

# Genome Variation Map: a worldwide collection of genome variations across multiple species

Cuiping Li<sup>1,2,†</sup>, Dongmei Tian<sup>1,2,†</sup>, Bixia Tang<sup>1,2,†</sup>, Xiaonan Liu<sup>1,2,3,4,†</sup>, Xufei Teng<sup>1,2,3,4,†</sup>, Wenming Zhao<sup>1,2,3,4</sup>, Zhang Zhang<sup>1,2,3,4,\*</sup> and Shuhui Song<sup>1,2,3,4,\*</sup>

<sup>1</sup>China National Center for Bioinformation, Beijing 100101, China, <sup>2</sup>National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, <sup>3</sup>CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and <sup>4</sup>University of Chinese Academy of Sciences, Beijing 100049, China

Received September 15, 2020; Revised October 10, 2020; Editorial Decision October 12, 2020; Accepted October 14, 2020

## ABSTRACT

The Genome Variation Map (GVM; <http://bigd.big.ac.cn/gvm/>) is a public data repository of genome variations. It aims to collect and integrate genome variations for a wide range of species, accepts submissions of different variation types from all over the world and provides free open access to all publicly available data in support of worldwide research activities. Compared with the previous version, particularly, a total of 22 species, 115 projects, 55 935 samples, 463 429 609 variants, 66 220 associations and 56 submissions (as of 7 September 2020) were newly added in the current version of GVM. In the current release, GVM houses a total of ~960 million variants from 41 species, including 13 animals, 25 plants and 3 viruses. Moreover, it incorporates 64 819 individual genotypes and 260 393 manually curated high-quality genotype-to-phenotype associations. Since its inception, GVM has archived genomic variation data of 43 754 samples submitted by worldwide users and served >1 million data download requests. Collectively, as a core resource in the National Genomics Data Center, GVM provides valuable genome variations for a diversity of species and thus plays an important role in both functional genomics studies and molecular breeding.

## INTRODUCTION

The Genome Variation Map (GVM; <https://bigd.big.ac.cn/gvm/>), as a core resource of the National Genomics Data Center (CNCB-NGDC) (1), part of the China National Center for Bioinformation (CNCB), is a public data repository of genome variations. Since its inception in 2017 (2),

GVM has served as a central public resource for genome variations and played an important role in both functional genomics studies and molecular breeding (3,4). For instance, variants and knowledge associations deposited in GVM have been used in several data resources (e.g. IC4R (5), SR4R (6), MBKbase for rice (7), GWAS Atlas (8) and Animal-ImputeDB (9)). Over the past several years, advances in high-throughput sequencing technologies have empowered large-scale population genome sequencing projects, leading to massive genome variations identified at unprecedented rates. Consequently, GVM has accepted >50 data submissions (10–12) from all over the world, and as of September 2020, accordingly housed a large number of genome variations from 41 species, including not only human, but also domesticated animals, cultivated plants and viruses, particularly SARS-CoV-2, a coronavirus provoking the ongoing global pandemic. Meanwhile, GVM has served >1 million data download requests (<https://bigd.big.ac.cn/gvm/statistics>). Importantly, to provide high-quality variant data and metadata and deliver user-friendly data services, GVM has been frequently updated in the past years by standardizing the curation model and process, improving the web functionalities for data submission, browse and download, providing the database tutorial in PPT and video, and adding external links to other public databases, such as dbSNP (13), GWAS Catalog (14), NCBI genome (15), ENSEMBL (16), JGI (17), maizedb (18) and DRDB (19). Here we present an updated release of GVM and briefly describe its recent updates and data growth.

## DATA COLLECTION AND METHODS

Whole-genome resequencing projects were collected from published literatures, and raw sequence data were downloaded from Sequence Read Archive (SRA) (20) and Genome Sequence Archive (GSA) (21,22). All collected

\*To whom correspondence should be addressed. Tel: +86 10 8409 7620; Fax: +86 10 8409 7720; Email: songshh@big.ac.cn  
Correspondence may also be addressed to Zhang Zhang. Tel: +86 10 8409 7261; Fax: +86 10 8409 7720; Email: zhangzhang@big.ac.cn

†The authors wish it to be known that, in their opinion, these authors should be regarded as joint First Authors.

