Justus-Liebig-University, Gießen

Institute of Agronomy and Plant Breeding I

Department of Plant Breeding

Maximizing information content of SNP arrays for genomic prediction

Inaugural Dissertation for a Doctorate Degree in Agricultural Sciences

(Dr.agr.)

in the Faculty of Agricultural Sciences, Nutritional Sciences and

Environmental Management

Examiners:

Prof. Dr. Rod J. Snowdon

Prof. Dr. Matthias Frisch

Submitted by

Sven Ernst Weber

Gießen, 2023

"The scariest moment is always just before you start. After that, things can only get better."

Stephen King on writing

Contents

1 Introduction

1.1 Seeds of Progress: The basic concept of the plant breeding process

Plant breeding is a critical component of agriculture and plays a pivotal role in ensuring food security, sustainability and adaptability in the face of global challenges such as climate change and population growth (Voss-Fels et al., 2019). Besides being an important source of nutrients, plants are also an increasingly important resource for fiber and fuel, especially with growing demand for renewable energy and other resources (Herrmann, 2013; Shelar et al., 2023; Visković et al., 2023). Overall, plant breeding involves the controlled and deliberate manipulation of plant genetics to develop improved crop varieties with desirable traits (Bernardo, 2020). This process is crucial for increasing crop yield, resilience and nutritional quality, as well as reducing the need for chemical inputs like fertilizers or pesticides. The latter is especially important in the face of strongly declining insect populations, which comes with unforeseeable consequences.

Plant breeding plays an essential role in that process, by combining different desirable characteristics of different plant individuals (and their combinations of genes) into superior genotypes. A plant breeder utilizes populations of different genotypes and mates genotypes with desirable characteristics, in the hope that the resulting offspring carries as many favorable characteristics as possible from the respective parents. This process is repeated multiple times to arrive at a genotype that has an advantage over its ancestors and can then be released to farmers to increase their output (Bernardo, 2014). In summary, plant breeding is a multi-step process that involves the strategic gathering of genetic diversity, recombination of genetic material and selection of superior genotypes (Bernardo, 2014). Through these steps, plant breeders develop crop varieties with the desired characteristics. While this is a strong simplification of the breeding process, the process can be summarized into four crucial steps, as summarized in the following four sections.

1.1.1 Identification and introgression of novel or optimized genetic diversity into a breeding population

In this initial step, plant breeders identify and select different genotypes with specific desirable characteristics to combine or add into a breeding population. These characteristics may include traits like high yield, disease resistance, improved nutritional content or adaptability to certain environmental conditions. The goal is to gather a diverse pool of genetic material

that carries favorable alleles for all of these valuable traits. This creates a population with a wide range of genetic diversity, in which each individual possesses a subset of all desired traits.

1.1.2 Recombination among crossing partners

In the second step, the different genotypes in the breeding population are allowed to crossbreed or are deliberately crossed. This involves controlled cross-pollination to mix and recombine their genetic material. The objective is to create offspring with a combination of the favorable characteristics present in the parental plants. This recombination step can be thought of as nature-inspired genetic shuffling, as it introduces genetic diversity and variability into the offspring, increasing the chances of obtaining superior performance in target traits.

1.1.3 Selection and fixation of superior genotypes

The third and final step involves the rigorous evaluation and selection of the offspring. Plant breeders carefully assess the plants in the population, looking for individuals that exhibit the desired characteristics to a high degree. These selected plants are considered superior genotypes.

Once superior genotypes are identified, the breeder works to "fix" these traits, meaning they aim to ensure that the desired characteristics are consistently present in future generations. In facultative inbreeding plant species, this can be achieved through repeated selfing of selected genotypes until the genome is (nearly) completely homozygous and traits are accordingly stabilized. With the advance of haploid and tissue culture technologies, the repeated selfing can be replaced by production of completely homozygous doubled haploids. In either cae, the overall aim of this process is to create a new genotype that combines the desirable traits of its ancestors, making it a superior and well-adapted crop variety.

In obligate outcrossing plant species, traits must be fixed in heterozygous plants, either by assembling a mixture of phenotypically homogeneous genotypes (population breeding) or by clone propagation (clone breeding) because selfing of obligate outcrossers (if it is even possible) results in a strong inbreeding depression and diminished yields (Adam-Blondon et al., 2011).

1.1.4 Seed multiplication and release

After seed multiplication, which can be a limiting step in some crops (Adhikari et al., 2021), newly developed varieties can then be released to farmers for cultivation, with the expectation that they will lead to increased crop yield, resilience, and other desirable qualities. In clonally propagated plants, vegetative clones (e.g. tubers, cuttings) are released. This thesis focusses on breeding methods for self-fertile, seed-propagated seed crops, hence the focus is placed on breeding schemes involving selfing (or doubled haploids) and inbred line/hybrid development.

1.2 The genotypic value: What makes a good genotype?

A breeding program aims to enhance its breeding population concerning yield, resilience, and other desirable traits, however an important question is how these traits are defined. Because of limited seed availability in early breeding stages, breeders rely on phenotypic data from genotypes tested with only limited replications or in few environments. However, this evaluation is susceptible to the influence of various environmental factors in which the plant was cultivated. Consequently, a selected candidate must exhibit robust performance across a diverse range of growing environments. Therefore, the selection process should prioritize the genotypic value of a specific genotype, representing the mean value of the trait of interest across all relevant environments (Lynch and Walsh, 1998). This group of environments is commonly known as the target population of environments (TPE).

As the breeding program progresses and larger quantities of seeds become available, the performance assessment of a genotype is typically conducted in multi-environmental trials comprising the TPE. The genotypic value of a genotype can then be derived from these multi-environmental trials with designated field designs (Cochran and Cox, 1992). A model describing the data in terms of the overall mean, genotype, environment, genotype × environment interaction, and a residual term is employed to extract the genotypic value. This model can generally be expressed in the following formula:

$$y_{ik} = \mu + G_i + E_k + G_i \times E_k + e_{ik}$$

where y_{ik} is the phenotypic observation of a given genotype G_i grown in environment E_k, $G_i \times E_k$ represents the interaction effect of genotype G_i and environment E_k, while e_{ik} is the

residual, which is composed of all variation not captured by the aforementioned effects. This formula can be expanded to encompass not only macro-environmental influences E_k but also the impact of field designs carefully selected by the breeder. These design-related factors may encompass effects related to rows, columns or blocks within the field. A detailed explanation of this procedure can be found in comprehensive sources such as Nyquist and Baker (1991), Cochran and Cox (1992), Holland et al. (2002), Falconer and Mackay (2009) and Littell et al. (2013).

These models are nowadays defined and solved in a mixed linear model framework, which combines random and fixed effects. Briefly, fixed effects refer to factors for which all levels are known and can be estimated, whereas random effects are assumed to represent a random sample of factors or factor levels from the entire population of factors and factor levels (Cockerham, 1980). The decision regarding whether to treat a factor as random or fixed is made by the breeder depending on the specific implications they wish to derive (Piepho et al., 2008). Typically, most environmental effects and interaction effects are considered random, while the treatment of genotype effects as fixed or random may vary depending on the task at hand.

1.3 Evaluating the genotypic value: Heritability

In addition to defining the genotypic value, it is essential to quantify the proportion of phenotypic variation accounted for by differences between genotypes (i.e. genotypic values), and the accuracy of the field trial. For several decades, breeders have employed the heritability (H^2) for this purpose, which represents the fraction of genetic variance relative to the relevant phenotypic variance. The heritability can be expressed using the following formula:

$$H^2_{standard} = \frac{\sigma^2_G}{\sigma^2_G + {\sigma^2_{GE}}/{n_E} + {\sigma^2_e}/{n_E * n_R}}$$

where σ^2_G and σ^2_{GE} represent the variance attributed to genotype and genotype × environment interaction, respectively. σ^2_e denotes the residual variance. These terms can be derived from a mixed linear model, as detailed in the description of the genotypic value above. It is important to note that to obtain an estimate for σ^2_G, the genotype effects must be treated as

random. The terms n_E and n_R refer to the number of environments in which the genotypes were assessed and the number of replicates per genotype in the respective environments.

