Fujisawa[1], Toshiaki Katayama[2], Mitsuteru Nakao[2], Shinobu Okamoto[2], Yasukazu Nakamura[1]

[1] National Institute of Genetics, Japan
[2] Database Center for Life Science, Japan

Presenter: Takatomo Fujisawa

We have been maintaining and expanding CyanoBase (http://genome.microbedb.jp/cyanobase), a genome database for cyanobacteria, and RhizoBase (http://genome.microbedb.jp/rhizobase), a genome database for rhizobia, nitrogen-fixing bacteria associated with leguminous plants since 1996. CyanoBase and RhizoBase have been accumulating reference genome annotations, which are continuously updated by manual curation. To enable this effort, we have been collaborating with domain experts who extracted names, products and functions of each gene reported in the literature. To ensure effectiveness of this procedure, we have developed and managed the TogoAnnotation (http://togo.annotation.jp) system offering a web-based user interface and a uniform storage of annotations for the curators of the CyanoBase and RhizoBase databases [1]. In these days, Semantic Web technologies have been deployed for integration of heterogeneous data in life sciences. A generic ontology [2] for semantically describing genomic annotations was developed by the DDBJ and the Database Center for Life Science (DBCLS). For improving interoperability, we have introduced this ontology for representing annotations stored in the CyanoBase and RhizoBase databases to export those database conetnts in the RDF (resource description framework) format. The result is accessible from our SPARQL (SPARQL Protocol and RDF Query Language) endpoint at http://genome.microbedb.jp/sparql. In this presentation, we will also describe the progress on database presented and our efforts to reduce data management costs.

References

1. Fujisawa T. et. al. (2014) CyanoBase and RhizoBase: databases of manually curated annotations for cyanobacterial and rhizobial genomes. Nucleic Acids Res., 42(1): D666-670. doi: 10.1093/nar/gkt1145.
2. Kodama Y. et. al. (2015) The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. Nucl. Acids Res. 43(1): D18-22 doi:10.1093/nar/gku1120.

Text mining and curation system for enzymatic and metabolism reactions: the TeBactEn tool.

Martin Krallinger[1], Andres Cañada[2], Alfonso Valencia[1]

[1] Structural Computational Biology Group, Spanish National Cancer Research Centre (CNIO), Spain
[2] National Bioinformatic Institute Unit, Spanish National Cancer Research Centre (CNIO), Spain

Presenter: Martin Krallinger

TeBactEn is a tool designed to facilitate the retrieval, extraction and annotation of bacterial enzymatic reactions and pathways from the literature. The system contains three different data collections, namely (a) a compilation of articles derived from the Microme database, i.e. articles (abstracts and full text articles) that had been used for manual annotation of bacterial pathways, (b) a set that covers abstracts from the entire PubMed database that are relevant to bacteria and finally (c) a collection of abstracts and full text articles that are relevant for a list of bacteria of special interest to metabolic reactions, facilitating a more exhaustive extraction of enzymes particularly for these bacteria. In case of all three TeBactEn data collections, an exhaustive recognition of mentions of all species and taxonomic entities was carried out. TeBactEn covers all the main steps relevant for the automatic extraction and ranking of metabolism relations from the literature and allows enhanced access and annotation of related information: 1.Identification of metabolism relevant articles. 2. Detection of the bio-entities involved in biochemical reactions: enzyme, compounds and organisms. 3. Extraction of weighted (ranked) relationships between these bio-entities. 4. An interface to browse this information and to construct a manually curated database of metabolism reactions. 5. Facilitate quick manual literature curation. 6. The option to normalize/ground bio-entity mentions to other knowledgebases like UniProt and ChEBI. The system is available at: http://tebacten.bioinfo.cnio.es

Poster Abstract #88

Wikidata: a central hub of linked open life science data

Andra Waagmeester[1,2], Elvira Mitraka[3], Benjamin Good[1], Sebastian Burgstaller-Muehlbacher[1], Paul Pavlidis[4], Gang Fu[5], Lynn M. Schriml[3], Andrew I. Su[1]

[1] Department of Molecular and Experimental Medicine, The Scripps Research Institute, USA
[2] Micelio, Belgium
[3] University of Maryland School of Medicine, USA
[4] Centre for High-Throughput Biology and Department of Psychiatry, University of British Columbia, Canada
[5] National Center for Biotechnology Information, USA

Presenter: Elvira Mitraka

Data in the life sciences are abundant, but dispersed over many different resources. However, for the onset of research these different resources need to be integrated. Although the Semantic Web has been proposed as a potential solution for rapid knowledge integration, most data remains in their different data silos, which expand continually, worsening the knowledge integration challenge. In the last decade, Wikipedia has been successful in becoming one of the most important sources of information on the web. Wikipedia thrives on the community for its curation. One of the partner projects currently is Wikidata, which is a public and free linked database using the same principles of community curation. Here we report on our effort to make Wikidata a central hub for linked open life science data. Doing so not only provides a linked data platform of said data, but also opens up the potential of the Wikidata community at large for curating and putting the different data sources under scrutiny. Our game plan is to (1) develop bots to publish knowledge from established data sources on genes, diseases and drugs on Wikidata, (2) harvest links between these entities and enrich the respective Wikidata entities with these relations, and (3) engage the community in curating the knowledge at hand by developing applications to disseminate the content to a wider audience. Here, we report the first milestones, being Wikidata entries on all human genes from Entrez Gene and the diseases from the Disease Ontology. Within days upon first publication of these entries, the curation power of Wikidata became visible by some valuable improvements made by the community. Our next goals are too add gene-disease, drug-disease and gene-drug relationships.

Poster Abstract #89

WEiGEM+, a mobile SNS platform for iGEM and Synthetic Biology

Hao-Ying Huang[1], Xing-Da Zheng[1], Yi-Jia Zhou[1], Nan Yao[1], Silei Zhu[1], Jun-Zhi WEN *[1]

[1] Bioinformatics Center, State Key Laboratory of BioControl, China

Presenter: Hao-Ying Huang

During the last decade, the mobile technology has developed a lot. An increasing number of people are joining the 'information age' via the mobile Internet, but websites designed for PCs are not always convenient to view with mobile devices. A better interface for Mobile Internet is needed. WEiGEM+, a special official account based on the popular SNS app WECHAT, has various applications. As an emerging discipline, synthetic biology is developing rapidly and it has been proved to be valuable in different areas. However, growing data makes it challenging for synthetic biologists to process data and find out what is needed, especially biobricks, registers of functional units that play an important role in synthetic biology. The number of biobricks is now over 25000 and keeps growing fast, which makes it hard to find suitable biobricks. So the first and foremost, WEiGEM+ is a search engine for biobricks, users can enter keywords to get information of a biobrick. And there is one more point, WEiGEM+ provides a classified catalogue of biobricks with different functions, thus users can look for the biobrick they desired in various ways. Another function of WEiGEM+ is that users can search as soon as a new idea emerge, and due to the comfortable chat history, remembering what they are thinking about when they make the search is not a hard work. Because we sometimes lose ideas which suddenly appear in our mind when we are busy, it can be helpful. The last but not the least, followers can keep up to date with latest correlative knowledge pushed by WEiGEM+. Also, due to the popularity of WeChat, push helps to introduce synthetic biology to the public. In addition, when users send a message in natural language, they will get replies, just like communicating with a chatterbot, from my own perspective it will improve the user experience. (Note: This is a MIT iGEM Best Software Tool project, which have more than 10 high school students interested in Biocuration within it.）

Poster Abstract #90

Proteome redundancy in UniProtKB: challenges and solutions

Ramona Britto[1], Claire O'Donovan[1], UniProt Consortium[1,2,3]

[1] EMBL-European Bioinformatics Institute, UK
[2] Swiss Institute of Bioinformatics, Switzerland
[3] Protein Information Resource, USA

Presenter: Ramona Britto

The UniProt Knowledgebase (UniProtKB) is a comprehensive resource for protein sequences and their annotation. It consists of SwissProt, the manually annotated section and TrEMBL, which offers translations of nucleotide sequences submitted to the INSDC supplemented with automatic annotation. As a result of the vast increase in genome sequencing projects in the last few years UniProtKB has doubled in size in the last year to nearly 90 million entries with a high level of redundancy. This is especially true for bacterial species where different strains of the same bacterium have been sequenced and submitted (e.g. 1,692 strains of Mycobacterium tuberculosis, 4,080 strains of Staphylococcus aureus). Redundancy leads to slow database searches, higher computational costs and an increasing bias in statistical analyses. Removing redundant sequences is also desirable to avoid highly repetitive search results for user queries that closely match an over-represented sequence. To deal with this issue, firstly, we identified highly redundant proteomes by performing pairwise comparisons of proteomes within species groups. Proteomes were then ranked based on several criteria including the level of manually curated information. A directed weighted graph was constructed and the lowest ranking proteomes in each group were eliminated in an iterative process until a minimal set was obtained. This procedure has been implemented for bacterial species and the sequences corresponding to redundant proteomes (approximately 47 million entries) were deprecated. All proteomes remain searchable through UniProt's Proteomes interface (http://www.uniprot.org/proteomes/) and redundant proteome sets are now available for download from the UniProt Archive (UniParc); users are also directed to non-redundant proteome(s) available for the same species. We will present the development methodology and discuss the impact and advantages of this approach for our users and the services that rely on this data.

IC4R: Information Commons for Rice

Lili Hao[1], Dawei Huang[1], Li Yang[1,2], Lin Xia[1,2], Dong Zou[1], Xingjian Xu[1,2], Songnian Hu[1], Zhang Zhang[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China

Presenter: Lili Hao

Rice is not only the key model organism for plant research community but also the main source of food for a large part of the world's population. Here we present Information Commons for Rice (IC4R; http://ic4r.org), a curated knowledgebase for the rice research community. IC4R integrates large volumes of omics data and large quantities of rice-related literature, with the aim to provide a reference genome for rice with standard and comprehensive annotations. IC4R utilizes a Web Service-based architecture, which integrates data from multiple IC4R committed projects through Web Application Programming Interfaces, thus bearing the potential to ease data integration, reusability, scalability, integrity, and community intelligence for collaborative knowledge curation. Currently, IC4R committed projects include RiceWiki (a wiki-based, publicly editable and open-content platform for community curation of rice genes), Rice Literature Miner (a comprehensive collection of rice-related publications with association with specific rice genes), Rice Expression Database (a complete integration of expression profiles derived entirely from RNA-Seq data), Rice Homology Database (an accurate summary of rice-centered homologous information based on several important plants), and Rice Variation Database (a comprehensive collection of rice genomic variations). Future directions include incorporation of other types of data (e.g., non-coding RNA, methylation) and association of agronomically important traits with omics data, aiming to build IC4R into a rice knowledgebase for not only molecular studies but also for rice production and improvement.

Poster Abstract #92

RegulonDB, a tool to decipher the regulation of bacterial complexity.

Julio Collado-Vides[1], Alberto Santos-Zavaleta[1], Socorro Gama-Castro[1], Mishael Sánchez-Pérez[1], Heladia Salgado[1], Hilda Solano-Lira[1], César Bonavides-Martínez[1], Luis José Muñiz-Rascado[1], Fabio Rinaldi[2]

[1] Genómica Computacional, Centro de Ciencias Genómicas, México
[2] Institute of Computational Linguistics, Switzerland

Presenter: Julio Collado-Vides

RegulonDB has historically been the most significant database on transcriptional regulation of the best-characterized organism, Escherichia coli. However, the knowledge of gene regulation is a small part of what represents the physiological universe of this bacterium. One strategy to solve or complete some pieces of the puzzle has been to curate elements beyond transcriptional regulation, such as riboswitches, attenuators, growth conditions, gensor units (GUs) and elements regulating transcription that do not bind to DNA, such as ppGpp and DksA. We have also included in our curation data from massive expression experiments, such as microarray analyses, in addition to computational predictions generated by our team, thus integrating our database into the "omics" era. Keeping this information up to date has involved a great effort from our curation team. As a result, we have initiated work to modify our manual curation strategies, taking advantage of the benefits of natural language-processing implementations to enhance our performance in expanding and enriching the quality and quantity of curated knowledge. Using the particular filters of the ODIN system, we have conducted an assisted biocuration for the growth conditions of the regulatory interactions of OxyR, SoxS, and SoxR proteins, as an initial test experiment. On the other hand, we have initiated an automatic process to construct the GUs of all transcription factors. In order to improve our initial searches for new articles, which we systematically perform to keep RegulonDB up to date with the literature, we have recently initiated an analysis of all curated papers by using the Knowledge Discovery in Databases (KDD) methodologies. This work will be connected with a parallel effort, contributing to the ontology of the regulation of gene expression in microbial genomes. Email contact: curatorsRegulon@ccg.unam.mx

Development of a collection of life stage ontologies

Anne Niknejad[1,2], Marc Robinson-Rechavi[1,2], Frederic B. Bastian[1,2]

[1] SIB Swiss Institute of Bioinformatics, Switzerland
[2] Department of Ecology and Evolution, University of Lausanne, Switzerland

Presenter: Anne Niknejad

In order to be able to study changes in phenotypes over developmental time, e.g., aging processes, it is essential to be able to accurately capture information related to life cycle stages, whether in human for biomedical applications, or in model and non-model organisms, to allow multi-species integrations. In the context of the Bgee project (see http://bgee.org/), we have developed life stage ontologies allowing to annotate expression data in 17 species. The strategy adopted is to use high-level terms to consistently organize the ontologies, then to create new terms as they are needed for annotations. For each species, the ontologies capture essential information such as duration of life stages, or state of sexual maturity. The high-level terms used are mapped to the multi-species ontology Uberon (see http://uberon.org/), allowing their automatic integration and mapping. The ontologies developed are hosted in a repository that also links to already-established resources, see https://code.google.com/p/developmental-stage-ontologies/.