raw sequence reads were subjected to quality control using Trimmomatic-0.36 and cleaned reads were aligned to the reference genomes using BWA-MEM (23). Those aligned reads were then merged into a single BAM file and sorted by SAMtools (19), and marked for duplicates using MarkDuplicates in GATK-4.0.5.0 (24). After removing duplicate reads, high-quality variants were identified by both GATK HaplotypeCaller and SAMtools mpileup, and base quality was recalibrated by Base Quality Score Recalibration (BQSR). Then, an intermediate genomic GVCF file for each sample was produced by running HaplotypeCaller in GVCF mode, and Genotype-GVCFs in GATK was applied to pool all GVCF files together to create a VCF file containing all raw variants. These raw variants were further filtered by using SelectVariants and VariantFiltration in GATK. Default parameters were used in the variant calling. The effects of all variants were annotated using VEP (25) as well as in-house pipelines. Functional annotation of variants were performed based on GO (26), UniProt (27) and Pfam (28). Furthermore, the genotype-to-phenotype (G2P) associations were manually curated from published GWAS literatures, and the relationships between sequence variants and phenotypic traits were established.

## DATA MODULES AND DATA GROWTH

Over the past several years, GVM has been significantly updated regarding data modules and data volume. To better present genomic variants, all relevant entities and metadata in GVM are organized into six modules in terms of species, project, sample, variant, association and submission. Moreover, the number of genomic variants hosted in GVM is growing rapidly from ~497 million in 19 species in August 2017 to ~960 million in 41 species in August 2020 (Table 1). An illustration of all collected species and data statistics is presented in Figure 1 (with details in Supplementary Table S1). Compared with the previous version, particularly, a total of 22 species, 115 projects, 55 937 samples, 463 429 609 variants, 66 220 associations and 56 submissions (as of 7 September 2020) were newly added in the current version of GVM.

The Species module provides a comprehensive overview on all collected species as well as their associated projects, samples, variants and associations (if available), which together are organized in a tabular table and linked to internal and external resources (Figure 2A). As of 7 September 2020, there are a total of 41 species, including 13 animals, 25 plants and 3 viruses. The newly updated species include three animals (cat, horse, and tarpan), 10 cultivated plants (carrot, cassava, common bean, cotton, cucumber, date palm, grape, apricot, rapeseed and wheat), five traditional Chinese medicine plants (*Catharanthus roseus*, Cannabis, Ganoderma, Jatropha and *Salvia miltiorrhiza*) and three coronaviruses (SARS-CoV-2, SARS-CoV and MERS-CoV). Since the outbreak of severe respiratory disease COVID-19 in late December 2019, SARS-CoV-2 has been rapidly spread as a global pandemic. To help worldwide researchers better understand the genome variation and transmission of SARS-CoV-2 (29), we analyzed genome sequences of SARS-CoV-2 as well as two close relatives (SARS-CoV and MERS-CoV) and made their genomic variants publicly available for the global research

community through 2019nCoV (19). As of 7 September 2020, there are a total of 16 934 variants identified from 52 466 high-quality SARS-CoV-2 assemblies as well as 477 and 1742 variants from 105 SARS-CoV and 248 MERS-CoV assemblies, respectively.

In the Project and Sample modules, we compiled the metadata of whole-genome resequencing projects (Figure 2B) and samples (Figure 2C), respectively. The Project module displays an overview of all resequencing projects, involving sequenced sample size, sampling material, sequencing technology, data type and average sequencing coverage. Besides, bibliographic details (e.g. title, year, journal, PubMed ID) and a short description for each publication are collectively summarized, which are helpful for users to quickly understand the outline of the sequencing project(s) of interest. Likewise, the Sample module provides a detailed description on sequenced samples, including sample name, cultivar/breeder, geographic information (from which sequenced samples were collected), etc. A unique accession ID was assigned for each sample, and the number of sampled materials for a specific geographic region was further mapped in a world map, providing a worldwide landscape on the distribution of samples for each species and accordingly facilitating researchers to evaluate the sample representativeness and genetic diversity.

