frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 23 October 2023
DOI 10.3389/fpls.2023.1221750

Check for updates

OPEN ACCESS

EDITED BY
Harsh Raman,
NSW Government, Australia

REVIEWED BY
Seth C. Murray,
Texas A and M University, United States
Li Li,
University of New England, Australia

*CORRESPONDENCE
Sven E. Weber
✉ Sven.E.Weber@agrar.uni-giessen.de

RECEIVED 12 May 2023
ACCEPTED 05 October 2023
PUBLISHED 23 October 2023

CITATION
Weber SE, Chawla HS, Ehrig L, Hickey LT, Frisch M and Snowdon RJ (2023) Accurate prediction of quantitative traits with failed SNP calls in canola and maize.
Front. Plant Sci. 14:1221750.
doi: 10.3389/fpls.2023.1221750

COPYRIGHT
© 2023 Weber, Chawla, Ehrig, Hickey, Frisch and Snowdon. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Accurate prediction of quantitative traits with failed SNP calls in canola and maize

Sven E. Weber[1]*, Harmeet Singh Chawla[2], Lennard Ehrig[1], Lee T. Hickey[3], Matthias Frisch[4] and Rod J. Snowdon[1]

[1]Department of Plant Breeding, Justus Liebig University, Giessen, Germany, [2]Department of Plant Science, University of Manitoba, Winnipeg, MB, Canada, [3]Centre for Crop Science, Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St Lucia, QLD, Australia, [4]Department of Biometry and Population Genetics, Justus Liebig University, Giessen, Germany

In modern plant breeding, genomic selection is becoming the gold standard to select superior genotypes in large breeding populations that are only partially phenotyped. Many breeding programs commonly rely on single-nucleotide polymorphism (SNP) markers to capture genome-wide data for selection candidates. For this purpose, SNP arrays with moderate to high marker density represent a robust and cost-effective tool to generate reproducible, easy-to-handle, high-throughput genotype data from large-scale breeding populations. However, SNP arrays are prone to technical errors that lead to failed allele calls. To overcome this problem, failed calls are often imputed, based on the assumption that failed SNP calls are purely technical. However, this ignores the biological causes for failed calls—for example: deletions—and there is increasing evidence that gene presence–absence and other kinds of genome structural variants can play a role in phenotypic expression. Because deletions are frequently not in linkage disequilibrium with their flanking SNPs, permutation of missing SNP calls can potentially obscure valuable marker–trait associations. In this study, we analyze published datasets for canola and maize using four parametric and two machine learning models and demonstrate that failed allele calls in genomic prediction are highly predictive for important agronomic traits. We present two statistical pipelines, based on population structure and linkage disequilibrium, that enable the filtering of failed SNP calls that are likely caused by biological reasons. For the population and trait examined, prediction accuracy based on these filtered failed allele calls was competitive to standard SNP-based prediction, underlying the potential value of missing data in genomic prediction approaches. The combination of SNPs with all failed allele calls or the filtered allele calls did not outperform predictions with only SNP-based prediction due to redundancy in genomic relationship estimates.

KEYWORDS

genomic selection, genome structural variants, presence-absence variations, machine learning, SNP markers

1 Introduction

Genomic prediction has become the gold standard to identify genetically superior accessions within breeding materials. Henderson (1975) was among the first breeders to use relatedness based on pedigree information, along with phenotypic data, for breeding value prediction in a mixed linear model framework. Based on recent advances in genome sequencing technologies, genomic data is used today to replace pedigree relationships in statistical prediction models (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008). The increasingly accurate genome sequencing today allows the identification of millions of polymorphisms across the genome with high quality and confidence. Along with phenotypic measurements, these genotypic profiles can be used to predict the breeding values of non-phenotyped individuals with statistical models (Lande and Thompson, 1990; Meuwissen et al., 2001; VanRaden, 2008). These statistical methods utilize phenotypic and genotypic information from some genotypes (training population) to predict genotypes with only genotypic information. Over the years, several mathematical models have been proposed for genomic prediction; the commonly used models include GBLUP (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008), Reproducing Kernel Hill Regression (RKHS) (de los Campos et al., 2009), and models from the Bayesian alphabet like Bayesian LASSO (Park and Casella, 2008) or Bayesian ridge regression (Pérez and de los Campos, 2014). These models differ in their assumption of variance components, marker effects, marker modes of action, and model assumptions. More recently, machine learning algorithms have also been implemented for genomic prediction (Azodi et al., 2019; Pérez-Enciso and Zingaretti, 2019).

Genotypic information utilized for genomic prediction normally comprises biallelic single-nucleotide polymorphisms (SNPs) that are enormously abundant in eukaryotic genomes (Rafalski, 2002; Frazer et al., 2007; Ganal et al., 2009). Besides their high frequency, SNPs are not always able to explain all of the genetic variations, particularly for more complex traits, which tend to be characterized by "missing heritability" (Manolio et al., 2009; Forer et al., 2010). Genome structural variants (SV) are another type of genomic polymorphism that might explain some of this missing heritability (Manolio et al., 2009; Génin, 2020; Theunissen et al., 2020; Zhou et al., 2022). Plant genomes exhibit widespread SV including copy number variations, deletions, or insertions (Eichten et al., 2011; Fuentes et al., 2019; Gabur et al., 2019; Schiessl et al., 2019; Yang et al., 2019; Chawla et al., 2021), and because these are not always in linkage disequilibrium with neighboring SNPs, their effects are not always captured by the surrounding SNP variants (Gabur et al., 2018). However, such polymorphisms have been shown to be associated with a wide range of agronomical important traits (Gabur et al., 2018; Vollrath et al., 2021a; Vollrath et al., 2021b; Yuan et al., 2021). Specifically, it was shown that SVs are associated with disease resistance and flowering time in canola (Gabur et al., 2018; Gabur et al., 2020; Vollrath et al., 2021a; Vollrath et al., 2021b), disease resistance and boron toxicity tolerance in barley (Sutton et al., 2007; Muñoz-Amatriaín et al., 2013), pathogen response and aluminum tolerance in maize (Beló et al., 2010; Maron et al., 2013), and plant height and heading date in wheat (Li et al., 2012; Nishida et al., 2013), for example [for a comprehensive review, see Gabur et al. (2019)].

In large-scale breeding populations, SNPs are usually assessed with SNP arrays; however, these platforms are prone to technical errors that result in failed allele calls. Markers with a very high failed call rate are commonly discarded from downstream genetic analyses (Zhao et al., 2013; Lehermeier et al., 2014; Werner et al., 2018a; Knoch et al., 2021). For the remaining markers, failed allele calls need to be imputed to avoid large numbers of missing data points for further genetic studies. There are numerous methods to impute missing allele calls, with the simplest being the population mean/median (Endelman, 2011; Covarrubias-Pazaran, 2016; Covarrubias-Pazaran, 2018) or more advanced algorithms like "BEAGLE", "SHAPEIT", and "IMPUTE2" (Browning and Browning, 2007; Howie et al., 2011; Delaneau et al., 2012; Browning et al., 2018) which rely on allele frequencies, haplotypes, and flanking marker information. Regardless of the approach, imputation assumes that each missing marker call represents a genuine technical error. However, using whole-genome sequencing and patterns of inheritance in structured populations, Gabur et al. (2018) have demonstrated that, in complex crop genomes, missing allele calls can often be caused by polymorphic presence–absence variations resulting from deletions of sequences spanning SNP loci. Omitting or imputing failed allele calls can hence obscure valuable marker–trait associations. Commonly, SNPs with excessive failed calls are frequently eliminated from new iterations of genotyping arrays because they are considered technically unreliable (Boichard et al., 2012; Bayer et al., 2017). This can lead to considerable loss of potentially important genotype information and false imputations.

Whole-genome long-read sequencing data can be used to accurately identify structural variants (Francia et al., 2015; Dumschott et al., 2020; Chawla et al., 2021), enabling the validation of presence–absence variations detected in SNP array data (Gabur et al., 2018). However, genotyping a whole breeding population with thousands of genotypes *via* whole-genome long-read sequencing is economically not feasible. Targeted long-read sequencing of agronomically interesting genomic regions using ReadUntil (Edwards et al., 2019) might provide an alternative, which is a financially viable approach to identify genome structural variations at the population scale. However, application at scale in a breeding program may still be challenging. Furthermore, SNP arrays are well established as one of the main methods of choice for breeders to genotype their populations, hence the detection of presence–absence variations using these arrays comes at no additional cost. Most published work, to date, linking structural variants to quantitative traits have focused on association studies (see Gabur et al. (2019) for a detailed review). Only few studies have investigated their use for genomic prediction (Hay et al., 2018; Lyra et al., 2019; Chen et al., 2021; Knoch et al., 2021; Lamb et al., 2021), most of which utilize structural variants called from long- or short-read sequencing data. The aim of this study was to examine the value of potential presence–absence variants in the form of failed allele calls from SNP arrays in genomic predictions. To our knowledge, previously, this has only been done in association studies (Gabur et al., 2018; Gabur et al., 2020; Vollrath et al., 2021a; Vollrath et al., 2021b), making this the first attempt to utilize failed allele calls in genomic prediction. Specifically, the following questions were addressed: (1) How predictive are failed allele calls in genomic prediction and (2) can the addition of failed allele call information to standard SNPs improve genomic prediction accuracy? To answer these questions, published datasets from maize and canola were utilized for genomic predictions based on failed allele calls and genome-wide SNP markers, respectively. Prediction accuracy from cross-validation was subsequently used to assess marker–trait associations. Genomic prediction was performed with GBLUP, Bayesian LASSO, EGBLUP, RKHS, and Gradient Boosting and Support Vector Machines. Furthermore, two naive methods were developed and deployed to select failed allele calls based on population information. Using failed allele calls as indicators for presence–absence events, we show that these are as predictive as standard SNP markers for agronomic traits, underlining the potential information content of missing data in SNP arrays.

2 Materials and methods

2.1 Datasets

Two previously published datasets were examined in this study. The first was a canola dataset from a spring-type canola hybrid breeding program (Jan et al., 2016). Here two male sterile lines were crossed to 475 doubled-haploid (DH) pollinators to create 950 test crosses. The test crosses were subsequently tested for seed yield, flowering time, field emergence, lodging, oil content, oil yield, and glucosinolate content in a multi-environment trial at four different locations in 2 years. All parental lines were genotyped with the Illumina *Brassica* 60 k SNP array (Clarke et al., 2016). In total, 910 test crosses with complete phenotypic and genotypic records are available. The phenotypic data was published on an adjusted trait mean per genotype.

