D8–D14 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 26 October 2018*
doi: 10.1093/nar/gky993

Database Resources of the BIG Data Center in 2019

BIG Data Center Members[†]

Received September 14, 2018; Revised October 08, 2018; Editorial Decision October 09, 2018; Accepted October 10, 2018

ABSTRACT

The BIG Data Center at Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences provides a suite of database resources in support of worldwide research activities in both academia and industry. With the vast amounts of multi-omics data generated at unprecedented scales and rates, the BIG Data Center is continually expanding, updating and enriching its core database resources through big data integration and value-added curation. Resources with significant updates in the past year include BioProject (a biological project library), BioSample (a biological sample library), Genome Sequence Archive (GSA, a data repository for archiving raw sequence reads), Genome Warehouse (GWH, a centralized resource housing genome-scale data), Genome Variation Map (GVM, a public repository of genome variations), Science Wikis (a catalog of biological knowledge wikis for community annotations) and IC4R (Information Commons for Rice). Newly released resources include EWAS Atlas (a knowledgebase of epigenome-wide association studies), iDog (an integrated omics data resource for dog) and RNA editing resources (for editome-disease associations and plant RNA editosome, respectively). To promote biodiversity and health big data sharing around the world, the Open Biodiversity and Health Big Data (BHBD) initiative is introduced. All of these resources are publicly accessible at http://bigd.big.ac.cn.

INTRODUCTION

The BIG Data Center (http://bigd.big.ac.cn) at Beijing Institute of Genomics (BIG) of the Chinese Academy of Sciences (CAS) provides open access to a suite of database resources, with the aim to support research activities for domestic and international users in both academia and industry to translate big data into big discoveries (1,2). During the past years, advanced sequencing technologies (3) (e.g. PacBio, Nanopore) have been widely applied in biomedical research studies, including genome sequencing of important/featured species (e.g. opium poppy (4), wheat (5), Mexican axolotl (6), goat (7)), metagenome sequencing (8), 3D genome organization (9), single-cell epigenomics (10) and integrative omics for precision medicine (11). Undoubtedly and consequently, huge amounts of multi-omics data have been and will also be produced at ever-greater scales and rates. As a corollary, the BIG Data Center, in close collaboration with partner institutions, is continually expanding, updating and enriching its database resources through big data integration and value-added curation, with significant improvements and advances over the previous version. Here, we provide a brief overview of core database resources of the BIG Data Center (Figure 1) and describe new and updated resources. All resources are publicly accessible through the home page of the BIG Data Center at http://bigd.big.ac.cn.

NEW RESOURCES

EWAS Altas

EWAS Atlas (http://bigd.big.ac.cn/ewas) is a curated knowledgebase of Epigenome-Wide Association Studies (EWAS). A growing body of EWAS studies have identified the associations between epigenetic variations and a wide range of traits, including not only phenotypes and diseases, but also behaviors and environmental exposures. Hence, EWAS Atlas is committed to comprehensively integrating EWAS associations through literature curation and making all curated information well organized and publicly available. In the current implementation, EWAS Atlas focuses on DNA methylation—one of the key epigenetic markers, and accordingly, integrates 329 172 high-quality EWAS associations manually curated from 649 studies in 401 publications, including 112 tissues/cell lines, 305 traits, 1830 cohorts and 390 ontology entities. Moreover, it provides an online tool for trait enrichment analysis, allowing users to explore trait–trait and trait–epigenome relationships. Together, EWAS Atlas is dedicated to the curation, integration and standardization of EWAS knowledge and aims to provide a valuable resource for the worldwide research community.

*To whom correspondence should be addressed: Zhang Zhang, Tel: +86 10 84097261; Fax: +86 10 84097720; Email: zhangzhang@big.ac.cn
Correspondence may also be addressed to Wenming Zhao. Email: zhaowm@big.ac.cn
Correspondence may also be addressed to Jingfa Xiao. Email: xiaojingfa@big.ac.cn
Correspondence may also be addressed to Yiming Bao. Email: baoym@big.ac.cn
[†]Full list provided in Appendix.

