Theoretical and Applied Genetics (2020) 133:1491–1502
https://doi.org/10.1007/s00122-019-03493-z

REVIEW

# Genome optimization for improvement of maize breeding

Shuqin Jiang[1] · Qian Cheng[2] · Jun Yan[1] · Ran Fu[1] · Xiangfeng Wang[1]

Received: 14 July 2019 / Accepted: 26 November 2019 / Published online: 6 December 2019


***Key message*** **We propose a new model to improve maize breeding that incorporates doubled haploid production, genomic selection, and genome optimization.**

**Abstract** Breeding 4.0 has been considered the next era of plant breeding. It is clear that the Breeding 4.0 era for maize will feature the integration of multi-disciplinary technologies including genomics and phenomics, gene editing and synthetic biology, and Big Data and artificial intelligence. The breeding approach of passively selecting ideal genotypes from designated genetic pools must soon evolve to virtual design of optimized genomes by pyramiding superior alleles using computational simulation. An optimized genome expressing optimal phenotypes, which may never actually be created, can function as a blueprint for breeding programs to use minimal materials and hybridizations to achieve maximum genetic gain. We propose a new breeding pipeline, "genomic design breeding," that incorporates doubled haploid production, genomic selection, and genome optimization and is facilitated by different scales of trait predictions and decision-making models. Successful implementation of the proposed model will facilitate the evolution of maize breeding from "art" to "science" and eventually to "intelligence," in the Breeding 4.0 era.

## Introduction

Along with rice and wheat, maize is a global staple cereal crop. Maize is not only an important nutrition source for humans, but is also a vital material in livestock feed and for bioenergy processes. Sustainable growth of maize yield per acre is critical for maintaining global food security. According to data released by the United Nations in 2016, the global human population will increase by 2 billion in the next 30 years and may exceed 11.2 billion by the end of this century (https://population.un.org). Based on predictions by the Food and Agriculture Organization (FAO) of the United Nations, cereal production needs to increase at least 70% by 2050 to accommodate this predicted world population growth (FAO 2011). To ensure food security, annual cereal production needs to increase from the current 2.1 billion tons to 3 billion tons, and annual meat production needs to increase more than 200 million tons. However, dramatic changes in the global climate have increased the frequency of extreme weather and natural disasters, both of which negatively influence crop production. Together, these global problems pose new challenges to the seed industry and necessitate revolutionary changes in crop breeding technology. These technologies must accelerate the cultivation of novel crop varieties that not only display high yield, superior quality, and stress resistance but are also ecologically and environmentally friendly. The second decade of the twenty-first century has been marked by the rapid advancement of artificial intelligence (A.I.) and its broad application in the life sciences (Webb 2018). These advancements offer an opportunity for crop breeding to enter a new era characterized by deep integration of modern information sciences and biotechnologies.

## Breeding 4.0: the intelligence era of maize breeding

In 2018, Wallace et al. suggested that the entire history of plant breeding could be divided into three major stages (Wallace et al. 2018). Using maize breeding as an example, the Breeding 1.0 stage ranges from the origin of maize

Communicated by Mingliang Xu.

✉ Xiangfeng Wang
xwang@cau.edu.cn

[1] National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100913, China

[2] Key Laboratory of Biology and Genetics Improvement of Maize in Arid Area of Northwest Region, Ministry of Agriculture, Northwest A&F University, Shaanxi, China

breeding in southern Mexico approximately 9000 years ago to the nineteenth century. During this stage, maize farmers cultivated landrace varieties by observing phenotypes and selecting desired trait variations purely based on their experience. This stage was characterized by "experience breeding." The Breeding 2.0 stage comprises the entire twentieth century. Professional maize breeders at seed companies used experimental breeding with predesigned breeding schemes that utilized current knowledge in genetics and statistics. This approach is also known as "experimental breeding." The Breeding 3.0 stage spans the last five decades, in which modern commercial breeding companies employed bioengineering technologies to genetically modify crop genomes to artificially create desired traits. This approach is also known as "biological breeding." Wallace et al. proposed that plant breeding technology is poised to evolve to the 4.0 stage; however, before successful transition from 3.0 to 4.0, three fundamental questions must be addressed. Each question is related to a major breeding objective: first, what are the biological mechanisms underlying environmental adaptation to fast climate change; second, how do we harness the correlation between genotypic and phenotypic variation to precisely utilize trait-regulatory genes; and third, how do we understand the roles of deleterious alleles to break the bottleneck in trait improvement. With a full understanding of these three fundamental questions, breeders will then be able to directionally modify or even de novo synthesize an optimal crop genome toward desired traits. The Breeding 4.0 era will be characterized by the integration of life science and informatics technologies to accelerate breeding cycles. Crop breeding in 4.0 will be the result of a Big Data-driven, A.I.-supported decision-making pipeline that includes integrative modeling of genotypic, phenotypic, environmental, and field management data. Thus, Breeding 4.0 may be considered the "intelligent breeding" stage.

A.I. is a branch of applied computer science and refers to the use of computer programs to simulate the human brain in understanding and processing information. The fields of A.I. include robotics, natural language processing, image recognition and processing, expert intelligent decision systems, and most importantly, machine learning analytical methodologies. The machine learning (ML) field, regarded as the "brain of A.I.," develops computational programs to build data-mining models to forecast the future, with the purpose of performing classifications and/or regressions. To build an ML system, the predictive model needs to be first trained by a set of training samples (training set) so that it can learn the sample features and establish innate correlations between features and outcomes. The trained ML model is then applied to a set of testing samples (testing set) to predict outcomes based on sample features. The most outstanding difference between ML and statistical models is that ML methods are more suitable for solving black-box questions because they automatically derive distribution parameters and feature importance. In contrast, statistical models must establish a hypothesis using prior knowledge of the distribution features of the data and then fit the data to the corresponding statistical model. Because of its 3V (volume, variety, and velocity) characteristics, Big Data commonly generates black-box problems that are suitable for ML methods to solve and build models for a variety of prediction goals.

Development of the Breeding 4.0 pipeline will require revolutionary innovation driven by Big Data and facilitated by biotechnology and informatics to facilitate decision making for professional breeders. Such a scenario may be imagined in the context of a future breeding company. The genotypes of germplasm lines will be acquired using high-throughput genotyping approaches such as next-generation sequencing (NGS) and SNP array platforms. Plant phenotypes will be acquired by field robotics and drones and dissected by deep learning algorithms to intelligently quantify trait measurements. Complex associations between genotypes, phenotypes, and the environment will be derived by ML models to direct germplasm selection and plan hybridization schemes. In a biotech facility, core trait-regulatory genes will be identified using bioinformatic approaches and genome-wide association studies (GWAS) to identify candidate genes for functional characterization and directional improvements of desired traits by gene editing, synthetic biology, and transgenic overexpression. With the rapid growth of "-omics" technologies, Big Data, and gene function knowledge in maize, it may soon be feasible to in silico assemble all possible superior alleles at quantitative trait loci (QTLs) into a single virtual, optimized genome to simulate optimal phenotypes. In contrast to "molecular design breeding," which mostly pyramids individual genes with explicit functions to facilitate introgression breeding, "genomic design breeding" is perhaps a more appropriate strategy for heterotic breeding, such as for maize, where yield hybrid vigor involves complex intra- and inter-genome interactions that are not easily characterized.

