Journal of Genetics and Genomics 44 (2017) 223–225

Contents lists available at ScienceDirect

Journal of Genetics and Genomics

Journal homepage: www.journals.elsevier.com/journal-of-genetics-and-genomics/

Editorial

Bioinformaticians wrestling with the big biomedical data

Database plays a critical role throughout the history of the development of bioinformatics and omics. In 1980s, a number of databases such as GenBank (NCBI Resource Coordinators, 2017), European Nucleotide Archive (ENA) (Toribio et al., 2017), and DNA Data Bank of Japan (DDBJ) (Mashima et al., 2017) were established and have still acted as major data resources for maintaining nucleotide sequences and other types of biological data until today. In early years, China had a relatively low step on the biomedical database construction, and most of Chinese biologists suffered from unstable and unreliable internet connections to access international biological databases, until official mirror sites of several major databases have been gradually established at Peking University since 1995 (Wei and Yu, 2008). During the past two decades, China had a rapid pace in the development of all aspects of biomedical researches, especially in bioinformatics. For example, the BIG Data Center was released at Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, as a central repository for the big biomedical data in China (BIG Data Center Members, 2017). Wrestling with the flood of biomedical data, Chinese bioinformaticians take the bull by the horns, focus on general or special topics of interest, and transform informative but noisy data into numerous knowledgebases. In this special issue, ten well-organized biomedical databases have been released, including five articles and five short letters (Table 1). These databases can be roughly classified into four types for their specific purposes, including for fundamental biological functions (HCSGD, RED and PLMD), for non-coding RNAs (lncRInter and mTD), for biomedical applications (ADMETNet, CMGene and SysFinder), and for infectious viruses (DRodVir and VirusMap) (Table 1).

For a better understanding of the basic biological functions, HCSGD, RED and PLMD were designed for their own purpose. As an anti-tumor program, human cellular senescence induces an irreversible cell cycle arrest to prevent cell proliferation, and is involved in a number of human diseases. Xiaowo Wang and colleagues at Tsinghua University first curated a high-quality network associated with cellular senescence by collecting 396 known human cellular senescence genes from the literature. They further 298 interacting partners of these known genes as a potential reservoir for further experimental consideration, and developed a knowledgebase of HCSGD, in which rich annotations such as protein-protein interactions, gene expression data, microRNA (miRNA) expression profiles and drug-target relations were also provided and integrated for a deep analysis of human cellular senescence (pp.227-234). Rice (*Oryza sativa*) is one of the most important cereal foods for feeding humans, and acts as an established model organism for studying monocotyledonous plants. Zhang Zhang and colleagues from Beijing Institute of Genomics (BIG) provided a rice gene expression database of RED, as a committed database of Information Commons for Rice (IC4R) (IC4R Project Consortium et al., 2016). RED not only collected, integrated and curated up to 284 quality-controlled RNA-Seq data sets for nine rice tissues, but also predicted potential housekeeping and tissue-specific genes in rice. In addition, gene co-expression networks can be visualized, whereas an easy-to-use web service was developed for using the data. RED can be a powerful data platform for further analyzing the rice functions (pp.235-241). Post-translational modifications (PTMs) occurring at specific lysine residues in proteins or protein lysine modifications (PLMs) have attracted much attention for their importance in the regulation of a large number of biological processes. Xu et al. from Huazhong University of Science and Technology constructed an integrative resource of protein lysine modification database (PLMD), containing 284,780 modification events in 53,501 proteins across 176 eukaryotes and prokaryotes for 20 types of PLMs. Further analyses demonstrated that different types of PLMs prefer to mutually crosstalk with each other by co-occurring at the same residues, although each type of PLM recognizes a distinct sequence motif. PLMD can be useful for further PLM studies (pp.243-250).