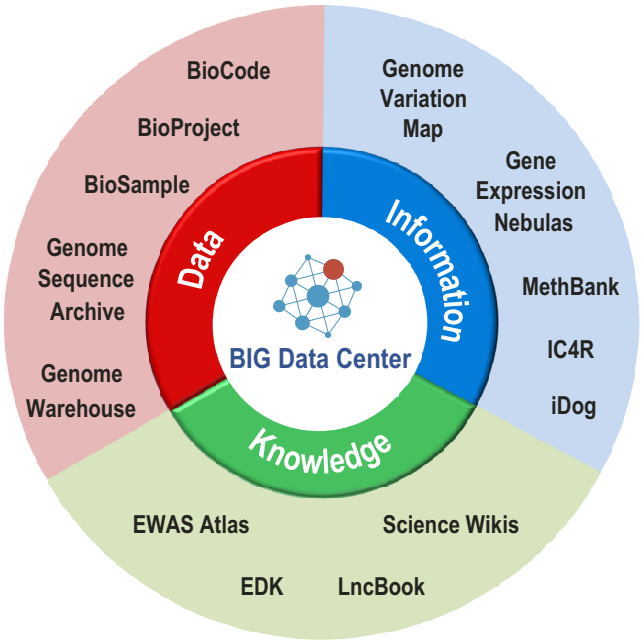

Figure 1. The BIG Data Center's core data resources. Three categories, viz. data, information and knowledge, are adopted to represent resources that are typically to deposit raw data/metadata (archives), house processed/analyzed data (libraries) and integrate validated knowledge (through literature curation; knowledgebases), respectively. A full list of data resources, which contains links to each resource, is available at http://bigd.big.ac.cn/databases.

iDog

iDog (http://bigd.big.ac.cn/idog), a component of Dog 10K Genomes Project (http://dog10k.big.ac.cn), is an integrated omics data resource for dog. Currently, iDog houses various types of omics data, primarily including 2 newly *de novo* assembled genomes, 42 871 184 genomic variations from 127 samples (12), a broad diversity of phenotypes, diseases and traits from 473 standardized breeds, more than 27 000 gene expression profiles from 7 public RNA-Seq projects as well as 6535 disease-related literatures/books. iDog provides friendly web interfaces for data browsing, retrieval and download. Moreover, genome browsers are adopted to provide data visualization services, displaying multiple omics data in a visualized manner. In addition, BLAST and BWA tools are integrated to provide online data analysis services. Together, iDog provides open access to all publicly available data and delivers a variety of data services for the *Canis* research community.

RNA editing resources

RNA editing, as an essential co-/post-transcriptional modification, plays critical roles in many biological processes and associates closely with human diseases and plant development and growth. Therefore, Editome Disease Knowledgebase (EDK; http://bigd.big.ac.cn/edk) and Plant Editosome Database (PED; http://bigd.big.ac.cn/ped) have been developed to integrate RNA editing data in human and plants, respectively. EDK is a curated knowledgebase of editome-disease associations, featuring comprehensive integration of abnormal RNA editing events and aberrant RNA editing enzyme activities associated with human diseases. Currently, EDK incorporates 65 diseases associated with 248 experimentally validated abnormal editing events located in 32 messenger RNAs, 16 microRNAs, 1 long non-coding RNAs (lncRNAs) and 11 viruses, and 44 aberrant activities involved in 6 editing enzymes, manually retrieved from more than 200 publications. PED is a curated database of RNA editosome in plants. Based on manual curation of related literatures and organelle genome annotations, PED integrates a complete collection of 98 RNA editing factors and 20 836 RNA editing events, which are identified in 205 targeted organelle genes and 1618 organisms. PED also features incorporation of molecular interactions between editing factors and editing events in eight model organisms, functional effects of editing factors regulating plant phenotypes as well as detailed experimental evidence. Collectively, EDK and PED integrate a comprehensive collection of curated RNA editing data and provide important resources for better understanding RNA editing machinery.

UPDATED RESOURCES

BioProject

The BioProject database (http://bigd.big.ac.cn/bioproject) is a public library of biological research projects, archiving a set of descriptive metadata on biological projects and providing a centralized access to all public projects. It supports various projects in terms of data types, ranging from genomic, transcriptomic, epigenomic and metagenomic sequencing projects to genome-wide association studies and variation analyses. In the past year, BioProject was enhanced by improving bilingual support in English and Chinese, adding hyperlinks to internal resources and providing more statistics in terms of data type, organism and funding agency. As of September 2018, BioProject has housed a total of 739 biological projects submitted by 421 users from 132 organizations, showing a rapid growth in project submission in the past one year (Figure 2A).

BioSample

The BioSample Database (http://bigd.big.ac.cn/biosample) is a public library of biological samples. It stores descriptive information about biological materials used for experiments, including sample type and attribute(s), and reciprocal links to data derived from it. In the past year, BioSample was significantly upgraded by adding metagenome and environmental samples and adopting the Genome Sequence Archive (GSC) Minimum Information about a MetaGenome Sequence (MIMS) standards to describe and standardize sample metadata from human-gut, soil and water. It was also improved by supporting bilingual web pages, updating batch-submission templates and adding reciprocal hyperlinks to BioProject as well as other resources. As of September 2018, BioSample has accommodated a total of 42 333 samples from 247 species (Figure 2A), presenting a dramatic increase in data submission compared to the previous release with 14 453 samples in last September.

