BMC Plant Biology

RESEARCH Open Access

CoreSNP: an efficient pipeline for core marker profile selection from genome-wide SNP datasets in crops

Tingyu Dou[1†], Chunchao Wang[1†], Yanling Ma[1], Zhaoyan Chen[1], Jing Zhang[1] and Ganggang Guo[1*]

Abstract

Background DNA marker profiles play a crucial role in the identification and registration of germplasm, as well as in the distinctness, uniformity, and stability (DUS) testing of new plant variety protection. However, selecting minimal marker sets from large-scale SNP dataset can be challenging to distinguish a maximum number of samples. Results: Here, we developed the CoreSNP pipeline using a "divide and conquer" strategy and a "greedy" algorithm. The pipeline offers adjustable parameters to guarantee the distinction of each sample pair with at least two markers. Additionally, it allows datasets with missing loci as input. The pipeline was tested in barley, soybean, wheat, rice and maize. A few dozen of core SNPs were efficiently selected in different crops with SNP array, GBS, and WGS dataset, which can differentiate thousands of individual samples. The core SNPs were distributed across all chromosomes, exhibiting lower pairwise linkage disequilibrium (LD) and higher polymorphism information content (PIC) and minor allele frequencies (MAF). It was shown that both the genetic diversity of the population and the characteristics of the original dataset can significantly influence the number of core markers. In addition, the core SNPs capture a certain level of the original population structure.

Conclusions CoreSNP is an efficiency way of core marker sets selection based on Genome-wide SNP datasets of crops. Combined with low-density SNP chip or genotyping technologies, it can be a cost-effective way to simplify and expedite the evaluation of genetic resources and differentiate different crop varieties. This tool is expected to have great application prospects in the rapid comparison of germplasm and intellectual property protection of new varieties.

Keywords Shannon index, Germplasm discrimination and management, Low-dense genotyping

†Tingyu Dou and Chunchao Wang contributed equally

*Correspondence:
Ganggang Guo
guoganggang@caas.cn
[1]Key Laboratory of Grain Crop Genetic Resources Evaluation and Utilization (MARA), The National Key Facility for Crop Gene Resources and Genetic Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (ICS-CAAS), Beijing 100081, China

Background

Accurate identification and registration of germplasm resources are essential for plant conservation efforts [1–3]. Traditional identification methods based on morphological or agronomical traits can be time-consuming due to the influence of environmental factors on most phenotypes [4]. Genetic molecular markers have been recommended by the International Union for the Protection of New Varieties of Plants (UPOV) as a more reliable and efficient approach for variety and cultivar identification, in addition to morphological characteristics [5–7].

With the increasing number of germplasm collections stored in genebanks, effective characterization and management have become major challenges [8]. In the USDA Soybean Germplasm Collection, it was discovered that over 30% of wild accessions and 23% of cultivated accessions were redundant, as their similarity exceeded 99.9% [9]. Similarly, an analysis of genetic profiles in the German ex situ gene bank revealed that approximately 33% of the 22,626 collections of barley (*Hordeum vulgare* L.) were potential duplicates [10]. By selecting a comprehensive core set of molecular markers, germplasm collections can be rapidly assessed and characterized. Assigning a unique "molecular passport" profile to each accession enables researchers to easily track and manage the collections, reducing duplication and facilitating targeted utilization [11].

Furthermore, core marker sets play a vital role in protecting new plant varieties. Due to the increasing number of crop varieties being released onto the market, the limited genetic diversity and close similarity among elite parental lines lead to fewer morphological differences that can be utilized for variety identification in modern breeding programs [12]. The preselected small set of markers can be cost-effectively utilized for DUS testing, identification of essential derived varieties (EDVs) and the verification of seed authenticity and purity. This benefits breeders by protecting their intellectual property rights [13–15].

Among various types of molecular markers, single nucleotide polymorphisms (SNPs) have gained significant importance due to their high reproducibility, locus specificity, and wide distribution throughout the genome [16]. The continuous advancements in high-throughput sequencing technology have facilitated the generation of a large number of SNPs. Through various genotyping platforms, researchers can obtain millions of SNPs that cover the entire genome, resulting in SNP sets with diverse characteristics [17–19]. The reduction of SNP density and the development of low-density SNP genotyping panels have gained prominence due to its cost-effectiveness, enabling the identification of large-scale germplasm and the assessment of their relatedness. A variety of genotyping methods and technologies, such as Taqman genotyping assays [20], Kompetitive allele-specific PCR (KASP) [21, 22], Amplification refractory mutation system PCR (ARMS-PCR) [23], as well as low density SNP chips [24] have been extensively used in genotyping specific SNPs of interest. Selecting the fewest and most representative SNPs from vast amounts of information that contain redundant data has become a concern for researchers.Currently, various methods for selecting core SNPs selection have been adopted in various crop species for variety identification and DNA fingerprinting [25–28]. In soybean, Liu et al. divided 4044 SNPs into 24 panels with varying numbers of SNPs based on polymorphic information content (PIC) values. A core set panel of 20 SNPs was selected to construct molecular IDs for 138 released soybean cultivars, resulting in the fewest number of indistinguishable pairs of accessions [29]. Using a combination of polymorphisms and principal component analysis, Li et al. selected 60 core SNPs distributed across all chromosomes to provide sufficient genetic information for 166 representative inbred lines of Chinese cabbage from a pool of 1167 SNPs [30]. However, these studies did not consider the discriminatory power of combinations of SNPs, and the manual screening process used in these studies may not be sufficient to meet the increasing demand for selecting core SNPs from large-scale sequenced data.

Automated methods based on computer programming have also been proposed to improve the efficiency of selecting core SNPs. These methods aim to construct relatively small marker sets capable of distinguishing a wide range of varieties. Hiroshi et al. (2013) employed an exhaustive method to develop MinimalMarker software for identifying minimal marker sets. They constructed a pairwise comparison matrix by calculating the number of differential alleles between each pair of varieties and consistently selected the marker with the highest discrimination to form the minimal set [31]. This algorithm was subsequently used for core SNP selection in the identification of pepper [32] and cucumber varieties [33]. In their study on rice, Yuan et al. introduced a method called conditional random selection (CRS) to specifically distinguish between EDV and non-EDV varieties [34]. The method follows a "divide and conquer" strategy, where specific haplotypes are initially constructed using randomly selected SNPs. Redundant SNPs were then eliminated by systematically shielding one SNP at a time while checking whether the remaining SNPs could still distinguish all varieties. Through this approach, the researchers selected a set of 390 SNPs that could distinguish between 3,024 rice varieties.