The second dataset represent two nested association mapping (NAM) populations of Flint and Dent maize. The population consists of 10 Dent and 11 Flint half-sib DH families. The lines were evaluated as test crosses, the DH Dent lines were all crossed to a single Flint tester line (UH007), and all DH Flint lines were crossed to a single Dent tester (F353). All DH lines were genotyped with the Illumina MaizeSNP50 SNP array (Ganal et al., 2011). This population was first described in Bauer et al. (2013), while Lehermeier et al. (2014) published phenotypic data from four locations for the Dent panel and at six locations for the Flint panel, including dry matter yield (DMY), dry matter content (DMC), plant height (PH), days till tasseling (DtTAS), and days till silking (DtSILK). The published field data was adjusted independently in the Flint and Dent pool, following the methods of the original publication. In total, complete phenotypic and genotypic data were available for test crosses from 847 Dent maternal lines and 918 Flint maternal lines.

2.2 Genotypic data

SNP matrices were filtered to remove markers with non-unique positions (multiple BLASTn hits of flanking sequences) on reference genomes. In canola, we utilized the *Brassica napus* Express 617 genome v2 (Lee et al., 2020) and in maize the B73 AGPv2 genome (Schnable et al., 2009). The genotypic data for the two maize pools was filtered jointly as one population. Compared to standard filtering pipelines, which removed SNPs with a certain proportion of failed calls, we treated failed SNP calls as third allele. In the first step, the coding for the original marker matrix was A/A, A/B, B/B, and F/F (“homozygous missing/failed allele”). Consequently, in this set, the markers were filtered according to an expected ≥0.095 (treating F as third allele), which corresponds to a minor allele frequency ≥0.05 in a biallelic case. From that, two copies of this matrix were created, one corresponding to the standard SNPs and one corresponding to the failed allele calls.

The copy corresponding to standard SNPs was then phased and imputed with the software “BEAGLE V5.2” (Browning and Browning, 2007; Browning et al., 2018). Subsequently, the markers were filtered for minor allele frequency ≥0.05 (to rule out monomorphic markers which could arise after imputation) and converted into numeric format (0, 1, 2 for A/A, A/B, and B/B).

The copy corresponding to the failed allele calls was recoded to successful call/successful call (regardless of allelic state) and F/F (“homozygous missing/failed allele”). This matrix was then also filtered for minor allele frequency ≥0.05 (to rule out monomorphic markers) and then converted into numeric format (0, 2 for successful call/successful call and F/F).

For canola, the processing resulted in 31,085 markers with successful allele calls and 7,169 markers with failed allele calls. In maize, we obtained 39,624 markers with successful allele calls and 8,024 markers with failed allele calls.

2.3 Population structure

For both datasets, the population structure was assessed by calculating the Euclidean distance between genotypes based on standard SNP markers and failed allele calls, respectively. Subsequently, the genotypes of each species were clustered into two subpopulations each using k-means clustering. A principal component analysis based on the genetic distance was conducted, and the first two principal components were utilized to visualize population stratification.

2.4 Methods to filter failed SNP calls with biological reasons

In the following two sections, we introduce two pipelines designed to distinguish between random failed allele calls and non-random systematic failed allele calls. This is done to strengthen the confidence that those failed allele calls stem from some biological reason, which hinders an allele call. These pipelines only rely on population measures and statistical tests.

2.4.1 Pool specificity

An important step in hybrid breeding is the creation of distinct genetic pools. Hence, the datasets assessed in this study naturally show a strong population structure corresponding to divergent genetic pools. In such populations, a proportion of alleles become pool-specific due to selection and genetic drift. On the other hand, technical errors can, by definition, not be pool specific; hence, they cannot show a bias between two different hybrid breeding pools. We thus assumed that there should be no relationship between subpopulation assignment and SNP call failure. In the breeding populations examined here, the populations for each species investigated split into two major gene pools. Hence, we expect that technical errors and successful allele calls should distribute equally in the two subpopulations. A χ^2 test of independence was utilized to test if there is an influence of subpopulation on allele call or failure. Pool assignment was based on k-means clustering with standard SNPs. Specifically, we tested for each failed allele call as follows:

- *H0*: failed allele call *versus* successful marker call and pool assignment is not related in the populations.
- *H1*: failed allele call *versus* successful marker call and pool assignment is related in the populations.

When *H0* is rejected, this is considered to be biological evidence for pool specificity of marker failure rather than a technical failure. Hence, we filter this failed allele call marker from the set of all failed allele calls and use it further in prediction models. After adjustment according to Benjamini and Hochberg (1995), the *p*-values were compared at a threshold of $\alpha = 0.05$.

2.4.2 Linkage disequilibrium

Linkage disequilibrium (LD) between markers on the same chromosome was calculated as r^2 (Hill and Robertson, 1968) in "SelectionTools" (http://population-genetics.uni-giessen.de/~software/), treating each failed allele call as an independent marker with the same genome position as its corresponding standard SNP.

If a failed marker call is purely due to a technical error, the failed call should not be in LD with any other marker. If the failed call is in considerable LD with markers on the same chromosome, we can assume that the failure is inherited together with other markers and the failure has a biological reason. Subsequently, a simple Student's *t*-test can be used to compare the LD patterns. If the LD of the failed marker with all other standard SNP calls on the same chromosome is considerably lower than its standard SNP counterpart, we can assume that the failure is due to a technical error. Specifically, for each failed marker call, we test the following hypotheses:

- *H0*: failed allele call and successful marker call show the same average LD to all standard markers on the same chromosome.
- *H1*: failed allele call and successful marker call show lower average LD to all standard markers on the same chromosome.

When *H0* is failed to reject, failed allele calls are considered to be in LD to markers on the same chromosome. Hence, we filter this failed allele call marker from the set of all failed allele calls and use it further in prediction models. After adjustment according to Benjamini and Hochberg (1995), the *p*-values were compared at a threshold of $\alpha = 0.05$.

2.5 Genomic prediction models

Six genomic prediction models were used to predict test cross performance. Two variations of GBLUP, two Bayesian methods, and two machine learning methods were used, covering parametric and non-parametric models. We applied standard GBLUP and extended GBLUP (EGBLUP) to account for second-order additive*additive epistasis (Jiang and Reif, 2015). Furthermore, we used the Bayesian LASSO model (Park and Casella, 2008) due to its capability for marker-specific shrinkage and the semiparametric model RKHS for modeling of higher-order epistasis (de los Campos et al., 2009). These approaches were complemented by the machine learning algorithms gradient boosting (Friedman, 2001) and support vector machines (SVM) (Boser et al., 1992).

In GBLUP and EGBLUP, the underlying mixed linear model is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_a\boldsymbol{a} + \boldsymbol{Z}_i\boldsymbol{i} + \boldsymbol{e}$$

where y is the vector of observations for a trait under consideration, $\boldsymbol{\beta}$ is the vector of fixed effects, a is the vector of random additive marker effects, i is the vector of random epistatic effects, and e is the random residual term. $\boldsymbol{Z}_a$ and $\boldsymbol{Z}_i$ are design matrices relating the random effects to the phenotypic records. X is the design matrix for fixed effects and, in the case of the canola dataset, a column of ones modeling the intercept and an additional column for the male sterile mother. In the maize datasets, X has a column of ones for the intercept and an additional 10 (Dent dataset) or 11 (Flint dataset) columns that assign individuals to half-sib families.

It is assumed that $\boldsymbol{a} \sim N(0, \boldsymbol{G}_a\sigma_a^2)$, $\boldsymbol{i} \sim N(0, \boldsymbol{G}_{aa}\sigma_{aa}^2)$ and $e \sim N(0, I\sigma_e^2)$, where σ_a^2, σ_{aa}^2, and σ_e^2 are additive genetic variance, epistatic genetic variance, and error variance, respectively. $\boldsymbol{G}_a$ and $\boldsymbol{G}_{aa}$ are the respective additive and epistatic relationship matrices, and $\boldsymbol{I}$ is an identity matrix. Depending on the inclusion of epistatic effects, the corresponding terms were included or omitted.

The additive genomic relationship matrix was calculated following VanRaden (2008):

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ}'}{2\sum \ p_i(1-p_i)}$$

In the case of prediction based on standard SNPs, the elements of $\boldsymbol{Z}$ are represented by (0-2p$_i$) for homozygous allele A, (1-2p$_i$) for the heterozygous state, and (2-2p$_i$) for homozygous allele B, with p$_i$ being the allele frequency of the B allele. For prediction based on all failed calls or filtered failed allele calls, the elements of Z are represented by ($0-2p_i$) for successful allele calls and ($2-2p_i$) for failed allele calls, with p_i being the allele frequency of the failed allele call. Furthermore, the combination of (i) SNPs and failed allele calls, (ii) SNPs and failed allele calls filtered by pool specificity, and (iii) SNPs and failed allele calls filtered by LD were considered.

A second-order (additive*additive) epistatic relationship matrix can be approximated with $\boldsymbol{G}_{aa} = \boldsymbol{G}\#\boldsymbol{G}$, where # denotes the pointwise (Hadamard) product operation (Henderson, 1985; Jiang and Reif, 2015).

All the mixed linear models described in this section were implemented and solved with the r package "sommer" (Covarrubias-Pazaran, 2016; Covarrubias-Pazaran, 2018), which also computes all model parameters including variance components.

The formula describing the Bayesian LASSO model, following Park and Casella (2008), is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Ma} + \boldsymbol{e}$$

where y is the vector of observations for a trait under consideration, β is the vector of fixed non-genetic effects, a is the vector of additive effects, X is the design matrix as described in the GBLUP section, and M is the incidence matrix relating phenotypic records with the respective marker. In standard SNP-based predictions, the elements of M are 0 for homozygous allele A, 1 for heterozygous, and 2 for homozygous allele B. In the case of prediction based on failed or filtered failed allele calls, the elements of M are 0 for a successful allele call and 2 for the failed allele call. Furthermore, we also considered the combination of (i) SNPs and failed allele calls, (ii) SNPs and failed allele calls filtered by pool specificity, and (iii) SNPs and failed allele calls filtered by LD. The coefficients of the fixed (β) effects are assigned flat priors, and the coefficients of the marker effects (a) are assigned double-exponential priors. This allows the shrinkage of some marker effects to effectively zero, introducing sparsity into the model. This model allows a stronger shrinkage of the marker effects, which may be useful especially for technical errors. Here e is the random residual term. This model was conducted in the r software with the package "BGLR" (Pérez and de los Campos, 2014), which computes all the model parameters. Default settings were utilized.