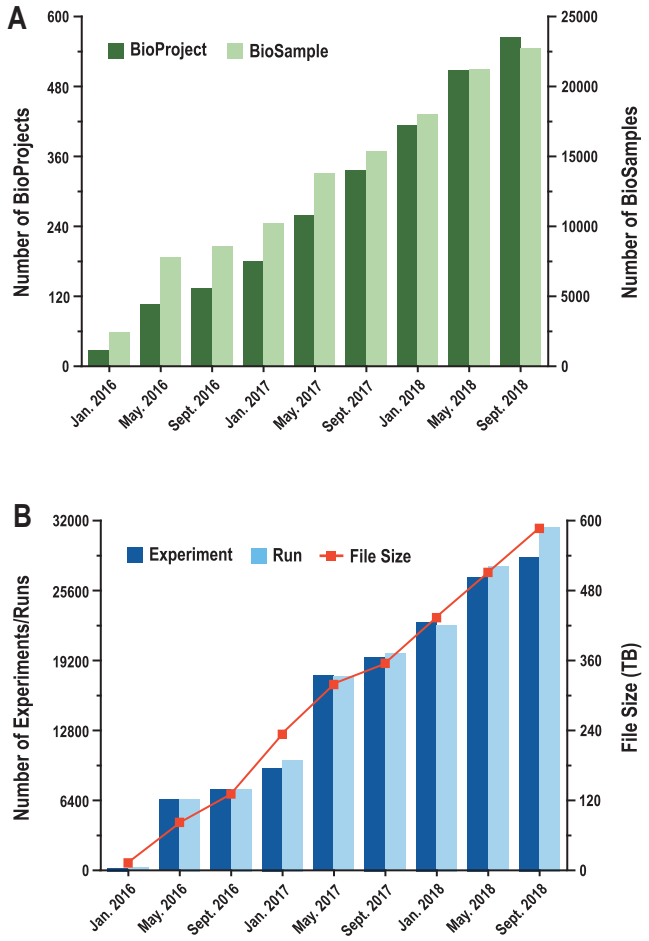

Figure 2. Statistics of data submissions to BioProject, BioSample and GSA. (**A**) Data statistics of BioProject and BioSample. (**B**) Data statistics of Experiments and Runs as well as file size in GSA. All statistics are frequently updated and publicly available at http://bigd.big.ac.cn/bioproject/, http://bigd.big.ac.cn/biosample/ and http://bigd.big.ac.cn/gsa/.

Genome Sequence Archive

The Genome Sequence Archive (GSA; http://bigd.big.ac.cn/gsa) (13) is a public data repository for archiving raw sequence reads. It accepts sequencing data submissions from all over the world and provides free access to all publicly available data for global scientific communities. In the past year, GSA has been greatly enhanced by supporting more sequencing platforms (such as BioNano, PacBio), improving bilingual support and documentation, and providing multiple statistics and charts for sample type, organism, platform, organization, download, etc. Till September 2018, GSA has archived a total of 28 660 Experiments and 31 348 Runs and housed more than 580 Terabytes of sequencing data (Figure 2B), exhibiting a rapid growth in data submission in contrast to the previous release (19 484 Experiments, 21 363 Runs, ~360 Terabytes). According to the statistics (http://bigd.big.ac.cn/gsa/statistics), data housed by GSA that were submitted by ~330 users from ~100 organizations (including research institutions, universities, hospitals and companies), have been frequently downloaded by worldwide users. In addition, data submitted to GSA have been reported in 43 journals, including Cell, Genome Research, Genomics Proteomics Bioinformatics, Nature, Plant Cell and PNAS. All released data in GSA are publicly accessible and downloadable at ftp://download.big.ac.cn/gsa/.

Genome Warehouse

The Genome Warehouse (GWH; http://bigd.big.ac.cn/gwh) is a public archival resource housing genome-scale data for a wide range of species. Compared to the previous release, GWH has been significantly upgraded by accepting more data types (including whole genome, chloroplast, mitochondrion and plasmid) and improving web services for data submission, release and sharing. Particularly, the current version of GWH is able to accept both online and offline submissions. Besides, it provides application programming interfaces to enable automatic data exchange. For each collected genome assembly, GWH incorporates detailed descriptive information, including biological sample metadata, genome assembly metadata, sequence data and genome annotation, and offers standardized quality control for genome sequence and annotation. In addition to 150 released genome assemblies integrated from NCBI (14), GWH has accepted 128 direct submissions from both domestic and international institutions and covered a broad diversity of species, viz. animals, plants, fungi, bacteria, archaea and viruses. Among all direct submissions, 19 genome assemblies have been publicly released before September 2018. Clearly, GWH has a rapid growth in data submission and thus bears the great promise to serve as an important resource for genome related studies. Future directions of GWH include improvement of web interfaces for data submission, presentation and visualization, standardization of genome reannotation pipelines and incorporation of more quality control tools.

