Plant Communications
Resource Article

CellPress
Partner Journal

HTPdb and HTPtools: Exploiting maize haplotype-tag polymorphisms for germplasm resource analyses and genomics-informed breeding

Yikun Zhao[1,7], Hongli Tian[1,7], Chunhui Li[2,7], Hongmei Yi[1,7], Yunlong Zhang[1], Xiaohui Li[3], Han Zhao[4], Yongxue Huo[1], Rui Wang[1], Dingming Kang[5], Yuncai Lu[6], Zhihao Liu[1], Ziyue Liang[1], Liwen Xu[1], Yang Yang[1], Ling Zhou[4], Tianyu Wang[2,*], Jiuran Zhao[1,*] and Fengge Wang[1,*]

[1]Maize Research Center, Beijing Academy of Agriculture & Forestry Sciences (BAAFS), Beijing Key Laboratory of Maize DNA Fingerprinting and Molecular Breeding, Shuguang Garden Middle Road No. 9, Beijing 100097, China

[2]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[3]Jilin Academy of Agricultural Sciences, Maize Research Institute, Gongzhuling 136100, China

[4]Provincial Key Laboratory of Agrobiology, Institute of Crop Germplasm and Biotechnology, Jiangsu Academy of Agricultural Sciences, Nanjing 210014, China

[5]College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China

[6]College of Advanced Agriculture and Ecological Environment, Heilongjiang University, Harbin 150080, China

[7]These authors contributed equally to this article.

***Correspondence: Tianyu Wang (wangtianyu@caas.cn), Jiuran Zhao (maizezhao@126.com), Fengge Wang (fenggewangmaize@126.com)**

https://doi.org/10.1016/j.xplc.2022.100331

ABSTRACT

Along with rapid advances in high-throughput-sequencing technology, the development and application of molecular markers has been critical for the progress that has been made in crop breeding and genetic research. Desirable molecular markers should be able to rapidly genotype tens of thousands of breeding accessions with tens to hundreds of markers. In this study, we developed a multiplex molecular marker, the haplotype-tag polymorphism (HTP), that integrates Maize6H-60K array data from 3,587 maize inbred lines with 6,375 blocks from the recombination block map. After applying strict filtering criteria, we obtained 6,163 highly polymorphic HTPs, which were evenly distributed in the genome. Furthermore, we developed a genome-wide HTP analysis toolkit, HTPtools, which we used to establish an HTP database (HTPdb) covering the whole genomes of 3,587 maize inbred lines commonly used in breeding. A total of 172,921 non-redundant HTP allelic variations were obtained. Three major HTPtools modules combine seven algorithms (e.g., chain Bayes probability and the heterotic-pattern prediction algorithm) and a new plotting engine named "BCplot" that enables rapid visualization of the background information of multiple backcross groups. HTPtools was designed for big-data analyses such as complex pedigree reconstruction and maize heterotic-pattern prediction. The HTP-based analytical strategy and the toolkit developed in this study are applicable for high-throughput genotyping and for genetic mapping, germplasm resource analyses, and genomics-informed breeding in maize.

Key words: haplotype tag, database, maize, pedigree reconstruction, heterotic-pattern prediction, genomics-informed breeding

Zhao Y., Tian H., Li C., Yi H., Zhang Y., Li X., Zhao H., Huo Y., Wang R., Kang D., Lu Y., Liu Z., Liang Z., Xu L., Yang Y., Zhou L., Wang T., Zhao J., and Wang F. (2022). HTPdb and HTPtools: Exploiting maize haplotype-tag polymorphisms for germplasm resource analyses and genomics-informed breeding. Plant Comm. **3**, 100331.

INTRODUCTION

Advances in genomics research have included the development of various technologies that facilitate the efficient and accurate translation of genetic variations into crop improvements. Molecular breeding accelerates the selection process by changing the genotypes and even the haplotypes of animals and plants. Obtaining a high-quality haplotype map of a species will promote

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

the development of industries related to that species while also enhancing the molecular breeding of animals and plants to accelerate the selection of new varieties and the prevention of human diseases.

Developing cost-effective genotyping platforms for breeding is still a major objective of breeders. To be useful for breeding, molecular markers should ideally be able to quickly genotype tens of thousands of breeding accessions with tens to hundreds of markers (Rasheed et al., 2017). Thus, a comprehensive haplotype map may be useful. An effective strategy for minimizing genotyping costs during breeding involves fully exploiting pre-existing genomic data. Because of linkage disequilibrium and the coinheritance of sequence variants within a single haplotype, a subset of representative variants (i.e., tag variants), including SNPs and insertions or deletions (InDels), may be sufficient for identifying haplotypes. Therefore, a haplotype map can extrapolate the data from existing whole-genome sequences or arrays and determine the genotypes of the individuals in a breeding group. Thus, it uses pre-existing whole-genome sequences or array data to minimize the amount of new genotype data that researchers must generate (Jensen et al., 2020; Torkamaneh et al., 2021).

In the present study, we developed a multiplex molecular marker, the haplotype-tag polymorphism (HTP), that integrates the Maize6H-60K array data from 3,587 maize inbred lines. On the basis of a rigorous evaluation, we demonstrated that the HTP is a high-resolution and highly efficient multiplex molecular marker. In addition, HTPs are genome-wide haplotype tags useful for efficiently assessing the whole-genome background of tens of thousands of samples at the same time. HTPs can simplify the characterization of whole genomes on the basis of seamless blocks resulting from recombination (i.e., the haplotype tag of the whole genome). These cosegregating blocks can be efficiently and accurately detected and analyzed. Furthermore, we developed a custom Python script called HTPtools as a genome-wide HTP analysis toolkit. We used HTPtools to establish an HTPdb (i.e., a haplotype-tag allelic variation and frequency database) covering the whole genomes of 3,587 maize inbred lines commonly used for breeding worldwide and generated 172,921 non-redundant HTP allelic variations. We also developed "Data prediction," "Group analysis," and "Data comparison" as HTPtools advanced application modules, which integrate seven algorithms (such as chain Bayes probability, the expectation maximization (EM) algorithm, and the heterotic-pattern prediction algorithm) and a new plotting engine named "BCplot" (Python Graphics) for rapidly visualizing the background information of a backcross group. We designed HTPtools to satisfy the increasing requirements for big-data analyses of germplasm resources and genomics-informed breeding, including the reconstruction of complex pedigrees, the prediction of maize heterotic patterns, and the assessment of the genomic background of a backcross group via the visualization of interactive data. Our results suggest that the HTP and HTPtools developed in this study will be useful for efficient and high-throughput genotyping, germplasm resource analyses, and genomics-informed breeding of maize. Because maize is a model genetic system, the HTP-based analytical strategy described herein may be relevant for investigating genetic variations and marker-assisted breeding in other crops. The robustness of HTPtools has been validated by research institutions, making it a useful tool for breeders.