Following de los Campos et al. (2009) with kernel averaging, the RKHS model has the following form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \sum\nolimits_{l=1}^{L} \boldsymbol{u}_l + \boldsymbol{e}$$

with

$$p(\boldsymbol{\beta}, \boldsymbol{u}_1, \ldots \boldsymbol{u}_L, \boldsymbol{e}) \propto \prod\nolimits_{l=1}^{L} N(\boldsymbol{u}|0,\ \boldsymbol{K}_{ul}\sigma_{ul}^2)\ N(\boldsymbol{e}|0,\ \boldsymbol{I}\sigma_e^2)$$

where y is the vector of observations, while $\boldsymbol{K}_{ul}$ represents an $n{*}n$ kernel calculated based on the Euclidean distance between genotypes called (a) standard SNPs, (b) failed allele calls, (c) failed allele calls filtered by pool specificity, and (d) failed allele calls filtered by LD or a combination of (a) with (b), (c), or (d). The kernel was chosen to be a Gaussian kernel with the lth value of the bandwidth parameter {0.1, 0.5, 2.5}. $X\beta$ is treated in a similar manner to the Bayesian LASSO, and u_l is assumed to be random. That way, the different random effects, i.e., the three kernel matrices from the three bandwidth parameters, are weighted by their variance components. Again, e is the random residual term. This model was also conducted in the r software with the package "BGLR" (Pérez and de los Campos, 2014), which computes all the model parameters using the default setting of the package.

Gradient boosting sequentially builds ensembles of decision trees. The algorithm starts with an intercept estimation. Subsequently, it sequentially fits models on the residual of its predecessor (Friedman, 2001). The goal of each model is to minimize the prediction error of the previous model. Generally, the model can be described with following formula:

$$\boldsymbol{y} = \boldsymbol{1}\mu + \sum\nolimits_{m=1}^{M} \eta f_m(\boldsymbol{X}) + \boldsymbol{e}$$

where y is the vector of observations, μ is the overall intercept, and f is the base learning function, i.e., a decision tree. η is a shrinkage parameter, controlling the overall contribution of each decision tree to the total prediction. X is a matrix of (a) standard SNPs, (b) failed allele calls, (c) failed allele calls filtered by pool specificity, and (d) failed allele calls filtered by LD or a combination of (a) with (b), (c), or (d). Furthermore, in the case of the canola dataset, an additional column for the male sterile mother was added. In the maize datasets, an additional 10 (Dent) or 11 (Flint) columns were added that assign individuals to half-sib families. This model was conducted with the r package "xgboost" (Chen and Guestrin, 2016). Hyperparameters "eta", "gamma", "max_depth", "min_child_weight", "subsample", and "colsample_bytree" were optimized *via* Bayesian hyperparameter optimization using the r package "rBayesianOptimization" (Yan, 2022).

The SVM model performs a form of nonlinear regression; specifically, the ϵ-support vector regression (Chang and Lin, 2011) is utilized. It performs non-linear regression by projecting the data into higher dimensional space with a kernel function. This model was conducted with the r package "kernlab" (Karatzoglou et al., 2004), using the radial basis function as kernel function. Hyperparameters epsilon and cost were optimized with Bayesian hyperparameter optimization using the r package "rBayesianOptimization" (Yan, 2022). Prediction was based on the matrix of (a) standard SNPs, (b) failed allele calls, (c) failed allele calls filtered by pool specificity, and (d) failed allele calls filtered by LD or a combination of (a) with (b), (c), or (d). Furthermore, in the case of the canola dataset, an additional column was added for the male sterile maternal line, whereas for maize an additional 10 (Dent dataset) or 11 (Flint dataset) columns were added, which assign individuals to half-sib families.

2.6 Evaluation of prediction accuracy

The prediction accuracy for the two datasets was evaluated using fivefold cross-validation. The population was randomly divided into five equal-sized sets. In each fold, the prediction models were trained on four sets (training population), and then these trained models were utilized to predict the remaining set (validation population) with masked phenotypic data. This process was repeated until each set served as the validation population once. The accuracy was measured using the Pearson correlation coefficient (*r*) between the observed and predicted phenotypic values of the validation set in each fold. To ensure robustness, this entire procedure was repeated 30 times.

2.7 Genomic relationship

To assess how well relationship based on standard SNPs is also captured by one of the failed allele call marker sets, we used the relationship coefficients obtained from the relationship matrix calculated following VanRaden (2008) (see above) and calculated the Pearson correlation between relationship coefficients from SNPs and those from the failed allele calls.

2.8 Simulation

To test how high prediction accuracy with failed allele calls can get by chance, i.e., random association between failed calls (due to random technical problems of the array), a simulation was conducted. The basis of the simulation was the genotypic data described in Section 2.2. Here we took the imputed marker matrices as “true” genotypic data and simulated marker effects. In total, 100, 1,000, and 10,000 markers were sampled to serve as QTL. Subsequently, marker effects were sampled from a normal distribution with mean = 0 and variance = 1. The phenotype was then obtained by adding a random residual term to the total additive value of the individual. The residuals were sampled from a normal distribution with mean = 0 and variance = V_e. V_e was calculated as $\frac{V_g}{H^2} - V_g$, where V_g is the total genetic variance, i.e., variance of the breeding values, and H^2 is the heritability calculated as $H^2 = \frac{V_g}{V_g+V_e}$. Three heritabilities ($H^2$ = 0.4, 0.6, and 0.8) were simulated for each number of QTL.

According to the number of failed calls observed before imputation, 658,730 entries of the marker matrix in canola and 3,712,821 entries of the marker matrix in maize were randomly sampled to be failed calls and treated as described in Sections 2.2. and 2.4. In each simulation, genomic prediction was conducted with the GBLUP model based on SNPs and failed allele calls. Prediction accuracy was then measured with fivefold cross-validation with 10 repetitions (see Section 2.6). For each combination of number of QTL and heritability, 100 simulations were conducted to obtain a robust result.

3 Results

3.1 Canola

In canola, k-means clustering based on standard SNP markers revealed a considerable population stratification into two subpopulations/pools which we designated as pool A and pool B, respectively (Figure 1). The lines in pool A had, on average, 686.80 (median = 618.5) failed allele calls, while the lines in pool B had, on average, 848.21 (median = 767) failed allele calls (Supplementary Figure S1). The first three principal components based on standard SNPs together explain 23.25% of the variance in the marker data. On the other hand, the population structure based on failed allele calls also shows a distinction into two subpopulations based on k-means clustering; however, clustering did not result in the same subpopulation assignment compared to the standard SNPs (Figure 1). Here the first three principal components together explain 10.56% of the variance in the failed marker set. A visual inspection of the first two components of the two respective marker sets show a considerable overlap of the subpopulations.

Each possible failed allele call was tested for pool specificity. In canola, 1,989 failed allele calls showed significant pool specificity. The lines in pool A carry, on average, 302.26 (median = 283) pool-specific failed allele calls, and the lines in pool B carry, on average, 398.93 (median = 409) (Supplementary Figure S1). The LD of each possible failed allele call was compared to its standard SNP counterpart in both datasets. This resulted in 1,084 failed allele calls showing considerable LD with standard SNPs on the same chromosome. The lines in pool A carry, on average, 206.72 (median = 202) failed allele calls filtered by LD, while the lines in pool B carry 274.77 (median = 301) failed allele calls on average (Supplementary Figure S1). Subsequently, the markers filtered by the two methods described were utilized for the following analysis. Combining SNPs and all failed allele calls yields a total of 38,254 markers. When SNPs are combined with failed allele calls filtered by pool specificity, there are 33,074 markers. The combination of SNPs with failed allele calls filtered by LD results in a set of 32,169 markers.

An analysis of genomic relationships showed a high correspondence between the estimates of relationship based on standard SNPs, failed allele calls, and the two filtering methods (Figure 2). Correlations between the relationships based on SNPs and the three failed allele call sets were generally high in canola (Figure 2). The lowest correlation (*r* = 0.604) was observed between the SNP-based relationship and the relationship based on failed alleles (Figure 2). In contrast, stronger correlations were found between the SNP-based relationships and the failed allele calls filtered by pool specificity (0.786) or the failed allele calls filtered by LD (0.779), respectively (Figure 2).

Genomic prediction based on standard SNPs resulted in prediction accuracies ranging from 0.174 with SVM for field emergence to 0.813 with XGB for oil content (Supplementary Figure S2). Considerable differences could be observed between traits, while the differences between marker sets or prediction models were only very small (Figure 3; Supplementary Figure S2). Only in the trait field emergence did all other models considerably outperformed the two machine learning models SVM and XGB (Supplementary Figure S2). Across all models with standard SNPs, the prediction accuracy was lowest for field emergence, followed by lodging, seed yield, glucosinolate content, days to flowering, oil yield, and oil content (Figure 4; Supplementary Figure S2). The prediction accuracy based on failed allele calls was generally similar to the accuracy of standard SNP-based predictions for all traits

