OXFORD

Briefings in Bioinformatics, 23(1), 2022, 1–13

https://doi.org/10.1093/bib/bbab447
Problem Solving Protocol

Genome optimization via virtual simulation to accelerate maize hybrid breeding

Qian Cheng[†], Shuqing Jiang[†], Feng Xu, Qian Wang, Yingjie Xiao, Ruyang Zhang, Jiuran Zhao, Jianbing Yan, Chuang Ma and Xiangfeng Wang

Corresponding authors: Xiangfeng Wang, Sanya Institute of China Agricultural University, Sanya 572000, China. National Maize Improvement Center of China Agricultural University, Beijing 100094, China. Tel.: +861062733399; Fax: 861062733404. E-mail: xwang@cau.edu.cn; Chuang Ma, State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, Yangling 712100, China. E-mail: cma@nwafu.edu.cn

[†]These authors contribute equally to this work.

Abstract

The employment of doubled-haploid (DH) technology in maize has vastly accelerated the efficiency of developing inbred lines. The selection of superior lines has to rely on genotypes with genomic selection (GS) model, rather than phenotypes due to the high expense of field phenotyping. In this work, we implemented 'genome optimization via virtual simulation (GOVS)' using the genotype and phenotype data of 1404 maize lines and their F_1 progeny. GOVS simulates a virtual genome encompassing the most abundant 'optimal genotypes' or 'advantageous alleles' in a genetic pool. Such a virtually optimized genome, although can never be developed in reality, may help plot the optimal route to direct breeding decisions. GOVS assists in the selection of superior lines based on the genomic fragments that a line contributes to the simulated genome. The assumption is that the more fragments of optimal genotypes a line contributes to the assembly, the higher the likelihood of the line favored in the F_1 phenotype, e.g. grain yield. Compared to traditional GS method, GOVS-assisted selection may avoid using an arbitrary threshold for the predicted F_1 yield to assist selection. Additionally, the selected lines contributed complementary sets of advantageous alleles to the virtual genome. This feature facilitates plotting the optimal

Qian Cheng is a PhD candidate of bioinformatics at the State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, China. His research focuses on developing novel algorithms and tools for genomic selection-assisted breeding in crops.
Shuqin Jiang is a postdoctoral researcher of biostatistics at the National Maize Improvement Center of China Agricultural University, Beijing, China. She develops software for phenotype analysis in crops and novel algorithms for genomic selection-assisted breeding in crops.
Feng Xu is a PhD candidate of bioinformatics at the National Maize Improvement Center of China Agricultural University, Beijing, China. His research focuses on developing software for visualizing large-scale genomic data and for analyzing third-generation isoform sequencing data.
Qian Wang is a postdoctoral researcher of bioinformatics at the National Maize Improvement Center of China Agricultural University, Beijing, China. Her research focuses on developing algorithms for analyzing gene co-expression network in maize.
Yingjie Xiao is an Associate Professor of the National Key Laboratory of Crop Genetic Improvement, College of Plant Sciences and Technology at Huazhong Agricultural University, Wuhan, China. His research focuses on quantitative genetics and genomic evolution in maize.
Ruyang Zhang is an Associate Principal Investigator of the Maize Research Center at Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His research focuses on population genetics and genetic breeding in maize.
JiuRan Zhao is a Principal Investigator of the Maize Research Center at Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. His team focuses on genetic breeding and molecular breeding of maize.
Jianbing Yan is a Professor of the National Key Laboratory of Crop Genetic Improvement, College of Plant Sciences and Technology at Huazhong Agricultural University, Wuhan, China. His team focuses on maize genomics and genetics, and develops molecular tools for maize breeding.
Chuang Ma is a Professor of Bioinformatics at the State Key Laboratory of Crop Stress Biology for Arid Areas, Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, China. His team develops data-mining tools for gene discovery and genomic selection models in crops.
Xiangfeng Wang is a Principal Investigator of plant breeding at the Sanya Institute of China Agricultural University, Hainan, China. He is also a member of the National Maize Improvement Center of China Agricultural University, Beijing, China. His team utilizes machine learning techniques to develop data-driven decision-making models to assist genomically designed breeding in plants.

Submitted: 5 August 2021; **Received (in revised form):** 24 September 2021

route for DH production, whereby the fewest lines and F_1 combinations are needed to pyramid a maximum number of advantageous alleles in the new DH lines. In summary, incorporation of DH production, GS and genome optimization will ultimately improve genomically designed breeding in maize.

Short abstract: Doubled-haploid (DH) technology has been widely applied in maize breeding industry, as it greatly shortens the period of developing homozygous inbred lines via bypassing several rounds of self-crossing. The current challenge is how to efficiently screen the large volume of inbred lines based on genotypes. We present the toolbox of genome optimization via virtual simulation (GOVS), which complements the traditional genomic selection model. GOVS simulates a virtual genome encompassing the most abundant 'optimal genotypes' in a breeding population, and then assists in selection of superior lines based on the genomic fragments that a line contributes to the simulated genome. Availability of GOVS (https://govs-pack.github.io/) to the public may ultimately facilitate genomically designed breeding in maize.

Key words: plant breeding; *Zea mays*; genomic selection; genotype-to-phenotype prediction; computational simulation; doubled haploid

Introduction

Crop improvement by selective breeding essentially depends on the selection of advantageous alleles that express phenotypes meeting the agricultural needs [1–3]. Pyramiding of the advantageous alleles is usually implemented during population development, in which breeding materials are hybridized with each other so that the genomic fragments of the parental lines are reshuffled via DNA recombination [4]. Thereby, the genotypes and phenotypes of new lines are more diversified compared to the original lines, offering genes for subsequent selection [5, 6]. Artificial selection of the desired phenotypes increases the frequencies of rare but advantageous alleles or allele combinations [7]. After multiple generations of selection, the maximum genetic gain is achieved when all possible advantageous alleles are pyramided [8, 9]. The development of inbred lines with homozygous genotypes is especially important for maize F_1-hybrid breeding, which mainly utilizes heterosis [10, 11].

Maize (*Zea mays*) is a worldwide cultivated staple crop as an important source not only for human nutrition, but also for livestock feed and biofuels. Sustainable growth of maize yield is critical to ensure global food security. However, dramatic climate change has become unpredictable, and increased frequency of extreme weather and natural disaster has greatly impacted maize production. The seed industry expects innovative informatics and biotechnology applied in maize breeding to face these global challenges. Maize is one of the most successful crops utilizing heterosis to cultivate F_1 hybrids exhibiting superior hybrid vigor in terms of grain yield, biomass and stress resistance [12]. In a conventional program of maize breeding, development of parental inbred lines is the major step that impedes breeding efficiency, as 6–8 generations of self-crossing are required to create pure inbred lines [13]. Owing to the speedy development of the doubled-haploid (DH) technology in maize, breeding cycles have greatly accelerated and the production of the DH lines ready for hybridization, now, costs only a single year [13, 14]. Thus, a mid-size breeding company may produce tens of thousands of inbred lines per year, and the challenge has shifted from developing lines to selecting lines. Although the large volume of DH lines offers a great opportunity of selecting superior lines pyramided with abundant advantageous alleles, phenotype-based screening of all the lines through field trials is time consuming, laborious and costly. To make line selection more affordable, one feasible solution is to apply genomic selection (GS) that utilizes various genotype-to-phenotype (G2P) predictive models to screen DH lines [15–22]. Then, only lines predicted with high trait performance, usually 5–10% of the total lines, are promoted to field trials.

In the modern seed industry, a standard pipeline of maize hybrid breeding consists of three major sections, as illustrated in Figure 1. The first section is developing a novel genetic pool or a base population using a panel of elite lines as founders, namely, population development. The goal of population development is to generate superior lines pyramided with the maximum advantageous alleles inherited from diverse founder parents [1, 23]. At the same time, the deleterious alleles should be removed as much as possible via the artificial selection of desired phenotypes or adaptive selection using specific environments [24, 25]. The second section is line selection, in which the candidate lines may be used as the maternal pool to cross with one or two paternal testers to examine the phenotypes of their F_1 progeny. By test-crossing, a panel of F_1 is generated and the ones exhibiting superior hybrid performances are identified. In the third section, the selected superior lines are crossed with a panel of paternal lines to obtain F_1 combinations, and the ones expressing the desired traits are selected. This helps in the identification of the optimal paternal lines suited to be paired with the maternal lines to produce the progeny with the desired traits. With the incorporation of GS based on various genotype-to-phenotype (G2P) predictive models into the above pipeline, breeding efficiency is expected to further accelerate along with a significant reduction in expenses of conducting field trials and phenotyping the F_1 hybrids.

