

# Plant-ImputeDB: an integrated multiple plant reference panel database for genotype imputation

Yingjie Gao<sup>1,†</sup>, Zhiquan Yang<sup>1,†</sup>, Wenqian Yang<sup>1,†</sup>, Yanbo Yang<sup>1</sup>, Jing Gong<sup>1,2,\*</sup>, Qing-Yong Yang<sup>1,3,\*</sup> and Xiaohui Niu<sup>1,\*</sup>

<sup>1</sup>Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P.R. China, <sup>2</sup>College of Biomedicine and Health, Huazhong Agricultural University, Wuhan 430070, P.R. China and <sup>3</sup>College of Agriculture, Shihezi University, Xinjiang 832003, P.R. China

Received August 04, 2020; Revised September 23, 2020; Editorial Decision October 06, 2020; Accepted October 08, 2020

## ABSTRACT

Genotype imputation is a process that estimates missing genotypes in terms of the haplotypes and genotypes in a reference panel. It can effectively increase the density of single nucleotide polymorphisms (SNPs), boost the power to identify genetic association and promote the combination of genetic studies. However, there has been a lack of high-quality reference panels for most plants, which greatly hinders the application of genotype imputation. Here, we developed Plant-ImputeDB ([http://gong\\_lab.hzau.edu.cn/Plant\\_imputeDB/](http://gong_lab.hzau.edu.cn/Plant_imputeDB/)), a comprehensive database with reference panels of 12 plant species for online genotype imputation, SNP and block search and free download. By integrating genotype data and whole-genome resequencing data of plants from various studies and databases, the current Plant-ImputeDB provides high-quality reference panels of 12 plant species, including ~69.9 million SNPs from 34 244 samples. It also provides an easy-to-use online tool with the option of two popular tools specifically designed for genotype imputation. In addition, Plant-ImputeDB accepts submissions of different types of genomic variations, and provides free and open access to all publicly available data in support of related research worldwide. In general, Plant-ImputeDB may serve as an important resource for plant genotype imputation and greatly facilitate the research on plant genetic research.

## INTRODUCTION

Natural variation as a primary resource to study the genetic basis for phenotypic differences among different individuals

of the same species, which mainly includes single nucleotide polymorphisms (SNPs) and genomic structural variations (1). In plants, SNPs are major variations widely used in genetic breeding and population evolution research (2–6). In recent years, with the development of sequencing and genotyping technologies, the cost of whole-genome resequencing (WGS) and genotyping has been declining (7), and large amounts of population genotype data from different species have been continuously released, facilitating the wide application of genetic linkage analysis or genome-wide association analysis (GWAS) in the research of different species (2–5). High-density markers of mass samples are conducive to increase statistical power, boost fine mapping of causal variants and facilitate the discovery of relationship between rare variants and traits (8,9). But due to the cost limitations, only a subset of SNPs is directly genotyped by SNP-chips or DNA sequencing in study samples (10). So, genotype imputation was developed to use the haplotypes and genotypes in a reference panel to estimate genotypes that not directly assayed in a sample of individuals and has been one of the key steps in preprocessing genetic data (10).

The basic idea of the genotype imputation methods is to explore and hunt for shared ‘identical by descent’ haplotypes that exhibit high linkage disequilibrium measured in  $r^2$  from a high-density typed reference panel of genotypes or haplotypes over a region of tightly linked markers, and use them to fill untyped SNPs (11). According to the idea, several imputation methods have been developed in recent years, such as Beagle (v5.1) (12) and Minimac3 (13) both based on common hidden Markov model framework (14,15), and Impute2 (16) based on a Markov chain Monte Carlo framework. Increasing evidence demonstrated the advantages of genotype imputation and it has become a standard step in GWAS and other genetic research because it is an economic and efficient way to acquire high-density population genotype data from SNP array, genotyping-by-sequencing (GBS) or reduced-representation sequenc-

\*To whom correspondence should be addressed. Tel: +86 027 87285085; Fax: +86 027 87285085; Email: niuxiaoh@mail.hzau.edu.cn  
Correspondence may also be addressed to Qing-Yong Yang. Tel: +86 027 87285085; Fax: +86 027 87285085; Email: yqy@mail.hzau.edu.cn  
Correspondence may also be addressed to Jing Gong. Tel: +86 027 87285085; Fax: +86 027 87285085; Email: gong.jing@mail.hzau.edu.cn

<sup>†</sup>The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

ing commonly used in plant research (17,18). For example, it is clear that the probability of detecting the phenotype associated SNPs with genotype imputation (8.9%) is much greater than that without genotype imputation (5.4%) at the significance level of  $P < 10^{-6}$  in  $\chi^2$  statistics, indicating that genotype imputation can greatly improve the power of GWAS (19). In an association analysis of the *indica* population, eight peaks for amylose content on chromosome 6 were detected using the imputed data, including the regions containing *Wx* and *SSII*, while three of these associations could not be detected using the original unimputed data (20). However, the challenges for genotype imputation methods will be in preparing large enough, diverse enough set of haplotypes available for constructing reference panel, and the imputation accuracy will decrease when new accessions that are not well-represented in the reference panel (21). In addition, it is still difficult to correctly impute rare variants under the current imputation framework and mainstream imputation methods (22). A high-quality reference panel is not only the essential prerequisite for genotype imputation but also play a crucial role for the imputation quality.