It is important to note that unbalanced data can arise from deliberate design choices or unforeseen environmental conditions, such as flooding, which might render certain field plots not harvestable. As a result, both n_R and n_E may not always be whole numbers and may vary among different genotypes. In such cases, applying the formula for $H^2_{standard}$ may prove impractical, leading to potentially misleading estimates. While a comprehensive exploration of this topic exceeds the scope of this thesis, one notable example can be highlighted, building on a method for estimating heritability in unbalanced settings, initially introduced by Holland et al. (2002) and further elucidated by Piepho and Möhring (2007). In this approach, heritability is calculated as:

$$H^2_{Piepho} = \frac{\sigma^2_G}{\sigma^2_G + {\bar{v}^{BLUE}_{\Delta..}}/{2}}$$

where $\bar{v}^{BLUE}_{\Delta..}$ denotes the mean variance of a difference between genotypic best linear unbiased estimates (BLUES), i.e. genotypic values or adjusted means, considering genotype as fixed effect in the mixed linear model described above. It can be shown that for the comparison of two genotypes, G1 and G2,

$$\bar{v}^{BLUE}_{\Delta G1-G2} = 2 * ({\sigma^2_{GE}}/{n_E} + {\sigma^2_e}/{n_E * n_R})$$

if the number of replicates is equal between genotypes, whereas

$$\bar{v}^{BLUE}_{\Delta G1-G2} = \left({\sigma^2_{GE}}/{n_E} + {\sigma^2_e}/{n_E * n_{R(G1)}}\right) + \left({\sigma^2_{GE}}/{n_E} + {\sigma^2_e}/{n_E * n_{R(G2)}}\right)$$

if the number of replicates differs between genotypes (Littell et al., 2013). Therefore, it is evident that in the balanced case, $H^2_{Piepho} = H^2_{standard}$. To calculate H^2_{Piepho}, both the mean variance of the differences between BLUES (i.e. genotypic values) and σ^2_G are required. Achieving this involves two distinct models, one considering the genotype as a random effect to obtain σ^2_G and the other treating the genotype as a fixed effect to derive $\bar{v}^{BLUE}_{\Delta..}$. A comprehensive review and comparison of heritability measures is provided by Schmidt et al. (2019).

Both heritability calculation methods highlight that an increase in n_E and/or n_R results in elevated heritability, enabling the acquisition of more accurate genotypic values and ultimately contributing to enhanced breeding success. However, expanding n_E or n_R implies additional plots in potentially diverse environments, but field plots often represent one of the costliest aspects in breeding programs. Therefore, increasing n_E or n_R may not always be a practical solution. A breeder must carefully evaluate whether augmenting environments or replications will genuinely enhance the effectiveness of the breeding program.

The emergence of high-throughput, high-resolution genotyping technologies provides an alternative avenue. Instead of measuring genotypes in field trials, breeders could leverage genomic information to infer implications for field performance and the genotypic value of a given genotype.

1.4 Incorporating genomics into breeding programs

The selection phase in plant breeding is of high significance as it functions as the cornerstone in the journey of enhancing breeding populations and developing superior crop varieties. Selection plays a pivotal role in shaping the genetic composition of future breeding populations, thus having substantial influence over the success of a breeding program. Traditionally, the selection decision was based solely on phenotypic observations, generally averaged across multiple environments and replications as described above. Collecting phenotype values can consume considerable resources in a breeding program. However, at least for some traits, the advent of DNA sequencing and marker technologies make it possible to change the way breeders select and improve crop varieties (Jannink et al., 2010; Hickey et al., 2014; Crossa et al., 2017). Breeders can leverage the entire genomic information available for a selection candidate, or a substantial portion of it, to make implications for specific traits. By doing so, genetic markers introduced a new era of precision and efficiency in plant breeding.

In the early 1980s, the first genetic markers were systematically employed to map genetic polymorphisms linked to trait expression, giving rise to the concept of quantitative trait locus/loci (QTL) (Davis and DeNise, 1998; Kumar, 1999; Collard et al., 2005). The introduction of restriction fragment length polymorphisms (RFLP) markers as tools for targeted trait modification through the identification of associated RFLP markers and the selection of

genotypes with desirable markers laid the foundation for marker-assisted selection (MAS) in plant breeding (Soller and Beckmann, 1983; Lande and Thompson, 1990; Bernardo, 1994, 1997). Indeed, MAS has proven successful in enhancing traits that are challenging or costly to phenotype, maintaining recessive alleles and combining multiple resistance genes against pests or diseases (Collard and Mackill, 2007; Xu and Crouch, 2008).

However, the majority of agronomic traits are inherently complex and influenced by a multitude of genes with small effects, making them subject to quantitative inheritance (Bernardo, 2020). Particularly in the case of quantitative traits, phenotypic expression is shaped by both genetic and non-genetic factors, as well as the intricate interplay between genotypes and the environment (Lynch and Walsh, 1998; Falconer and Mackay, 2009; Bernardo, 2020). With ongoing advancements in sequencing technology, especially through next-generation sequencing or third-generation sequencing, it has become feasible to sequence millions of polymorphisms across the entire genome and develop cost-effective technologies for high-throughput genotyping (Edwards and Batley, 2010; Yu et al., 2011, Edwards et al., 2013). This capability allows for high-resolution genotyping of entire breeding populations within a reasonable timeframe. However, conventional methods for mapping QTL are limited in their ability to identify a comprehensive set of responsible QTL, falling short of elucidating the extensive genetic basis underlying phenotypic expression (Manolio et al., 2009).

This constraint indicates that the identified QTL can account for only a portion of the overall genetic variance (Lande and Thompson, 1990; Meuwissen et al., 2001). Consequently, alternative approaches for establishing connections between genotypic markers and phenotypic traits were needed.

1.5 From QTL mapping to genomic prediction

As scientific capabilities transitioned from managing information on just a few genetic markers to handling thousands up to several million genetic markers simultaneously, there emerged a necessity for the development of new statistical pipelines. These pipelines are designed to efficiently cope with these extensive datasets and identify as much genetic variance as possible. A significant challenge emerges as one shifts from scenarios with more individuals than markers (often referred to as n>p problems) to more markers than individuals (often

referred to as n<p), rendering a least squares solution for effect estimation impractical. Additionally, the abundance of markers introduces the risk of multicollinearity, a statistical concept that arises when two or more independent variables in a regression model are highly correlated, complicating the determination of each variable's individual effect on the dependent variable (Barrie Wetherill et al., 1986).

One solution might be to utilize only significant markers identified through some form of QTL mapping. However, this approach nevertheless encounters the problems described earlier. Consequently, there arose a necessity to employ all available markers simultaneously in a unified regression model, aiming to capture as much genetic variance as possible.

Adopting all available markers in a mixed linear model and treating them as random variables was one of the initial solutions, commonly credited to multiple authors (Lande and Thompson, 1990; Bernardo, 1994; Whittaker et al., 2000; Meuwissen et al., 2001; Visscher et al., 2006; Habier et al., 2007; VanRaden, 2008). Essentially, the model can be outlined as follows:

$$y = X\beta + Wu + e$$

here, y is the vector of response variables (i.e. phenotypes, often genotypic values), X is the design matrix for the fixed effects (in the simplest case only a vector of effects modeling the overall mean), β represents the fixed effects associated with X, while e is the random residual term. W is a matrix of marker genotypes with genotypes in rows and markers in columns, whereas u is the associated random effect of each marker. With biallelic single nucleotide polymorphism (SNP) markers for example, W typically represents genetic information with numbers, for example 0 for genotype AA , 1 for genotype AB and 2 for genotype BB. This numbering is based on a randomly chosen reference allele (here B) at a given SNP locus. Alternatively, some representations of W use contrast coding, where AA is represented as -1, AB as 0 and BB as 1.

In general, it is assumed that y follows a multivariate normal distribution. Therefore, as u and e are considered the random effects, it is assumed that u follows a normal distribution with mean 0 and variance σ_u^2 (i.e. $u \sim N(0, \sigma_u^2)$), while e follows a normal distribution with mean 0 and variance-covariance matrix $I\sigma_e^2$ (i.e., $e \sim N(0, I\sigma_e^2)$).