Genotype-assisted selection of DH lines has become more affordable nowadays, than it was in the past, due to the invention of multiple cost-effective genotyping technologies, among which genotyping by target sequencing (GBTS) is the most promising technology suitable for plant breeding [26]. GBTS adopted two main strategies—the first one is based on multiplexing PCR (GenoPlexs) that may capture up to 5000 SNPs based on a panel of predesigned primers to amplify the DNA fragments containing the target SNPs [3, 27]; the second one uses a liquid chip to capture target DNAs in solution (GenoBaits) that may cover up to 50 000 SNPs [3, 26]. Considering the balance of genotyping cost, marker universality, and prediction accuracy, a panel of 3000–5000 target SNPs is the optimal size for both genetic analysis and G2P prediction in maize [26]. Additionally, the coverage of 500 Kb to 1 Mb per SNP in the maize genome is sufficient to trace back the exchanged fragments among the DH lines generated from each one of the F_1 or F_2 hybrids [28, 29]. This is because the average rate of recombination events incurred in each DH line is only half of that of the crossovers incurred in an F_2 hybrid [30]. Because of the same reason, genotypic and phenotypic variants in a DH population may not be as rich as those in a population developed by several generations of self-crossings. Thus, screening of DH lines should not merely rely on phenotype

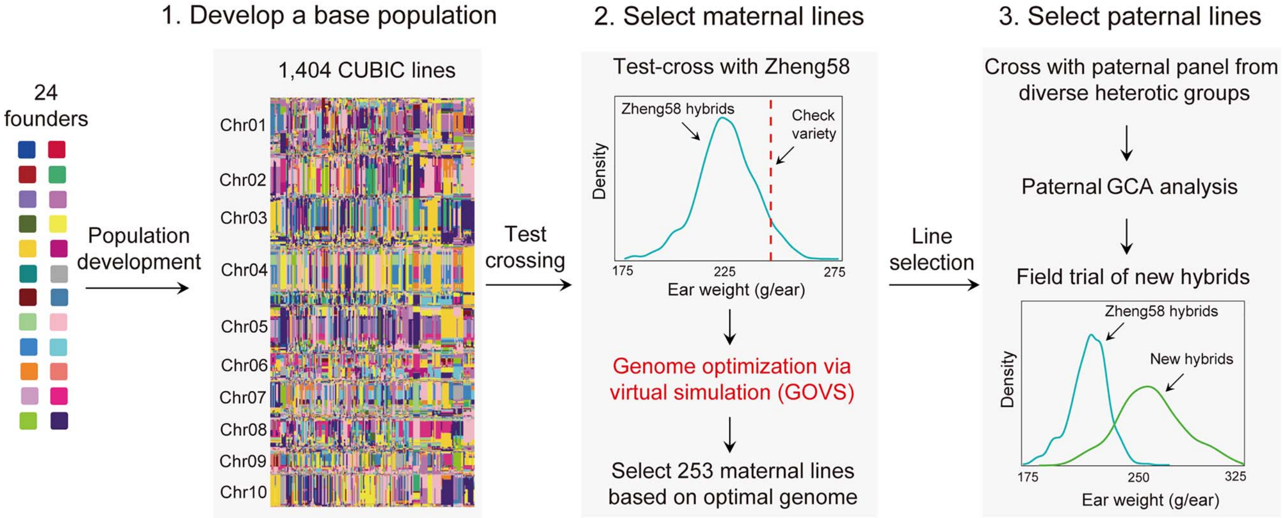

Figure 1. The model of GOVS-assisted genomically designed breeding. The model consists of three major steps: first, novel base population called the CUBIC population was developed from 24 founder lines; second, the 1404 F_1 hybrids were generated by crossing the 1404 CUBIC lines with the tester line Zheng58, and GOVS was applied to select superior lines; third, the selected superior lines were crossed with a panel of paternal lines to determine the optimal F_1 combinations, finally resulting in a significant improvement in grain yield compared to the original 24 founders.

prediction but should also include the estimation of recombination rate, tracing of the chromosomal crossover events and determination of the optimal haplotypes of trait-associated fragments. Using a combination of filtering criteria, the large amount of DH lines may be narrowed down to a set of superior lines affordable to perform field trial to assess the phenotypes of their F_1 progeny.

To meet these demands, a novel computational model other than the G2P prediction is needed for efficient, large-scale screening of DH lines. The model should be able to perform the 'genome optimization' that we previously proposed incorporating DH production, G2P prediction and genome optimization to direct genomically designed breeding [8]. GOVS presented in the current work refers to the employment of computational algorithms to generate a virtually optimized genome by assembling the genomic fragments featuring the optimal genotypes of a desired trait. The optimal genome is presumed to express optimal phenotypes. Although the so-called 'optimal genome' may never be developed in reality, it may function as an assembly of the favorable genotypes of all trait-associated genes to direct breeding decisions in terms of selecting superior lines and planning the next cycles of DH production. It is worth noting that GOVS is not a substitute for, but a complement to GS, and has multiple advantages over GS. Firstly, as DH lines originate from the F_1 or F_2 hybrids of the two parental lines, GOVS adopts the identity-by-descent (IBD) analysis to infer the recombination events based on the haplotypes of the SNPs within an exchanged fragment [8, 23]; then, the optimal genotype of each fragment contributing to a target trait of F_1 progeny is inferred. Secondly, GOVS assists in the selection of superior lines based on the genomic fragments that a line contributes to the virtually optimized genome, rather than merely basing it on the predicted phenotypes. Thirdly, GOVS plots the optimal route to pyramid the maximum advantageous alleles from minimum number of lines and times of crossing, since the superior lines selected by GOVS contribute complementary sets of advantageous alleles in known proportions to the optimal genome. Due to these advantages, the genetic gain may be rapid using the fewest breeding materials and the fewest events of hybridization, thus, greatly accelerating the breeding efficiency. GOVS is publically available at https://govs-pack.github.io/.

Methods

Construction of the *bin* map for the CUBIC population

The development of the complete-diallel plus unbalanced breeding-derived inter-cross (CUBIC) population from the 24 founders, SNP calling from the resequencing of the 1458 inbred lines, and the inference of the initial 24 720 *bins* via IBD analysis were previously described [23, 31]. The data were obtained from http://zeamap.hzau.edu.cn/ftp/99_MaizegoResources/01_CUBIC_related/. The original 4.5 million SNPs were highly redundant and may have decreased the computing efficiency. We, then, screened the SNPs in five steps—first, the function of 'indep' in PLINK [32] was applied and a total of 45 576 tag SNPs were identified; second, the SNPs located in the highly repetitive regions of the maize genome according to Tarailo-Graovac and Chen [33] were removed; third, the SNPs with minor allele frequency (MAF) between 0.15 and 0.35 were retained; fourth, 50-nt flanking sequence on each side of an SNP (101-nt in total) were aligned to the reference genomes of multiple elite inbred lines that are available on MaizeGDB [34], and only the SNPs residing in highly conserved regions were retained. As a result, the final set obtained after stringent filtration contained 4903 SNPs, which may be used as a universal panel of SNPs for the GenoPlexs platform to reduce the expense of genotyping. Based on the 4903 SNPs, the phylogenetic tree of the 24 founders was constructed using the unweighted neighbor-joining method using MEGA software [35]; the population structure of the 900 F_1 hybrids (30 maternal × 30 paternal lines) was analyzed using principal component analysis (PCA) function of the 'sklearn' package [36] of Python (https://www.python.org/). The 4903 SNPs were further mapped to the original 24 720 unique bins identified using IBD analysis, resulting in the identification of 3515 bins that represented the recombinant fragments in the maize genome.

Collection and processing of phenotype data of the F_1 hybrids

The three main agronomic traits, namely the days to tasseling (DTT), plant height (PH) and ear weight (EW), were phenotyped for the 1404 F_1 progeny of Zheng58, and the 900 F_1 combinations at five different locations, which were Yushu (Jilin Province, N43°42′, E125°18′), Shenyang (Liaoning Province, N42°03′, E123°33′), Beijing (N40°10′, E116°21′), Baoding (Hebei Province, N38°39′, E115°51) and Xinxiang (Henan Province, N35°27′, E114°01′) in Northern China. The phenotypes were collected from unreplicated trial with checks in which 20 individual plants of each F_1 hybrid were planted in a row. The check variety of ZhengDan958 resulting from the crossing of Chang7-2 and Zheng58 was planted in every 50th rows, which was used for the correction of spatial heterogeneity in the field. The average phenotype of five well-pollinated ears in the middle of each row was calculated to avoid the edge effect of the plot. To remove the environmental influence, the mixed linear model from the R package 'lme4' [37] was used to estimate the best linear unbiased prediction (BLUP) value for each F_1 hybrid summarized from the five locations using the following the model: $y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where y_{ij} is the observed value of the ith line in the jth environment, μ is the overall average, α_i is the effect of ith line, β_j is the effect of jth environment and ε_{ij} is the random error. Here, both α_i and β_j are assumed to be random effects, in which $\alpha_i \sim N\left(0, \sigma_\alpha^2\right)$ and $\beta_j \sim N\left(0, \sigma_\beta^2\right)$. The BLUP values for each phenotype were then used for subsequent analyses. The general combining ability (GCA) of the 30 paternal lines was calculated using the BLUP values of the 900 F_1 hybrids using the formula: $GCA_i = \bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet}$, where $\bar{y}_{i.}$ is the mean phenotype of the F_1 progeny resulting from crossing with the paternal line i, and $\bar{y}_{\bullet\bullet}$ is the mean phenotype of all the 900 F_1 progeny. The field trial of the 80 selected F_1 combinations was performed in Sanya in the Hainan Province. Seven individuals of each F_1 combination were planted in 3 rows, and, thus, 21 phenotype data points were collected for each F_1 combination. Three commercial cultivars, namely ZhengDan958, XianYu335, and JingKe968 were used as check varieties, which were planted in every 15 rows and 105 repeats to calibrate the bias in the same plot of the 80 F_1 combinations.

Detection of superior lines by the G2P and GOVS methods

We adopted two ways to validate the availability of GOVS-assisted selection and to compare the precision between the G2P and GOVS methods. In the first way, the 1404 lines were evenly divided into training and testing sets. Superior lines were defined as the maternal lines that were crossed with the paternal tester Zheng58 and generated F_1 EWs surpassing the EW (243.98 g/ear) of the check (CK) ZhengDan958 variety. Thus, detection rate of superior lines was used to represent the precision of selection. To minimize the bias potentially caused by different enrichments of superior lines in training and testing sets, the 1404 lines were randomly split to 702 training and 702 testing samples for 20 times, and the number of superior lines was counted at each time. Then, the time of sample division generating the number of superior lines closest to the average number of the 20 times was used as the final division of 702 training and 702 testing samples. To test the precision of G2P, the F_1 EWs and genotypes of the 702 training lines were used to train the rrBLUP model, followed by predicting the F_1 EWs of testing lines based on their genotypes. In the second way, we adopted a repeated holdout method of cross-validation with a different way of sample division to compare the precision of G2P and GOVS-assisted selection. In this way, the 1404 lines were randomly divided to training and testing sets with a ratio of 2:1 (936 versus 468). Then, the F_1 EWs of the 468 testing lines crossed with Zheng58 were predicted using the rrBLUP model trained with the F_1 EWs and genotypes of the 936 training lines, followed by applying GOVS on predicted F_1 EWs of testing lines to build the virtually optimized genome. Selection of superior lines assisted GOVS was based on the number of contributed *bins*. To compare the detection rates of superior lines by the two methods, we counted the numbers of actual superior lines among the top 5% (23), 10% (47) and 15% (70) lines out of the 468 testing lines sorted based on the predicted F_1 EWs and contributed *bins* by the G2P and GOVS methods, respectively. The above process was repeated for 10 times to generate 10 detection rates for each method. Then, a paired t-test was performed to test the significance of the difference between the 10 detection rates by G2P-assisted and GOVS-assisted selection of superior lines.