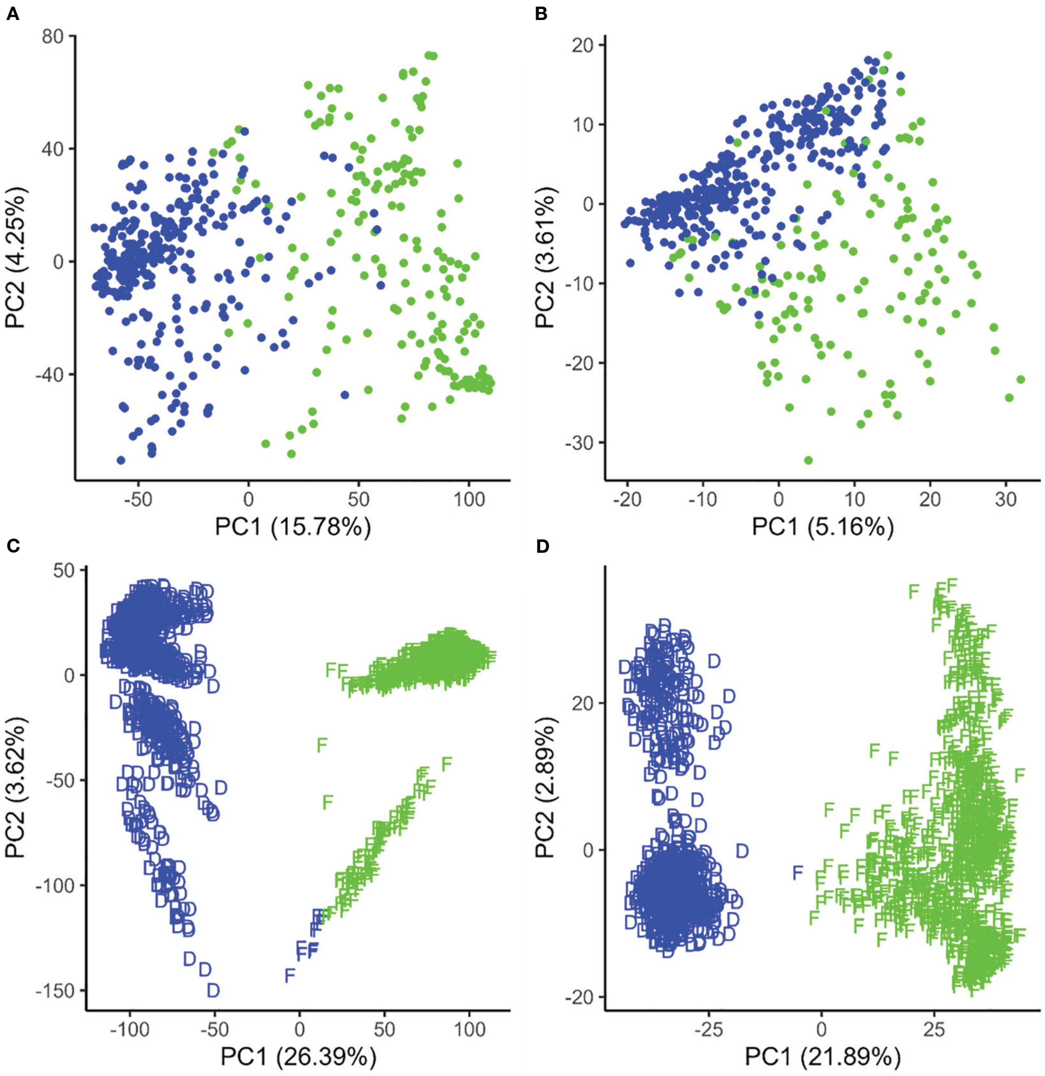

FIGURE 1
Population structure displayed by the first two principal components of the genetic distance in canola **(A, B)** and maize **(C, D)** based on standard single-nucleotide polymorphisms **(A, C)** and failed allele calls **(B, D)**. The colors (red and blue) represent clusters based on k-means clustering, while the D and F shapes in maize represent true Dent and Flint clusters.

(Figure 3; Supplementary Figure S2). When using markers from one of the methods to filter failed allele calls, the prediction accuracy did not improve compared to the prediction based on all failed allele calls. However, we also observed no further decrease in prediction accuracy (Figure 3; Supplementary Figure S2). When combining both (i) SNPs and failed allele calls, (ii) SNPs and failed allele calls filtered by pool specificity, and (iii) SNPs and failed allele calls filtered by LD, genomic prediction did not change compared to standard SNP-based prediction (Figure 3; Supplementary Figure S2).

3.2 Maize

In maize, k-means clustering based on standard SNP markers revealed a strong population stratification into two major groups that more or less correspond to the respective Flint and Dent pools (Figure 1). The lines in the Dent pool had, on average, 1,796.76 (median = 1,756) failed allele calls, while the lines in the Flint pool had on average 2,088.72 (median = 2,100) failed allele calls (Supplementary Figure S1). k-means clustering based on standard SNP markers assigned 10 genotypes of the Flint pool wrongly to the Dent pool (Figures 1, 3). Here the first three principal components together explain 33.03% of the variance in the marker data. The population structure based on failed allele calls also shows a strong distinction into two subpopulations. Clustering based on failed allele calls assigned only one genotype of the Flint pool incorrectly to the Dent pool (Figures 1, 4). The first three principal components cumulatively explain 27.38% of the variance in the failed marker set. A visual inspection of the first two principal components of the two respective marker sets did not show any overlap between the Flint and Dent pools (Figure 1).

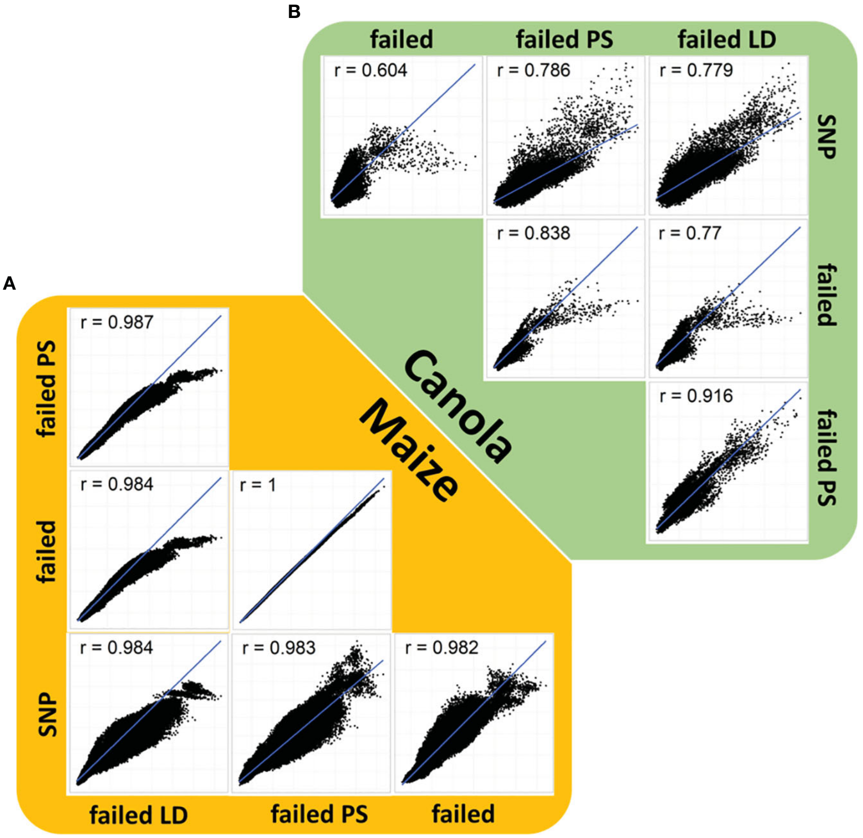

FIGURE 2
Correlation plot of genomic relationship coefficients based on single-nucleotide polymorphisms, failed allele calls (failed), failed allele calls filtered by pool specificity (failed PS), and failed allele calls filtered by LD (failed LD) in **(A)** canola (green) and **(B)** maize (orange).

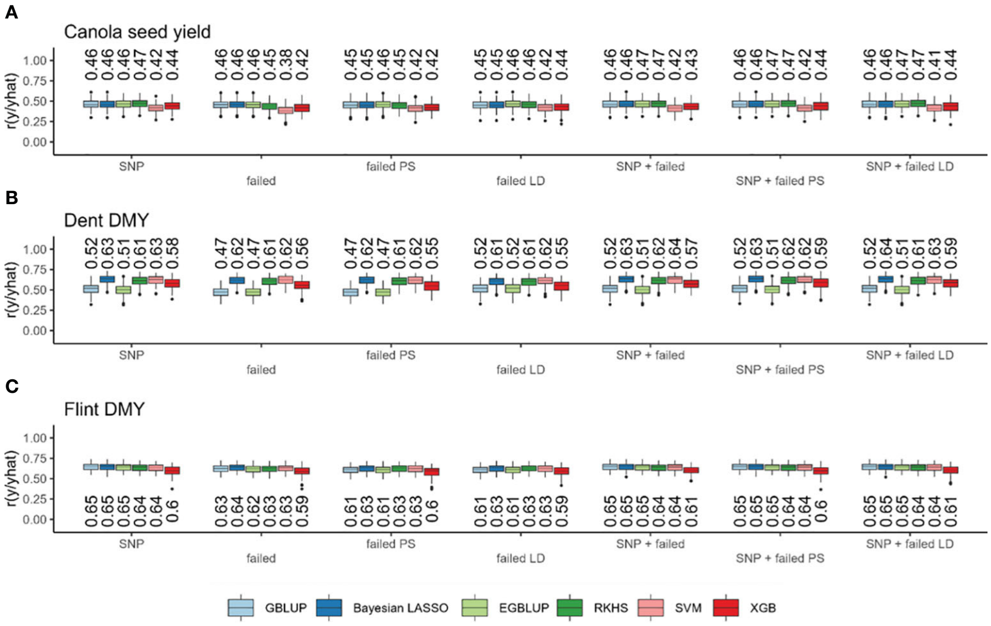

FIGURE 3
Prediction accuracy (*r*) based on standard single-nucleotide polymorphisms (SNPs), failed SNP calls (failed), failed SNP calls filtered by pool specificity (failed PS), and failed SNP calls filtered by LD (failed LD) as well as their combination with GBLUP (light blue), Bayesian Lasso (dark blue), EGBLUP (light green), RKHS (dark green), SVM (pink), and XGB (red). In canola seed yield **(A)**, maize Dent dry matter yield **(B)** and maize Flint dry matter yield **(C)**. Values above the boxplots represent median values across all cross-validation runs.

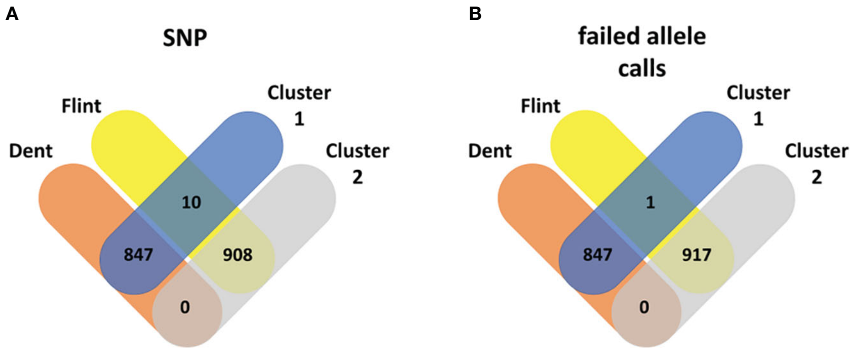

FIGURE 4
Venn diagram: Maize pool assignment to Dent (red) and Flint (yellow) subpools vs. pool assignment based on k-means clustering into cluster 1 (blue) and cluster 2 (gray) based on the genetic distance from standard single-nucleotide polymorphisms **(A)** and failed allele calls **(B)**.

Further subclusters could be seen in both the Flint and Dent pools which likely correspond to different families of the NAM population within the Dent and Flint material (Figure 1).

Testing each possible failed allele call for pool specificity showed that 7,286 markers with failed allele calls show pool specificity. The lines in the Dent pool carry, on average, 1,647.95 (median = 1,614) pool-specific failed allele calls, whereas the lines in the Flint pool carry, on average, 1,962.51 (median = 1,996) (Supplementary Figure S1). The LD-based method, on the other hand, filtered 2,156 failed allele calls that show considerable LD with standard SNPs on the same chromosome. Here the lines in the Dent pool carry, on average, 650.34 (median = 661) failed allele calls filtered by LD, while the lines in the Flint pool carry, on average, 913.88 (median = 949) (Supplementary Figure S1). Subsequently, the markers filtered by these two methods were utilized for the following analysis. The combination of SNPs and all failed allele calls yields a total of 47,648 markers. When we merge SNPs with failed allele calls filtered by pool specificity, there are 46,910 markers. Meanwhile, the combination of SNPs with failed allele calls filtered by LD produces a set of 41,780 markers.