Despite the progress made in selecting core SNPs, these studies are relatively time-consuming and not user-friendly, especially when dealing with data that contain missing loci. Here, we propose CoreSNP, a novel

and customized pipeline for developing minimal core SNPs. This pipeline can largely reduce the number of SNPs essential for identifying different varieties from large-scale genotyping data. Based on publicly available genotype datasets, the pipeline has been proven to be robust in different crops and across various sequencing platforms.

Materials and methods

Description of the coreSNP pipeline

A greedy search algorithm was applied to the core SNP selection and the specific workflow of the pipeline is summarized in Fig. 1. To provide a clear illustration of the selection process, we took the dataset with 8 individuals and 8 markers as an example with a schematic diagram (Fig. 2). In the initial step, missing SNPs were imputed by major homozygous SNP genotype. The Shannon index was calculated for each SNP was subsequently calculated for each SNP in the dataset, and the SNP with the highest value was randomly selected as the first solution. The pipeline then proceeded by iteratively analyzing the remaining SNPs in the dataset and constructing the haplotypes using the selected SNPs. Next, the SNP that maximized the Shannon entropy index was selected as

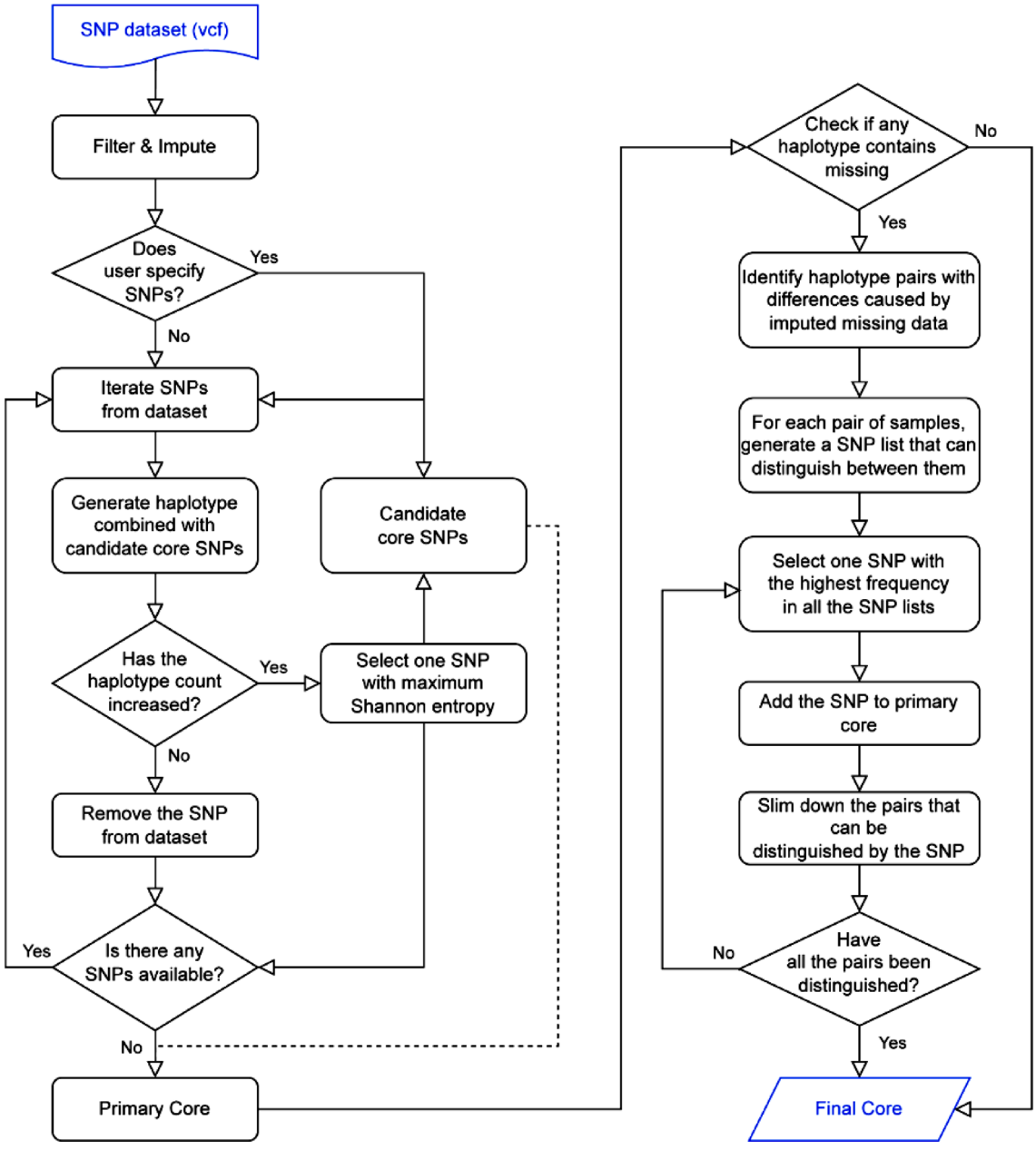

Fig. 1 The basic workflow of the CoreSNP pipeline

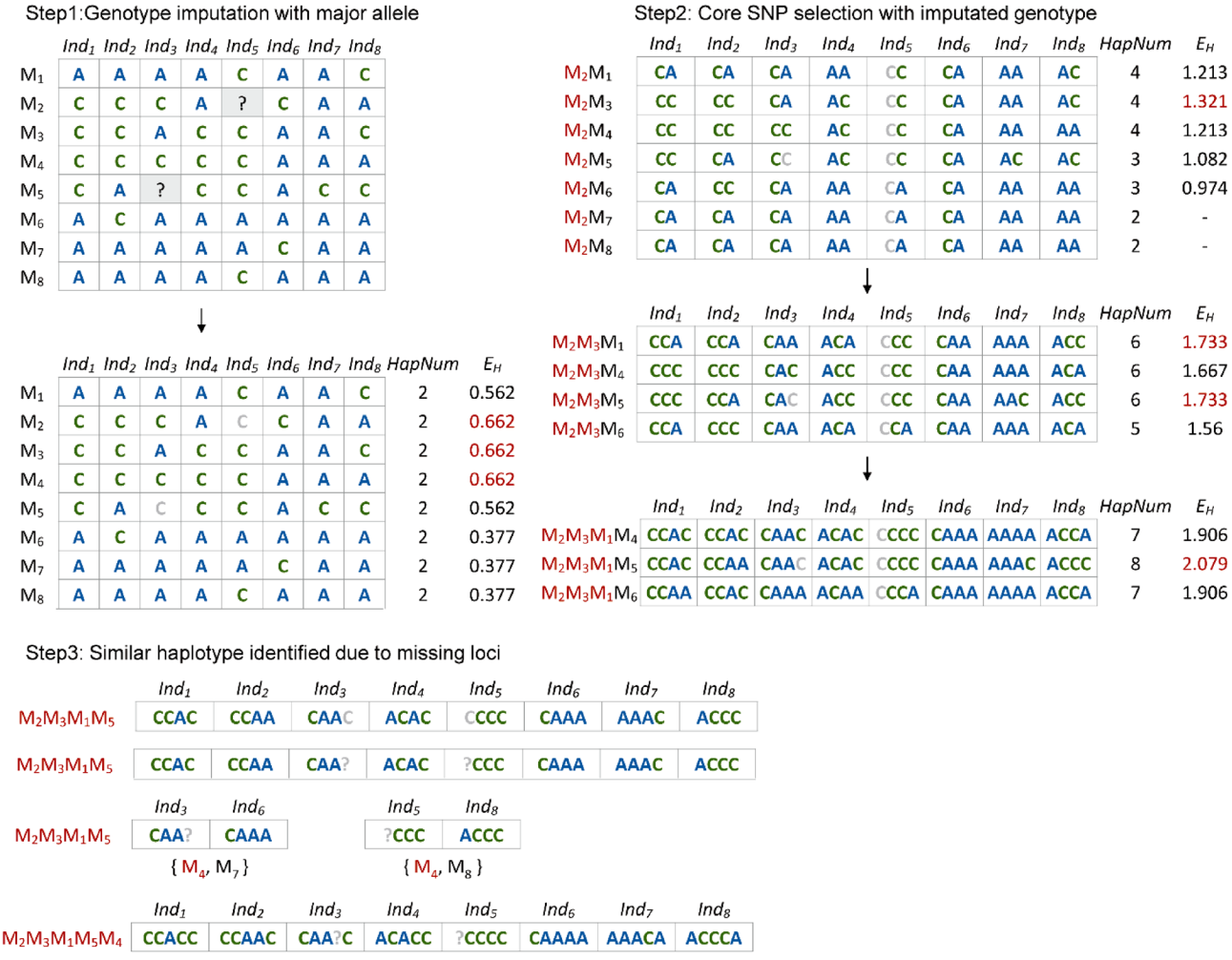

Fig. 2 Schematic diagram of the CoreSNP pipeline

the second solution and accordingly removed from the dataset correspondingly. The iterative process continues until no additional SNPs can be added. During this procedure, a crucial step involvesassessing whether the newly selected SNP contributes to an increase in the number of observed haplotypes before calculating the Shannon index. If not, the SNP is skipped, thereby expediting the pipeline. As depicted in Fig. 2, SNPs such as '*Marker7*' and '*Marker*8' were used as examples. In this step, we exclude SNPs without polymorphisms and nearby SNPs linked to the selected core SNPs. Consequently, at each selection step, the process achieved the local optimal solution to maximize sample differentiation and haplotype diversity.

After employing major allele imputation, the haplotypes without any missing values were outputted, and the chosen SNPs were designated as the primary core markers. Recognizing that missing sites could potentially result from sequencing errors, we retrieved haplotypes with missing loci based on the positions of imputed values. Subsequently, the pipeline identified pairs of haplotypes with differences caused by imputed missing data. Through an iterative process, SNPs with the highest frequency were selected and added to the primary list. Eventually, the final core sets were generated after achieving complete differentiation across all sample pairs (Fig. 2).