RESULTS AND DISCUSSION

Development of a database consisting of genome-wide HTPs

We previously converted SNPs generated by genotyping by sequencing into effective recombination blocks (Li et al., 2015). We then updated our Maize6H-60K array (Tian et al., 2021) with 66,905 loci (60,830 SNPs and 6,007 new InDels), including 68 loci (65 SNPs and 3 InDels) from the chloroplast genome, based on the Affymetrix Axiom platform. Next, we updated the recombination block map that we developed using the B73 AGP_v3 reference genome and integrated 6,375 blocks from the recombination block map and 66,905 loci to generate a preliminary version of the haplotype tag using the same reference genome. During this process, cosegregating SNP and InDel markers were combined in a recombination block to be used as an effective haplotype tag (i.e., HTP). The high quality of the SNPs and InDels from the Maize6H-60K array has been confirmed. Thus, they were used as the core set to represent each haplotype block (i.e., to "tag" each block). Finally, we obtained 6,163 HTPs (6,163 loci) (Supplemental Table 1), which are multiple seamlessly connected blocks covering 98.85% of the B73 reference genome. The HTP distribution density indicated that despite a few large gaps, the tags were uniformly distributed across the maize genome, with an average genetic distance of about 0.23 cM between adjacent HTPs (Figure 1D). Their genome-wide coverage was an important feature of the HTPs that enabled the reconstruction of complex pedigrees, the prediction of maize heterotic patterns, and the visualization of background information of backcross groups.

We subsequently collected maize germplasm resources from various regions (e.g., China and the US). We obtained a total of 3,587 accessions, including important elite inbred lines from each region (e.g., B73, Mo17, HZS, Dan340, and Zheng58), and genotyped them using the 6,163 HTPs. We then constructed an HTPdb comprising 172,921 non-redundant HTP allelic variations from the 3,587 accessions (Supplemental Table 2). Accordingly, we comprehensively evaluated the database and divided the 3,587 accessions into 11 groups on the basis of the 6,163 HTPs (Figure 1C). The number of pan-HTP (total HTP allelic variations) increased as groups were added (i.e., 85,546 allelic variations in one group to 172,911 allelic variations in 11 groups) (Figure 1E). By contrast, the number of core HTPs (core HTP allelic variations) decreased as groups were added (i.e., 85,546 allelic variations in one group to 20,986 allelic variations in 11 groups). Statistical analyses revealed that the 11 groups included 20,986 core allelic variations and 14,654 private allelic variations (Supplemental Table 1). When the group number increased to nine groups, the total number of non-redundant HTP allelic variations nearly plateaued, reflecting the representativeness of these 3,587 maize accessions (11 groups). We also determined that the proportions of non-redundant HTP allelic variations increased with every 10 inbred lines, with 100 random selections during each sampling. Our results indicated that a plateau was approached when the sample size reached 1,450, and at least 95% of the total allelic variations were included (Figure 1F).

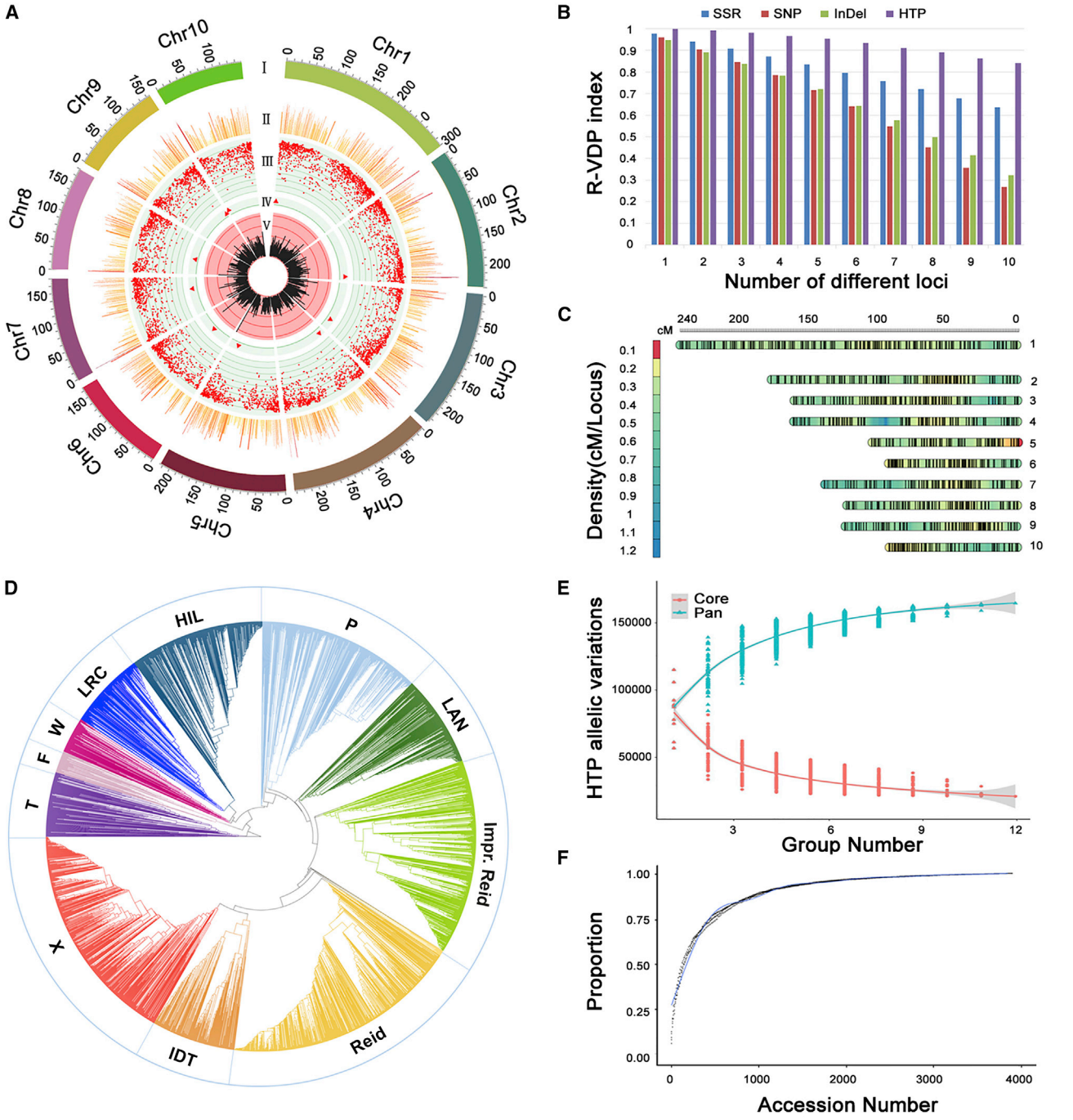

Figure 1. Evaluation of the HTPs.