For the analysis of non-coding RNAs, two databases including lncRInter and mTD have been constructed. Long noncoding RNAs (lncRNAs) are transcribed from noncoding regions of the genome (Johnsson and Morris, 2014; Li and Wang, 2015), and play a critical roles in regulating gene/protein functions by interacting with different types of bio-molecules, such as DNAs, RNAs and proteins. For a better understanding of the functional significance of lncRNAs, An-Yuan Guo and colleagues from Huazhong University of Science and Technology manually collected 922 lncRNA interaction pairs between 276 lncRNAs and 597 partners of 15 organisms, from over 500 publications, and developed a well-annotated database of lncRInter, which can serve as a useful resource for the lncRNA research community (pp.265-268). Also, as small noncoding RNAs with 19–24 nucleotides, miRNAs play essential roles in the regulation of almost all of biological processes (Liang and Wang, 2013). Recent findings demonstrated that a considerable number of miRNAs can interact with drug response genes to affect the efficacy of the therapy. In this issue, Xing-Ming Zhao and colleagues at Tongji University developed a new database of microRNAs affecting therapeutic effects of drugs (mTD), which maintains 893 miRNA-drug interactions among 208 miRNAs and 157 drugs. The mTD database not only provides a useful resource

http://dx.doi.org/10.1016/j.jgg.2017.05.002

Table 1
A summary of the ten biomedical databases.

Database	Web link	Content
HCSGD	http://bioinfo.au.tsinghua.edu.cn/member/xwwang/HCSGD/	Collects 396 known human cellular senescence genes with 298 interacting partners
RED	http://expression.ic4r.org/	Incorporates 284 high-quality RNA-Seq data sets for rice
PLMD	http://plmd.biocuckoo.org/	Integrates 284,780 sites for 20 types of protein lysine modifications
lncRInter	http://bioinfo.life.hust.edu.cn/lncRInter/	Provides 922 interactions between 276 lncRNAs and 597 partners
mTD	http://mtd.comp-sysbio.org/	Contains 893 miRNA-drug interactions among 208 miRNAs and 157 drugs
ADMETNet	http://bioinf.xmu.edu.cn/ADMETNet/	Has 1974 proteins associated with drug bioavailability and toxicity
CMGene	http://cmgene.bioinfo-minzhao.org/	Maintains 2006 well-curated human cancer metastasis genes
SysFinder	http://lifecenter.sgst.cn/SysFinder	To find appropriate animal models for translational research
DRodVir	http://www.mgc.ac.cn/DRodVir	Contains 7690 sequences of 5491 rodent-associated animal viruses
VirusMap	http://virusmap.renlab.org/	For analyzing the distribution of influenza A viruses and integrating a couple of software packages for the phylogenetic analysis

for an explicit description of miRNAs that affect the drug sensitivity or resistance, but also can be fundamental for understanding the molecular mechanisms of miRNAs in cells (pp.269-271).

Three databases including ADMETNet, CMGene and SysFinder were developed mainly for the purpose of biomedical usage. The drug Absorption, Distribution, Metabolism, and Excretion (ADME) orchestrated by numerous ADME-associated proteins (ADME-APs) is important for understanding the drug toxicity and benefiting new drug discovery. ADME-APs interact with drugs/metabolites to form networks, and the aberrance of these proteins results in poor pharmacokinetics, low efficacy or toxicity in human bodies. In this issue, Zhiliang Ji's group at Xiamen University presented a comprehensive resource of ADMETNet for a better understanding of the pharmacokinetics and toxicity. Totally, ADMETNet contains 5905 ADMET pathways, 1541 approved small-molecule drugs and 1974 ADME-APs and 13,459 drug-protein pairs. This database can be highly useful for a precision drug design with better pharmacokinetic features and less toxicity (pp.273-276). Cancer metastasis is a complicated multistep process mainly involved in spreading cancer cells into other tissues through either the blood or the lymphatic system, and causes massive death of cancer patients. Previously, Zhao et al. (2015) manually curated 194 experimentally identified metastasis suppressor genes, while in this issue they further expanded their scope and developed the CMGene database by a total collection of 2006 human cancer metastasis-related genes (CMGs). Rich annotations were provided for each CMG entry, and multiple options were implemented for a convenient usage of the database. CMGene provides a unique resource for the cancer research (pp.277-279). In addition, accumulative research demonstrated that human and animal models exhibit considerably differences in a variety of biological functions, and thus the selection of appropriate animal models for studying human biology has emerged to be a great challenge. In this issue, Yang et al. developed SysFinder to prioritize optimal models for analyzing disease mechanisms and drug effects that may benefit for the translational research in humans, by the multi-level comparison of human genes implicated in a certain disease, drug response or biological pathway with their cognates of animal models. Besides maintaining species-specific information in human and animal models, SysFinder also provides a sub-option of Topic2Strain for the genome editing in mice (pp.251-258).

