Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa1022

Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021

CNCB-NGDC Members and Partners[*,†]

Received September 14, 2020; Revised October 13, 2020; Editorial Decision October 14, 2020; Accepted October 16, 2020

ABSTRACT

The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB), provides a suite of database resources to support worldwide research activities in both academia and industry. With the explosive growth of multi-omics data, CNCB-NGDC is continually expanding, updating and enriching its core database resources through big data deposition, integration and translation. In the past year, considerable efforts have been devoted to 2019nCoVR, a newly established resource providing a global landscape of SARS-CoV-2 genomic sequences, variants, and haplotypes, as well as Aging Atlas, BrainBase, GTDB (Glycosyltransferases Database), LncExpDB, and TransCirc (Translation potential for circular RNAs). Meanwhile, a series of resources have been updated and improved, including BioProject, BioSample, GWH (Genome Warehouse), GVM (Genome Variation Map), GEN (Gene Expression Nebulas) as well as several biodiversity and plant resources. Particularly, BIG Search, a scalable, one-stop, cross-database search engine, has been significantly updated by providing easy access to a large number of internal and external biological resources from CNCB-NGDC, our partners, EBI and NCBI. All of these resources along with their services are publicly accessible at https://bigd.big.ac.cn.

INTRODUCTION

The National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB) officially founded in November 2019, was built based on the BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (CAS), with joint efforts and collaborations from two CAS institutions, viz., Institute of Biophysics (IBP) and Shanghai Institute of Nutrition and Health (SINH) as well as several partners (https://bigd.big.ac.cn/partners). Powered by higher-throughput and lower-cost genomics sequencing technologies, large-scale sequencing projects for precision medicine and biodiversity studies have been conducted around the world, leading to large amounts of multi-omics data that are still generated at ever-growing rates and scales. Therefore, CNCB-NGDC is dedicated to advancing life and health sciences by providing open access to a suite of data resources and services in support of global research activities on big data archive, storage, management and public sharing as well as multi-disciplinary data-driven research (1–4).

During the past year of 2020, an ongoing pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in more than 27 million infected cases and 897 000 deaths (as of 9 September 2020). To provide SARS-CoV-2 genome sequences and variants publicly available for the global research community (5), in the past year, CNCB-NGDC has made considerable efforts to build a SARS-CoV-2 information resource (6) by genomic data collection, curation and deep-mining with extensive updates on a daily basis. Additionally, CNCB-NGDC has continued to expand and update other resources through data deposition, integration and curation. In terms of database property, database resources of CNCB-NGDC can be generally grouped into three layers: Data—raw data

[*]To whom correspondence should be addressed. Tel: +86 10 84097261; Fax: +86 10 84097720; Email: ybxue@big.ac.cn
Correspondence may also be addressed to Yiming Bao. Email: baoym@big.ac.cn
Correspondence may also be addressed to Zhang Zhang. 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 Shunmin He. Email: heshunmin@ibp.ac.cn
Correspondence may also be addressed to Guoqing Zhang. Email: gqzhang@picb.ac.cn
Correspondence may also be addressed to Yixue Li. Email: yxli@sibs.ac.cn
Correspondence may also be addressed to Guoping Zhao. Email: gpzhao@sibs.ac.cn
Correspondence may also be addressed to Runsheng Chen. Email: crs@ibp.ac.cn
[†]Full list is provided in the Appendix.

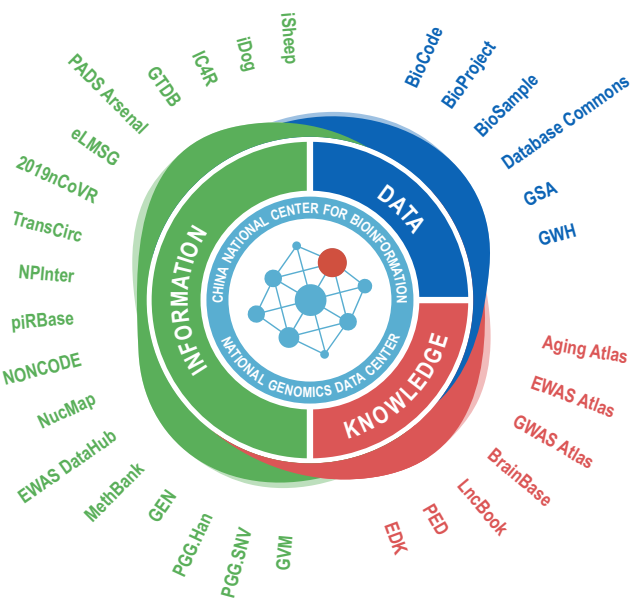

Figure 1. Core data resources of CNCB-NGDC. Three categories, viz., data, information and knowledge, are adopted to represent resources that are typically to deposit raw data/metadata (archives), house value-added information (databases) and integrate validated knowledge through literature curation (knowledgebases), respectively. A full list of data resources, which contains links to each resource, is available at https://bigd.big.ac.cn/databases.

and affiliated metadata, Information—standardized information and analyzed results, and Knowledge—curated associations and value-added knowledge. Here we provide a brief overview of new databases and recent updates to existing databases in CNCB-NGDC and describe its core resources and services (Figure 1). All these resources, along with their services, are publicly accessible through the home page of CNCB-NGDC at https://bigd.big.ac.cn.

NEW DATABASES

2019nCoVR

The 2019 Novel Coronavirus Resource (2019nCoVR, https://bigd.big.ac.cn/ncov/) (6) is an open-accessed SARS-CoV-2 information resource. It contains a comprehensive collection of genome sequences and clinical information for all publicly available SARS-CoV-2 isolates, which are manually curated with quality evaluation and value-added annotations by our in-house automated pipeline. Consequently, it houses a dynamic landscape of SARS-CoV-2 genomic variants and haplotypes on a global scale. Specifically, 2019nCoVR identifies all variants from complete and high-quality genomes, visualizes the spatiotemporal change for each variant, and constructs haplotype network maps and phylogenetic trees for the course of the outbreak. Moreover, 2019nCoVR offers a set of online tools covering various needs for SARS-CoV-2 genomic data analysis. In addition, it provides a full collection of literatures on COVID-19, including published papers from PubMed as well as preprints from bioRxiv and medRxiv through Europe PMC. Collectively, all SARS-CoV-2 genome sequences, variants, haplotypes and literatures are integrated and updated daily since January 2020, making 2019nCoVR a valuable resource for the global research community.

Aging Atlas

The Aging Atlas (https://bigd.big.ac.cn/aging; detailed in (7) in this issue) is an integrative database in support of aging research. It provides open access to large-scale multi-omics datasets generated by a variety of high-throughput sequencing technologies, involving genomics, epigenomics, transcriptomics, proteomics, metabolomics, pharmacogenomics and single-cell omics. The current implementation includes five modules: RNA sequencing, epigenomic regulation, single-cell sequencing, protein interactions and geroprotective compounds.