An analysis of genomic relationships in maize showed high correlations between estimates of relationship based on standard SNPs, failed allele calls, and the two filtering methods (Figure 2). In maize, the lowest correlation (r = 0.982) detected was observed between the SNP-based relationship and failed allele calls (Figure 2). However, the difference to the correlations between standard SNPs and failed allele calls filtered by pool specificity (r = 0.984) or failed allele calls filtered by LD (r = 0.983) was considerably lower than the corresponding differences in canola (Figure 2). In all correlation plots of relationhip estimates, there were observable clusters corresponding to the strong distinction into genetically distinct pools (Figure 2).

3.2.1 Dent pool

Within the maize Dent pool, genomic prediction based on standard SNPs resulted in prediction accuracies in the range from 0.505 with EGBLUP for DMY to 0.850 with SVM for DMC. There were considerable differences between traits and models, while the differences between marker sets were only very small. With standard SNPs, the prediction accuracy across all models was lowest for DMY, followed by PH, DtSILK, DtTAS, and DMC (Figure 3; Supplementary Figure S3). Interestingly, GBLUP, EGBLUP, and XGB showed lower prediction accuracies compared to all other models across all traits, with the exception of PH (Figure 3; Supplementary Figure S3), for which XGB showed slightly higher prediction accuracies than GBLUP and EGBLUP (Supplementary Figure S3). Across traits, there was no consistent ranking between the remaining models Bayesian LASSO, RKHS, and SVM, with Bayesian LASSO yielding the highest prediction accuracy for DMY, PH, and DtTAS, whereas SVM yielded the highest prediction accuracy for DMC and DtSILK. Using all failed allele calls reduced the prediction accuracy only marginally, while the two alternative methods to filter failed allele calls gave a similar prediction accuracy compared to the use of all failed allele calls (Figure 3; Supplementary Figure S3). The combination of both (i) SNPs and failed allele calls, (ii) SNPs and failed allele calls filtered by pool specificity, and (iii) SNPs and failed allele calls filtered by LD in genomic prediction did not change the prediction accuracy compared to standard SNP-based prediction (Figure 3; Supplementary Figure S3).

3.2.2 Flint pool

Within the maize Flint pool, genomic prediction based on standard SNPs resulted in prediction accuracies in the range from 0.598 with XGB for DMY to 0.909 with GBLUP for DtSILK (Figure 3; Supplementary Figure S3). There were considerable differences again between traits and models. The differences between marker sets were only very small (Figure 3; Supplementary Figure S4). Across all models, the prediction accuracy based on standard SNPs was the lowest for DMY, followed by PH, DtSILK, DtTAS, and DMC (Figure 3; Supplementary Figure S4). Generally, the prediction accuracies obtained from XGB were among the worst across all traits, while GBLUP and EGBLUP showed considerably lower prediction accuracies only for DtTAS and PH (Figure 3; Supplementary Figure S4). Generally, the differences between models were much smaller in scale than the differences in prediction accuracy between traits (Figure 3; Supplementary Figure S3). The prediction based on failed allele calls reduced the prediction accuracy again only marginally. The two methods to filter failed allele calls did not improve the prediction accuracy compared to the prediction based on all failed allele calls. However, no large decrease in prediction accuracy could be observed. Combining both (i) SNPs and failed allele calls, (ii) SNPs and failed allele calls filtered by pool specificity, and (iii) SNPs and failed allele calls filtered by LD in genomic prediction did not change the prediction accuracy compared to standard SNP-based prediction (Figure 3; Supplementary Figure S3).

3.3 Simulation

Applying the filtering methods to the random failed allele calls within each simulation repetition only rarely yielded any failed allele call after filtering. If failed allele calls were left in the simulations, there were only up to two failed allele calls left after filtering. Consequently, we applied genomic prediction only with the complete set of failed allele calls in each simulation. Generally, the prediction accuracies based on SNPs for all simulated traits in both crops followed closely the simulated heritability, independent of the number of QTL. With failed allele calls, on the other hand, the prediction accuracy was close to zero across all simulation runs (Supplementary Figures S5–S7). It is worth to mention that, in many simulation cross-validation combinations, no genetic variance could be attributed to failed allele calls; hence, here only the intercept of the model contributed to the prediction (Supplementary Figures S5–S7).

4 Discussion

Utilizing data from three populations in two important crops, we show that failed allele calls can be informative to identify valuable genotype-trait associations in the context of genomic prediction. While the marker number was considerably decreased with failed allele calls compared to standard SNPs, the prediction accuracy was comparable. We developed two alternative pipelines to distinguish failed allele calls with a genuine biological cause from random technical errors. The markers obtained from those two pipelines yielded similar prediction accuracies compared to standard SNPs and to all failed allele calls despite a lower marker density. Therefore, regarding prediction accuracy in genomic prediction, there is no necessity for additional analysis of failed allele calls. Nevertheless, the two pipelines provided enhance the confidence that these failed allele calls arise from a non-random event, possibly attributable to a biological reason. The combinations of the different marker sets did not improve the prediction accuracy, which is likely due to the highly redundant estimation of genomic relationship. However, in cases where failed calls are caused by deletions that are not in LD with neighboring SNPs, it is plausible that they could contribute to improved trait prediction, just as they have been shown to do for QTL analysis [e.g., Gabur et al. (2018); Gabur et al. (2019)].

In both datasets investigated here, failed allele calls were very useful in identifying population structure and relationship, indicating a high relevance of presence–absence variation for population differentiation. Due to different marker filtering and distance calculation, the PCA and the clustering yielded different results in canola than in a previous study using the same dataset (Jan et al., 2016). Interestingly, the failed allele calls were more effective at the identification of present Flint and Dent maize material based on clustering. Sun et al. (2018) and Beló et al. (2010) revealed strong differences between genetically distant maize genotypes in the frequency of copy number variations. Furthermore, in both datasets, one of the two pools had higher average numbers of failed allele calls per line, which can also be observed with the two methods described to filter failed allele calls. This indicates a role of structural variation events underlying failed SNP calls in subpopulation (Gabur et al., 2018) or pool development.

There are several pipelines to detect copy number variations from SNP arrays relying on light intensity signals generated during a single base extension (Colella et al., 2007; Wang et al., 2007; Greenman et al., 2010; Xu et al., 2014; Grandke et al., 2017). However, in case of zero light signal, these pipelines cannot distinguish a genomic deletion from a technically failed allele call. Gabur et al. (2018) provide an alternate strategy to reliably identify genomic deletions using SNP array data. They used segregation patterns of failed allele calls in a nested association mapping population of *Brassica napus* to validate real deletions from technical artifacts of the SNP arrays. Several studies implemented this pipeline to filter and use large numbers of failed allele calls (Gabur et al., 2018; Gabur et al., 2020; Vollrath et al., 2021a; Vollrath et al., 2021b), which are normally removed from downstream analyses by a standard filtering process. However, the pipeline described in those studies cannot be applied in the present study since it relies on deviations from expected allele frequencies in segregating families, whereas the populations investigated here are genetically diverse breeding populations. Therefore, we used pool assignment and LD to filter failed allele calls. These two approaches can be applied to a wider range of populations as they do not need clear family structures while being simple and straightforward to implement. In canola, these two alternative methods delivered similar results: 1,989 failed allele calls filtered based on pool specificity and 1,084 failed allele calls filtered *via* analysis of LD. A pipeline to place markers with unknown chromosomal positions based on LD accurately placed 5,920 out of 21,251 unplaced markers (Yadav et al., 2021). Here with the LD-based filtering method, marker alleles are filtered rather than unplaced markers. The key advantage is that, rather than setting an arbitrary threshold, LD between markers on the same chromosome is used to set a dynamic threshold. Generally, the two pipelines that we developed consider any non-random cause for the allele call failure; however, they cannot classify the cause. While the cause for the allele call failure can have high importance in the detection of major QTL and causal genes, for genomic prediction of quantitative traits, the cause is less relevant as a single marker usually has only a small effect on the prediction (Tayeh et al., 2015;

van Binsbergen et al., 2015; Werner et al., 2018a; Werner et al., 2018b; e Sousa et al., 2019).

With the advancements in genotyping technology and the decreasing costs associated with it, genotyping by sequencing (GBS) has emerged as a promising alternative to SNP arrays for genotyping breeding populations (Poland and Rife, 2012; Kim et al., 2016; Chung et al., 2017). Unlike the closed architecture of SNP arrays, which typically only allows the identification of two alleles, GBS has the added advantage of detecting other variants, such as small deletions (Poland and Rife, 2012). This capability offers a potential solution to the aforementioned limitations by directly identifying the true variant at a given locus.

In the canola analysis, the genomic prediction accuracy based on all marker sets roughly corresponded to the original results of Jan et al. (2016). However, for all traits, a small improvement in prediction accuracy could be observed. Compared to Jan et al. (2016), we filtered for SNP markers with a fixed position on the reference genome Express 617 (Lee et al., 2020). Furthermore, we applied a different filtering method for allelic diversity; these together resulted in an additional 2,799 markers. The prediction accuracy across traits and marker sets generally did not deviate considerably from prediction accuracies reported in previous studies, although minor differences can be observed in field emergence and glucosinolate content (Würschum et al., 2014; Jan et al., 2016; Werner et al., 2018a; Werner et al., 2018b; Knoch et al., 2021).

In the maize analysis, the genomic prediction accuracy obtained from all marker sets corresponded to the original results of Lehermeier et al. (2014). The differences can be attributed to the considerably different cross-validation scheme that we used in comparison with the previous study. Furthermore, the different filtering, especially for allelic diversity, resulted in 5,508 more markers compared to the original publication. The accuracies were generally higher than in the canola analysis. As seen in the high prediction accuracies reported in other studies of hybrid prediction in maize (Technow et al., 2012; Crossa et al., 2014; Technow et al., 2014; Millet et al., 2019), we also observed generally high prediction accuracies for all traits and marker sets. Interestingly, the prediction accuracies varied between Flint and Dent datasets. For the traits DtSILK and DtTAS, the prediction accuracy was higher in the test crosses with Dent maternal lines than in the hybrids with Flint maternal lines. Moreover, the two models implemented in a frequentist framework, i.e., GBLUP and EGBLUP, delivered poorer predictions than the remaining models for all traits with the Dent test crosses. This behavior was not observed in the Flint or canola test crosses.

