Published online 3 October 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D121–D126***
doi: 10.1093/nar/gkx875

ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization

Jian Sang[1,2,3,†], Zhennan Wang[3,4,†], Man Li[1,2,3], Jiabao Cao[1,2,3], Guangyi Niu[1,2,3], Lin Xia[1,2,3], Dong Zou[1,2], Fan Wang[1,2], Xingjian Xu[1,2], Xiaojiao Han[5], Jinqi Fan[6], Ye Yang[7], Wanzhu Zuo[7], Yang Zhang[1,2,3], Wenming Zhao[1,3], Yiming Bao[1,2,3], Jingfa Xiao[1,2,3,8], Songnian Hu[2,3,8], Lili Hao[1,2,*] and Zhang Zhang[1,2,3,8,*]

[1]BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [3]University of Chinese Academy of Sciences, Beijing 100049, China, [4]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China, [5]Key Laboratory of Tree Genomics, The Research Institute of Subtropical Forestry, Chinese Academy of Forestry, Hangzhou 311400, China, [6]School of Ocean, Yantai University, Yantai 264005, China, [7]School of Life Sciences, Zhengzhou University, Zhengzhou 450001, China and [8]Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai 200438, China

Received August 15, 2017; Revised September 04, 2017; Editorial Decision September 17, 2017; Accepted September 23, 2017

ABSTRACT

Real-time quantitative PCR (RT-qPCR) has become a widely used method for accurate expression profiling of targeted mRNA and ncRNA. Selection of appropriate internal control genes for RT-qPCR normalization is an elementary prerequisite for reliable expression measurement. Here, we present ICG (http://icg.big.ac.cn), a wiki-driven knowledgebase for community curation of experimentally validated internal control genes as well as their associated experimental conditions. Unlike extant related databases that focus on qPCR primers in model organisms (mainly human and mouse), ICG features harnessing collective intelligence in community integration of internal control genes for a variety of species. Specifically, it integrates a comprehensive collection of more than 750 internal control genes for 73 animals, 115 plants, 12 fungi and 9 bacteria, and incorporates detailed information on recommended application scenarios corresponding to specific experimental conditions, which, collectively, are of great help for researchers to adopt appropriate internal control genes for their own experiments. Taken together, ICG serves as a publicly editable and open-content encyclopaedia of internal control genes and accordingly bears broad utility for reliable RT-qPCR normalization and gene expression characterization in both model and non-model organisms.

INTRODUCTION

Real-time quantitative PCR (RT-qPCR) is one of the most powerful molecular techniques for accurate expression profiling of targeted nucleic acid in a wide range of biological research (1,2). To reduce experimental bias and produce accurate expression levels, several variables (like operator variability, amount of RNA extraction yield and variation) need to be taken into account for normalization (3–5). Currently, the most frequently used approach for RT-qPCR normalization is the use of internal control genes (or reference genes) (6) that ideally should have relatively stable expression levels across all samples from different tissues, during all developmental stages and in response to distinct experimental treatments (7,8). Thus, housekeeping genes, such as *ACT*, *18S rRNA* and *GAPDH*, were frequently used for RT-qPCR normalization (9). However, evidence has accumulated that some traditional housekeeping genes used to control for experimental bias are expressed at relatively constant levels only for certain conditions (10,11). It is clearly that internal control genes are condition-specific and accordingly there is no universal gene that can be used for internal control for all application scenarios (12), strongly indicating the necessity of proper selection of internal control gene(s) before performing any RT-qPCR experiment.

Over the past decade, with the ever-increasing RT-qPCR expression analyses carried out in both model and non-

*To whom correspondence should be addressed. Email: haolili@big.ac.cn
Correspondence may also be addressed to Zhang Zhang. Tel: +86 10 84097261; Fax: +86 10 84097720; Email: zhangzhang@big.ac.cn
†These authors contributed equally to this work as first authors.