BrainBase

BrainBase (https://bigd.big.ac.cn/brainbase) is a curated knowledgebase for brain diseases. Based on manual curation of published articles and related databases, BrainBase features comprehensive integration of disease associations from multiple omics levels and its current version houses a total of 4248 associations covering 113 brain diseases and 3996 genes/CpG sites. In addition, based on bioinformatic analysis on expression datasets, BrainBase collects 655 brain-specific genes, 575 brain-region-specific genes and 1128 cerebrospinal fluid (CSF)-detectable genes. With a particular focus on glioma, BrainBase integrates 22 glioma-related omics datasets (genome, transcriptome, epigenome and proteome) and provides multi-omics molecular profiles for glioma, which are of great utility to identify potential biomarkers for glioma diagnosis, prognosis and treatment prediction. Thus, BrainBase bears great promise to serve as a valuable knowledgebase for brain studies.

CGIR

The Chloroplast Genome Information Resource (CGIR; http://bigd.big.ac.cn/cgir) is a curated resource of chloroplast genome information through comprehensive integration and value-added annotation. The current release of CGIR contains 4709 chloroplast genomes of 4485 species; 4290 are retrieved from NCBI, and the rest 419 are from CNCB-NGDC Genome Warehouse, among which 403 genome assemblies of 247 species are sequenced by National Resource Center for Chinese Materia Medica and publicly released for the first time. Based on expert curation, we standardize taxonomic classification for each chloroplast (including families, genera, and species) and present a comprehensive high-quality collection of chloroplast genomes that belong to 1887 genera and 441 families and cover 1165 featured plants with one or more associated category (namely, medicinal, edible, energy and wood). Considering the importance of photosynthesis, we further investigate presence/absence variation (PAV) of photosynthesis genes among all collected genomes and detect the strength of selective pressure acting on photosynthesis genes by comparing nonsynonymous and synonymous substitution rates. Moreover, we identify potential molecular markers for all collected assemblies and obtain a total of 120,152 DNA sequence signatures (DSSs) and 1 770

546 simple sequence repeats (SSRs), which are of broad utility to identify species in Chinese Pharmacopoeia (2020 edition), and to develop SNP markers and PCR methods for species identification. In conclusion, CGIR is capable to help users easily access chloroplast genome information.

GTDB

The Glycosyltransferases Database (GTDB; https://www.biosino.org/gtdb/) (8) is an integrated resource for glycosyltransferase annotations, incorporating comprehensive information of protein classification families, catalytic reactions and metabolic pathways, etc. In the current version, GTDB contains 520 179 glycosyltransferases from 21 647 taxonomy nodes and 394 kinds of enzymatic reactions. In addition, GTDB provides: (i) a powerful search to retrieve the complete details of a query by combining multiple identifiers and data sources; (ii) an interactive browser to visualize data by different classifications and download data in batches; (iii) a BLAST tool (9) to search against pre-defined sequences, facilitating the annotation of biological function of glycosyltransferases and lastly, (iv) GTdock (8), which uses AutoDock Vina to perform docking simulations of several glycosyltransferases with the same single acceptor.

LncExpDB

LncExpDB (https://bigd.big.ac.cn/lncexpdb; detailed in (10) in this issue) is an expression database of human long non-coding RNAs (lncRNAs). Based on our previous work on LncBook (11), LncExpDB houses abundant expression profiles of 101 293 non-redundant, manually-curated lncRNA genes across 337 biological conditions, which can be further classified into nine important biological contexts, namely, normal tissue/cell line, cancer cell line, subcellular localization, exosome, cell differentiation, preimplantation embryo, organ development, circadian rhythm, and virus infection. Among them, 92 016 lncRNA genes (90.8%) are supported with reliable transcriptional evidence and more than one third of lncRNAs (31249) have the capacity to be highly expressed under certain conditions. Most importantly, LncExpDB provides a collection of featured lncRNAs and their interacting partners and thus is of great significance to help users conduct functional studies on lncRNAs.

scMethBank

Single-cell bisulfite sequencing methods are widely used to assay epigenomic heterogeneity in cell states. Large amounts of data have been generated over the past several years, bearing great promises in deeper understanding of the epigenetic regulation of key biological processes. scMethBank (https://bigd.big.ac.cn/methbank/scm) is an integrated database of single-cell methylation maps. It is dedicated to the collection, integration, analysis and visualization of single-cell methylation data and metadata. The current release of scMethBank includes 3166 single-cell methylation profiles as well as curated metadata, covering two species (human and mouse), 14 projects, 26 cell types and two diseases, and provides user-friendly web interfaces for data browsing, search and download.

TransCirc

TransCirc (https://www.biosino.org/transcirc/) is a specialized database that provides evidence of translation potential for circular RNAs (circRNAs) (detailed in (12) in this issue). It integrates seven types of direct and indirect evidence of coding potential for human circRNAs and their putative translation products, including ribosome/polysome binding evidence, internal ribosomal entry sites, *N*-6-methyladenosine modification data, sequence composition scores, mass spectrometry data, etc. TransCirc can serve as an important resource for investigating the translation capacity of circRNAs and will be expanded to add new evidence or additional species in the future.

UPDATES TO EXISTING DATABASES

BioProject & BioSample

BioProject (https://bigd.big.ac.cn/bioproject) and BioSample (https://bigd.big.ac.cn/biosample) are two public repositories of biological research projects and samples, respectively. They collect descriptive metadata on biological projects and samples investigated in experiments, and provide centralized accesses to all public projects and samples, as well as cross links to their related data resources. BioProject organizes and classifies a huge volume of projects in terms of various data types, ranging from genomic, transcriptomic, epigenomic and metagenomic sequencing efforts to genome-wide association studies and variation analyses. BioSample supports a wide scope of sample types, including human, plant, animal, microbe, virus, pathogen and metagenome. Up to August 2020, there are a total of 2288 biological projects and 176 288 biological samples submitted by 1341 users from 364 organizations (Figure 2A).

Genome Sequence Archive

The Genome Sequence Archive (GSA; https://bigd.big.ac.cn/gsa) (13) is a public data repository for archiving raw sequence reads. GSA accepts multi-omics data submissions from all over the world and provides free access to all publicly available data for global scientific communities. In April 2020, GSA-Human (https://bigd.big.ac.cn/gsa-human), a sub-database of GSA, was further established, with the specific aim to provide a set of services for secure management of human genetic data with controlled access. Particularly, any data submission to GSA-Human is affiliated with a Data Administration Committee (DAC) that is responsible for authorizing/declining data access to data requestor. As of August 2020, GSA (together with GSA-Human), has archived a total of 181 123 experiments and 198 262 runs and housed >4600 Terabytes of sequencing data (Figure 2B), exhibiting the nearly quadruple volume compared to the previous release last August (~1200 TB).