## Omics data foundations of genome design breeding

In order to achieve intelligent breeding in maize using genome optimization, it is necessary to characterize the correlations between traits and superior alleles, inferior alleles, and their complex interactions. The integration of maize omics data, including genomic, phenomic, epigenomic, transcriptomic, proteomic, and metabolomic data, will form an essential foundation for ML methods to model the relationships between various genetic elements as a network (Cooper et al. 2014; Ma et al. 2014). Recently, a series

of GWAS analyses have facilitated identification of numerous major-effect genes responsible for trait domestication, regulation, and improvement in maize (Jiao et al. 2012; Li et al. 2013; Ma et al. 2018; Wang et al. 2016). Nevertheless, the "omnigenic model" suggests that complex traits, such as height in humans and yield in crops, are products of many genes contributing to the traits with minor effects (Wray et al. 2018). Thus, phenotype prediction must account for every genetic component to fully describe the heritability of quantitative traits, regardless of the major or minor effectiveness of each QTL (Wray et al. 2013). In addition, genetic interactions between superior and inferior alleles are of particular interest, because trait improvement during breeding is partially a consequence of how superior and inferior alleles are balanced by artificial selection (Li et al. 2015). Therefore, the roles of superior and inferior alleles and how they have been fixed in modern breeding germplasms to contribute to heterosis performance and environmental adaptation require an in-depth investigation.

Due to the dramatic reduction in the cost of NGS, it is feasible to generate large-scale genotype data from actual, modern breeding populations that are comprised of different heterotic groups. High-throughput phenotyping focused on traits that display heterosis in the corresponding $F_1$ hybrid populations is also required. Such genotypic and phenotypic data describing modern breeding populations may help elucidate the role of genetic interactions between superior and inferior alleles and how such interactions have been fixed and utilized to generate hybrid vigor in maize production. For the maize basic research community, spatial and temporal omics data, including transcriptomic, proteomic, epigenomic, and metabolomic data, need to be continuously generated in multiple inbred reference lines (i.e., B73, Mo17, W22, B104, PH207, and CML247). Acquisition of multidimensional omics data in maize will facilitate identification of key regulatory genes and dissection of the molecular networks underlying important agronomic traits and other biological processes (Luo 2015).

Phenomics is a young and growing interdisciplinary field (Houle et al. 2010). A rapid development of A.I.-supported crop phenomics is inseparable from the wide application of field robotics systems, optical imaging systems, and deep learning-based image recognition and processing algorithms. A.I.-powered phenotyping systems are not only used for basic research but also represent the future of intelligence agriculture and precision farming. To collect precise environmental parameters, Internet of Things (IoT) devices equipped with various electronic sensors continuously record field data, including meteorological, soil, insect, and disease conditions (Xu 2016). These environmental factors may be integrated into the genotype-to-phenotype (G2P) predictive models to enhance prediction accuracy, especially for traits heavily influenced by the environment (Li et al. 2018). In addition, correlation of environmental data with genotypic and phenotypic data may facilitate modeling of genotype–environment interactions, identification of environment-responsive genes, and prediction of optimal ecological locations for a given maize variety.

Physiological phenotyping of crops is an upcoming area in phenomics (Ghanem et al. 2015). Under stressful conditions, such as pest damage, disease infection, nutrient deprivation, and other abiotic stresses, crops undergo a series of physiological and biochemical cellular changes (Cooper et al. 2014; Xu 2016). Although invisible to the human eye, changes in physiological phenotypes can be captured by various optical imaging modalities, including 3D laser scanning imaging, hyperspectral imaging, multispectral imaging, thermal imaging, near-infrared imaging, radar imaging, and kinetic imaging of chlorophyll fluorescence (Ghanem et al. 2015). The resulting image data are then dissected by deep learning algorithms to derive effective digital indicators that precisely reflect physiological changes (Ubbens and Stavness 2017). As stress response-related traits are complex phenotypes, it is difficult to genetically map major-effect genes using association populations. However, the detection of physiological changes inside plant cells may help break down these complex stress-responsive features into specific indicators to enhance the mapping power of GWAS or linkage analyses to identify individual causative genes (Chen et al. 2016).

## Pitfalls of genomic selection-assisted maize breeding

In marker-assisted selection (MAS), trait-linked DNA variations, mostly SNPs (single nucleotide polymorphisms), are used as markers to identify individual plants carrying desired alleles in introgression breeding programs (Bouchez et al. 2002). Application of MAS breeding requires cloning and characterization of the regulatory gene that displays a major effect on the target qualitative trait. MAS breeding has been most widely used for genetic improvement in rice, likely because many agronomically important rice genes have been cloned, allowing for successful implementation of molecular design breeding (Wing et al. 2018). Compared to rice, most maize genes are functionally uncharacterized, especially those involved in heterosis. In addition, maize breeding mostly depends on utilization of heterosis and involves genome-wide allelic interactions, QTL interactions, and inter-genomic interactions when the two parental genomes merge in the $F_1$ hybrid. Thus, genomic design breeding that considers whole-genome markers is a feasible and promising solution for maize breeding. Genomic selection (GS) is one form of genomic design breeding where it is not necessary to know the exact function of genes or

accurately evaluate the effectiveness of each individual marker (Voss-Fels et al. 2019). GS has been very successful in livestock breeding and has been gradually introduced into crop breeding, but fundamental differences between these two types of organisms may influence its efficacy (Hickey et al. 2017). This is especially true for predicting crop heterosis that involves a considerable proportion of nonadditive genetic effects (Chen 2010). As crop traits are more heavily influenced than livestock by the field environment, it will be necessary to integrate environmental factors into the GS model (Heslot et al. 2014).