The solution for this model can be obtained with the famous mixed model equations of Henderson (1950), which simultaneously estimates fixed effects and predicts random effects:

$$\begin{bmatrix}\widehat{\beta}\\ \hat{u}\end{bmatrix} = \begin{bmatrix}X^TX & X^TW\\ W^TX & W^TW + \lambda I_M\end{bmatrix}^{-1}\begin{bmatrix}X^Ty\\ W^Ty\end{bmatrix}$$

here, λ, referred to as the ridge parameter, is defined as the ratio between σ_e^2 and σ_u^2, both of which are nowadays estimated from the data using some form of restricted maximum likelihood (REML) procedure (Lynch and Walsh, 1998). In the literature, this procedure is referred to as ridge regression best linear unbiased prediction (rrBLUP), or sometimes also as random regression BLUP.

It is evident that the size of the matrix to be inverted here is directly proportional to the number of fixed effects and the number of markers. Consequently, solving this system computationally could become problematic if the marker count surpasses the number of individuals. Another historical approach, commonly known among plant breeders as the animal model due to its origins in animal breeding, is the genomic best linear unbiased prediction (GBLUP) (Henderson, 1975; Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008). In this method, rather than predicting marker effects, genotype effects are predicted. The linear mixed model mentioned earlier can thus be reformulated as follows:

$$y = X\beta + g + e$$

in which all notations are similar to those in the rrBLUP. However, g represents the random genotype effects with $g \sim N(0, G\sigma_g^2)$. G is a relationship matrix based on the marker information and is in the simplest case $G = WW'$, where W is the marker matrix as described above. In the literature, W and the simple form of G are usually subjected to some scaling and centering, using common methods like those of VanRaden (2008), Endelman and Jannink (2012) or Vitezica et al. (2017). Again, this mixed model can be solved with mixed model equations:

$$\begin{bmatrix}\widehat{\beta}\\ \hat{g}\end{bmatrix} = \begin{bmatrix}X^TX & X^TI\\ I^TX & I + G^{-1}\lambda\end{bmatrix}^{-1}\begin{bmatrix}X^Ty\\ y\end{bmatrix}$$

here, the parameter λ is defined as the ratio between σ_e^2 and σ_g^2, and both parameters are estimated from the data using a REML procedure. With GBLUP, the matrix which needs to be

inverted is proportional to the number of fixed effects and the number of individuals in y. Hence, GBLUP is preferrable when one encounters problems with more markers than individuals, whereas rrBLUP is preferable in situations with more individuals than markers.

It can be demonstrated that GBLUP and rrBLUP yield identical results and are indeed analogous (Habier et al., 2007). Moreover, employing basic algebra allows the derivation of marker effects from the GBLUP model using the following formula:

$$\hat{u} = W'G^{-1}\hat{g}$$

If the relationship matrix undergoes scaling, this adjustment must also be considered when transforming genotypes into marker effects. After solving the model and the prediction of marker or genotype effects, the model can be utilized to predict the genotypic values of individuals when only genotypic data is available.

In the context of GBLUP, one method for establishing the genetic relationship between individuals involves calculating genomic relationships based on marker information. Another approach relies on tracing the pedigree within a population, starting with a set of assumed non-related founder individuals. Intriguingly, a procedure similar to GBLUP was already employed by Henderson (1975), prior to availability of genetic marker data, who utilized relatedness based on pedigree information, along with phenotypic data, for breeding value prediction within a mixed linear model framework.

Over the years, various other mathematical models have been proposed for genomic prediction. Commonly utilized models include those previously mentioned, such as rrBLUP and GBLUP (Lande and Thompson, 1990; Bernardo, 1994; Whittaker et al., 2000; Meuwissen et al., 2001; Visscher et al., 2006; Habier et al., 2007; VanRaden, 2008). Additionally, alternative models like Reproducing Kernel Hill Regression (RKHS) (de los Campos et al., 2009) and Bayesian models, including Bayesian least absolute shrinkage and selection operator (LASSO) (Park and Casella, 2008) and Bayesian ridge regression (Pérez and de los Campos, 2014) have been introduced. These models distinguish themselves through differing assumptions about variance components, marker effects, marker modes of action and other model-related considerations. Furthermore, machine learning algorithms have also recently

been introduced and applied to genomic prediction (Azodi et al., 2019; Pérez-Enciso and Zingaretti, 2019).

1.6 The workflow in genomic prediction

Figure 1 describes the standard workflow in genomic prediction, which generally relies on three components: A training population, a statistical prediction model and a prediction population. The training population comprises phenotypic records (i.e. genotypic values) of various individuals, accompanied by their genotypic profiles, which are commonly composed of SNP markers. This population serves as the foundation for training a statistical prediction model, wherein model parameters are estimated based on observations within the training population. The prediction population may encompass the entire germplasm at a breeder's disposal, exclusively comprising the genotypic profiles of individuals for whom the breeder seeks phenotypic information. However, no phenotyping is carried out in the prediction population, enabling breeders to overcome financial constraints or limited seed availability.

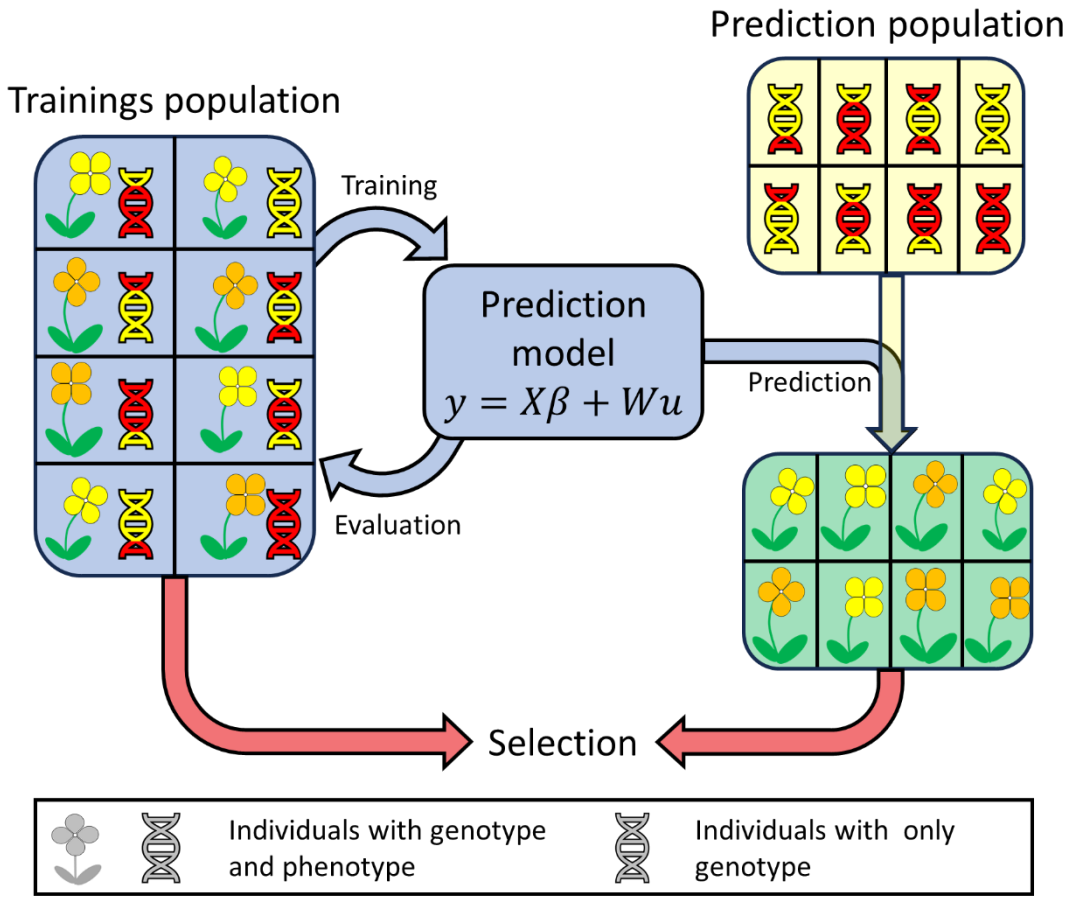

Figure 1 Schematic overview of the genomic prediction workflow. The training population with genotype and phenotype information is used to train a statistical prediction model. The trained statistical model is then utilized to predict the phenotypic information of individuals with available genotypic information. Boxes with blue background indicate information of the training population, and yellow information of the prediction population. After prediction, a breeder can select in both populations to further improve their breeding program (red).

