D816–D826 *Nucleic Acids Research, 2023, Vol. 51, Database issue* *Published online 27 October 2022*
https://doi.org/10.1093/nar/gkac954

Animal-SNPAtlas: a comprehensive SNP database for multiple animals

Yingjie Gao[1,†], Guanghui Jiang[1,†], Wenqian Yang[1], Weiwei Jin[1], Jing Gong[1,2,*], Xuewen Xu[3,*] and Xiaohui Niu[1,*]

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P. R. China, [2]College of Biomedicine and Health, Huazhong Agricultural University, Wuhan 430070, China and [3]Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education & College of Animal Science and Technology, Huazhong Agricultural University, Wuhan 430070, China

Received August 11, 2022; Revised October 06, 2022; Editorial Decision October 08, 2022; Accepted October 26, 2022

ABSTRACT

Single-nucleotide polymorphisms (SNPs) as the most important type of genetic variation are widely used in describing population characteristics and play vital roles in animal genetics and breeding. Large amounts of population genetic variation resources and tools have been developed in human, which provided solid support for human genetic studies. However, compared with human, the development of animal genetic variation databases was relatively slow, which limits the genetic researches in these animals. To fill this gap, we systematically identified ∼ 499 million high-quality SNPs from 4784 samples of 20 types of animals. On that basis, we annotated the functions of SNPs, constructed high-density reference panels and calculated genome-wide linkage disequilibrium (LD) matrixes. We further developed Animal-SNPAtlas, a user-friendly database (http://gong_lab.hzau.edu.cn/Animal_SNPAtlas/) which includes high-quality SNP datasets and several support tools for multiple animals. In Animal-SNPAtlas, users can search the functional annotation of SNPs, perform online genotype imputation, explore and visualize LD information, browse variant information using the genome browser and download SNP datasets for each species. With the massive SNP datasets and useful tools, Animal-SNPAtlas will be an important fundamental resource for the animal genomics, genetics and breeding community.

INTRODUCTION

A central goal of genetics is to pinpoint the functional variants that contribute to diseases or population traits (1). Over the past dozen years, large-scale international collaborative efforts have successively constructed larger and more ethnically diverse genetic variation resources in humans, and several of them include thousands of samples and millions of single nucleotide polymorphisms (SNPs), such as International HapMap Project Phase3 (2), 1000 Genomes Project Phase 3 (3) and Haplotype Reference Consortium (4). Besides, hundreds of SNP-related resources and tools have been developed for genetic analyses, such as VARAdb (5), TOP-LD (6), LDtrait (7). These large-scale genetic variation resources and support tools constitute a key component for human genetic studies, which have elucidated the properties and distribution of genetic variants, provided insights into the processes that shape genetic diversity and advanced understanding of disease biology (8–12). However, in contrast to humans, animal genetic variation databases were developed relatively slowly, especially the comprehensive databases for the systematical characterization of SNPs, function annotation and genome-wide LD calculation, which limits the development of animal genomics, genetics and breeding research.

In recent years, the advances in high-throughput sequencing technologies have notably reduced the cost of whole-genome sequencing (13), and large amounts of population genotype data from different species have been continuously released. However, whole genome sequencing is still expensive for the vast majority of studies. Accordingly, low-coverage sequencing or SNP arrays integrating with genotype imputation (14), which could increase the SNP density, boost the power to detect significant associations and facilitate to identify causal variants of complex traits (15), have been widely applied in animal genetic studies (16–19).

*To whom correspondence should be addressed. Tel: +86 27 87285085; Fax: +86 27 87285085; Email: gong.jing@mail.hzau.edu.cn
Correspondence may also be addressed to Xuewen Xu. Tel: +86 27 87285085; Fax: +86 27 87285085; Email: xuewen_xu@mail.hzau.edu.cn
Correspondence may also be addressed to Xiaohui Niu. Tel: +86 27 87285085; Fax: +86 27 87285085; Email: niuxiaoh@mail.hzau.edu.cn
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

But, to perform genotype imputation, high-quality reference panels constructed from large sample size of population and high-density SNP datasets are crucial (15). Although the whole-genome sequence data of animals have shown a staggering increase, the high-quality SNP datasets and reference panels for the animals remain inadequate to date.

In addition, the specific background information of SNPs, such as linkage disequilibrium (LD), functional annotation and so on, are also critical for the downstream genetic analyses. Due to the LD phenomenon in the genome, direct inference from significantly associated SNPs rarely yields causal variants (20). High-resolution and accurate LD calculation with high-density SNPs can significantly improve the power of association study, facilitate the prioritization of the candidate causal variants (21–23), and provide more insights into the underlying biological mechanisms of the complex traits (24,25). For example, a genome-wide LD-based fine-mapping identified pleiotropic and functional variants that could predict many traits across global cattle populations (26). Another study using LD analysis based on whole-genome sequence data has revealed three candidate causal variants for backfat within intronic regions and downstream of the *CCND2* gene in pig (27).

Prioritizing causal SNPs with fine-mapping requires integrating not only regional LD patterns but also functional prediction of associated SNPs, which is an important process before biological experiments in the post-GWAS study (28,29). SNP functional annotation, which assigns a likely biological function to genetic variants based on their position and roles in sequence change, can assist the prioritization of causal SNPs in follow-up functional studies (30).

With the rapid progress in animal genomic research in recent years, some animal-related databases have been developed, such as Animal-APAdb (31), Animal-eRNAdb (32), AnimalTFDB 3.0 (33) and Animal QTLdb (34). In the field of animal genetics, several databases were developed for specific domestic animals and aquaculture species. For example, iDog (35), iSwine (36), BGVD (37) and Galbase (38) collected genetic variation for dog, pig, cattle and chicken, respectively. iSheep (39) and Goat Genome Variation Database (GGVD) (40) provided genetic variants and phenotypic data or selection signature patterns for sheep and goat. For aquaculture species, AMBP (41) was developed for genotype imputation and genetic analysis in aquaculture. In addition, the Genome Variation Map (42) database collected 18 animal species, and provide basic information of genetic variants. These databases provide useful data for animal genetic studies. However, databases that provide comprehensive information of SNP annotation, imputation panels and LD information for multiple species of animals, are still relatively rare. Furthermore, a comprehensive database containing multiple animal species could greatly facilitate the cross-species genetic analysis, which could provide insights into the origin and evolution of species and specific genes (43,44), narrow down the range of quantitative trait loci (45) and help decipher the functions of candidate genes or variants (46–48). To fill these gaps, we curated and analyzed the public whole-genome sequencing data or raw genotype data of 4784 samples from 20 animal species. We systematically identified ∼499 million high-quality SNPs, constructed high-density reference panels, and performed SNP annotation and LD calculation. Finally, we developed a comprehensive database named as Animal-SNPAtlas (Figure 1). Animal-SNPAtlas allows users to search SNP annotated information, perform online genotype imputation directly, browse and visualize LD information, apply genome browser for SNP visualization and download high-quality SNP datasets with VCF or M3VCF format. Overall, Animal-SNPAtlas will be a useful resource for animal genetic studies.