Genome Variation Map

The Genome Variation Map (GVM; http://bigd.big.ac.cn/gvm) (15) is a public database of genome variations, including single nucleotide polymorphisms and small insertions and deletions. Unlike NCBI dbSNP and dbVar (which phased out support for non-human data and stopped accepting non-human data submissions from September 2017), GVM aims to collect, integrate and visualize genome variations for a wide range of species, including not only human but also cultivated plants (e.g. rice, maize), domesticated animals (e.g. chicken, goat) and featured species (e.g. giant panda, moso bamboo), and accepts submissions of different types of genome variations from all over the world. The current release of GVM houses a total of ~7.5 billion variants (including 6.4 billion single nucleotide polymorphism (SNPs) and 1.1 billion Indels) for 30 species; in contrast to the previous version, 2.5 billion variants of 1527 individuals for 11 species, including 2 animals (horse and tarpan) and 9 plants (carrot, cassava, common bean, cotton, cucumber, date palm, grape, Japanese apricot and rapeseed), are newly integrated. In addition, GVM has been greatly enriched by integrating 97 426 high-quality genotype-to-phenotype (G2P) associations for 13 non-human species through literature curation. Moreover,

GVM has accepted 12 genome variation dataset submissions, involving 18 597 samples primarily from human and duck. Meanwhile, protein domain information obtained from Pfam (16) have been incorporated into GVM in order to enhance the annotation for functionally relevant variants. Collectively, GVM dedicates to integrating genomic variation data and G2P associations, with the aim to become a valuable resource for better understanding population genetic diversity and deciphering complex mechanisms associated with different phenotypes, especially for domesticated animals and cultivated plants.

Science Wikis

Science Wikis (http://bigd.big.ac.cn/sciencewikis) is a catalog of biological knowledge wikis that are built based on MediaWiki (a free and open-source software wiki package) or wiki concept, with the aim to harness community intelligence in knowledge integration and curation. The current release of Science Wikis has six bio-wikis, including LncRNAWiki (17), RiceWiki (18), ESND (19), WikiCell (20), ICG (21) and one new resource Database Commons that is based on wiki concept. In the past year, the major updates of Science Wikis are as follows. LncRNAWiki (http://lncrna.big.ac.cn) has been updated by curating more experimentally validated human lncRNAs and associating lncRNAs with human diseases; a total of 1867 featured lncRNAs have been manually community-curated based on published literatures and 1502 of them have been associated with cancer and other diseases. Since LncRNAWiki, built based on MediaWiki, has significant limitations on managing structured data and providing customized functionalities, we developed an expert-curation-based resource, LncBook (http://bigd.big.ac.cn/lncbook), as a complement to community-curation-based LncRNAWiki. LncBook houses a large number of 270 044 lncRNAs, integrates an abundance of multi-omics data from expression, methylation, genome variation and lncRNA–miRNA interaction, and identifies 97 998 lncRNAs that are putatively associated with diseases. Database Commons (http://databasecommons.org) is a curated catalog of global biological databases that provides open access to a comprehensive collection of publicly available databases encompassing different data types and spanning diverse organisms. Currently, it contains descriptive metadata for a total of 4478 databases, involving ∼500 organisms and ∼2100 organizations throughout the world. RiceWiki (http://ricewiki.big.ac.cn) is a wiki-based, publicly editable and open-content platform for community curation of rice genes. The current release of RiceWiki includes community annotations for more than 600 genes, which were obtained through literature curation of more than 1000 published articles.

Information Commons for Rice

Information Commons for Rice (IC4R; http://ic4r.org) (22) is an integrated resource containing multiple omics data for rice. With the rapid advances in high-throughput sequencing technologies, the availability of massive RNA-Seq data in rice offers great opportunities to refine rice gene models. Based on large-scale RNA-Seq datasets, IC4R has been upgraded by reannotating the rice genome (*Oryza sativa* L. ssp. japonica), providing more complete and accurate characterization of rice gene models and accordingly releasing a new annotation system – IC4R-2.0. By comparison with the previous annotation systems, IC4R-2.0 significantly improves the completeness of gene structure, identifies a number of novel genes, lncRNAs and circular RNAs, and integrates a variety of functional annotations. As a consequence, IC4R-2.0 achieves higher integrity and quality primarily attributable to massive RNA-Seq data adopted in genome reannotation. Additionally, IC4R has been considerably updated by providing more friendly web interfaces and implementing a series of useful online tools, which together makes it a valuable resource for comparative and functional genomic studies in rice and other monocotyledonous species.