Benefit from the construction of large reference panel for genotype imputation and the development of genotype imputation methods, genotype imputation is widely used in human genetic studies (21,23–26). In human, the commonly used public reference panels mainly include International HapMap Project Phase3 (27), 1000 Genomes Project Phase 3 (1) and Haplotype Reference Consortium (28). International HapMap Project Phase3 comprises 1011 samples and 1.4 million variants (27); 1000 Genomes Project Phase 3 includes 81.7 million variants and 2504 samples of 26 populations (1); and Haplotype Reference Consortium integrates 20 studies to develop a human reference panel that includes 32 470 samples and 40.4 million variants (28). In animals, Animal-ImputeDB comprises 2565 samples of 13 species and over 400 million variants (29). Construction of these large reference panels makes it possible to acquire high-density genetic markers from low-density data, and untyped variants can be accurately imputed at low minor allele frequencies (MAFs), provided that they are first observed in the reference population (30). Recently, an imputation platform has been established for rice, which allows online genotype imputation (20). However, there has been no database that provides reference panels of multiple species for plant genotype imputation to the best of our knowledge.

With the increasing availability of massive genotype data in plants and mature tools, it is possible to construct a comprehensive database with multiple plant reference panels and online imputation tools. Here, we developed the Plant Imputation database (Plant-ImputeDB, [http://gong.lab.hzau.edu.cn/Plant\\_imputeDB/](http://gong.lab.hzau.edu.cn/Plant_imputeDB/)), which comprises a collection of high-quality reference panels derived from publicly available plant genomic sequencing or genotype data, for the browsing, searching and downloading of reference panels and its related information. Through data curation, sample filtering, genotype calling and haplotype phasing, a total of 12 high-quality plant reference panels were finally built using 34 244 resequencing samples. The database includes the plants of arabidopsis, oilseed rape, common bean, cotton, cucumber, zucchini, maize, muskmelon, rice,

soybean, watermelon and bread wheat. In addition, the database offers a user-friendly online tool with the option of two popular tools to support the genotype imputation.

## DATA COLLECTION AND PROCESSING

### Data collection

With the rapid development of sequencing technology in recent years, genomic datasets of a large number of species have been constantly released and updated. In order to include the representative species as many as possible, we collected the high-quality raw sequencing and SNP datasets of 12 species from widely studied plant databases such as 1001genomes (31) (<https://www.1001genomes.org/>), Rice SNP-seek database (32) (<https://snp-seek.irri.org/>), Maize HapMap (33) (<https://www.panzea.org/>), SoyBase (34) (<https://soybase.org/snps/>), 1000 wheat exomes project (<http://wheatgenomics.plantpath.ksu.edu/1000EC/>) (35) and Cucurbit Genomics Database (36) (<http://www.cucurbitgenomics.org/>), as well as the original sequencing data published in recent years (37–39).

For 10 of 12 species, raw genotype files (VCF format) were downloaded from database or research. Among them, samples of five species (arabidopsis, common bean, maize and watermelon) were genotyped using WGS (31–33,37,40); samples of three species (cucumber, muskmelon and zucchini) were genotyped with high-throughput GBS (36,41); for bread wheat, samples were genotyped using exome capture sequencing technology (35); for soybean, samples were genotyped with SoySNP50K Illumina Infinium II BeadChip (34). For the other two species, oilseed rape and cotton, the raw sequencing datasets were downloaded from the NCBI database under accession SRP155312 and SRP115740, respectively (38,39).

Detailed information of the species, such as NCBI taxonomy ID, assembly version and SNP number, is presented in Table 1. The data sources, genotyping methods and population summaries of 12 species are presented in Supplementary Table S1.

### Data processing

With the raw sequencing data, high-quality SNPs were identified using the Sentieon pipeline (42). First, the raw reads were mapped to the current standard reference genome by the Burrows–Wheeler Alignment mem algorithm (43), and then the BAM files of reads with quality greater than 10 were retained by SAMtools (44). Alignment summary, GC bias, base quality by sequencing cycle, base quality score distribution and insert size metrics were collected, and the duplicate reads were removed with the Sentieon driver. Then, the indels were realigned, and the base quality was recalibrated using the Sentieon driver. The SNP data of each sample were identified using Sentieon's Haplotyper algorithm. Then, the variant data of all samples were merged into VCF files using Sentieon GVCFTyper algorithm. The raw SNPs of all samples were filtered using the GATK VariantFiltration module with the parameter `-filterExpression 'QUAL < 30.0 || MQ < 50.0 || QD < 2' -clusterSize 3 -clusterWindowSize 10`. Subsequently, the SNPs with a call rate < 0.5 or an MAF < 0.01 were removed. Finally, all the

**Table 1.** Data summary in Plant-ImputeDB

Species	NCBI taxonomy ID	Assembly version	Number of chromosomes	Reference panel	
				Number of samples	Number of SNPs
<i>Arabidopsis thaliana</i> (Arabidopsis)	3702	TAIR10	5	2029	2 963 242
<i>Brassica napus</i> (Oilseed rape)	3708	ZS11 v0	19	991	9 141 089
<i>Phaseolus vulgaris</i> (Common bean)	3885	PhaVulg1.0	11	628	4 811 097
<i>Gossypium hirsutum</i> (Cotton)	3635	TM-1 UTX_v2.0	26	686	3 149 846
<i>Cucumis sativus</i> (Cucumber)	3659	Cucumber (Gy14) v2	7	1234	21 154
<i>Cucurbita pepo</i> (Zucchini)	3664	Cucurbita pepo v4.1	20	830	41 888
<i>Zea mays</i> (Maize)	4577	AGPv3	10	1210	35 073 758
<i>Cucumis melo</i> (Muskmelon)	3656	Melon (DHL92) v3.5.1	12	2084	26 011
<i>Oryza sativa</i> Japonica (Rice)	39 947	IRGSP-1.0	12	3240	4 897 277
<i>Glycine max</i> (Soybean)	3847	Wm82.a2	20	20 087	39 636
<i>Citrullus lanatus</i> (Watermelon)	3654	Watermelon (97103) v2	11	414	8 816 591
<i>Triticum aestivum</i> (Bread wheat)	4565	IWGSC v1.0	21	811	942 041

high-quality SNPs that had passed the filtering were used to construct the reference panel (Figure 1). The detailed statistics of the genetic variants and sample data of each species in the final dataset are listed in Table 1. In addition, the genomic blocks of each species were also identified using Plink with the parameter `-blocks` (45).