(A) Circos graph presenting the distribution of all HTPs across the maize genome. I, physical position of each HTP on 10 chromosomes; II, number of SNPs and InDels in each HTP; III, PIC value of each HTP; IV, red triangles indicate the 10 HTPs with the highest PIC values on each chromosome; V, number of allelic variations for each HTP.

(B) The selection method is based on the fixed number of different loci (abscissa). The ordinate represents the R-VDP index. Forty SSR, SNP, InDel, and HTP markers were selected as DNA fingerprints to determine the varieties of 2,800 samples. The identification results are displayed according to the R-VDP.

(C) Heatmap presenting the HTP distribution density across the 10 maize chromosomes.

(D) Using 6,163 HTPs, 3,587 samples were genetically clustered into the following 11 groups: tropical (T), flint inbred lines (F), waxy (W), LvDa Red Cob (LRC), HZS-improved line (HIL), Improved Reid (Impr. Reid), Reid, Lancaster (LAN), P, Iodent (IDT), and X.

(E) Numbers of pan (total HTP allelic variations) and core (core HTP allelic variations) HTPs as maize groups were added. The 3,587 maize inbred lines were grouped into 11 groups according to the HTPs.

(F) Proportions of HTP allelic variations. Ten samples were randomly selected each time. The shaded area represents 100 repetitions for each sampling time-point.

An HTP-based analytical strategy can divide the whole genome into blocks with haplotype tags during analyses of the genomic background of germplasm resources. A single HTP or a few HTPs may be selected as independent molecular markers for variety identification, which may be relevant for protecting intellectual property rights. A survey of all HTPs in the maize B73 reference genome revealed that our 6,163 nuclear HTPs were evenly distributed in the whole genome and were highly polymorphic (Figure 1A). Each HTP contained several SNPs and InDels. Among the 6,163 HTPs, 2,811 had more than 10 variations (SNPs and InDels combined). In addition, our HTPs were highly informative. More specifically, 5,549 of the markers had a polymorphism information content (PIC) value exceeding 0.5, and 576 had a PIC value greater than 0.9. Furthermore, regarding their polymorphism, 811 HTPs had more than 50 allelic variations, and 12 HTPs had more than 100 allelic variations (Supplemental Table 1). We also compared the variety discrimination power of SNPs, simple sequence repeats (SSRs) (Wang et al., 2014, 2017), InDels, and HTPs using VDPtools (Figure 1B) (Yang et al., 2021). Our results indicated that the HTPs were better for distinguishing between maize varieties than the other three molecular markers, making them potentially useful for protecting crop-related intellectual property rights.

The summary statistics for the HTPs and HTPdb are available online (https://htp.plantdna.site/database/nucleus-haplotype).

HTPtools and the online analysis platform

We developed an HTP analysis toolkit called HTPtools based on the Python3 environment. More specifically, we used an interactive command line tool built using the "Click" module. HTPtools can run on Linux and Windows platforms where Python3 has been installed, with good cross-platform features. The basic function of HTPtools involves the conversion of Maize6H-60K array data into HTP data (data format conversion or preprocessing). Using HTPtools, we established an HTPdb covering 3,587 maize inbred lines commonly used for breeding worldwide. In addition to its basic function, HTPtools has the following three advanced application modules: Data comparison, Data prediction, and Group analysis. The Data comparison module integrates an efficient genome-wide haplotype-tag sequence comparison algorithm and an HTP comparison algorithm (inbred line and hybrid).

The core aspect of an HTP-based analytical strategy involves the integration of available information on whole genomes of tens of thousands of samples and the subsequent analysis using haplotype tags. Thus, we developed the Data prediction and Group analysis advanced application modules that exploit the characteristics of HTPs to generate whole-genome haplotype tags. First, we developed an inbred-line pedigree analysis (ILPA) module and integrated it into the Data prediction module, which can be used to identify candidate inbred lines for "Inbred X" and for reconstructing pedigree breeding histories, especially those of incomplete pedigrees. We also developed a heterotic-pattern analysis (HPA) module according to the HTPs and then integrated it into the Data prediction module. The HPA module can be used to infer heterotic patterns. Second, the Group analysis module was designed to incorporate a genome-wide background assessment pipeline for maize groups and our newly developed plotting engine, BCplot (Python Graphics), into HTPtools, enabling the rapid visualization of background information for a group with mass samples.

We also developed an algorithm for predicting HTP loci on the basis of the haplotype-tag dictionary (HTPdb) and integrated it into the HTP predicting module, which was also integrated into the Data prediction module. The main theory underlying this method is chain Bayes probability. We used group data to construct a haplotype-tag dictionary according to the EM algorithm. Frequency information was included for each haplotype tag in the dictionary. This dictionary can be used to predict the full haplotype tag of a specific sample with individual SNPs/InDels within an HTP (each HTP contained several SNPs and InDels). Accordingly, even with very few HTP loci, we can obtain the complete haplotype-tag information using the generated haplotype-tag dictionary. This method can substantially decrease the number of HTP loci needed for a particular application.

HTPtools is available online (https://github.com/plantdna/htp). We developed a freely accessible online platform using Next.js technology (https://htp.plantdna.site/). This platform enables researchers and breeders to identify varieties and conduct genetic analyses.

Inferring Inbred X using HTPtools

We developed a method that supports complex pedigree reconstructions via the genome-wide HTPs and HTPtools along with the maize HTPdb (from 3,578 accessions). Details regarding the parents of several maize varieties generated via breeding have been lost. In a previous study, the unknown parent was designated as Inbred X (Lai et al., 2010). In the current study, we used our maize HTPdb and the Data prediction module of HTPtools to identify candidate inbred lines for Inbred X. To verify the reliability of our method, we inferred the identity of Inbred X in a known pedigree. Jing2416 and Jing24 are elite inbred lines of the Chinese HIL heterotic group, whose pedigree history is well known. The parents of Jing2416 are Jing24 and "5237." Jing24 was considered to be Inbred X in this pedigree.