BioCreative V a new challenge in text mining for biocuration

Cecilia N. Arighi[1,2], Andrew Chatr-aryamontri[3], Cohen B. Kevin[4], Donald Comeau[5], Juliane Fluck[6], Rezarta Islamaj Dogan[5], Lynette Hirschman[7], Sun Kim[5], Martin Krallinger[8], Florian Leitner[9], Zhiyong Lu[5], Julen Oyarzabal[10], Obdulia Rabal[10], Fabio Rinaldi[11], Catalina O. Tudor[1], Alfonso Valencia[8], Thomas Wiegers[12], John Wilbur[5], Cathy H. Wu[1,2]

[1] Center for Bioinformatics and Computational Biology, and Department of Computer and Information Sciences, University of Delaware, USA; [2] Protein Information Resource, USA; [3] BioGrid, Canada; [4] University of Colorado, USA; [5] NCBI, National Institutes of Health, USA; [6] Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Germany; [7] MITRE, USA; [8] CNIO, Spanish National Cancer Centre, Spain; [9] Universidad Politecnica de Madrid, Spain; [10] Center for Applied Medical Research (CIMA), University of Navarra, Spain; [11] Institute of Computational Linguistics, University of Zurich, Switzerland; [12] North Carolina State University, USA

Presenter: Cecilia N. Arighi

BioCreative: Critical Assessment of Information Extraction in Biology is a community-wide effort for evaluating text mining and information extraction systems applied to the biological domain. For the past ten years BioCreative challenges have spanned a number of tasks particularly important to biocuration. BioCreative V is underway and consists of five tracks: 1-Collaborative Biocurator Assistant Task for development of BioC-compatible modules which complement each other for an integrated system that assists BioGRID curators. This will be a non-competitive, cooperative task in which participants work together to build a better system; 2-CHEMDNER patents to address the automatic extraction of chemical and biological data from medicinal chemistry patents. Currently, the identification and integration of all information contained in these patents is an extremely hard task not only for database curators, but also for life sciences researches and biomedical text mining experts; 3-Chemical-disease relation task for automatic detection of mechanistic and biomarker chemical-disease relations from the biomedical literature, in support of biocuration, new drug discovery and drug safety surveillance; 4-Extraction of causal network information in Biological Expression Language for text mining solutions to develop and test novel approaches for relation extraction in the context of pathway networks. The goal is to assess the utility of such tools for the automated annotation and network expansion, and their suitability as supporting tools for assisted curation; 5-Interactive Curation for the demonstration and evaluation of web-based systems addressing user-defined tasks, evaluated by curators on performance and usability. BioCreative tracks are aimed to advance text-mining research and provide practical benefits to biocuration, therefore many tracks rely on curator's participation. We will describe past and current efforts. Information is available at www.biocreative.org.

Poster Abstract #95

Curation and classification of inherited disease variants in a high-throughput clinical-grade genetic screening laboratory environment

Kambiz Karimi[1], Peter Kang[2], Imran Haque[3], Eric A. Evans[4]

[1] Curation Team Lead, Counsyl Inc., 94080
[2] Laboratory Director, Counsyl Inc., 94080
[3] Research Director, Counsyl Inc., 94080
[4] CSO, Co-founder, Counsyl Inc., 94080

Presenter: Kambiz Karimi

Counsyl is a technology-driven medical laboratory that offers clinically comprehensive, affordable, and high quality genetic screening and genetic counseling services. Our physician-prescribed genetic screens provide vital information on a panel of >100 rare autosomal recessive diseases, as well as determining risk for inherited breast, ovarian and prostate cancer. Counsyl has a robust system of variant interpretation and classification utilizing customized software analysis that gathers information from multiple sources including: patient data (case reports, and patient databases), population data, molecular functional data, mutational co-occurrence, protein structural analysis, conservation, in-silico predictors, and internal data. All variants are reviewed by a team of genetic counselors and PhD-level scientists, as well as our laboratory directors. Manual curation of case reports and molecular functional data plays an important role in variant classification at Counsyl. To enable correlations between variants and disease phenotype, the following details are curated from case reports for each variant: Patient inclusion/exclusion criteria, variant allele frequency among unrelated patients and controls, patient zygosity/genotype, patient diagnosis/phenotype, ethnicity, gender, age of disease onset, and disease severity. Similarly, the following parameters are recorded from molecular functional studies for each variant: Experimental system, effect of variant on protein/mRNA expression, splicing, protein activity, intracellular localization, folding/processing and post-translational modifications. To date, over 4100 variants have been manually curated and classified according to ACMG guidelines. All variant interpretations are periodically submitted to public databases (E.g. ClinVar) to enable peer review of the classifications and to allow for updates to classifications when there is new evidence for variant pathogenicity or neutrality.

Collection and curation of whole genome studies of budding yeast at the Saccharomyces Genome Database (SGD)

Rama Balakrishnan[1], Edith Wong[1], Janos Demeter[1], Shuai Weng[1], J. Michael Cherry[1]

[1] Department of Genetics, Stanford University, USA

Presenter: Rama Balakrishnan

The Saccharomyces Genome Database (SGD; www.yeastgenome.org) is a comprehensive resource of curated molecular and genetic information on the genes and proteins of Saccharomyces cerevisiae. The emergence of large-scale, genome-wide technologies such as expression microarrays, RNA-seq, and high-throughput sequencing have widened the scope of functional annotation beyond that of individual genes to entire genomes. These new data allow us to identify shared and divergent features between genomes. We have collected published data from whole genome studies that employ a diverse set of modern techniques, including tiling arrays, cDNA clone libraries, TIF-seq, single and paired end RNA-seq, and serial analysis of gene expression (SAGE). These divergent methodologies target different genomic regions, such as ncRNA, transcription start sites (TSS), transcripts, poly-A sites, and antisense RNA. Metadata were curated from more than 1000 publications using datasets from NCBI's GEO repository (Gene Expression Omnibus). These data will be available for easy, straightforward querying at SGD via a faceted/aggregated search tool, which will facilitate user access to yeast genomic data.

From High-throughout Analysis, Comprehensive Analysis of downstream genes of Human chromatin remodeling INO80 complex

Liguo Dong[1]

[1] JiLin University, China

Presenter: Liguo Dong

Human INO80 complex is a conservative ATP-dependent chromatin remodeling complex,belong to the SWI,SNF family1.hINO80 complex play important roles in variety of biological processes2.To analyze the function of INO80 in gene regulation,we knocked down five conservative subunits1,4 compared with other species of human INO80 complex.Then we respectively applied statistical method and GO annotation without statistical analysis to explore the core gene clusters that the complex regulates.Besides we got twenty high different expression genes that hINO80 complex may regulate directly.For future understanding the regulated genes of hINO80 complex,we integrated the protein interaction with hINO80.From comprehensive understanding of INO80 downstream genes,we can clearly illustrate the regulation network of Human INO80 complex.

A large-scale identification of prokaryotic replication origins and its applications

Feng Gao[1,2,3]

[1] Department of Physics, Centre of Bioinformatics, Tianjin University, Tianjin, China
[2] Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China
[3] SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), China

Presenter: Feng Gao

Replication of chromosomes is one of the central events in the cell cycle. The core machineries that initiate DNA replication are conserved in all three domains of life: bacteria, archaea, and eukaryotes. Clarification of the archaeal replication mechanism is particularly important, as it may reveal the evolutionary history between bacteria and eukaryotes. Prokaryotic DNA replication begins at specific sites, called replication origins (oriCs). A large-scale identification of prokaryotic replication origins in silico will speed up the experimental confirmation and facilitate the functional analysis by comparative genomics approaches. The availability of increasing complete prokaryotic genomes has created challenges and opportunities for identification of their replication origins on a large scale. Based on the Z-curve method, with the means of comparative genomics, a web-based system, Ori-Finder (http://tubic.tju.edu.cn/Ori-Finder/), has been developed to find oriCs in prokaryotic genomes with high accuracy and reliability. The oriC regions identified by in silico analyses, as well as in vivo experiments have been organized into DoriC (http://tubic.tju.edu.cn/doric/), a database of oriC regions in bacterial and archaeal genomes, which contributes to revealing the regulatory mechanisms of the initiation step in DNA replication and understanding the molecular mechanisms of strand-specific biases. For example, it has been widely used in the comparative genomics analysis of oriC regions, and the strand-specific biases analysis of nucleotide composition, codon usage, and protein-coding or essential genes distributions. The application of the rules from the database will also be helpful to develop new prediction algorithms of replication origins and speedup the experimental confirmation and functional analysis of oriCs in bacterial or archaeal genomes. In addition, it has important significance in practical use, such as in synthetic biology.

Building a Unified Mouse Gene Catalog

Yunxia S. Zhu[1], Joel Richardson[1], Carol J. Bult[1]

[1] Mouse Genome Informatics, The Jackson Laboratory, USA

Presenter: Yunxia S. Zhu

We report here a semi-automated process by which the mouse genome feature predictions (e.g., genes, pseudogenes, functional RNAs, etc.) from Ensembl, NCBI and Havana are reconciled with genes represented in Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org) into a single comprehensive and non-redundant catalog. The MGI unified gene catalog serves as the foundation for the annotation of biological attributes of genome features (i.e., phenotype, function, expression, and pathway membership) by MGI curators and bio-data analysts. Our gene unification method employs an algorithm called fjoin for efficient comparisons of overlaps among genome features based on their genome coordinates. Following overlap detection, features are binned into six possible categories based on the relationships of features to one another. The categories are then prioritized for targeted assessment of annotation anomalies. Currently there are 59062 genome features in MGI genome feature catalog: 22599 protein-coding coding genes, 12455 pseudogenes, and 24007 other types of genome features (e.g., lincRNA genes, miRNAs, snoRNAs, etc.). The majority of genome features (55967) in MGI are associated with predictions from at least one of the major annotation providers. Less than half (23174) of the genome features have equivalent predictions in all three providers. Many of the genome features in MGI's catalog are unique to only one annotation provider. Researchers can quickly obtain a comprehensive list of mouse genome features from MGI and visualize the gene and transcript details using MGI's mouse genome browser (http://jbrowse.informatics.jax.org). The catalog is updated on a regular basis when new versions of annotations are released. In addition to the three major genome annotation providers (NCBI, Ensembl, Havana), we also integrate genome features from specialized databases (e.g., mirBase, GtRNAdb, etc.). MGI is supported by NIH grants HG000330-P1.

A system for pathological comparison between human and animal models

Shuang Yang[1], Hong Sun[2], Guoqing Zhang[2]

[1] College of Life Science and Biotechnology, Shanghai Jiaotong University, China
[2] Shanghai Center for Bioinformation Technology, China

Presenter: Hong Sun

Background Most of the data on disease mechanism and drug effects generated from animal models are not transferable to human, which calls into question the use of better animal models in the study of human diseases and the development of clinical treatments. In addition, the different phenotypes due to genetic variation create a challenge for the selection or development of appropriate animal strains. Currently existing systems are rarely designed for the similarity search in the context of human pathology, so that scientists need to link diverse datasets through searching many detached resources to compare molecular pathology between human patients and animal models. Results To address these issues, we developed a system for scientists to investigate disease associated mammalian species specific information (SysDAMS). We imported compendium of human genes and genetic phenotypes from OMIM, repository of detailed drug data with comprehensive drug target and drug action information from DrugBank, side effect resource from SIDER, annotations of gene products from the Gene Ontology, and collection of manually drawn pathways representing the molecular interaction and reaction networks from KEGG and PID database. To investigate the relationships of molecular pathology between human and animal models, orthologs, paralogs and several other features were generated by computational methods and were deposited in the system. The values of betweenness were calculated to measure the relative importance of each node in a KEGG network. Conclusions SysDAMS has the power in the pathological comparison between human and animal models, and it's a quite useful platform to help researchers select better animal models or mouse strains for conditional investigation.