Implementation of GOVS in R

We developed an R package—GOVS (https://govs-pack.github.io/), which stands for Genome Optimization via Virtual Simulation—to assist genomically designed breeding for maize, including three modules for virtual simulation of the optimal genome and one module for G2P prediction. Four input files are required for GOVS, including phenotype and genotype files of a population, and two *bin* files containing the recombinant hotspots of the population. One of the two *bin* files should contain the genomic fragments identified using the IBD algorithm to trace the origins of the *bins* back to the parental lines [23, 38], and the other one is an annotation file of the *bins* used to indicate the recombination hotspots. The first module 'genomeOptimization' simulates the optimal genome statistically using the 'optimal genotypes' associated with the 'optimal phenotypes (i.e., grain yield of F_1 hybrids)' using LSMEANS on the haplotype of each *bin* instead of the genotype of individual SNP. Based on the results from the 'genomeOptimization' module, the optimal genotypes of the SNPs mapped to the *bins* were extracted from the population, and were consecutively assembled to simulate an optimal genome using the second module 'extractGenome.' It has to be noted that, the optimal genome only represents a theoretical assembly of advantageous alleles but cannot be developed in reality, due to genetic drag, meaning that linkage of certain advantageous and deleterious alleles may never be disrupted [39]. The resulting optimal genome and the associated information is then made available to the third module 'statDesign' to perform statistical analysis on the lines contributing genomic fragments. The lines are ranked based on their contribution of genomic fragments to the assembly of the optimal genome. The top-ranked lines are selected on priority for further crossing with more paternal lines apart from the tester line. The GOVS package also includes two popular G2P models, ridge regression best linear unbiased prediction (rrBLUP) and genomic best linear unbiased prediction (gBLUP) [40], which can be called using the functions of 'SNPrrBLUP' and 'GBLUP,' respectively. Additionally, GOVS supplies a convenient toolkit to convert the character format of genotype data to numeric codes of 0, 1 and 2 representing the homozygous major alleles, heterozygous alleles and homozygous minor alleles, respectively.

Results

The rationale of genome optimization via virtual simulation

To test genomically designed breeding using the proposed model of GOVS, we obtained the previously published dataset of the 1404 CUBIC lines and considered it as the maternal pool and their F_1 progeny resulting from crossing with Zheng58—a paternal tester commonly used in China [41]. The details of the CUBIC population and the related data on the F_1 progeny of the CUBIC lines and Zheng58 were presented elsewhere [23, 42]. The 24 founders used to develop the CUBIC population were selected from three main heterotic groups widely used in China, namely the SiPingTou (SPT), LvDaHongGu (LDHG) and ZI330 (Z330) (Figure S1). Therefore, the CUBIC population may be regarded as a pool of advantageous alleles adapted to the different local environments in China. In addition, ZhengDan958, which is the F_1 progeny of Zheng58 and Chang7-2 (one of the 24 founders), was used as the check variety in the current work [41]. The EW of ZhengDan958 was used as the threshold to determine the superior F_1 hybrids, which is a stringent criterion commonly used in real-world breeding practices [41, 43]. The yield of ZhengDan958 was used for the evaluation of the precision of prediction by identifying the F_1s with observed EWs surpassing the threshold among the highly ranked F_1s predicted by GS or GOVS. The crossings of the 24 founders with Zheng58 showed only four F_1 combinations surpassing the threshold of EW of ZhengDan958 (Figure S2). More F_1 combinations with the potential of being new cultivars surpassing the threshold of the EW of ZhengDan958 were expected since the CUBIC population contained superior lines pyramided with more advantageous alleles than the 24 founders. Thus, the purpose of this work was to identify these lines via the proposed model of GOVS.

The availability of the genotypes of the 24 founders and 1404 CUBIC lines allowed us to computationally trace the inheritance of a designated genomic fragment back to the founders. Using phenotypes, the genetic effect of the genomic fragment may be further quantitatively determined in terms of its contribution to its F_1's yield. In the current article, the genomic fragments are defined as '*bins,*' which are segmented regions of a chromosome, resulting from chromosomal recombination during population development. The method of *bin* inference from the CUBIC population was described in Liu *et al.* [23]. Using a genome-wide scan, the positive and negative associations between *bins* and F_1 yields were inferred statistically. The *bins* contributing positive effects to F_1 yield were defined as advantageous alleles with optimal genotypes and those exhibiting negative effects were defined as deleterious alleles. Subsequently, the haplotype of a *bin* exhibiting the highest positive association with EW was extracted from the corresponding F_1 hybrid. The optimal genotypes of all the *bins* were assembled to simulate a virtually optimized F_1 genome. The simulated genome encompassed all the advantageous alleles favoring F_1 yield, thus, theoretically expressing the optimal phenotype. It has to be noted that the simulated genome may never be developed in reality but may represent the maximum potential of genetic gain that a genetic pool of advantageous alleles can possess.

The algorithm of GOVS used to assemble a virtually optimized genome

We used the genotype and phenotype data of the 1404 F_1 progeny of the CUBIC lines crossed with Zheng58 tester to implement GOVS. The genotype data included 4903 SNPs selected from the 4.5 million SNPs detected from the previously published whole-genome resequencing of the 1404 CUBIC lines, 24 founder lines and 30 paternal testers [23, 42]. The 4903 SNPs represented a variation atlas of the 3515 *bins* in the maize genome with the least redundancy and ensured an optimal balance of genotyping expense and variation coverage. Detailed methods of SNP filtration and the construction of the *bin* map are described in the Methods. The phenotype data included DTT, PH and EW representing the three main agronomic traits of flowering time, plant stature and grain yield, respectively, at the three important developmental stages.

The algorithm of GOVS is illustrated in Figure 2. A *bin* map dividing the maize genome into 3515 segments was first constructed using IBD analysis based on the genotypes of the 1404 CUBIC lines and 24 founder lines (Figure 2a). The *bin* map essentially represented the hotspots of chromosomal recombination incurred during the development of the CUBIC population. Each *bin* of a line was coded by one of 24 colors to indicate its origin (one of the 24 founders). This means that the origin of a *bin* contributing positive effect to the F_1's phenotype, such as the maximizing EW used as an example for the current work, may be traced back to determine the corresponding founder as well as the optimal haplotype associated with the *bin*. This procedure is illustrated using the *Bin*-1758 located at Chromosome 5: 3 670 776–3 770 999 in Figure 2b. There were 22 groups of haplotypes associated with *Bin*-1758, indicating that 22 founders contributed fragments to the CUBIC population, except TY4 and QI1261. One possible reason is that the genotypes of *Bin*-1758 in TY4 and QI1261 might be deleterious to the fitness, and thus, were eliminated via artificial selection. Using the 22 *bin* haplotypes, the 1404 lines were then categorized into 22 groups, which were subsequently used to infer the optimal genotypes positively affecting the EWs of F_1s. The inference was based on the results of least-squares means (LSMEANS) across the 22 haplotypic groups from a general linear model. The 22 groups were then sorted from high to low based on the LSMEANS of the corresponding group. The genotypes of the SNPs in *Bin*-1758 from the line exhibiting the highest EW in F_1 in the topmost group were extracted and then placed at the corresponding *bin* location in the maize genome (Figure 2c). After genome-wide scanning using LSMEANS, the optimal genotypes of the 3515 *bins* were determined and assembled as a virtually optimized genome according to their locations in the maize genome. Similarly, GOVS may also be used to select the genotypes of *Bin*-1758 with moderate and the least effects on the F_1's EW. Therefore, three types of virtual genomes may be simulated with optimal, moderate and poor genotypes. Using the genotypes and observed phenotypes of the 1404 F_1 hybrids as a training set, the optimal, moderate and poor phenotypes of the three virtual genomes can be predicted using the G2P model (Figure 2d).