For most existing GS tools, linear mixed models, such as the ridge-regression best linear unbiased prediction (rrBLUP) model, are used to perform regression-based predictions (Endelman 2011; Piepho et al. 2012). The GS model first needs to be trained using a reference population (training set), in which the genotypes and phenotypes of each individual sample are precisely measured. During training, the GS model infers the correlation between genotypes and phenotypes in the population and derives necessary parameters. Subsequently, the trained GS model is applied to a candidate population (testing set), and the genotype is used as the input to predict the phenotype outcome for each sample. In commercial maize breeding pipelines, GS has become an important decision-making step to assist in selection of inbred lines for single-cross hybridization based on the predicted $F_1$ phenotypes of all possible hybrid combinations that can be derived from the candidate population (Guo et al. 2019). However, based on the authors' knowledge and experience in employing GS in maize, two major pitfalls need to be cautioned when designing a GS experiment.

## Small training set, big testing set

In a single-cross hybridization program, maternal and paternal inbred lines are selected from two heterotic pools, between which hybridizations frequently generate strong heterosis performance. If 500 lines were selected from each pool, hybridizations could generate approximately 250,000 theoretical $F_1$ combinations. In a commercial breeding pipeline, 15–20% of the total combinations are field-tested to obtain phenotypes for the training population. The phenotypes of the remaining 80–85% of combinations will be predicted by the GS model. However, this unbalanced proportion of combinations used in the training and testing sets results in insufficient coverage of genotypes in the testing set, especially for low-frequency ($0.05 \leq \text{MAF} \leq 0.15$; Minor Allele Frequency) or rare alleles ($\text{MAF} < 0.05$), and thus reduces its predictive power.

We have evaluated the influence of insufficient sample coverage in a breeding population containing 1428 $F_1$ hybrids (unpublished data). The total population was partitioned into 207 (14.5%) training samples and 1221 (85.5%) testing samples. The rrBLUP model was used to predict DTT (days to tasseling), PH (plant height), and EW (ear weight), three representative maize traits with different degrees of heritability. The overall predictive powers, as evaluated by Pearson correlation, for the three traits by rrBLUP were DTT: $r = 0.371$, PH: $r = 0.344$, and EW: $r = 0.269$. The 1221 testing samples were than ranked in ascending order by the absolute difference between the observed and predicted trait values computed for each sample ($D_i = |y_i - \hat{y}_i|$) for comparison with the phenotype distribution of the 207 $F_1$ hybrids in training set. As shown in Fig. 1a, scissor-shaped distributions were observed for all three traits. Although the testing samples with high prediction accuracy (lowest 25% $D_i$) corresponded to the center of the training sample distribution, the testing samples with low prediction accuracy (highest 25% $D_i$) corresponded to the outliers of the training sample distribution. The remaining 50% of testing samples with moderate prediction accuracies were found between the center and outliers of the training sample distribution. This result indicates that the genotypes and phenotypes that comprise a small training set do not sufficiently represent the samples in the testing set, resulting in low predictive power.

A small training set may result in insensitive outlier prediction, unless the training set is large enough to contain low-frequency genotypes (rare alleles). This is perhaps the biggest challenge in heterosis prediction for maize, because the outstanding heterosis performance is generally defined by the outcome of specific compatible ability (SCA) due to low-frequency alleles and/or occasional nonadditive interactions with large over-dominance effects in the hybrid genome. Therefore, the optimization of GS models to enhance the sensitivity of outlier prediction or detect and weigh contributions from low-frequency alleles is expected future developments.

## Population stratification

Population stratification is another important factor that influences GS prediction. Population stratification has become an important issue because modern germplasms display complicated kinship due to crossing of lines from different heterotic subpopulations. If samples with discrepant genetic backgrounds are not proportionally distributed among the training and testing sets, population stratification may result in overfitting. Under such circumstances, it is necessary to adjust the model to remove any bias caused by population stratification. We tested the influence of population stratification using mixed $F_1$ hybrid populations (Fig. 1b). The training set was composed of 4140 $F_1$ hybrids generated by crossing 207 maternal lines with 20 paternal testers from different heterotic groups, such as the Reid, Lancaster, Iodent, and Tropical lines. Three testing sets were composed. The first testing set included