### Reference panel construction

Beagle, Minimac3 and Impute2 are the most popular tools for genotype imputation. A comparison among the three tools shows that despite of the similarity in accuracy, they vary greatly in memory requirements and computation time. Beagle and Minimac3 are superior to Impute2 in computation time and memory efficiency (46), and support the genotype imputation of polyploid plants (47). Therefore, Beagle and Minimac3 were chosen for the construction of reference panels in this study. The reference panels of 12 species were constructed by Beagle using clean SNP data ( $MAF > 0.01$ , call rate  $> 0.5$ ) with the default parameters, and then converted from VCF to M3VCF format by Minimac3.

### Evaluation of the reference haplotype libraries

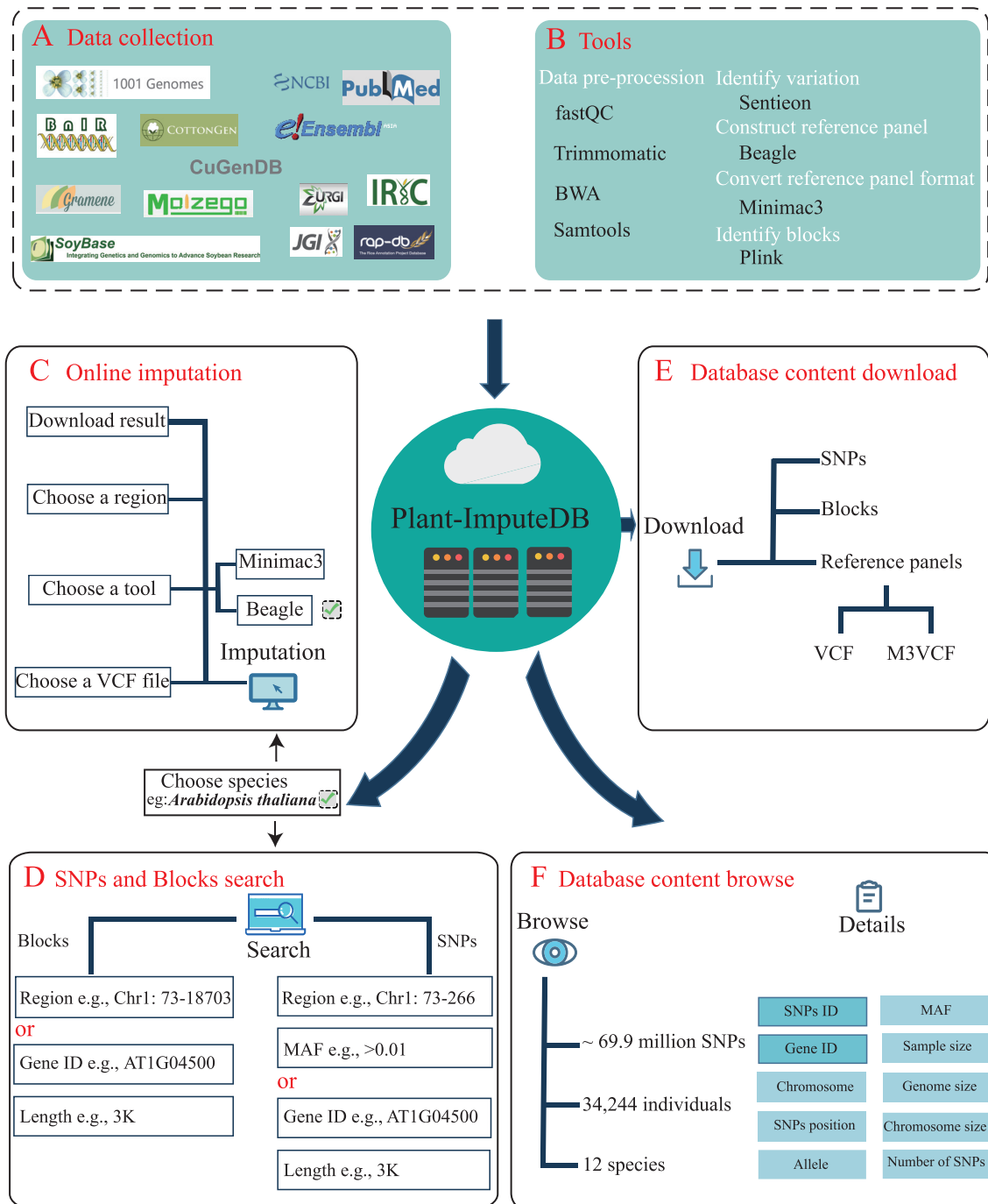
Reliable haplotypes are important for genotype phasing and imputation (48). Therefore, we followed the method of Marchini, J. *et al.* and applied switch accuracy as an index to evaluate the reliability of haplotypes (49). For simulating haplotype blocks, we referred to the method of Osabe, D. *et al.* (50). Firstly, we randomly selected 100 contiguous haplotype blocks, and all the SNPs located in them were extracted for the evaluation. Then, 100 genotyping datasets with the same population size were selected by re-sampling with replacement from original samples in reference panels. Their haplotype blocks were identified using Plink (45). The switch accuracies were obtained based on the simulation data. The average switch accuracies of the 12 species ranged from 0.92 for maize to 0.99 for watermelon, indicating the reliability of the haplotypes in our panels (Supplementary Figure S1). In addition, we calculated haplotype blocks and frequency in each species and summarized the block sizes and SNP numbers in blocks (Supplementary Table S2).