BIG Search

BIG Search (http://bigd.big.ac.cn/search), a scalable text search engine based on ElasticSearch (a highly scalable open-source full-text search and analytics engine based on Apache Lucene), delivers a one-stop search service that executes a query against massive indexed data provided by multiple different resources. In the era of big data where the quantity of biological data and the number of database resources continue to grow, BIG Search provides cross-domain search and facilitates users to gain access to a variety of database resources, inside and outside of the BIG Data Center. In the current version, BIG Search has integrated data indexes from 20 partner databases, among which representative partners are AnimalTFDB (23) that is a database of genome-wide transcription factors and cofactors in 97 animals, iUUCD (24) that is an integrated database of regulators for ubiquitin and ubiquitin-like conjugations, LncRNADisease (25) that is a database of lncRNA-disease associations, lncRNASNP (26) that is a database of functional SNPs and mutations in human and mouse lncRNAs, PceRBase (27) that is a database of plant competing endogenous RNAs, PlantTFDB (28) that is a database of plant transcriptional factors, RhesusBase (29) that is a knowledgebase for the monkey research community, SEGreg (30) that is a database for human specifically expressed genes and their regulations in cancer and normal tissues and THANATOS (31) that is an integrative resource of proteins and post-translational modifications in the regulation of autophagy and cell death. A full list of partner databases containing their corresponding links is available at http://bigd.big.ac.cn/partners. Accordingly, BIG Search has been significantly updated by incorporating a huge number of data indexes and providing inter-domain navigation to a wide range of biomedical data in many database resources.

BIG Submission

BIG Submission (http://bigd.big.ac.cn/gsub), previously named as Gsub, is a unified submission portal that provides submission services for a series of database resources in the BIG Data Center, including BioProject, BioSample, BioCode, GSA, GWH and GVM. BIG Submission has un-

dergone significant updates in the past year; its backend infrastructure has been upgraded by increasing the network bandwidth and expanding the storage and computing resources, with the purpose to meet the needs of the rapid growth of data submissions. Additionally, it has been updated by improving submission services with more friendly web interfaces and providing better bilingual support and documentation.

BIG SSO

The BIG Single Sign-On (SSO; http://bigd.big.ac.cn/sso) is a user access control system that enables users to access a family of web systems in the BIG Data Center while just providing their credentials only once. Users just sign-on any system once and then grant access to the rest of systems without using different usernames or passwords. Till September 2018, there are eight database resources equipped with BIG SSO, greatly facilitating users to submit data or provide community annotations. Meanwhile, BIG SSO provides personal profiles for all registered users; for each user, it not only contains registration information but also provides data statistics of all relevant submissions. Ongoing developments are equipment of SSO in other database resources of the BIG Data Center where authentication is required.

CONCLUDING REMARKS

The BIG Data Center provides open access to a suite of database resources, with the aim to support worldwide research activities in both academia and industry. In the past year, it has been significantly updated by accepting more data submissions, integrating different types of data, conducting value-added curation, developing new database resources and improving web interfaces and services. Albeit relatively young, the BIG Data Center has gained an increasing number of data submissions as well as funding support from the government and CAS. Considering the exponentially growing volume of biological data powered by advanced sequencing technologies toward higher throughput and lower cost, we initialized to set up the Global Biodiversity and Health Big Data (BHBD) Alliance (http://bhbd-alliance.org), a non-profit, non-governmental organization under the framework of 'Open Biodiversity and Health Big Data Initiative' by the International Union of Biological Sciences. Based on the BIG Data Center, the BHBD Alliance is committed to promoting biodiversity and health big data sharing in the world. Taken together, the BIG Data Center will continue to grow to offer a wide range of data services in aid of worldwide research activities for big data deposition, integration and translation.

ACKNOWLEDGEMENTS

We thank a number of users for submitting data, providing annotations, sending suggestions and reporting bugs. The BIG Data Center is indebted to its funders, including the Ministry of Science and Technology of China, the Natural Science Foundation of China, the CAS and BIG.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) [XDA19050302, XDB13040500, XDA08020102]; National Key Research & Development Program of China [2018YFC0310602, 2017YFC0907502, 2017YFC0907503, 2017YFC0908403, 2017YFC1201200, 2016YFC0901603, 2016YFB0201702, 2016YFC0901903, 2016YFE0206600]; National Natural Science Foundation of China [31771465, 31671360, 31571358]; International Partnership Program of the CAS [153F11KYSB20160008]; 13th Five-year Informatization Plan of CAS [XXH13505-05]; Key Program of the Chinese Academy of Sciences [KJZD-EW-L14]; Key Technology Talent Program of the CAS; The 100 Talent Program of the Chinese Academy of Sciences; The Youth Innovation Promotion Association of the CAS [2017141, 2018134]; The Special Project on Precision Medicine under the National Key R&D Program [SQ2017YFSF090210]; The Open Biodiversity and Health Big Data Initiative of IUBS. Funding for open access charge: Strategic Priority Research Program of the CAS [XDA19050302].
Conflict of interest statement. None declared.