3CDB: A 3C Database

Xiaoxiao Yun[1], Lili Xia[1], Bixia Tang[1], Hui Zhang[1], Zhihua Zhang[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China

Presenter: Zhihua Zhang

It has been more than 10 years since the invention of chromosome conformation capture (3C) technology. With the help of next generation sequencing technology, 3C and it genomic variants have enabled the study of nuclear organization reached an unprecedented level. However, all its high-throughput variants are facing inherent issues, such as low resolution, experimental bias and cross-linking problem, which in turn limits its application. In this paper, we describe a manually mined and refined database containing all chromatin-chromatin interactions revealed by traditional 3C assay from current literature. Instead of using literature mining tools, we manually searched literature with carefully designed key word combinations. The 3C signal intensities at the interaction locus, as well as the signals at the flanking anchor sites, were retrieved together with their associated experimental settings. In total, about 1700 interactions passed our quality filter in 15 species. As the 3C technology is continuing evolve, to better assess the reliability of each interactions, which may be detected by a slightly modified 3C protocols, we developed a systematic evaluation scheme. According to the scheme, each interaction was classified into one of the four categories. All scoring details in the scheme is accessible to customs. A simple web-based visualization tool was also integrated into the database. We believe the 3C database (3CDB) can bridge the researchers who are working on particular genome locus to the systems biologists who are suffering with the noisy high-throughput data, and therefore became a valuable resource to serve the whole community.

Poster Abstract #102

NONCODE, a noncoding RNA database with emphasis on long noncoding RNAs

Chaoyong Xie[1], Jiao Yuan[2], Runsheng Chen[2], Yi Zhao[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, China
[2] Institute of Biophysics, Chinese Acadamy of Sciences, China

Presenter: Jiao Yuan

Non-coding RNAs (ncRNAs) have been implied in diseases and identified to play important roles in various biological processes. NONCODE (http://www.bioinfo.org/noncode/) is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs). By collection of newly identified ncRNAs from published literature and integration of the latest version of RefSeq and Ensembl, NONCODE provides a comprehensive ncRNA data set, in which there are 210,831 lncRNAs. To help systematic understanding of lncRNAs, which show similar alternative splicing pattern to mRNAs, we put forward the concept of lncRNA genes, and generated 56,018 and 46,475 lncRNA genes from 95,135 and 67,628 lncRNA transcripts for human and mouse, respectively. Additionally, based on public RNA-seq data, we presented expression profiles of lncRNA genes by graphs, as well as predicted functions of these lncRNA genes. The convenience of the database also includes an incorporation of an ID conversion tool from RefSeq or Ensembl ID to NONCODE ID and a service of lncRNA identification.

NPInter, a database of noncoding RNA related interactions

Jiao Yuan[1], Wei Wu[1], Yi Zhao[2], Runsheng Chen[1]

[1] Institute of Biophysics, Chinese Academy of Sciences, China
[2] Institute of Computing Technology, Chinese Academy of Sciences, China

Presenter: Jiao Yuan

NPInter (http://www.bioinfo.org/NPInter) is a database that integrates experimentally verified functional interactions between non-coding RNAs (excluding tRNAs and rRNAs) and other biomolecules (proteins, RNAs and genomic DNAs). With the development of high-throughput biotechnology, such as crosslinking immunoprecipitation and high-throughput sequencing (CLIP-seq), the number of known ncRNA interactions, especially those formed by protein binding, has grown rapidly in recent years. By collecting ncRNA interactions from literature and related databases, NPInter currently covers 201,107 entries from 18 species. In addition, NPInter incorporated a service for the BLAST alignment search as well as visualization of interactions.

NaviCell Web Service for network-based data visualization and analysis

Inna Kuperstein[1], Eric Bonnet[1], Eric Viara[1], Laurence Calzone[1], David Cohen[1], Emmanuel Barillot[1], Andrei Zinovyev[1]

[1] Bioinformatics and Computational Systems Biology of Cancer, Institut Curie – U900 INSERM - Mines ParisTech, France

Presenter: Inna Kuperstein

Data visualization is an essential element of biological research, required for obtaining insights and formulating new hypotheses on mechanisms of health and disease. NaviCell Web Service is a tool for network-based visualization of “omics” data which implements several data visual representation methods and tools for combining them together. NaviCell Web Service uses Google Maps and semantic zooming to browse large biological network maps, represented in various formats, together with different types of the molecular data mapped on top of them. The input data for NaviCell Web Service are various omics data as mRNA, microRNA or proteins expression, mutation landscapes, copy-number genomic profiles. NaviCell Web Service is also suitable for computing aggregate values for sample groups and protein families and mapping this data onto the maps. A table with sample annotations can be uploaded in order to define biologically or clinically relevant groups of samples. The tool provides standard heatmaps, barplots and glyphs as well as the novel map staining technique for grasping large-scale trends in numerical values (such as whole transcriptome) projected onto a pathway map. The web service provides a server mode, which allows automating visualization tasks and retrieve data from maps via RESTfull (standard HTTP) calls. Bindings to different programming languages are provided (Python, R, Java). The novelty of NaviCell Web Service is in the combination of these flexible features that provides an opportunity to adjust the modes of visualization to the data type and achieve the most meaningful picture. We illustrate the features of the tool with several case studies using pathway maps created by different research groups, in which data visualization provides new insights into molecular mechanisms involved in systemic diseases such as cancer and neurodegenerative diseases and beyond.

Applying the Ontology of Biological and Clinical Statistics (OBCS) to standardize statistical analysis of literature mined vaccine investigation data

Jie Zheng[1], Yongqun He[2]

[1] Department of Genetics, University of Pennsylvania, USA
[2] Department of Microbiology and Immunology, University of Michigan Medical School, USA

Presenter: Jie Zheng

Statistics are critical to protocol-driven biomedical research. However, the statistical methods used in biological and clinical research have thus far not been described in a consistent way, in part due to the absence of ontology resources that could be used for this purpose. Ontologies are human- and computer-interpretable representations of the types of entities in a specific scientific domain and of the relations between these types. The Ontology of Biological and Clinical Statistics (OBCS; https://code.google.com/p/obcs/) is a community-based ontology that extends the Ontology of Biomedical Investigations (OBI; http://obi-ontology.org/) and targets on the domain of statistics in the biological and clinical fields. The live attenuated Yellow Fever vaccine 17D is very effective against Yellow Fever viral infection. It has been used as a vaccine model for analysis of protective vaccine immune mechanism. Over 300 journal articles have been focused on the 17D vaccine from different aspects, such as basic DNA/RNA/protein information, immune mechanism, and various clinical human responses to the vaccine. OBCS, together with OBI and other ontologies, is used to represent metadata, experimental data, and statistical data analyses of example data sets from 17D vaccine literature reports. The ontology-based metadata representation is applied to: (1) serve as a semantic framework that links all pieces of data and data analysis methods together in a logic manner; (2) identify inconsistent results and whether the inconsistency is due to different analysis methods; and (3) build up standardized statistical methods to analyze various data types. This vaccine case study demonstrates that OBCS plays an important role in standardizing, representing, and analyzing data, facilitating automatic and reproducible biological and clinical research.

Plastid-LCGbase: a collection of evolutionarily conserved plastid-associated gene pairs

Dapeng Wang[1,2], Jun Yu[1]

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, China
[2] Stem Cell Laboratory, UCL Cancer Institute, University College London, UK

Presenter: Dapeng Wang

Plastids carry their own genetic material that encodes a variable set of genes that are limited in number but functionally important. Aside from orthology, the lineage-specific order and orientation of these genes are also relevant. Here, we develop a database, Plastid-LCGbase (http://lcgbase.big.ac.cn/plastid-LCGbase/), which focuses on organizational variability of plastid genes and genomes from diverse taxonomic groups. The current Plastid-LCGbase contains information from 470 plastid genomes and exhibits several unique features. First, through a genome-overview page generated from OrganellarGenomeDRAW, it displays general arrangement of all plastid genes (circular or linear). Second, it shows patterns and modes of all paired plastid genes and their physical distances across user-defined lineages, which are facilitated by a step-wise stratification of taxonomic groups. Third, it divides the paired genes into three categories (co-directionally-paired genes or CDPGs, convergently-paired genes or CPGs and divergently-paired genes or DPGs) and three patterns (separation, overlap and inclusion) and provides basic statistics for each species. Fourth, the gene pairing scheme is expandable, where neighboring genes can also be included in species-/lineage-specific comparisons. We hope that Plastid-LCGbase facilitates gene variation (insertion-deletion, translocation and rearrangement) and transcription-level studies of plastid genomes.

Reference: Wang, D. and Yu, J. (2015) Plastid-LCGbase: a collection of evolutionarily conserved plastid-associated gene pairs, Nucleic Acids Res, 43, D990-995.

Ontology-based framework for mass spectrum data standard and analysis

Jinmeng Jia[1], Tielilu Shi[1]

[1] Center for Bioinformatics and Computational Biology, and Shanghai Key Laboratory of Regulatory Biology, School of Life Science, East China Normal University, China

Presenter: Jinmeng Jia

Data standardizing is a basic procedure to facilitate the integration of multiple distributed and autonomous data sources. Our objective is to develop a standardizing framework for organizing LC-MS/MS experiment-related information and data from existing laboratory reports, electronic records and publications. Based on MIAPE and the data generated from the laboratories in BPRC, we created an LC-MS/MS Data-related Standard (LMMDS) to confirm the expression of each concept as well as unify the definitions. To make the data integration and analysis automatically, we suggested that computer-recognizable and computer-operable needs should be considered. We developed an application-oriented ontology system named LC-MS/MS Data-related Ontology (LMMDO) as a computerized approach to illustrate the semantic relatedness between different processes and parameters listed in LMMDS. Our proposed framework is based on the experiment records submitted to Firmiana - a MS data processing platform - and the LMMDO system. We adopted Protégé as the ontology development environment to create structured representations of the LMMDS text. Our framework not only can be used as a tool to unify and analyze the LC-MS/MS data, but also may be extended to build the knowledge base in specific domain to realize knowledge storage, knowledge sharing, alternative knowledge transmission, knowledge reuse and knowledge discovery in the future.

Challenges and Solutions to Incorporating and Using Orthology Data in Model Organism Databases

Li Ni[1], Mary Dolan[1], Sue Bello[1], Cynthia Smith[1], James Kadin[1], Joel Richardson[1], Carol J. Bult[1], Janan Eppig[1], Judith Blake[1]

[1] Bioinformatics and Computational Biology, The Jackson Laboratory, USA

Presenter: Li Ni

The fundamental mission of the Mouse Genome Database (MGD) is to facilitate use of mouse as a model system for studying human biology and disease. MGD has long exploited mammalian orthology to relate human and mouse data and infer function of human genes from experimental knowledge about mouse genes. Recently, MGD implemented many-to-many orthology assertions to better reflect current understanding about relationships among genes of vertebrate organisms. MGD moved from curated one-to-one orthology assertions to using algorithmically determined orthology representations such as HomoloGene. Although one-to-one orthology assertions between mouse/human/rat genes still hold for 95% of protein-coding genes, MGI can now more clearly represent loci that include a more complex sequence of speciation and gene duplication events. In addition, MGD extended orthology sets captured to be vertebrate-inclusive, now including orthology assertions from well-studied species such as zebrafish (Danio rerio) and chicken (Gallus gallus). MGD is exploring phylogenetic-tree based orthology predictions (e.g. PANTHER gene families) and links to orthology sets defined by sequence-based and phylogeny-based algorithms (e.g. Human Comparative Orthology Prediction (HCOP)). While most orthology assertions using different algorithms are concordant, there are important differences. The complexities of relationships between mouse and human genome features are reflected in MGD's mouse and human comparative genotype and phenotype data. In particular, clinical and molecular geneticists look to MGD to provide access to functional and phenotypic data about genes implicated in human diseases. Mouse models for diseases are provided using comparative phenotype data through the Human Mouse Disease Connection (HMDC, www.diseasemodel.org). HMDC views support complex relationships between multiple genes as related to human diseases. MGD is supported by NIH grant NHGRI HG000330.