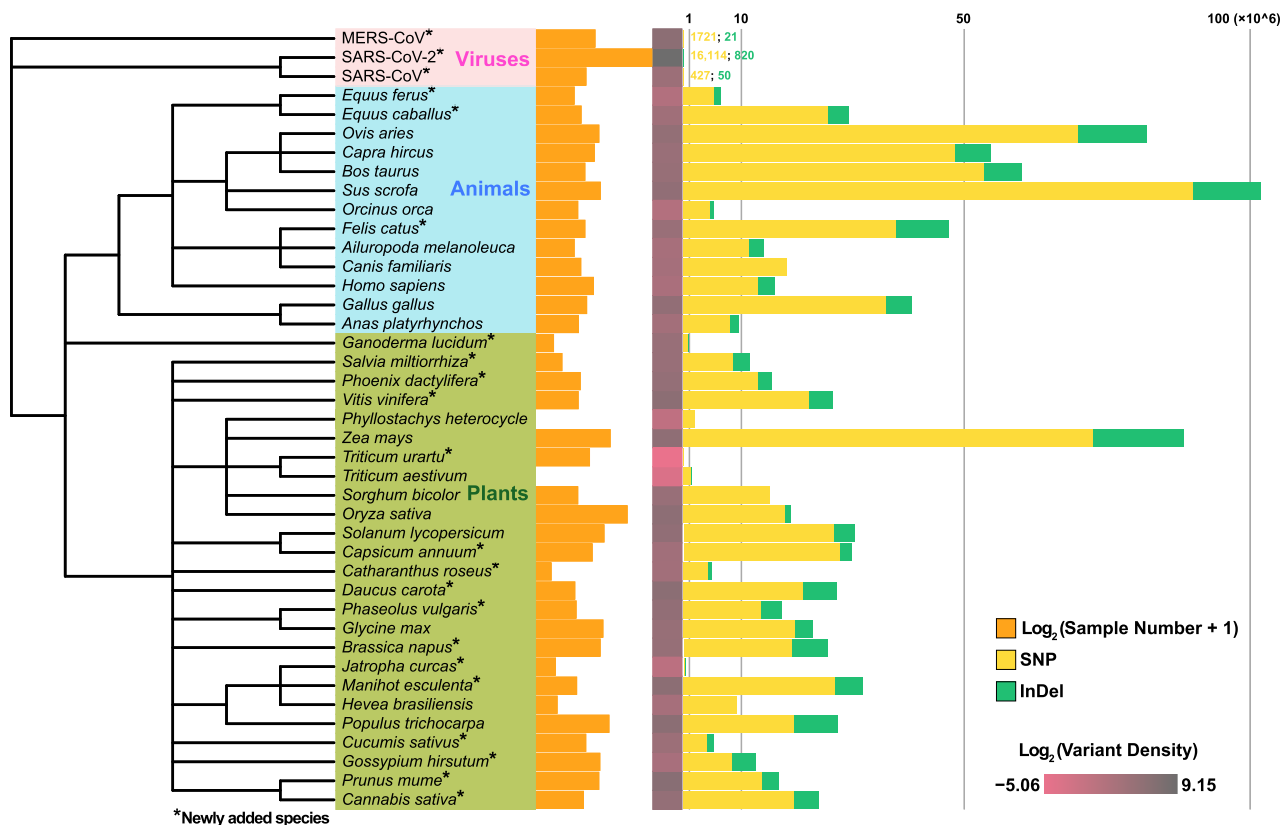
The Variant module (Figure 2D) provides a catalog of genome variations, including SNPs and indels, identified from a diversity of species (details see methods). For each variant, a unique identifier was assigned and its related details including variant coordinate, reference and alternative alleles, minor allele frequency, and hyperlinks to external databases (e.g. dbSNP, ClinVar) were provided. To help users prioritize the potentially functional SNPs, GVM provides comprehensive annotations for each variant, including consequence type, variant effect, population frequency and phenotype association, and also incorporates the functional domain information from UniProt and Pfam. Moreover, with the rapid accumulation of huge amounts of genomic variants, we further calculated the SNP density for each species and found that the number of SNP markers ranged from 1 to 64 per kb, with an averaged distance of 131 bp between adjacent SNP loci (Figure 1). In short, the SNP-based high-density genetic map for each species is critically important for a wide range of functional studies.