MATERIALS AND METHODS

Data collection

In order to curate as many samples as possible in this study, we collected raw sequencing data and SNP datasets of 20 representative animal species from public databases by a consistent filtering pipeline. As for raw sequencing data, the WGS data of cat, cattle, chicken, chimpanzee, duck, goat, honey bee, large yellow croaker, mouse, rabbit, rat, rhesus monkey, rock pigeon, salmon, sheep and water buffalo were collected from the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) (49). Then, the detailed information including bioproject, tissue and breed of sequencing data was downloaded and manually curated. Available bioprojects were further checked according to the sample size, the sequencing depth, the experimental condition and the research purpose. The corresponding samples were retained and applied for subsequent analyses. As for genotype data, the raw SNP datasets of horse and vervet monkey were gathered from the European Institute of Bioinformatics (EBI, https://www.ebi.ac.uk/) (50). The raw genotype data of dog and pig were obtained from the National Human Genome Research Institute (NHGRI, https://www.genome.gov/) (51) and the Genome Variation Map (GVM, https://bigd.big.ac.cn/gvm/home) (42), respectively. In addition, the reference genome files and genome annotation files were downloaded from the Ensembl (https://www.ensembl.org) (52).

Data Processing

For the raw sequencing data, high-quality SNPs were identified using the Sentieon pipeline (53) (Figure 1). First, the raw reads were subjected to quality control using fastp (54) and cleaned with Trimmomatic (55). Next, the clean reads were aligned to the current standard reference genome by the Burrows–Wheeler Alignment (BWA) mem algorithm with default parameters (56), and the BAM files of reads with quality greater than 10 were retained by SAMtools (57). Then, the duplicate reads were removed, the indels were realigned, and the base quality was recalibrated using the Sentieon driver. Subsequently, the SNPs of each sample were identified using Sentieon's Haplotyper algorithm, and the variant data of all samples were merged into VCF files using the Sentieon GVCFtyper algorithm. Finally, SNPs and samples with low quality were filtered out. Specifically, SNPs with the threshold 'QD $<$ 5.0 or FS $>$ 15.0 or MQ $<$ 50.0 or ReadPosRankSum $<$ −8.0 or MQRankSum $<$ −12.5 or SOR $>$ 3.0' were filtered by Genome Analysis Toolkit (GATK) (58) VariantFiltration. Additionally,

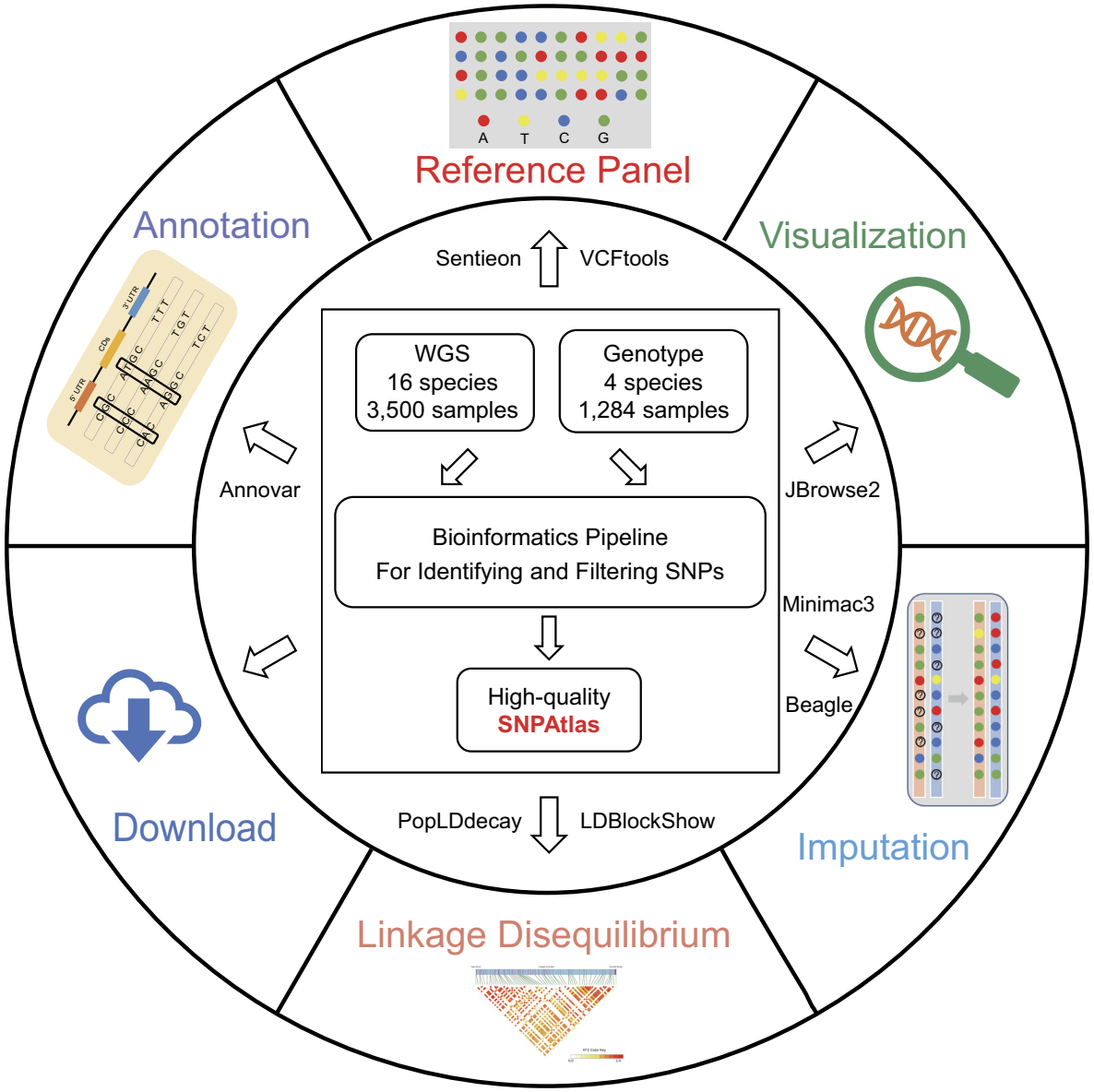

Figure 1. Flow chart of the Animal-SNPAtlas database.

SNPs with minor allele frequency (MAF) < 0.01 or calling rate < 0.9 were removed by VCFtools (59). Besides, samples with high missing rate were also removed by VCFtools. For the genotype data, the same filtering method was adopted to construct high-quality SNP datasets.

The detailed statistics of the genetic variants and sample data of each species in the final dataset are listed in Table 1.

Population structure analyses

Since the population structure of each species may influence the accuracy of the genotype imputation and LD calculation, we provided more information about the subtypes / breeds of samples. First, PLINK was used to convert the SNP datasets in VCF format to binary BED format and perform LD pruning. Then, several strategies were used to explore the genetic composition and population structure for each species. The admixture analysis was performed with ADMIXTURE v1.3.0 software (60). Cross-validation errors were estimated for each *K*-value. The *K*-value with the lowest CV error was regarded as the optimal value for estimating the level of admixture in each sample. The phylogenetic tree of the samples based on the neighbor-joining algorithm was constructed by TreeBeST (https://github.com/Ensembl/treebest). The Principal Component Analysis (PCA) was carried out by GCTA (61).