Given the diverse options for selecting a statistical model, cross-validation is frequently employed to evaluate different models and to identify the most effective prediction model. During cross-validation, the training population is split into a part used for model training and a validation part reserved for evaluating the model accuracy. The trained model is then applied to predict the phenotypes of individuals in the validation part, relying only on their genotypic profiles. Prediction accuracy is subsequently assessed by comparing the actual phenotypes of individuals in the validation part with the predictions of the model. This process is frequently repeated with various allocations of individuals between the training and validation parts to guarantee an accurate estimation of prediction accuracy.

Following the identification of the model with the highest prediction accuracy, this model is trained on the entire training population. Subsequently, it is employed to predict the phenotypes of individuals within the prediction population. With phenotypic observations from the training population and predictions for the prediction population, a breeder can now utilize this information to select individuals from both populations to intercross for the next breeding cycle or to go into variety development.

1.7 The genotypic data base for genomic prediction

The most prevalent type of polymorphism across a genome are SNP variants, representing single DNA base variations at a common locus among different individuals. SNPs are abundantly present in eukaryotic genomes (Rafalski, 2002; Frazer et al., 2007; Ganal et al., 2011). Typically, SNPs are restricted to two alleles. However, because the heterozygous state can also be measured, individuals can be classified into three distinct groups for each locus. The simplicity, widespread distribution throughout the genome and low mutation rate of SNPs have established them as the marker of choice for various applications. Hence, in genomic prediction, the primary source of genotypic information still centers on SNPs (Crossa et al., 2017; Hickey et al., 2017; Werner et al., 2018; Jighly et al., 2023). However, despite their prevalence, SNPs do not always account for all the genetic variance, especially in the context of more complex traits that often manifest in what is sometimes referred to as "missing heritability" (Manolio et al., 2009; Eichler et al., 2010). Missing heritability can arise, at least in part, because biallelic SNPs may fall short in capturing all the variants and allelic

combinations of genes contributing to a specific trait, given that most genes exhibit multiple sequence polymorphisms.

Moreover, the accuracy of genomic prediction tends to be higher for closely related individuals (VanRaden, 2008; Hayes et al., 2009) and diminishes as the validation individuals become more distantly related (Habier et al., 2010; Wolc et al., 2011). This phenomenon suggests that SNPs may not always be in linkage disequilibrium (LD) with causal QTL. Furthermore, this means that the prediction accuracy relies, at least in part, on implicitly capturing relationships among individuals.

One strategy to enhance prediction accuracy involves increasing marker density. Thanks to advancements in whole-genome sequencing technologies, generating extensive marker datasets is now feasible for most crops (Edwards and Batley, 2010; Yu et al., 2011). However, it is important to note that increasing marker density does not consistently translate to improved prediction accuracies (Solberg et al., 2008; Druet et al., 2014; Hayes et al., 2014; Norman et al., 2018).

The standard tool for high throughput and low cost genotyping is a SNP array, usually manufactured by the commercial biotechnology companies Illumina Inc. (San Diego, CA, USA) or Affymetrix Inc. (Santa Clara, CA, USA). SNP arrays from these providers are readily accessible across a diverse array of crops and with various marker densities. In Germany, the genotyping service provider SGS INSTITUT FRESENIUS GmbH - TraitGenetics Section (Gatersleben) offers an array of genotyping services encompassing SNP arrays with varying marker numbers for diverse crops including wheat, brassicas, maize, barley, cotton, faba bean, sunflower, soybean, triticale and rye.

Due to the cost-effectiveness and high reproducibility of SNP arrays, they are commonly employed in large-scale breeding populations. The choice of which array and the number of markers to use depends on multiple factors, including the specific plant, its genome size, population characteristics and, of course, the breeders' preferences and budget constraints, as more markers generally result in higher genotyping costs. However, in the context of genomic prediction, the challenge of missing heritability may be exacerbated if the chosen SNP array fails to effectively tag all QTL, or if the SNP markers do not exhibit LD with relevant

chromosome segments. In such scenarios, the available markers limit the achievable prediction accuracy through genomic prediction.

1.8 Scope and objectives

1.8.1 Getting more out of a SNP array

Genomic prediction offers a great potential to rapidly accelerate crop genetic gain and improve populations in terms of yield and other traits (Bernardo, 2009, 2010; García-Ruiz et al., 2016; Gaynor et al., 2017). Despite the availability of alternative technologies such as DArT-seq (Jaccoud et al., 2001), genotyping by sequencing (Davey et al., 2011) and targeted sequencing (Bybee et al., 2011), SNP arrays are the method of choice in many commercial breeding programs, particularly for major crops, and they are used in most genomic prediction studies.

As previously mentioned, SNP markers assessed using arrays do not comprehensively represent all genetic variations. Consequently, breeders must optimize the information content of the existing SNP array for all relevant traits in the breeding program. This includes identifying more informative variants for predictions without necessarily increasing the marker density *per se*, especially when faced with budget constraints or limited marker options. With this in mind, the objective of this thesis is to evaluate methods which try to maximize the information content of SNP arrays for genomic prediction. The work specifically emphasizes and evaluates haplotype blocks (Figure 2A), structural variations identified through SNP arrays (Figure 2B), and the imputation of whole-genome sequencing marker data (Figure 2C) as options to enhance the information derived from SNP arrays for genomic prediction. This is accomplished by employing various genomic prediction models, including parametric, semi-parametric and machine learning models. Different cross-validation schemes are utilized to evaluate the prediction accuracy achieved with SNPs, along with methods aimed to increase the information content of SNP arrays across four distinct crops.

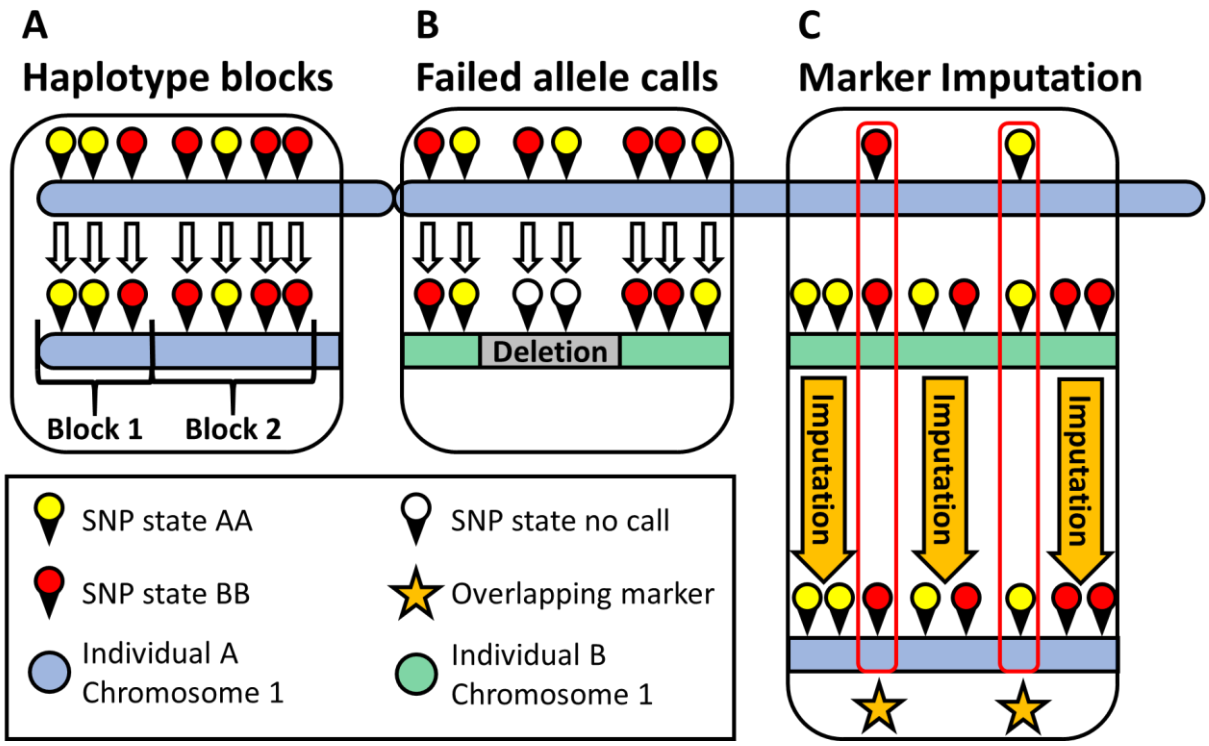

Figure 2 Graphical summary of methods to increase the information content of SNP arrays for genomic prediction. **A** Summarizing adjacent SNP marker into haplotype blocks based on some algorithm, **B** detecting structural variations, such as deletions, based on failed allele calls and **C** imputation of markers from whole-genome sequencing based on overlapping marker patterns between individuals genotyped with differing marker density

The following three paragraphs will provide a brief introduction to these methods, with more comprehensive insights and detailed results and discussions presented in the studies outlined in Chapters 2, 3, and 4.