### Imputation accuracy using reference panels in Plant-ImputeDB

Performance of the reference panels and imputation process were evaluated based on three strategies. First, we calculated the imputation accuracy of all species using a 5-fold cross-validation strategy. For each species, all the samples in the reference panel were randomly divided into five folds, with one fold being selected as the study population, and the remaining folds being used as the reference panels for each time. Since most commercial SNP arrays of plants contain about 50–100 k probes (51), we randomly selected 100 000 SNPs from the whole genome of the study population and masked other SNPs. Considering that four species had a relatively small number of SNPs ( $\leq 100\ 000$ ), we randomly selected 5000 SNPs from the whole genome for these four species (Table 1). Then, Beagle and Minimac3 were used to impute the genotypes with the default parameters.

In this way, both the true and imputed genotypes were obtained, and the imputed SNPs with  $MAF \geq 0.01$  and estimated squared correlation  $\geq 0.3$  were retained as properly imputed variants and used 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 through 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 folds was taken as the accuracy of the imputation for each species, and the results are summarized in Table 2. Moreover, the corresponding boxplots are shown in Supplementary Figure S2. The number of SNPs increased by an average 34.47 folds in the study population after imputation. The average CR for all test species was greater than 0.88. The average  $R^2$  of Beagle ranged from 0.76 for melon to 0.96 for cotton, and that of Minimac3 ranged from 0.76 for melon to 0.97 for common bean.

In addition, imputation accuracies with the reference panels were assessed using simulated datasets with different densities and independent datasets respectively. First, as for 12 species in our database, we randomly selected 100 samples with 10 different percentages of masked SNPs from 50 to 95% following the simulation method of Friedrich, J. *et al.* (52). Imputation accuracy was calculated by comparing imputation results and raw genotypes. As for two



**Figure 1.** Construction of plant reference panels in Plant-ImputeDB. (A) Data collection. (B) Data processing. (C–F) Database content and web interface.

imputation tools Beagle and Minimac3, the average accuracy of all simulation datasets ranged from 0.83 to 0.99 (Supplementary Figures S3 and 4). Second, nine independent validation sets for the corresponding species in our database, including rice (53), arabidopsis (54), maize (Maize 282) (55), oilseed rape (56), cotton (57), soybean (58), cucumber (59), muskmelon (60) and bread wheat (61) were collected for assessment of imputation accuracy. These raw sequencing datasets were processed following the same Sentieon pipeline and parameters, and the missing genotypes

were imputed by Beagle with default parameters. Then, the common SNPs in independent populations and our reference panels were retained to validate imputation accuracy. The validation datasets were constructed with 10 different percentages of masked SNPs from 50 to 95%. Finally, these independent datasets were imputed using Beagle and Minimac3 with default the parameters respectively. Imputation accuracies were achieved with the true and imputed genotypes. Similarly, the average accuracy ranged from 0.77 to 0.99, and the detailed results are interpreted in Supplemen-

**Table 2.** Imputation accuracy using reference panels in Plant-ImputeDB

	Beagle imputation results			Mimic3 imputation results				
	Number of imputed SNP (mean ± SD)	Increased fold	CR (mean ± SD)	R <sup>2</sup> (mean ± SD)	Number of imputed SNPs (mean ± SD)	Increased fold	CR (mean ± SD)	R <sup>2</sup> (mean ± SD)
Arabidopsis	2 792 659 ± 5127	27.93	0.9906 ± 0.0002	0.9411 ± 0.0014	2 884 385 ± 5631	28.84	0.9912 ± 0.0002	0.9448 ± 0.0010
Oilseed rape	4 604 327 ± 69 131	46.04	0.8857 ± 0.0016	0.7717 ± 0.0037	1 412 928 ± 58 660	14.31	0.9286 ± 0.0022	0.8135 ± 0.0023
Common bean	3 289 257 ± 13 766	32.89	0.9584 ± 0.0012	0.8973 ± 0.0017	4 152 965 ± 76 476	41.53	0.9798 ± 0.0017	0.9717 ± 0.0018
Cotton	2 927 154 ± 76 601	29.27	0.9810 ± 0.0032	0.9615 ± 0.0057	2 935 382 ± 751 456	29.35	0.9848 ± 0.0084	0.9588 ± 0.0095
Maize	21 336 638 ± 142 290	213.37	0.9396 ± 0.0017	0.7996 ± 0.0069	7 827 635 ± 266 095	78.28	0.9502 ± 0.0015	0.8363 ± 0.0077
Rice	4 996 975 ± 1960	49.97	0.9538 ± 0.0009	0.9416 ± 0.0011	3 570 124 ± 64 495	35.70	0.9655 ± 0.0010	0.9420 ± 0.0016
Watermelon	8 058 314 ± 510 335	80.58	0.9861 ± 0.0040	0.8675 ± 0.0398	7 628 587 ± 468 864	76.29	0.9903 ± 0.0032	0.9102 ± 0.0375
Bread wheat	496 703 ± 121 523	4.97	0.9890 ± 0.0019	0.9534 ± 0.0036	580 923 ± 129 173	5.81	0.9878 ± 0.0019	0.9560 ± 0.0034
Cucumber	6090 ± 59	1.52	0.9332 ± 0.0021	0.8099 ± 0.0045	13 350 ± 193	3.34	0.9413 ± 0.0010	0.8210 ± 0.0066
Zucchini	17 729 ± 326	3.55	0.9081 ± 0.0027	0.7588 ± 0.0025	27 853 ± 458	5.57	0.9171 ± 0.0026	0.7712 ± 0.0030
Muskmelon	6856 ± 48	1.37	0.9043 ± 0.0007	0.7582 ± 0.0030	10 387 ± 86	2.08	0.9277 ± 0.0003	0.7602 ± 0.0014
Soybean	33 808 ± 15	6.76	0.9697 ± 0.0008	0.9099 ± 0.0024	39 453 ± 37	7.89	0.9788 ± 0.0007	0.9419 ± 0.0023