After compiling the HTP data for Jing2416 and 5237, we used the ILPA module to infer that Jing2416 inherited part of its genome only from 'Inbred X' (2,936 HTPs, 'Inbred X' private HTPs) (orange lines in Figure 2A, 2B and 2C). The remaining genomic regions were filtered out in the next step (3,227 HTPs, Filtered HTPs) which were likely inherited from two ways, one was '5237' only, another was '5237' and 'Inbred X' shared. (gray lines in Figure 2A and 2B) (Supplemental Table 3). The ILPA module was then used to identify the outliers (details are in Methods) in the data for single and continuous HTPs ($\geq$2 seamlessly connected HTPs) in the genomic regions from Inbred X private HTPs (Figure 2C) and to decrease the background noise. We detected 34 long continuous fragments (LCFs) in the Jing2416 genome that contained multiple seamlessly connected HTPs, with an average length of approximately 13 Mb. An earlier investigation revealed that meiotic recombination is predictable across diverse maize hybridizations (Rodgers-Melnick et al., 2015). Therefore, LCFs were used to infer Inbred X. We next sorted the LCFs on Jing2416 by length (number of

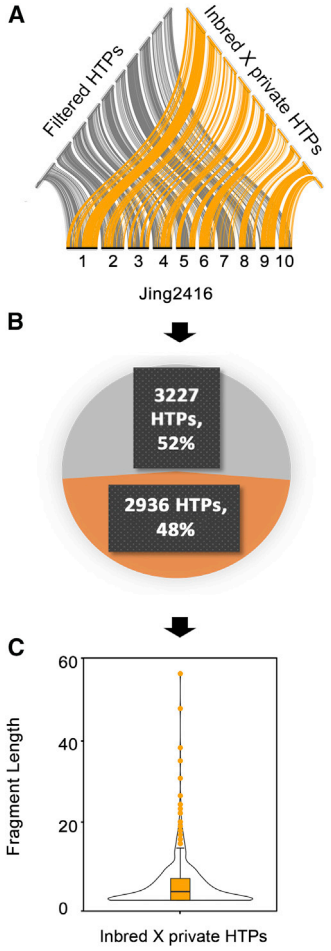

Figure 2. Inferring Inbred X using HTPs during maize breeding.
Inferring the HTPs in Inbred X in the Jing2416 breeding pedigree.
(A) Jing24 was considered to be Inbred X during the breeding of Jing2416. The gray lines represent the Jing2416 inherited part of its genomic regions which were likely inherited from two ways, one was '5237' only, another was '5237' and 'Inbred X' shared. The orange lines represent Jing2416 inherited part of its genome only from 'Inbred X'.
(B) After completing the genome-wide HTP comparison, the proportions of the Jing2416 genome inherited from 'Filtered HTPs' and 'Inbred X private HTPs' were inferred.
(C) "Fragment Length" represents the number of HTPs in each fragment.

continuous HTPs). Then, the longest LCF (HTP5611–HTP5713) was used to identify the matched fragment in the database (https://htp.plantdna.site/analysis/ilpa) using a developed LCF matching algorithm. On the basis of breeding history, we excluded varieties that were apparently bred later than Jing2416, enabling us to infer that the unknown parent was Jing24 (LCF matching degree: 0.9318). We then analyzed the second (HTP3839–HTP3937) and third (HTP5498–HTP5594) longest LCFs. The results showed that all three LCFs were accurately inferred as being from Jing24 (mean LCF matching degree: 0.917). One of the two parents of Jing2416 is known. Our findings suggested that the unknown parent is Jing24.

Maize heterotic patterns revealed by HTPtools

Heterosis is a common phenomenon that has been exploited to improve the production of many crops, including rice (*Oryza sativa*), cotton (*Gossypium hirsutum*), and maize (*Zea mays*), and hybrids are commercially available. Analyses of heterotic patterns can assist government departments with the regulation of commercial hybrids, but some of the archived information about hybrid pedigrees is inaccurate and/or incomplete. The method for predicting heterotic patterns we developed relies on big data validation to obtain information on hybrid heterotic patterns (i.e., information about the heterotic group of the hybrid parents). This method can be used to cross-reference the archival data or to validate data, which is important for future maize breeding and related research.

In this study, we developed an HPA module on the basis of HTPs and an heterotic-pattern prediction (HPP) module, which were subsequently integrated into HTPtools. Available information on the HTPs in the maize group can be used to complete virtual hybridizations and obtain virtual heterotic patterns. Researchers can select known maize groups with Maize6H-60K array data or HTP data in HTPtools and determine the combination of heterotic patterns that must be constructed. Once the information is collected, the HPP module with the heterotic-pattern prediction algorithm can infer the hybrid heterotic patterns.

To further clarify the utility of the HPA and HPP modules in HTPtools, we selected 674 elite inbred lines, including important inbred lines used in China for breeding (e.g., Jing2416, Zheng58, Chang7-2, and Jing92), and 100 Chinese state-approved maize hybrids with known heterotic patterns (e.g., ZD958, ND108, and XD20, which are widely cultivated in China, and their parents as triplets) as representative samples (Figure 3). First, we developed an HTPdb for the 674 inbred lines, which were divided into six groups. Second, accessions sampled from the six groups were hybridized within and between groups to generate virtual hybrids using the HPA module. We generated 21 heterotic-pattern prediction models (from 21 crosses) (details are in Methods). Information on the two groups that included the parents of the virtual hybrid was recorded and used for analyses in the HPP module. In this study, the private, core, and dispensable HTP allelic variations of each group were extracted using the HPP module and used for model evaluation (Figure 3A and 3C) (Supplemental Table 4). HTP allelic variations present in all 6 groups were defined as core HTP allelic variations, those present in 2 to 5 groups were defined as dispensable HTP allelic variations, and those present in only one group were defined as private HTP allelic variations. We then used the HPA function to randomly construct 2,100 virtual hybrids (100 in each heterotic pattern) (Supplemental Table 6) to evaluate the reliability of the HPP module. The results indicated that the predictions made by the algorithm were 100% accurate. Next, the heterotic patterns of the 100 Chinese state-approved hybrids were inferred by the HPP module with the heterotic-pattern prediction algorithm. Our results indicated that the predicted parental mating patterns of 98 hybrids were correct (98% correctly inferred) (Figure 3B; Supplemental Table 4), implying that our method for inferring heterotic patterns is reliable.