The Association module (Figure 2E) integrates a total of 78 950 high-quality ( $P < 0.001$ ) genotype-to-phenotype (G2P) GWAS associations for 12 non-human species that were manually curated from 304 publications. These G2P associations account for 735 traits across seven cultivated plants (cotton, Japanese apricot, maize, rapeseed, rice, sorghum and soybean) and five domesticated animals (chicken, cattle, goat, pig and sheep). More importantly, all associations and traits have been further annotated and organized based on a suite of ontologies (Plant Trait Ontology, Animal Trait Ontology for Livestock, etc.) in GWAS Atlas (8), and these G2P associations are of great significance for genetic research on important traits and breeding application.

The Submission module offers online data submission services and accepts multiple data formats including VCF, GVCF and HapMap. It allows variation submission for any species and from any particular genome (e.g. mito-

**Table 1.** Statistics and comparison between the two versions of GVM

	GVM 2018 (As of September 2017)	GVM 2021 (As of September 2020)
Species	19	41
Projects	87	202
Samples	8884	64 819
Number of SNPs	434 525 449	818 769 204
Number of indels	62 454 393	141 640 247
Associations	194 173	260 393
Submitted samples	NA	43 754
Variant annotation	GO/UniProt/ClinVar/OMIM	GO/UniProt/ClinVar/OMIM/Pfam
Global distribution of samples	NA	Available
FTP	NA	Available



**Figure 1.** A skeleton view of all collected species in GVM and their corresponding samples and variants.

chondria and chloroplast), and supports batch online submission from all over the world. A detailed instruction for data submission is available on <https://bigd.big.ac.cn/gvm/instruction>. Besides, the complete collection of those released variants can be downloaded directly via FTP at <ftp://download.big.ac.cn/GVM>. As of 7 September, 2020, GVM contains 56 submissions describing variations from 23 species (e.g. human, soybean and maize). According to our data statistics in GVM, more than half of submitted samples (43 754) belong to human (26 956, 61.6%) and the rest are animals (8195, 18.7%) and plants (8601, 19.7%).

## DOWNLOAD


GVM provides open access to all publicly available variants, which are downloadable in both VCF and FASTA formats at <https://bigd.big.ac.cn/gvm/download>. The brief VCF file contains genomic position, reference and alternative alle-

les and variant quality, and the FASTA file provides 50nt flanking sequences for each variant. According to the user's feedback, we newly added the detailed VCF file containing the genotype information for all samples, which would be more useful for users to conduct in-depth GWAS functional analysis.

## FUTURE PLANS

GVM, as a public data repository of genomic variants, features comprehensive integration of different types of genome variations for a wide range of species. With the development of high-throughput sequencing technology, GVM is expected to continue to grow rapidly over the next following years. As GVM offers high-density variation map for each species, these variants are of critical significance for population genetics, evolutionary analysis, association studies and genomic breeding. Thus, future developments

**A Species overview**

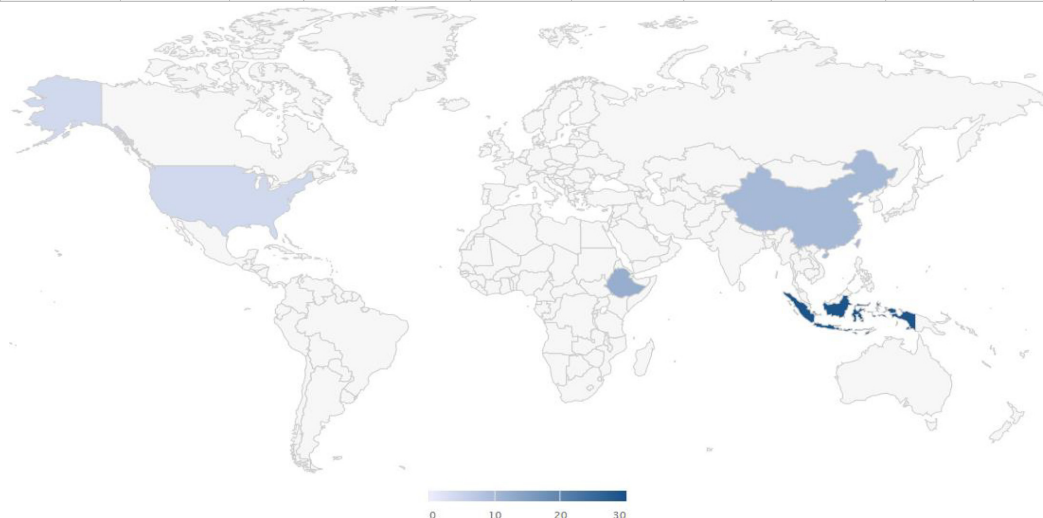
Species	Taxonomy ID	Assembly Version	Genome Size	SNP	INDEL	#Project	#Sample	Association	Variants Density(kb)
 Chicken <i>Gallus gallus</i>	9031	Gallus_gallus-5.0	1.23G	36,174,851	4,619,064	9	112	1,233	33.16