Genome Warehouse

The Genome Warehouse (GWH; https://bigd.big.ac.cn/gwh) is a public resource archiving genome-scale data of

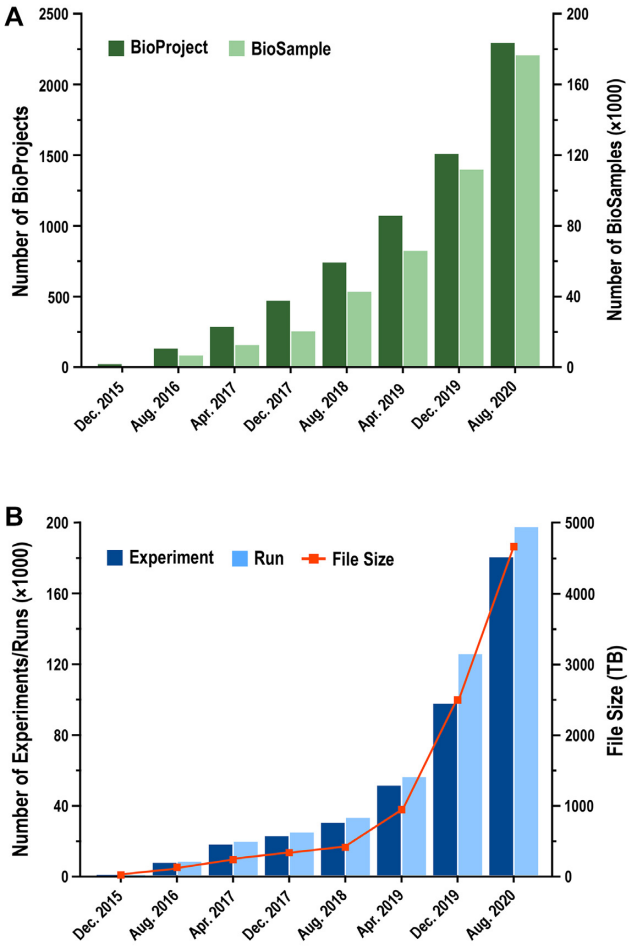

Figure 2. Statistics of data submissions to BioProject, BioSample and GSA. (**A**) Data statistics of BioProject and BioSample. (**B**) Data statistics of Experiments and Runs as well as file size in GSA. All statistics are frequently updated and publicly available at https://bigd.big.ac.cn/bioproject,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

a wide range of species. GWH accepts worldwide submissions of genome assemblies and incorporates detailed descriptive information for each assembly. It offers standardized quality control for genome assembly and equips with a genome browser (14) for genome visualization. By August 2020, GWH has received 9337 direct submissions covering a broad diversity of species. Among them, 1491 genome assemblies have been publicly released and reported in 52 journal articles. Particularly, in collaboration with 2019nCoVR, GWH has received the submission of 815 SARS-CoV-2 genome assemblies with standardized genome annotations (15). So far, 78 of the genomes have been publicly released and 25 have been shared, with the submitters' permission, in GenBank (16) through a data exchange mechanism established with NCBI. In this model, GWH accessions are represented as secondary accessions in GenBank records, which are retrievable by the Entrez system. Collectively, the rapid growth of genome-scale data submissions demonstrates the great potential of GWH as an important resource for accelerating the worldwide genomic research.

Genome Variation Map

The Genome Variation Map (GVM; https://bigd.big.ac.cn/gvm) (17) is a public repository of genome variations, including single nucleotide polymorphisms (SNP) and small insertions and deletions (indel). Unlike NCBI dbSNP (dedicated only for human genome variations since September 2017), GVM features data collection for a wide range of species and accepts data submissions from all over the world. During the past year, GVM has been significantly updated by reorganizing data entities and metadata into six modules in terms of species, project, sample, variation, association, and submission. In addition, it has received 56 genome variation data submissions involving 43 754 samples from 26 species. Till August 2020, GVM houses a total of ∼960 million variants derived from 191 projects and 64 820 samples and covering 13 animals, 25 plants and 3 viruses.

GWAS Atlas

GWAS Atlas (https://bigd.big.ac.cn/gwas) (18) is a curated resource of genome-wide variant-trait associations in plants and animals. In the current version, GWAS Atlas has been updated by integrating 78 950 associations across seven cultivated plants and five domesticated animals that were manually curated from 1088 studies in 304 publications. As a result, a total of 31 684 genes and 735 traits were annotated and presented based on a set of ontologies. Together, GWAS Atlas provides high-quality curated GWAS associations for plants and animals, and accordingly serves as a valuable resource for genetic research of important traits and breeding application.

Gene Expression Nebulas

The Gene Expression Nebulas (GEN) (https://bigd.big.ac.cn/gen/) is a comprehensive data portal of gene expression profiles across various biological conditions. Based on a set of ontologies on disease, tissue and cell type, GEN integrates large-scale publicly available bulk and single-cell RNA sequencing datasets with strict criteria from raw sequence repositories such as CNCB-NGDC GSA (13) and NCBI SRA (19). All high-quality sequencing data are processed with standardized pipeline and manually curated based on meta information from GSA, NCBI GEO (20) as well as publications. In the current version, GEN has integrated human expression profiles across 25 631 experiments and 99 tissues from 141 studies, including 22 128 single-cell experiments that cover 410 149 cells in 31 diseases and 47 development stages. In addition, GEN has also integrated plant expression profiles in 35 organs from 50 studies, including 945 experiments for rice, 506 for soybean, 462 for sorghum and 78 for wheat, respectively. GEN provides convenient and user-friendly web interfaces for data browsing, search, visualization and batch downloading, and also

equips with a suite of analysis tools for differential gene expression, functional enrichment, regulatory network, and cell type annotation.

Editome Disease Knowledgebase

Editome Disease Knowledgebase (EDK; http://bigd.big.ac.cn/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 (21). In the past year, the curated associations in EDK have been updated, including 36 diseases associated with 582 experimentally validated abnormal editing events in 143 messenger RNAs, 4 microRNAs, 47 viruses and 79 aberrant activities involved in three editing enzyme families. Moreover, based on controlled vocabulary for viral classification, EDK has integrated virus-RNA editing disease associations from more than 200 publications.

NONCODE

NONCODE is a comprehensive database that hosts the most complete collection of noncoding RNAs and their annotations (22). Particularly, it is dedicated to providing the full landscape of long non-coding RNAs (lncRNAs). In the current version (v6), lncRNAs in human and mouse were greatly updated, and the number of lncRNAs has been increased from 548 640 to 644 509. Moreover, NONCODE summarized a total of 13 749 lncRNA-cancer associations from public databases and literature. For plants, NONCODE housed a set of 94 697 lncRNAs and also introduced two important new features: (i) tissue expression profiles and function prediction of lncRNAs in five common plants; (ii) conservation annotation of lncRNAs for 23 plants. Collectively, NONCODE is a comprehensive portal of lncRNAs for both plants and animals and is freely available at http://v6.noncode.org/.

SmProt

SmProt is a dedicated database that provides the scientific community with valuable information about small proteins (23). Here, we introduce the update of SmProt, which emphasizes the reliability of the translated sORF, the genetic variation in the translated sORF, the translation event or sequence of the disease-specific sORF, and the significant increase in data volume. The updated SmProt also includes more components, such as non-AUG translation initiation, functions and new resources. Totally, the current version of SmProt incorporated 802,906 unique small proteins curated from 3 695 141 primary records. These proteins were calculated from 419 Ribo-seq data sets and collected from literature and other sources, including 370 cell lines or tissues of 8 species (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila*, *Danio rerio*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Escherichia coli*). In addition, small protein families identified from human microbiomes were also collected. All datasets in SmProt are publicly available for browse, search and bulk downloads at http://bigdata.ibp.ac.cn/SmProt/.