'Next-Generation-Biocuration' at the European Nucleotide Archive (ENA)

Clara Amid[1], Ana M. Cerdeño-Tárraga[1], Richard Gibson[1], Marc Rosello[1], Petra Ten Hoopen[1], Ana Luisa Toribio[1], Guy Cochrane[1]

[1] EMBL-EBI

Presenter: Clara Amid

The European Nucleotide Archive, ENA (http://www.ebi.ac.uk/ena/), is a globally comprehensive data resource covering sequence reads, assemblies, alignments and functional annotation. ENA provides more than three decades of expertise in data storage (in pre- and post-publication stage), analysis and representation of published data. Biocuration at the European Nucleotide Archive is currently in a transit position and Biocurators are moving forward to a model of 'Next-Generation-Biocuration' that is sustainable, given the dramatic growth in sequencing that has occurred over the last few years. In this new approach ENA Biocuration moves away from the point of submission to add value by applying expert knowledge when and where it is most impactful. Biocurators in this role are being increasingly more involved in a number of tasks that require their expert knowledge, these include: Development of structured annotation checklists Development of validation rules and fixes Development and maintenance of community reporting standards Development of dictionaries and structured vocabularies for query resolution Data management and coordination for various projects Here we present ENA and its new approach of Biocuration and give examples of different areas of the team's activities that ensure ENA's delivery as a foundational resource for the scientific community.

Poster Abstract #110

Functionality and Adequacy of the Biocuration Text-mining Pipeline

Dina Vishnyakova[1], Patrick Ruch[2]

[1] Division of Medical Information Sciences, University and University Hospitals of Geneva, Switzerland
[2] Information Science Department, HES-SO/University of Applied Science Geneva, Switzerland

Presenter: Dina Vishnyakova

The proposed by the NLP community tools are standalone applications, developed by researchers with little interest in infrastructure building. On one hand, their impact in improving the quality of biocuration or accelerating the overall biocuration process has not been fully investigated. On the other hand the available research literature did not reach a consensus on a methodology to measure this impact. We have analysed inter-annotation agreement of the manual curation. Thereafter, the assistant curation was done with a support of the system, which tentatively covers the main steps of biocuration. This system is a text-mining pipeline - KiCat. It was developed in order to adapt to the annotation model of biological databases. The KiCat exploits information retrieval technologies, which have the advantage to be relatively data-poor as compared to other classification or information extraction strategies. The main idea of the KiCat system is to combine triage and annotation tasks to represent results to the end-user. The system ranks documents according to their relevance to a given subject. For the sake of evaluation the biocuration process was split on the following sub-processes: classification and annotation. The evaluation of the classification/annotation capability of the system was achieved from a user-centric perspective. Thus, the results returned by the experts for the classification task were compared to results of the system. In contrast, for the annotation task, the manual results of the experts were compared to results returned by the experts, which were assisted by the system. Finally, we have quantified how experts faced some divergence of opinion when performing their curation works.

Poster Abstract #111

Mouse Genome Informatics (MGI) GO Annotations in Context: Who, What, When and Where

Li Ni[1], David Hill[1], Karen Christie[1], Harold Drabkin[1], Dmitry Sitnikov[1], Judith Blake[1]

[1] Bioinformatics and Computational Biology, The Jackson Laboratory, USA

Presenter: Li Ni

The Mouse Genome Informatics (MGI) (http://www.informatics.jax.org) system presents both a consensus view and an experimental view of the knowledge concerning the genetics and genomics of the laboratory mouse. These data are collected from literature curation as well as downloads of large data sets. To maximize integration of all data, multiple standard vocabularies and ontologies are applied, including nomenclature, Gene Ontology (GO), Mammalian Phenotype Ontology, mouse anatomies (embryonic and adult), and disease terms. MGI is one of the founding members of the GO and uses the GO for functional annotation of genes. The GO Consortium provides an annotation workflow that includes the ability to incorporate contextual data into basic GO annotations (Huntley et al, BMC Bioinformatics, 2014). The MGI GO team group has undertaken a large-scale expert curation effort to capture these types of annotations and to provide them to the Gene Ontology Consortium. Here we report on the status of MGI contextual annotations and provide an example of how this representation can be useful by illustrating curation that describes the coordination of several gene products to regulate transcription in a tissue-specific manner. Contextual annotation can be searched and viewed using the AmiGO2 ontology search tool provided by the Gene Ontology Consortium. These enhanced annotations will also support sophisticated queries and reasoning, and will provide curated, directional links between many gene products to support pathway and network reconstruction. MGD is supported by NIH grant NHGRI HG000330; GO is supported by grant NHGRI HG002273.

The BioGRID Interaction Database and Yeastphenome: Featuring New Curation of Yeast Models of Human Disease, Full Coverage of Yeast Interactions, and Comprehensive Curation of Yeast Phenotypes

Rose Oughtred[1], Song Sun[2,3], Chandra Theesfeld[1], Jodi Hischman[1], Christie Chang[1], Michael S. Livstone[1], Jennifer Rust[1], Sven Heinicke[1], Anastasia Baryshnikova[1], Frederick P. Roth[2,3], Mike Tyers[4], Kara Dolinski[1]

[1] Lewis-Sigler Institute for Integrative Genomics, Princeton University, USA
[2] Department of Molecular Genetics, University of Toronto, Canada
[3] Department of Computer Science, University of Toronto, Canada
[4] Institute for Research in Immunology and Cancer, Université de Montréal, Canada

Presenter: Rose Oughtred

The Biological General Repository for Interaction Datasets (BioGRID) (www.thebiogrid.org) curates genetic and protein interactions for human and major model organisms. One model organism, the budding yeast Saccharomyces cerevisiae, has proved invaluable not only due to its genetic manipulability, but also due to the relatively strong conservation of basic biological processes across eukaryotes. BioGRID maintains full coverage of the literature for this organism, and as of February 2015, over 340,000 budding yeast interactions have been curated from high- and low-throughput studies found in the literature. Recent curation efforts have also focused on identifying potential models of human diseases in S. cerevisiae. To this end, human disease genes with a putative yeast ortholog were identified using the Princeton Protein Orthology Database (P-POD: ppod.princeton.edu), as well as the InParanoid Database (inparanoid.sbc.su.se). Further curation of the literature was carried out to identify human genes shown to functionally complement their orthologous S. cerevisiae mutants. These data have already proven useful to researchers using yeast as a model for human diseases. For example, the Roth lab used our curated results to evaluate their platform to assay human-to-yeast complementation experiments, where functional complementation was assessed not only for the wild type human genes, but also their known disease variants. BioGRID has also begun an additional collaboration with the Baryshnikova lab that involves curating budding yeast phenotypic data derived from high-throughput systematic screens. This yeast phenotype curation is being shared with the research community through the newly established resource, Yeastphenome.org.

CNPHD：Physique and Health Database of Chinese Nationals

Xiaolin Yang[1], Sangang Xu[1], Yanhong Wang[2], Guangliang Shan[1], Guangjin Zhu[1], Heng Wang[1], Weimin Zhu[1]

[1] Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and School of Basic Medicine, Peking Union Medical College, China
[2] Taicang Institute of Life Sciences Information, China

Presenter: Xiaolin Yang

Physique and Health Database of Chinese Nationals (CNPHD) archives in average 200 physiological & health parameters per individual from 1/10000 geographically & ethnically representative natural Chinese populations (140,000 in total). It is a product of a 14-year-old survey effort in establishing Physiological Constant System of Chinese Nationals. Two survey data types collected in CNPHD include 1) interview information on demographics (such as age, gender and ethnicity), socioeconomics (education level, marital status and occupation), lifestyle (e.g., diet, drinking and smoking habits), and past & family medical histories; 2) physical examination results including measures of anthropometrics, blood biochemical & protein profiles, electrocardiogram, bone density, vision, hemodynamics and spirometry. Blood DNA samples have been collected since last two years of the surveys for the elucidation of genetic background (genotypes) of characteristic physiological parameters (phenotypes). CNPHD is carefully curated to ensure the data quality and interoperability between different survey phases, regularly updated whenever survey data become available (typically 2-3 times/year). It provides free data query and analysis service to the public through a user-friendly web interface at http://cnphd.bmicc.cn/chs/en/.

GPNS: Gene & Protein Name-Mapping Service

Sangang Xu[1], Yin Huang[2], Weimin Zhu[1,2], Xiaolin Yang[1]

[1] Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and School of Basic Medicine, Peking Union Medical College, China
[2] Beijing Proteome Research Center, China

Presenter: Sangang Xu

Genes and their corresponding expression products (transcripts & proteins) often have several names and synonyms, which is collectively referred as naming heterogeneity. It is a big barrier for information integration, search, and retrieval of diverse sources of biomedical data. Gene & Protein Name-Mapping Service (GPNS) is developed specifically to address the challenge. The MySQL database under GPNS is constructed by using HUGO gene nomenclature as the core gene naming system, which is expanded to map gene, transcript and protein names of human and other model organisms from more than 10 representative biomedical databases such as Ensembl, UniProtKB, ENA and Refseq. The database is carefully curated and regularly updated. GPNS search service is built by using Apache Solr technology, scalable, well performed and supporting sophisticated queries, available at http://59.108.16.237/gpns/gpns.html. Besides a useful web tool, programmatic access of GPNS service as part of semantic approach for database integration is under active development. More databases will be added to GPNS in the future.

Poster Abstract #115

Novel Software Tools for Crowdsourcing Mitochondrial Protein Knowledge in Gene Wiki

Jennifer S. Polson[1,2], Anders O. Garlid[1,2], Tevfik Umut Dincer[1,2], Jessica M. Lee[1,2], Sarah B. Scruggs[1,2], Ding Wang[1,2], Andrew I. Su[1,3], Peipei Ping[1,2]

[1] NIH BD2K Center of Excellence at UCLA, USA
[2] Physiology, Medicine, Bioinformatics, University of California, Los Angeles, USA
[3] Molecular and Experimental Medicine, The Scripps Research Institute, USA

Presenter: Jennifer S. Polson

Mitochondrial biology is integral to understanding the pathophysiology of many human diseases. A wealth of knowledge is available to experienced biomedical scientists, but its access and comprehension remain unavailable to the general public. The Gene Wiki project, an effort within Wikipedia, was established to bridge the gap between esoteric and common knowledge. However, many mitochondrial proteinsÃ‚Â remain inadequately represented and poorly annotated. Concerted efforts have been undertaken to identify and address these deficiencies. The first is an assessment tool that systematically analyzes gene pages within Wikipedia by examining the quality of biologically relevant content, as well as the number of semantic web links and peer-reviewed references. Two-thirds of the 556 core mitochondrial proteins are either missing pages or they are grossly incomplete. Only 5 of the 556 proteins have relatively complete pages. Thus, significant challenges remain to address this paucity of knowledge in Wikipedia. To enhance productivity and prioritize crowdsourcing efforts, a second tool was developed to curate the PubMed database. Users input two lists of keywords to be queried in PubMed. The tool populates a contingency table, which indicates the total number of articles for the main search term as well as the number of articles that result from the crosstab query. Users can access each PubMed search query and download relevant abstracts. For the purposes of the Cardiac Gene Wiki project, this tool enables dedicated researchers and crowdsourced participants to populate Gene Wiki articles more efficiently. Beyond the Gene Wiki project, this tool will minimize the time-intensive tasks associated with academic writing. The Cardiac Gene Wiki team will employ these tools and crowdsourcing mechanisms to aggregate unstructured knowledge in biomedical literature and organize the information into a structured, user-friendly format for a broad community of users.

Comprehensive transcriptome and improved genome annotation of Bacillus licheniformis WX-02

Jing Guo[1], Gang Cheng[2], Xiang-Yong Gou[3], Feng Xing[1], Sen Li[1], Yi-Chao Han[1], Long Wang[1], Jia-Ming Song[1], Cheng-Cheng Shu[1], Shou-Wen Chen[3], Ling-Ling Chen[1]

[1] Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Informatics, Huazhong Agricultural University, China
[2] College of Life Science, South-Central University for Nationalities, China
[3] State Key Laboratory of Agricultural Microbiology, College of Life Science and Technology, Huazhong Agricultural University, China

Presenter: Jing Guo

Bacillus licheniformis WX-02 (*B. licheniformis* WX-02) is a Gram-positive bacterium that is widely used in industrial production of poly-γ-glutamic acid (γ-PGA). The updated genome of *B. licheniformis* WX-02 comprises a circular chromosome of 4,286,821 base-pairs (bp) containing 4,518 protein-coding genes. We applied strand-specific RNA-sequencing (ssRNA-seq) to explore the transcriptome profiles of *B. licheniformis* WX-02 under normal conditions and high-salt treatment (NaCl 6%). Based on this analysis, we identified 2,382 co-expressed gene pairs constituting 872 operon structures. In addition, we detected 1,169 antisense transcripts and 94 small RNAs (sRNAs). Systematic comparison of the differentially expressed genes under normal and high-salt conditions revealed that genes involved in cell motility were significantly repressed under salt stress. Genes related to the promotion of glutamic acid synthesis were activated by 6% NaCl, potentially explaining the high yield of γ-PGA under salt stress. This analysis provides an improved genome annotation and dynamic transcriptome profile for *B. licheniformis* WX-02, which will be useful for the optimization of crucial metabolic activities in this important bacterium.