For a better understanding of the virome diversity of rodents and widespread zoonotic diseases that severely threaten human health, Chen et al. developed a comprehensive database DRodVir, containing 7690 nucleotide sequences of 5491 well-curated rodent-associated viruses classified into 26 viral families from 194 rodent species distributed in 93 countries. DRodVir not only integrates detailed information on related samples and host rodents, but also provides multiple online options for querying and using the data, and can serve as a useful platform to monitor the epidemiological and geographical distribution of zoonotic diseases (pp.259-264). As a highly virulent pathogen, influenza A virus caused several pandemic events in the history, and still acts as a severe threat to human health at present. Thus, the investigation of the epidemiological and geographical distribution of influenza A viruses can facilitate the determination of the origin of infectious viruses and the design of new treatments. Jian Ren's group at Sun Yat-sen University presented a data platform of VirusMap, which contained the host, serotype and sampling information of 583,052 protein and 448,495 nucleotide sequences for influenza A viruses. VirusMap not only realizes a Google Maps API for the data visualization, but also integrates a couple of software packages for the phylogenetic analysis of viruses of interest (pp.281-284).

Finally, the last three decades have witnessed a rapid accumulation of biomedical data, which contain both information and noise. How to retrieve useful knowledge from the highly noisy big biomedical data has emerged to be a great challenge for all bioinformaticians, and greatly hampered our understandings on the basis of life. Although too many computational or experimental analyses can be performed by playing with the big data, the development and maintenance of biomedical databases with well-organized contents are undoubtedly the most fundamental efforts for the scientific community.

References

BIG Data Center Members, 2017. The BIG Data Center: from deposition to integration to translation. Nucleic Acids Res. 45, D18–D24.
IC4R Project Consortium, Hao, L., Zhang, H., Zhang, Z., Hu, S., Xue, Y., 2016. Information Commons for rice (IC4R). Nucleic Acids Res. 44, D1172–D1180.
Johnsson, P., Morris, K.V., 2014. Expanding the functional role of long noncoding RNAs. Cell Res. 24, 1284–1285.
Li, Y.P., Wang, Y., 2015. Large noncoding RNAs are promising regulators in embryonic stem cells. J. Genet. Genomics 42, 99–105.
Liang, Z., Wang, X.J., 2013. Rising from ashes: non-coding RNAs come of age. J. Genet. Genomics 40, 141–142.
Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., Takagi, T., 2017. DNA Data Bank of Japan. Nucleic Acids Res. 45, D25–D31.
NCBI Resource Coordinators, 2017. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 45, D12–D17.
Toribio, A.L., Alako, B., Amid, C., Cerdeno-Tarraga, A., Clarke, L., Cleland, I., Fairley, S., Gibson, R., Goodgame, N., Ten Hoopen, P., Jayathilaka, S., Kay, S., Leinonen, R., Liu, X., Martinez-Villacorta, J., Pakseresht, N., Rajan, J., Reddy, K., Rosello, M., Silvester, N., Smirnov, D., Vaughan, D., Zalunin, V., Cochrane, G., 2017. European nucleotide archive in 2016. Nucleic Acids Res. 45, D32–D36.
Wei, L., Yu, J., 2008. Bioinformatics in China: a personal perspective. PLoS Comput. Biol. 4, e1000020.
Zhao, M., Li, Z., Qu, H., 2015. An evidence-based knowledgebase of metastasis suppressors to identify key pathways relevant to cancer metastasis. Sci. Rep. 5, 15478.

Yu Xue*
MOE Key Laboratory of Molecular Biophysics, College of Life Science and Technology and the Collaborative Innovation Center for Brain Science, Huazhong University of Science and Technology, Wuhan 430074, China

Xiu-Jie Wang**
The State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

* Corresponding author.

** Corresponding author.
E-mail address: xueyu@hust.edu.cn (Y. Xue).
E-mail address: xjwang@genetics.ac.cn (X.-J. Wang).

14 April 2017
Available online 18 May 2017