1.8.2 Haplotype blocks

At its core, a eukaryotic genome is always structured into chromosomes, which serve as the primary organizational units of the genome. These long, thread-like structures are composed of DNA molecules arranged in the form of compact chromatin. In most eukaryote species, including humans and many plants, chromosomes exist in pairs, with one inherited from each parent. Loci on the chromosome are typically inherited in block-like structures along the genome (Gabriel et al., 2002) (Figure 2A). These structures are defined by only a few recombination within the blocks but with recombination hotspots at the block border. The resulting blocks are what is known as haplotype blocks (Figure 2A) (Daly et al., 2001; Jeffreys et al., 2001; Reich et al., 2001).

Defining haplotype blocks from marker data can be done in several ways. After assigning markers to chromosomal positions on a designated reference genome, they can be defined as a fixed window of adjacent markers or a fixed window of adjacent base pairs. These methods are straightforward but may not precisely represent the biological occurrence and borders of haplotype blocks. More sophisticated methods to construct haplotype blocks are designed to model the LD structure on the chromosome.

Haplotype blocks are assumed to be in stronger LD with QTL than SNPs. Furthermore, haplotype blocks can have potentially more alleles (known as haplotypes) than a biallelic SNP marker. Further, it has been demonstrated that haplotype blocks can effectively capture local epistasis among markers in close proximity (Jiang et al., 2018). Furthermore, they offer a potential solution to the issue of apparent or phantom epistasis, which can arise when markers and QTL are in incomplete LD (Wood et al., 2014; de los Campos et al., 2019). This suggests that haplotype blocks may have the capacity to improve genomic prediction by providing a more accurate representation of the genome across the chromosomes. The utilization of haplotype blocks in genomic prediction is evaluated and discussed in Chapter 2 (Weber et al., 2023b) along with a variety of methods and parameters for the construction of these blocks.

1.8.3 Structural variations

Genome structural variants (SVs) represent another category of genomic polymorphisms that may help to cope with “missing heritability” (Manolio et al., 2009; Génin, 2020; Theunissen et al., 2020; Zhou et al., 2022). Plant genomes, in particular, are known for their widespread SVs including copy number variants, deletions and 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). Importantly, SVs do not always consistently exhibit LD with neighboring SNPs, hence their effects cannot always be detected by surrounding SNPs (Gabur et al., 2018). Nonetheless, these kind of polymorphisms have been linked to a broad spectrum of agricultural traits (Sutton et al., 2007; Beló et al., 2010; Maron et al., 2013; Muñoz-Amatriaín et al., 2013; Gabur et al., 2018; Vollrath et al., 2021b; Yuan et al., 2021).

Although whole-genome long-read sequencing data can accurately pinpoint structural variants (Francia et al., 2015; Dumschott et al., 2020; Chawla et al., 2021), genotyping an entire breeding population with thousands of genotypes using whole-genome long-read sequencing

remains economically challenging. SNP arrays, on the other hand, are not inherently capable of directly identifying such variants. Furthermore, these platforms are susceptible to technical errors that can result in failed allele calls. Nevertheless, Gabur et al. (2018) demonstrated that in crop genomes, missing allele calls could often be attributed to polymorphic presence-absence variations stemming from large-scale deletions spanning SNP loci (Figure 1B). Thus, one strategy to potentially increase the information content derived from a SNP array is to utilize failed allele calls as proxies for structural variations, such as deletions, in genomic prediction. This approach has the potential to enhance prediction accuracy and will be further evaluated in Chapter 3 (Weber et al., 2023a).

1.8.4 Imputation of whole-genome sequencing data

As previously discussed, modern sequencing technologies enable the identification of millions of genetic variations, including genome-wide SNPs and SVs, within a single genome. However, when it comes to breeding populations, sequencing the entire population with sufficient coverage and accuracy is often cost-prohibitive. Nonetheless, as sequencing technologies have advanced, so too have the statistical pipelines used to process sequencing data. In particular, there are tools available today that can impute marker data for whole-genome sequencing in a population, even if only a subset of representative genotypes within that population were sequenced at the whole-genome level (Browning and Browning, 2007; Howie et al., 2009; Delaneau et al., 2012; Browning et al., 2018) (Figure C). The rest of the breeding population may have been genotyped with limited marker density, using for example a SNP array (Figure 1C). These imputation tools rely on various factors like chromosomal position, allele frequencies, haplotypes, LD and/or information from flanking markers to infer the allelic state of all markers in a genotype, even if those markers were not directly genotyped in every individual (Figure 2C). Commonly used tools for this imputation task include the software packages "BEAGLE" (Browning and Browning, 2007; Browning et al., 2018), "SHAPEIT" (Delaneau et al., 2012) and "IMPUTE2" (Howie et al., 2009). While it may seem that no new information is introduced through this process, since the imputation relies on existing data from the SNP array, this approach can still provide valuable insights in scenarios where a SNP on the array exhibits very low LD with an adjacent QTL. In contrast, it might be in substantial LD with a neighboring marker not present on the array, which, in turn, is in LD with the QTL. Consequently, this strategy has the potential to enhance prediction accuracy in genomic prediction, and its efficacy will be further assessed in the manuscript presented in Chapter 4 (Weber et al., submitted).

2 Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets

Weber, S. E., Frisch, M., Snowdon R. J. and Voss-Fels, K.-P. (2023). *Front. Plant Sci. 14:1217589. doi: 10.3389/fpls.2023.1217589*

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 05 September 2023
DOI 10.3389/fpls.2023.1217589

OPEN ACCESS

EDITED BY
Lewis Lukens,
University of Guelph, Canada

REVIEWED BY
José Marcelo Soriano Viana,
Universidade Federal de Viçosa, Brazil
Valerio Hoyos-Villegas,
McGill University, Canada

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

RECEIVED 05 May 2023
ACCEPTED 21 August 2023
PUBLISHED 05 September 2023

CITATION
Weber SE, Frisch M, Snowdon RJ and Voss-Fels KP (2023) Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Front. Plant Sci.* 14:1217589. doi: 10.3389/fpls.2023.1217589

Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets

Sven E. Weber[1]*, Matthias Frisch[2], Rod J. Snowdon[1] and Kai P. Voss-Fels[3]

[1]Department of Plant Breeding, Justus Liebig University, Giessen, Germany, [2]Department of Biometry and Population Genetics, Justus Liebig University, Giessen, Germany, [3]Institute for Grapevine Breeding, Hochschule Geisenheim University, Geisenheim, Germany