In this study, additional command options were introduced to configure the running parameters, thereby enhancing the applicability of the pipeline. Users can define input files using the "-v" option and customize which markers should be included or excluded using the "-i" and "-e" options, respectively. Since some SNP combinations (haplotypes) have close Shannon index values, the "-x" option was provided to allow users to specify the number of candidate SNP combinations to be selected in each round. This option provides users with more feasible solutions for downstream analysis. Furthermore, core SNPs were randomly and independently selected each round, and users have the flexibility to define the number of repetitions using the "-c" option. The "-m" option

allows users to specify that at least two markers are required to effectively discriminate between every pair of varieties. This criterion can indeed mitigate potential errors or noise introduced by the raw sequencing data, thereby enhancing the reliability and robustness of the results (Table S1).

Genotype dataset collection

To validate the CoreSNP pipeline, we collected the public SNP genotype dataset of barley, soybean, wheat, rice, and maize from widely studied plant databases or research studies (Table S2). The SNP data were obtained from different genotyping and sequencing platforms, including SNP microarray, high-throughput genotyping by sequencing (GBS), and whole-genome sequencing (WGS).

For the barley analysis, genotypic information of 1,000 accessions was downloaded from the Germinate Barley SNP Platforms [35]. The collection of 1,000 genotypes is a representative subset of the global 22,626 barley accessions from the German Federal ex situ GenBank. By integrating our unpublished data, we curated a dataset consisting of 1081 barley accessions with 42,520 SNPs genotyped by the 50 K iSelect SNP Array, referred to as Barley**I**. Additionally, we obtained a GBS SNP matrix of 1,297 barley accessions from the IPK data repository, referred to as Barley**II** [10]. We merged two different datasets by using a shared sample set of 1081 barley accessions. The resulting combined dataset, designated Barley **I** & **II**, contains a total of 185,508 SNPs.

The soybean dataset comprises 817 soybean accessions with 158,959 SNPs, including 77 parental lines, 169 non-parental lines, and 571 progenies. Genotyping was performed using the ZDX1 array genotyping platform [36]. The wheat genotypic dataset consists of 178,803 SNPs and includes genotypic information for 271 Chinese wheat landraces that were genotyped using the 660 K wheat SNP array [37].