Importantly, predictions based on one of the three marker sets including failed allele calls always gave prediction accuracies competitive with standard SNP-based predictions. The simulation study indicates that this prediction accuracy seems to be not occurring by chance as the randomly sampled failed allele calls in the simulations resulted in a prediction accuracy close to zero. While failed allele calls were observed to be equally predictive as standard SNPs, it is essential to note that this might not directly translate to the entire germplasm of the given crop. This is because SNP arrays usually undergo thorough validation before being released for use. Of course, SNPs are influenced and linked to structural variations like deletions and insertions (Hinds et al., 2006; McCarroll et al., 2006; Redon et al., 2006; Gabur et al., 2018). Our analyses indicated that at least a proportion of the failed allele calls stem from structural variants. The two hybrid breeding crops maize and canola are known to be highly influenced by structural variants (Schnable et al., 2009; Springer et al., 2009; Beló et al., 2010; Lai et al., 2010; Swanson-Wagner et al., 2010; He et al., 2017; Samans et al., 2017; Hurgobin et al., 2018; Sun et al., 2018; Chawla et al., 2021). Furthermore, it is well known that structural variations like deletions, insertions, or inversions can be associated with agronomical traits (Würschum et al., 2015; Gabur et al., 2018; Gabur et al., 2019; Schiessl et al., 2019; Gabur et al., 2020; Vollrath et al., 2021a; Vollrath et al., 2021b) and differential gene expression (Shen et al., 2006; McHale et al., 2012; Tan et al., 2012; Chiang et al., 2017; Alonge et al., 2020). Hence, it can be assumed that the inclusion of SV data can improve the genomic prediction accuracy for some traits in crops; however, just like what is shown here, an improvement is not consistently observed (Hay et al., 2018; Lyra et al., 2019; Knoch et al., 2021). Furthermore, in cattle, only a marginal improvement in prediction accuracy was observed for important milk traits when accounting for structural variations from whole-genome sequencing (Chen et al., 2021).

Although machine learning has promising capabilities in genomic prediction (Montesinos-López et al., 2018; Pérez-Enciso and Zingaretti, 2019; Montesinos-López et al., 2021; Montesinos López et al., 2022), with encouraging results in human (Bellot et al., 2018; Lello et al., 2018), animal (González-Recio et al., 2010; Long et al., 2010; Gianola et al., 2011), and plant research (Heslot et al., 2012; Crossa et al., 2017; Montesinos-López et al., 2018; Azodi et al., 2019; Bayer et al., 2021), we failed to observe any fundamental advantage of two tested machine learning algorithms for any trait, population, or marker set. In contrast to the findings of González-Recio et al. (2010); Li et al. (2018), and Abdollahi-Arpanahi et al. (2020), we did not observe a competitive prediction accuracy of the boosting algorithm XGB in comparison to the other prediction models for 14 out of the 17 examined traits. This corresponds to the findings of Perez et al. (2022). Hyperparameter tuning is crucial for machine learning (Pérez-Enciso and Zingaretti, 2019; Zingaretti et al., 2020; Montesinos López et al., 2022). In this study, we applied a Bayesian hyperparameter optimization which, based on a given set of hyperparameter starting values, optimizes the hyperparameters sequentially with the objective of reducing the mean squared prediction error. It is possible that this optimization algorithm becomes obstructed in a local optimum, resulting in low prediction accuracies. However, it seems unrealistic that this would have occurred in every cross-validation run. Alternatively, the size of the training datasets that we used might be too small for machine learning models, which usually cope with $n > p$ problems (Azodi et al., 2019).

Incomplete LD between markers and QTL can lead to apparent or phantom epistasis. This can cause statistically significant marker interactions in association studies (Wood et al., 2014; de los Campos et al., 2019) and improved prediction accuracies with models considering epistasis (Schrauf et al., 2020). For predictions using only one of the failed marker sets, we need to assume the occurrence of considerable phantom epistasis due to the considerably lower marker number, which tends to result in lower LD between markers and QTL (Wood et al., 2014; de los Campos et al., 2019). For this reason, we extended the prediction portfolio from GBLUP and Bayesian LASSO to also include EGLUP and RKHS regression for explicit modeling of epistasis and the two machine learning methods SVM and XGB for modeling of nonlinear effects. However, models considering epistasis or nonlinear effects did not consistently outperform simple GBLUP or Bayesian LASSO in any of the failed marker sets. A possible explanation could be that, despite the reduced marker density, a sufficient proportion of QTL can nevertheless be covered by these markers. Indeed marker density can often be reduced without a considerable loss of prediction accuracy (de Roos et al., 2009; Zhang et al., 2019; Kriaridou et al., 2020). Besides co-segregation or LD between markers and QTL, another important factor impacting genomic prediction is the accurate estimation of relationship (Habier et al., 2010; Daetwyler et al., 2013; Habier et al., 2013). In fact, accurate pedigree information can already yield prediction accuracies that are comparable to predictions based on genomic information (Burgueño et al., 2012; Crossa et al., 2014; Deomano et al., 2020). The high correlations between relationship coefficients obtained from SNP markers and the three marker sets from failed allele calls show that information about failure of allele calls can be a good estimate for relationships between genotypes. The correlations between SNP markers and the three respective marker sets from failed allele calls were considerably lower in canola than in maize; however, losses in prediction accuracies were on a similar level in both species. Since SNPs are still only a fraction of all genetic information present on the genome, even SNPs are only able to "sample" a true relationship (Goddard et al., 2011), which could explain the comparable loss of prediction accuracy between the two datasets. However, the high correlation between relationship coefficients also explains the lack of gain in prediction accuracy, indicating that the information added by the failed SNP calls is at least partly redundant. In populations in Hardy–Weinberg equilibrium, this redundant information likely corresponds to SNPs within older deletions that are in LD with surrounding SNPs, whereas more recent structural variants leading to deletions (and failed SNP calls) are not always in LD with redundant SNPs and more likely to contribute additional information to predictions.

While we only observed marginal to no increases in prediction accuracy based on combinations of SNPs with failed marker calls, they may be especially beneficial in the context of association studies, where it has been shown that previously undetected QTL can be identified with the inclusion of failed SNP allele calls (Gabur et al., 2018). Furthermore, the analytical approaches applied here are straightforward to implement with no additional cost.

5 Conclusion

Our study confirms that failed allele calls from SNP array data can be highly predictive for agronomical traits in canola and maize. Based on population structure (pool specificity) and LD, we were able to distinguish random errors from systematic allele call failure, enabling the filtering of presence–absence marker data representing deletions with potential impacts on traits. In all examined traits and datasets, genomic prediction using presence–absence markers filtered from failed SNP calls was nearly as accurate as SNP-based prediction. This is likely due to the following: (a) capture of previously overlooked genomic regions, (b) accurate estimation of relationships (similar to SNP-based relationship), and (c) capture of dominance effects caused by deletions which differentiate between heterotic pools in hybrid breeding. However, prediction accuracy did not improve when combining SNP information with failed allele calls, which can be attributed to the high redundancy between estimates of genomic relationship. Nevertheless, we recommend the inclusion of information of allele call failure into genomic prediction, as it adds information that is potentially highly predictive for agronomic traits not always in LD with neighboring SNPs and is available to plant breeders using SNP array datasets for genotyping at no additional cost.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Author contributions

SW and RS designed the study. SW conceived the analysis, MF developed the software for LD calculation and supervised the statistical analysis. LE assisted with the statistical analysis. SW wrote the manuscript. RS, LH, and HC revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The work was funded by grant FKZ 031B0890A from the German Federal Ministry of Education and Research (BMBF) to MF and RS. The informatics infrastructure was provided by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics (de.NBI).

Acknowledgments

The authors thank Christian Obermeier and Philipp Heilmann for discussions on the potential applications of failed allele calls in plant breeding.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1221750/full#supplementary-material

References

Abdollahi-Arpanahi, R., Gianola, D., and Peñagaricano, F. (2020). Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genet. Sel Evol.* 52, 12. doi: 10.1186/s12711-020-00531-z

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., et al. (2020). Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* 182, 145–161.e23. doi: 10.1016/j.cell.2020.05.021

Azodi, C. B., Bolger, E., McCarren, A., Roantree, M., de los Campos, G., and Shiu, S.-H. (2019). Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3 Genes|Genomes|Genetics* 9, 3691–3702. doi: 10.1534/g3.119.400498

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., et al. (2013) Intraspecific variation of recombination rate in maize. In: *Genome Biology*. Available at: http://prodinra.inra.fr/record/256105 (Accessed June 28, 2022).

Bayer, M. M., Rapazote-Flores, P., Ganal, M., Hedley, P. E., Macaulay, M., Plieske, J., et al. (2017). Development and evaluation of a barley 50k iSelect SNP array. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01792

Bayer, P. E., Petereit, J., Danilevicz, M. F., Anderson, R., Batley, J., and Edwards, D. (2021). The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome* 14, e20112. doi: 10.1002/tpg2.20112

Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? *Genetics* 210, 809–819. doi: 10.1534/genetics.118.301298

Beló, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B., and Rafalski, A. (2010). Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* 120, 355–367. doi: 10.1007/s00122-009-1128-9

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Society. Ser. B (Methodological)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bernardo, R. (1994). Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Sci*. 34. doi: 10.2135/cropsci1994.0011183X003400010003x

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., et al. (2012). Design of a bovine low-density SNP array optimized for imputation. *PLoS One* 7, e34130. doi: 10.1371/journal.pone.0034130

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). "A training algorithm for optimal margin classifiers," in *COLT '92: Proceedings of the fifth annual workshop on Computational learning theory*, (New York, NY, Unite States: Association for Computing Machinery). 144–152. doi: 10.1145/130385.130401

Browning, S. R., and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* 81, 1084–1097. doi: 10.1086/521987

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015

Burgueño, J., de los Campos, G., Weigel, K., and Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Sci.* 52, 707–719. doi: 10.2135/cropsci2011.06.0299

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2 (27), 1–27:27. doi: 10.1145/1961189.1961199

Chawla, H. S., Lee, H., Gabur, I., Vollrath, P., Tamilselvan-Nattar-Amutha, S., Obermeier, C., et al. (2021). Long-read sequencing reveals widespread intragenic structural variants in a recent allopolyploid crop plant. *Plant Biotechnol. J.* 19, 240–250. doi: 10.1111/pbi.13456

Chen, T., and Guestrin, C. (2016). "XGBoost: A scalable tree boosting system," in *KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* (New York, NY, USA: Association for Computing Machinery). 785–794. doi: 10.1145/2939672.2939785

Chen, L., Pryce, J. E., Hayes, B. J., and Daetwyler, H. D. (2021). Investigating the effect of imputed structural variants from whole-genome sequence on genome-wide association and genomic prediction in dairy cattle. *Animals* 11, 541. doi: 10.3390/ani11020541

Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., Kim, Y., et al. (2017). The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699. doi: 10.1038/ng.3834

Chung, Y. S., Choi, S. C., Jun, T.-H., and Kim, C. (2017). Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Hortic. Environ. Biotechnol.* 58, 425–431. doi: 10.1007/s13580-017-0297-8