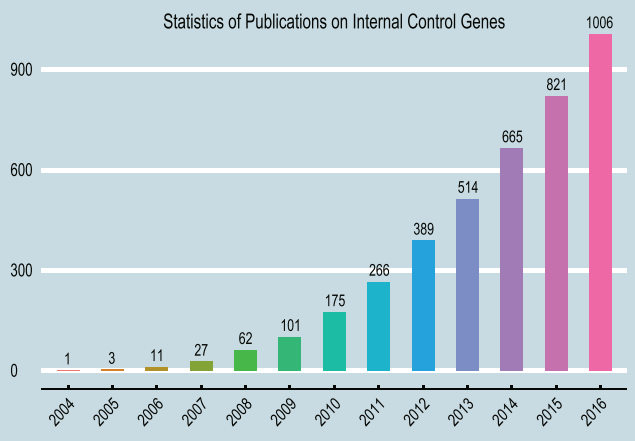

Figure 1. Cumulative numbers of relevant publications on internal control genes from 2004∼2016. All statistics were extracted from NCBI PubMed by using the search terms: ('internal control genes' OR 'reference genes') AND ('qPCR' OR 'qRT-PCR') occurring in Title/Abstract.

model organisms, advancements have been made in identification and validation of appropriate internal control genes under specific tissues, developmental stages and experimental treatments (Figure 1). However, characterizing internal control genes is an onerous task requiring well-designed molecular experiments followed with a series of elaborate computational analyses (3,13,14). Therefore, it is extremely necessary to comprehensively integrate experimentally validated internal control genes from published literature and make these genes and their associated experimental conditions well-organized and public accessible to the whole scientific community. Although valuable efforts have been made in building related databases including RTPrimerDB (15), PrimerBank (16), qPrimerDepot (17) and GETPrime (18), they merely focus on RT-qPCR primers in model organisms (mainly human and mouse), ignoring collection of internal control genes as well as their associated application scenarios. To date, there still lacks a unified knowledgebase that integrates internal control genes adjusting for various tissues, developmental stages and experimental treatments across a wide variety of species.

In order to fill this gap and provide molecular biologists with informative guidance on selecting internal control genes to customize their RT-qPCR experiments, here we present ICG (http://icg.big.ac.cn), a wiki-based, publicly editable and open-content resource for community curation of internal control genes across a diversity of species. Unlike extant relevant databases, ICG features harnessing collective intelligence in collaborative integration of experimentally validated internal control genes as well as their associated application scenarios in both model and non-model organisms, accordingly bearing great utility for proper selection of internal control genes and reliable gene expression normalization and characterization.

IMPLEMENTATION

ICG is built based on MediaWiki (http://www.mediawiki.org; version 1.28.2), which is one of popular open-source wiki engines, originally providing a collaborative framework for use on Wikipedia. The majority of contents in ICG is stored as wiki-markup text, which is organized by MediaWiki concepts such as 'template scheme' and 'content page'. Additionally, *Category*, as a software feature of MediaWiki, is extensively used for automatic indexes and classifications of content pages in ICG. To increase the usability and searchability, a series of extensible plugins are installed in aid of content presentation and customized functionalities (http://icg.big.ac.cn/index.php/Special:Version). ICG is implemented with Apache (https://httpd.apache.org; an open-source HTTP server; version 2.2.15), PHP (http://www.php.net; a widely-used general-purpose scripting language; version 7.0.19) and MySQL (http://www.mysql.org; a free and popular relational database management system; Version 5.7.13) on a CentOS release 6.5 Linux Server. Powered by MediaWiki, therefore, ICG allows any registered user to edit any content simply via a web browser and enables internal control genes to be edited and updated by multiple users. For each page, ICG records all revisions and their associated users who are responsible for each revision, and most importantly, each history revision can be easily recovered, with the purpose to minimise invalid/incorrect edits.

DATABASE CONTENT AND USAGE

To facilitate appropriate selection of internal control genes for accurate RT-qPCR normalization, ICG integrates >750 experimentally validated internal control genes manually curated from 283 publications, corresponding to a wide range of specific tissues, development stages and experiment treatments and covering a wide variety of species including 73 animals, 115 plants, 12 fungi and 9 bacteria. Consequently, ICG provides two major categories, namely, Species and Genes, to allow users to access internal control genes and their associated specific experimental conditions.

ICG organizes experimentally validated internal control genes as well as their associated experimental conditions in terms of 'Species' (http://icg.big.ac.cn/index.php/Species), where each species corresponds to a wiki page (Figure 2). Specially, the content of a species page is structured into multiple sections, namely, basic description, experimental condition(s), reference(s) and category. For each experimental condition, ICG incorporates an abundance of information, involving internal control genes (e.g. gene symbol, full name, accession number), primers (e.g. validated primer sequence, amplicon size) and RT-qPCR conditions (e.g. recommended application scopes, detection chemistry, annealing temperature), which, collectively, are helpful for researchers to select appropriate internal control genes for their own experiments. Additionally, ICG specifies evaluation methods that are used for identification of internal control genes and provides relevant publications, citations and contact information of their corresponding authors.