In modern plant breeding, genomic selection is becoming the gold standard for selection of superior genotypes. The basis for genomic prediction models is a set of phenotyped lines along with their genotypic profile. With high marker density and linkage disequilibrium (LD) between markers, genotype data in breeding populations tends to exhibit considerable redundancy. Therefore, interest is growing in the use of haplotype blocks to overcome redundancy by summarizing co-inherited features. Moreover, haplotype blocks can help to capture local epistasis caused by interacting loci. Here, we compared genomic prediction methods that either used single SNPs or haplotype blocks with regards to their prediction accuracy for important traits in crop datasets. We used four published datasets from canola, maize, wheat and soybean. Different approaches to construct haplotype blocks were compared, including blocks based on LD, physical distance, number of adjacent markers and the algorithms implemented in the software "*Haploview*" and "*HaploBlocker*". The tested prediction methods included Genomic Best Linear Unbiased Prediction (GBLUP), Extended GBLUP to account for additive by additive epistasis (EGBLUP), Bayesian LASSO and Reproducing Kernel Hilbert Space (RKHS) regression. We found improved prediction accuracy in some traits when using haplotype blocks compared to SNP-based predictions, however the magnitude of improvement was very trait- and model-specific. Especially in settings with low marker density, haplotype blocks can improve genomic prediction accuracy. In most cases, physically large haplotype blocks yielded a strong decrease in prediction accuracy. Especially when prediction accuracy varies greatly across different prediction models, prediction based on haplotype blocks can improve prediction accuracy of underperforming models. However, there is no "best" method to build haplotype blocks, since prediction accuracy varied considerably across methods and traits. Hence, criteria used to define haplotype blocks should not be viewed as fixed biological parameters, but rather as hyperparameters that need to be adjusted for every dataset.

KEYWORDS

genomic selection, SNP markers, haploblocks, haplotype blocks, genomic prediction

1 Introduction

Genomic prediction has greatly improved animal and plant breeding (Hickey et al., 2017) and has the potential to improve genetic gain even in crops with complex genomes (Voss-Fels et al., 2021). In the past, predictions based on linear mixed models used relatedness to borrow information on target phenotypes of relatives. Henderson (1975) derived this relationship from pedigrees *via* the numerator relationship matrix with the expectation that each parent contributes exactly 50% of its genome to its offspring. With the advance of sequencing technology nowadays, genomic data is used to replace the pedigree relationship with realized relationships calculated from dense marker maps. Furthermore, with the inclusion of genetic markers, information about linkage disequilibrium and cosegregation is available for genomic prediction (Habier et al., 2013). Today, individuals in breeding populations of major crops can be sequenced with high quality at low costs, enabling the identification of millions of genome-wide single nucleotide polymorphism (SNP) markers that can be easily screened in large populations using high-throughput genotyping technologies. Together with phenotype measurements, genome-wide marker profiles can be used to predict breeding values of non-phenotyped individuals (Lande and Thompson, 1990; Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008). This can assist breeders in the accurate identification of superior genotypes within their breeding material without the need for additional phenotyping. Moreover, it can facilitate the decision-making process for selecting which genotypes should undergo phenotyping, leading to reduced phenotyping costs and improved accuracy in estimating breeding values. Hence, genomic selection has the potential to considerably increase genetic gain and profit in many crops (Voss-Fels et al., 2021).

There are a variety of statistical methods for genome-based predictions (e.g. VanRaden, 2008; de los Campos et al., 2009; Zhang et al., 2010; Gianola, 2013; Hofheinz and Frisch, 2014; Werner et al., 2018a; Millet et al., 2019), differing in their assumptions of variance components, marker effects or marker modes of action. Examples for genomic prediction models are ridge regression BLUP, GBLUP (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008), Reproducing Kernel Hilbert Space Regression (RKHS) (de los Campos et al., 2009), as well as Bayesian models like Bayesian LASSO (Park and Casella, 2008) or Bayesian ridge regression (Pérez and de los Campos, 2014).

However, biallelic SNPs are sometimes unable to identify all variants and allelic combinations of genes that contribute to a particular trait, since most genes carry multiple sequence polymorphisms. Furthermore, accurate genomic prediction is often obtained based on close relatives (VanRaden, 2008; Hayes et al., 2009) while this accuracy decreases as the validation individuals get more unrelated (Habier et al., 2010; Wolc et al., 2011). This implies that SNPs are not necessarily in LD with causal QTL and the prediction accuracy is at least partly driven by implicitly capturing relationship among individuals. Hence, one strategy to improve predictions is increasing marker density. With the advance of whole genome sequencing technologies, increasingly large and dense marker datasets can today be generated for most major crops (Edwards and Batley, 2010; Yu et al., 2011). However, increasing marker density does not consistently improve prediction accuracies (Solberg et al., 2008; Druet et al., 2014; Hayes et al., 2014; Norman et al., 2018) and often improvements are only observed following pre-selection of markers (van Binsbergen et al., 2015; Ni et al., 2017; Raymond et al., 2018). Furthermore, prediction accuracy is influenced by trait heritability (Zhang et al., 2017) and the number of genotypes with phenotypic records available for genomic selection. Hence, another approach to enhance prediction accuracy is by increasing the number of phenotyped lines used for model training (VanRaden et al., 2009; Combs and Bernardo, 2013). However, due to the high costs associated with phenotyping, this may not always be feasible, particularly when sparse testing methods (Jarquin et al., 2020; Crespo-Herrera et al., 2021; Atanda et al., 2022; Terraillon et al., 2022) are not applicable. Hence, one strategy to address low prediction accuracy could be to identify more informative variants for predictions without necessarily increasing the marker density *per se*.

Loci along the genome are usually inherited in a block-like structure, with only few recombination hotspots (Daly et al., 2001; Jeffreys et al., 2001; Reich et al., 2001) defining the so-called haplotype blocks. There are several ways to define a haplotype block, for example as a fixed window of adjacent markers, as a fixed window of adjacent base pairs, or based on a statistical measure of LD. While the first two are straightforward and simple, they may not represent haplotype blocks in a true biological sense. More sophisticated approaches may model the true haplotype blocks better. Commonly, LD based measures like D′ or r^2 are used for construction of haplotype blocks (Devlin and Risch, 1995). Furthermore, prior information of interaction between adjacent markers may help model local epistasis (Liu et al., 2019), however, difficulties in computing higher order interactions limits the size of haplotype blocks of that type. Haplotype blocks are assumed to be in higher linkage disequilibrium with QTL, and it was proven that haplotype blocks are able to capture local epistasis of markers in close proximity (Jiang et al., 2018). Furthermore, it has been suggested that the problem of apparent or phantom epistasis, which occurs between markers and QTL in incomplete LD, can be overcome with haplotype blocks (Wood et al., 2014; de los Campos et al., 2019). Hence it can be assumed, that haplotype blocks may improve genomic prediction.

In genomic selection, there is evidence that markers grouped to haplotype blocks can improve genomic prediction (Cuyabano et al., 2014; Jiang et al., 2018; Ballesta et al., 2019), while other studies delivered evidence against improving predictions (Solberg et al., 2008). Even with the methods described above for construction of haplotype blocks, it is always necessary to set appropriate hyperparameters like window size or an LD threshold to define block boundaries. Most previous studies in this area investigated a small range of LD thresholds, adjacent markers or window sizes in association studies and genomic prediction (Cuyabano et al., 2014; Hess et al., 2017; Maldonado et al., 2019). However, in terms of genomic prediction for plant breeding the huge variety of options and hyperparameters possible to construct haplotype blocks were not assessed in detail. Hence, the present study sought to investigate the following questions: 1.) How does the method of building haplotype

blocks and its parameters affect the number of haplotypes? 2.) Are haplotype block predictions different from SNP predictions in terms of prediction accuracy? 3.) Is there a preferable haplotype construction method to improve genomic prediction?

These questions were addressed by employing various methods for constructing haplotypes, which are commonly discussed in the literature. The methods range from simple approaches such as marker adjacency (Villumsen and Janss, 2009; Villumsen et al., 2009; Jiang et al., 2018; Liang et al., 2020) and physical distances (Hess et al., 2017; Liang et al., 2020) to more sophisticated methods based on LD thresholds (Cuyabano et al., 2014; Voss-Fels et al., 2019; Bayer et al., 2021; Li et al., 2022) the confidence intervals of *D´* method described by Gabriel et al. (2002), the *Four-gamete Rule* method described by Wang et al. (2002) the *Solid Spine of LD* method (Barrett et al., 2005) and *"HaploBlocker"* (Pook et al., 2019), using four example datasets from canola, maize, wheat and soybean. To assess prediction accuracy, genomic prediction was performed using GBLUP, Bayesian LASSO, EGBLUP and RKHS models.