CR: concordance rate between true and imputed genotypes.  
R<sup>2</sup>: squared correlation between true and imputed genotypes.

tary Figures S5 and 6. All of these validation results indicate that the reference panels and the imputation tools can be used for genotype imputation from different population with relatively high accuracy.

## IMPLEMENTATION

Plant-ImputeDB ([http://gong.lab.hzau.edu.cn/Plant\\_imputeDB/](http://gong.lab.hzau.edu.cn/Plant_imputeDB/)) was built based on the Flask (version 1.1.1) framework with AngularJS (version 1.6.1) as the JavaScript library, and runs on the Apache 2 web server (version 2.4.18) with MongoDB (version 3.4.2) as its database engine. The database is available online without registration and optimized for Chrome (recommended), Internet Explorer, Opera, Firefox, Windows Edge and macOS Safari.

## DATABASE CONTENT AND THE WEB INTERFACE

### Samples of 12 species in Plant-ImputeDB


The current version of Plant-ImputeDB contains a total of ~69.9 million SNPs from 12 species covering 34 244 individuals. The detailed statistics of the number of samples per species, the number of chromosomes, genome version, NCBI taxonomy ID and the number of SNPs are displayed and maintained online at the home page of Plant-ImputeDB and summarized in Table 1. Besides, the basic introduction, genome size and chromosome number of each species are presented in the ‘Species information’ module, and users can access to this module by clicking the plant photo on the ‘Home’. The detailed sample information of each species is provided in the ‘Sample information’ module. The introduction of samples, population structure and the list of accessions are provided. In addition, we have provided two advanced search boxes for different species. The users can browse the information of accessions for each species according to the sub-population or country and obtain the specific accession of interest. Finally, the sample information, including the PubMed ID, publication journal, publication year of the article, the sample number, material, technology, platform, data type and coverage of the sequencing of the project, was listed as supplemental information (Supplemental Table S1).

### Web interface

A user-friendly web interface for Plant-ImputeDB was constructed, and users can access to three main modules, including Module1: ‘Imputation’ for online genotype imputation, Module2: ‘Reference Panel’ for SNP and block search based on genomic region information or gene ID, and sample information of the reference panels, and Module3: ‘Download’ for reference panel download in two formats (VCF and M3VCF). Specifically, users can access to the three modules by clicking the corresponding buttons in the navigation menu on the ‘Home’ page or by clicking the corresponding plant photo (Figure 2A). These modules provide species information as well as realize online genotype imputation, SNP search, and genomic block search (Figure 2B–E). Plant-ImputeDB provides detailed supporting documentation on the ‘Help’ page, and is open to any feedback with email address listed on the ‘Contact’ page.

**A**

**Plant-ImputeDB** Home Imputation Reference Panel Download Help Contact



**B**

Species information Online imputation SNP search Block search

**Step 1: Choose a VCF file**

In this step, users can upload a VCF file or input text to impute ungenotyped markers. The input-file or text must be in VCF format Example file format: filename.vcf. The maximum file size is 50 MB. [example](#)

Choose File

**Step 2: Choose software and chromosome region**

In this step, users can select **Minimac3** or **Beagle** to phase the observed genotypes and impute the missing genotypes in the file or text you uploaded. In the next box, users need input the chromosome region to perform imputation. If users want to impute the entire genome, we recommend them to download our reference panel and run the software locally.

Software  e.g., Beagle [Advanced software options](#)

Region  e.g., Chr1:1-1000000

**Step 3: Submit and download results**

Click the "Submit" button to perform imputation. Imputation with Beagle and Minimac3 may take 3 minutes and 12 minutes respectively. Results can be downloaded as either vcf/vcf.gz for Beagle imputation or as a tar/tar.gz file for Minimac3.

Submit

**C**

Species information Online imputation SNP search Block search

**Search SNPs of *Arabidopsis thaliana***

Region  e.g., "Chr1:73-266"

MAF  e.g., ">0.05"

Submit

Download

Chromosome	Position	Ref. allele	Alt. allele	Minor allele frequency
1	73	C	A	0.0635
1	92	A	C	0.3519
1	110	G	T	0.0645
1	125	G	T	0.0458
1	253	T	C	0.0828
1	266	G	A	0.0295

**D**

Region search Gene ID search

**Search gene ID of *Arabidopsis thaliana***

Gene ID  e.g., "AT1G04500"

Length  e.g., "3K(default)"

Submit

Download

Chr	Up region	Down region	BlockID	KB
1	1217101	1219863	ATChr1block895	2.763
1	1219980	1220415	ATChr1block896	0.436
1	1220453	1220469	ATChr1block897	0.017
1	1220638	1220819	ATChr1block898	0.182
1	1220901	1220909	ATChr1block899	0.009
1	1220933	1221893	ATChr1block900	0.961

**E**

**Download the interested region of the reference panel**

Region  e.g., "Chr1:73-2227"

VCF format  e.g., "VCF"