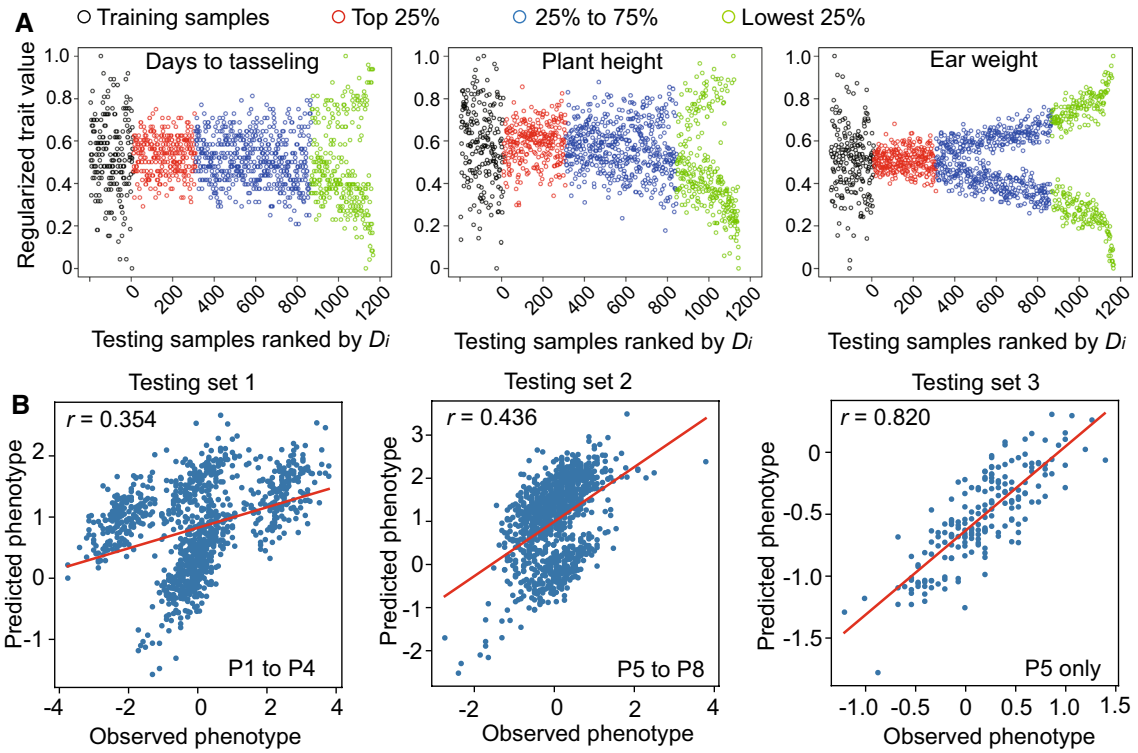


**Fig. 1** Pitfalls of genomic selection in phenotype prediction. **a**. A small training set and a large testing set can cause insufficient coverage of low-frequency alleles and outlier phenotypes, which may result in low prediction power. The maize trait values of days to tasseling (DTT), plant height (PH), and ear weight (EW) were regularized to a normal distribution with values between 0 and 1. A total of 1400 samples were used in this analysis. The first 200 samples comprised the training set and are indicated by black circles. The remaining 1200 samples comprised the testing set and were ranked in ascending order by the absolute difference between the observed and predicted trait values computed for each sample ($D_i = |y_i - \hat{y}_i|$), which corresponded to increasing prediction accuracy. Red, blue, and green circles represented testing samples with prediction accuracies in the following groups: top 25%, 25–75%, and lowest 25%. **b**. Population stratification reduces model robustness. The influence of population stratification on prediction accuracy was evaluated in three testing sets of $F_1$ hybrids. The first testing set included 800 $F_1$ hybrids generated by individually crossing 200 maternal lines with 4 paternal testers (*P*1–*P*4). *P*1 and *P*2 are Reid lines, and *P*3 and *P*4 are Tropical lines. The second testing set included 800 $F_1$ hybrids generated by individually crossing 200 maternal lines with four paternal testers (all Reid lines). The third testing set included 200 $F_1$ hybrids generated by crossing 200 maternal lines with one Reid paternal tester. DTT was used as the phenotype, and the absolute values of DTT were normalized to z-scores

207 maternal lines hybridized with four testers with discrepant genetic backgrounds, including two Reid and two Tropical testers. The scatterplot of the observed DTT phenotype versus the predicted DTT phenotype revealed a clearly stratified distribution, and the predictive power was low ($r = 0.354$). The second testing set included the same 207 maternal lines crossed with four Reid testers, and the stratification problem lessened, resulting in greater predictive power ($r = 0.436$). The third testing set included only hybrids generated from crossing 207 maternal lines with one Reid tester, and increase in predictive power due to stratification totally disappeared ($r = 0.820$). These results indicate that mixed genetic backgrounds in the training set may not influence the GS model, but if the testing set is composed of samples from mixed genetic backgrounds, model evaluations based on Pearson correlation may be biased.

## G2P prediction using machine learning methods

With the rapid development of high-throughput genotyping and phenotyping technology, Big Data analytics assisted by high-performance parallel computing are considered a promising approach to utilize millions of markers and super-large sample sets (Ma et al. 2014). As heterosis of yield-related traits is partly caused by nonadditive effects and influenced by the environment, nonlinear ML algorithms are presumed to be superior to linear GS models. Application of different ML algorithms in building GS

models has been found to produce ideal prediction performance (Crossa et al. 2017). However, one of the pitfalls of ML methods is that inappropriate training or over-tuning of ML parameters may result in substantial overfitting that is difficult to evaluate. Thus, to ensure model robustness, evaluation of ML should be conducted using only one testing set, but should include additional validation sets that are absolutely independent from the training and testing sets. Moreover, validation sets should not participate in cross-validation when training models.

To objectively compare nonlinear ML methods with the linear GS model, we conducted phenotype predictions using six ML methods and the rrBLUP model on an $F_1$ population with mixed genetic backgrounds from different heterotic groups. Unfortunately, none of the ML methods were able to surpass rrBLUP in their predictive power (Fig. 2a). One possible reason for this result is that ML methods may be advantageous for solving black-box problems without the need to know the data distribution characteristics, but for white-box problems in which statistical parameters are transparent and derivable, statistical models are more robust than ML methods. In addition, in our evaluation of published GS software, rrBLUP ranks highly for its precision, efficiency, and robustness (Fig. 2b). Thus, rrBLUP is an ideal choice for GS analysis.

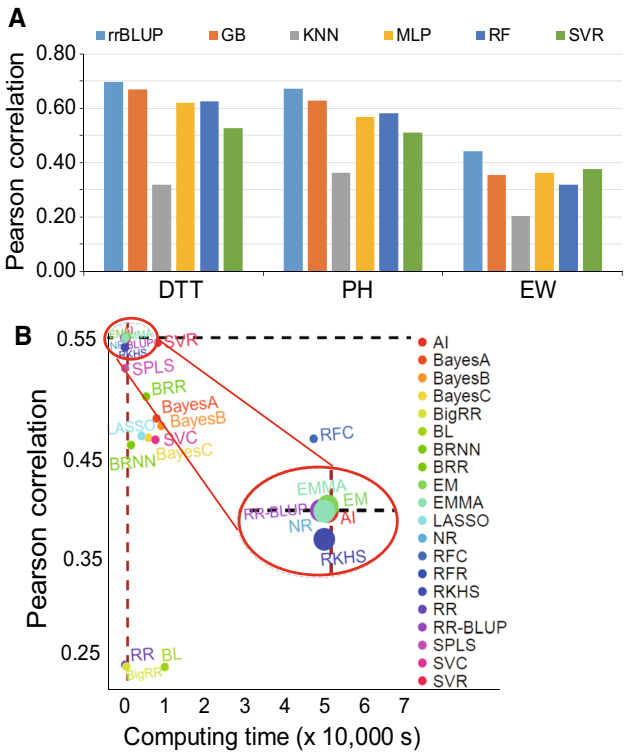


**Fig. 2** Machine learning models versus the rrBLUP model for phenotype prediction. **a**. Six machine learning (ML) models were used to build genomic selection models to predict DTT phenotypes. These ML models included convolutional neural network (CNN), gradient boosting (GB), random forest (RF), *K*-nearest neighbors (KNN), support vector regression (SVR), and multilayer perceptron (MLP) models. Evaluation of the seven methods was performed on the same training and testing sets. The total population was partitioned into 30 groups, and for each evaluation, 29 groups were used for training and the remaining group was used for testing. The rrBLUP model outperformed the six machine learning methods in 29 of 30 evaluations. **b**. Comparison of rrBLUP with other previously published genomic prediction tools revealed that rrBLUP is a superior algorithm in terms of prediction precision, training efficiency, and model stability

However, ML methods are not useless for G2P prediction. In fact, their utility for human risk assessment has been demonstrated using disease-related marker panels (Libbrecht and Noble 2015). In contrast to rrBLUP, which predicts phenotypes based on the kinship of the "genetic background" using whole-genome markers, ML methods are more effective for inferring the correlation between "genetic foreground" and traits using a panel of major-effect markers. Especially for certain deep learning methods such as deep neural networks, convolutional neural networks, and recurrent neural networks, the number of features needs to be far more less than the number of samples. Then, the perquisite of appropriately using these models is accurate identification of major-effect QTL haplotypes and development of universal marker panels. Otherwise, gradient vanishing and gradient exploding may lead to the failure of training process. Therefore, rrBLUP and ML methods may be complementary when used to address different prediction goals. For instance, rrBLUP may be first used for the first round of selection of candidates with top-ranked grain yield based on the genomic relationship of samples, and then, ML methods can be used for the second round to select specific combinations of beneficial genotypes with a small panel of SNPs associated with desired traits. In the past decade, numerous GWAS analyses targeting different important traits have been performed, and similar GWAS analyses targeting additional traits are ongoing. In the coming years, a large number of maize QTLs may be available so that universal, major-effect QTL haplotypes from germplasm banks may be used to accomplish trait prediction goals (Yu et al. 2016).