MethBank

The Methylation Bank (MethBank; http://bigd.big.ac.cn/methbank) (24,25) is a comprehensive database that integrates consensus reference methylomes (CRMs) and single-base resolution methylomes (SRMs) across a variety of species, with a particular focus on human health and aging, animal embryonic development, and plant growth and development. In the current version, MethBank presents 163 CRMs and 5 687 344 methylation profiles of corresponding genes from 80 normal tissues/cells of human (deduced from 22,775 publicly available DNA methylation 450K data). In addition to CRMs, it provides 394 SRMs, 19 701 343 methylation profiles of genes, 1 258 420 methylated CpG Islands and 304 884 differentially methylated promoters in different genomic contexts based on whole-genome bisulfite sequencing data from normal human tissues, different developmental stages in five economically important plants, and multi-stage gametes and early embryos in two model animals. Moreover, MethBank is armed with online tools to predict human methylation age and identify differentially methylated promoters via Fisher's exact test (26) with FDR correction. In addition, MethBank provides useful information on 421 methylation data analysis tools, helpful for users to easily find any tool of interest.

EWAS Atlas

EWAS Atlas (https://bigd.big.ac.cn/ewas) (27) is a curated knowledgebase of epigenome-wide association studies. In the past year, it has been enriched by adding a total of 126 393 EWAS associations manually curated from 324 publications. Taking advantage of massive high-quality DNA methylation data, EWAS toolkit (https://bigd.big.ac.cn/ewas/toolkit), was greatly enhanced for a wide range of EWAS analyses (trait enrichment, GO enrichment, motif analysis, chromatin enrichment, etc.). Till August 2020, EWAS Atlas has integrated 577 267 high-quality EWAS associations derived from 1216 studies in 725 publications, including 3124 cohorts, 155 tissues/cell lines, 498 traits and 435 ontology entities. As a data portal of EWAS Atlas, EWAS Data Hub (https://bigd.big.ac.cn/ewas/datahub) (28) houses 95 783 samples of standardized DNA methylation array data and metadata, and provides DNA methylation profiles for a list of 485 512 probes in association with 36 397 genes.

Biodiversity Resources

Biodiversity resources are dedicated for specific species, including economically important crops, domesticated animals and livestock. Currently, there are four major biodiversity resources in CNCB-NGDC, namely, iDog, iSheep, Information Commons for Rice (IC4R) and SorGVD. iDog (https://bigd.big.ac.cn/idog) is an integrated omics data resource for dog, including eight data modules and one analysis module (29). As a dedicated resource for the ongoing Dog10K Project (30), iDog has been considerably updated by integrating more data and deploying new online tools. In the current version, iDog mainly houses two *de novo* assembly genomes, 42 871 184 non-redundant SNPs from 127 samples, 783 curated diseases, 473 standardized

breeds for phenotype traits, 594 genotype-to-phenotype (G2P) pairs and 27 534 gene profiles from public RNA-seq projects. iSheep (https://bigd.big.ac.cn/isheep) is a specialized resource dedicated to integrating omics data for sheep. Currently, it contains 82 689 498 genomic variations (including 70 370 968 SNPs and 12 318 530 Indels) from 2778 samples, 26 802 genes and 1417 breed information of worldwide sheep. Moreover, it includes 922 genome-wide variant-trait associations linked with 922 variants and 110 traits. IC4R (http://ic4r.org) (31,32) is a curated database that provides rice genome sequences, gene annotations and multi-omics data profiles. It was updated by incorporating a new gene annotation system with improved gene structure and completeness (33). Meanwhile, SnpReady for Rice (SR4R; http://sr4r.ic4r.org) (34), a committed sub-database of IC4R, was built based on a collection of 18 million SNPs identified from 5152 rice accessions. Accordingly, SR4R delivers four reference SNP panels (2 097 405 hapmapSNPs, 156 502 tagSNPs, 1180 fixedSNPs and 38 barcodeSNPs), offering a highly efficient rice variation map for different needs. SorGVD (https://bigd.big.ac.cn/sorgvd) (35), is a comprehensive database for sorghum genomic variations and phenotypes. The updated version of SorGSD provides curated information of 39 547 621 genomic variations (including 33 825 236 SNPs and 5 722 385 small INDELs) from resequencing data and phenotypes of 289 sorghum accessions.

Plant Resources

Plants are the basis of our Earth's ecosystems, providing the world's molecular oxygen and serving as basic human foods and medicines. Currently, CNCB-NGDC has two major resources developed from different aspects, viz., Plant Editosome Database (PED) and Leaf Senescence Database (LSD). PED (https://bigd.big.ac.cn/ped) (36) is a curated database of plant RNA editing factors. In the past year, it has been updated by integrating 94 RNA editing factors, 78 edited genes, and 1,796 RNA editing events from 34 organelles of 29 species manually curated from 39 publications. Most editing factors and genes are related to plant growth and development, among which 43 RNA editing factors and 7 edited genes are newly added in PED. LSD (https://bigd.big.ac.cn/lsd) (37) is a comprehensive database for the leaf senescence research community. It currently incorporates 5853 senescence-associated genes and 617 mutants from 68 species.

Database Commons

Database Commons (https://bigd.big.ac.cn/databasecommons) is a catalogue of worldwide biological databases. It provides easy access to a global landscape of all publicly available databases and their descriptive metadata manually curated from their publications. Currently, it catalogues a total of 5064 databases, involving 7595 publications and 1944 organizations throughout the world. In the past year, in addition to more database entries and publications, web interfaces have been greatly improved, allowing users to access and browse databases by country, institution, category, data type and object. Furthermore, powered by Europe PMC APIs (38), citations to all collected databases are added in an automated manner and updated weekly. To promote the incorporation of more databases and indexed data, Database Commons is open to accept data entry from the global research community.

BIG Search

BIG Search (https://bigd.big.ac.cn/search) is a distributed and scalable full-text search engine built on Elasticsearch (a highly scalable search and analytics engine, https://www.elastic.co/). It features cross-database search and provides uniform interfaces for retrieving information from a wide range of biological databases in real-time. In the current version, BIG Search has been significantly updated by incorporating data indexes from internal and external biological resources, including all resources in CNCB-NGDC and 38 partner resources (see details at https://bigd.big.ac.cn/partners). Followed by the integration of EBI resources using the EBI Search RESTful API (39) last year, NCBI resources were added to BIG Search powered by NCBI Entrez (40). In summary, BIG Search offers easy access to a large number of biological resources and provides one-stop cross-database search services for the global research community.

Education

The interdisciplinary nature of bioinformatics, coupled with rapid advances in genomics, artificial intelligence and data science, has made bioinformatics an increasingly data-intensive and data-driven field, bearing great promise to translate big data into big discovery in life and health sciences. To provide bioinformatics education services to our users, this year we established our online education platform (https://bigd.big.ac.cn/education/) that provides a series of educational materials including online courses, tutorials and training documents. As a starting point, we currently offered two courses (Bioinformatics and Genomics) and online tutorials for briefly introducing our core databases and services. In addition, we delivered training offerings nationally and internationally, particularly in coordination with the Global Biodiversity and Health Big Data (BHBD) Alliance. Over the past year, we have conducted training and outreach programs for international researchers in China and over 100 people in Pakistan. We plan to establish worldwide collaborations with peers who have common interests in developing and enriching our educational materials and contents.