Download

**Download the reference panel in VCF format**

arabidopsis\_impute\_Chrl.vcf.gz (373 M) arabidopsis\_impute\_Chrl2.vcf.gz (265 M)

arabidopsis\_impute\_Chrl3.vcf.gz (355 M) arabidopsis\_impute\_Chrl4.vcf.gz (262 M)

arabidopsis\_impute\_Chrl5.vcf.gz (335 M)

**Download the reference panel in M3VCF format**

arabidopsis\_impute\_Chrl.m3vcf.gz (42 M) arabidopsis\_impute\_Chrl2.m3vcf.gz (28 M)

arabidopsis\_impute\_Chrl3.m3vcf.gz (38 M) arabidopsis\_impute\_Chrl4.m3vcf.gz (30 M)

arabidopsis\_impute\_Chrl5.m3vcf.gz (38 M)

**Figure 2.** Overview of the Plant-ImputeDB database. (A) Main modules in Plant-ImputeDB, including 'Imputation', 'Reference Panel' and 'Download' modules. (B) Online genotype imputation in the Plant-ImputeDB database. (C) Browsing of SNPs based on genomic region. (D) Browsing of genomic blocks based on gene ID. (E) 'Download' function of Plant-ImputeDB.

### Online genotype imputation in Plant-ImputeDB

Plant-ImputeDB supports two popular imputation tools (Beagle and Minimac3). The users can access the 'Imputation' module by either clicking 'Imputation' in the 'Home' page navigation menu or clicking the hyperlink in the corresponding species photo on the 'Home' page. Then, the genotype data of normal VCF format are entered into the text box or uploaded directly through the 'Choose File' button. Besides, an example of genotype data in the VCF format is provided and can be accessed by clicking the 'Example' button above the input box. After uploading of the candidate genotype data, users should select one of the two tools, enter the chromosome region and click the 'Submit' button to finish the query (Figure 2B).

### Searching and browsing of SNPs and genomic blocks in Plant-ImputeDB

The 'Reference Panel' module provides an advanced search box for different species, and users can search and browse SNPs based on the genomic region or gene ID. SNPs can be browsed by inputting the specific chromosomal region (e.g. Chr1:73–266) and MAF (e.g. >0.01). In addition, users can also input the gene ID (e.g. AT1G04500) and choose different lengths of upstream and downstream regions (e.g. 3K) to search for SNPs. 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. For example, when users select '*A. thaliana*' and enter 'Chr1:73–266' in the 'Region' box, the

query results will be returned as shown in Figure 2C. The returned tables can be sorted by clicking a specific column header. In addition, the query results can be exported as a tab-separated file and saved by clicking the 'Download' button.

Similarly, Plant-ImputeDB also supports the searching and browsing of genomic blocks based on genomic region or gene ID. The query results are displayed in a table with the basic genomic block information, including the chromosome, upstream region, downstream region, block ID and the length of block region. For example, when users select '*A. thaliana*' and enter 'AT1G04500' in the 'Gene ID' box, the query results will be returned as shown in Figure 2D. The returned tables can be sorted by clicking a specific column header. In addition, the query results can also be exported as a tab-separated file and saved by clicking the 'Download' button.

### Free download of reference panels in Plant-ImputeDB

The reference panels for 12 species are publicly available on the 'Download' page of Plant-ImputeDB (Figure 2E). Users can enter the genomic region of interest in the 'Region' box to obtain the corresponding VCF file. In addition, users can also download the reference panels of different chromosomes and carry out genotype imputation on the local server for GWAS or meta-GWAS analysis. These 12 reference panels support both VCF and M3VCF file formats (text and binary). Thus, users can download a reference panel in either VCF format or M3VCF format according to their own tool requirements. The database provides a total of ~538 G data for users to download.

### SUMMARY AND FUTURE DIRECTIONS

Recent decades have witnessed rapid progress in plant genetic research. Some plant-related databases including PMDBase (62) and PlantTFDB (63) have been widely used in plant research. However, they are mostly related to plant transcription factors and microsatellite DNA. Reference panels play an important role in genotype imputation for plant genetic research and breeding programs. In animal studies, Animal-ImputeDB (29) is a database that integrates high-quality reference panels from 13 species, while there is no high-quality reference panel database for plant genotype imputation. Therefore, we developed the Plant-ImputeDB database by collecting publicly available data, constructing reference panels of 12 selected species and offering an easy-to-use online genotype imputation tool with the option of two popular tools. Different from the existing related databases, Plant-ImputeDB is characterized by the comprehensive integration of genotype data for a wide range of species and supports two ways of search for SNPs and genomic blocks. It accepts submissions of plant genotype data, and provides free open access to all publicly available data to support the related research all over the world. Moreover, it is equipped with friendly web interfaces for data browse, search, imputation and download. Taken together, Plant-ImputeDB may achieve the archiving of plant genotype data at a global scale, and help the full capture of population genetic diversity and a better understanding of

the complex mechanisms associated with different phenotypes.

It can be expected that the advancement of the next-generation sequencing technology and imputation algorithms will greatly facilitate the wide applications of genotype imputation. With the continually collecting available data in the field of plant population studies, we will update the database annually by incorporating more reference panels of new species (e.g. tomato, sorghum, foxtail millet, etc.) and increasing the number of representative accessions for existing species. Overall, we will maintain Plant-ImputeDB as an informative and valuable resource for plant genetic research.

### SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

### FUNDING

Fundamental Research Funds for the Central University (Huazhong Agricultural University) [2662018PY068 to Q.Y.Y.]; National Key Research and Development Plan, China [2017YFE0104800 to Q.Y.Y.]; Huazhong Agricultural University Scientific & Technological Self - innovation Foundation [11041810351 to J.G.]. Funding for open access charge: Huazhong Agricultural University Scientific & Technological Self - innovation Foundation [11041810351]. *Conflict of interest statement.* None declared.