## G2P prediction in genomic design breeding

Decision making for genomic design breeding in modern maize breeding pipelines will be driven by three aspects of G2P prediction. Each aspect uses different scales of markers to achieve distinct trait improvement goals (Fig. 3). The first aspect is GS prediction to assist the selection of parental lines based on the phenotypes of $F_1$ hybrids whose genotypes are inferred from combining the two crossed parental genotypes. GS prediction uses whole-genome markers to infer correlations between genetic backgrounds and general agronomic traits of $F_1$ hybrids, such as maturity and

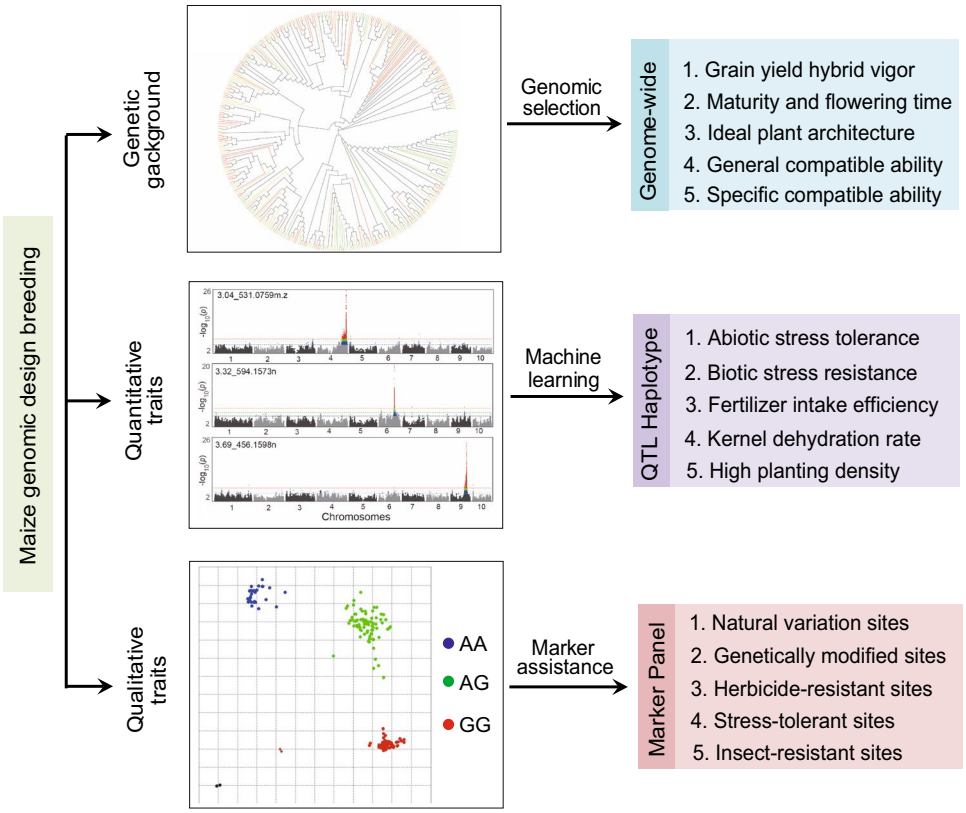


**Fig. 3** Genomic design breeding encompasses three aspects of G2P prediction. The proposed genomic design breeding pipeline includes three aspects of phenotype prediction. First, genomic selection is based on the kinship inferred from the relationship between the overall genetic backgrounds of the inbred lines. Thus, genomic selection should consider genome-wide markers. The goal is to the phenotypes contributed by a collection of all possible minor-effect loci, such as yield, heterosis, GCA, and SCA, which may also involve interactions between the two parental genomes. For quantitative traits that are determined by multiple major-effect QTLs, the haplotypes associated with the QTLs are identified using an association or linkage population. Then, dozens of high-efficacy tagging SNPs for each desired trait are fed into the machine learning models to predict specific quantitative traits. Qualitative traits are caused by natural functional variations or genetically modified variations, such as sites for insect resistance, stress tolerance, and herbicide resistance

flowering time, ideal plant architecture, grain yield heterosis, general compatible ability (GCA), and SCA. By this means, candidate combinations of parental lines with high GCA and SCA of yield are selected for field trials. The second aspect is ML-based prediction to facilitate selection of specific quantitative traits unrelated to the genetic backgrounds. ML-based prediction uses a series of QTL haplotype panels and major-effect tag SNPs to infer correlations between genetic foregrounds and target traits, such as abiotic stress tolerance, pest and disease resistance, fertilizer intake efficiency, kernel dehydration rate, and high planting density. The third aspect is marker-assisted selection (MAS), which uses a small panel of SNP markers to screen for qualitative traits that are caused by natural functional variations or genetically modified variations. These qualitative traits include herbicide and insect resistance, high amylopectin content, thermo-sensitive male sterility, and drought stress tolerance created by CRISPR/Cas9 gene editing (Zhang et al. 2018).

GS has been successfully applied in modern breeding pipeline to predict $F_1$ phenotypes as the genotype of a $F_1$ hybrid can be inferred from the two parental genotypes. Thus, genotyping cost is manageable to be balanced with phenotyping cost to partition modeling and predicting populations. To employ GS or MAS on a $F_2$ population, each of $F_2$ seed harvested from $F_1$ plants needs to be genotyped, implemented by the seed-chipping technologies in which a small part of endosperm is sampled to extract DNA without affecting normal germination (Gao et al. 2008). With the synthetic use of high-throughput seed chipping, DNA extraction, and SNP genotyping automation platforms, seed-DNA genotyping system has become a popular solution to facilitate the GS or MAS to select seeds with desired genotypes before planting in multi-national seed companies.

## From genomic prediction to genome optimization

Crop breeding essentially consists of two steps of trait improvement. The first step is to increase genotype and phenotype diversity by population development, so that new phenotypes are created for artificial selection. During population development, exchange of chromosomal fragments can combine superior alleles, disrupt linkage between superior and inferior alleles, and amplify the frequency of rare superior alleles. As a result, a new genetic pool is formed that contains candidates with combinations of abundant superior alleles. These candidates are utilized in the second step of crop breeding, line selection. Line selection can be accomplished by either phenotype observation according to breeders' experience or by genomic selection based on G2P prediction. Both methods guide breeders in the creation of new inbred lines. "Genomic design breeding," assisted by varying scales of genomic prediction, aims to create a virtual blueprint so that breeders can perform the minimal number of hybridizations to create only candidate materials that combine the best alleles.