REFERENCES

1. BIG Data Center Members (2018) Database Resources of the BIG Data Center in 2018. *Nucleic Acids Res.*, **46**, D14–D20.
2. BIG Data Center Members (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, **45**, D18–D24.
3. Goodwin,S., McPherson,J.D. and McCombie,W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333–351.
4. Guo,L., Winzer,T., Yang,X., Li,Y., Ning,Z., He,Z., Teodor,R., Lu,Y., Bowser,T.A., Graham,I.A. *et al.* (2018) The opium poppy genome and morphinan production. *Science*, doi: 10.1126/science.aat4096.
5. Ling,H.Q., Ma,B., Shi,X., Liu,H., Dong,L., Sun,H., Cao,Y., Gao,Q., Zheng,S., Li,Y. *et al.* (2018) Genome sequence of the progenitor of wheat A subgenome Triticum urartu. *Nature*, **557**, 424–428.
6. Nowoshilow,S., Schloissnig,S., Fei,J.F., Dahl,A., Pang,A.W.C., Pippel,M., Winkler,S., Hastie,A.R., Young,G., Roscito,J.G. *et al.* (2018) The axolotl genome and the evolution of key tissue formation regulators. *Nature*, **554**, 50–55.
7. Bickhart,D.M., Rosen,B.D., Koren,S., Sayre,B.L., Hastie,A.R., Chan,S., Lee,J., Lam,E.T., Liachko,I., Sullivan,S.T. *et al.* (2017) Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat. Genet.*, **49**, 643–650.
8. Meng,C., Bai,C., Brown,T.D., Hood,L.E. and Tian,Q. (2018) Human gut microbiota and gastrointestinal cancer. *Genomics Proteomics & Bioinformatics*, **16**, 33–49.
9. Bonev,B. and Cavalli,G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661–678.
10. Kelsey,G., Stegle,O. and Reik,W. (2017) Single-cell epigenomics: recording the past and predicting the future. *Science*, **358**, 69–75.
11. Karczewski,K.J. and Snyder,M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.*, **19**, 299–310.
12. Bai,B., Zhao,W.M., Tang,B.X., Wang,Y.Q., Wang,L., Zhang,Z., Yang,H.C., Liu,Y.H., Zhu,J.W., Irwin,D.M. *et al.* (2015) DoGSD: the dog and wolf genome SNP database. *Nucleic Acids Res.*, **43**, D777–D783.
13. Wang,Y., Song,F., Zhu,J., Zhang,S., Yang,Y., Chen,T., Tang,B., Dong,L., Ding,N., Zhang,Q. *et al.* (2017) GSA: genome sequence archive. *Genomics Proteomics & Bioinformatics*, **15**, 14–18.
14. Coordinators,N.R. (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
15. Song,S., Tian,D., Li,C., Tang,B., Dong,L., Xiao,J., Bao,Y., Zhao,W., He,H. and Zhang,Z. (2018) Genome Variation Map: a data