Poster Abstract #117

Comprehensive resources for sweet orange genome

Ling-Ling Chen[1]

[1] Agricultural Bioinformatics Key Laboratory of Hubei Province, College of Informatics, Huazhong Agricultural University, China

Presenter: Ling-Ling Chen

Sweet orange is one of the most important and widely grown fruit crop in the world. We sequenced the draft genome of a double-haploid sweet orange (Citrus sinensis) [1]. The assembled sequence covers 87.3% of the estimated orange genome with 29,445 protein-coding genes, half of which are in the heterozygous state. We provide evidence to suggest that sweet orange is originated from a backcross hybrid between pummelo and mandarin [1]. We also investigated the heterozygosity of the sweet orange genome and examine how this heterozygosity affected gene expression [2]. In addition, we employed ortholog identification and domain combination methods to predict the protein-protein interaction (PPI) network for sweet orange. The final predicted PPI network, CitrusNet, contained 8,195 proteins with 124,491 interactions [3]. All the above information was stored in a database, Citrus sinensis annotation project (CAP, http://citrus.hzau.edu.cn/orange/), which provides comprehensive resources beneficial for the researchers of sweet orange and other woody plants [4]. Finally, we constructed a web application tool, CRISPR-P, for CRISPR sgRNA design in more than 30 plant species including sweet orange. CRISPR-P allows users to search for highly specific Cas9 target sites within DNA sequences of interest, which also provides off-target loci prediction for further analysis and marks restriction enzyme cutting sites [5]. CRISPR-P is freely available at http://cbi.hzau.edu.cn/crispr/.

References:

1. Qiang Xu#, Ling-Ling Chen#, Xiaoan Ruan#, Di-Jun Chen, Andan Zhu, et al. The draft genome of sweet orange (Citrus sinensis). Nat. Genetics, 2013, 45(1): 59-66.
2. Wen-Biao Jiao, Ding Huang, Feng Xing, Yibo Hu, Xiu-Xin Deng, Qiang Xu*, Ling-Ling Chen*. Genome-wide characterization and expression analysis of genetic variants in sweet orange. Plant J., 2013, 75(6): 954-964.
3. Yu-Duan Ding#, Ji-Wei Chang#, Jing Guo, Di-Jun Chen, Sen Li, Qiang Xu, Xiu-Xin Deng, Yun-Jiang Cheng*, Ling-Ling Chen*. Prediction and functional analysis of sweet orange protein-protein interaction network. BMC Plant Biology, 2014, 14(1): 213.
4. Jia Wang#, Di-Jun Chen#, Yang Lei#, Ji-Wei Chang, Bao-Hai Hao, Feng Xing, Sen Li, Qiang Xu, Xiu-Xin Deng, Ling-Ling Chen*. Citrus sinensis Annotation Project (CAP): a comprehensive database for sweet orange genome. PLoS One, 2014, 9(1): e87723.
5. Yang Lei#, Li Lu#, Hai-Yang Liu, Sen Li, Feng Xing, Ling-Ling Chen*. CRISPR-P: A web tool for synthetic single-guide RNA design of CRISPR-system in plants. Mol. Plant, 2014, 7(9): 1494-1496.

PlantTFDB: A portal for the functional and evolutionary study of plant transcription factors

Jinpu Jin[1], He Zhang[1], Anyuan Guo[1], Kun He[1], Xin Chen[1], Qihui Zhu[1], Ge Gao[1], Jingchu Luo[1]

[1] Collge of Life Sciences, Peking University, China

Presenter: Jinpu Jin

PlantTFDB: A portal for the functional and evolutionary study of plant transcription factors Jinpu Jin, He Zhang, Anyuan Guo, Kun He, Xin Chen, Qihui Zhu, Ge Gao#, Jingchu Luo# State Key Laboratory of Protein and Plant Gene Research, College of Life Sciences and Center for Bioinformatics, Peking University, Beijing, 100871, P.R. China With the aim to provide a resource for functional and evolutionary study of plant transcription factors (TFs), we constructed a plant transcription factor database PlantTFDB (http://planttfdb.cbi.pku.edu.cn) in 2007 and updated it in 2010 and 2013. In the latest version of this database, we systematically identified 129 288 TFs from 83 species, of which 67 species have genome sequences, covering main lineages of green plants. Besides the abundant annotation provided in the previous version, we generated more annotations for identified TFs, including expression, regulation, interaction, conserved elements, phenotype information, expert-curated descriptions derived from UniProt, TAIR and NCBI GeneRIF, as well as references to provide clues for functional studies of TFs. To help identify evolutionary relationship among identified TFs, we assigned 69 450 TFs into 3 924 orthologous groups, and constructed 9 217 phylogenetic trees for TFs within the same families or same orthologous groups respectively. In addition, we set up a TF prediction server in this version for users to identify TFs from their own sequences.

Poster Abstract #119

RhesusBase: Evolutionary Interrogation of Human Biology in Well-Annotated Genomic Framework of Rhesus Macaque

Shi-Jian Zhang[1], Xiaoming Zhong[1], Chu-Jun Liu[1], Chuan-Yun Li[1]

[1] Institute of Molecular Medicine, Peking University, China

Presenter: Shi-Jian Zhang

With genome sequence and composition highly analogous to human, rhesus macaque represents a unique reference for evolutionary studies of human biology. Here, we developed a comprehensive genomic framework of rhesus macaque, the RhesusBase2, for evolutionary interrogation of human genes and the associated regulations. A total of 1,667 next generation sequencing (NGS) data sets were processed, integrated, and evaluated, generating 51.2 million new functional annotation records. With extensive NGS annotations, RhesusBase2 refined the fine-scale structures in 30% of the macaque Ensembl transcripts, reporting an accurate, up-to-date set of macaque gene models. On the basis of these annotations and accurate macaque gene models, we further developed an NGS-oriented Molecular Evolution Gateway to access and visualize macaque annotations in reference to human orthologous genes and associated regulations (www.rhesusbase.org/molEvo). We highlighted the application of this well-annotated genomic framework in generating hypothetical link of human-biased regulations to human-specific traits, by using mechanistic characterization of the DIEXF gene as an example that provides novel clues to the understanding of digestive system reduction in human evolution. On a global scale, we also identified a catalog of 9,295 human-biased regulatory events, which may represent novel elements that have a substantial impact on shaping human transcriptome and possibly underpin recent human phenotypic evolution. Taken together, we provide an NGS data-driven, information-rich framework that will broadly benefit genomics research in general and serves as an important resource for in-depth evolutionary studies of human biology.

SorGSD: a Sorghum Genome SNP Database

Hong Luo[1], Wenming Zhao[2], Yanqing Wang[2], Jingchu Luo[3], Haichun Jing[1]

[1] Institute of Botany, Chinese Academy of Sciences, China
[2] Beijing Institute of Genomics, Chinese Academy of Sciences, China
[3] Peking University, China

Presenter: Hong Luo

We constructed a database of sorghum genome SNPs, SorGSD (http://sorgsd.big.ac.cn). Sorghum is one of the most important global crops and a source of food, feed, fiber and fuel. Recently, the genetic basis of some phenotypical features such as high photosynthetic efficiency, drought resistance and heat tolerance, has been elucidated. Nevertheless, many adaptively and agronomically important traits of different sorghum accessions vary greatly including plant architecture, stem sugar, juice accumulation, and biomass production. For high efficient genomic selection and molecular breeding, a genome-wide spectrum of the variations among various sorghum accessions is required. Based on the newly assembled and annotated genome sequence of sorghum bicolor and the recently published sorghum re-sequencing data, we constructed a sorghum genome SNP database SorGSD (http://sorgsd.big.ac.cn). SorGSD collects ~62.9M SNPs from a diverse panel of 48 sorghum accessions divided into four groups, including improved inbreds, landraces, wild and weedy sorghums, and a wild relative Sorghum propinquums. It has a web-based query interface to search and show SNPs from individual accessions, as well as to display and compare SNPs among several accessions. The query results can be visualized as text format in tables, or rendered as graphics in a genome browser. Users may find useful information from query results including type of SNPs such as non-synonymous and start/stop gain, chromosome locations of these SNPs, links to the Phytozome sorghum genome database. It also gives geographic origins and breed information of individual sorghum samples. In addition, general information related to sorghum research such as online sorghum resource and literature reference can also be found in the website. All the SNP data and annotations ca

LSD: A leaf senescence database

Yi Zhao[1], Zhonghai Li[2], Junying Peng[2], Hongwei Guo[2], Jingchu Luo[2]

[1] College of Life Sciences, Peking University, China
[2] Peking University, China

Presenter: Yi Zhao

We constructed a leaf senescence database (LSD, http://www.eplantsenescence.org/) in 2010 and ypdated it in 2013. LSD provides comprehensive information concerning senescence-associated genes (SAGs) and their corresponding mutants. We have made extensive annotations for these SAGs through both manual and computational approaches. The updated version contains 5356 genes and 322 mutants from 44 species. The updated version includes several new features: (i) Primer sequences retrieved based on experimental evidence or designed for high-through analysis were added; (ii) More than 100 images of Arabidopsis SAG mutants were added; (iii) Arabidopsis seed information obtained from The Arabidopsis Information Resource (TAIR) was integrated; (iv) Subcellular localization information of SAGs in Arabidopsis mined from literature or generated from the SUBA3 program was presented; (v) QTL information was added with links to the original database; (vi) New options such as primer and miRNA search for database query were implemented. The updated database will be a valuable and informative resource for basic research of leaf senescence and for the manipulation of traits of agronomically important plants.

AnnoLnc: a web server for integratively annotating novel human lncRNAs

Mei Hou[1], Jinpu Jin[1], Ge Gao[1]

[1] College of Life Sciences, Peking University, China

Presenter: Mei Hou

While the number of newly identified long noncoding RNAs (lncRNAs) is increasing continuously, their function are still largely elusive. Driven by the demand for timely studying these versatile molecules, we developed the AnnoLnc (http://annolnc.cbi.pku.edu.cn), an online web server for systematic annotating human lncRNAs. Designed as a one-stop portal, AnnoLnc generates full spectrum of annotations for input sequences, covering their sequence and structure features, regulation, expression, protein interaction, genetic association and evolution. To help users grasping the essence quickly, abundant text annotations are summarized as brief synopsis and heterogeneous annotations are combined and visualized as intuitive figures. Besides intuitive and mobile-friendly Web interface, AnnoLnc allows batch process as well as programmatic analysis through JSON-based Web Service APIs.

Author Index

Author	Abstract #
Shuang Yang	100
Xiaolin Yang	113, 114
Yiya Yang	70
Nan Yao	89
Jasmine Young	34
Chunlei Yu	72
Jun Yu	11, 20, 45, 106
Jiao Yuan	102, 103
Xiaoxiao Yun	101
Monique Zahn	17
Guoqing Zhang	81, 100
Haichen Zhang	24
He Zhang	118
Hongmei Zhang	62
Hui Zhang	101
Huixiong Zhang	55
Jian Zhang	69, 85
Jing Zhang	44
Shi-Jian Zhang	119
Xuegong Zhang	53

*Same name but different person.