To test the computing efficiency of the GOVS algorithm, we performed a series of benchmark tests by simulating different numbers of samples, SNPs and *bins* on a server with configuration of four 16-core CPUs (E7-4850 v4 @ 2.10GHz) and 1 TB memory (Table S1). When genotype data were fixed by 1000 SNPs and 1000 *bins*, GOVS spent only 35 min and consumed 22.5 GB memory to finish computing on 50 000 samples. When 1000 *bins* and 1000 samples were fixed, GOVS spent 4 and 5 min on 1000 SNPs and 100 000 SNPs, respectively. It indicates that GOVS is not sensitive to the number of SNPs, because the SNP data are only used when assembling the virtual genome with the optimal genotypes of SNPs. When 1000 samples and 30 000 SNPs were fixed, CPU times and memory usage gradually increased along with the increased number of *bins* from 4 min and 1.2 GB on 1000 *bins* to 155 min and 16.2 on 30 000 *bins*. Thus, the number of *bins* is the limiting factor of computing efficiency, as GOVS

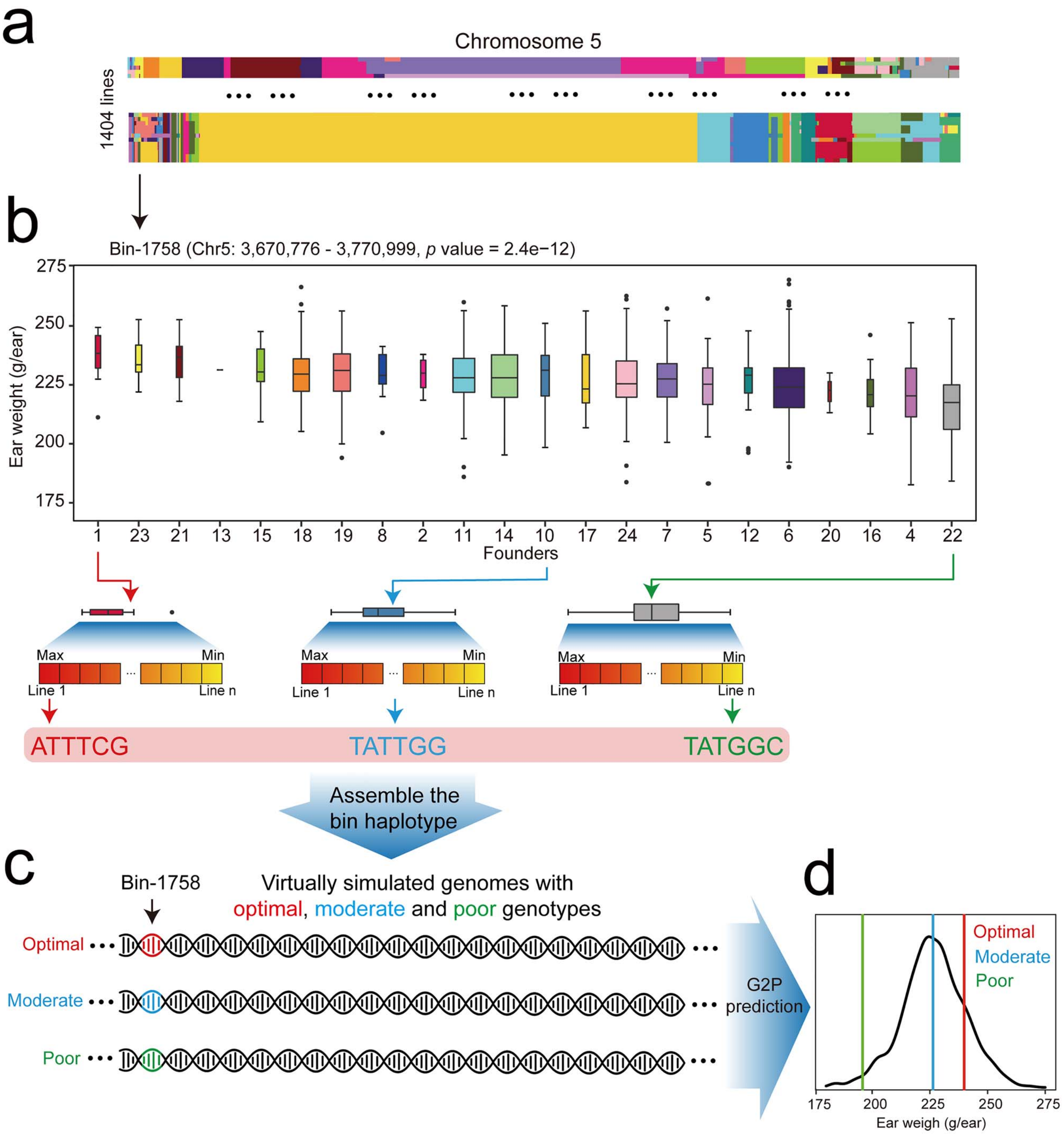

Figure 2. Schematic illustration of the algorithm of GOVS. (a) The *bin* map of the 1404 CUBIC lines exemplified by chromosome 5. Each of the 24 founders was represented by a color and each bin was then coded by one of the 24 colors. (b) The identification of the optimal haplotype of *Bin*-1758 associated with a higher yield in the F_1 hybrids. The Bin-1758 possessed 22 types of haplotypes that could be traced back to the 22 founder lines. Thus, the 1404 lines were categorized into 22 groups, which were sorted by the average EW of their F_1 progeny. The topmost group showed the highest average EW compared to the rest of the groups. The genotypes of the SNPs mapped to *Bin*-1758 of the line with the highest EW in this group were defined as the optimal genotypes. (c) The optimal genotype of *Bin*-1758 was extracted and assembled with the optimal genotypes of other bins to generate a simulated genome. This procedure may also be used to generate moderate and poor genomes by extracting the moderate and poor genotypes of the *Bin*-1758. (d) G2P prediction was then used to predict the phenotypes of the simulated optimal, moderate and poor genomes using the actual genotypes and phenotypes of the 1404 CUBIC lines to train the model.

bin-by-*bin* scans the genome to infer the genetic contribution of a *bin* towards the trait.

G2P-assisted selection based on predicted phenotypes

One of the functions of GOVS is to assist in selecting superior lines based on the *bins* contributed by a line, rather than merely relying on the predicted phenotype derived from a G2P model. We first tested the precision of G2P-assisted selection of superior lines whose F_1 progenies exhibited high EWs using the rrBLUP model [40]. The 1404 F_1s were divided into training and test datasets, each containing 702 samples (Methods). Using the EW (243.98 g/ear) of ZhengDan958 as the threshold, 133 (9.47%), 62

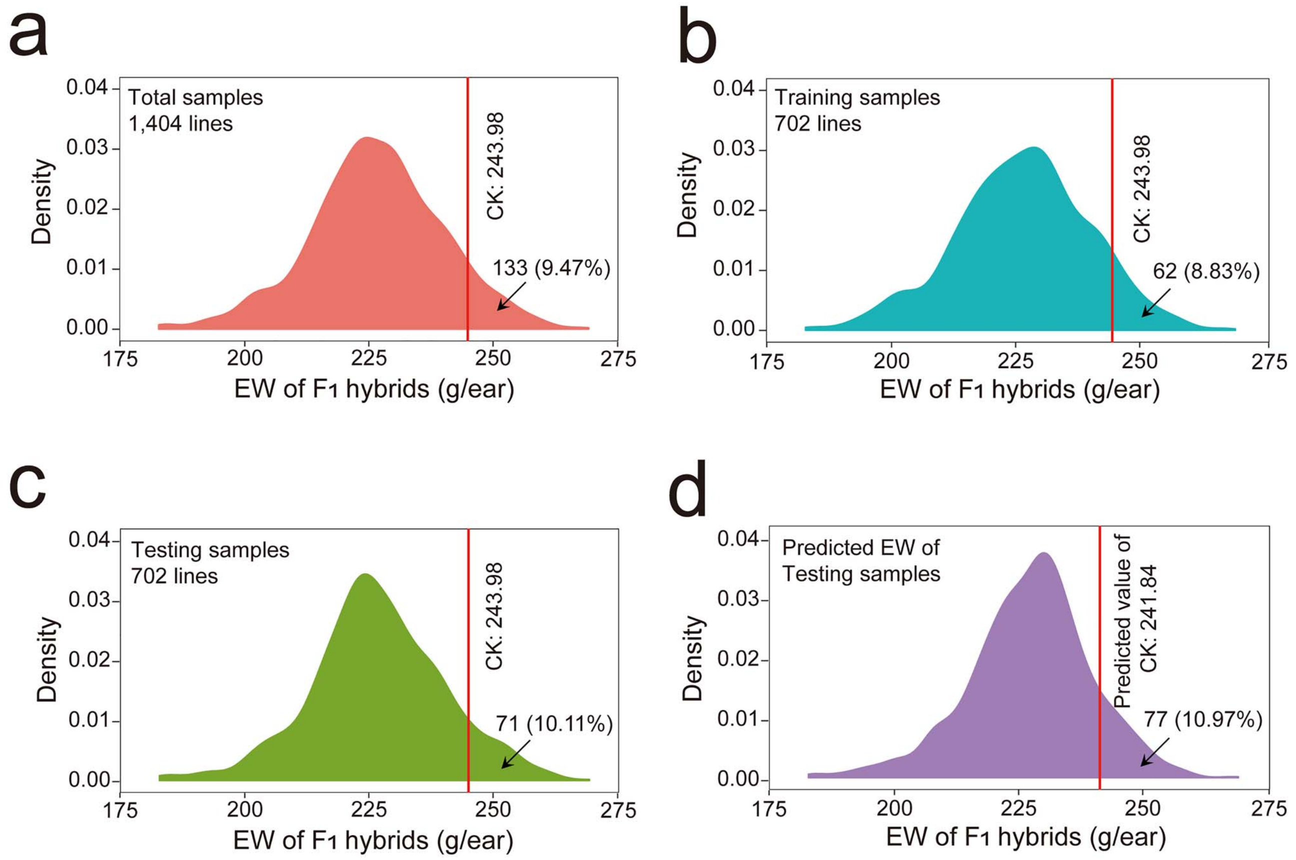

Figure 3. Partition of training and testing datasets. (a) The distribution of the observed EW of the total 1404 F_1 hybrids of which 9.47% F_1s surpassed the threshold of the EW of ZhengDan958. (b) The distribution of the observed EW of the 702 F_1 hybrids from the training set of which 8.83% F_1s surpassed the threshold of EW of ZhengDan958. (c) The distribution of the observed EW of the 702 F_1 hybrids from the test set of which 10.11% F_1s surpassed the threshold of EW of ZhengDan958. (d) The distribution of the predicted EW of the 702 F_1 hybrids from the test set of which 10.97% F_1s surpassed the threshold of EW of ZhengDan958.