repository of genome variations in BIG Data Center. *Nucleic Acids Res.*, **46**, D944–D949.
16. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
17. Ma,L.N., Li,A., Zou,D., Xu,X.J., Xia,L., Yu,J., Bajic,V.B. and Zhang,Z. (2015) LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res.*, **43**, D187–D192.
18. Zhang,Z., Sang,J., Ma,L., Wu,G., Wu,H., Huang,D., Zou,D., Liu,S., Li,A., Hao,L. *et al.* (2013) RiceWiki: a wiki-based database for community curation of rice genes. *Nucleic Acids Res.*, **42**, D1222–D1228.
19. Dai,L., Xu,C., Tian,M., Sang,J., Zou,D., Li,A., Liu,G., Chen,F., Wu,J., Xiao,J. *et al.* (2013) Community intelligence in knowledge curation: an application to managing scientific nomenclature. *PLoS One*, **8**, e56961.
20. Zhao,D., Wu,J., Zhou,Y., Gong,W., Xiao,J. and Yu,J. (2012) WikiCell: a unified resource platform for human transcriptomics research. *OMICS*, **16**, 357–362.
21. Sang,J., Wang,Z., Li,M., Cao,J., Niu,G., Xia,L., Zou,D., Wang,F., Xu,X., Han,X. *et al.* (2018) ICG: a wiki-driven knowledgebase of internal control genes for RT-qPCR normalization. *Nucleic Acids Res.*, **46**, D121–D126.
22. IC4R Project Consortium. (2016) Information Commons for Rice (IC4R). *Nucleic Acids Res.*, **44**, D1172–D1180.
23. Hu,H., Miao,Y.R., Jia,L.H., Yu,Q.Y., Zhang,Q. and Guo,A.Y. (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res.*, doi: 10.1093/nar/gky1822.
24. Zhou,J., Xu,Y., Lin,S., Guo,Y., Deng,W., Zhang,Y., Guo,A. and Xue,Y. (2018) iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations. *Nucleic Acids Res.*, **46**, D447–D453.
25. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
26. Miao,Y.R., Liu,W., Zhang,Q. and Guo,A.Y. (2018) lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.*, **46**, D276–D280.
27. Yuan,C., Meng,X., Li,X., Illing,N., Ingle,R.A., Wang,J. and Chen,M. (2017) PceRBase: a database of plant competing endogenous RNA. *Nucleic Acids Res.*, **45**, D1009–D1014.
28. Jin,J., Tian,F., Yang,D.C., Meng,Y.Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.
29. Zhang,S.J., Liu,C.J., Shi,M., Kong,L., Chen,J.Y., Zhou,W.Z., Zhu,X., Yu,P., Wang,J., Yang,X. *et al.* (2013) RhesusBase: a knowledgebase for the monkey research community. *Nucleic Acids Res.*, **41**, D892–D905.
30. Tang,Q., Zhang,Q., Lv,Y., Miao,Y.R. and Guo,A.Y. (2018) SEGreg: a database for human specifically expressed genes and their regulations in cancer and normal tissue. *Brief Bioinform.*, doi: 10.1093/bib/bbx1173.
31. Deng,W., Ma,L., Zhang,Y., Zhou,J., Wang,Y., Liu,Z. and Xue,Y. (2018) THANATOS: an integrative data resource of proteins and post-translational modifications in the regulation of autophagy. *Autophagy*, **14**, 296–310.

Corresponding author: Zhang Zhang[1,2,3,4,*]
Co-corresponding authors: Wenming Zhao[1,2,3,4,*], Jingfa Xiao[1,2,3,*], Yiming Bao[1,2,3,4,*]

BIG DATA CENTER MEMBERS (Arranged by project role and then by contribution except for Team Leader (TL), as indicated)