Author	Abstract #
Yaping Zhang	84
Ying Zhang	37
Zhang Zhang	44, 45, 72, 91
Zhihua Zhang	101
Guoping Zhao	1
Wenming Zhao	84, 120
Yi Zhao *	121
Yi Zhao *	102, 103
Jie Zheng	27, 105
Xing-Da Zheng	89
Xiaoming Zhong	119
Peng Zhou	55
Yi-Jia Zhou	89
Guangjin Zhu	113
Junwei Zhu	84
Qihui Zhu	118
Silei Zhu	89
Weimin Zhu	33, 113, 114
Yunxia Zhu	99
Andrei Zinovyev	104
Dong Zou	44, 45, 72, 91

List of Participants

A

Sam Ansari
PMI
Switzerland
sam.ansari@pmi.com

Cecilia Arighi
University of Delaware
United States
arighi@dbi.udel.edu

B

Hui Bai
Beijing Institute of Radiation Medicine
China
huibai13@hotmail.com

Amos Bairoch
SIB Swiss Institute of Bioinformatics and University of Geneva
Switzerland
ab@isb-sib.ch

Vladimir Bajic
King Abdullah University of Science and Technology (KAUST)
Saudi Arabia
vladimir.bajic@kaust.edu.sa

Rama Balakrishnan
Stanford University
United States
ramab@stanford.edu

Frederic Bastian
SIB Swiss Institute of Bioinformatics - University of Lausanne
Switzerland
frederic.bastian@unil.ch

Alex Bateman
EMBL-EBI
United Kingdom
agb@ebi.ac.uk

Mei Bigliardi-Qi
A*STAR
Singapore
qimei1@gmail.com

Xiaochen Bo
Beijing Institute of Radiation Medicine
China
boxc@bmi.ac.cn

Ramona Britto
EMBL-EBI
United Kingdom
rbritto@ebi.ac.uk

Gregory Butler
Concordia University
Canada
gregb@cs.concordia.ca

C

Yinghao Cao
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
China
yhcao@genetics.ac.cn

Zhiwei Cao
Tongji University
China
zwcao@tongji.edu.cn

Esther Chan
Stanford University
United States
etchan@stanford.edu

Jiwei Chang
Huazhong Agricultural University
China
longkaichang@163.com

Chunyan Chen
Institute of Zoology, Chinese Academy of Sciences
China
shashaverygood@163.com

Guankun Chen
China Academy of Chinese Medical Sciences
China

chenguangkun2008@126.com

Ling-ling Chen
Huazhong Agricultural University
China
llchen@mail.hzau.edu.cn

Han Cheng
Huazhong University of Science and Technology
China
chenghan@hust.edu.cn

Mike Cherry
Stanford University
United States
cherry@stanford.edu

Julio Collado-Vides
Center for Genomic Sciences, UNAM
Mexico
colladojulio@gmail.com

Melanie Courtot
Simon Fraser University
Canada
mcourtot@gmail.com

D

Jean Davidson
Stanford University
United States
jean2@stanford.edu

Wankun Deng
Huazhong University of Science and Technology
China
dengwankun@gmail.com

Myra Derbyshire
NIH/NLM/NCBI
United States
derbyshi@ncbi.nlm.nih.gov

Tevfik Umut Dincer
UCLA
United States
umutdincer@gmail.com

Nan Ding
Beijing Institute of Genomics, Chinese Academy of Sciences
China
dingnan@big.ac.cn

Liguo Dong
Jilin University
China
donglg13@mail.163.com

Xunong Dong
Beijing Institute of Genomics, Chinese Academy of Sciences
China
dongxn@big.ac.cn

E

Stacia Engel
Stanford University
United States
stacia@stanford.edu

F

Takatomo Fujisawa
National Institute of Genetics
Japan
tf@nig.ac.jp

G

Ren Gan
Beijing Institute of Genomics, Chinese Academy of Sciences
China
rzgedu@163.com

Feng Gao
Centre of Bioinformatics, Tianjin University
China
fgao@tju.edu.cn

Ge Gao
Peking University
China
gaog@mail.cbi.pku.edu.cn

Qiang Gao
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
China

qgao@genetics.ac.cn

Anders Garlid
NIH BD2K Center of Excellence at UCLA
United States
aogarlid@gmail.com

Pascale Gaudet
SIB Swiss Institute of Bioinformatics
Switzerland
pascale.gaudet@isb-sib.ch

Takashi Gojobori
King Abdullah University of Science and Technology
Saudi Arabia
takashi.gojobori@kaust.edu.sa

Jin Gu
Tsinghua University
China
jgu@tsinghua.edu.cn

Jian Guan
Peking Union Medical College Hospital
China
xusangang@163.com

An-Yuan Guo
Huazhong University of Science and Technology
China
guoay@mail.hust.edu.cn

Jing Guo
Huazhong Agricultural University
China
gj30501@163.com

Jingwen Guo
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
gjw_0418@sina.com

H

Lu Han
Beijing Institute of Radiation Medicine
China
hanl06@mails.tsinghua.edu.cn

Lili Hao
Beijing Institute of Genomics, Chinese Academy of Sciences
China
haolili@big.ac.cn

Bifang He
University of Electronic Science and Technology of China
China
herb@immunet.cn

Min (Max) He
Marshfield Clinic
United States
he.max@mcrf.mfldclin.edu

Shunmin He
Institute of Zoology, Chinese Academy of Sciences
China
heshunmin@gmail.com

Song He
Beijing Institute of Radiation Medicine
China
809848790@qq.com

Henning Hermjakob
EMBL-EBI
United Kingdom
hhe@ebi.ac.uk

Ursula Hinz
SIB Swiss Institute of Bioinformatics
Switzerland
ursula.hinz@isb-sib.ch

Robert Hoehndorf
King Abdullah University of Science and Technology
Saudi Arabia
robert.hoehndorf@kaust.edu.sa

Li Hou
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China

Academy of Military Medical Science
China
lijiangyu@bmi.ac.cn

Hongxing Lei
Beijing Institute of Genomics, Chinese Academy of Sciences
China
leihx@big.ac.cn

Thomas Lemberger
EMBO
Germany
thomas.lemberger@embo.org

Suzanna Lewis
Lawrence Berkeley National Laboratory
United States
selewis@lbl.gov

Donghui Li
Phoenix Bioinformatics
United States
donghui@arabidopsis.org

Fei Li
Beijing Institute of Radiation Medicine
China
pittacus@qq.com

Haotian Li
Huazhong University of Science and Technology
China
767265653@qq.com

Hong Li
China Agricultural University
China
lihong_cau@sina.com

Jiao Li
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China
li.jiao@imicams.ac.cn

Jinbi Li
The General Hospital of People's Liberation Army
China
xusangang@163.com

Pengchao Li
Chongqing Police College
China
tllipc@sina.com

Tian Li
Southwest university
China
914578898@qq.com

Wen Li
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
China
liwen@genetics.ac.cn

Xixi Li
Beijing Normal University
China
cmblixx@gmail.com

Xue Li
Beijing Institute of Genomics, Chinese Academy of Sciences
China
lixueliuwei@163.com

Xuefei Li
MRC Human Nutrition Research
United Kingdom
xuefei.li@mrc-hnr.cam.ac.uk

Yixue Li
Shanghai Center of Bioinformation Technology
China
yxli@sibs.ac.cn

Yuanbai Li
China Academy of Chinese Medical Sciences
China
2442028249@qq.com

Yuling Li
California Institute of Technology
United States
liyuling@caltech.edu

Chaojie Lian
China Academy of Chinese Medical Sciences
China
2442028249@qq.com

Michal Linial
The Hebrew University
Israel
michall@mail.huji.ac.il

Ailin Liu
Institute of Materia Medica, Chinese Academy of Medical Sciences
China
xusangang@163.com

Bin Liu
University of Texas MD Anderson Cancer Center
United States
bliu1@mdanderson.org

Fangzhou Liu
China Academy of Chinese Medical Sciences
China
2442028249@qq.com

Lei Liu
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China
liu.lei@imicams.ac.cn

Tianfei Liu
Guangdong Academy of Agricultural Sciences
China
liutfei@qq.com

Wenjuan Liu
Academy of Military Medical Sciences
China
xusangang@163.com

Yang Liu
Beijing Institute of Radiation Medicine
China
liuyang@bmi.ac.cn

Zexian Liu
Huazhong University of Science and Technology
China
lzx@hust.edu.cn

Meng Lu
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
792459817@163.com

Chenglong Luo
Institute of Animal Science, Guangdong Academy of Agricultural Sciences
China
13427662693@163.com

Wei Luo
The General Hospital of People's Liberation Army
China
xusangang@163.com

M

Lili Ma
Huazhong University of Science and Technology
China
lili_ma@hust.edu.cn

Lina Ma
Beijing Institute of Genomics, Chinese Academy of Sciences
China
malina@big.ac.cn

Mingyue Ma
Beijing Normal University
China
mamy@mail.bnu.edu.cn

Venkat Malladi
Stanford University
United States
vmalladi@stanford.edu

Raja Mazumder
George Washington University

United States
mazumder@gwu.edu

Johanna McEntyre
EMBL-EBI
United Kingdom
mcentyre@ebi.ac.uk

Huaiyu Mi
University of Southern California
United States
huaiyumi@usc.edu

Xiao Ming
Southwest University
China
182265163@qq.com

Elvira Mitraka
University of Maryland/IGS
United States
elvira.mitraka@gmail.com

Ilene Mizrachi
NIH/NLM/NCBI
United States
mizrachi@ncbi.nlm.nih.gov

Monica Munoz-Torres
Lawrence Berkeley National Laboratory
United States
McMunozT@lbl.gov

O

Claire O'Donovan
EMBL-EBI
United Kingdom
odonovan@ebi.ac.uk

Sandra Orchard
EMBL-EBI
United Kingdom
orchard@ebi.ac.uk

Francis Ouellette
Ontario Institute for Cancer Research
Canada
francisadmin@oicr.on.ca

Rose Oughtred
Princeton University
United States
oughtred@princeton.edu

P

Yang Pan
The George Washington University
United States
panyang1989@gwmail.gwu.edu

Chong Peng
Tianjin University
China
pengchong@tju.edu.cn

Wentao Peng
Unilever Discover Shanghai
China
wen-tao.peng@unilever.com

Jennifer Polson
NIH BD2K Center of Excellence at UCLA
United States
jpolson@mednet.ucla.edu

Sylvain Poux
SIB Swiss Institute of Bioinformatics
Switzerland
sylvain.poux@isb-sib.ch

Kim Pruitt
NIH/NLM/NCBI
United States
pruitt@ncbi.nlm.nih.gov

Xun Pu
Southwest University
China
puxun@swu.edu.cn

Q

Hongzhu Qu
Beijing Institute of Genomics, Chinese Academy of Sciences
China
quhongzhu@big.ac.cn

R

Valentine Rech de Laval
SIB Swiss Institute of Bioinformatics

Switzerland
valentine.rechdelaval@isb-sib.ch

Lorna Richardson
eMouseAtlas
United Kingdom
lorna.richardson@igmm.ed.ac.uk

Marc Robinson-Rechavi
University of Lausanne
Switzerland
marc.robinson-rechavi@unil.ch

Patrick Ruch
Swiss Institute of Bioinformatics & HEG
Switzerland
patrick.ruch@hesge.ch

S

Reza Salek
EMBL-EBI
United Kingdom
reza.salek@ebi.ac.uk

Sarah Scruggs
University of California, Los Angeles
United States
sarahbscruggs@gmail.com

Chen Shao
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
scshaochen@126.com

Yi Shao
Institute of Zoology, Chinese Academy of Sciences
China
shaoyicharlie@126.com

Lingling Shen
Novartis Institutes for BioMedical Research (China)
China
lingling.shen@novartis.com

Xin Sheng
Beijing Institute of Genomics, Chinese Academy of Sciences
China
shengxin@big.ac.cn

Xing Shi
Beijing Institute of Genomics, Chinese Academy of Sciences
China
sxwy10@163.com

Yuefeng Shi
Journal of Genetics and Genomics
China
yfshi@genetics.ac.cn

Randeep Singh
SAP Labs India Pvt. Ltd.
India
randeep.singh01@sap.com

Luke Slater
King Abdullah University of Science and Technology
Saudi Arabia
luke.slater@kaust.edu.sa

Cricket Sloan
Stanford University
United States
crickets@stanford.edu

Shuhui Song
Beijing Institute of Genomics, Chinese Academy of Sciences
China
songshh@big.ac.cn

Kimchi Strasser
Concordia University
Canada
kimchi.strasser@concordia.ca

Hanfei Sun
Tongji University
China
hfsun.tju@gmail.com

Wei Sun
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
sunwei1018@sina.com

Justyna Szostak
PMI R&D
Switzerland
Justyna.Szostak@contracted.pmi.com

T

Todd Taylor
RIKEN Center for Integrative Medical Sciences
Japan
taylor@riken.jp

Mark Thomas
Wellcome Trust Sanger Institute
United Kingdom
mt4@sanger.ac.uk

Paul Thomas
University of Southern California
United States
pdthomas@usc.edu

V

Madhura Vipra
Athena Consulting
India
madhvi125@gmail.com

W

Hengtao Wang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
wanght@big.ac.cn

Jia Wang
Huazhong Agricultural University
China
wang.jia@mail.hzau.edu.cn

Jianyong Wang
Huazhong Agricultural University
China
wjy01@mail.hzau.edu.cn

Jinyue Wang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
qq-772390865@163.com

Junhui Wang
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China
informatics@imicams.ac.cn

Kai Wang
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
China
kaiwang@genetics.ac.cn

Lei Wang
CapitalBio Technology
China
leiwang@capitalbiotech.com

Liguo Wang
Mayo Clinic
United States
wang.liguo@mayo.edu

Ming Wang
Institute of Biophysics, Chinese Academy of Sciences
China
wangm08@yeah.net