Global maize production mainly involves the cultivation of hybrid lines (Masuka et al., 2017a, b). Therefore, heterosis serves as the foundation of modern crop breeding (Birchler, 2016). The analytical strategy based on HTPs and HTPtools described

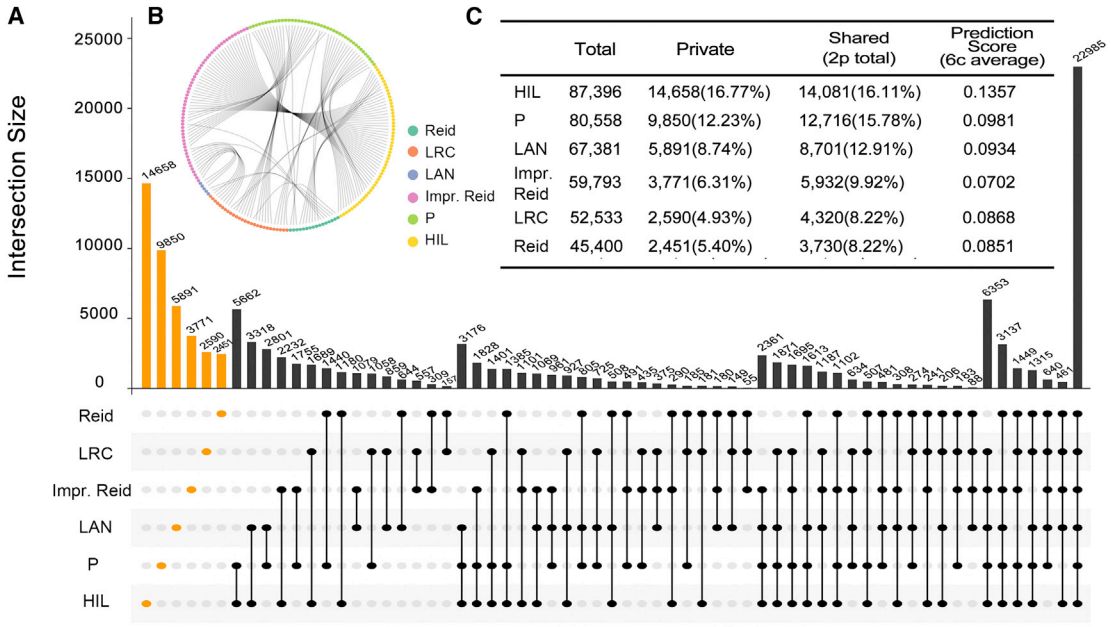

Figure 3. Crossing patterns during maize breeding exploiting heterosis revealed using HTPs.

(A) Private, core, and dispensable HTP allelic variations for each maize group. Individual dots (yellow) at the bottom of the graph represent the private HTP allelic variations (non-redundant) in each group. Two connected dots indicate HTP allelic variations (non-redundant) shared by two groups. Similarly, six connected dots indicate HTP allelic variations (non-redundant) shared by six groups.

(B) Gray curves within the circle represent the crossing patterns (e.g., HIL × Impr. Reid and P × LRC).

(C) Total number of HTP allelic variations, private HTP allelic variations, and HTP allelic variations shared by two groups (Shared (2p total)), as well as the corresponding proportion in each group. The prediction score (6c average) represents the average score of the heterotic-pattern prediction model for the six crossing patterns for each group (e.g., for HIL: HIL × HIL, HIL × P, HIL × LAN, HIL × Impr. Reid, HIL × LRC, and HIL × Reid). See Supplemental Table 4 for the complete data.

herein is potentially useful for predicting maize heterotic patterns. In addition to being an important crop species, maize is also a model genetic system. Thus, it is likely that the HTPs and HTPtools developed in this study will facilitate the discovery of new genetic variations and marker-assisted breeding in other crops.

Background selection using HTPs during backcross breeding

Regardless of whether they use traditional backcross-breeding methods or biotechnology-based methods involving transgenic plants and gene editing, breeders must determine the similarity in the genetic backgrounds of the progeny group material and the recipient (i.e., the genetic background recovery rates). Accurate genetic background analyses should enable the visualization of the overall background recovery in the recombination exchange blocks throughout the whole genome in all group samples. The HTP markers proposed in this study are distributed throughout the genome. Moreover, they can reflect the dynamic changes in the recombination process and can be used to visualize the recovery of the genetic background, which is ideal for molecular-marker-assisted breeding.

To verify the reliability of an HTP-based analytical strategy, we used HTPs to conduct backcross breeding tests (details in Methods). Using the HTP comparison algorithm and BCplot, HTPtools generated a genetic background recovery map for all samples (Figure 4A; (Supplemental Data 1; Supplemental Table 7). The Group analysis module can help breeders accurately screen out individual samples with a high background recovery rate, thereby shortening the generation time.

The Group analysis module can also generate graphs presenting additional data for the backcross groups, including the proportion of samples with different genetic background recovery rates (Figure 4B). The background recovery rates of the samples in the BC_1 generation exhibited a normal distribution trend. The background recovery rates of these samples were mostly concentrated between 0.4 and 0.6 (about 82% of all samples), whereas the background recovery rates of the BC_2 and BC_3 samples were higher and more concentrated. More specifically, the background recovery rates of the BC_2 and BC_3 samples were mainly 0.8–0.9 (about 71.8% of all samples) and 0.9–1.0 (about 57.3% of all samples), respectively.