CONCLUDING REMARKS

The year of 2020 was very special. For one thing, CNCB-NGDC has been significantly reinforced by joint efforts from BIG, IBP and SINH, close collaborations from our partners, and long-term, continuous support from the whole research community. For another, to deal with the pandemic caused by SARS-CoV-2, CNCB-NGDC has developed 2019nCoVR, a SARS-CoV-2 information resource, with daily updates on data integration, curation, and analysis. More importantly, the COVID-19 outbreak accelerated

our collaboration in data sharing with the INSDC through SARS-CoV-2 genome sequence exchange with NCBI. We will be using this model to expand data sharing to genome sequences of other organisms and other data types. Meanwhile, growth of multi-omics data, particularly in human, is explosive. Consequently, database resources of CNCB-NGDC have been enriched and updated by accepting data submissions from all over the world, performing value-added curation and annotation and also improving web interfaces and data services.

Ongoing efforts include, but not limited to, optimization of curation models and processes, improvement of web functionalities and database usage statistics, upgrade of infrastructure capability for big data storage and transfer, integration of more datasets from different resources, and continuous development of new resources and tools in aid of data-driven studies. We will also put in more efforts to establish and improve underlying links between our database resources, with the aim to fully realize the findability, accessibility, interoperability and reusability (FAIR) of different levels of data. In addition, CNCB-NGDC heavily engages in the BHBD Alliance (http://bhbd-alliance.org) in order to accelerate the translation of big data into knowledge discovery by global collaborations in data sharing and mining. With more stable support, CNCB-NGDC will continue to grow and deliver a family of data resources and services in support of both domestic and international research activities.

ACKNOWLEDGEMENTS

We thank our users for submitting data, sending suggestions, reporting bugs and getting involving in community curation. CNCB-NGDC is indebted to its funders, including the Ministry of Science & Technology and the Ministry of Finance of the People's Republic of China as well as Chinese Academy of Sciences. We also thank the whole bioinformatics community in China, particularly the late Prof. Bailin Hao, who advocated the establishment of CNCB since the 1990s.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030200, XDA19050302, XDA19090116, XDA24040201, XDB38050300, XDB38030100, XDB38030400, XDA12030100, XDB38040300]; National Key Research & Development Program of China [2019YFA0801801, 2018YFA0801405, 2018YFD1000505, 2018YFC2000100, 2018YFC1406902, 2018YFC0910400, 2018YFC0310602, 2018YFA0903700, 2018YFA0900704, 2017YFC1201200, 2017YFC0908405, 2017YFC0908404, 2017YFC0908403, 2017YFC0907505, 2017YFC0907503, 2017YFC0907502, 2016YFE0206600, 2016YFC0906403, 2016YFC0903003, 2016YFC0901904, 2016YFC0901903, 2016YFC0901702, 2016YFC0901604, 2016YFC0901603, 2016YFB0201702, 2016YFA0501704]; National Natural Science Foundation of China [91731303, 81670462, 31970565, 31871328, 31871294, 31701117, 31970647, 31801104, 31771465, 31771410, 31771388, 31671360, 81701567, 31571358, 31525014, 1470330, 31961130380, 31711530221, 31771477, 31571366, 31822030, 31801113, 31801154, 31771458, 91940303, 91940306, 31661143031, 31730110, 31871281, 31970634, 31930021, 31970633]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008, 153D31KYSB20170121]; 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05]; Genomics Data Center Construction of Chinese Academy of Sciences [XXH-13514-0202]; Fundamental Research Funds for the Central Universities [2019kfyRCPY043]; UK Royal Society-Newton Advanced Fellowship [NAF\R1\191094]; Key Program of the Chinese Academy of Sciences [KJZD-EW-L14]; Key Research Program of Frontier Sciences of the Chinese Academy of Sciences [QYZDJ-SSW-SYS009]; Key Technology Talent Program of the Chinese Academy of Sciences; The 100 Talent Program of the Chinese Academy of Sciences; K.C. Wong Education Foundation; The Youth Innovation Promotion Association of the Chinese Academy of Sciences [2019104, 2018134, 2017141]; The Special Project on Precision Medicine under the National Key R&D Program [SQ2017YFSF090210]; China Postdoctoral Science Foundation [2019M652623, 2018M632830]; The Open Biodiversity and Health Big Data Program of IUBS; The Professional Association of the Alliance of International Science Organizations [ANSO-PA-2020-07]; Funds for Basic Resources Investigation Research of the Ministry of Science and Technology [2018FY10080002]; Special Project on National Science and Technology Basic Resources Investigation [2019FY100102]; CAS Pioneer 100-Talent program; Key Research Program of the Chinese Academy of Sciences [KFZD-SW-219-5]; Zhangjiang special project of national innovation demonstration zone [ZJ2018-ZD-013]; Science and Technology Service Network Initiative of Chinese Academy of Sciences. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences.
Conflict of interest statement. None declared.

REFERENCES

1. National Genomics Data Center Members and Partners. (2020) Database Resources of the National Genomics Data Center in 2020. *Nucleic Acids Res.*, **48**, D24–D33.
2. BIG Data Center Members. (2019) Database Resources of the BIG Data Center in 2019. *Nucleic Acids Res.*, **47**, D8–D14.
3. BIG Data Center Members. (2018) Database Resources of the BIG Data Center in 2018. *Nucleic Acids Res.*, **46**, D14–D20.
4. BIG Data Center Members. (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, **45**, D18–D24.
5. Zhang,Z., Song,S., Yu,J., Zhao,W., Xiao,J. and Bao,Y. (2020) The Elements of Data Sharing. *Genomics Proteomics Bioinformatics*, **18**, 1–4.
6. Zhao,W.M., Song,S.H., Chen,M.L., Zou,D., Ma,L.N., Ma,Y.K., Li,R.J., Hao,L.L., Li,C.P., Tian,D.M. *et al.* (2020) The 2019 novel coronavirus resource. *Yi chuan = Hereditas / Zhongguo yi chuan xue hui bian ji*, **42**, 212–221.
7. Aging Atlas Consortium. (2021) Aging Atlas: a multi-omics database for aging biology. *Nucleic Acids Res.*, doi:10.1093/nar/gkaa894.
8. Zhou,C., Xu,Q., He,S., Ye,W., Cao,R., Wang,P., Ling,Y., Yan,X., Wang,Q. and Zhang,G. (2020) GTDB: an integrated resource for