### REFERENCES

- Genomes Project, C., 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. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
- Huang, X., Zhao, Y., Wei, X., Li, C., Wang, A., Zhao, Q., Li, W., Guo, Y., Deng, L., Zhu, C. *et al.* (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32–39.
- Buckler, E.S., Holland, J.B., Bradbury, P.J., Acharya, C.B., Brown, P.J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J.C. *et al.* (2009) The genetic architecture of maize flowering time. *Science*, **325**, 714–718.
- Zhao, K., Tung, C.W., Eizenga, G.C., Wright, M.H., Ali, M.L., Price, A.H., Norton, G.J., Islam, M.R., Reynolds, A., Mezey, J. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.*, **2**, 467.
- Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., Yang, R., Zhang, M., Liu, S., Lu, S. *et al.* (2018) Parallel selection on a dormancy gene during domestication of crops from multiple families. *Nat. Genet.*, **50**, 1435–1441.
- Wang, B., Lin, Z., Li, X., Zhao, Y., Zhao, B., Wu, G., Ma, X., Wang, H., Xie, Y., Li, Q. *et al.* (2020) Genome-wide selection and genetic improvement during modern maize breeding. *Nat. Genet.*, **52**, 565–571.
- 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.
- Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J.L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H.F. *et al.* (2015) Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.*, **6**, 8111.
- Das, S., Abecasis, G.R. and Browning, B.L. (2018) Genotype imputation from large reference panels. *Annu. Rev. Genomics Hum. Genet.*, **19**, 73–96.

10. Wu, Y., Eskin, E. and Sankaraman, S. (2020) A unifying framework for imputing summary statistics in genome-wide association studies. *J. Comput. Biol.*, **27**, 418–428.
11. Wang, Y., Lin, G., Li, C. and Stothard, P. (2017) Genotype imputation methods and their effects on genomic predictions in cattle. *Springer Sci. Rev.*, **4**, 79–98.
12. 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.
13. Das, S., Forer, L., Schonherr, 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.
14. Browning, B.L. and Browning, S.R. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.
15. Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R., Altshuler, D. and Daly, M.J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.
16. Howie, B.N., Donnelly, P. and Marchini, J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
17. Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., Li, C., Zhu, C., Lu, T., Zhang, Z. *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.*, **42**, 961–967.
18. Golicz, A.A., Bayer, P. E. and Edwards, D. (2015) Plant genotyping: methods and protocols. In: Bately, J. (ed). *Methods in Molecular Biology*, Springer, NY, Vol. **1245**, pp. 257–270.
19. Clark, A.G. and Li, J. (2007) Conjoining SNPs to detect associations. *Nat. Genet.*, **39**, 815–816.
20. Wang, D.R., Agosto-Perez, F.J., Chebotarov, D., Shi, Y., Marchini, J., Fitzgerald, M., McNally, K.L., Alexandrov, N. and McCouch, S.R. (2018) An imputation platform to enhance integration of rice genetic resources. *Nat. Commun.*, **9**, 3519.
21. Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, **11**, 499–511.
22. Wang, Z. and Chatterjee, N. (2017) Increasing mapping precision of genome-wide association studies: to genotype and impute, sequence, or both? *Genome Biol.*, **18**, 118.
23. Chang, D., Nalls, M.A., Hallgrimsdottir, I.B., Hunkapiller, J., van der Brug, M., Cai, F. and International Parkinson's Disease Genomics, C. International Parkinson's Disease Genomics, C., and Me Research, T., Kerchner, G.A., Ayalon, G. *et al.* (2017) A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.*, **49**, 1511–1516.
24. Spencer, C.C., Su, Z., Donnelly, P. and Marchini, J. (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.*, **5**, e1000477.
25. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
26. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N. *et al.* (2018) Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.*, **50**, 1505–1513.
27. International HapMap, C., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.
28. McCarthy, S., Das, S., Kretschmar, 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.
29. 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.
30. Tam, V., Patel, N., Turcotte, M., Bosse, Y., Pare, G. and Meyre, D. (2019) Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.*, **20**, 467–484.
31. Koch, L. (2016) 1001 genomes and epigenomes. *Nat. Rev. Genet.*, **17**, 503–503.
32. Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., Fuentes, R.R., Ulat, V.J., Chebotarov, D., Zhang, G., Li, Z. *et al.* (2014) SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.*, **43**, D1023–D1027.
33. Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C. *et al.* (2017) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, 1–12.
34. Grant, D., Nelson, R.T., Cannon, S.B. and Shoemaker, R.C. (2009) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **38**, D843–D846.
35. He, F., Pasam, R., Shi, F., Kant, S., Keeble-Gagnere, G., Kay, P., Forrester, K., Fritz, A., Hucl, P., Wiebe, K. *et al.* (2019) Exome sequencing highlights the role of wild-relative introgression in shaping the adaptive landscape of the wheat genome. *Nat. Genet.*, **51**, 896–904.
36. Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S., Zhao, K., Blanca, J., Zhang, Z., Huang, S. *et al.* (2018) Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Res.*, **47**, D1128–D1136.
37. Wu, J., Wang, L., Fu, J., Chen, J., Wei, S., Zhang, S., Zhang, J., Tang, Y., Chen, M., Zhu, J. *et al.* (2020) Resequencing of 683 common bean genotypes identifies yield component trait associations across a north-south cline. *Nat. Genet.*, **52**, 118–125.
38. Ma, Z., He, S., Wang, X., Sun, J., Zhang, Y., Zhang, G., Wu, L., Li, Z., Liu, Z., Sun, G. *et al.* (2018) Resequencing a core collection of upland cotton identifies genomic variation and loci influencing fiber quality and yield. *Nat. Genet.*, **50**, 803–813.
39. Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., Zhou, G., Lohwasser, U., Hua, S., Wang, H. *et al.* (2019) Whole-Genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Mol. Plant.*, **12**, 30–43.
40. Guo, S., Zhao, S., Sun, H., Wang, X., Wu, S., Lin, T., Ren, Y., Gao, L., Deng, Y., Zhang, J. *et al.* (2019) Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.*, **51**, 1616–1623.
41. Wang, X., Bao, K., Reddy, U.K., Bai, Y., Hammar, S.A., Jiao, C., Wehner, T.C., Ramirez-Madera, A.O., Weng, Y., Grumet, R. *et al.* (2018) The USDA cucumber (*Cucumis sativus* L.) collection: genetic diversity, population structure, genome-wide association studies, and core collection development. *Hortic. Res.*, **5**, 64.
42. 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 DNaseSeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.*, **10**, 736.
43. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
44. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup. (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
45. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M. and Lee, J.J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, **4**, 7.
46. Browning, B.L. and Browning, S.R. (2016) Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.*, **98**, 116–126.
47. Brandariz, S.P., Gonzalez Reymundez, A., Lado, B., Malosetti, M., Garcia, A.A., Quincke, M., von Zitzewitz, J., Castro, M., Matus, I., Del Pozo, A. *et al.* (2016) Ascertainment bias from imputation methods evaluation in wheat. *BMC Genomics*, **17**, 773.
48. Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhuri, S., Xiao, M., Peterson, A.S., Kwok, P.Y., Seshagiri, S. and Wall, J.D. (2019) Evaluating the quality of the 1000 genomes project data. *BMC Genomics*, **20**, 620.
49. Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E., Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R. *et al.* (2006) A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.*, **78**, 437–450.
50. Osabe, D., Tanahashi, T., Nomura, K., Shinohara, S., Nakamura, N., Yoshikawa, T., Shiota, H., Keshavarz, P., Yamaguchi, Y., Kunika, K. *et al.* (2007) Evaluation of sample size effect on the identification of haplotype blocks. *BMC Bioinformatics*, **8**, 200.