The plotting engine BCplot can graphically present the proportion of group samples in different background recovery-rate intervals. This information is important for establishing the maize backcross group size and may be used to prevent breeders from blindly increasing or decreasing the group size. The distribution of the samples that exchanged genetic material from each chromosome (Figure 4C) and the frequency of the exchange of the whole-genome HTPs on 10 chromosomes (Figure 4D) were determined. As the number of generations increased, the average number of recombinations and exchanges of whole

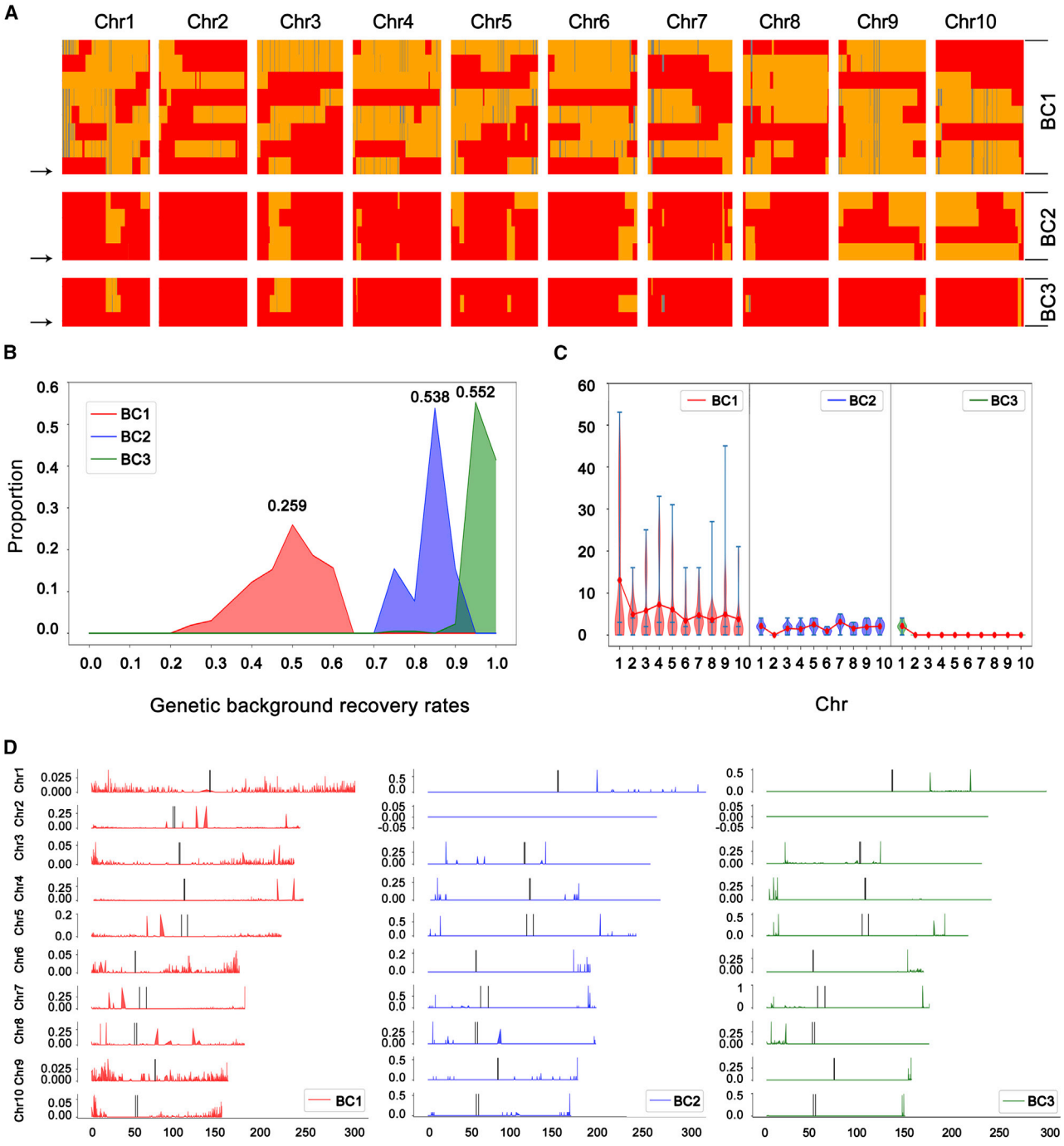

Figure 4. Background selection using HTP markers during backcross breeding.

(A) Genetic background recovery map of individual backcross groups. Red represents the same genotype (HTP) as the recipient, gray represents the same genotype as the donor, and yellow represents a heterozygous genotype. The sample with the highest recovery rate in different recovery intervals is presented (BC1: 0.2–0.25, 0.25–0.3, 0.3–0.35, 0.35–0.4, 0.4–0.45, 0.45–0.5, 0.5–0.55, 0.55–0.6; BC2: 0.7–0.75, 0.75–0.8, 0.8–0.85, 0.85–0.9; BC3: 0.7–0.75, 0.75–0.8, 0.85–0.9, 0.9–0.95, 0.95–1.0). See Supplemental Table 5 for the complete data. The black arrow indicates the recommended candidates.

(B) Proportion of samples with different genetic background recovery rates.

(C) Distribution of the samples that exchanged genetic material from each chromosome. The red line represents the average value.

(D) Frequency of the exchange of the whole-genome HTPs on 10 chromosomes. Black bars indicate the centromere regions.

chromosomes of the samples decreased, as did the volatility. Outliers were undetectable in all generations. The average number of exchanges for all chromosomes, except for chromosome 1 in the BC_3 generation, was 0 (Figure 4C). Figure 4D presents the proportions of HTPs on each chromosome that were exchanged in three generations for all samples. There were almost no exchanges involving HTPs in the centromere regions of the 10 chromosomes. The frequency

of exchanges was low for the HTPs within a certain distance from the centromere (on both sides), in contrast to the relatively high frequency of exchanges for HTPs near the telomeres. As the number of generations increased, the number of loci altered by recombinations decreased.

Finally, this module can sort the recovery rates and directly recommend candidates. Using the HTPs and the Group analysis module, the background recovery rate of each sample in the backcross groups was calculated and displayed. Moreover, the exchange frequency in each chromosomal block was intuitively determined, which is useful information for breeding. HTPtools can decrease the background selection time during backcross breeding, especially for large groups comprising thousands of samples. We developed HTPtools to provide breeders with an efficient and convenient solution to problems associated with data handling. Furthermore, HTPtools can display the background of all samples very intuitively and recommend the best candidate.

METHODS

Maize materials, DNA extraction, and genotyping

We collected 3,587 maize inbred lines from the Maize Research Center, Beijing Academy of Agriculture and Forestry Sciences (BAAFS) for this study. All accessions were grown in the experimental field of the BAAFS. Genomic DNA was extracted from the fresh leaves of individual plants by the cetyltrimethyl ammonium bromide (CTAB) method (Wang et al., 2011a). The quality and quantity of the extracted genomic DNA were evaluated using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA) and by 0.8% agarose gel electrophoresis, respectively, to ensure that the DNA was appropriate for genotyping in the Maize6H-60K array.

All maize inbred lines were genotyped on the Affymetrix GeneTitan platform according to the procedure recommended by Affymetrix (Axiom 2.0 Assay for 384HT Array Format Automated Workflow User Guide rev.5).