Genotype imputation and construction of reference panel

Using the high-quality SNP datasets, Beagle (v5.4) (62) was first used to phase the haplotype with the default parame-

Table 1. Data summary of Animal-SNPAtlas

Species	Annotation version	Number of chromosomes	Reference panel		
			Number of samples	Number of SNPs	Number of annotated genes
Anas platyrhynchos (Duck)	ZJU_duck1.0	33	358	19 097 802	13 828
Apis mellifera (Honey bee)	Amel_HAv3.1	16	290	1 278 406	9 657
Bos taurus (Cattle)	ARS-UCD 1.2	30	264	15 515 352	20 615
Bubalus bubalis (Water buffalo)	UOA_WB_1	25	207	32 474 303	18 300
Canis familiaris (Dog)	CanFam3.1	39	658	61 065 811	25 587
Capra hircus (Goat)	ARS1	29	261	23 732 005	21 036
Chlorocebus pygerythrus (Vervet monkey)	ChlSab1.1	30	163	47 180 225	22 948
Columba livia (Rock pigeon)	Cliv_2.1	15 057 scaffolds	307	17 198 341	13 988
Equus caballus (Horse)	EquCab3.0	32	87	18 517 526	24 503
Felis catus (Cat)	Felis_catus_9.0	19	247	29 598 513	25 010
Gallus gallus (Chicken)	GRCg6a	34	284	16 572 439	20 237
Larimichthys crocea (Large yellow croaker)	L_crocea_2.0	24	198	17 141 732	19 100
Macaca mulatta (Rhesus monkey)	Mmul_10	22	256	23 331 766	26 475
Mus musculus (Mouse)	GRCm39	21	40	40 784 198	44 983
Oryctolagus cuniculus (Rabbit)	OryCun2.0	22	74	27 332 887	16 830
Ovis aries (Sheep)	ARS-UI_Ramb_v2.0	27	263	37 875 937	24 872
Pan troglodytes (Chimpanzee)	Clint_PTRv2	25	49	20 950 687	28 747
Rattus norvegicus (Rat)	mRatBN7.2	22	60	16 740 211	24 076
Salmo salar (Atlantic salmon)	ICSASG_v2	29	342	6 913 798	35 305
Sus scrofa (Pig)	Sscrofa10.2	19	376	25 530 833	16 802

ters. Then Minimac3 (63) was applied to convert the format of the reference panel to M3VCF. In the end, two popular types of reference panels for all species were provided in our database (Figure 1).

Evaluation of the performance of reference panels

To validate the performance of reference panels and the imputation process, we calculated imputation accuracy for each species using a 5-fold cross-validation strategy. For each species, all the samples in the reference panel were randomly divided into five parts, with one part being selected as the study population, and the remaining parts being used as the reference panels for each time. Since SNP arrays with about 50k probes are widely used in animal genetic researches, we randomly selected 50k SNPs from the whole genome of the study population and masked other SNPs for all species. At the same time, considering the numbers of probes in SNP arrays vary depending on the species. We selected three representative species based on different genome sizes, including chicken (1.1 G), cattle (2.6 G) and Atlantic salmon (3.4 G), and simulated six different SNP densities from 2K to 600K SNPs (i.e. 2K, 10K, 50K, 100K, 300K and 600K), to test the accuracy of our panels. Then, we used Beagle and Minimac3 to impute the masked genotypes with default parameters, respectively.

In this way, we have both the true and imputed genotypes. The imputed SNPs with MAF $\geq$0.01 and estimated squared correlation $\geq$0.3 were remained as properly imputed variants and applied for the following evaluation. The concordance rate (CR) and the squared correlation (R^2) were used to validate the accuracy of the imputation. CR was calculated by dividing the number of correctly imputed genotypes by the total number of imputed genotypes per species, and R^2 was the squared correlation between true and imputed genotypes, the mean of CR or R^2 across five-parts was taken as the accuracy of the imputation for each species. The number of SNPs increased by 28.31 folds on average in the study population after imputation. The average CR of Beagle ranged from 0.862 for rat to 0.963 for cat, and that of Minimac3 ranged from 0.889 for honey bee to 0.970 for horse. The average R^2 of Beagle ranged from 0.674 for duck to 0.925 for cat, and that of Minimac3 ranged from 0.671 for honey bee to 0.935 for horse (Table 2). Additionally, we simulated different SNP densities and performed 5-fold cross-validation for three species with different genome sizes. The results showed that the accuracies of genotype imputation increased with the SNP densities and the average R^2 was more than 0.8 when the number of SNPs is more than 50K for both imputation tools, indicating the relative robustness of our panels (Supplementary Figure S1).

SNP annotation and Linkage disequilibrium calculation

For species with gene annotation files, we converted the VCF format of the reference panels to the standard format of ANNOVAR input files with the ANNOVAR package (30). We obtained their predicted functional consequences such as missense, nonsense, intronic, splice-site, UTRs, exonic and intergenic variants by ANNOVAR and showed them on the website.

Linkage disequilibrium in the whole genome was calculated using imputed genotypes for each species using PopLDdecay (64) with default parameters (Figure 1). The SNP pairs with $R^2 < 0.2$ in the LD matrix were removed. Furthermore, LDBlockShow (65) was used to visualize LD haplotype blocks. In addition, samples of each species were grouped according to the analysis of population structure. The reference panels of different groups for each species and