(8.83%) and 71 (10.11%) F_1s out of the total 1404 samples, 702 training samples and 702 test samples, respectively, surpassed the threshold (Figure 3a-c). These percentages indicated that the average selection rate of the superior lines from the CUBIC population was ~10%. We then predicted the EWs of the 702 F_1s in the test dataset using the genotype of ZhengDan958 as a spike-in sample for determining the threshold. Out of the 702 test samples, the predicted EWs of 77 (10.97%) samples surpassed the threshold (Figure 3d). The overall precision was 0.488, which was the Pearson's correlation coefficient (*r*) between the predicted and observed EWs (Figure S3a). However, only 20 samples were common when comparing the top 77 and top 71 samples with the predicted and observed EWs, respectively (Figure S3b). Therefore, the actual selection precision was only 28.17% (20/71), indicating that G2P-assisted selection may enrich the superior lines by about ~2.8 times compared to a random selection of 70 out of the 702 test samples.

GOVS-assisted selection based on contributed *bins*

GOVS employs an alternative strategy to assist in line selection, using prediction results from G2P, and does not merely rely on an arbitrary threshold to select top predictions. Simply speaking, GOVS determines whether a line is superior based on the number of *bins* it contributes to the simulated genome, by assuming that the number of *bins* with optimal genotypes of a line is positively correlated with the abundance of advantageous alleles it possesses; thus, the F_1 progenies of the lines contributing a high proportion of *bins* have a high likelihood of expressing optimal phenotypes. To test this assumption, we first applied GOVS to the 702 training samples. Statistics on the *bin* composition of the simulated genome showed that the top 3 lines contributed complementary sets of *bins* or genomic fragments accounting for 8.43%, 8.21% and 5.74% of the total *bins*, respectively. The EWs of their F_1 progeny were also ranked as the top three among the 702 training samples. When the number of top lines ranked by the contribution of *bins* increased to 12, 38 and 170 samples, the cumulative percentage of *bins* achieved 10%, 80% and 100% of the simulated genome, respectively. When the selection threshold was 10% (70/702), the cumulative percentage achieved was 91.85% (Figure 4a). The examination of the haplotypes of the *bins* that the 170 superior lines contributed allowed us to trace their origins to the founders. The top 8 founders included all four lines from Z330 and four from SPT groups, contributing *bins* accounting for 65.84% of the genomic coverage of the 170 lines. The top two founders, namely HUANGC and ZONG31 from the Z330 group, contributed 15.09% and 13.14% *bins*, respectively.

The F_1 progenies of the 170 lines identified using GOVS mostly exhibited higher EWs than the rest of the 532 lines contributed zero *bins* (Figure 4b). It was, therefore, reasonable to presume that GOVS may assist in the selection of superior lines, and it was tested using the 702 test samples. The EWs of the F_1 progeny of the 702 test samples were first predicted using the rrBLUP model trained with the genotypes and

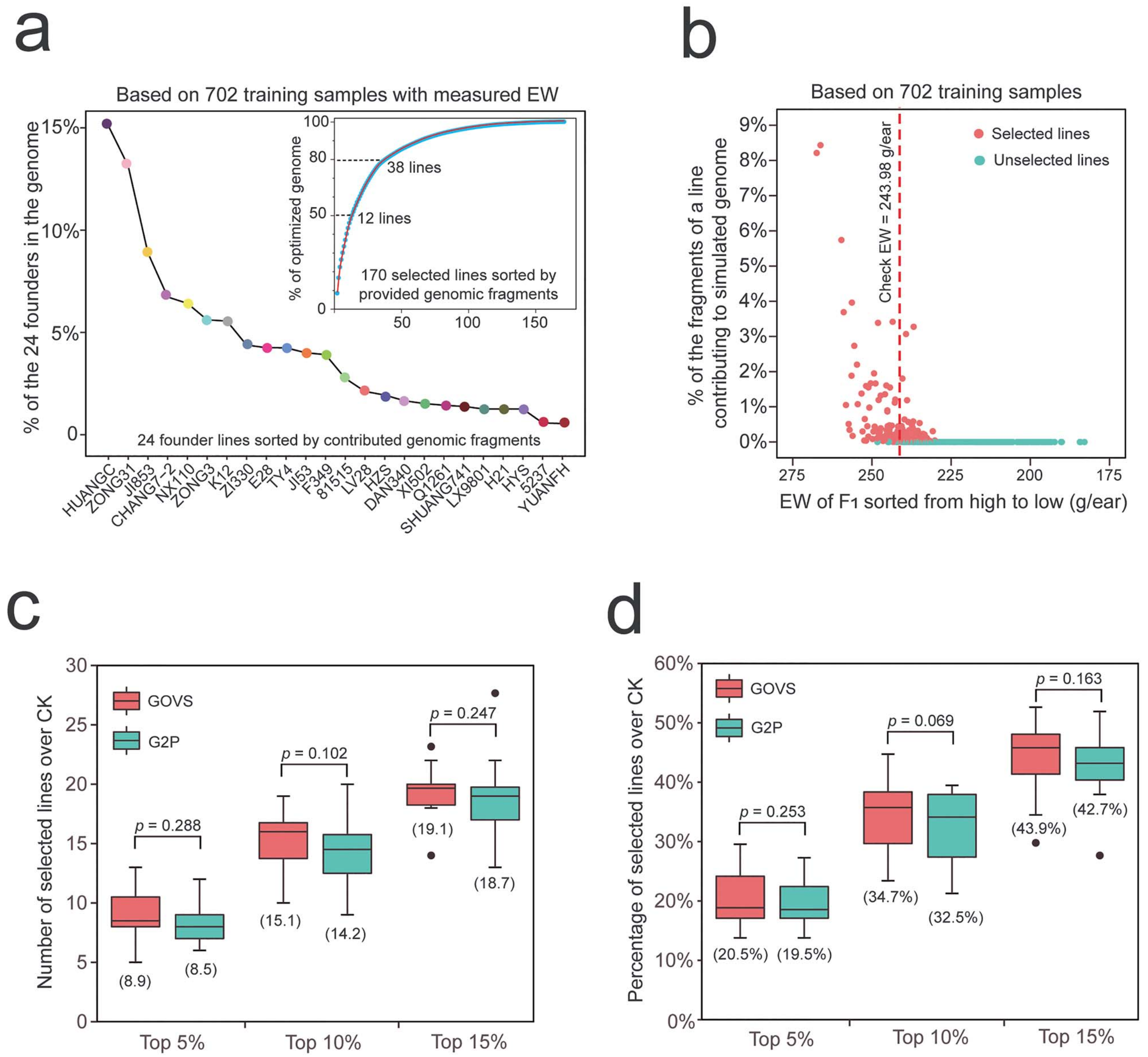

Figure 4. GOVS-assisted selection based on contributed *bins*. (a) The percentage of the 24 founders contributing genomic fragments with optimal genotypes to the optimal genome simulated using the genotypes of the F_1 progeny of the 702 training samples crossed with Zheng58. The inner box shows that 12, 38 and 170 lines contributed 50%, 80% and 100% fragments, respectively, to the simulated genome. (b) The lines contributing higher fractions of fragments also showed a higher yield in their F_1 progeny based on the analysis of the 702 training samples. (c) Comparison of the numbers of superior lines surpassing checks identified by GOVS and G2P methods. The numbers in the brackets are the average numbers of corrected superior lines among the top 5%, 10% and 15% lines with predicted values resulted from the 10 times of holdouts. (d) Comparison of the average percentages of superior lines surpassing checks identified by GOVS and G2P methods.

observed EWs of the 702 training samples. GOVS, then, assembled the virtually optimized genome based on the predicted EWs. Out of the 702 test samples, 158 contributed *bins* with optimal genotypes, among which the top 8, 26 and 158 lines cumulatively contributed 50%, 80% and 100% *bins* to the simulated genome, respectively. Using the selection threshold of 10%, 95.74% of the simulated genome was achieved (Figure S4a). Similar to the results from the training samples, the F_1 progenies of the selected lines mostly showed higher EWs compared to those of the unselected lines (Figure S4b). Additionally, out of the top 8 founders identical in both training and test samples, 38.23% of the genomic context of the superior lines was traced back to the four Z330 founders indicating the importance of the Z330 group in forming the superior hybrids in the CUBIC population. In contrast, the four founders in LDHG and 16 founders in STP contributed 11.95% and 49.82% of the genomic context in the CUBIC population, respectively.

Finally, we compared the precision of selecting superior lines assisted by G2P and GOVS. A repeated holdout method was performed for cross-validation (Methods). Averaged numbers and percentages of superior lines surpassing checks from the 10 time of holdouts were computed at selection rates of top 5%, 10% and 15% from the list of testing samples sorted by the G2P and GOVS methods based on predicted phenotypes and contributed *bins*, respectively. As shown in Figure 4c and d, selection precision of GOVS was slightly better than that of G2P, but the difference was not significant according to the comparison of the two methods by paired *t*-test. Therefore, we may conclude that precision of

GOVS-assisted selection of superior lines is comparable to that of G2P-assisted selection.

Since the feasibility of GOVS-assisted selection of superior lines was validated, we applied GOVS using the EWs of the 1404 F_1 progeny to generate a virtually optimized genome. Out of the 1404 CUBIC lines, 253 (18.02%) superior lines were selected using GOVS that contributed *bins* with optimal genotypes to the simulated genome. The 253 superior lines selected by GOVS were then considered as the new genetic pool enriched with advantageous alleles expressing a higher yield than the unselected lines, among which 45.8% (116 out of 253) featuring high F_1 EWs surpassing the check (Figure S4c).