Shihan Wang
Peking Union Medical College Hospital
China
xusangang@163.com

Shur-Jen Wang
Medical College of Wisconsin
United States
sjwang@mcw.edu

Weiye Wang
Xinhua Hospital,School of Medicine, Shanghai Jiao Tong University
China
wcwangw13@163.com

Xiaodong Wang
California Institute of Technology
United States
xdwang@caltech.edu

Xiu-Jie Wang
Institute of Genetics and Developmental Biology, Chinese Academy of Sciences
China
xjwang@genetics.ac.cn

Xuncheng Wang
School of life science, Peking University
China
xuncheng_wang@163.com

Yan Wang
Institute of Animal Science, Guangdong Academy of Agricultural Sciences
China
wynew2004@163.com

Yongbo Wang
Huazhong University of Science and Technology
China
ybwang@hust.edu.cn

Zhigang Wang
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
wangzg.pumc@outlook.com

Liu Wei
Huazhong University of Science and Technology
China
liuweiathust@foxmail.com

Jun-Zhi Wen
State Key Laboratory of BioControl, Sun Yat-sen University
China
JamesRNA@gmail.com

Ulrike Wittig
Heidelberg Institute for Theoretical Studies
Germany
ulrike.wittig@h-its.org

Chunlei Wu
The Scripps Research Institute
United States
cwu@scripps.edu

Jiayan Wu
Beijing Institute of Genomics, Chinese Academy of Sciences
China
wujy@big.ac.cn

Sizhu Wu
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China
wu.sizhu@imicams.ac.cn

Wei Wu
Institute of Biophysics, Chinese Academy of Sciences
China
412917426@qq.com

X

Zhang Xi
Tianjin University
China
zxwinner@yeah.net

Lili Xia
Beijing Institute of Genomics, Chinese Academy of Sciences
China
xiall@big.ac.cn

Yuan Xia
Southwest University
China
8242497@qq.com

Sizhe Xiao
Gigascience
Hong Kong SAR China
jesse@gigasciencejournal.com

Bingbing Xie
Beijing Institute of Genomics, Chinese Academy of Sciences
China
xiebb2015@gmail.com

Dafei Xie

Beijing Institute of Radiation Medicine
China
xiedafei@sina.com

Jialei Xie
Genomics, Proteomics & Bioinformatics
China
xiejl@big.ac.cn

Feng Xing
Huazhong Agricultural University
China
xfengr@hotmail.com

Sangang Xu
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
xusangang@163.com

Yu Xue
Huazhong University of Science and Technology
China
xueyu@hust.edu.cn

Y

Cheng Yan
George Wanshington University
United States
chengyan@gwu.edu

Jian Yang
Institute of Pathogen Biology, Chinese Academy of Medical Sciences
China
yangj@ipbcams.ac.cn

Lin Yang
Institute of Medical Information and Library, Chinese Academy of Medical Sciences
China
yang.lin@imicams.ac.cn

Shiping Yang
China Agricultural University
China
shipingyang@cau.edu.cn

Xiaolin Yang
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
yangxl74@gmail.com

Yadong Yang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
yangyd@big.ac.cn

Yang Yang
China Academy of Chinese Medical Sciences
China
2442028249@qq.com

Yehong Yang
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
jihuiyang@126.com

Yiyan Yang
Tongji University
China
tsubasayyy@163.com

Tian Ye
China Academy of Chinese Medical Sciences
China
2442028249@qq.com

Jasmine Young
RCSB Protein Data Bank
United States
jasmine@rcsb.rutgers.edu

Daqi Yu
Institute of Zoology, Chinese Academy of Sciences
China
1208902543@qq.com

Jun Yu
Beijing Institute of Genomics, Chinese Academy of Sciences

China
junyu@big.ac.cn

Jiao Yuan
Institute of Biophysics, Chinese Academy of Sciences
China
yuan2006jiao@163.com

Xiaoxiao Yun
Beijng Institute of Genomics, Chinese Academy of Sciences
China
yunxx@big.ac.cn

Z

Hongbo Zeng
Institute of National Health and Family Plan
China
xusangang@163.com

Guoqing Zhang
Shanghai Center for Bioinformation Technology
China
gqzhang@scbit.org

Le Zhang
Southwest University
China
zhanglcq@swu.edu.cn

Liwei Zhang
China Agricultural University
China
lwzhanghz@163.com

Qian Zhang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
zhangqian@big.ac.cn

Shi-Jian Zhang
Peking University
China
zsjsky@pku.edu.cn

Wen Zhang
Chinese Center for Disease Control and Prevetion
China
zhangwen@icdc.cn

Yanyan Zhang
Institute of Plant Protection, Chinese Academy of Agricultural Sciences
China
zyy980926@aliyun.com

Yingjie Zhang
Chinese Center for Disease Control and Prevention
China
xusangang@163.com

Zhang Zhang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
zhangzhang@big.ac.cn

Zhihua Zhang
Beijing Institute of Genomics, Chinese Academy of Sciences
China
zhangzhihua@big.ac.cn

Guoping Zhao
Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
China
gpzhao@sibs.ac.cn

Wenming Zhao
Beijing Institute of Genomics, Chinese Academy of Sciences
China
zhaowm@big.ac.cn

Yisong Zhen
Fuwai Hospital, Chinese Academy of Medical Sciences
China
zhenyisong@fuwaihospital.org

Chunqiu Zheng
Southwest University
China
1174336109@qq.com

Jie Zheng
University of Pennsylvania

United States
jiezheng@upenn.edu

Xiaoming Zhong
Peking University
China
lamz138138@163.com

Rui Zhou
HuaZhong University of Science and Technology
China
zhourui@hust.edu.cn

Weimin Zhu
Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences
China
wmzhuworld@gmail.com

Yunxia Zhu
The Jackson Laboratory
United States
sophia.zhu@jax.org

Dong Zou
Beijing Institute of Genomics, Chinese Academy of Sciences
China
zoud@big.ac.cn

About Beijing

Beijing, capital of the People's Republic of China, is the nation's political, economic, cultural, educational and international trade and communication center. The city is renowned for its opulent palaces, temples, parks and gardens, tombs, walls and gates, and its art treasures and universities have made it a center of culture and art in China. Now Beijing has become one of the most popular tourist destinations in the world, with about 140 million Chinese tourists and 4.4 million international visitors in a year. As of 1 January 2013, tourists from 45 countries are permitted a 72-hour visa-free stay in Beijing. The 45 countries include Singapore, Japan, the United States, Canada, all EU and EEA countries (except Norway and Liechtenstein), Switzerland, etc.

- **Great Wall**

The Great Wall, one of the greatest wonders of the world, winds up and down across deserts, grasslands, mountains and plateaus, stretching approximately 8,851.8 kilometers (5,500 miles) from east to west of China. With a history of more than 2000 years, some of the sections are now in ruins or have disappeared. Several walls were being built as early as the 7th century BC; these, later joined together and made bigger and stronger, are now collectively referred to as the Great Wall. Especially famous is the wall built between 220-206 BC by the first Emperor of China, Qin Shi Huang. The Great Wall is a series of fortifications made of stone, brick, tamped earth, wood, and other materials, generally built along an east-to-west line across the historical northern borders of China in part to protect the Chinese Empire or its prototypical states against intrusions by various nomadic groups or military incursions by various warlike peoples or forces.

- **Forbidden City**

The Forbidden City was the Chinese imperial palace for twenty-four emperors during the Ming dynasty (1368 - 1644) and the Qing dynasty (1644 - 1912). It was first built throughout 14 years (1406 - 1420) during the reign of Emperor Chengzu in the Ming Dynasty. Ancient Chinese Astronomers believed that the Purple Star (Polaris) was in the center of heaven and the Heavenly Emperor lived in the Purple Palace. The Palace for the emperor on earth was so called the Purple City. It was forbidden to enter without special permission of the emperor. Hence its name, 'The Purple Forbidden City', is usually 'The Forbidden City'. For almost 500 years, it served as the home of emperors and their households, as well as the ceremonial and political center of Chinese government. The Forbidden City consists of 980 buildings, covers 72 ha (180 acres) and exemplifies traditional Chinese palatial architecture, and has influenced cultural and architectural developments in East Asia and elsewhere.

More information about Beijing can be found at: http://en.wikipedia.org/wiki/Beijing

Transport Information

Beijing is a major hub for the national highway, expressway, railway, and high-speed rail networks. The Beijing Capital International Airport is the second busiest in the world by passenger traffic.

Taxi

Taking a taxi is the most convenient way for newcomers to travel around a metropolis like Beijing. There are over 66,000 taxis running in every corner of the city. Most of the taxi drivers in Beijing can speak some simple English, which offers visitors a great convenience of being able to communicate with them.

Price items and standards of charge:

- Minimum Fare: CNY 13 for the first 3 kilometers (1.86 miles).
- Basic Unit Price: CNY 2.3 per kilometer (0.62 mile) above 3 kilometers in the daytime.
- Low-Speed Drive and Waiting Fare: Waiting and traffic jam time when the speed is lower than 12 kilometers (7.5 miles) per hour is charged extra the fare of 2 kilometers for each 5 minutes during the rush hours, and the fare of 1 kilometer for each 5 minutes during other time period.
- Night Fare: From 23:00 to 5:00, the charge per kilometer rises 20%.
- CNY 6 for reservation over 4 hours in advance; CNY 5 for reservation within 4 hours.

Please note:

- The morning rush hours are from 7:00 to 9:00, and the evening rush hours is from 17:00 to 19:00.
- The price is calculated exactly by every 500 meters (547 yards) and every 2.5 minutes.
- The price will be rounded to the whole number of Chinese Yuan. For example, CNY 15.4 will be rounded down to CNY 15, and CNY 15.6 will be rounded up to CNY 16.
- The toll for an expressway or a bridge should be extra paid by the passengers.

Useful tips:

- Please make sure that the taximeter is turned on and remember to ask for a printed receipt from the taxi driver. It is highly because the details of the car are listed in the receipt and this is useful in case you have any problems such as leaving property behind.
- Usually, you can hail a taxi anywhere in the city. However, if there is a solid white line with polices around in the prosperous area, such as Tiananmen Square, the drivers will not stop for your hailing. In this case, you should look for the taxi stands or wait at the side streets.

- Make sure you use an official taxi with a sign on the roof and with the driver's registration card.
- Ask someone to write down your destination in Chinese beforehand in case of getting lost.

Subway

Subways are the fastest transportation in Beijing and they are a good way to avoid frequent traffic jams. By the end of 2014, there are 17 subway lines plus one airport express line in operation. The price of the subway is charged according to the distance. The stops are announced in both Chinese and English. Beijing Friendship Hotel is within walking distance of two subway stations on **Subway Line 4**, RENMIN University station and Weigongcun station.

- RENMIN University station: ~300 meters north of the Hotel's East Gate.
- Weigongcun station: ~350 meters south of the Hotel's East Gate, passing Beijing Institute of Technology.

Please note that a child shorter than 4.3 feet (1.3m) cannot ride the subway alone. If an adult takes one child below 4.3 feet high on the subway, the child is free of charge; if two, one of the children is free of charge. Tickets are issued for one-day use only.

Train

Beijing is the center of China's railway net, which has trains to most of cities of the country. In recent years, the inter-city railway and the high-speed rails to many cities (including Shanghai, Xi'an, Hong Kong, etc.) have also been built. For example, the Beijing-Shanghai High-speed Railway runs at a highest speed of 186mi/h (300km/h), shorten the 819 miles (1,318km) journey between the two cities to 5 hours.

To the Airport

Beijing Capital International Airport (BCIA) is located in the northeast part of the city, ~32 km (~20 mile) to Beijing Friendship Hotel.

- By **Airport Shuttle Line 4**: CNY 24 per person from 4:30~22:00. Depart every 30 minutes and without traffic take around 1 hour 10 minutes.
- By **Subway**: CNY 30 per person. **Line 4** (RENMIN University –> HaiDianHuangZhuang) –> **Line 10** (HaiDianHuangZhuang –> SanYuanQiao) –> **Airport Express Line** (SanYuanQiao –> BCIA). Take around 1 hour.
- By **Taxi:** without traffic, CNY ~100 during the daytime and CNY ~120 at night, taking around 35 minutes.