Table 2. Imputation accuracy using reference panels in Animal-SNPAtlas

	Beagle imputation result				Minimac3 imputation results			
	Number of imputed SNP (mean ± SD)	Increased fold	CR (mean ± SD)	R^2 (mean ± SD)	Number of imputed SNP (mean ± SD)	Increased fold	CR (mean ± SD)	R^2 (mean ± SD)
Atlantic salmon	300 812 ± 5 358	6.02	0.898 ± 0.004	0.786 ± 0.006	358 996 ± 8 325	7.18	0.917 ± 0.003	0.823 ± 0.006
Cat	2 606 967 ± 86 996	52.14	0.963 ± 0.005	0.925 ± 0.010	2 606 797 ± 87 847	52.13	0.964 ± 0.004	0.927 ± 0.010
Cattle	844 012 ± 24 298	16.88	0.922 ± 0.008	0.848 ± 0.015	807 480 ± 14 490	16.15	0.929 ± 0.007	0.858 ± 0.015
Chicken	3 027 639 ± 10 464	60.55	0.865 ± 0.003	0.757 ± 0.007	563 230 ± 20 365	11.26	0.900 ± 0.002	0.802 ± 0.005
Chimpanzee	528 060 ± 168 837	10.56	0.939 ± 0.011	0.834 ± 0.036	721 530 ± 188 603	14.42	0.942 ± 0.010	0.847 ± 0.021
Dog	600 387 ± 5 745	12.01	0.919 ± 0.002	0.818 ± 0.004	419 551 ± 13 003	8.40	0.916 ± 0.002	0.827 ± 0.003
Duck	1971 592 ± 24 294	39.43	0.874 ± 0.002	0.674 ± 0.008	441 533 ± 11 707	8.83	0.904 ± 0.003	0.754 ± 0.013
Goat	1 385 117 ± 26 340	27.70	0.941 ± 0.002	0.880 ± 0.005	1 357 492 ± 28 762	27.15	0.950 ± 0.002	0.894 ± 0.006
Honey bee	654 235 ± 43 314	13.08	0.892 ± 0.013	0.680 ± 0.023	1 039 021 ± 42 151	20.78	0.889 ± 0.016	0.671 ± 0.033
Horse	1 032 480 ± 7 339	20.65	0.959 ± 0.004	0.912 ± 0.010	1 073 550 ± 6 620	21.47	0.970 ± 0.004	0.935 ± 0.010
Large yellow croaker	812 871 ± 9 196	16.26	0.882 ± 0.003	0.706 ± 0.008	131 457 ± 1 908	2.63	0.955 ± 0.001	0.882 ± 0.003
Mouse	2 168 790 ± 50 637	43.38	0.914 ± 0.012	0.871 ± 0.009	1 661 101 ± 79 140	33.22	0.923 ± 0.014	0.884 ± 0.014
Pig	2 188 342 ± 34 196	43.77	0.941 ± 0.004	0.877 ± 0.009	21 88 129 ± 34 957	43.76	0.940 ± 0.006	0.874 ± 0.011
Rabbit	1 712 166 ± 184 206	34.24	0.875 ± 0.043	0.827 ± 0.075	839 562 ± 162 247	16.79	0.893 ± 0.022	0.845 ± 0.039
Rat	645 725 ± 31 965	12.91	0.873 ± 0.017	0.758 ± 0.044	436 016 ± 31 637	8.72	0.901 ± 0.021	0.810 ± 0.051
Rhesus monkey	729 231 ± 13 042	14.58	0.915 ± 0.002	0.821 ± 0.007	995 381 ± 168 223	19.91	0.905 ± 0.004	0.809 ± 0.008
Rock pigeon	1 214 581 ± 12 933	24.29	0.903 ± 0.010	0.801 ± 0.023	-	-	-	-
Sheep	1 720 612 ± 362 379	34.41	0.892 ± 0.012	0.750 ± 0.010	1 014 694 ± 188 819	20.29	0.937 ± 0.012	0.822 ± 0.009
Vervet monkey	1 705 331 ± 72 461	34.11	0.947 ± 0.005	0.886 ± 0.014	1 517 875 ± 65 250	30.36	0.948 ± 0.005	0.892 ± 0.013
Water buffalo	2 445 894 ± 22 610	48.92	0.917 ± 0.003	0.877 ± 0.003	1 604 571 ± 35 710	32.09	0.927 ± 0.002	0.894 ± 0.006

-: the genome of Rock pigeon is at scaffold level, which Minimac3 is not compatible.

local pipeline were provided for LD calculating and visualization.

Database architecture

All processed data in the Animal-SNPAtlas was stored in MongoDB (version 3.6.8). The Animal-SNPAtlas website was built with Flask (version 1.1.1) framework and AngularJS (version 1.6.1), hosting on the Apache 2 webserver (version 2.4.18). Animal-SNPAtlas is freely available and does not require registration or login for access. Users can access it via http://gong_lab.hzau.edu.cn/Animal_SNPAtlas/.

DATABASE CONTENT AND THE WEB INTERFACE

Samples in Animal-SNPAtlas

In Animal-SNPAtlas, ∼499 million high-quality SNPs across 4784 samples of 20 animal species were included. Ranging from 1.28 million SNPs in honey bee to 61.06 million SNPs in dog, and from 40 samples in mouse to 658 samples in dog (Table 1). The detailed statistics of the sample size per species, reference genome versions, the number of SNPs and annotated genes are displayed on the 'Help' page. Moreover, the basic species information, including genome size, and chromosome number of each species, are presented in the 'Species information' module on the 'Imputation' page. Furthermore, the detailed sample information of each species is provided in the 'Sample information' section of the 'SNP' page. The introduction of samples and population structure for each species are provided. Specifically, the sample information, including the PubMed ID, publication journal, the publication year of the article, the bioproject included in NCBI, technology, platform, data type, coverage of the sequencing of the project and breed information was listed.

Website interface of Animal-SNPAtlas

The Animal-SNPAtlas database provides a user-friendly interface. It contains five main modules, namely, 'SNP' for searching and browsing the functional annotation of SNPs; 'Imputation' for running online genotype imputation by two imputation tools: Beagle and Minimac3; 'Linkage Disequilibrium' for searching, browsing, plotting and downloading the LD matrix of the corresponding gene or region; 'Visualization' for visualizing detailed information of variants with the genome browser; and 'Download' for downloading high-quality SNPs in two different formats, VCF format and M3VCF format (Figure 2A). Users can access the five modules by clicking the corresponding buttons in the navigation menu on the 'Home' page. Animal-SNPAtlas provides detailed supporting documentation and is open to any feedback with the email address listed on the 'Help' page.

SNP functional annotation in Animal-SNPAtlas

The 'SNP' module provides an advanced search box for different species, and users can search and browse SNP functional annotation based on the genomic region. SNPs

can be browsed by inputting the specific chromosomal region and selecting the MAF. Fuzzy queries are applied in the search procedure, and the query results are displayed in a table with the basic SNP information, including the chromosome position, allele, and MAF. By clicking the 'Details' button, the users can obtain SNP annotation information, including the location of each variant relative to genes, such as exonic, intronic, intergenic, splicing site, 5′/3′-UTR, upstream/downstream of genes, and functions, such as amino acid changes that may be caused by the variant.

For example, when users select 'Cattle' and enter 'Chr1:138654–217226' in the 'Region' box, the query results will be returned and they can be exported as a tab-separated file and saved by clicking the 'Download' button (Figure 2B). In addition, SNP annotation information will be displayed by clicking the 'Details' button as shown in Figure 2C.

Genotype imputation online in Animal-SNPAtlas

Genotype imputation can dramatically increase the SNP density, boost the power to detect significant associations and facilitate to identify causal variants of complex traits. In Animal-SNPAtlas, users can access the 'Imputation' module by clicking 'Imputation' in the 'Home' page navigation menu. The 'Imputation' module provides two popular imputation tools (Beagle v5.4 and Minimac3), and the genotype data with normal VCF format are entered into the text box or uploaded directly through the 'Choose File' button. After uploading the candidate genotype data, users are required to select one of the two tools, enter the chromosome region, and click the 'Submit' button to finish the query (Figure 2D). Then, the imputation results can be downloaded freely.

LD matrix search and visualization in Animal-SNPAtlas

The LD matrixes among SNPs are important information in population-based genetic studies. In Animal-SNPAtlas, 'Linkage Disequilibrium' module was developed to display the LD pattern among SNPs. On the 'Linkage Disequilibrium' page, users can input an Ensembl ID (e.g. ENSBTAG00000009188) or gene ID (e.g. *ADRB3*) to search for the LD matrix. Besides, users can also browse LD Information based on the genomic region (e.g. Chr1: 883311–883320) (Figure 3A). The query results are displayed in two tables, which include 'Gene Information' and 'LD Information', respectively. The 'Gene Information' table contains the 'chromosome position', 'Ensembl ID', 'Gene Name ID' and 'Gene Biotype'. The 'LD Information' table shows the chromosome position of SNP pairs, D', R2, and the distance of SNP pairs.