**B Project metadata**

Name	Samples	Material	Technology	Platform	Data Type	Coverage	Description	Author	Journal	Year	PMID
DRP003109	51	Whole blood	WGS	ILLUMINA	PAIRED	~6x	Genetic features of red and green junglefowls and relationship with Indonesian native chickens Sumatera and Kedu Hitam	Ulfah M et al.	BMC Genomics	2016	27142387
SRP022583	2	Peripheral venous blood	WGS	ILLUMINA	PAIRED	~25x	Genome-Wide Patterns of Genetic Variation in Two Domestic Chickens	Wen-Hsiung Li et al.	Genome Biology and Evolution	2013	23814129

**C Sample metadata and global distribution**

GVMSampleID	SampleName	Breed	Geographic	Material	ProjectAcc	SampleAcc	PubmedID	Instrument	AssayType	CenterName
gga.s1	Black_Java_1	Black Java	NA	Whole blood	DRP003109	DRS042905	27142387	Illumina HiSeq 2500	WGS	TUAGRI
gga.s2	Black_Java_10	Black Java	NA	Whole blood	DRP003109	DRS042914	27142387	Illumina HiSeq 2500	WGS	TUAGRI



**D Genomic variants**

Variants for Chicken search results [SNP: 36,174,851](#) [InDel: 4,619,064](#)

Advanced Search

Item 1 - 10 of 36174851  Items per page [First](#) [Prev](#)  of 3804790 [Next](#) [Last](#) [GOTO](#)

<input type="checkbox"/> All	VarID	Position	Alleles	MAF	Consequence Type   Effect	Gene	dbSNP
<input type="checkbox"/>	<a href="#">gga19715</a>	chr1:463380	G/C	G:0	upstream_gene_variant MODIFIER;	<a href="#">ENSGALG00000033149</a> (GCC1)	-
<input type="checkbox"/>	<a href="#">gga19716</a>	chr1:463386	G/C	G:0	upstream_gene_variant MODIFIER;	<a href="#">ENSGALG00000033149</a> (GCC1)	-
<input type="checkbox"/>	<a href="#">gga19718</a>	chr1:463409	G/T	G:0	upstream_gene_variant MODIFIER;	<a href="#">ENSGALG00000033149</a> (GCC1)	-

**E Genotype-to-phenotype associations**

VarID	Traits	Species	Position	P-values	R2(%)	Mapped Gene(s)	Consequence Type(s)	PMID
<a href="#">gga15019603</a>	antibody response to sheep red blood cells	Chicken	chr2:78425099	3.60E-06	-	<a href="#">ENSGALG00000038328</a>	intron_variant	28100171
<a href="#">gga18034797</a>	infectious bursal disease virus antibody titre	Chicken	chr20:13601844	2.82E-08	15	<a href="#">ENSGALG00000017366</a> <a href="#">ENSGALG00000007986</a>	upstream_gene_variant synonymous_variant	27687164

**Figure 2.** Screenshots of data modules in terms of species, project, sample, variation, and association. (A) Species overview. (B) Project metadata. (C) Sample metadata and global distribution. (D) Genomic variants. (E) Genotype-to-phenotype association.

are to generate different reference SNP panels, including hapmapSNPs after data filtration and genotype imputation, tagSNPs after removing linkage disequilibrium-based redundancy SNPs, fixedSNPs selected from genes exhibiting selective sweep signatures and barcodeSNPs selected from DNA fingerprinting simulation. In fact, it has been implemented in the 3000 Rice Genome Project (30) and SNP Ready for Rice (SR4R, <http://sr4r.ic4r.org/>) (19). Additionally, these SNP datasets will be readily for optimal design of low-density (LD), medium-density (MD) or high-density (HD) SNP chip, which would be helpful to develop a rapid, accurate and efficient method for genotyping several hundred or thousand polymorphisms in large numbers of individuals. Furthermore, ongoing efforts will also include optimization of curation models and processes, integration of more variation datasets, enhancement of genomic variant annotation, and improvement of web interfaces and data analysis pipelines.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank our colleagues, students, and a number of users for reporting bugs and sending comments.

## FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDA24040201 to S.S., XDA19090116 to S.S., XDA19050302 to Z.Z.]; National Key R&D Program of China [2020YFC0848900, 2018YFD1000505, 2017YFC0907502]; 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05]; Genomics Data Center Construction of Chinese Academy of Sciences [XXH-13514-0202]; K. C. Wong Education Foundation (to Z.Z.); International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; Youth Innovation Promotion Association of Chinese Academy of Science [2017141 to S.S.]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences.

*Conflict of interest statement.* None declared.

## REFERENCES

- National Genomics Data Center Members and Partners. (2020) Database resources of the national genomics data center in 2020. *Nucleic Acids Res.*, **48**, D24–D33.
- 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.
- He,G., Wang,Z., Guo,J., Wang,M., Zou,X., Tang,R., Liu,J., Zhang,H., Li,Y., Hu,R. *et al.* (2020) Inferring the population history of Tai-Kadai-speaking people and southernmost Han Chinese on Hainan Island by genome-wide array genotyping. *Eur. J. Hum. Genet.*, **28**, 1111–1123.
- Liu,S., Li,C., Wang,H., Wang,S., Yang,S., Liu,X., Yan,J., Li,B., Beatty,M., Zastrow-Hayes,G. *et al.* (2020) Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biol.*, **21**, 163.
- 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.
- 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.
- Peng,H., Wang,K., Chen,Z., Cao,Y., Gao,Q., Li,Y., Li,X., Lu,H., Du,H., Lu,M. *et al.* (2020) MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. *Nucleic Acids Res.*, **48**, D1085–D1092.
- 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.
- Yang,W., Yang,Y., Zhao,C., Yang,K., Wang,D., Yang,J., Niu,X. and Gong,J. (2020) Animal-ImputeDB: a comprehensive database with multiple animal reference panels for genotype imputation. *Nucleic Acids Res.*, **48**, D659–D667.
- Zhou,Z., Li,M., Cheng,H., Fan,W., Yuan,Z., Gao,Q., Xu,Y., Guo,Z., Zhang,Y., Hu,J. *et al.* (2018) An intercross population study reveals genes associated with body size and plumage color in ducks. *Nat. Commun.*, **9**, 2648.
- Zhang,Z., Jia,Y., Chen,Y., Wang,L., Lv,X., Yang,F., He,Y., Ning,Z. and Qu,L. (2018) Genomic variation in Pekin duck populations developed in three different countries as revealed by whole-genome data. *Anim. Genet.*, **49**, 132–136.
- Liu,Y., Du,H., Li,P., Shen,Y., Peng,H., Liu,S., Zhou,G.A., Zhang,H., Liu,Z., Shi,M. *et al.* (2020) Pan-Genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.
- Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
- Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
- Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
- Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
- Nordberg,H., Cantor,M., Dusheyko,S., Hua,S., Poliakov,A., Shabalov,I., Smirnova,T., Grigoriev,I.V. and Dubchak,I. (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, **42**, D26–D31.
- Polacco,M., Coe,E., Fang,Z., Hancock,D., Sanchez-Villeda,H. and Schroeder,S. (2002) MaizeDB - a functional genomics perspective. *Comp. Funct. Genomics*, **3**, 128–131.
- 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*, **42**, 212–221.
- Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database, C. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
- 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.
- Zhang,S.S., Chen,T.T., Zhu,J.W., Zhou,Q., Chen,X., Wang,Y.Q. and Zhao,W.M. (2018) [GSA: Genome Sequence Archive]. *Yi Chuan*, **40**, 1044–1047.
- Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
- McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

25. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R., Thormann,A., Flicek,P. and Cunningham,F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
26. The Gene Ontology, C. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.
27. UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **46**, 2699.
28. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. et al. (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
29. 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.
30. Wang,W., Mauleon,R., Hu,Z., Chebotarov,D., Tai,S., Wu,Z., Li,M., Zheng,T., Fuentes,R.R., Zhang,F. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.