BioCode: Fan Wang[1], Lili Hao[1,2] (TL)
BioProject, BioSample & GSA: Junwei Zhu[1,#], Tingting Chen[1,#], Sisi Zhang[1,#], Xu Chen[1,#], Bixia Tang[1,3], Qing Zhou[1,3], Zhonghuang Wang[1,3], Lili Dong[1], Yanqing Wang[1,#] (TL)
GWH: Yingke Ma[1,2,#], Fan Wang[1,2], Zhewen Zhang[1,2], Zhonghuang Wang[1,2,3], Meili Chen[1,2,#] (TL)
GVM: Dongmei Tian[1,#], Cuiping Li[1,#], Lili Dong[1,#], Xufei Teng[1,2,3,#], Bixia Tang[1,3,#], Zhenglin Du[1], Na Yuan[1], Jingyao Zeng[1], Zhewen Zhang[1,2], Jinyue Wang[1,2,3], Shuo Shi[1,2,3], Yadong Zhang[1,2,3], Qi Wang[1,2,3], Mengyu Pan[1,2,3], Qiheng Qian[1,2,3], Shuhui Song[1,2,#] (TL)
GEN & RNA Editing: Guangyi Niu[1,2,3,#], Man Li[1,2,3,#], Lin Xia[1,2,3,#], Dong Zou[1,2,#], Yuansheng Zhang[1,2,3], Jian Sang[1,2,3], Mengwei Li[1,2,3], Yang Zhang[1,2,3], Pei Wang[1,2,3], Fan Wang[1], Yadong Zhang[1,2,3], Qianwen Gao[1,2,3], Jingfa Xiao[1,2,3,4],Lili Hao[1,2] (TL)
MethBank: Fang Liang[1], Mengwei Li[1,2,3], Dong Zou[1,2], Rujiao Li[1,2] (TL)
Science Wikis: Lin Liu[1,2,3,#], Jiabao Cao[1,2,3,#], Jian Sang[1,2,3,#], Dong Zou[1,2,#], Mengwei Li[1,2,3], Amir A. Abbasi[6], Huma Shireen[6], Pei Wang[1,2,3], Yang Zhang[1,2,3], Zhao Li[1,2,3], Qi Wang[1,2,3], Lin Xia[1,2,3], Zhuang Xiong[1,2,3], Meiye Jiang[1,2,3], Tongkun Guo[1,2,3], Zhaohua Li[1,2,3,4], Hao Zhang[1,2,3], Mengyu Pan[1,2,3], Lina Ma[1,2,#] (TL)
EWAS Atlas: Mengwei Li[1,2,3,#], Guangyi Niu[1,2,3], Lin Xia[1,2,3], Dong Zou[1,2], Yuansheng Zhang[1,2,3], Jian Sang[1,2,3], Zhaohua Li[1,2,3,4], Ran Gao[3,5], Rujiao Li[1,2], Tao Zhang[1,2,3], Yiming Bao[1,2,3,4], Zhang Zhang[1,2,3,4] (TL)
iDog: Bixia Tang[1,3,#], Qing Zhou[1,3,#], Lili Dong[1,#], Wulue Li[7], Xiangquan Zhang[7], Li Lan[1], Shuang Zhai[1], Yiming Bao[1,2,3], Yaping Zhang[7], Guodong Wang[7] (TL), Wenming Zhao[1,3,4,*] (TL)
IC4R: Jian Sang[1,2,3,#], Zhennan Wang[3,8,#], Dong Zou[1,2,#], Yuansheng Zhang[1,2,3], Lili Hao[1,2,#] (TL)
BHBD: Yiming Bao[1,2,3,4], Zhang Zhang[1,2,3,4], Wenming Zhao[1,2,3,4], Jingfa Xiao[1,2,3], Li Lan[1], Yongbiao Xue[3,4,5] (Project Leader)
Hardware & System Administration: Yubin Sun[1], Lei Yu[1], Shuang Zhai[1], Mingyuan Sun[1], Huanxin Chen[1] (TL)
Writing Group: Zhang Zhang[1,2,3,4,*], Wenming Zhao[1,2,3,4,*], Jingfa Xiao[1,2,3,*], Yiming Bao[1,2,3,4,*], Shuhui Song[1], Lili Hao[1,2], Rujiao Li[1], Lina Ma[1,2], Yanqing Wang[1], Bixia Tang[1,3], Meili Chen[1,2]

BIG DATA CENTER PARTNERS (Listed in alphabetical order by database resources)

AnimalTFDB: Hui Hu[9], An-Yuan Guo[9]
dbPAF & WERAM: Shaofeng Lin[9], Yu Xue[9]
dbPPT: Chenwei Wang[9], Yu Xue[9]
dbPSP: Wanshan Ning[9], Yu Xue[9]
CGDB: Ying Zhang[9], Yu Xue[9]
DEG & DoriC: Hao Luo[10,11,12], Feng Gao[10,11,12]
EKPD: Yaping Guo[9], Yu Xue[9]
hTFtarget: Qiong Zhang[7], An-yuan Guo[7]
iUUCD: Jiaqi Zhou[9], Yu Xue[9]
LncRNADisease: Zhou Huang[13], Qinghua Cui[13]
lncRNASNP: Ya-Ru Miao[9], An-Yuan Guo[9]
MiCroKiTS: Chen Ruan[9], Yu Xue[9]
PceRBase: Chunhui Yuan[14], Ming Chen[14]
PlantTFDB: Jin Jinpu[15], Ge Gao[15]
PLMD: Haodong Xu[7], Yu Xue[7]
RhesusBase: Yumei Li[16], Chuan-Yun Li[16]
SEGreg: Qing Tang[9], An-Yuan Guo[9]

THANATOS: Di Peng[9], Wankun Deng[9]
[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]School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[5]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[6]National Center for Bioinformatics, Programme of Comparative and Evolutionary Genomics, Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[7]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
[8]State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[9]Department of Bioinformatics and Systems Biology, Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[10]Department of Physics, Tianjin University, Tianjin 300072, China
[11]Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China
[12]SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin University, Tianjin 300072, China
[13]Department of Biomedical Informatics, MOE Key Laboratory of Cardiovascular Sciences, School of Basic Medical Sciences, Peking University, Beijing 100191, China
[14]Department of Bioinformatics, State Key Laboratory of Plant Physiology and Biochemistry, Institute of Plant Science, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
[15]Center for Bioinformatics, Peking University, Beijing 100871, China
[16]Institute of Molecular Medicine, Peking University, Beijing 100871, China
[*]To whom correspondence should be addressed: Zhang Zhang (zhangzhang@big.ac.cn). Correspondence may also be addressed to Wenming Zhao (zhaowm@big.ac.cn), Jingfa Xiao (xiaojingfa@big.ac.cn) and Yiming Bao (baoym@big.ac.cn).
[#]The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.