For example, when users select 'Cattle' and enter 'ENSBTAG00000009188' in the 'Ensembl ID' box (Figure 3A), the query results will be returned as shown in Figure 3B. In addition, the 'Gene Information' table also provides an 'LD plot' button, and users can click on the button to view the heatmap of LD and haplotype blocks. The triangular correlation heatmap was generated using LDBlockShow (65) with the mean of R^2 values (Figure 3C).

SNP visualization in Animal-SNPAtlas

Animal-SNPAtlas has embedded an interactive and user-friendly genome browser server (JBrowse), on which users can visualize detailed information of variants in the curated species by inputting a region of the genome (Figure 4A). Within the JBrowse browser window (right part), the 'Available tracks' menu gives access to additional data concerning each species, including reference genome, reference genome annotation, and reference panels. The appropriate resolution is automatically chosen by JBrowse depending on the size of the visualized genomic region, and tracks are distinguished by colors. Unfiltered transcripts are shown in yellow, and genes are displayed in blue. By left-clicking on the specific gene, users can gain access to the detailed functional annotation of the corresponding gene. A yellow site indicates a SNP, users can access the detailed information of SNP by left-clicking on the yellow site (Figures 4B to C).

High-quality SNPs download in Animal-SNPAtlas

The high-quality SNPs for each species are publicly available on the 'Download' page of Animal-SNPAtlas. Users can enter the genomic region of interest in the 'Region' box to obtain the corresponding VCF and M3VCF file formats (text and binary) according to their own tool requirements. Animal-SNPAtlas provides a total of ∼678G of data for users to download.

SUMMARY AND FUTURE DIRECTIONS

Rapid progress has been achieved in animal genomic research in recent decades. Some animal-related databases have been widely used by animal researchers (31–34). However, studies on the systematical characterization of SNPs, function annotation and genome-wide LD calculation in most animals are missing. In this study, we systematically identified ∼ 499 million high-quality SNPs of 4784 samples from 20 animal species, construct high-density reference panels, and perform SNP annotation and LD calculation. We further developed Animal-SNPAtlas, a versatile animal genetic variant database: (i) it presents ∼499 million SNP annotation information and ∼446 thousand annotated genes, which is convenient for researchers to study the function of candidate causal variants and their related genes; (ii) it provides 20 high-quality reference panels, users can download them for local imputation or perform online genotype imputation directly; (iii) it calculates genome-wide LD using the constructed reference panels for each species, and provides a heatmap of LD, which could assist the user to delineate candidate causal variants and increase the power of fine-mapping studies and (iv) it integrates a Genome Browser service, which provides all information about variants and genes in a specific region of genome. We believe that these functions will greatly facilitate users to conduct animal genomics, genetics and breeding studies.

Although Animal-SNPAtlas collected and analyzed the large-scale WGS data of multiple animals, the sample sizes of some species are still relatively small and only common SNPs (MAF $\geq$ 1%) have been studied. Due to limited sample size of some breeds, we did not include all the available breeds, which may decrease the imputation accuracy when

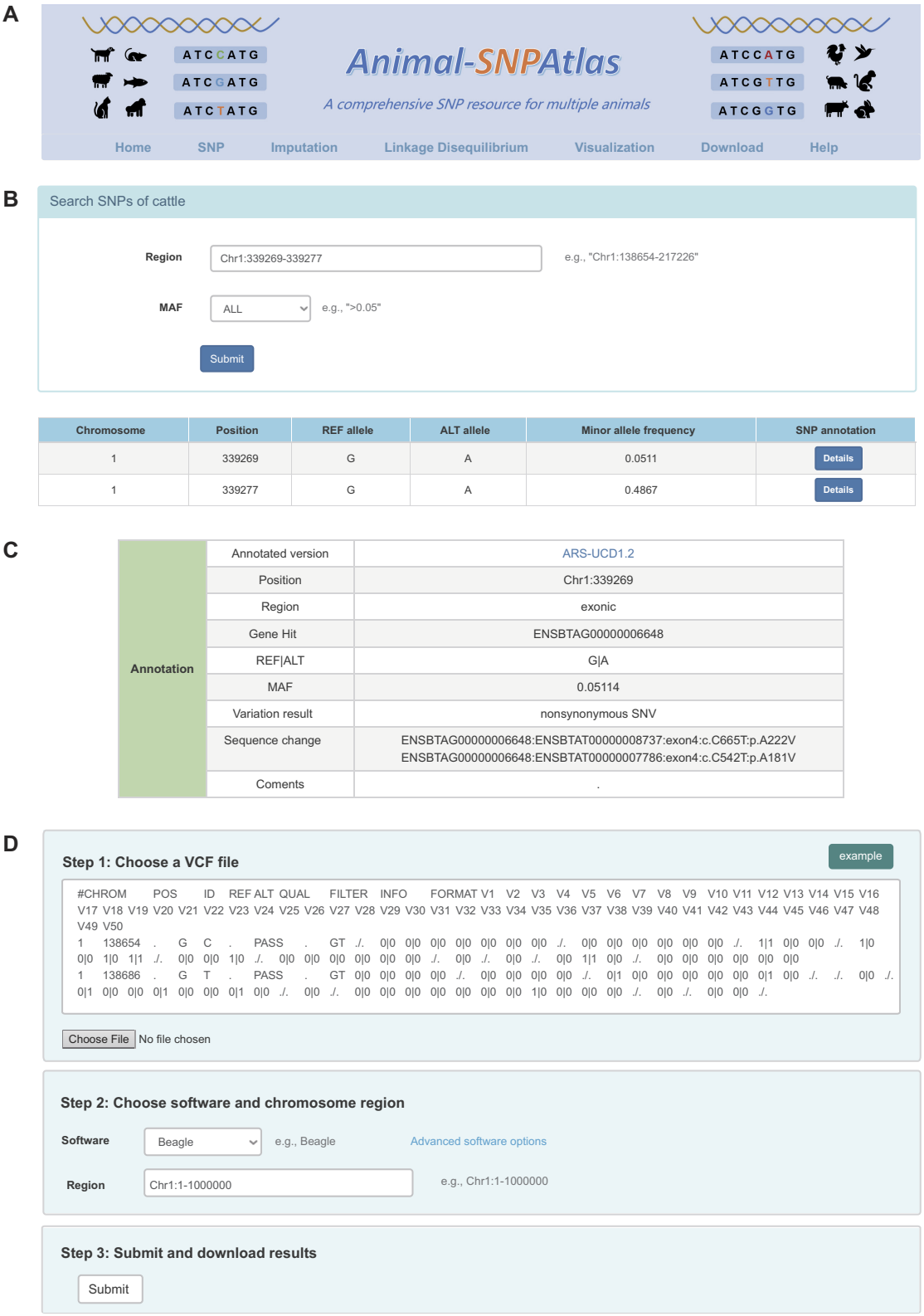

Figure 2. Overview of the Animal-SNPAtlas database. (**A**) Main modules in Animal-SNPAtlas, including 'SNP', 'Imputation', 'Linkage Disequilibrium', 'Visualization' and 'Download' modules. (**B**) The search results of SNPs based on genomic region. (**C**) SNP annotation information. (**D**) Online genotype imputation in the Animal-SNPAtlas database.