Meanwhile, considering that one gene is most likely used for internal control in multiple species, ICG sets up a specific page for each collected gene (http://icg.big.ac.cn/index.php/ICG:Genes). For any given gene, ICG integrates a wide range of related information, including its synonyms, applicable species, recommended application scenarios, sequence from representative species, conserved domains and

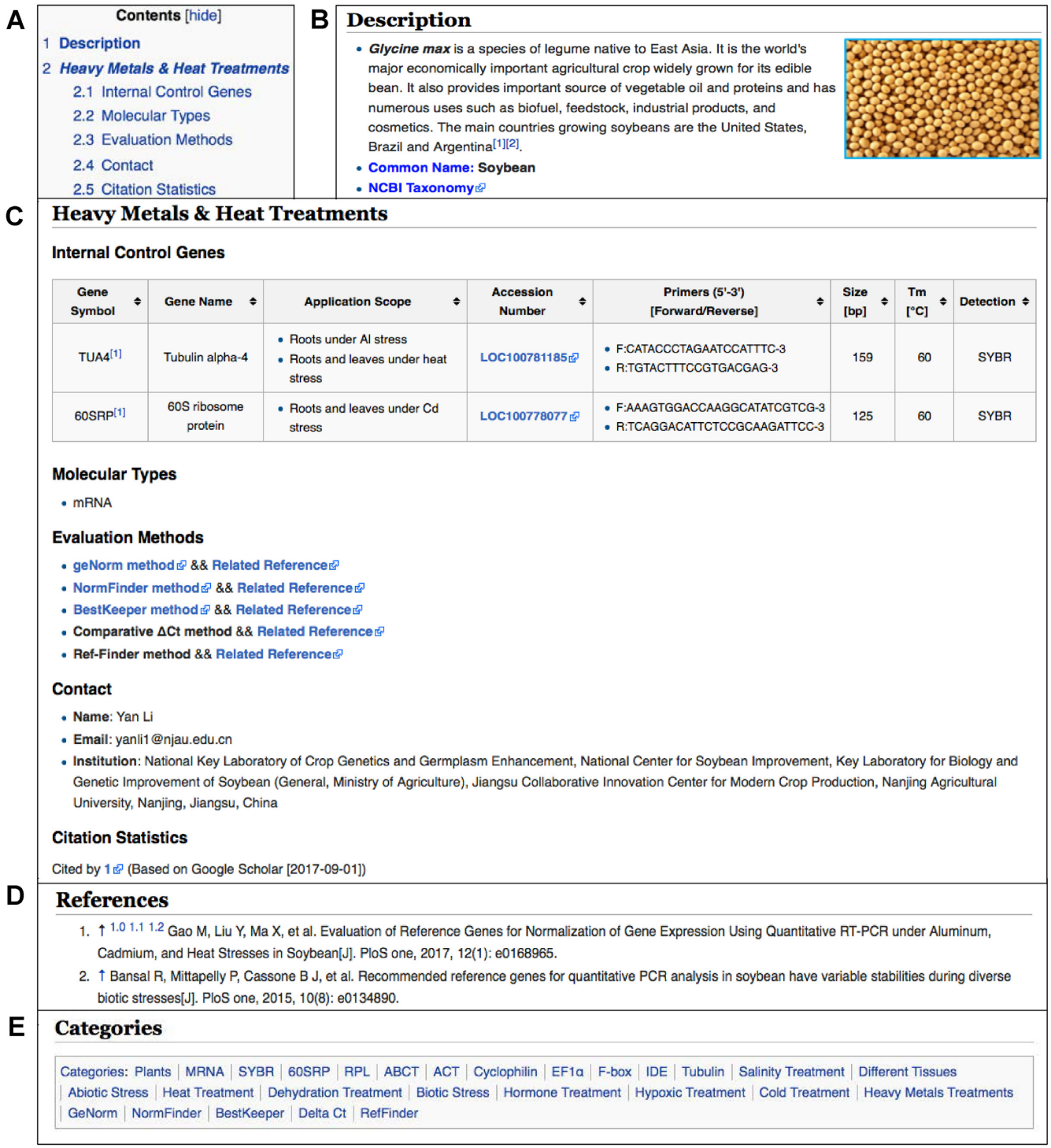

Figure 2. Screenshots of a species page for *Glycine max* (http://icg.big.ac.cn/index.php/Glycine_max). (**A**) Table of contents; (**B**) Description of the species as well as its common name and a hyperlink to NCBI Taxonomy; (**C**) Detailed information on internal control genes for a specific experimental condition; (**D**) References associated with this species; (**E**) Categories associated with this species.

Figure 3. Screenshots of a gene page for *Actin* (http://icg.big.ac.cn/index.php/Gene:ACT). (**A**) Synonymous names; (**B**) A tabulated form summarizing its utilization as internal control in all relevant species as well as recommended application scenarios and related references; (**C**) Sequence for *Actin* in a representative species; (**D**) Gene structure; (**E**) Links to external resources.

external hyperlinks (Figure 3), which on the whole provides a whole picture for the utilization of this gene across different species and thus greatly facilitates users to perform systematic investigations on this gene. For instance, according to the statistics as of 10 August 2017 (http://icg.big.ac.cn/index.php/ICG:Statistics), the most popular internal control gene collected in ICG is *EF1α* (Elongation factor 1-alpha), which has been widely adopted for controlling experimental bias in 79 species (http://icg.big.ac.cn/index.php/Gene:EF1A). Additionally, ICG collects internal control genes for non-coding RNAs normalization, which can be accessed through a specific category of non-coding RNA (http://icg.big.ac.cn/index.php/Category:Non-coding_RNA).

In the era of big data, community curation bears the potential in dealing with the flood of data (19). Based on MediaWiki, ICG enables users to be easily involved in an ongoing process of collaboration that adds newly identified internal control genes and frequently updates the contents for all collected genes. Thus, ICG can significantly ease the process of data collection, curation and sharing, befitting the exploding volume of biological knowledge. Moreover, ICG features user-friendly web interfaces for data search and retrieval just by specifying a gene name or a species name. To get an