2 Materials and methods

2.1 Datasets

The datasets examined in this study are all publicly available. The canola dataset is from a spring-type canola hybrid breeding program (Jan et al., 2016). Briefly, 475 double haploid (DH) pollinators were crossed with two male sterile lines to create 950 F_1 test hybrids. The hybrids where subsequently tested for seed yield, flowering time, field emergence, lodging, oil yield and glucosinolate content. For 910 test hybrids the complete phenotypic records were available, and all parental lines were genotyped with the Illumina *Brassica* 60k SNP array (Clarke et al., 2016). The maize dataset is derived from 847 test hybrids from a diverse dent nested association mapping population described by Bauer et al. (2013) consisting of 10 half-sib DH families. Double haploid lines were all crossed to the common flint line UH007 and F1 hybrids were phenotypically analyzed for dry matter yield (DMY), dry matter content (DMC), plant height (PH), days till tasseling (DtTAS) and days till silking (DtSILK), as described by Lehermeier et al. (2014). All DH lines were genotyped with the Illumina MaizeSNP50 SNP array (Clarke et al., 2016). The wheat dataset, described in Voss-Fels et al. (2019), consists of 191 released wheat varieties from 1966 to 2013 that were tested under three agrichemical treatments for a wide range of agronomic traits including yield, biomass yield, falling number, days till heading, plant height, harvest index kernel spike^{-1}, nitrogen use efficiency (NUE), powdery mildew resistance, protein content, protein yield sedimentation value spike m^{-2}, stripe rust and thousand kernel weight (TKW). All lines were genotyped with the Illumina 15k wheat SNP array described in Soleimani et al. (2020). The soybean dataset consisted out of 1000 lines from the USDA Soybean Germplasm Collection (Grant et al., 2010) with phenotypic records for protein and oil content (PC, OC) (Bandillo et al., 2015). For all lines, genotypic information from the Illumina Infinium SoySNP50K BeadChip (Song et al., 2013) was available.

With the exception of the maize dataset, all phenotypic data represented adjusted trait means per genotype. The published field data from the maize population was adjusted following methods used for phenotypic data analyses from the original publication.

2.2 Genotypic data

With exception of the canola dataset, physical SNP marker positions were obtained from the respective reference genome assemblies used in the original publications, namely the *Brassica napus* Express 617 genome (Lee et al., 2020), the maize B73 AGPv2 genome (Schnable et al., 2009), the wheat Chinese Spring IWGCS reference Sequence v1.0 (Zimin et al., 2017) and the soybean Glyma1.01 reference (Schmutz et al., 2010). In general, only markers with a unique physical position on the reference genome, a minor allele frequency ≥ 0.05 and a maximum of 10% missing values in each population were used for further analyses. This left a total of 29385, 32363, 8710 and 35821 markers for the canola, maize, wheat and soybean datasets, respectively. This corresponds to a marker density of 31.78, 15.63, 0.57 and 37.48 SNPs mbp^{-1} in canola, maize, wheat and soy respectively. After filtering, markers were imputed with the software "*BEAGLE*" V5.2 (Browning and Browning, 2007; Browning et al., 2018).

2.3 Haplotype block construction

We considered seven haplotype block construction methods based on (i) pre-determined LD thresholds, (ii) fixed windows of adjacent markers, (iii) fixed windows of adjacent base pairs, (iv) *"HaploBlocker"* (Pook et al., 2019), (v) the confidence intervals of *D´* method described by Gabriel et al. (2002), (vi) the *Four-gamete Rule* method described by Wang et al. (2002) and (vii) the *Solid Spine of LD* method (Barrett et al., 2005). The first three methods were implemented in the r package *"SelectionTools"* (downloadable at http://population-genetics.uni-giessen.de/~software/), while the latter three are implemented in the software "*Haploview*" v4.1 (Barrett et al., 2005). The different approaches are described in detail below. These methods were selected for their widespread use in haplotype block formation and their distinct characteristics. Methods such as the pre-determined LD threshold, confidence intervals of *D'*, the *Four-gamete Rule*, and the *Solid Spine of LD* are based on linkage disequilibrium (LD) and gamete frequency. They aim to model historical recombination hotspots and generate meaningful blocks within populations. However, these blocks do not necessarily represent functional groups. Therefore, we also included methods based on fixed windows to assess blocks that would not be constructed based on population-based measures alone. Additionally, while most methods consider block borders across the entire population, it is important to note that subpopulations or genotypes may have different recombination patterns. To account for this, we utilized the method *"HaploBlocker"* described in Pook et al. (2019) to construct haplotype blocks specific to different groups.

2.3.1 LD threshold

LD between markers on the same chromosome was calculated as r^2 (Hill and Robertson, 1968) in *"SelectionTools"*. Haplotype blocks were built by starting with the two neighboring markers with the highest LD. If the pairwise LD exceeded a certain threshold, those markers were then assigned to a haplotype block. In the next step, if the LD between the next immediately adjacent markers and the markers at the block border again exceeded the threshold, the block was extended. This was done until no more markers fulfilled this criterion and the algorithm started over again with new markers. To account for misplaced markers, a tolerance parameter of 1 was used, meaning that one marker that did not fulfill the LD threshold was accepted if the next flanking marker fulfilled the LD criterion. Thresholds were set sequentially from 0.01 to 1 with a step size of 0.01, resulting in 100 different LD thresholds. Using very high thresholds to form blocks effectively eliminates redundant information, making these scenarios similar to LD pruning, which has been shown to improve prediction accuracy (Ye et al., 2019). On the other hand, very low thresholds result in the formation of large blocks commonly observed in introgression breeding, where recombination is sometimes very limited (Hao et al., 2020).

2.3.2 Fixed windows of adjacent markers

Starting at the beginning of each chromosome, haplotype blocks consisting of m neighboring markers were constructed until all markers on a chromosome were assigned to blocks. We considered $\lceil 2^x \rceil$ markers with x being {1, 1.5, 2, 2.5 …}, until in the most excessive case all markers of a chromosome represented a haplotype block containing all markers of that chromosome. We chose to create blocks of such large size to address scenarios where entire chromosomes or large segments play an important role in traits, as well as scenarios related to introgression breeding, where recombination is limited (Hao et al., 2020).

2.3.3 Fixed windows of adjacent base pairs

Starting at the beginning of each chromosome, haplotype blocks of m consecutive base pairs were constructed until the whole chromosome was partitioned into blocks. We considered $\lceil 2^x \rceil$ base pairs with x being {10, 10.5, 11, 11.5 …} until in the most excessive case a whole chromosome represented a block. Similar to the approach using fixed windows of adjacent markers, we selected to construct blocks of considerable size to accommodate scenarios where entire chromosomes or large segments influence traits, as well as situations related to introgression breeding characterized by limited recombination (Hao et al., 2020).

2.3.4 HaploBlocker

Since different subpopulations might result in different block borders, we also built haplotype blocks with the algorithm of Pook et al. (2019). This algorithm relies on linkage instead of linkage disequilibrium to construct haplotype blocks. Here blocks are defined as consecutive sequence of genetic markers with a predefined frequency, a sequence of haplotype merging and splitting steps is applied to construct subgroup-specific haplotype blocks. This algorithm allows subgroup specific haplotype block borders. The algorithm was conducted with default settings with the r package *"HaploBlocker"* (Pook et al., 2019).

2.3.5 Gabriel algorithm

The algorithm developed by Gabriel et al. (2002) (GAB) for the Human Haplotype Map generates 95% confidence bounds on D' between all intrachromosomal marker pairs. Marker pairs are considered in "strong LD" if the one-sided upper 95% D' confidence bound is higher than 0,98 and the lower bound is higher than 0.7. Markers in "strong LD" are consequently grouped into blocks. Blocks are extended until the outermost marker pairs don't fulfill this criterion anymore.

2.3.6 Four gamete rule

The *Four Gamete Rule* (GAM) described by Wang et al. (2002) groups consecutive markers into haplotype blocks if no evidence for a historical recombination event can be found between all marker pairs of a block. A historical recombination is defined if all four haplotypes of the new marker and any other previous marker are found with at least 1% frequency. If this is the case, a block border is created between those markers and the algorithm starts with a new block.