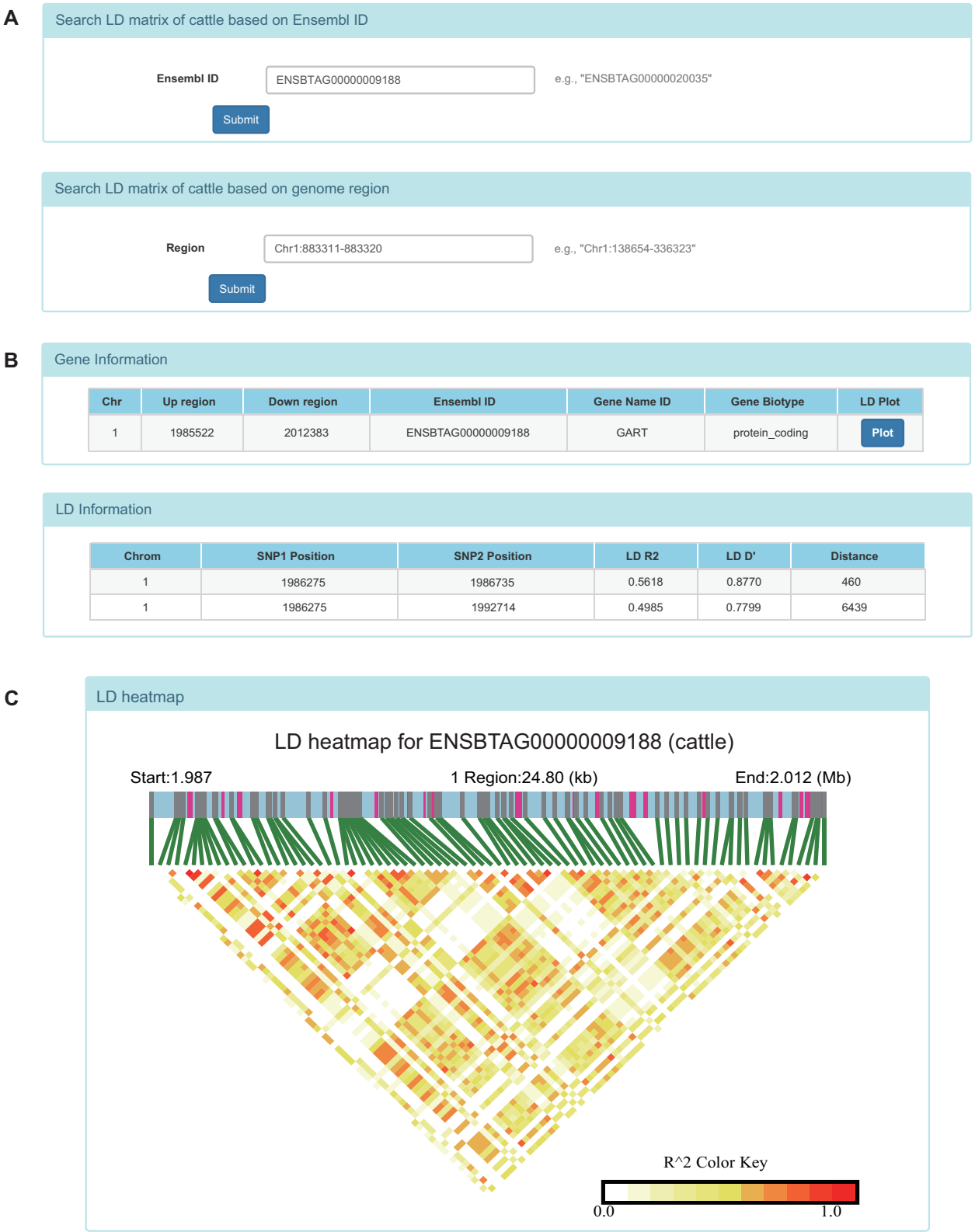

Figure 3. Visualization of the 'Linkage Disequilibrium' module. (**A**) The search box of Linkage Disequilibrium (LD) based on gene ID and genomic region. (**B**) The search results of LD based on gene ID. (**C**) The triangular correlation heatmap of the LD matrix.

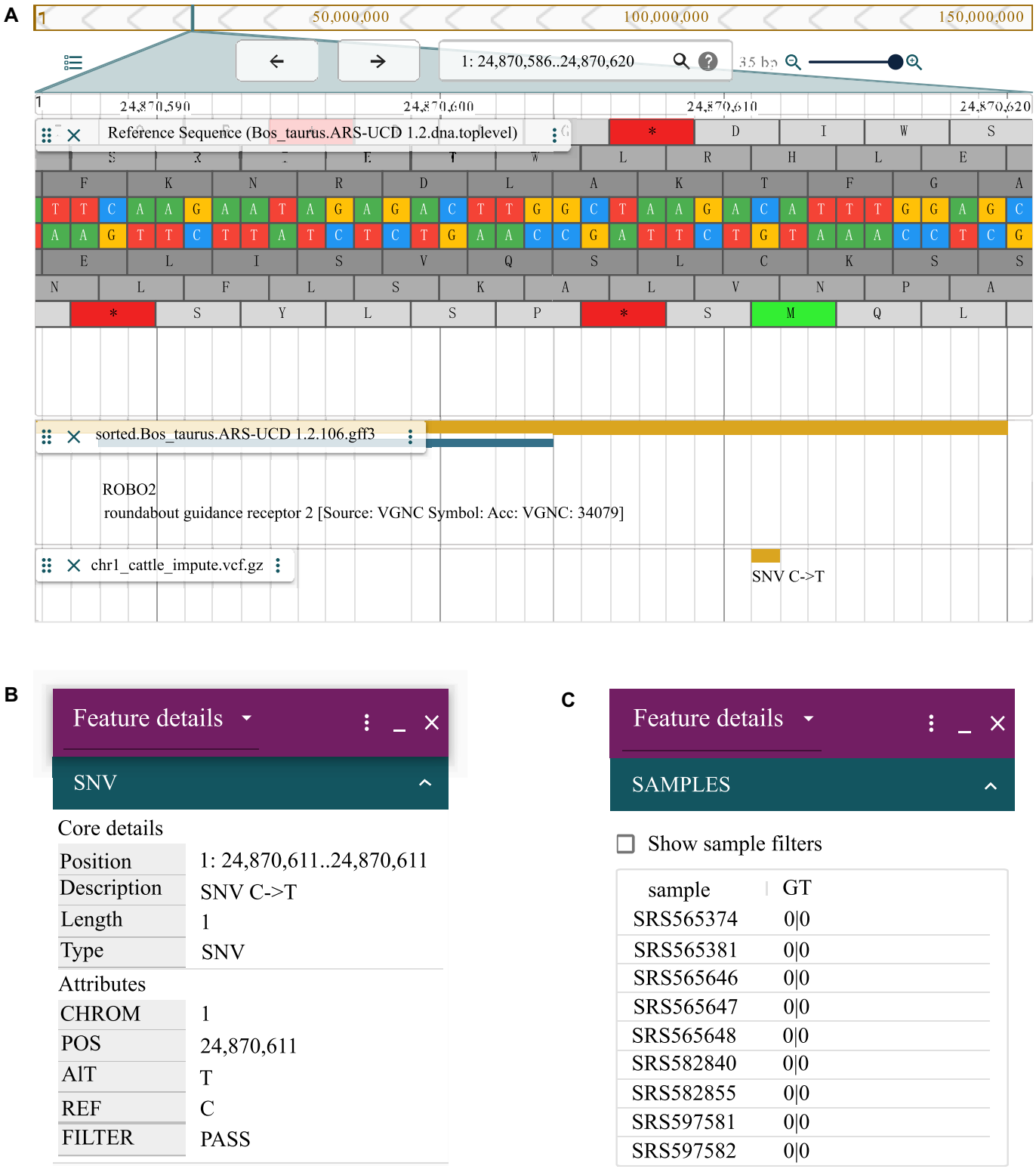

Figure 4. JBrowse Visualization of SNPs in the Animal-SNPAtlas database. (**A**) Display the context of JBrowse in specific genomic regions. (**B**, **C**) Information of the specific SNV.