51. Rasheed,A., Hao,Y., Xia,X., Khan,A., Xu,Y., Varshney,R.K. and He,Z. (2017) Crop Breeding chips and genotyping platforms: progress, challenges, and perspectives. *Mol. Plant.*, **10**, 1047–1064.
52. Friedrich,J., Antolin,R., Edwards,S.M., Sanchez-Molano,E., Haskell,M.J., Hickey,J.M. and Wiener,P. (2018) Accuracy of genotype imputation in Labrador Retrievers. *Anim. Genet.*, **49**, 303–311.
53. Yang,M., Lu,K., Zhao,F.J., Xie,W., Ramakrishna,P., Wang,G., Du,Q., Liang,L., Sun,C., Zhao,H. *et al.* (2018) Genome-wide association studies reveal the genetic basis of ionomic variation in rice. *Plant Cell*, **30**, 2720–2740.
54. Arouisse,B., Korte,A., van Eeuwijk,F. and Kruijjer,W. (2020) Imputation of 3 million SNPs in the Arabidopsis regional mapping population. *Plant J.*, **102**, 872–882.
55. Flint-Garcia,S.A., Thuillet,A.C., Yu,J., Pressoir,G., Romero,S.M., Mitchell,S.E., Doebley,J., Kresovich,S., Goodman,M.M. and Buckler,E.S. (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J.*, **44**, 1054–1064.
56. Song,J.-M., Guan,Z., Hu,J., Guo,C., Yang,Z., Wang,S., Liu,D., Wang,B., Lu,S., Zhou,R. *et al.* (2020) Eight high-quality genomes reveal pan-genome architecture and ecotype differentiation of *Brassica napus*. *Nat. Plants*, **6**, 34–45.
57. Wang,M., Tu,L., Lin,M., Lin,Z., Wang,P., Yang,Q., Ye,Z., Shen,C., Li,J., Zhang,L. *et al.* (2017) Asymmetric subgenome selection and cis-regulatory divergence during cotton domestication. *Nat. Genet.*, **49**, 579–587.
58. Lu,S., Dong,L., Fang,C., Liu,S., Kong,L., Cheng,Q., Chen,L., Su,T., Nan,H., Zhang,D. *et al.* (2020) Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. *Nat. Genet.*, **52**, 428–436.
59. Qi,J.J., Liu,X., Shen,D., Miao,H., Xie,B.Y., Li,X.X., Zeng,P., Wang,S.H., Shang,Y., Gu,X.F. *et al.* (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.*, **45**, 1510–1515.
60. Zhao,G., Lian,Q., Zhang,Z., Fu,Q., He,Y., Ma,S., Ruggieri,V., Monforte,A.J., Wang,P., Julca,I. *et al.* (2019) A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat. Genet.*, **51**, 1607–1615.
61. Pont,C., Leroy,T., Seidel,M., Tondelli,A., Duchemin,W., Armisen,D., Lang,D., Bustos-Korts,D., Goue,N., Balfourier,F. *et al.* (2019) Tracing the ancestry of modern bread wheats. *Nat. Genet.*, **51**, 905–911.
62. Yu,J., Dossa,K., Wang,L., Zhang,Y., Wei,X., Liao,B. and Zhang,X. (2017) PMDBase: a database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Res.*, **45**, D1046–D1053.
63. Jin,J., Tian,F., Yang,D.-C., Meng,Y.-Q., Kong,L., Luo,J. and Gao,G. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*, **45**, D1040–D1045.