glycosyltransferase sequences and annotations. *Database (Oxford)*, **2020**, baaa047.
9. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
10. Li,Z., Liu,L., Jiang,S., Li,Q., Feng,C., Du,Q., Zou,D., Xiao,J., Zhang,Z. and Ma,L. (2021) LncExpDB: an expression database of human long non-coding RNAs. *Nucleic. Acids. Res.*, doi:10.1093/nar/gkaa850.
11. Ma,L., Cao,J., Liu,L., Du,Q., Li,Z., Zou,D., Bajic,V.B. and Zhang,Z. (2019) LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic. Acids. Res.*, **47**, D128–D134.
12. Huang,W., Ling,Y., Zhang,S., Xia,Q., Cao,R., Fan,X., Fang,Z., Wang,Z. and Zhang,G. (2021) TransCirc: an interactive database for translatable circular RNAs based on multi-omics evidence. *Nucleic. Acids. Res.*, doi:10.1093/nar/gkaa823.
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. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
15. Ren,L.L., Wang,Y.M., Wu,Z.Q., Xiang,Z.C., Guo,L., Xu,T., Jiang,Y.Z., Xiong,Y., Li,Y.J., Li,X.W. *et al.* (2020) Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chin. Med. J. (Engl.)*, **133**, 1015–1024.
16. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
17. 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.
18. Tian,D., Wang,P., Tang,B., Teng,X., Li,C., Liu,X., Zou,D., Song,S. and Zhang,Z. (2020) GWAS Atlas: a curated resource of genome-wide variant-trait associations in plants and animals. *Nucleic Acids Res.*, **48**, D927–D932.
19. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, 19–21.
20. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
21. Niu,G., Zou,D., Li,M., Zhang,Y., Sang,J., Xia,L., Li,M., Liu,L., Cao,J., Zhang,Y. *et al.* (2019) Editome Disease Knowledgebase (EDK): a curated knowledgebase of editome-disease associations in human. *Nucleic Acids Res.*, **47**, D78–D83.
22. Fang,S., Zhang,L., Guo,J., Niu,Y., Wu,Y., Li,H., Zhao,L., Li,X., Teng,X., Sun,X. *et al.* (2018) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.*, **46**, D308–D314.
23. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F. *et al.* (2018) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, **19**, 636–643.
24. Li,R., Liang,F., Li,M., Zou,D., Sun,S., Zhao,Y., Zhao,W., Bao,Y., Xiao,J. and Zhang,Z. (2018) MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Res.*, **46**, D288–D295.
25. Zou,D., Sun,S., Li,R., Liu,J., Zhang,J. and Zhang,Z. (2015) MethBank: a database integrating next-generation sequencing single-base-resolution DNA methylation programming data. *Nucleic Acids Res.*, **43**, D54–58.
26. Sprent,P. (2011) Fisher Exact Test. In: LovricM (ed). *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_253.
27. Li,M., Zou,D., Li,Z., Gao,R., Sang,J., Zhang,Y., Li,R., Xia,L., Zhang,T., Niu,G. *et al.* (2019) EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. *Nucleic Acids Res.*, **47**, D983–D988.
28. Xiong,Z., Li,M., Yang,F., Ma,Y., Sang,J., Li,R., Li,Z., Zhang,Z. and Bao,Y. (2020) EWAS Data Hub: a resource of DNA methylation array data and metadata. *Nucleic Acids Res.*, **48**, D890–D895.
29. Tang,B., Zhou,Q., Dong,L., Li,W., Zhang,X., Lan,L., Zhai,S., Xiao,J., Zhang,Z., Bao,Y. *et al.* (2019) iDog: an integrated resource for domestic dogs and wild canids. *Nucleic Acids Res.*, **47**, D793–D800.
30. Ostrander,E.A., Wang,G.D., Larson,G., vonHoldt,B.M., Davis,B.W., Jagannathan,V., Hitte,C., Wayne,R.K., Zhang,Y.P. and Dog,K.C. (2019) Dog10K: an international sequencing effort to advance studies of canine domestication, phenotypes and health. *Natl. Sci. Rev.*, **6**, 810–824.
31. IC4R Project Consortium. (2016) Information Commons for Rice (IC4R). *Nucleic Acids Res.*, **44**, D1172–D1180.
32. Xia,L., Zou,D., Sang,J., Xu,X., Yin,H., Li,M., Wu,S., Hu,S., Hao,L. and Zhang,Z. (2017) Rice Expression Database (RED): an integrated RNA-Seq-derived gene expression database for rice. *J Genet Genomics*, **44**, 235–241.
33. Sang,J., Zou,D., Wang,Z., Wang,F., Zhang,Y., Xia,L., Li,Z., Ma,L., Li,M., Xu,B. *et al.* (2020) IC4R-2.0: rice genome reannotation using massive RNA-seq data. *Genomics Proteomics Bioinformatics*, **18**, 161–172.
34. Yan,J., Zou,D., Li,C., Zhang,Z., Song,S. and Wang,X. (2020) SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genomics Proteomics Bioinformatics*, **18**, 173–185.
35. Luo,H., Zhao,W., Wang,Y., Xia,Y., Wu,X., Zhang,L., Tang,B., Zhu,J., Fang,L., Du,Z. *et al.* (2016) SorGSD: a sorghum genome SNP database. *Biotechnol. Biofuels*, **9**, 6.
36. Li,M., Xia,L., Zhang,Y., Niu,G., Li,M., Wang,P., Zhang,Y., Sang,J., Zou,D., Hu,S. *et al.* (2019) Plant editosome database: a curated database of RNA editosome in plants. *Nucleic. Acids. Res.*, **47**, D170–D174.
37. Li,Z., Zhang,Y., Zou,D., Zhao,Y., Wang,H.L., Zhang,Y., Xia,X., Luo,J., Guo,H. and Zhang,Z. (2020) LSD 3.0: a comprehensive resource for the leaf senescence research community. *Nucleic Acids Res.*, **48**, D1069–D1075.
38. Levchenko,M., Gou,Y., Graef,F., Hamelers,A., Huang,Z., Ide-Smith,M., Iyer,A., Kilian,O., Katuri,J., Kim,J.H. *et al.* (2018) Europe PMC in 2017. *Nucleic Acids Res.*, **46**, D1254–D1260.
39. Madeira,F., Park,Y.M., Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
40. Gibney,G. and Baxevanis,A.D. (2011) Searching NCBI Databases Using Entrez. *Curr. Protoc. Hum. Genet.*, doi:10.1002/0471142905.hg0610s71.

APPENDIX

Corresponding author: Yongbiao Xue[1,2,3,*]
Co-corresponding authors: Yiming Bao[1,2,3,4,*], Zhang Zhang[1,2,3,4,*], Wenming Zhao[1,2,3,4,*], Jingfa Xiao[1,2,3,4,*], Shunmin He[3,5,6,*], Guoqing Zhang[3,7,*], Yixue Li[3,7,*], Guoping Zhao[3,7,8,9,*], Runsheng Chen[6,10,*]
CNCB-NGDC MEMBERS (Arranged by project role and then by contribution except for Team Leader (TL), as indicated)
2019nCoVR: Shuhui Song[1,2,3,4,#], Lina Ma[1,2,4,#], Dong Zou[1,2,4,#], Dongmei Tian[1,2,4,#], Cuiping Li[1,2,4,#], Junwei Zhu[1,2,4,#], Zheng Gong[1,2,3,4,#], Meili Chen[1,2,4], Anke Wang[1,2,4], Yingke Ma[1,2,4], Mengwei Li[1,2,3,4], Xufei Teng[1,2,3,4], Ying Cui[1,2,3,4], Guangya Duan[1,2,3,4], Mochen Zhang[1,2,4,15], Tong Jin[1,2,3,4], Chengmin Shi[1,11], Zhenglin Du[1,2,4], Yadong Zhang[1,2,3,4], Chuandong Liu[1,11], Rujiao Li[1,2,4], Jingyao Zeng[1,2,4], Lili Hao[1,2,4], Shuai Jiang[1,2,4], Hua Chen[1,11], Dali Han[1,11], Jingfa Xiao[1,2,3,4], Zhang Zhang[1,2,3,4,*] (TL), Wenming Zhao[1,2,3,4,*] (TL), Yongbiao Xue[1,2,3,*] (TL), Yiming Bao[1,2,3,4,*] (TL)
Aging Atlas: Tao Zhang[1,2,3,4,#], Wang Kang[1,3,11,#], Fei Yang[1,2,3,4,#], Jing Qu[3,12,13], Weiqi Zhang[2,3,11,12,#] (TL), Yiming Bao[1,2,3,4,*] (TL), Guang-Hui Liu[3,12,14,#] (TL)