the breed of target panel do not include in our reference panel. Furthermore, more than half of the species selected by Animal-SNPAtlas were mammals, which limits the application of Animal-SNPAtlas in aquaculture and livestock farming. In the future, we will continue to collect available data, and keep the database updated annually by expanding variation types and new species such as aquaculture animals, increasing more representative samples of current species, and developing additional functions, especially epigenetic data from methylation, CHIP-seq and ATAC-seq. Overall, we will maintain Animal-SNPAtlas as an important fundamental resource for the animal genetic research community.

DATA AVAILABILITY

Animal-SNPAtlas is freely available to the public without registration or login requirements (http://gong_lab.hzau.edu.cn/Animal_SNPAtlas/).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Key R&D Program of China [2021YFF0703703]; Scientific & Technological Self-innovation Foundation [11041810351]; National Natural Science Foundation of China [32072698]. Funding for open access charge: National Key R&D Program of China [2021YFF0703703 to J.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Sachidanandam,R., Weissman,D., Schmidt,S.C., Kakol,J.M., Stein,L.D., Marth,G., Sherry,S., Mullikin,J.C., Mortimore,B.J., Willey,D.L. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.
2. Altshuler,D.M., Gibbs,R.A., Peltonen,L., Altshuler,D.M., Gibbs,R.A., Peltonen,L., Dermitzakis,E., Schaffner,S.F., Yu,F., Peltonen,L. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
3. Auton,A., Brooks,L.D., Durbin,R.M., Garrison,E.P., Kang,H.M., Korbel,J.O., Marchini,J.L., McCarthy,S., McVean,G.A. and Abecasis,G.R. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
4. McCarthy,S., Das,S., Kretzschmar,W., Delaneau,O., Wood,A.R., Teumer,A., Kang,H.M., Fuchsberger,C., Danecek,P., Sharp,K. *et al.* (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.*, **48**, 1279–1283.
5. Pan,Q., Liu,Y.J., Bai,X.F., Han,X.L., Jiang,Y., Ai,B., Shi,S.S., Wang,F., Xu,M.C., Wang,Y.Z. *et al.* (2021) VARAdb: a comprehensive variation annotation database for human. *Nucleic Acids Res.*, **49**, D1431–D1444.
6. Huang,L., Rosen,J.D., Sun,Q., Chen,J., Wheeler,M.M., Zhou,Y., Min,Y.I., Kooperberg,C., Conomos,M.P., Stilp,A.M. *et al.* (2022) TOP-LD: a tool to explore linkage disequilibrium with TOPMed whole-genome sequence data. *Am. J. Hum. Genet.*, **109**, 1175–1181.
7. Lin,S.H., Brown,D.W. and Machiela,M.J. (2020) LDtrait: an online tool for identifying published phenotype associations in linkage disequilibrium. *Cancer Res.*, **80**, 3443–3446.
8. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
9. Abecasis,G.R., Altshuler,D., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
10. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Fritz,M.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
11. Piñero,J., Bravo,À., Queralt-Rosinach,N., Gutiérrez-Sacristán,A., Deu-Pons,J., Centeno,E., García-García,J., Sanz,F. and Furlong,L.I. (2017) DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.*, **45**, D833–D839.
12. Battle,A., Brown,C.D., Engelhardt,B.E. and Montgomery,S.B. (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
13. 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.
14. Wu,Y., Eskin,E. and Sankararaman,S. (2020) A unifying framework for imputing summary statistics in genome-wide association studies. *J. Comput. Biol.*, **27**, 418–428.
15. Marchini,J. and Howie,B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.
16. Nosková,A., Bhati,M., Kadri,N.K., Crysnanto,D., Neuenschwander,S., Hofer,A. and Pausch,H. (2021) Characterization of a haplotype-reference panel for genotyping by low-pass sequencing in swiss large white pigs. *BMC Genomics*, **22**, 290.
17. Fernandes Júnior,G.A., Carvalheiro,R., de Oliveira,H.N., Sargolzaei,M., Costilla,R., Ventura,R.V., Fonseca,L.F.S., Neves,H.H.R., Hayes,B.J. and de Albuquerque,L.G. (2021) Imputation accuracy to whole-genome sequence in nellore cattle. *Genet. Sel. Evol.*, **53**, 27.
18. Yang,X., Sun,J., Zhao,G., Li,W., Tan,X., Zheng,M., Feng,F., Liu,D., Wen,J. and Liu,R. (2021) Identification of major loci and candidate genes for meat production-related traits in broilers. *Front. Genet.*, **12**, 645107.
19. Yoshida,G.M. and Yáñez,J.M. (2021) Multi-trait GWAS using imputed high-density genotypes from whole-genome sequencing identifies genes associated with body traits in nile tilapia. *BMC Genomics*, **22**, 57.
20. Uffelmann,E., Huang,Q.Q., Munung,N.S., de Vries,J., Okada,Y., Martin,A.R., Martin,H.C., Lappalainen,T. and Posthuma,D. (2021) Genome-wide association studies. *Nat. Rev. Methods Primers*, **1**, 59.
21. Rosenberg,N.A., Huang,L., Jewett,E.M., Szpiech,Z.A., Jankovic,I. and Boehnke,M. (2010) Genome-wide association studies in diverse populations. *Nat. Rev. Genet.*, **11**, 356–366.
22. Zaitlen,N., Paşaniuc,B., Gur,T., Ziv,E. and Halperin,E. (2010) Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.*, **86**, 23–33.
23. Wang,X., Liu,X., Sim,X., Xu,H., Khor,C.C., Ong,R.T., Tay,W.T., Suo,C., Poh,W.T., Ng,D.P. *et al.* (2012) A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations. *Eur. J. Hum. Genet.*, **20**, 469–475.
24. Schaid,D.J., Chen,W. and Larson,N.B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491–504.
25. Broekema,R.V., Bakker,O.B. and Jonkers,I.H. (2020) A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biol.*, **10**, 190221.
26. Xiang,R., MacLeod,I.M., Daetwyler,H.D., de Jong,G., O'Connor,E., Schrooten,C., Chamberlain,A.J. and Goddard,M.E. (2021) Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. *Nat. Commun.*, **12**, 860.
27. Oliveira,H.C., Derks,M.F.L., Lopes,M.S., Madsen,O., Harlizius,B., van Son,M., Grindflek,E.H., Gòdia,M., Gjuvsland,A.B., Otto,P.I. *et al.* (2022) Fine mapping of a major backfat QTL reveals a causal regulatory variant affecting the CCND2 gene. *Front. Genet.*, **13**, 871516.
28. Li,M.J., Sham,P.C. and Wang,J. (2012) Genetic variant representation, annotation and prioritization in the post-GWAS era. *Cell Res.*, **22**, 1505–1508.