The rationale of "genome optimization" is to employ computational algorithms to simulate a virtual genome that possesses "optimal genotypes" composed of most of superior alleles to produce "optimal phenotypes." It is worth noting that the simulated, optimized genome may never be created in reality, considering the fact that many superior alleles and deleterious alleles may reside in the same linkage disequilibrium block which are difficult to break. The optimal phenotype for a target trait is the theoretical upper limit that a designated breeding population may generate. In single-cross hybrid breeding, the optimized genome represents the benchmark of the maximum potential heterosis utilization of grain yield that crossing of two heterotic pools may generate. In other words, a computationally optimized genome is an assembly of all most of superior alleles pyramided together to express superior phenotypes. Figure 4 outlines how genome optimization is implemented to efficiently achieve maximal yield improvement using a designated pool of breeding materials.

For example, assume a pool of 5000 maternal lines developed from multiple elite founder lines are crossed with one paternal tester to generate 5000 $F_1$ hybrids. The yield distribution of the 5000 $F_1$ hybrids is compared with the yield of a check hybrid, usually an elite variety with high yield. Of the 5000 hybrids, 250 (top 5%) exhibit higher yield than the check. Genome optimization is performed in five steps. In the first step, IBD (Identity-By-Descent) analysis is performed on the genotype data of the 5000 maternal lines, so that recombination hotspots are identified, and a recombination frequency map is generated. According to this map,

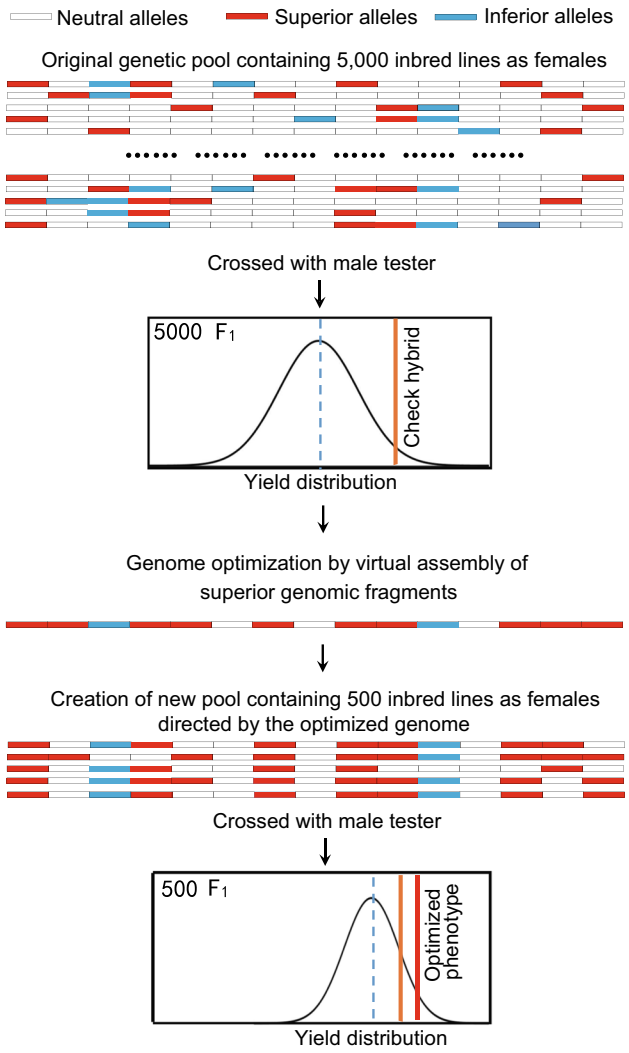


**Fig. 4** Genome optimization to facilitate maize breeding. The details of the genome optimization approach to direct line selection and population development are described in the main text

the maize genome is partitioned into $n$ IBD bins. Each bin represents one putative chromosomal fragment with a high frequency of genetic exchanges, in which $m$ tag SNPs are identified to represent the haplotype of the bin. In the second step, a genome-wide scan is performed to identify the haplotype most closely associated with the highest yield for each bin. The process is similar to a GWAS scan but is performed on each bin instead of on single SNPs. The IDs of the lines that contributed the best haplotype for the bin are recorded during scanning. In the third step, the best haplotypes identified from the $n$ bins are consecutively assembled based on their genomic positions to produce a simulated, optimized genome. In the fourth step, the 5000 maternal lines are ranked based on the number of bins that each maternal line donates to the optimized genome. Therefore, the more bins that a line donated to the genome, the more

superior alleles it likely possesses, and thus, it displays a greater potential of generating high-yield $F_1$ hybrids. In the fifth step, the genotypes and yield phenotypes of the 5000 $F_1$ hybrids are used as a training set to predict the optimal phenotype of the optimized genome using a GS model. As the optimized maternal genome is an assembly of the $n$ bins that donated the best haplotypes associated with high yield, virtually crossing the maternal line with the paternal tester is expected to produce an optimized genome with a yield greater than that of the check.

How is the optimized genome used to direct the next round of population development? The most important information contained in the optimized genome is the number of bins, or the percentage of genomic fragments in the optimized genome, that each maternal line donates to the assembly. The breeding value of each line is then evaluated based on the number of donated bins instead of the yield of its offspring. Most importantly, because the top-ranked maternal lines donate complementary sets of bins, breeders can use this information to select the top 5–10% of the maternal lines to generate new lines with minimal hybridization and with the maximal number of superior alleles for the next round of population development. Assuming that the newly developed population contains 500 new maternal lines, crossing with the same paternal tester may positively shift the overall yield distribution. In addition, the new group of 500 $F_1$ hybrids is expected to contain a higher proportion of hybrids whose yields surpass the check and perhaps have achieved the ideal yield phenotype.

## A proposed new breeding model incorporating DH production, GS, and genome optimization

Generation of pure inbred lines is essential for single-cross heterosis breeding. Doubled haploid (DH) breeding has become a popular approach in maize that has greatly accelerated creation of inbred lines and expansion of genetic pools (Longin et al. 2007; Prigge and Melchinger 2012; Ren et al. 2017). However, one obstacle restricting wide application of DH breeding is large-scale screening of DH lines for desired traits, because one round of DH production for one hybrid combination may generate hundreds or thousands of DH lines. However, only a small portion of DH lines are expected to carry the combination of target alleles from the two parents. An ideal solution would be to apply GS to screen DH lines using low-depth genotyping by sequencing (GBS) and select 5–10% of the lines as candidate lines for the subsequent test-crossing. If the cost of GBS is lower than five dollars per line (and possibly even less in the future), GS may eventually replace phenotype screening in the field. With a greatly reduced cost for screening of DH lines, multiple groups of GS-selected DH lines produced from different combinations may be test-crossed in parallel, which will greatly enhance breeding efficiency.