We extracted genotypic information from 453 high-coverage rice accessions obtained from the 3,000 Rice Genome Project [38]. The maize genotypic dataset includes genotype information from 1,210 maize lines based on whole-genome sequencing data, comprising approximately 83 million raw SNPs. The data were downloaded from the maize HapMapV3 study [39].

Data preprocessing

Heterozygous sites were treated as missing data and processed using BCFtools version 1.10.2 [40]. Genotype imputation of missing sites was performed using FILLIN with default parameters [41]. Sample and SNP filtration as well as LD-based SNP pruning were performed using PLINK version 1.90 [42].

Polymorphism analysis

In this study, the Shannon indices [43] and PIC values [44] were calculated using the following formulas.

$$H = -\sum_{i=1}^{n} P_i ln\left(P_i\right) \tag{1}$$

where H is the diversity index, n is the population size and P_i is the frequency of the combined haplotypes.

$$PIC = 1 - \sum_{i=1}^{n} P_i^2 \tag{2}$$

$$PIC = 1 - \sum_{i=1}^{n} P_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2P_i^2 P_j^2 \tag{3}$$

where n is the population size and P_i and P_j are the frequencies of two SNP alleles among all samples. Equation 2 suggests the simplified formula, while Eq. 3 suggests the full formula for PIC calculation.

Genetic diversity analysis

Allele counts and frequencies and IBS information were calculated by PLINK version 1.90 [42]. LD estimation was measured as parameter r^2 with a maximum distance of 10 Mb using PopLDdecay version 3.40 [45]. A Mantel test for correlation between distance matrices was performed using an in-house python script [46]. Principal component analysis (PCA) of the barley samples was performed using EIGENSOFT/smartPCA software version 6.1.4 [47]. All figures were generated using R software version 4.1.2 using the packages ggplot [48] and RIdeogram [49].

Data availability

The CoreSNP pipeline was developed using the Python programming language and accepts compressed Variant Call Format (VCF) files or uncompressed VCF files as input. The cost-free program is readily executed from the command line, relying on specific dependencies, such as Python (version 3.6 or higher), Numpy and PLINK 1.9. Both the source code and the essential dependencies are accessible on GitHub (https://github.com/admy55/CoreSNP) or Gitee (https://gitee.com/admy55/CoreSNP).

Results

Performance test of the CoreSNP pipeline

Initially, imputation and filtration were performed on the raw dataset. Samples with a genotype missing rate of ≥0.5 and SNPs with a missing rate of ≥0.2 were removed. The final datasets used for testing the pipeline are shown in Table S2. Using default parameters, the pipeline successfully differentiated 1081 barley samples using 21 core

SNPs, 1297 barley samples using 32 core SNPs, a merged dataset of 1081 barley samples using 19 SNPs, 800 soybean varieties using 52 SNPs, 271 wheat varieties using 24 SNPs, 453 rice accessions using 18 SNP markers, and 1206 maize varieties using 60 SNPs (Fig. 3, Table S2). The data represent the minimum set of markers obtained after running the pipeline ten times (Other results are not shown).

We compared CoreSNP with a Random Selection (RS) method using the Barley I & II dataset to evaluate their performance in selecting core SNPs. In the RS process, two markers were randomly selected from the long and short arms of each chromosome, forming a combination of 28 SNPs with aMAF greater than 0.3 (RS1) and MAF greater than 0.4 (RS2). After 20 iterations of selection, the randomly selected 28 SNPs from RS1 and RS2 identified accessions ranging from 977 to 1061. Through saturation curve analysis, we found that the CoreSNP approach exhibited significantly higher efficiency in SNP selection than the RS method, as it efficiently identified the same number of germplasms with a smaller set of selected SNPs (Fig. S1).

Characteristics of the core SNPs selected from different datasets

To assess the utility of the core SNPs, we calculated the genetic diversity parameters, including MAF and PIC, as indicators of the markers' discriminatory ability. Values closer to 0.5 for MAF and 0.375 (depending on the formula) for PIC of biallelic markers indicate better discriminatory properties.

The results showed that more than 70% of the core SNPs had an MAF greater than 0.3 and PIC greater than 0.35, except for the soybean dataset (Fig. S2). In the case of the soybean and maize datasets, there are 6 and 8 markers, respectively, with a minor allele frequency (MAF) below 0.1. These markers were specifically chosen to distinguish individuals with high genetic similarity. Notably, all markers selected from the Barley I & II dataset exhibited MAF values greater than 0.3 and PIC values greater than 0.35, except for two markers located on chromosome 1 H and chromosome 5 H, which displayed MAF values of 0.23 and 0.19, respectively.

The 24 core markers were cover 13 chromosomes in the wheat genome, and in addition to these, the selected markers are distributed across nearly every chromosome in the genome. The majority of the core SNPs were not located in close proximity to each other, suggesting a lower LD distance between them (Supplementary Data 2).

Genetic diversity analysis of the test datasets

Significant variations were noted in the count of the ultimate core sets during the evaluation of the CoreSNP pipeline across different datasets. To investigate the drivers underlying these disparities, we assessed several factors, encompassing MAF, missing rate, linkage