BrainBase: Lin Liu[1,2,3,4,#], Yang Zhang[1,2,3,4,#], Guangyi Niu[1,2,3,4,#], Tongtong Zhu[1,2,4,15], Changrui Feng[1,2,3,4], Xiaonan Liu[1,2,4,15], Yuansheng Zhang[1,2,3,4], Zhao Li[1,2,3,4], Ruru Chen[1,2,4,16], Qianpeng Li[1,2,3,4], Xufei Teng[1,2,3,4], Lina Ma[1,2,4,#] (TL)
CGIR: Zhongyi Hua[17,#], Dongmei Tian[1,2,4,#], Chao Jiang[17,#], Ziyuan Chen[17], Fangshu He[17], Yuyang Zhao[17], Yan Jin[17], Zhang Zhang[1,2,3,4,*], Luqi Huang[17], Shuhui Song[1,2,3,4,#] (TL), Yuan Yuan[17,#] (TL)
GTDB: Chenfen Zhou[7], Qingwei Xu[18], Sheng He[7,19], WeiYe[7], Ruifang Cao[7], Pengyu Wang[7], Yunchao Ling[7], Xing Yan[8], Qingzhong Wang[7], Guoqing Zhang[3,7,*]
LncExpDB: Zhao Li[1,2,3,4,#], Lin Liu[1,2,3,4,#], Shuai Jiang[1,2,4], Qianpeng Li[1,2,3,4], Changrui Feng[1,2,3,4], Qiang Du[1,2,3,4], Lina Ma[1,2,4,#] (TL)
scMethBank: Wenting Zong[1,2,3,4,#], Hongen Kang[1,2,3,4,#], Mochen Zhang[1,2,4,15], Zhuang Xiong[1,2,3,4], Rujiao Li[1,2,4,#] (TL)
TransCirc: Wendi Huan[3,7,#], Yunchao Ling[7,#], Sirui Zhang[3,7], Qiguang Xia[3,7], Ruifang Cao[7], Xiaojuan Fan[7], Zefeng Wang[3,7,20,#], Guoqing Zhang[3,7,*]
BioProject & BioSample & GSA & BIG Submission: Xu Chen[1,2,4,#], Tingting Chen[1,2,4,#], Sisi Zhang[1,2,4,#], Bixia Tang[1,2,4,#], Junwei Zhu[1,2,4,#], Lili Dong[1,2,4], Zhewen Zhang[1,2,4], Zhonghuang Wang[1,2,3,4], Hailong Kang[1,2,3,4], Yanqing Wang[1,2,4,#] (TL)
GWH: Yingke Ma[1,2,4,#], Song Wu[1,2,3,4#], Hongen Kang[1,2,3,4], Meili Chen[1,2,4,#] (TL)
GVM: Cuiping Li[1,2,4,#], Dongmei Tian[1,2,4,#], Bixia Tang[1,2,4,#], Xiaonan Liu[1,2,3,4,#], Xufei Teng[1,2,3,4,#], Shuhui Song[1,2,3,4,#] (TL)
GWAS Atlas: Dongmei Tian[1,2,4,#], Xiaonan Liu[1,2,3,4,#], Cuiping Li[1,2,4], Xufei Teng[1,2,3,4], Shuhui Song[1,2,3,4,#] (TL)
GEN: Yuansheng Zhang[1,2,3,4,#], Dong Zou[1,2,4,#], Tongtong Zhu[1,2,4,15,#], Ming Chen[1,2,4,15], Guangyi Niu[1,2,3,4], Chang Liu[1,2,3,4], Yujia Xiong[21,22], Lili Hao[1,2,4,#] (TL)
EDK: Guangyi Niu[1,2,3,4,#], Dong Zou[1,2,4,#], Tongtong Zhu[1,2,4,15], Xueying Shao[23], Lili Hao[1,2,4,#] (TL)
SmProt: Yanyan Li[6,24,#], Honghong Zhou[6,#], Xiaomin Chen[3,6,#], Yu Zheng[6,24], Quan Kang[6], Di Hao[6], Lili Zhang[3,6], Huaxia Luo[6], Yajing Hao[6], Runsheng Chen[6,10,*], Peng Zhang[6,#], Shunmin He[3,5,6,*]
MethBank: Dong Zou[1,2,4,#], Mochen Zhang[1,2,4,15,#], Zhuang Xiong[1,2,3,4], Zhi Nie[1,2,3,4], Shuhuan Yu[1,2,3,4], Rujiao Li[1,2,4,#] (TL)
EWAS Atlas: Mengwei Li[1,2,3,4,#], Rujiao Li[1,2,4], Yiming Bao[1,2,3,4,*] (TL)
EWAS Data Hub: Zhuang Xiong[1,2,3,4,#], Mengwei Li[1,2,3,4,#], Fei Yang[1,2,3,4,#], Yingke Ma[1,2,4], Jian Sang[1,2,3,4], Zhaohua Li [1,2,4,15], Rujiao Li[1,2,4,#] (TL)
iDog: Bixia Tang[1,2,4,#], Xiangquan Zhang[25,#], Lili Dong[1,2,4,#], Qing Zhou[1,2,3,4], Ying Cui[1,2,3,4], Shuang Zhai[1,2,4], Yaping Zhang[25], Guodong Wang[25,#] (TL), Wenming Zhao[1,2,3,4,*] (TL)
iSheep: Zhonghuang Wang[1,2,3,4,#], Qianghui Zhu[3,26,#], Xin Li[26], Junwei Zhu[1,2,4], Dongmei Tian[1,2,4], Hailong Kang[1,2,3,4], Cuiping Li[1,2,4], Sisi Zhang[1,2,4], Shuhui Song[1,2,3,4], Menghua Li (TL)[26,27], Wenming Zhao[1,2,3,4,*] (TL)
IC4R: Jun Yan[28,#], Jian Sang[1,2,3,4,#], Dong Zou[1,2,4,#], Chen Li[29], Zhennan Wang[3,30], Yuansheng Zhang[1,2,3,4], Tongtong Zhu[1,2,4,15], Shuhui Song[1,2,3,4,#] (TL), Xiangfeng Wang[28,#] (TL), Lili Hao[1,2,4,#] (TL)
SorGSD: Yuanming Liu[3,31,#], Zhonghuang Wang[1,2,3,4,#], Hong Luo[31], Junwei Zhu[1,2,4], Xiaoyuan Wu[31], Dongmei Tian[1,2,4], Cuiping Li[1,2,4], Wenming Zhao[1,2,3,4,*] (TL), Hai-Chun Jing[3,31,32,#] (TL)
PED: Ming Chen[1,2,3,4,#], Dong Zou[1,2,4,#], Lili Hao[1,2,4,#] (TL)
NONCODE: Lianhe Zhao[3,5,#], Jiajia Wang[6,24,#], Yanyan Li[6,24,#], Tinrui Song[6], Yu Zheng[6,24], Runsheng Chen[6,10,*], Yi Zhao[5,#], Shunmin He[3,6,*]
Database Commons: Dong Zou[1,2,4,#], Furrukh Mehmood[33], Shahid Ali[33], Amjad Ali[34], Shoaib Saleem[33], Irfan Hussain[33], Amir A. Abbasi[33], Lina Ma[1,2,4,#] (TL)
BIG Search: Dong Zou[1,2,4,#] (TL)
Education: Dong Zou[1,2,4,#], Shuai Jiang[1,2,4], Zhang Zhang[1,2,3,4,*] (TL)
Writing Group: Shuai Jiang[1,2,4,#], Wenming Zhao[1,2,3,4,*], Jingfa Xiao[1,2,3,4,*], Yiming Bao[1,2,3,4,*], Zhang Zhang[1,2,3,4,*]
CNCB-NGDC PARTNERS (Listed in alphabetical order by database names)
BBCancer: Zhixiang Zuo[35], Jian Ren[35]
CancerSEA: Xinxin Zhang[36], Yun Xiao[36], Xia Li[36]
CellMarker: Xinxin Zhang[36], Yun Xiao[36], Xia Li[36]
CGDB: Yiran Tu[37], Yu Xue[37]
circAtlas: Wanying Wu[38], Peifeng Ji[38], Fangqing Zhao[38]
CircFunBase: Xianwen Meng[39], Ming Chen[39]
dbPSP & THANATOS: Di Peng[37], Yu Xue[37]
DEG & DoriC: Hao Luo[40,41,42], Feng Gao[40,41,42]
DiseaseEnhancer: Xinxin Zhang[36], Yun Xiao[36], Xia Li[36]
DrLLPS: Wanshan Ning[37], Yu Xue[37]
EPSD & WERAM: Shaofeng Lin[37], Yu Xue[37]
EVmiRNA: Teng Liu[37], An-Yuan Guo[37]
GenTree: Hao Yuan[43,44], Yong E. Zhang[3,43,44]
iEKPD: Xiaodan Tan[37], Yu Xue[37]
iUUCD: Weizhi Zhang[37], Yu Xue[37]
lncAR: Yubin Xie[35], Jian Ren[35]
MiCroKiTS: Chenwei Wang[37], Yu Xue[37]
miRNASNP: Chun-Jie Liu[37], An-Yuan Guo[37]
PlantRegMap: De-Chang Yang[45], Feng Tian[45], Ge Gao[45]
PLMD: Dachao Tang[37], Yu Xue[37]
PTMD: Lan Yao[37], Yu Xue[37], Qinghua Cui[46,47]
RhesusBase: Ni A. An[48], Chuan-Yun Li[48]
RMVar: XiaoTong Luo[35], Jian Ren[35]
SEECancer: Xinxin Zhang[36], Yun Xiao[36], Xia Li[36]
* To whom correspondence should be addressed: Yongbiao Xue (ybxue@big.ac.cn).
Correspondence may also be addressed to Yiming Bao (baoym@big.ac.cn), Zhang Zhang (zhangzhang@big.ac.cn), Wenming Zhao (zhaowm@big.ac.cn), Jingfa Xiao (xiaojingfa@big.ac.cn), Shunmin He (heshunmin@ibp.ac.cn), Guoqing Zhang (gqzhang@picb.ac.cn), Yixue Li (yxli@sibs.ac.cn), Guoping Zhao (gpzhao@sibs.ac.cn) and Runsheng Chen (crs@ibp.ac.cn).
#The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.