Genome optimization is particularly applicable for DH breeding, as one group of DH lines is developed from two parents, and it is easy to trace the exchanged fragments by IBD analysis and evaluate the contribution of each bin to the target trait. In addition, because the parental lines used for DH production are from one complete genetic pool, the original pool can be used as a general reference population to train the GS model. It is not difficult to imagine that in the future, the genomic design breeding pipeline will incorporate DH production, GS, and genome optimization, which will significantly shorten the breeding cycle and reduce

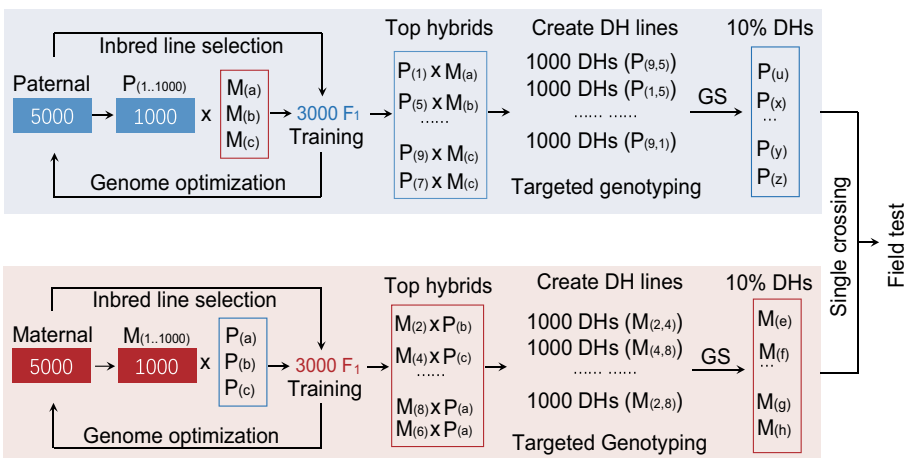


**Fig. 5** Combining DH breeding, GS, and genome optimization into a pipeline for improvement of maize breeding. The details of the proposed maize breeding pipeline are described in the main text

breeding costs. Such a model is especially suitable for small breeding teams, which account for ~85% of the breeding industry in China, to implement a small-scale breeding program with focused goals to solve specific, local problems. This proposed new breeding pipeline is shown in Fig. 5.

Assume hybridizations between one paternal pool and one maternal pool developed from two heterotic groups frequently generate superior heterosis performance for grain yield, and each pool contains 5000 candidate lines for selection. During the first round of line selection, 1000 paternal ($P_{(1\ldots1000)}$) and 1000 maternal ($M_{(1\ldots1000)}$) lines from the two heterotic pools are randomly selected to be hybridized with three elite maternal ($M_{(a,\,b,\,c)}$) testers and three elite paternal ($P_{(a,\,b,\,c)}$) testers, respectively. The two sets of 3000 $F_1$ hybrids form two initial training populations with known genotypes and measured phenotypes to build GS models and assemble optimized genomes within each heterotic pool. Based on the GCAs of yield computed from the three test-crosses in each training set, optimized genomes are assembled to identify the lines that donated a high percentage of the fragments in the optimized genome. The top-ranked lines that contributed abundant, complementary, superior fragments are hybridized with each other to generate hybrid offspring for DH production. Assume that each hybrid generates 500 DH lines, but not all of them are worth test-crossing in the field. The 500 DH lines are genotyped by GBS to predict their traits using the GS model trained by the two initial training populations. Based on the GS prediction, 5–10% of the DH lines are selected from each heterotic pool, and hybridization between the two sets of DH lines generates new $F_1$ hybrids for subsequent field testing.

As only 20% of the 5000 lines were sampled, the original pool may still harbor 80% of the unexploited superior alleles, thus requiring a second round of selection. The optimized genome simulated in the first round offers two layers of information to direct the second round: (1) the approximate number of superior lines remaining in the original pool and (2) the ideal phenotype predicted from the optimized genome establishes a "finish line" for the second round of selection and also represents the maximum potential of the original pool that can be exploited to improve the target trait. The second round begins with GS prediction of the remaining 4000 lines in each pool, trained by the 3000 $F_1$ hybrids. The trait for prediction may use either yield value or yield GCA computed from the three sets of test-crosses. Using the ideal phenotype as a reference, the top-ranked lines are further selected for the second round of hybridization with the three testers. The field-measured yield phenotype of the new $F_1$ hybrids is then merged with the previous training set to generate a new optimized genome. The result from the new assembly is then used to direct the second round of DH production. Two to three cycles of selection are expected to allow for the maximum potential of the two heterotic pools each containing 5000 lines to be fully exploited.

It is worth noting that the optimized genome virtually assembled by bin haplotypes positively correlated with yield of $F_1$ hybrids only reflect the accumulated additive effects of beneficial alleles. As illustrated in Fig. 4, the predicted $F_1$ yield based on optimized genotypes is approximately positioned at the 10% percentile in field-measured yield distribution. That means, additive accumulation of beneficial alleles may maximally explain 90% of genetic effects contributing to heterosis. The rest 10% of effects may attribute to complicated epistasis interactions and genotype by environment ($G \times E$) interactions that are difficult to be precisely modeled. Therefore, the optimized genome essentially represents an optimized genetic background to generate high GCA, but at the same time maximize the opportunity to generating superior SCA and adaptive G × E interaction. Furthermore, the proposed breeding pipeline represents a naïve model that only incorporates DH production using $F_1$ hybrids, and this pipeline is modifiable to be incorporated with rapid cycle recurrent selection and forward breeding. For instance, DH production may be also applied on $F_2$ hybrids to enhance recombination rates, or utilize a small panel of trait-associated markers in order to reduce the original DH lines.