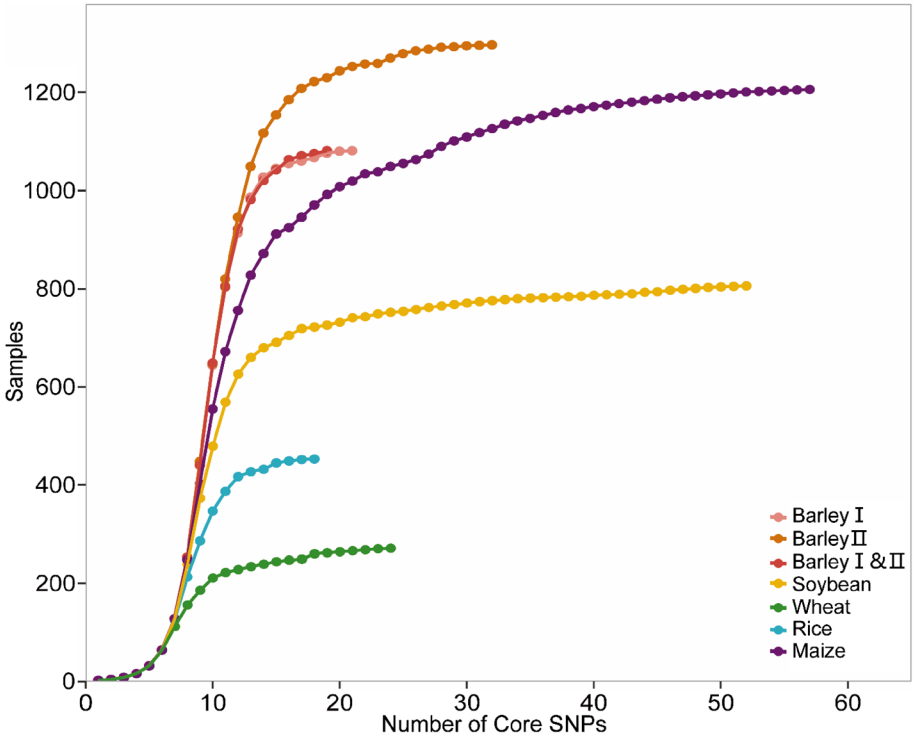

Fig. 3 Discrimination saturation curve of core SNPs selected from raw dataset

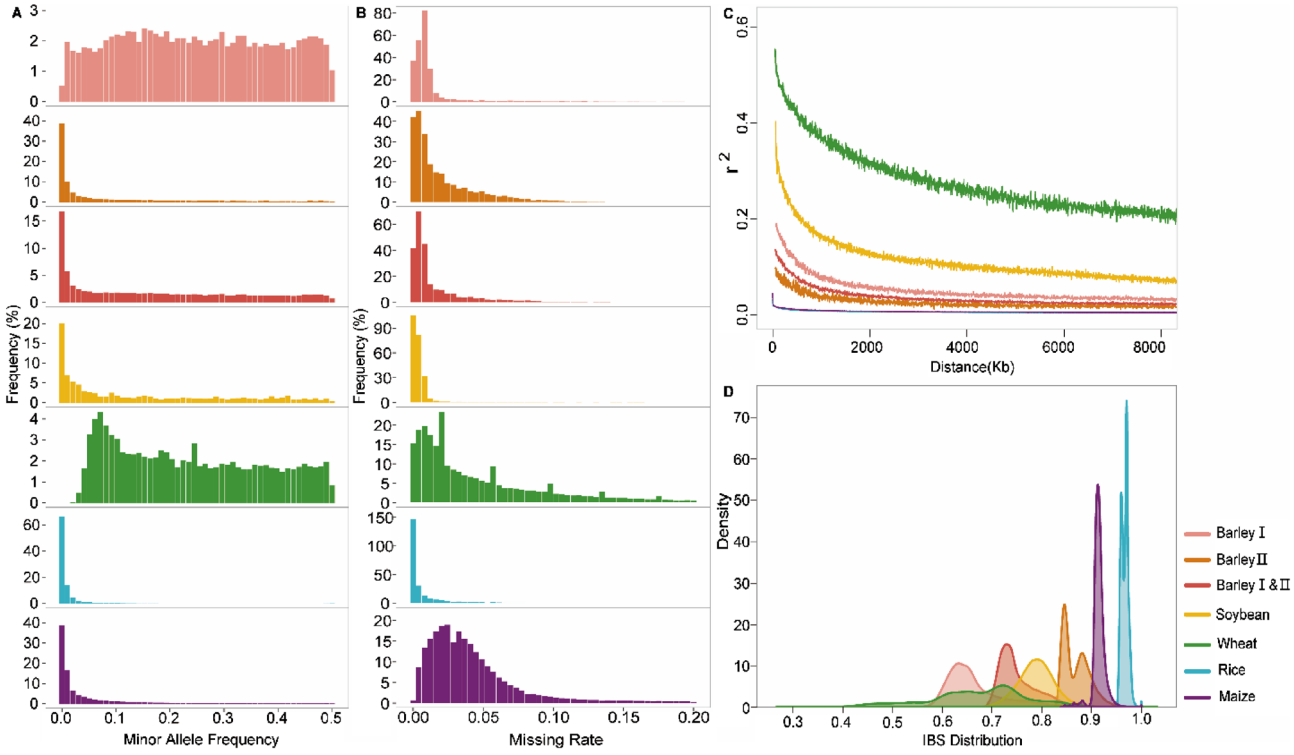

Fig. 4 Genetic diversity of the raw test datasets. **A**, Frequency distribution of MAF values among raw dataset. **B**, Frequency distribution of PIC values among raw dataset. **C**, Linkage disequilibrium (LD) decay patterns of different species. The decays of LD (r2) with physical distance (kilobases) for SNPs in five crops are shown. **D**, Identity by states (IBS) distribution across five crops

disequilibrium decay distance (LDD), and identity by state (IBS) information utilizing the test dataset (Fig. 4).

No specific regularity was observed in the distribution of MAF or missing rate in soybean experiments (Fig. 4A-B). However, the soybean dataset exhibited a slower LD decay and consisted of 817 individuals, including both parental populations and their descendants (Fig. 4C). Finally, 52 markers capable of differentiating 800 samples were selected. In contrast, the maize dataset with a faster LD decay exhibited a higher missing rate and rare genetic variations. Approximately 28% of the SNPs exhibited missing rates between 0.05 and 0.2, and 20% of the SNPs had an MAF below 0.01. Consequently, during the process of selecting core SNPs in maize, we initially selected 26 markers to form the primary core, which was later expanded to include 60 SNPs after accounting for missing loci.

The wheat and barley I dataset showed a relatively uniform pattern in MAF distribution (Fig. 4A). However, the wheat dataset exhibited a higher proportion of markers with missing rates ranging from 0.05 to 0.2, slower LD decay and higher IBS peaks compared to the barley dataset (Fig. 4B-D). Consequently, the efficiency of core SNP selection was found to be lower in the wheat dataset. Under similar conditions, we were able to distinguish 1081 barley samples using 21 core markers, whereas only 271 wheat samples could be differentiated using 24 core markers (see Table S2). Overall, both the genetic diversity of the population and the specific characteristics of the sequencing data significantly influence the number of core markers.

Flexible application of CoreSNP for barley accessions

The fertility of spikelet florets, commonly referred to as row type, and the presence or absence of grain hulls are two crucial infraspecific morphological traits in barley. In this study, we assessed the feasibility of the CoreSNP pipeline with three SNP markers (rs7_527405910 on chromosome 7 H, JHI-Hv50k-2016-107445 on chromosome 2 H, and JHI-Hv50k-2016-230985 on chromosome 4 H) that are tightly linked to these two traits as included markers for core SNP selection. By establishing a criterion of a minimum of at least two markers distinguishing each pair of samples, we successfully identified 31 core SNP markers that differentiated 1081 barley samples using the merged dataset. These core SNP markers were distributed across all seven chromosomes (Fig. S3).