Clarke, W. E., Higgins, E. E., Plieske, J., Wieseke, R., Sidebottom, C., Khedikar, Y., et al. (2016). A high-density SNP genotyping array for Brassica napus and its ancestral diploid species based on optimised selection of single-locus markers in the allotetraploid genome. *Theor. Appl. Genet.* 129, 1887–1899. doi: 10.1007/s00122-016-2746-7

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., et al. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* 35, 2013–2025. doi: 10.1093/nar/gkm076

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* 11, e0156744. doi: 10.1371/journal.pone.0156744

Covarrubias-Pazaran, G. (2018). Software update: Moving the R package *sommer* to multivariate mixed models for genome-assisted prediction. *Genetics.* doi: 10.1101/354639

Crossa, J., Pérez, P., Hickey, J., Burgueño, J., Ornella, L., Cerón-Rojas, J., et al. (2014). Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity* 112, 48–60. doi: 10.1038/hdy.2013.16

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. *Trends Plant Sci.* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., and Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics* 193, 347–365. doi: 10.1534/genetics.112.147983

Delaneau, O., Marchini, J., and Zagury, J.-F. (2012). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

de los Campos, G., Gianola, D., and Rosa, G. J. M. (2009). Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation1. *J. Anim. Sci.* 87, 1883–1887. doi: 10.2527/jas.2008-1259

de los Campos, G., Sorensen, D. A., and Toro, M. A. (2019). Imperfect linkage disequilibrium generates phantom epistasis (& Perils of big data). *G3 Genes|Genomes|Genetics* 9, 1429–1436. doi: 10.1534/g3.119.400101

Deomano, E., Jackson, P., Wei, X., Aitken, K., Kota, R., and Pérez-Rodríguez, P. (2020). Genomic prediction of sugar content and cane yield in sugar cane clones in different stages of selection in a breeding program, with and without pedigree information. *Mol. Breed.* 40, 38. doi: 10.1007/s11032-020-01120-0

de Roos, A. P. W., Hayes, B. J., and Goddard, M. E. (2009). Reliability of genomic predictions across multiple populations. *Genetics* 183, 1545–1553. doi: 10.1534/genetics.109.104935

Dumschott, K., Schmidt, M. H.-W., Chawla, H. S., Snowdon, R., and Usadel, B. (2020). Oxford Nanopore sequencing: new opportunities for plant genomics? *J. Exp. Bot.* 71, 5313–5322. doi: 10.1093/jxb/eraa263

Edwards, H. S., Krishnakumar, R., Sinha, A., Bird, S. W., Patel, K. D., and Bartsch, M. S. (2019). Real-time selective sequencing with RUBRIC: read until with basecall and reference-informed criteria. *Sci. Rep.* 9, 11475. doi: 10.1038/s41598-019-47857-3

Eichten, S. R., Foerster, J. M., de Leon, N., Kai, Y., Yeh, C.-T., Liu, S., et al. (2011). B73-mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol.* 156, 1679–1690. doi: 10.1104/pp.111.174748

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. *Plant Genome* 4, 250–255. doi: 10.3835/plantgenome2011.08.0024

e Sousa, M. B., Galli, G., Lyra, D. H., Granato, Í.S.C., Matias, F. I., Alves, F. C., et al. (2019). Increasing accuracy and reducing costs of genomic prediction by marker selection. *Euphytica* 215, 18. doi: 10.1007/s10681-019-2339-z

Forer, L., Schönherr, S., Weissensteiner, H., Haider, F., Kluckner, T., Gieger, C., et al. (2010). CONAN: copy number variation analysis software for genome-wide association studies. *BMC Bioinf.* 11, 318. doi: 10.1186/1471-2105-11-318

Francia, E., Pecchioni, N., Policriti, A., and Scalabrin, S. (2015). "CNV and structural variation in plants: prospects of NGS approaches," in *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*. Eds. G. Sablok, S. Kumar, S. Ueno, J. Kuo and C. Varotto (Cham: Springer International Publishing), 211–232. doi: 10.1007/978-3-319-17157-9_13

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449, 851–861. doi: 10.1038/nature06258

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Stat* 29, 1189–1232. doi: 10.1214/aos/1013203451

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., Hoz, J. F. D., Mohiyuddin, M., et al. (2019). Structural variants in 3000 rice genomes. *Genome Res.* 29, 870–880. doi: 10.1101/gr.241240.118

Gabur, I., Chawla, H. S., Liu, X., Kumar, V., Faure, S., von Tiedemann, A., et al. (2018). Finding invisible quantitative trait loci with missing data. *Plant Biotechnol. J.* 16, 2102–2112. doi: 10.1111/pbi.12942

Gabur, I., Chawla, H. S., Lopisso, D. T., von Tiedemann, A., Snowdon, R. J., and Obermeier, C. (2020). Gene presence-absence variation associates with quantitative Verticillium longisporum disease resistance in Brassica napus. *Sci. Rep.* 10, 4131. doi: 10.1038/s41598-020-61228-3

Gabur, I., Chawla, H. S., Snowdon, R. J., and Parkin, I. A. P. (2019). Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.* 132, 733–750. doi: 10.1007/s00122-018-3233-0

Ganal, M. W., Altmann, T., and Röder, M. S. (2009). SNP identification in crop plants. *Curr. Opin. Plant Biol.* 12, 211–217. doi: 10.1016/j.pbi.2008.12.009

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., et al. (2011). A large maize (Zea mays L.) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PloS One* 6, e28334. doi: 10.1371/journal.pone.0028334

Génin, E. (2020). Missing heritability of complex diseases: case solved? *Hum. Genet.* 139, 103–113. doi: 10.1007/s00439-019-02034-4

Gianola, D., Okut, H., Weigel, K. A., and Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12, 87. doi: 10.1186/1471-2156-12-87

Goddard, M. E., Hayes, B. J., and Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *J. Anim. Breed. Genet.* 128, 409–421. doi: 10.1111/j.1439-0388.2011.00964.x

González-Recio, O., Weigel, K. A., Gianola, D., Naya, H., and Rosa, G. J. M. (2010). L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet. Res.* 92, 227–237. doi: 10.1017/S0016672310000261

Grandke, F., Snowdon, R., and Samans, B. (2017). gsrc: an R package for genome structure rearrangement calling. *Bioinformatics* 33, 545–546. doi: 10.1093/bioinformatics/btw648

Greenman, C. D., Bignell, G., Butler, A., Edkins, S., Hinton, J., Beare, D., et al. (2010). PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* 11, 164–175. doi: 10.1093/biostatistics/kxp045

Habier, D., Fernando, R. L., and Garrick, D. J. (2013). Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics* 194, 597–607. doi: 10.1534/genetics.113.152207

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., and Thaller, G. (2010). The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel Evol.* 42, 5. doi: 10.1186/1297-9686-42-5

Hay, E. H. A., Utsunomiya, Y. T., Xu, L., Zhou, Y., Neves, H. H. R., Carvalheiro, R., et al. (2018). Genomic predictions combining SNP markers and copy number variations in Nellore cattle. *BMC Genomics* 19, 441. doi: 10.1186/s12864-018-4787-6

He, Z., Wang, L., Harper, A. L., Havlickova, L., Pradhan, A. K., Parkin, I. A. P., et al. (2017). Extensive homoeologous genome exchanges in allopolyploid crops revealed by mRNAseq-based visualization. *Plant Biotechnol. J.* 15, 594–604. doi: 10.1111/pbi.12657

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423–447. doi: 10.2307/2529430

Henderson, C. R. (1985). Best linear unbiased prediction of nonadditive genetic merits in noninbred populations. *J. Anim. Sci.* 60, 111–117. doi: 10.2527/jas1985.601111x

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic selection in plant breeding: A comparison of models. *Crop Sci.* 52, 146–160. doi: 10.2135/cropsci2011.06.0297

Hill, W. G., and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoret. Appl. Genet.* 38, 226–231. doi: 10.1007/BF01245622

Hinds, D. A., Kloek, A. P., Jen, M., Chen, X., and Frazer, K. A. (2006). Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nat. Genet.* 38, 82–85. doi: 10.1038/ng1695

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 Genes|Genomes|Genetics* 1, 457–470. doi: 10.1534/g3.111.001198

Hurgobin, B., Golicz, A. A., Bayer, P. E., Chan, C.-K. K., Tirnaz, S., Dolatabadian, A., et al. (2018). Homoeologous exchange is a major cause of gene presence/absence variation in the amphidiploid Brassica napus. *Plant Biotechnol. J.* 16, 1265–1274. doi: 10.1111/pbi.12867

Jan, H. U., Abbadi, A., Lücke, S., Nichols, R. A., and Snowdon, R. J. (2016). Genomic prediction of testcross performance in canola (Brassica napus). *PLoS One* 11, e0147769. doi: 10.1371/journal.pone.0147769

Jiang, Y., and Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics* 201, 759–768. doi: 10.1534/genetics.115.177907

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab - an S4 package for kernel methods in R. *J. Stat. Software* 11, 1–20. doi: 10.18637/jss.v011.i09

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., and Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Sci.* 242, 14–22. doi: 10.1016/j.plantsci.2015.04.016

Knoch, D., Werner, C. R., Meyer, R. C., Riewe, D., Abbadi, A., Lücke, S., et al. (2021). Multi-omics-based prediction of hybrid performance in canola. *Theor. Appl. Genet.* 134, 1147–1165. doi: 10.1007/s00122-020-03759-x

Kriaridou, C., Tsairidou, S., Houston, R. D., and Robledo, D. (2020). Genomic prediction using low density marker panels in aquaculture: performance across species, traits, and genotyping platforms. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00124

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

Lamb, H. J., Hayes, B. J., Randhawa, I. A. S., Nguyen, L. T., and Ross, E. M. (2021). Genomic prediction using low-coverage portable Nanopore sequencing. *PloS One* 16, e0261274. doi: 10.1371/journal.pone.0261274

Lande, R., and Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743–756. doi: 10.1093/genetics/124.3.743

Lee, H., Chawla, H. S., Obermeier, C., Dreyer, F., Abbadi, A., and Snowdon, R. (2020). Chromosome-scale assembly of winter oilseed rape Brassica napus. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00496

Lehermeier, C., Krämer, N., Bauer, E., Bauland, C., Camisan, C., Campo, L., et al. (2014). Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. *Genetics* 198, 3–16. doi: 10.1534/genetics.114.161943

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D. H. (2018). Accurate genomic prediction of human height. *Genetics* 210, 477–497. doi: 10.1534/genetics.118.301267

Li, Y., Xiao, J., Wu, J., Duan, J., Liu, Y., Ye, X., et al. (2012). A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. *New Phytol.* 196, 282–291. doi: 10.1111/j.1469-8137.2012.04243.x

Li, B., Zhang, N., Wang, Y.-G., George, A. W., Reverter, A., and Li, Y. (2018). Genomic prediction of breeding values using a subset of SNPs identified by three machine learning methods. *Front. Genet.* 9. doi: 10.3389/fgene.2018.00237