## Concluding remarks

The history of crop breeding over the past 10,000 years can be described by three major eras: the 1.0 era of experience breeding, the 2.0 era of experimental breeding, and the 3.0 era of biological breeding. With rapid advances in biotechnology and information sciences, crop breeding is poised to evolve into the 4.0 era (Wallace et al. 2018). Breeding 4.0 will be defined as the era of "intelligent breeding," characterized by the integration of modern genomics, phenomics, gene editing, and synthetic biology, combined with A.I. technology to form a Big Data-driven, A.I.-supported, decision-making pipeline. In this review, we summarized the rationale for optimized genome design, which may be implemented as a new breeding model beyond genomic selection. In this approach, an optimized genome is virtually designed to contain all possible superior alleles donated from one designated genetic pool. The most significant advantage of the designed genome is its ability to guide breeders in carrying out a new round of population development in addition to selection of superior lines. The new round will utilize fewer founder lines that all carry superior alleles that are complementary with each other, such that the minimal number of hybridizations are performed to combine the maximum number of

superior alleles. Two or three rounds of small-scale population development are sufficient to rapidly achieve the trait improvement goal while exploiting the maximum potential of the genetic pool. We also propose a "genomic breeding design" pipeline for maize. This new pipeline will incorporate doubled haploid production, genomic selection, and optimized genome design, and represents a potential ideal solution for small-scale breeding focused on improvement of specific traits or local problems. Such a breeding model may be especially suitable for the Chinese maize breeding industry, which is mostly composed of small individual breeding teams. Future implementation of the suggested breeding pipeline will promote a revolutionary change in maize breeding from "art" to "science" and eventually to "intelligence" in the Breeding 4.0 era.

**Acknowledgements** This work was supported by the Key Research and Development Program of China (2018YFA0901003) and the National Science Foundation of China (31871706).

**Authors contribution** JY performed the comparative analysis of machine learning methods. QC and SJ performed the rrBLUP analysis of population stratification and population size. SJ, RF, and XW wrote the manuscript.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

Bouchez A, Hospital F, Causse M, Gallais A, Charcosset A (2002) Marker-assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162:1945–1959

Chen JZ (2010) Molecular mechanisms of polyploidy and hybrid vigor. Trends Plant Sci 15:57–71

Chen W, Wang W, Peng M, Gong L, Gao Y, Wan J, Wang S, Shi L, Zhou B, Li Z, Peng X, Yang C, Qu L, Liu X, Luo J (2016) Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat Commun 7:12767

Cooper M, Messina C, Podlich D, Totir R, Baumgarten A, Hausmann N, Wright D, Graham G (2014) Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci 65:311–336

Crossa J, Perez-Rodriguez P, Cuevas J, Montesinos-Lopez O, Jarquin D, delosCampos G, Burgueno J, Gonzalez-Camacho JM, Perez-Elizalde S, Beyene Y, Dreisigacker S, Singh R, Zhang X, Gowda M, Roorkiwal M, Rutkoski J, Varshney RK (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255

FAO (2011) Looking ahead in the world food and agriculture: perspective to 2050. FAO, Rome

Gao S, Martinez C, Skinner DJ, Krivanek AF, Crouch JH, Xu Y (2008) Development of a seed DNA-based genotyping system for marker-assisted selection in maize. Mol Breed 22:477–494

Ghanem ME, Marrou H, Sinclair TR (2015) Physiological phenotyping of plants for crop improvement. Trends Plant Sci 20:139–144

Guo T, Yu X, Li X, Zhang H, Zhu C, Flint-Garcia S, McMullen MD, Holland JB, Szalma SJ, Wisser RJ, Yu J (2019) Optimal designs for genomic selection in hybrid crops. Mol Plant 12:390–401

Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480

Hickey JM, Chiurugwi T, Mackay I, Powell W, Implementing Genomic Selection in CBPWP (2017) Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nat Genet 49:1297–1303

Houle D, Govindaraju DR, Omholt S (2010) Phenomics: the next challenge. Nat Rev Genet 11:855–866

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J, Wang B, Liu Z, Chen J, Li W, Zhang M, Xie S, Lai J (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44:812–815

Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, Han Y, Chai Y, Guo T, Yang N, Liu J, Warburton ML, Cheng Y, Hao X, Zhang P, Zhao J, Liu Y, Wang G, Li J, Yan J (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet 45:43–50

Li X, Li XR, Fridman E, Tesso TT, Yu J (2015) Dissecting repulsion linkage in the dwarfing gene Dw3 region for sorghum plant height provides insights into heterosis. Proc Natl Acad Sci USA 112:11823–11828

Li X, Guo T, Mu Q, Li X, Yu J (2018) Genomic and environmental determinants and their interplay underlying phenotypic plasticity. Proc Natl Acad Sci USA 115:6679–6684

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16:321–332

Longin CF, Utz HF, Reif JC, Wegenast T, Schipprack W, Melchinger AE (2007) 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 115:519–527

Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24:31–38

Ma C, Zhang HH, Wang X (2014) Machine learning for Big Data analytics in plants. Trend Plant Sci 19:798–808

Ma H, Li G, Wurschum T, Zhang Y, Zheng D, Yang X, Li J, Liu W, Yan J, Chen S (2018) Genome-wide association study of haploid male fertility in maize (*Zea mays* L.). Front Plant Sci 9:974

Piepho HP, Ogutu JO, Schulz-Streeck T, Estaghvirou B, Gordillo A, Technow F (2012) Efficient computation of ridge-regression best linear unbiased prediction in genomic selection in plant breeding. Crop Sci 52:1093–1104

Prigge V, Melchinger AE (2012) Production of haploids and doubled haploids in maize. Methods Mol Biol 877:161–172

Ren J, Wu P, Trampe B, Tian X, Lubberstedt T, Chen S (2017) Novel technologies in doubled haploid line development. Plant Biotech J 15:1361–1370

Ubbens JR, Stavness I (2017) Deep plant phenomics: a deep learning platform for complex plant phenotyping tasks. Front Plant Sci 8:1190

Voss-Fels KP, Cooper M, Hayes BJ (2019) Accelerating crop genetic gains with genomic selection. Theor Appl Genet 132:669–686

Wallace JG, Rodgers-Melnick E, Buckler ES (2018) On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. Annu Rev Genet 52:421–444

Wang XL, Wang HW, Liu SX, Ferjani A, Li JS, Yan JB, Yang XH, Qin F (2016) Genetic variation in ZmVPP1 contributes to drought tolerance in maize seedlings. Nat Genet 48:1233–1241

Webb S (2018) Deep learning for biology. Nature 554:555–557

Wing RA, Purugganan MD, Zhang Q (2018) The rice genome revolution: from an ancient grain to Green Super Rice. Nat Rev Genet 19:505–517

Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14:507–515

Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM (2018) Common disease is more complex than implied by the core gene omnigenic model. Cell 173:1573–1580

Xu Y (2016) Envirotyping for deciphering environmental impacts on crop plants. Theor Appl Genet 129:653–673

Yu X, Li X, Guo T, Zhu C, Wu Y, Mitchell SE, Roozeboom KL, Wang D, Wang M, Pederson GA, Tesso TT, Schnable PS, Bernardo R, Yu J (2016) Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat Plants 2:16150

Zhang Y, Massel K, Godwin ID, Gao C (2018) Applications and potential of genome editing in crop improvement. Genome Biol 19(1):210