Selection of the optimal paternal lines to generate F_1 combinations

Since the selection of the 253 superior lines assisted by GOVS was based on test-crossing with only one paternal line Zheng58, the next step was to select optimal paternal lines that may generate superior F_1 combinations to further improve the grain yield of hybrid maize. This may be implemented via the selection of a subset of lines from the 253 superior lines to cross with a panel of paternal lines from diverse genetic backgrounds, so that optimal heterotic progeny from different heterotic groups may be obtained. Therefore, 30 out of the 253 selected CUBIC lines were crossed with 30 paternal lines from the six major heterotic groups to generate 900 F_1 combinations (Figure S5). The three traits DTT, PH and EW of the 900 F_1 combinations were measured at five different locations. With the phenotype data, the GCA of EW for each paternal line was computed based on the corresponding set of the 30 F_1 progeny, and the resulting paternal GCA was used as a reference for selecting optimal paternal lines (Methods). The thirty sets of F_1 progeny were ranked using the paternal GCAs, and a line exhibiting high paternal GCA indicated a high probability of generating superior F_1 combinations if it were to be crossed with the 253 CUBIC lines (Figure 5a). The top five paternal lines were MG1533, MG1534, MG1543, MG1546 and MG1548 that generated 20, 17, 18, 14 and 9 F_1 combinations, respectively, out of the 30 F_1s in each set that surpassed the EW of ZhengDan958. GCA of the tester Zheng58 was ranked at the 11th position, and 10 of its 30 F_1s surpassed the threshold. It is worth noting that ZhengDan958 is a compact cultivar featuring a short plant stature and early flowering time suited for mechanical harvesting [44]. Thus, the yield per plant is not the top priority for ZhengDan958, as increasing the density of plants may improve the overall yield per unit due to its compact structure. Among the top 5 paternal lines, F_1s of MG1548 from the Reid group exhibited a significantly reduced PH ($P = 0.023$), slightly shortened DTT ($P = 0.365$), and an improved EW ($P = 0.049$) compared to the F_1s of Zheng58 (Figure 5b). Thus, MG1548 was selected for generating F_1 combinations in the field trial. F_1s of MG1533 showed average PH and DTT but the highest EW, compared to the other five sets of F_1s. Thus, MG1533 was also selected. MG1534 and MG1543 with similar GCAs showing comparable DTT, PH and EW, belonged to the X-population. Nevertheless, MG1534 showed better pathogen resistance than MG1543 according to empirical evidence from breeders. Thus, MG1534 was selected. Finally, 40, 20 and 20 lines were randomly selected from the 223 untested superior lines to cross with MG1533, MG1534 and MG1548, respectively, resulting in a total of 80 F_1 hybrids selected for validation through field trial.

Results from the field trials of the 80 F_1 hybrids

The 80 F_1 hybrids were successfully harvested and the EW, grain weigh and water content measured. Three commercial cultivars widely planted in China, namely ZhengDan958, XianYu335 and JingKe968, were used as spike-in check varieties for the 80 F_1 hybrids during the field trial. As for the phenotype of EW, 73, 16 and 7 F_1 hybrids out of the 80 combinations surpassed the EWs of ZhengDan958, XianYu335 and JingKe968, respectively (Figure 6a). As for the grain weight, the numbers of F_1s surpassing the three thresholds were 72, 9 and 1, respectively. (Figure S6). By far, the pipeline of selecting maternal lines, paternal lines, and generating superior F_1 combinations via the model of GOVS was demonstrated using the CUBIC population.

The 253 selected inbred lines exhibited insignificant but slightly higher EWs than the 1151 unselected lines (Figure S7); whereas, their F_1 progeny with Zheng58 exhibited significantly elevated EWs compared to the F_1 progeny of the unselected lines (Figure 6b). Furthermore, compared to the average EW (223.2 g/ear) of the 24 F_1 hybrids resulting from crossing the 24 founders with Zheng58, a 9.18% increment in the F_1 yield was achieved (average EW 244.1 g/ear) for the 253 superior lines selected from the 1404 samples (Figure 6b). It is worth noting that the EWs of the 1151 F_1s showed almost no differences when compared to the F_1s of the 24 founders. It explicitly suggested that the genes contributed by the 253 selected lines played an important role in generating superior hybrids for yield, and GOVS successfully identified these heterotic *bins*. The selection of paternal lines via GCA analysis further validated the importance of the *bins* in generating heterotic varieties, reflected by the significantly elevated EWs of the 80 F_1 progenies of MG1548, MG1533 and MG1534 compared to the F_1 progenies with Zheng58. Therefore, the effectiveness of the GOVS strategy in terms of accelerating genetic gains and establishing heterotic patterns was validated.

Discussion

GOVS improves genomically designed breeding in maize

The fast advancement in the DH technology has vastly accelerated the production of inbred lines of maize [14, 45, 46]. For each of the F_1 and F_2 combinations, two to five hundreds of inbred DH lines need to be generated to ensure sufficient exchanges of parental alleles [8]. A mid-sized seed company may regularly produce several tens to hundreds of thousands of DH lines per year, as DH production is applied on hundreds of hybrid combinations in parallel. Thus, precise and efficient screening of DH lines at a lower expense is in great demand. GS based on various G2P predictive models have been incorporated in the modern breeding pipeline to assist the screening of lines but the precision of prediction of the yield is still not satisfactory since crop yield is a complicated polygenic trait influenced by several factors [1, 8, 16, 18, 47–49]. In the current work, we implemented the previously proposed model of GOVS [8]. GOVS was developed and tested using the previously published genotype and phenotype data of the CUBIC population and their 1404 F_1 progeny [23, 42]. The results from the field trial of 80 F_1 combinations selected using GOVS further verified the feasibility of this strategy. It also should be noted that GOVS is not a substitute for, but a complement to GS-assisted selection, as it may better interpret and utilize the results from G2P prediction to direct breeding decisions. One of the main differences between GS and GOVS is that, while GS-assisted selection is based on the predicted phenotype using an arbitrary threshold, GOVS selects lines based on the number of *bins* that a line contributes to the assembly of optimal genome. Moreover, the lines selected by GOVS possess complementary sets of *bins*, and this information may be used to direct next round of crossing to further pyramid

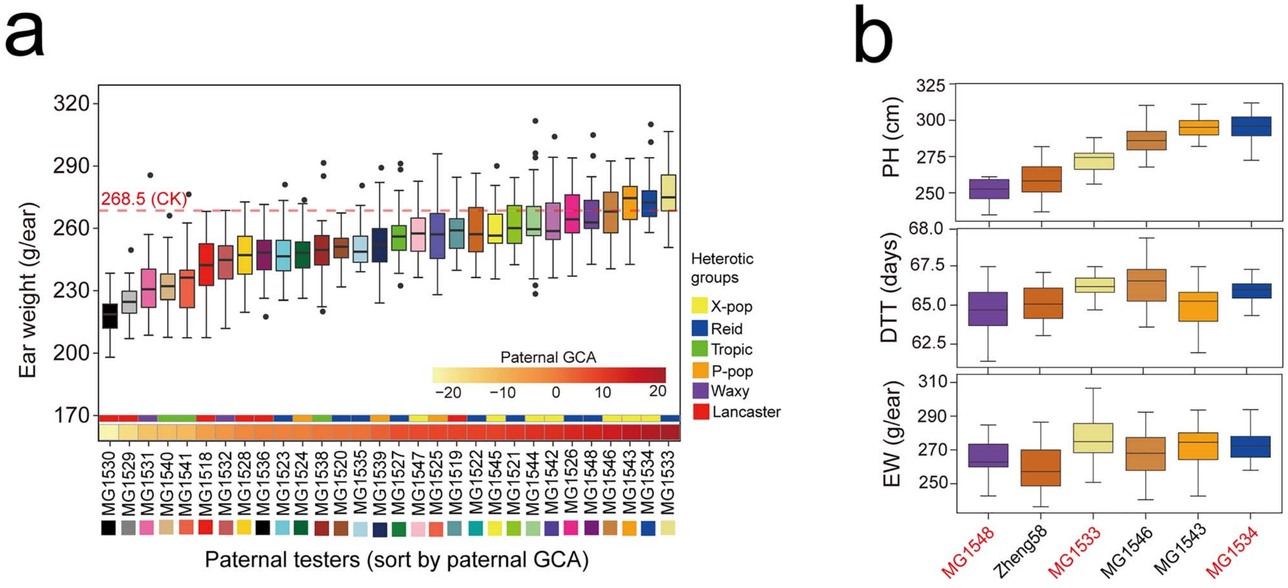

Figure 5. Determination of the optimal paternal lines. (a) The thirty sets of F_1 hybrids were ranked according to the paternal GCA of EW computed based on the 30 sets of F_1 combinations resulting from crossing the 30 CUBIC lines with each one of the 30 paternal lines. (b) Comparison of DTT, PH and EW of the six sets of F_1 combinations to facilitate the determination of the optimal paternal lines.

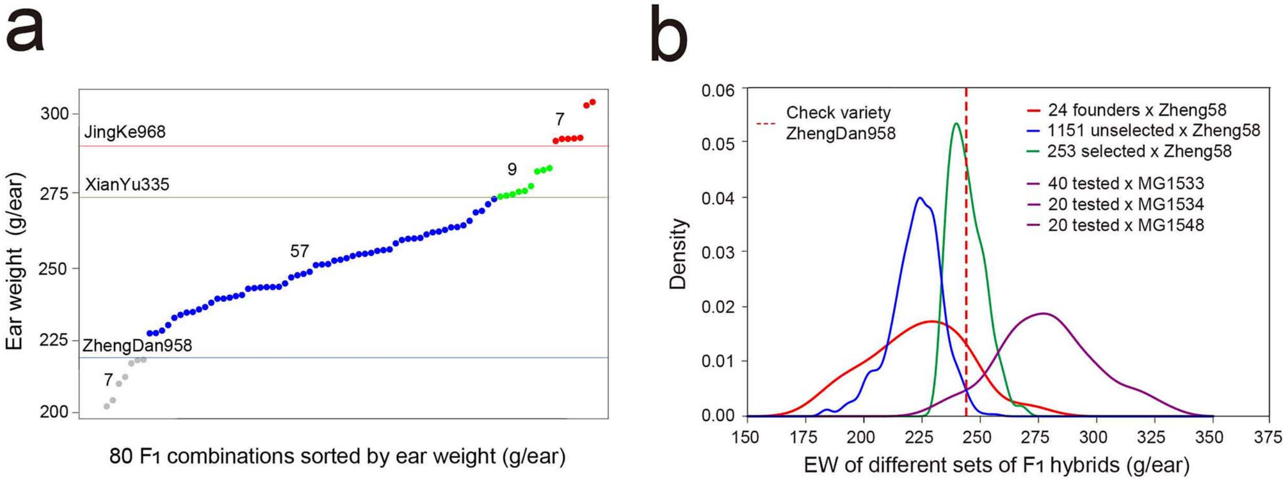

Figure 6. The result of the field trial of the 80 F_1 combinations. EW of 73, 16 and 7 out of the 80 F_1 combinations surpassed the threshold of EW of cultivars Zheng958, XianYu335 and JingKe968, respectively. (a) Genetic improvement in EW reflected by the four populations of the F_1 progeny.

advantageous alleles. Therefore, GOVS may theoretically accelerate the progress of genetic gain with fewer numbers of lines and cycles of hybridizations compared to traditional GS method.