The genetic structure of 1081 barley collections was analyzed using both the core SNP set and the complete set of genome-wide SNPs. Principal component analysis

(PCA) revealed that some samples with relatively distant genetic distances within three clusters were still well separated along the PC1 and PC2 dimensions. This observation highlights that the core SNPs capture a certain level of the original population structure (Fig. S4).

The correlation coefficient (r) values of 0.4024 indicate a significant correlation between the genetic distance matrices of the core SNP panels and the total genome-wide SNPs. This affirms the robustness and reliability of our core SNP selection process (Table S10).

Discussion

Development of the CoreSNP pipeline

Appraising the discriminatory power of loci combinations is a necessary step in marker screening [50]. When assessing the performance of multiple loci combinations, MAF values have shown lower sensitivity compared to PIC values and the Shannon Diversity index, as demonstrated through statistical analysis. In the case of the frequency distribution presented in Table S11, it was observed that the Shannon index displayed a similar trend to that of the complete PIC formula. However, the computational complexity associated with calculating the full PIC formula is relatively higher than that associated with calculating the Shannon index, especially as the diversity of loci combinations expands.

In addition, we conducted a comparative analysis with the Conditional Random Selection (CRS) method using the data released in their research [34]. The results demonstrate that, for the same fractal dataset, our approach can differentiate 1000 samples with 44 SNPs, while the CRS method requires 54–59 SNPs. Additionally, due to the random sampling nature of their method, the final number of labels exhibited significant fluctuations. Furthermore, it was noted that their method has limitations in handling missing data. Thus, in the present study, we introduced CoreSNP, an screening pipeline that utilizes the Shannon index to create core sets of SNPs. Overall, the pipeline initiates from the input VCF dataset and proceeds through a fast and thorough process. CoreSNP achieved discrimination of over 1000 barley samples in just five minutes of runtime on our machine (OS: Windows 11, CPU: Intel Core i7-9850 H 2.6 GHz, RAM: 32 GB), outperforming other traditional methods [31, 32, 34].

Polymorphic analysis of the core SNPs selected based on the different datasets

The discrimination curve of the markers was plotted based on the haplotype count obtained at each step (Fig. 2). This curve demonstrated that the SNP panels generated by this pipeline possess a remarkable discriminatory capacity. The selected core SNP markers distinguished 100% of the test collections, with the exception of 17 soybean samples and two maize samples. This discrepancy in these samples can be attributed to the MAF filtration applied to the raw dataset obtained from public sources. These particular samples represent closely related accessions, as evidenced by the calculated identical by state (IBS) information, where pairs of samples exhibited an IBS value of 1.

The selected core SNPs exhibited a high degree of polymorphism and were distributed across the genome. In fact, within the pipeline, the evaluation of an increase in haplotype numbers involved removing redundant linked markers, ensuring that adjacent pairs of SNPs were not positioned too closely to each other. The number of core SNPs is influenced by the genetic diversity of different populations. However, it is also significantly influenced by the missing rate in the original dataset. In terms of population structure, certain closely related samples are still unable to be distinctly separated in the reduced-dimensional data. This limitation could be attributed to the insufficient number of markers in comparison to the original dataset, which hindered their ability to adequately capture the genetic variations present.

Applications of CoreSNP pipeline in the future

The cost of genome sequencing tools remains high for routine large-scale germplasm identification. However, with the development of low-density sequencing techniques, such as KASP and ARMS-PCR, our CoreSNP pipeline provides a cost-effective solution for germplasm identification and genetic relationship analysis. By constructing a reference library using a combination of core SNP alleles (haplotypes), it becomes possible to establish a comprehensive catalog of DNA profiles and DNA fingerprints for each accession. This will greatly facilitate the analysis of variety traceability and genetic backgrounds. However, it is important to note that this reference library is not a one-size-fits-all solution, thereby necessitating additional efforts in collecting more genotypes.

Moreover, these SNPs can also be utilized for parent combination selection and progeny screening during the breeding process, thus enhancing breeding efficiency and precision. Additionally, the processing of many samples inevitably introduces the issue of sample mix-up. It becomes challenging when samples contain mixtures with few identifiable characteristics. The combination of genetic profiles and core SNP alleles can be used to identify accidental sample mix-ups, ensuring the authenticity and purity of seeds. In summary, the CoreSNP pipeline takes into account sequencing platform constraints and user-specific preferences. By potentially saving time and reducing costs, it simplifies and streamlines the process of genomic identification. This tool will serve as a valuable foundation for modern breeding efforts and future germplasm management and preservation endeavors.

Conclusion

In conclusion, we developed the CoreSNP pipeline for evolution of the discrimination power of SNP combinations. We validated it using diverse genotype files from various crops and found that it exhibited high efficiency. This tool can be efficiently used for selecting core SNPs capable of representing genome-wide variation to identify similarity and redundancy within germplasm resources.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12870-023-04609-w.

Supplementary Material 1: **Fig. S1**. Comparison of the CoreSNP pipeline and Random Selection(RS) method for SNP selection. **Fig. S2**. Frequency distribution of MAF and PIC values among the selected core SNPs. **Fig. S3**. Distribution of core SNPs selected from barley merged dataset with specific parameters. **Fig. S4**. The comparison of the Principle Component Analysis (PCA) analysis based on the raw genome-wide SNPs and core SNPs. **Table S1**. Multiple options for running the core SNP pipeline. **Table S2**. Datasets description and core SNP selection in various crop species. **Table S10**. Mantel's test for comparisons among genetic distance matrices calculated using the core SNPs and the original dataset. **Table S11**. Comparison of MAF, PIC and Shannon index among different haplotypes

Supplementary Material 2: **Table S3**. Description of Core SNPs Selected from the Barleyldataset with default parameters. **Table S4**. Description of Core SNPs Selected from the Barleyldataset with default parameters. **Table S5**. Description of Core SNPs Selected from the Barleyl&lldataset with default parameters. **Table S6**. Description of Core SNPs Selected from the Soybean dataset with default parameters. **Table S7**. Description of Core SNPs Selected from the Wheat dataset with default parameters. **Table S8**. Description of Core SNPs Selected from the Rice dataset with default parameters. **Table S9**. Description of Core SNPs Selected from the Maize dataset with default parameters

Acknowledgements
Not applicable.