overview of all collected data contents, ICG also provides statistics for species, internal control genes, and experiment conditions and generates a word cloud for visualizing the most prominent terms (http://icg.big.ac.cn/index.php/ICG:Statistics). In addition, molecular sequences of validated internal control genes are collected and publicly available at http://icg.big.ac.cn/index.php/Downloads.

DISCUSSION AND FUTURE DEVELOPMENTS

ICG, to our knowledge, is the first knowledgebase integrating a comprehensive collection of experimentally validated internal control genes as well as their associated application scenarios across a wide range of species. Currently, it has integrated >750 experimentally validated internal control genes covering 209 species, accordingly providing valuable guidance for researchers to choose proper genes for their own RT-qPCR experiments. Considering the continuous accumulation of newly characterized internal control genes from subsequently published literature, ICG will continue to regularly update the experimentally verified genes for newly studied species and/or conditions, not only for linear RNAs but also circular RNAs (20). As a core resource of BIG Data Center (http://bigd.big.ac.cn) (21), ICG serves as a publicly editable and open-content encyclopedia of internal control genes and thus bears broad utility for reliable RT-qPCR normalization and gene expression characterization in both model and non-model organisms. Future directions of ICG include integration of more internal control genes through literature curation and development of new functionalities for inviting authors of recent relevant publications to get involved in community curation. We will also develop tools in aid of literature mining and community curation in order to facilitate automatic information retrieval and improve the reliability of community-provided contents.

ACKNOWLEDGEMENTS

We thank a number of users for reporting bugs and providing suggestions as well as the anonymous reviewers for their valuable comments on this work.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDB13040500, XDA08020102 to Z.Z.; XDA08020102 to S.H.]; National Programs for High Technology Research and Development [2015AA020108, 2012AA020409 to Z.Z.]; National Key R&D Program of China [2017YFC0907502 to Z.Z.]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; The 100-Talent Program of Chinese Academy of Sciences [to Y.B. Z.Z.]; National Natural Science Foundation of China [31100915 to L.H.]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences.

Conflict of interest statement. None declared.