29. Hou,L. and Zhao,H. (2013) A review of post-GWAS prioritization approaches. *Front. Genet.*, **4**, 280.
30. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
31. Jin,W., Zhu,Q., Yang,Y., Yang,W., Wang,D., Yang,J., Niu,X., Yu,D. and Gong,J. (2021) Animal-APAdb: a comprehensive animal alternative polyadenylation database. *Nucleic Acids Res.*, **49**, D47–D54.
32. Jin,W., Jiang,G., Yang,Y., Yang,J., Yang,W., Wang,D., Niu,X., Zhong,R., Zhang,Z. and Gong,J. (2022) Animal-eRNAdb: a comprehensive animal enhancer RNA database. *Nucleic Acids Res.*, **50**, D46–D53.
33. 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.*, **47**, D33–D38.
34. Hu,Z.L., Park,C.A. and Reecy,J.M. (2022) Bringing the animal QTLdb and CorrDB into the future: meeting new challenges and providing updated services. *Nucleic Acids Res.*, **50**, D956–D961.
35. 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.
36. Fu,Y., Xu,J., Tang,Z., Wang,L., Yin,D., Fan,Y., Zhang,D., Deng,F., Zhang,Y., Zhang,H. *et al.* (2020) A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model. *Commun Biol*, **3**, 502.
37. Chen,N., Fu,W., Zhao,J., Shen,J., Chen,Q., Zheng,Z., Chen,H., Sonstegard,T.S., Lei,C. and Jiang,Y. (2020) BGVD: an integrated database for bovine sequencing variations and selective signatures. *Genomics Proteomics Bioinformatics*, **18**, 186–193.
38. Fu,W., Wang,R., Xu,N., Wang,J., Li,R., Asadollahpour Nanaei,H., Nie,Q., Zhao,X., Han,J., Yang,N. *et al.* (2022) Galbase: a comprehensive repository for integrating chicken multi-omics data. *BMC Genomics*, **23**, 364.
39. Wang,Z.H., Zhu,Q.H., Li,X., Zhu,J.W., Tian,D.M., Zhang,S.S., Kang,H.L., Li,C.P., Dong,L.L., Zhao,W.M. *et al.* (2021) iSheep: an integrated resource for sheep genome, variant and phenotype. *Front. Genet.*, **12**, 714852.
40. Fu,W., Wang,R., Yu,J., Hu,D., Cai,Y., Shao,J. and Jiang,Y. (2021) GGVD: a goat genome variation database for tracking the dynamic evolutionary process of selective signatures and ancient introgressions. *J. Genet. Genomics*, **48**, 248–256.
41. Zeng,Q., Zhao,B., Wang,H., Wang,M., Teng,M., Hu,J., Bao,Z. and Wang,Y. (2022) Aquaculture molecular breeding platform (AMBP): a comprehensive web server for genotype imputation and genetic analysis in aquaculture. *Nucleic Acids Res.*, **50**, W66–W74.
42. 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.
43. Hu,Y.B., Wu,Q., Ma,S., Ma,T.X., Shan,L., Wang,X., Nie,Y.G., Ning,Z.M., Yan,L., Xiu,Y.F. *et al.* (2017). Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 1081–1086.
44. Poot,M., Badea,A., Williams,R.W. and Kas,M.J. (2011). Identifying human disease genes through cross-species gene mapping of evolutionary coserved process. *PLoS One*, **6**, e18612.
45. Malsen,A.M., Vinkers,C.H., Peterse,D.P., Olivier,B. and Kas,M.J. (2011). Cross-species behavioural genetics: a starting point for unravelling the neurobiology of human psychiatric disorder. *Prog. Neuropsychopharmacol. Biol. Psychiatry*, **35**, 1381–1390.
46. Johnson,R.A., Wright,K.D., Poppleton,H., Mohankumar,K.M., Fnkelstein,D., Pounds,S.B., Rand,V., Leary,S.E., White,E., Eden,C. *et al.* (2010). Cross-species genomics matches driver mutations and cell compartments to model ependymoma. *Nature*, **466**, 632–636.
47. Wong,K., Weyden,L.V., Schott,C.R., Foote,A., Constantino-Casas,F., Smith,S., Dobson,J.M., Murchison,E.P., Wu,H., Yeh,I. *et al.* (2019). Cross-species genomic landscape comparison of human mucosal melanoma with canine oral and equine melanoma. *Nat. Commun.*, **10**, 353.
48. Graeber,T.G. and Sawyers,C.L. (2005). Cross-species comparisons of cancer signaling. *Nat. Genet.*, **37**, 7–8.
49. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
50. Cook,C.E., Lopez,R., Stroe,O., Cochrane,G., Brooksbank,C., Birney,E. and Apweiler,R. (2019) The european bioinformatics institute in 2018: tools, infrastructure and training. *Nucleic Acids Res.*, **47**, D15–D22.
51. Plassais,J., Kim,J., Davis,B.W., Karyadi,D.M., Hogan,A.N., Harris,A.C., Decker,B., Parker,H.G. and Ostrander,E.A. (2019) Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat. Commun.*, **10**, 1489.
52. Howe,K.L., Achuthan,P., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J. *et al.* (2021) Ensembl 2021. *Nucleic Acids Res.*, **49**, D884–D891.
53. Kendig,K.I., Baheti,S., Bockol,M.A., Drucker,T.M., Hart,S.N., Heldenbrand,J.R., Hernaez,M., Hudson,M.E., Kalmbach,M.T., Klee,E.W. *et al.* (2019) Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.*, **10**, 736.
54. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.
55. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
56. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.
57. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
58. 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.
59. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
60. Alexander,D.H., Novembre,J. and Lange,K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
61. Yang,J., Lee,S.H., Goddard,M.E. and Visscher,P.M. (2013) Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol. Biol.*, **1019**, 215–236.
62. Browning,B.L., Zhou,Y. and Browning,S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, **103**, 338–348.
63. Das,S., Forer,L., Schönherr,S., Sidore,C., Locke,A.E., Kwong,A., Vrieze,S.I., Chew,E.Y., Levy,S., McGue,M. *et al.* (2016) Next-generation genotype imputation service and methods. *Nat. Genet.*, **48**, 1284–1287.
64. Zhang,C., Dong,S.S., Xu,J.Y., He,W.M. and Yang,T.L. (2019) PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics*, **35**, 1786–1788.
65. Dong,S.S., He,W.M., Ji,J.J., Zhang,C., Guo,Y. and Yang,T.L. (2021) LDBlockShow: a fast and convenient tool for visualizing linkage disequilibrium and haplotype blocks based on variant call format files. *Brief. Bioinform.*, **22**, bbaa227.