Maize is one of the crops that mainly utilizes heterosis [50, 51]. The heterotic pattern established between the maternal and paternal groups is extremely important and should not be ignored when incorporating GOVS and GS into the DH-empowered breeding pipeline. Thus, pyramiding of the advantageous alleles in the maternal and paternal groups needs to be performed separately, so that the fixed heterotic patterns are maintained. In other words, DH production should be strictly applied to the F_1 and F_2 hybrids generated by crossing the lines within each group but not between groups. To minimize the times of crossing and to achieve maximum pyramided advantageous alleles using the fewest lines, the precise selection of the lines carrying complementary sets of advantageous alleles is essential. GOVS offered such precision using the statistics on the genomic context of the virtually optimized genome, which may facilitate the selection of lines with complementary advantageous alleles to generate hybrids to be sent for DH production. According to our analysis, the selection of dozens of top lines contributing a high proportion of *bins* may help achieve an ideal coverage of the genome (Figure 4a). Briefly, with the advantages of GOVS illustrated in the current work, it is foreseeable that the integration of DH production, G2P prediction and GOVS may ultimately improve genomically designed breeding for maize.

GOVS reveals basic principles of maize hybrid breeding

GOVS is not only a useful tool to assist data-driven decision-making but may also reveal multiple basic principles of maize hybrid breeding. Tracing the origins of the optimal haplotypes of *bins* to the 24 founders revealed that the top eight founders contributed almost two thirds (66.03%) of the *bins* to the 253 superior lines, among which, four lines from the Z330 group

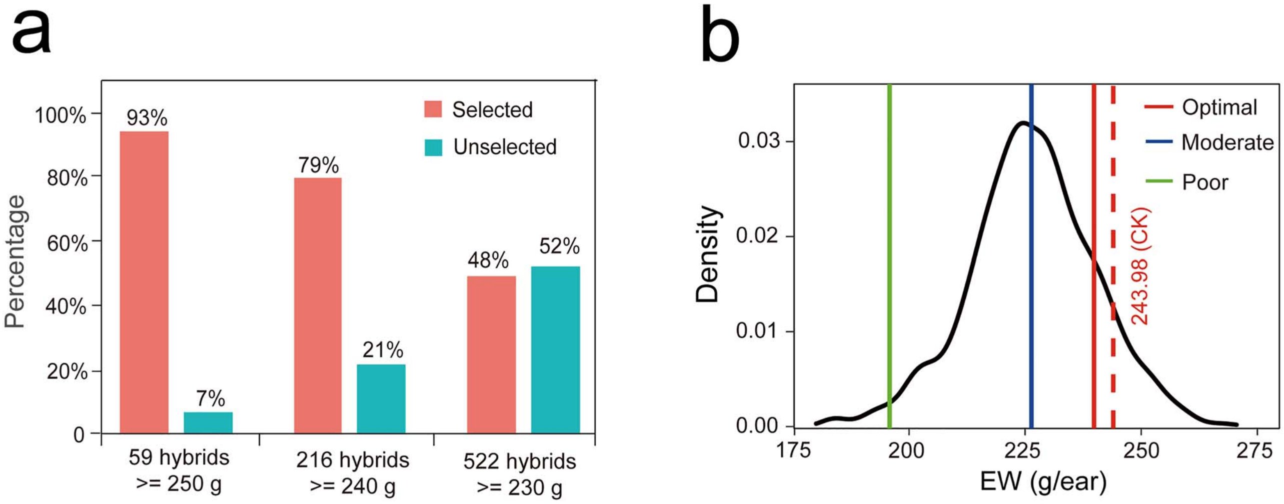

Figure 7. Basic principles of maize hybrid breeding. (a) The percentages of the selected and unselected lines whose EW of F_1 progenies are larger than the cutoffs of 250, 240 and 230 g/ear. (b) The predicted phenotypes of EW using the genotypes of the simulated optimal, moderate and poor genomes. The G2P model was trained using the observed phenotypes and the actual genotypes of the 1404 F_1 hybrids.

contributed 39.11% of the *bins* (Figure S8). These four Z330 lines were developed between the 70s and 90s in China, and they are rarely used nowadays in China. Except for HUANGC, whose F_1 progeny with Zheng58 surpassed the EW of ZhengDan958 (Chang7-2 × Zheng58), the progeny of the other three lines did not (Figure S2). The possible explanation is that the genetic contexts of the Z330 group may represent a general fitness in the local environments in China, although they do not show superior F_1 yield [52]. Nevertheless, due to the general fitness they showed in the new CUBIC population, pyramiding of the advantageous alleles from other founders may lead to better performance and form a better genetic complement to the paternal alleles [23]. Thus, it potentially suggested a basic principle for selecting founder lines when planning on creating novel germplasm, which is to include elite lines adapted to the local environment rather than basing your results only on yield. In addition, it also suggested that the indigenous germplasm is of great value offering broader adaptability to the environment when breeding new cultivars. Thus, even though the germplasm resource may not be used, nowadays, in the breeding industry, it is still worth using it to exploit the advantageous alleles for the improvement of maize.

Overall, the F_1 yields of the selected lines were significantly greater than those of the unselected lines, especially for those top-ranking lines contributing a substantial amount of *bins*. For instance, among the 59 hybrids with EWs over 250 g/ear, 55 (93%) lines contributed *bins* to the simulated genome; among the 216 hybrids with EWs over 240 g/ear, 171 (79%) contributed *bins* (Figure 7a). The 55 and 171 lines accounted for only 3.9% and 12.2%, respectively, of the 1404 lines. Thus, it indicated that the pyramiding of the selected superior lines with maximum combinations of the advantageous alleles is a low-probability event, similar to breeding, which is an arduous and time-consuming task. So far, only yield has been considered and the other traits and stress resistance have not. To further pyramid advantageous alleles that cover other traits, the only solution is to continuously enlarge population size, so that recombination may break the linkage between the advantageous and deleterious QTLs [4, 5, 42].

A significant correlation (Pearson's correlation coefficient, $r = 0.713$) was observed between the F_1 yields of the 58 superior lines and the number of *bins* they contributed (Figure S9). This correlation suggested an important role of the additive effect of the advantageous alleles, consistent with the previous report in rice that heterosis is a result of pyramiding advantageous alleles contributed from both parental genomes [5, 39, 53, 54]. However, the simulated genome representing an assembly of advantageous alleles favoring a higher yield explained about 84% of the total genetic effect of heterosis. This proportion was deduced based on the optimal phenotype predicted using the optimal genotype of the simulated genome, which was positioned at the 83.9th percentile of the observed EW distribution of the 1404 F_1 hybrids (Figure 7b). The rest 16.1% may be explained by the genotype-by-environment (G × E) interaction [55, 56]. Thus, the optimal phenotype of the simulated genome is unable to surpass the threshold as the estimation of the contribution of G × E to the hybrid yield is difficult using a linear model. In contrast, the genetic effects contributing to heterosis of the flowering time and plant stature are mostly additive effects, as indicated by the positions of the 98th and 95th percentiles of the predicted DTT and PH among the 1404 F_1 hybrids, respectively (Figure S10).

Concluding remarks

Modern breeding is characterized by the adoption of multidisciplinary science and technologies into breeding, including plant biology and quantitative genetics, genome editing and synthetic biology, bioinformatics and machine learning, as well as high-throughput genotyping and phenotyping, which enable the rapid development of genome-wide combinations of advantageous alleles that express the desirable traits [57, 58]. Although the virtually simulated genomes with optimal genotypes expressing optimal phenotypes may never be developed in reality, it may accelerate maize breeding using the minimum numbers of breeding materials and times of hybridization, to achieve the maximum genetic gain with the fewest breeding cycles. Furthermore, the DH technology has been available in 43 plant species with mature production protocols so far, and it is widely applied in major cereals like maize, rice and wheat in the seed industry [59, 60]. Due to the greatly accelerated process of line development, population of inbred or hybrid lines subjected to selection has been much larger than previous years. With the advantages of GOVS illustrated in maize in this study, it is foreseeable that GOVS has the potential to incorporate the strategy of simulation of virtually optimized genome as part of the DH breeding to accelerate genomically designed breeding in other plant species.

Data availability

Source codes, scripts and demo data are publically available at the website of GOVS https://govs-pack.github.io/.

Author contributions

X-F.W., M.C. and J-B.Y. conceived and supervised the project, F.X. constructed the bin map, Q.C. and S-Q.J. performed genome optimization analysis and developed the GOVS package, Q.W. performed G2P prediction, Y-J.X. processed the genotype and phenotype data of 1404 F_1 hybrids, R-Y.Z. and J-R.Z. collected the data from the field trial, and X-F.W. and Q.C. wrote the manuscript.