REFERENCES

1. Erickson,H.S., Albert,P.S., Gillespie,J.W., Rodriguez-Canales,J., Linehan,W.M., Pinto,P.A., Chuaqui,R.F. and Emmert-Buck,M.R. (2009) Quantitative RT-PCR gene expression analysis of laser microdissected tissue samples. *Nat. Protoc.*, **4**, 902–922.
2. Mestdagh,P., Van Vlierberghe,P., De Weer,A., Muth,D., Westermann,F., Speleman,F. and Vandesompele,J. (2009) A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.*, **10**, R64.
3. Andersen,C.L., Jensen,J.L. and Orntoft,T.F. (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Resa*, **64**, 5245–5250.
4. Bustin,S.A., Benes,V., Nolan,T. and Pfaffl,M.W. (2005) Quantitative real-time RT-PCR - a perspective. *J. Mol. Endocrinol.*, **34**, 597–601.
5. Remans,T., Keunen,E., Bex,G.J., Smeets,K., Vangronsveld,J. and Cuypers,A. (2014) Reliable gene expression Analysis by reverse transcription-quantitative PCR: reporting and minimizing the uncertainty in data accuracy. *Plant Cell*, **26**, 3829–3837.
6. Schmittgen,T.D. and Livak,K.J. (2008) Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.*, **3**, 1101–1108.
7. Pfaffl,M.W. (2001) A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.*, **29**, e45.
8. Udvardi,M.K., Czechowski,T. and Scheible,W.R. (2008) Eleven golden rules of quantitative RT-PCR. *Plant Cell*, **20**, 1736–1737.
9. Sang,J., Han,X., Liu,M., Qiao,G., Jiang,J. and Zhuo,R. (2013) Selection and validation of reference genes for real-time quantitative PCR in hyperaccumulating ecotype of Sedum alfredii under different heavy metals stresses. *PLoS One*, **8**, e82927.
10. Czechowski,T., Stitt,M., Altmann,T., Udvardi,M.K. and Scheible,W.R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in Arabidopsis. *Plant Physiol.*, **139**, 5–17.
11. Rubie,C., Kempf,K., Hans,J., Su,T.F., Tilton,B., Georg,T., Brittner,B., Ludwig,B. and Schilling,M. (2005) Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol. Cell. Probe*, **19**, 101–109.
12. Bustin,S.A., Benes,V., Garson,J.A., Hellemans,J., Huggett,J., Kubista,M., Mueller,R., Nolan,T., Pfaffl,M.W., Shipley,G.L. *et al.* (2009) The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.*, **55**, 611–622.
13. Pfaffl,M.W., Tichopad,A., Prgomet,C. and Neuvians,T.P. (2004) Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper—excel-based tool using pair-wise correlations. *Biotechnol. Lett.*, **26**, 509–515.

14. Vandesompele,J., De Preter,K., Pattyn,F., Poppe,B., Van Roy,N., De Paepe,A. and Speleman,F. (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.*, **3**, RESEARCH0034.
15. Pattyn,F., Robbrecht,P., De Paepe,A., Speleman,F. and Vandesompele,J. (2006) RTPrimerDB: the real-time PCR primer and probe database, major update 2006. *Nucleic Acids Res.*, **34**, D684–D688.
16. Spandidos,A., Wang,X.W., Wang,H.J. and Seed,B. (2010) PrimerBank: a resource of human and mouse PCR primer pairs for gene expression detection and quantification. *Nucleic Acids Res.*, **38**, D792–D799.
17. Cui,W., Taub,D.D. and Gardner,K. (2007) qPrimerDepot: a primer database for quantitative real time PCR. *Nucleic Acids Res.*, **35**, D805–D809.
18. Gubelmann,C., Gattiker,A., Massouras,A., Hens,K., David,F., Decouttere,F., Rougemont,J. and Deplancke,B. (2011) GETPrime: a gene- or transcript-specific primer database for quantitative real-time PCR. *Database (Oxford)*, **2011**, bar040.
19. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
20. Gao,Y., Wang,J.F. and Zhao,F.Q. (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.*, **16**, doi:10.1186/s13059-014-0571-3.
21. Zhang,Z., Zhao,W.M., Xiao,J.F., Song,S.H., Hao,L.L., Li,R.J., Ma,L.N., Sheng,X., Sang,J., Wang,Y.Q. *et al.* (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, **45**, D18–D24.