Author Contributions
GGG and JZ designed and supervised the research. TYD and CCW performed the pipeline and the bioinformatics analysis. TYD drafted the manuscript. YLM and ZYC approved the revision of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding
This work was supported by grants from the National Natural Science Foundation of China (U20A2026), National Key R&D Program of China (2022YFD2301300), Agricultural Science and Technology Innovation Program of CAAS, and China Agriculture Research System (CARS-05).

Data Availability
The original contributions presented in the study are included in the article and Supplementary materials.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 14 September 2023 / Accepted: 14 November 2023
Published online: 21 November 2023

References

1. Korir NK, Han J, Shangguan LF, Wang C, Kayesh E, Zhang YY, Fang JG. Plant variety and cultivar identification: advances and prospects. Crit Rev Biotechnol. 2013;33(2):111–25.
2. He SP, Sun GF, Geng XL, Gong WF, Dai PH, Jia YH, Shi WJ, Pan ZE, Wang JD, Wang LY, et al. The genomic basis of geographic differentiation and fiber improvement in cultivated cotton. Nat Genet. 2021;53(6):916–24.
3. Couasnet G, Abidine MZE, Laurens F, Dutagaci H, Rousseau D. Machine learning meets distinctness in variety testing. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). 2021;1303–1311.
4. Dreisigacker S, Sharma RK, Huttner E, Karimov A, Obaidi MQ, Singh PK, Sansaloni C, Shrestha R, Sonder K, Braun HJ. Tracking the adoption of bread wheat varieties in Afghanistan using DNA fingerprinting. BMC Genomics. 2019;20(1):660.
5. Al-Samarai F, Al-Kazaz A. Molecular markers: an introduction and applications. Eur J Mol Biotechnol. 2015;9.
6. Grover A, Sharma PC. Development and use of molecular markers: past and present. Crit Rev Biotechnol. 2016;36(2):290–302.
7. Elakhdar A, Kumamaru T, Qualset CO, Brueggeman RS, Amer K, Capochichi L. Assessment of genetic diversity in Egyptian barley (*Hordeum vulgare* L.) genotypes using SSR and SNP markers. Genet Resour Crop Evol. 2018;65(7):1937–51.
8. De Beukelaer H, Davenport GF, Fack V. Core Hunter 3: flexible core subset selection. BMC Bioinf. 2018;19(1):203.
9. Song QJ, Hyten DL, Jia GF, Quigley CV, Fickus EW, Nelson RL, Cregan PB. Fingerprinting soybean germplasm and its utility in genomic research. G3 (Bethesda). 2015;5(10):1999–2006.
10. Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, Weise S, Knüpffer H, Basterrechea M, König P, et al. Genebank genomics highlights the diversity of a global barley collection. Nat Genet. 2019;51(2):319–26.
11. Jaiswal S, Sheoran S, Arora V, Angadi UB, Iquebal MA, Raghav N, Aneja B, Kumar D, Singh R, Sharma P, et al. Putative microsatellite DNA marker-based wheat genomic resource for Varietal Improvement and Management. Front Plant Sci. 2017;8:2009.
12. Wang YY, Lv HK, Xiang XH, Yang AG, Feng QF, Dai PG, Li Y, Jiang X, Liu GX, Zhang XW. Construction of a SNP Fingerprinting Database and Population Genetic Analysis of Cigar Tobacco Germplasm Resources in China. Front Plant Sci. 2021;12:618133.
13. Portis E, Lanteri S, Barchi L, Portis F, Valente L, Toppino L, Rotino GL, Acquadro A. Comprehensive characterization of simple sequence repeats in Eggplant (Solanum melongena L.) Genome and Construction of a web resource. Front Plant Sci. 2018;9:401.
14. Kimura T, Sugisawa T, Taira M. International Union for the Protection of New varieties of plants. Report on Developments of a Software Tool for Marker Selection Using the Traveling Salesman Algorithm; 2019.
15. International Union for the Protection of New Varieties of Plants. Guidelines for DNA-Profiling: Molecular Marker Selection and Database Construction ("BMT Guidelines") 2020.
16. Schulthess AW, Kale SM, Liu F, Zhao Y, Philipp N, Rembe M, Jiang Y, Beukert U, Serfling A, Himmelbach A, et al. Genomics-informed prebreeding unlocks the diversity in genebanks for wheat improvement. Nat Genet. 2022;54(10):1544–52.
17. Hiremath PJ, Kumar A, Penmetsa RV, Farmer A, Schlueter JA, Chamarthi SK, Whaley AM, Carrasquilla-Garcia N, Gaur PM, Upadhyaya HD, et al. Large-scale development of cost-effective SNP marker assays for diversity assessment and genetic mapping in chickpea and comparative mapping in legumes. Plant Biotechnol J. 2012;10(6):716–32.
18. Zeng XQ, Guo Y, Xu QJ, Mascher M, Guo GG, Li SC, Mao LK, Liu QF, Xia ZF, Zhou JH, et al. Origin and evolution of qingke barley in Tibet. Nat Commun. 2018;9(1):5433.
19. Pankin A, Altmüller J, Becker C, von Korff M. Targeted resequencing reveals genomic signatures of barley domestication. New Phytol. 2018;218(3):1247–59.
20. Broccanello C, Chiodi C, Funk A, McGrath JM, Panella L, Stevanato P. Comparison of three PCR-based assays for SNP genotyping in plants. Plant Methods. 2018;14:28.