Key Points

- GOVS traces the inheritance of a genomic fragment back to the parental lines used to produce DH lines, and quantitatively assesses its genetic contribution to F_1 yield.
- GOVS determines the optimal genotype of each genomic fragment positively contributing to F_1 yield, and the optimal genotype is subsequently selected to assemble the virtual genome.
- GOVS assists in selection of superior lines based on the number of fragments that a line contributes to the virtually optimized genome, rather than basing it on predicted phenotypes.
- GOVS plots the optimal route to pyramid the maximum advantageous alleles, since the GOVS-selected lines contribute complementary sets of advantageous alleles in known proportions to the optimal genome.
- GOVS accelerates the progress of genetic gain using the fewest breeding materials and fewest events of hybridization though genomically designed breeding.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Funding

This work was supported by the National Key Research and Development Program of China (2018YFA0901000), the 2020 Research Program of Sanya Yazhou Bay Science and Technology City (SKJC-2020-02-003), the National Science Foundation of China (31871706), the China Postdoctoral Science Foundation (2020TQ0355) and the 2115 Talent Development Program at China Agricultural University.

References

1. Hickey LT, Hafeez AN, Robinson H, *et al.* Breeding crops to feed 10 billion. *Nat Biotechnol* 2019;**37**:744–54.
2. Voss-Fels KP, Stahl A, Hickey LT. Q&A: modern crop breeding for future food security. *BMC Biol* 2019;**17**:1–7.
3. Xu Y, Liu X, Fu J, *et al.* Enhancing genetic gain through genomic selection: from livestock to plants. *Plant Commun* 2020;**1**:100005.
4. Wijnker E, de Jong H. Managing meiotic recombination in plant breeding. *Trends Plant Sci* 2008;**13**:640–6.
5. Huang X, Yang S, Gong J, *et al.* Genomic architecture of heterosis for yield traits in rice. *Nature* 2016;**537**:629–33.
6. Studer AJ, Wang H, Doebley JF. Selection during maize domestication targeted a gene network controlling plant and inflorescence architecture. *Genetics* 2017;**207**: 755–65.
7. Fernie AR, Yan J. De novo domestication: an alternative route toward new crops for the future. *Mol Plant* 2019;**12**:615–31.
8. Jiang S, Cheng Q, Yan J, *et al.* Genome optimization for improvement of maize breeding. *Theor Appl Genet* 2020;**133**:1491–1502.
9. Xu Y, Li P, Zou C, *et al.* Enhancing genetic gain in the era of molecular breeding. *J Exp Bot* 2017;**68**:2641–66.
10. Birchler JA, Yao H, Chudalayandi S, *et al.* Heterosis. *Plant Cell* 2010;**22**:2105–12.
11. Hochholdinger F, Baldauf JA. Heterosis in plants. *Curr Biol* 2018;**28**:R1089–92.
12. Birchler JA. Heterosis: the genetic basis of hybrid vigour. *Nat Plants* 2015;**1**:15020.
13. Ren J, Wu P, Trampe B, *et al.* Novel technologies in doubled haploid line development. *Plant Biotechnol J* 2017;**15**: 1361–70.
14. Prigge V, Xu X, Li L, *et al.* New insights into the genetics of in vivo induction of maternal haploids, the backbone of doubled haploid technology in maize. *Genetics* 2012;**190**:781–93.
15. Heffner EL, Lorenz AJ, Jannink JL, *et al.* Plant breeding with genomic selection: gain per unit time and cost. *Crop Sci* 2010;**50**:1681–90.
16. Jannink J-L, Lorenz AJ, Iwata H. Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 2010;**9**:166–77.
17. Desta ZA, Ortiz R. Genomic selection: genome-wide prediction in plant improvement. *Trends Plant Sci* 2014;**19**:592–601.
18. Crossa J, Pérez-Rodríguez P, Cuevas J, *et al.* Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci* 2017;**22**:961–75.
19. Yu X, Li X, Guo T, *et al.* Genomic prediction contributing to a promising global strategy to turbocharge gene banks. *Nature Plants* 2016;**2**:1–7.
20. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 2001;**157**:1819–29.
21. Millet EJ, Kruijer W, Coupel-Ledru A, *et al.* Genomic prediction of maize yield across European environmental conditions. *Nat Genet* 2019;**51**:952–6.
22. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, *et al.* Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* 2012;**44**:217.
23. Liu H-J, Wang X, Xiao Y, *et al.* CUBIC: an atlas of genetic architecture promises directed maize improvement. *Genome Biol* 2020;**21**:1–17.
24. Kremling KAG, Chen S-Y, Su M-H, *et al.* Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 2018;**555**:520–3.
25. Kono TJY, Fu F, Mohammadi M, *et al.* The role of deleterious substitutions in crop genomes. *Mol Biol Evol* 2016;**33**: 2307–17.

26. Guo Z, Wang H, Tao J, *et al*. Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. *Mol Breed* 2019;**39**:1–12.
27. Stevanato P, Broccanello C, Pajola L, *et al*. Targeted next-generation sequencing identification of mutations in disease resistance gene analogs (RGAs) in wild and cultivated beets. *Genes* 2017;**8**:264.
28. Ren W, Gong X, Li K, *et al*. Recombination pattern characterization via simulation using different maize populations. *Int J Mol Sci* 2020;**21**:2222.
29. Tulsieram L, Compton WA, Morris R, *et al*. Analysis of genetic recombination in maize populations using molecular markers. *Theor Appl Genet* 1992;**84**:65–72.
30. Liu N, Liu J, Li W, *et al*. Intraspecific variation of residual heterozygosity and its utility for quantitative genetic studies in maize. *BMC Plant Biol* 2018;**18**:1–15.
31. Luo J, Wei C, Liu H, *et al*. MaizeCUBIC: a comprehensive variation database for a maize synthetic population. *Database* 2020;**2020**:1–8.
32. Purcell S, Neale B, Todd-Brown K, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75.
33. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* 2009;**25**:4.10.1–4.10.14.
34. Lawrence CJ, Dong Q, Polacco ML, *et al*. MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res* 2004;**32**:D393–7.
35. Tamura K, Dudley J, Nei M, *et al*. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* 2007;**24**:1596–9.
36. Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011;**12**:2825–30.
37. Bates D, Machler M, Bolker BM, Walker SC. Fitting linear mixed-effects models using lme4. *J Stat Softw* 2015;**67**(1):1–48.
38. Mott R, Talbot CJ, Turri MG, *et al*. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci* 2000;**97**:12649–54.
39. Wei X, Qiu J, Yong K, *et al*. A quantitative genomics map of rice provides genetic insights and guides breeding. *Nat Genet* 2021;**53**:243–53.
40. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 2011;**4**:250–5.
41. Lai J, Li R, Xu X, *et al*. Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 2010;**42**:1027–30.
42. Xiao Y, Jiang S, Cheng Q, *et al*. The genetic mechanism of heterosis utilization in maize improvement. *Genome Biol* 2021;**22**:1–29.
43. Li H, Yang Q, Fan N, *et al*. Quantitative trait locus analysis of heterosis for plant height and ear height in an elite maize hybrid zhengdan 958 by design III. *BMC Genet* 2017; **18**:36.
44. Wang X, Wang X, Xu C, *et al*. Decreased kernel moisture in medium-maturing maize hybrids with high yield for mechanized grain harvest. *Crop Sci* 2019;**59**:2794–805.
45. Longin CFH, Utz HF, Reif JC, *et al*. Hybrid maize breeding with doubled haploids: III. Efficiency of early testing prior to doubled haploid production in two-stage selection for testcross performance. *Theor Appl Genet* 2007;**115**:519–27.
46. Wang B, Zhu L, Zhao B, *et al*. Development of a haploid-inducer mediated genome editing system for accelerating maize breeding. *Mol Plant* 2019;**12**:597–602.
47. Heffner EL, Sorrells ME, Jannink JL. Genomic selection for crop improvement. *Crop Sci* 2009;**49**:1–12.
48. Lorenz AJ, Chao S, Asoro FG, *et al*. Genomic selection in plant breeding: knowledge and prospects. *Adv Agron* 2011;**110**:77–123.
49. Wray NR, Yang J, Hayes BJ, *et al*. Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* 2013;**14**:507–15.
50. Sprague GF. Heterosis in maize: theory and practice. In: Frankel R (ed). *Heterosis: Reappraisal of Theory and Practice*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983, 47–70.
51. Schnable PS, Springer NM. Progress toward understanding heterosis in crop plants. *Annu Rev Plant Biol* 2013;**64**: 71–88.
52. C-f L, Z-q T, Liu P, *et al*. Increased grain yield with improved photosynthetic characters in modern maize parental lines. *J Integr Agric* 2015;**14**:1735–44.
53. Huang X, Yang S, Gong J, *et al*. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* 2015;**6**:1–9.
54. Zhou G, Chen Y, Yao W, *et al*. Genetic composition of yield heterosis in an elite rice hybrid. *Proc Natl Acad Sci* 2012;**109**:15847–52.
55. Abdulai M, Adu G, Akromah R, *et al*. Assessment of genotype by environment interactions and grain yield performance of extra-early maize (*Zea mays* L.) hybrids. 2013.
56. Doust AN, Lukens L, Olsen KM, *et al*. Beyond the single gene: how epistasis and gene-by-environment effects influence crop domestication. *Proc Natl Acad Sci* 2014;**111**: 6178–83.
57. Ramstein GP, Jensen SE, Buckler ES. Breaking the curse of dimensionality to identify causal variants in breeding 4. *Theor Appl Genet* 2019;**132**:559–67.
58. Wallace JG, Rodgers-Melnick E, Buckler ES. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Annu Rev Genet* 2018;**52**: 421–44.
59. Segui-Simarro JM, Moreno JB, Fernandez MG, *et al*. Species with haploid or doubled haploid protocols. *Methods Mol Biol* 2021;**2287**:41–103.
60. Weyen J. Applications of doubled haploids in plant breeding and applied research. *Methods Mol Biol* 2021;**2287**:23–39.