Long, N., Gianola, D., Rosa, G. J. M., Weigel, K. A., Kranis, A., and González-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genet. Res.* 92, 209–225. doi: 10.1017/S0016672310000157

Lyra, D. H., Galli, G., Alves, F. C., Granato, Í.S.C., Vidotti, M. S., Bandeira e Sousa, M., et al. (2019). Modeling copy number variation in the genomic prediction of maize hybrids. *Theor. Appl. Genet.* 132, 273–288. doi: 10.1007/s00122-018-3215-2

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Maron, L. G., Guimarães, C. T., Kirst, M., Albert, P. S., Birchler, J. A., Bradbury, P. J., et al. (2013). Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci.* 110, 5241–5246. doi: 10.1073/pnas.1220766110

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., et al. (2006). Common deletion polymorphisms in the human genome. *Nat. Genet.* 38, 86–92. doi: 10.1038/ng1696

McHale, L. K., Haun, W. J., Xu, W. W., Bhaskar, P. B., Anderson, J. E., Hyten, D. L., et al. (2012). Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159, 1295–1308. doi: 10.1104/pp.112.194605

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819–1829. doi: 10.1093/genetics/157.4.1819

Millet, E. J., Kruijer, W., Coupel-Ledru, A., Alvarez Prado, S., Cabrera-Bosquet, L., Lacube, S., et al. (2019). Genomic prediction of maize yield across European environmental conditions. *Nat. Genet.* 51, 952–956. doi: 10.1038/s41588-019-0414-y

Montesinos López, O. A., Montesinos López, A., and Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Cham, Switzerland: Springer Nature). doi: 10.1007/978-3-030-89010-0

Montesinos-López, A., Montesinos-López, O. A., Gianola, D., Crossa, J., and Hernández-Suárez, C. M. (2018). Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 Genes|Genomes|Genetics* 8, 3813–3828. doi: 10.1534/g3.118.200740

Montesinos-López, O. A., Montesinos-López, A., Pérez-Rodríguez, P., Barrón-López, J. A., Martini, J. W. R., Fajardo-Flores, S. B., et al. (2021). A review of deep learning applications for genomic selection. *BMC Genomics* 22, 19. doi: 10.1186/s12864-020-07319-x

Muñoz-Amatriaín, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., et al. (2013). Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol.* 14, R58. doi: 10.1186/gb-2013-14-6-r58

Nishida, H., Yoshida, T., Kawakami, K., Fujita, M., Long, B., Akashi, Y., et al. (2013). Structural variation in the 5′ upstream region of photoperiod-insensitive alleles Ppd-A1a and Ppd-B1a identified in hexaploid wheat (Triticum aestivum L.), and their effect on heading time. *Mol. Breed.* 31, 27–37. doi: 10.1007/s11032-012-9765-0

Park, T., and Casella, G. (2008). The bayesian lasso. *J. Am. Stat. Assoc.* 103, 681–686. doi: 10.1198/016214508000000337

Perez, B. C., Bink, M. C. A. M., Svenson, K. L., Churchill, G. A., and Calus, M. P. L. (2022). Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. *G3 Genes|Genomes|Genetics* 12, jkac039. doi: 10.1093/g3journal/jkac039

Pérez, P., and de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442

Pérez-Enciso, M., and Zingaretti, L. M. (2019). A guide on deep learning for complex trait genomic prediction. *Genes* 10, 553. doi: 10.3390/genes10070553

Poland, J. A., and Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* 5, 92–102. doi: 10.3835/plantgenome2012.05.0005

Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5, 94–100. doi: 10.1016/S1369-5266(02)00240-6

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., et al. (2006). Global variation in copy number in the human genome. *Nature* 444, 444–454. doi: 10.1038/nature05329

Samans, B., Chalhoub, B., and Snowdon, R. J. (2017). Surviving a genome collision: genomic signatures of allopolyploidization in the recent crop species brassica napus. *Plant Genome* 10, plantgenome2017.02.0013. doi: 10.3835/plantgenome2017.02.0013

Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H. S., and Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop J.* 7, 127–140. doi: 10.1016/j.cj.2018.07.006

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Schrauf, M. F., Martini, J. W. R., Simianer, H., de los Campos, G., Cantet, R., Freudenthal, J., et al. (2020). Phantom epistasis in genomic selection: on the predictive ability of epistatic models. *G3 Genes|Genomes|Genetics* 10, 3137–3145. doi: 10.1534/g3.120.401300

Shen, J., Araki, H., Chen, L., Chen, J.-Q., and Tian, D. (2006). Unique evolutionary mechanism in R-genes under the presence/absence polymorphism in arabidopsis thaliana. *Genetics* 172, 1243–1250. doi: 10.1534/genetics.105.047290

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., et al. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PloS Genet.* 5, e1000734. doi: 10.1371/journal.pgen.1000734

Sun, S., Zhou, Y., Chen, J., Shi, J., Zhao, H., Zhao, H., et al. (2018). Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.* 50, 1289–1295. doi: 10.1038/s41588-018-0182-0

Sutton, T., Baumann, U., Hayes, J., Collins, N. C., Shi, B.-J., Schnurbusch, T., et al. (2007). Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 318, 1446–1449. doi: 10.1126/science.1146853

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D., et al. (2010). Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res.* 20, 1689–1699. doi: 10.1101/gr.109165.110

Tan, S., Zhong, Y., Hou, H., Yang, S., and Tian, D. (2012). Variation of presence/absence genes among Arabidopsis populations. *BMC Evolutionary Biol.* 12, 86. doi: 10.1186/1471-2148-12-86

Tayeh, N., Klein, A., Le Paslier, M.-C., Jacquin, F., Houtin, H., Rond, C., et al. (2015). Genomic prediction in pea: effect of marker density and training population size and composition on prediction accuracy. *Front. Plant Sci.* 6. doi: 10.3389/fpls.2015.00941

Technow, F., Riedelsheimer, C., Schrag, T. A., and Melchinger, A. E. (2012). Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *Theor. Appl. Genet.* 125, 1181–1194. doi: 10.1007/s00122-012-1905-8

Technow, F., Schrag, T. A., Schipprack, W., Bauer, E., Simianer, H., and Melchinger, A. E. (2014). Genome properties and prospects of genomic prediction of hybrid performance in a breeding program of maize. *Genetics* 197, 1343–1355. doi: 10.1534/genetics.114.165860

Theunissen, F., Flynn, L. L., Anderton, R. S., Mastaglia, F., Pytte, J., Jiang, L., et al. (2020). Structural variants may be a source of missing heritability in sALS. *Front. Neurosci.* 14. doi: 10.3389/fnins.2020.00047

van Binsbergen, R., Calus, M. P. L., Bink, M. C. A. M., van Eeuwijk, F. A., Schrooten, C., and Veerkamp, R. F. (2015). Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Selection Evol.* 47, 71. doi: 10.1186/s12711-015-0149-x

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91, 4414–4423. doi: 10.3168/jds.2007-0980

Vollrath, P., Chawla, H. S., Alnajar, D., Gabur, I., Lee, H., Weber, S., et al. (2021a). Dissection of quantitative blackleg resistance reveals novel variants of resistance gene Rlm9 in elite Brassica napus. *Front. Plant Sci.* 12. doi: 10.3389/fpls.2021.749491

Vollrath, P., Chawla, H. S., Schiessl, S. V., Gabur, I., Lee, H., Snowdon, R. J., et al. (2021b). A novel deletion in FLOWERING LOCUS T modulates flowering time in winter oilseed rape. *Theor. Appl. Genet.* 134, 1217–1231. doi: 10.1007/s00122-021-03768-4

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F. A., et al. (2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674. doi: 10.1101/gr.6861907

Werner, C. R., Qian, L., Voss-Fels, K. P., Abbadi, A., Leckband, G., Frisch, M., et al. (2018a). Genome-wide regression models considering general and specific combining ability predict hybrid performance in oilseed rape with similar accuracy regardless of trait architecture. *Theor. Appl. Genet.* 131, 299–317. doi: 10.1007/s00122-017-3002-5

Werner, C. R., Voss-Fels, K. P., Miller, C. N., Qian, W., Hua, W., Guan, C.-Y., et al. (2018b). Effective genomic selection in a narrow-genepool crop with low-density markers: Asian rapeseed as an example. *Plant Genome* 11, 170084. doi: 10.3835/plantgenome2017.09.0084

Wood, A. R., Tuke, M. A., Nalls, M. A., Hernandez, D. G., Bandinelli, S., Singleton, A. B., et al. (2014). Another explanation for apparent epistasis. *Nature* 514, E3–E5. doi: 10.1038/nature13691

Würschum, T., Abel, S., and Zhao, Y. (2014). Potential of genomic selection in rapeseed (Brassica napus L.) breeding. *Plant Breed.* 133, 45–51. doi: 10.1111/pbr.12137

Würschum, T., Boeven, P. H. G., Langer, S. M., Longin, C. F. H., and Leiser, W. L. (2015). Multiply to conquer: Copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. *BMC Genet.* 16, 96. doi: 10.1186/s12863-015-0258-0

Xu, L., Cole, J. B., Bickhart, D. M., Hou, Y., Song, J., VanRaden, P. M., et al. (2014). Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. *BMC Genomics* 15, 683. doi: 10.1186/1471-2164-15-683

Yadav, S., Ross, E. M., Aitken, K. S., Hickey, L. T., Powell, O., Wei, X., et al. (2021). A linkage disequilibrium-based approach to position unmapped SNPs in crop species. *BMC Genomics* 22, 773. doi: 10.1186/s12864-021-08116-w

Yan, Y. (2022). rBayesianOptimization: bayesian optimization of hyperparameters.

Yang, N., Liu, J., Gao, Q., Gui, S., Chen, L., Yang, L., et al. (2019). Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. *Nat. Genet.* 51, 1052–1059. doi: 10.1038/s41588-019-0427-6

Yuan, Y., Bayer, P. E., Batley, J., and Edwards, D. (2021). Current status of structural variation studies in plants. *Plant Biotechnol. J.* 19, 2153–2163. doi: 10.1111/pbi.13646

Zhang, H., Yin, L., Wang, M., Yuan, X., and Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Front. Genet.* 10. doi: 10.3389/fgene.2019.00189

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic prediction of hybrid wheat performance. *Crop Sci.* 53, 802–810. doi: 10.2135/cropsci2012.08.0463

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., et al. (2022). Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* 606, 527–534. doi: 10.1038/s41586-022-04808-9

Zingaretti, L. M., Gezan, S. A., Ferrão, L. F. V., Osorio, L. F., Monfort, A., Muñoz, P. R., et al. (2020). Exploring deep learning for complex trait genomic prediction in polyploid outcrossing species. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00025