21. Shikari AB, Najeeb S, Khan G, Mohidin FA, Shah AH, Nehvi FA, Wani SA, Bhat NA, Waza SA, Subba Rao LV, et al. KASP™ based markers reveal a population sub-structure in temperate rice (Oryza sativa L.) germplasm and local landraces grown in the Kashmir valley, north-western Himalayas. Genet Resour Crop Evol. 2020;68(3):821–34.
22. Sejake T, Shargie N, Christian R, Amelework AB, Tsilo TJ. Genetic diversity in sorghum (Sorghum bicolor L. Moench) accessions using SNP based Kompetitive allele-specific (KASP) markers. Aust J Crop Sci. 2021;15(06):890–8.
23. Li X, Guo Y, Huang F, Wang Q, Chai J, Yu F, Wu J, Zhang M, Deng Z. Authenticity identification of Saccharum officinarum and Saccharum spontaneum germplasm materials. Agronomy. 2022;12(4).
24. Fan H, Wang T, Li Y, Liu H, Dong Y, Zhang R, Wang H, Shang L, Xing X. Development and validation of a 1 K sika deer (Cervus nippon) SNP chip. BMC Genom Data. 2021;22(1):35.
25. Kuang M, Wei SJ, Wang YQ, Zhou DY, Ma L, Fang D, Yang WH, Ma ZY. Development of a core set of SNP markers for the identification of upland cotton cultivars in China. J Integr Agric. 2016;15(5):954–62.
26. Wang Y, Wu XH, Li YW, Feng ZS, Mu ZH, Wang J, Wu XY, Wang BG, Lu ZF, Li GJ. Identification and validation of a core single-nucleotide polymorphism marker set for genetic Diversity Assessment, Fingerprinting Identification, and Core Collection Development in Bottle Gourd. Front Plant Sci. 2021;12:747940.
27. Varshney RK, Thiel T, Sretenovic-Rajicic T, Baum M, Valkoun J, Guo P, Grando S, Ceccarelli S, Graner A. Identification and validation of a core set of informative genic SSR and SNP markers for assaying functional diversity in barley. Mol Breed. 2007;22(1):1–13.
28. Wu XY, Wang BG, Wu SQ, Li SJ, Zhang Y, Wang Y, Li YW, Wang J, Wu Xh, Lu ZF et al. Development of a core set of single nucleotide polymorphism markers for genetic diversity analysis and cultivar fingerprinting in cowpea. Legume Sci. 2021;3(3).
29. Liu ZX, Li J, Fan XH, Htwe NMPS, Wang SM, Huang W, Yang JY, Xing LL, Chen LJ, Li YH, et al. Assessing the numbers of SNPs needed to establish molecular IDs and characterize the genetic diversity of soybean cultivars derived from Tokachi nagaha. Crop J. 2017;5(4):326–36.
30. Li PR, Su TB, Yu SC, Wang HP, Wang WH, Yu YJ, Zhang DS, Zhao XY, Wen CL, Zhang FL. Identification and development of a core set of informative genic SNP markers for assaying genetic diversity in Chinese cabbage. Hortic Environ and Biotechnol. 2019;60(3):411–25.
31. Fujii H, Ogata T, Shimada T, Endo T, Iketani H, Shimizu T, Yamamoto T, Omura M. Minimal marker: an Algorithm and Computer Program for the identification of minimal sets of discriminating DNA markers for efficient Variety Identification. J Bioinf Comput Biol. 2013;11:02.
32. Du HS, Yang JJ, Chen B, Zhang XF, Zhang J, Yang K, Geng SS, Wen CL. Target sequencing reveals genetic diversity, population structure, core-SNP markers, and fruit shape-associated loci in pepper varieties. BMC Plant Biol. 2019;19(1):578.
33. Zhang J, Yang JJ, Zhang L, Luo J, Zhao H, Zhang JN, Wen CL. A new SNP genotyping technology target SNP-seq and its application in genetic analysis of cucumber varieties. Sci Rep. 2020;10(1):5623.
34. Yuan X, Li ZR, Xiong LW, Song SF, Zheng XF, Tang ZH, Yuan ZM, Li LZ. Effective identification of varieties by nucleotide polymorphisms and its application for essentially derived variety identification in rice. BMC Bioinf. 2022;23(1):30.
35. Darrier B, Russell J, Milner SG, Hedley PE, Shaw PD, Macaulay M, Ramsay LD, Halpin C, Mascher M, Fleury DL, et al. A comparison of mainstream genotyping platforms for the evaluation and use of Barley Genetic resources. Front Plant Sci. 2019;10:544.
36. Sun RJ, Sun BC, Tian Y, Su SS, Zhang Y, Zhang WH, Wang JS, Yu P, Guo BF, Li HH, et al. Dissection of the practical soybean breeding pipeline by developing ZDX1, a high-throughput functional array. Theor Appl Genet. 2022;135(4):1413–27.
37. Jia MM, Yang LJ, Zhang W, Rosewarne G, Li JH, Yang E, Chen L, Wang WX, Liu Y, Tong HW, et al. Genome-wide association analysis of stripe rust resistance in modern Chinese wheat. BMC Plant Biol. 2020;20(1):491.
38. Wang WS, Mauleon R, Hu ZQ, Chebotarov D, Tai SS, Wu ZC, Li M, Zheng TQ, Fuentes RR, Zhang F, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557(7703):43–9.
39. Bukowski R, Guo XS, Lu YL, Zou C, He B, Rong ZQ, Wang B, Xu DW, Yang BC, Xie CX, et al. Construction of the third-generation *Zea mays* haplotype map. Gigascience. 2018;7(4):1–12.
40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
41. Swarts K, Li HH, Romero Navarro JA, An D, Romay MC, Hearne S, Acharya C, Glaubitz JC, Mitchell S, Elshire RJ et al. Novel methods to optimize Genotypic Imputation for Low-Coverage, Next-Generation sequence data in crop plants. The Plant Genome. 2014;7(3).
42. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.
43. Shannon CE, Weaver W. The mathematical theory of communication. University of Illinois Press; 1949.
44. Lander ES, Green P. Construction of multilocus genetic linkage maps in humans, Proc. Natl. Acad. Sci. U. S. A. 1987;84:2363–2367.
45. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35(10):1786–8.
46. Mantel N. The detection of Disease Clustering and a generalized Regression Approach. Cancer Res. 1967;27:209–20.
47. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
48. Wickham H. ggplot2. WIREs Comp Stat. 2011;3:180–185.
49. Hao ZD, Lv DK, Ge Y, Shi JS, WeiJers D, Yu GC, Chen JH. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. PeerJ Comput Sci. 2020;6.
50. Yang Y, Tian HL, Wang R, Wang L, Yi HM, Liu YW, Xu LW, Fan YM, Zhao JR, Wang FG. Variety discrimination power: an Appraisal Index for loci Combination Screening Applied to Plant Variety discrimination. Front Plant Sci. 2021;12:566796.

Publisher's Note