Development of HTPdb

First, all samples were genotyped using the Maize6H-60K array (Tian et al., 2021) on the Affymetrix GeneTitan platform. We removed accessions with a missing data rate greater than 10% after genotyping using the Maize6H-60K array. Finally, the data from 3,587 accessions were divided according to the chromosomes and 6,375 blocks from the recombination block map. To obtain a set of high-quality HTPs, we filtered the SNPs that satisfied one of the following criteria: cumulative heterozygosity and missing data rate >0.7 or minor allele frequency (MAF) = 0. Next, we constructed a haplotype phased block map using the EM algorithm (3,587 accessions were used to complete the imputation of the HTPs) and generated an HTPdb.

HTPtools

HTP predicting algorithm development

We considered that each HTP has n loci (each locus in the HTP is an SNP or InDel) and m HTP haplotype tag sequences (HTP allelic variation sequences). Now, we have an incomplete haplotype tag sequence (IHTS). If there is one locus in the IHTS that is the same as the locus with the same location of one haplotype-tag sequence from the haplotype-tag dictionary (HTS-HTD), then the following Bayes formula applies:

$$P(H|E) = P(E|H)\,P(H)/P(E),$$

where P(H) is the probability that the IHTS is in the HTS-HTD, P(E) is the probability that one locus in the IHTS is the same as the locus in the HTS-HTD, and P(E|H) is the probability that when the IHTS is in the HTS-HTD, the locus in the IHTS is the same as the locus in the HTS-HTD. If only one locus in the IHTS is the same as the locus in the HTS-HTD, this is designated as E_1, whereas if two loci in the IHTS are the same as loci in the HTS-HTD, this is designated as E_2, and so on. Thus, we need to calculate $P(H|E_1, E_2, E_3, \ldots, E_n)$, which we abbreviate as $P(H|E_{1:n})$. On the basis of the chain rule of probability,

$$P\left(\bigcap_{k=1}^{n} E_k\right) = \prod_{k=1}^{n} P\left(E_k \middle| \bigcap_{j=1}^{k-1} E_j\right),$$

we can obtain the chain of Bayes:

$$P(H|E_{1:n}) = \frac{P(E_n|H)}{P(E_n)} P(H|E_{1:n-1}).$$

Now, we can calculate the result with the recursive method, which means that we can calculate $P(H|E_1)$ first, then use the result to calculate $P(H|E_2)$, etc. But, in each calculation iteration, there is difficulty in calculating $P(E_i)$, so we use another format of the Bayes calculation formula:

$$P(H|E) = P(E|H)P(H^*) / (P(E|H)P(H^*) + P(E|\neg H)(1 - P(H^*))),$$

where the mathematical logic symbol ¬ means "not."

In this formula format, we do not need to calculate P(E), and P(H*) is the result of the previous calculation.

After we calculate each P(H|E) in the m HTP haplotype-tag sequences, we then pick the sequence with the maximum value as the predicted result.

Inferring Inbred X using HTPs

Using the inferred Jing2416 pedigree as an example, we compared the HTPs in the whole genome to identify HTPs in the Jing2416 genome that were not from 5237. We speculated that these different HTPs may have been inherited only from Inbred X. These different HTPs were extracted and sorted according to fragment length, from single HTPs (whose fragment length was 1) through continuous HTPs (i.e., ≥2 seamlessly connected HTPs, whose fragment length was equal to the number of HTPs). If the length of the fragment exceeded the upper limit (Q3 + 1.5 interquartile range [IQR]), it was designated as an outlier. For determining the outliers, Q3 represents the third quartile of the total number of single and continuous HTPs, and IQR represents the difference between the third quartile and the first quartile (i.e., IQR = Q3 − Q1), which is indicated by the width of the box. Outliers were 1.5× the width of the box. We selected the fragments belonging to the outlier, and the ILPA module was used to reduce the background noise. The LCFs of Inbred X were then generated from the ILPA module. Finally, we uploaded the LCF data (in CSV format) to the HTP analysis platform (https://htp.plantdna.site/analysis/ilpa), which was used to search for the best match in the database.

Maize heterotic patterns revealed by HTPtools

We generated 21 heterotic-pattern prediction models (from 21 crosses) using the HPA module. Then, the private, core, and dispensable HTP allelic variations of each group were extracted using the HPP module (Figure 3A and 3C; Supplemental Table 4). In addition, the proportions of private and shared (by two groups) HTP allelic variations for each virtual hybrid sample in the corresponding heterotic pattern were determined. We then formulated the evaluation parameter (threshold) of each model (i.e., prediction score) on the basis of the weight of the private HTP allelic variations and the HTP allelic variations shared by two groups for all virtual hybrids from each cross. For the prediction score, we first calculated the proportion (G) of the private HTP allelic variations and the HTP allelic variations shared by two groups (corresponding to each heterotic pattern) in the genome of all virtual

hybrids in each heterotic pattern. The prediction score (P) formula was expanded as follows:

$$P = \overline{G} + 3 \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G_i - \overline{G})^2}$$

The predictions were unacceptable if the minimum threshold ($\geq$0.05) was not satisfied (there was no upper limit). Therefore, the representativeness of each group sample was important (Supplemental Table 4). The 21 models that satisfied the minimum threshold were analyzed using the heterotic-pattern prediction algorithm to infer the hybrid heterotic patterns.

Background selection using HTPs during backcross breeding

We selected 400 individual samples in the BC_1 generation. After a foreground-selection step (target gene detection), the 264 remaining individual samples were analyzed using HTPtools. Finally, one BC_1 sample was selected to produce the BC_2 generation. A total of 400 individual samples in the BC_2 generation were selected, but only 12 individual samples remained after the foreground-selection step. Following the same analysis using HTPtools, one sample was finally selected to produce the BC_3 generation. After 400 individual samples in the BC_3 generation were selected and screened, the remaining 181 individual samples were evaluated using HTPtools. Finally, one plant with a recovery rate of 99.4% was selected for self-purification. The final sample with the target gene was obtained following a field verification step.

We collected individual samples from each generation of the backcross group after the foreground-selection step (target gene detection) and used HTPtools for genotyping. The HTPtools basic function and the Group analysis advanced application module were used for the data format conversion or preprocessing, noise reduction, and fitting, which was completed on the basis of polynomial curve fitting and partial least squares. This module can efficiently analyze the whole-genome haplotype-tag data from thousands of samples at the same time, thereby saving time and resources. Subsequently, the whole-genome HTPs of each sample were superimposed on the backgrounds of the donor and recipient.