Notes

Notes

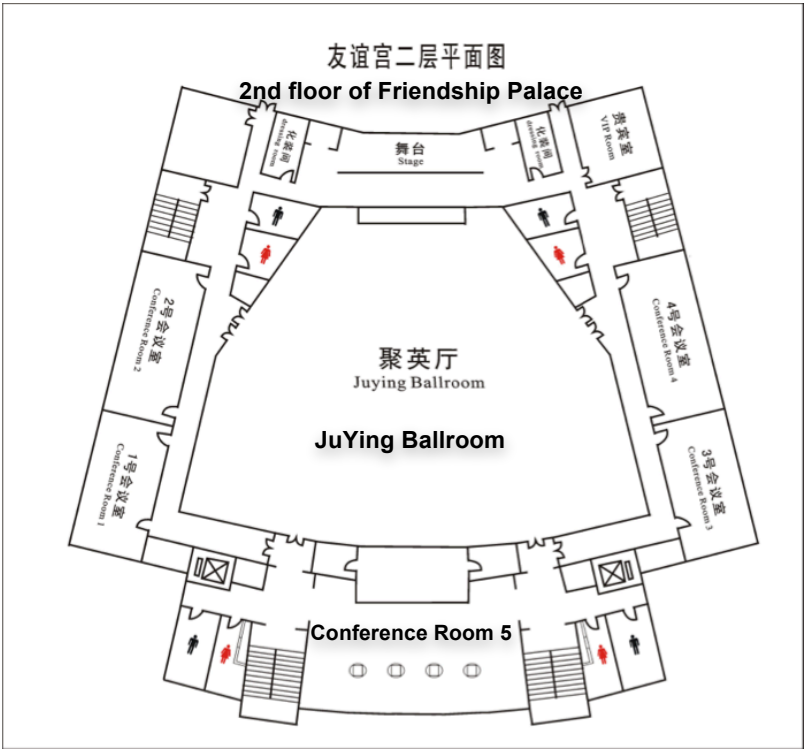
Notes

Notes

Venue Map

Catering Locations:

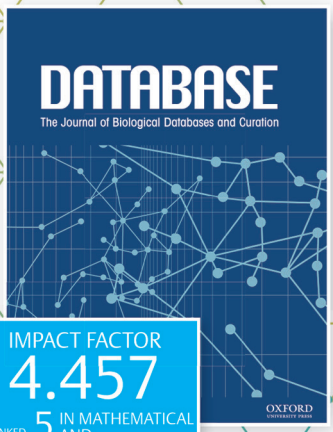
April 23: Buffet dinner at Building 1
April 24: Buffet lunch at Ju He Yuan, 1st floor of Friendship Palace
April 25: Buffet lunch at Ju He Yuan, 1st floor of Friendship Palace
April 26: Buffet lunch at Ju He Yuan, 1st floor of Friendship Palace



Oral Session Locations:
April 23: Function Room, Building 1
April 24: JuYing Ballroom, Friendship Palace
April 25: JuYing Ballroom, Friendship Palace
April 26: Ya Shi Ting, Building 1, except Workshop 6 and Special Session II at Conference Room 4 in Friendship Palace

Poster Session Locations:
April 24 and 25: Conference Rooms 1, 2 and 3, Friendship Palace

Computational Biology from Oxford Journals

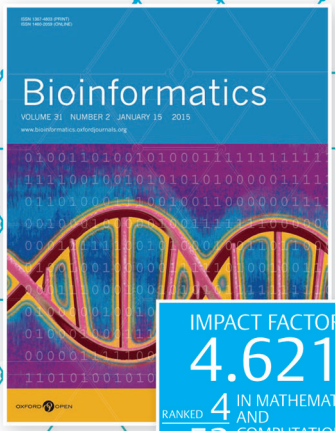


IMPACT FACTOR
4.457
RANKED 5 IN MATHEMATICAL AND 52 COMPUTATIONAL BIOLOGY

Official Journal of the International Society of Biocuration

DATABASE: The Journal of Biological Databases and Curation, the official journal of the International Society for Biocuration, provides an open access platform for the presentation of novel ideas in database research and biocuration and aims to help strengthen the bridge between database developers, curators, and users.

database.oxfordjournals.org



IMPACT FACTOR
4.621
RANKED 4 IN MATHEMATICAL AND 52 COMPUTATIONAL BIOLOGY

A leading journal in its field, *Bioinformatics* publishes the highest quality scientific papers and review articles of interest to academic and industrial researchers. Its main focus is on new developments in genome bioinformatics and computational biology.

bioinformatics.oxfordjournals.org

View OUP's full selection of Computational Biology titles, **oxford.ly/Comp-Bio**

OXFORD OPEN

OXFORD UNIVERSITY PRESS

中源协和
VCANBIO

中源协和细胞基因工程
股份有限公司

中源协和细胞基因工程股份有限公司（以下简称“中源协和”）是中国最早投资生物资源储存项目的企业，是目前国内沪深两市中唯一一家以细胞工程和基因工程为主营业务，双核驱动发展的上市公司（股票代码：600645）。

中源协和承接国家级干细胞基因信息临床转化基地建设项目并运营管理着世界最大的干细胞库平台之一——天津市脐带血造血干细胞库，同时构建了一个覆盖了全国人口四分之三地区的细胞资源储存库网络平台。截至目前，细胞储存量已达30余万份，堪称全球规模最大的细胞生物资源库之一。

经过十五年的经营，中源协和成功发展成为集细胞存储、基因检测及临床试剂、化妆品美容保健品抗衰老、肿瘤生物治疗、生命银行卡、中源药业等项目为一体的全产业链生命科技公司。

2014年9月11日，中源协和细胞基因工程股份有限公司投资成立专注于基因检测服务的中源协和基因科技有限公司。

中源协和基因科技有限公司是一家引领个性化医学发展，为广大医疗机构和民众提供全方位基因序列分析和基因信息解读服务的专业机构。秉承“准确检测、科学解读、合理引导”的理念，致力于打造中国乃至全球顶尖的检测中心。

地址：天津市空港经济区东九道45号　网址：www.vcanbio.com　电话：022-58775766　邮编：300304

todaysoft 华生基因

江苏华生恒业科技股份有限公司成立于2010年7月，位于盐城经济技术开发区，是专业从事肿瘤个体化医疗基因大数据分析平台开发及信息服务的国家高新技术企业，其控股公司北京华生恒业科技有限公司成立于2000年5月，位于北京中关村，十余年来一直致力于DNA测序软件、DNA实验室系统LIMS以及DNA数据库等系统的研发。公司自2010年起，已先后与中科院北京基因组研究所、河海大学、江苏省中医院、宁波第二医院等科研和医疗机构合作，开展临床基因测序及个体化医疗产品研发及咨询服务。

国际领先的DNA测序数据分析软件开发者

华生恒业与美国Softgenetics公司合作十余年至今，先后研发了GeneMarker、Mutation Surveyor、NextGENe等多款DNA测序分析软件。目前已有3000家以上国际基因检测实验室及研究机构应用这一系列软件，包括美国国家癌症中心、约翰.霍普金斯基梅尔癌症研究中心、GeneDX公司、日本东京大学等，有超过300篇文章和会议报告声明使用了该系列软件产品。

国家法庭科学DNA数据库服务系统承建和运维服务者

华生恒业于2007年中标北京奥运安保项目，负责承建北京市公安局法医中心DNA实验室项目。华生恒业为全球最大的公安部国家法庭科学DNA数据库的承建和运维者，此系统在快速侦破刑侦案件、打击拐卖妇女儿童、重大灾难事件伤亡人员鉴定方面提供了技术支撑的动力。

江苏省重大科技成果转化项目承接者

华生恒业“基于新一代测序技术的个性化医疗云计算平台”已获江苏省重大科技成果转化项目立项。该平台可为国内重大疾病患者自动、快速的提供基因测序分析报告，提供个体化诊疗咨询建议；为健康人群提供重大疾病早期预警基因测序分析；为机构和个人提供基因测序数据分析存储和分析服务。该平台的建立可满足未来公众大量的基因测序和个性化医疗服务的需求，提升整个社会的医疗及健康服务水平。

致力于搭建肿瘤基因大数据实时采集与应用服务平台

通过华生基因移动互联网服务平台，包括位于盐城总部的云计算中心、移动互联网终端（App端、微信端以及Pad端等）以及遍布江苏及华东地区的线下服务网络，搭建肿瘤基因大数据的采集系统及网络，利用大数据挖掘技术，为肿瘤医生、患者及医药研发机构提供肿瘤基因大数据的存储、挖掘、分析服务，最终为终端用户提供相应的癌症个体化医疗咨询服务。

联系方式

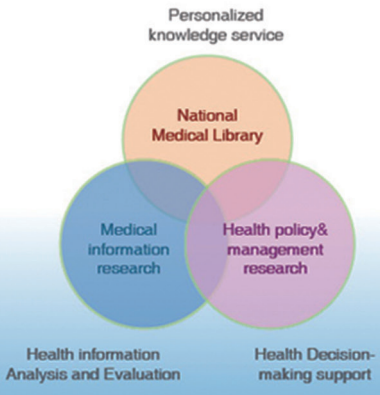
地址:江苏省盐城市希望大道南路5号（国际软件园）　　免费服务电话:400-860-2099

Introduction to Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College

The Institute of Medical Information (IMI) & Library, Chinese Academy of Medical Sciences and Peking Union Medical College (CAMS & PUMC), is designated as the National Center of Medical Information Research and Biomedical Information Resources, the Medical Center of National Science and Technology Library (NSTL), the Center for Health Policy and Management of CAMS & PUMC, the Engineering & Technology Center for National Population and Health Scientific Data Sharing Platform, the International MEDLARS Center in China, the WHO Collaborating Center for Health and Biomedical Information, the Chair of China Board for WHO Global Health Library, and the Office of Medical Information Management affiliated to Medical Information Administrative Committee of Ministry of Health. It carries out its pivotal responsibilities as national medical library and conducts research on medical information and health policy, providing information service and decision making support for the national health system reform and medical scientific innovation.

IMI was developed from the former Medical Information Department of CAMS founded in 1958. It has carried out its institutional mission since the integration of the Library of CAMS & PUMC in 1974. In the past decades, IMI has played an important role in supporting national health system reform and medical scientific innovation by conducting meaningful research in the fields of medical informatics, medical competitive intelligent analysis, health policy and management, health information system, and etc. IMI has participated in national key health related development roadmaps making, including medical science and technology development programs, national science and technology roadmaps, and 'Health China 2020'. To provide effective and efficient information services, IMI has developed medical information systems serving medical researchers, health professionals, and health policy makers, including Chinese biomedical literature database (CBM), Chinese biomedical literature service system (SinoMed), population and health scientific data sharing system, consumer health information service (CHealth) and health policy and decision support system. IMI has also hosted medical journals, including Journal of Medical Informatics, Chinese Journal of Health Policy, Journal of Medical Research, China Medical Herald, and China Modern Doctor.

The Library was developed from the former Peking Union Medical College Library founded in 1917. It been recognized as China's most historic medical library with the largest medical literature collection and was designated as the First National Medical Library by the State Council in 1957. It has served as the MOH's (Ministry of Health) National Center for Medical Literature Resources Sharing Network since 1990, the WHO's Depository Library in China since 1990, and the NSTL's Medical Center. The Library has more than 2.75 million medical books, journals, and other forms medical information, where 6900 foreign journals, 1400 Chinese journals, 84 databases and 10,000 postgraduate thesis, 1000 ancient books on Chinese traditional medicine, 1000 monographs on foreign medical history, and 20,000 WHO publications. In the digital era, the Library takes advantage of information technology to improve its service capability and scope, providing user-centered and subject-oriented information services for medical innovation and decision making consultation.



The Center for Health Policy and Management, an independent research institution affiliated to IMI, carries its mission of conducting in-depth and meaningful health policy and management related research with developing novel theories and methods to solve the frontier and strategic issues. It has focused on addressing key issues in China's health reform including health policy analysis and evaluation, health system and global health, health information management, health decision-making support system, primary healthcare, maternal and child healthcare, and health economics. It has built up an architecture for health policy analysis and decision-making supports, providing government and related agencies with policy-making evidence and consultation. In recent years, the center has participated in health reform policy making and evaluation, national health service system roadmap making, and 'Health China 2020' roadmap making. It has developed decision-making support systems including national new rural cooperative medical information platform, national population and health policy and regulation evaluation platform, and medical science and technology decision-making support system. The center also founded and hosted the widely recognized journal, Chinese Journal of Health Policy. Nowadays, the center has become an important part in China's health policy research and played an important role in government' scientific decision-making support.

Biocuration 2016

Geneva, Switzerland
April 10-13, 2016

The 9th International Biocuration Conference will be organized by the SIB Swiss Institute of Bioinformatics in Geneva, Switzerland

Join us in the hometown of the Swiss-Prot database!

SIB Swiss Institute of Bioinformatics

ISB International Society for Biocuration

Biocuration 2015

Beijing, China