[1]China National Center for Bioinformation, Beijing 100101, China
[2]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[3]University of Chinese Academy of Sciences, Beijing 100049, China
[4]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[5]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China
[6]National Genomics Data Center & Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China
[7]National Genomics Data Center & Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Xuhui, Shanghai 200031, China
[8]CAS-Key Laboratory of Synthetic Biology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Chinese Academy of Sciences, 300 Fenglin Road, Xuhui, Shanghai 200032, China
[9]Center for Quantitative Synthetic Biology, Institute of Synthetic Biology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[10]Guangdong Geneway Decoding Bio-Tech Co. Ltd, Foshan 528316, China
[11]CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[12]Institute for Stem cell and Regeneration, CAS, Beijing 100101, China
[13]State Key Laboratory of Stem Cell and Reproductive Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[14]State Key Laboratory of Membrane Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[15]School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China
[16]Sino-Danish College, University of Chinese Academy of Sciences, Beijing 100049, China
[17]National Resource Center for Chinese Materia Medica, Chinese Academy of Chinese Medical Sciences (CACMS), China
[18]College of Computer, Hubei University of Education, 129 Second Gaoxin Road, Wuhan Hi-Tech Zone, WuHan 430205, China
[19]School of Life Science and Technology, Shanghai Tech University, 393 Middle Huaxia Road, Pudong, Shanghai 201210, China
[20]CAS Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China
[21]Beijing Neurosurgical Institute, Beijing, China
[22]Capital Medical University, Beijing, China
[23]School of Computer Science and Engineering, South China University of Technology, China
[24]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[25]State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China
[26]CAS Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[27]College of Animal Science and Technology, China Agricultural University, Beijing 100193, China
[28]Department of Crop Genomics and Bioinformatics, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100094, China
[29]Rice Research Institute, Guangdong Academy of Agricultural Sciences, Guangzhou 510640, China
[30]Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[31]Key Laboratory of Plant Resources, Institute of Botany, Chinese Academy of Sciences, Beijing 100093, China
[32]Engineering Laboratory for Grass-Based Livestock Husbandry, Chinese Academy of Sciences, Beijing 100093, China
[33]Faculty of Biological Sciences, Quaid-i-Azam University, Islamabad 45320, Pakistan
[34]Atta-ur-Rahman School of Applied Biosciences (ASAB), National University of Sciences & Technology (NUST), Islamabad 44000, Pakistan
[35]State Key Laboratory of Oncology in South China, Cancer Center, Collaborative Innovation Center for Cancer Medicine, School of Life Sciences, Sun Yat-sen University, Guangzhou 510060, China
[36]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150081, China
[37]Key Laboratory of Molecular Biophysics of Ministry of Education, Hubei Bioinformatics and Molecular Imaging Key Laboratory, Center for Artificial Intelligence Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China
[38]Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China
[39]Zhejiang University, Hangzhou, 310027, China
[40]Department of Physics, School of Science, Tianjin University, Tianjin 300072, China
[41]Frontiers Science Center for Synthetic Biology and Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin 300072, China
[42]SynBio Research Platform, Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), Tianjin 300072, China
[43]Key Laboratory of Zoological Systematics and Evolution and State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[44]CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[45]Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), and State Key Laboratory of Protein and Plant Gene Research at School of Life Sciences, Peking University, Beijing 100871, China
[46]Department of Biomedical Informatics, School of Basic Medical Sciences, MOE Key Lab of Cardiovascular Sciences, Center for Noncoding RNA Medicine, Peking University, Beijing 100190, China
[47]Center of Bioinformatics, Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan 610054, China
[48]Beijing Key Laboratory of Cardiometabolic Molecular Medicine, Institute of Molecular Medicine, Peking University, Beijing, China