Evaluation of the HTPs

The R-VDP values were calculated using VDPtools (v.1.1.1.0) (Yang et al., 2021). A total of 2,800 accessions were analyzed using 40 SSRs (NY/T 1432-2014 Technical Regulations for Identification of Maize Varieties in China), 40 SNPs, 40 InDels, and 40 HTPs. The SSR analysis was performed as previously described (Wang et al., 2011b). Markers that were evenly distributed on the 10 chromosomes with a high PIC value were selected.

Online analysis platform

The HTP online platform was built using the Next.js framework, with the front end comprising Ant Design components powered by Rect.js. The Next.js framework is a versatile router that connects web pages to the APIs. The built-in Webpack package made it convenient to build and deploy the released version. The Prettier tool standardized the code format. The Eslint tool was used to eliminate obvious bugs from the code. Babel maintained the compatibility between the normal JS and ES6.

Data availability

We developed a freely accessible online platform using Next.js technology (https://htp.plantdna.site/). The summary statistics of the HTPs and HTPdb are available online (https://htp.plantdna.site/database/nucleus-haplotype). HTPtools is available online (https://github.com/plantdna/htp).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at *Plant Communications Online*.

FUNDING

This work was supported by grants from the 13th Five-Year National Key R&D Program of China (2017YFD0102001).

AUTHOR CONTRIBUTIONS

F.W., J.Z., and T.W. conceived the project; Y. Zhao, H.T., C.L., and H.Y. wrote and modified the manuscript; Y.Z., Y. Zhang, Z. Liu, and C.L. designed the functions of the toolkit; Y.H. and H.Z. designed the homepage of the HTP online platform. X.L., R.W., D.K., and Y.L. tested the functions; Z. Liang, L.X., Y.Y., and L.Z. prepared the figures. All authors read and approved the final manuscript.

DECLARATION OF INTERESTS

No conflict of interests declared.

Received: November 25, 2021
Revised: March 11, 2022
Accepted: April 28, 2022
Published: May 5, 2022

REFERENCES

Birchler, J.A. (2016). Plant science: Hybrid vigour characterized. Nature, 620–621, In press. https://doi.org/10.1038/nature19433.

Jensen, S.E., Charles, J.R., Muleta, K., Bradbury, P.J., Casstevens, T., Deshpande, S.P., Gore, M.A., Gupta, R., Ilut, D.C., Johnson, L., et al. (2020). A sorghum practical haplotype graph facilitates genome-wide imputation and cost-effective genomic prediction. Plant Genome **13**, e20009. https://doi.org/10.1002/tpg2.20009.

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., Xiang, Z., Song, W., Ying, K., Zhang, M., et al. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. Nat. Genet. **42**:1027–1030. https://doi.org/10.1038/ng.684.

Li, C., Li, Y., Bradbury, P.J., Wu, X., Shi, Y., Song, Y., Zhang, D., Rodgers-Melnick, E., Buckler, E.S., Zhang, Z., et al. (2015). Construction of high-quality recombination maps with low-coverage genomic sequencing for joint linkage analysis in maize. BMC Biol. **13**:78. https://doi.org/10.1186/s12915-015-0187-4.

Masuka, B., Atlin, G.N., Olsen, M., Magorokosho, C., Labuschagne, M., Crossa, J., Banziger, M., Pixley, K.V., Vivek, B.S., Biljon, A., et al. (2017a). Gains in maize genetic improvement in Eastern and Southern Africa : I. CIMMYT hybrid breeding pipeline. Crop Sci. **57**:168–179. https://doi.org/10.2135/cropsci2016.05.0343.

Masuka, B., Magorokosho, C., Olsen, M., Atlin, G.N., Bänziger, M., Pixley, K.V., Vivek, B.S., Labuschagne, M., Matemba-Mutasa, R., Burgueño, J., et al. (2017b). Gains in maize genetic improvement in Eastern and Southern Africa: II. CIMMYT open-pollinated variety breeding pipeline. Crop Sci. **57**:180–191. https://doi.org/10.2135/cropsci2016.05.0408.

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. https://doi.org/10.1016/j.molp.2017.06.008.

Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y., and Buckler, E.S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. U S A. **112**:3823–3828. https://doi.org/10.1073/pnas.1413864112.

Tian, H., Yang, Y., Yi, H., Xu, L., He, H., Fan, Y., Wang, L., Ge, J., Liu, Y., Wang, F., et al. (2021). New resources for genetic studies in maize (Zea mays L.): A genome-wide Maize6H-60K single nucleotide polymorphism array and its application. Plant J. **105**:1113–1122. https://doi.org/10.1111/tpj.15089.

Torkamaneh, D., Laroche, J., Valliyodan, B., O'Donoughue, L., Cober, E., Rajcan, I., Vilela Abdelnoor, R., Sreedasyam, A., Schmutz, J., Nguyen, H.T.P., et al. (2021). Soybean (Glycine max) Haplotype Map (GmHapMap): a universal resource for soybean translational and functional genomics. Plant Biotechnol. J. **19**:324–334. https://doi.org/10.1111/pbi.13466.

Wang, F., Zhao, J., Sun, S., Zhi, J., Yi, H., Tian, H., and Yang, G. (2011a). Maize Varieties DNA Fingerprint Technology-The Research and Application of SSR Marker (China's Agricultural Science and Technology Press), pp. 140–142.

Wang, F.G., Tian, H.L., Zhao, J.R., Yi, H.M., Wang, L., and Song, W. (2011b). Development and characterization of a core set of SSR markers for fingerprinting analysis of Chinese maize varieties. Maydica **56**:7–17.

Wang, F., Yi, H., Zhao, J., Liu, P., Zhang, X., Tian, H., and Du, Y. (2014). Protocol for the Identification of Maize Varieties-SSR Marker Method (Agricultural Industry Standards of the People's Republic of China NY/T), pp. 1432–2014.

Wang, F., Yang, Y., Yi, H., Zhao, J., and Hou, Z. (2017). Construction of an SSR-based standard fingerprint database for corn variety authorized in China. Sci. Agric. Sin. **50**:1–14.

Yang, Y., Tian, H., Wang, R., Wang, L., Yi, H., Liu, Y., Xu, L., Fan, Y., Zhao, J., and Wang, F. (2021). Variety discrimination power: an appraisal index for loci combination screening applied to plant variety discrimination. Front. Plant Sci. **12**:566796. https://doi.org/10.3389/fpls.2021.566796.