2.3.7 Solid spine of LD

The *Solid Spine of LD* method (SPI), introduced by the developers of "*Haploview*" (Barrett et al., 2005), searches for a spine of strong LD by calculation of LD between all intrachromosomal marker pairs. In this method, two markers on the same chromosome form a block border if the pairwise D' is higher than 0.8. All markers in that window form the block. This allows for intermediate markers to not be in LD.

2.4 Genomic prediction models

In total, four genomic selection models were used to predict testcross (maize, canola) and inbred line (soybean, wheat) performance, respectively. The models represent two variations of the GBLUP and two models implemented in a Bayesian framework. The frequentist models were GBLUP (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008) and extended GBLUP to account for second-order additive*additive epistasis, following the EGBLUP model of Jiang and Reif (2015). The Bayesian model included the Bayesian LASSO model (Park and Casella, 2008) which offers the capability of marker-specific shrinkage, and the semiparametric RKHS regression model (de los Campos et al., 2009) which allows modeling of higher order epistasis.

In the GBLUP and EGBLUP the underlying model is assumed to be:

$$y = X\beta + Z_a a + Z_i i + e$$

where y is a vector of observations for a trait under consideration, β is a vector of fixed non-genetic effects, a is a vector of random additive effects, i is a vector of random epistatic effects and e is the random residual term. Z_a and 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 and soybean datasets, a vector of ones modeling the intercept ($1_n\mu$). In the wheat dataset, two additional fixed effects for N fertilization and fungicide treatment were added, while in the maize dataset an additional 10 columns were added to assign individuals to half-sib families.

It is assumed that

$$a \sim N(0, G\sigma_a^2),\ \ i \sim N(0, G_{aa}\sigma_{aa}^2)\ \ and\ \ e \sim N(0, I\sigma_e^2)$$

where $\sigma_a^2, \sigma_{aa}^2$ and σ_e^2 are additive genetic variance, epistatic genetic variance and residual variance respectively. G and G_{aa} are the additive and epistatic relationship matrices, respectively. I is an identity matrix. Depending on inclusion of epistatic effects the epistasis terms were included or omitted.

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

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

with the elements of Z being (0-2p_i) for genotype H_iH_i, (1-2p_i) for genotype H_iH_j and (2-2p_i) for genotype H_jH_j, where H_j is the haplotype (treated as a single marker) within a haplotype block, H_i is any other haplotype within that haplotype block except H_i, and p_i is the frequency of the ith haplotype in a haplotype block. Haplotype blocks with only two haplotypes were treated like standard biallelic markers. For the canola dataset, prior to construction of the genomic relationship matrix, parental genotypes were crossed *in silico* to derive hybrid genotypes, as described by Werner et al. (2018a).

According to Henderson (1985) and Jiang and Reif (2015), the second order (additive*additive) epistatic relationship matrix can be approximated with $G_{aa} = G\#G$, with # denoting the pointwise (hadamard) product operation.

GBLUP and EGBLUP were implemented and solved with the R package *sommer* (Covarrubias-Pazaran, 2016; Covarrubias-Pazaran, 2018).

The general formula describing the model Bayesian LASSO model of Park and Casella (2008) is:

$$y = X\beta + Mf + e$$

where y is the vector of observations for a trait under consideration, β is a vector of fixed non-genetic effects, a is a vector of additive effects. X is the design matrix as described in the GBLUP section. M is an incidence matrix relating phenotypic records with the respective marker/haplotype profiles coded 0, 1, 2. The coefficients of the fixed (β) effects are assigned flat priors, while the coefficients of the marker/haplotype effects (f) are assigned double-exponential priors. This allows the shrinkage of some marker/haplotype effects to effectively zero, introducing sparsity into the model. This model was tested because we assumed that some marker variants and particularly some haplotypes would have no effect on some traits. Here, e is the random residual term. In the Bayesian LASSO, only additive effects were modeled, because additional effects in this framework would increase the computational burden to an unacceptable degree. This model was conducted in the r software with the package *BGLR* (Pérez and de los Campos, 2014) using the default parameters.

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

$$y = X\beta + \sum_{l=3}^{L} u_l + e$$

with

$$p(\beta, u_1, \ldots u_L, e) \propto \prod_{l=1}^{L} N(u \mid 0,\ K_l\sigma_{ul}^2)\ \ N(e \mid 0,\ I\sigma_e^2)$$

where K_l is an $n \star n$ kernel. It is calculated from the Euclidean distance between genotypes based on their marker/haplotype profile. We selected 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 the random genomic effect. That way the different random effects, i.e. the three kernel matrices from the three bandwidth parameters, are weighted by their variance components. Here, e is the random residual term. As for the Bayesian LASSO, the RKHS model was conducted in the r software with the package *BGLR* (Pérez and de los Campos, 2014) using the default parameters.

2.5 Genomic relationship

Generally, constructing haplotype blocks applies a transformation to the original marker data. To assess how well the marker data is also captured by haplotype blocks, we used the relationship coefficients obtained from the relationship matrix calculated following VanRaden (2008) (see above) and calculated the Pearson correlation between relationship coefficients obtained from SNPs and those obtained from haplotype blocks.

2.6 Evaluation of prediction accuracy

For all the four datasets, model performance was assessed by running 100 cross-validation runs, where each cycle consisted of splitting the population into 80% training population and 20% validation population. Each model was trained on the training population and then this model was used to predict the validation population with masked phenotypic data. Furthermore, in the maize dataset, a family wise cross validation was conducted. This was done to test how predictive haplotype blocks are to predict genetically distant individuals. Here, the dataset was split according to the family assignment of the nested association mapping population and each family served once as validation set. In both cross validation schemes, the Pearson correlation coefficient (r) between observed and predicted phenotypic values of the validation population was used as a measure of prediction accuracy.

3 Results

3.1 Haplotype block properties

In all the datasets analyzed, haplotypes of varying sizes were examined. The haplotype blocks had average physical sizes ranging

from 1.02 kbp to 47453.13 kbp, 379625.06 kbp, 1073741.82 kbp, and 47453.13 kbp, respectively, for canola, maize, wheat and soybean. A summary of the average size distributions can be found in Table 1. Notably, the fixed window approaches allowed for the construction of both the smallest haplotype blocks (1.02 kbp) and the largest haplotype blocks (Table 1).

Within all datasets using the methods implemented in *"Haploview"* (GAB, GAM, SPI), the number of haplotype blocks was consistently lower than the number of total SNPs (Figures 1A, E, I, M). However, a significant portion of those blocks consisted of only a single SNP (unblocked SNPs) (Figures 1A, E, I, M). Moreover, the total number of haplotypes available for genomic prediction (excluding single SNP blocks) increased in the canola and soybean datasets, remained similar to the number of SNPs in wheat, and decreased in maize (Figures 1A, E, I, M). Across all datasets, the number of blocks based on LD increased with higher LD thresholds. Additionally, in the case of maize, the number of haplotypes exhibited a similar pattern. With LD-based haplotype blocks, the number of haplotypes (excluding single SNP blocks) exceeded the total number of SNPs across all LD thresholds in soybean and was lower across all thresholds in maize (Figures 1B, F, J, N). In canola, thresholds above r^2 = 0.75 resulted in fewer haplotypes than SNPs, while lower thresholds yielded higher numbers. Conversely, in wheat, only relatively small blocks ($r^2 \leq$ 0.10) increased the number of haplotypes compared to the number of SNPs (Figure 1B). With fixed window blocks, the number of haplotype blocks generally decreased with increasing block size (Figures 1C, D, G, H, K, L, O, P). Here, the number of haplotypes was the highest with relatively small blocks, with increasing block size, the number of haplotypes decreased (Figures 1C, D, G, H, K, L, O, P). Notably, in comparison to SNPs, the number of haplotypes was higher for blocks smaller than 1024, 6, 128, and 1449 SNPs, or 23726.57 kbp, 92.68 kbp, 134217.73 kbp, and 33554.43 kbp in the canola, maize, wheat, and soybean datasets, respectively (Figures 1C, D, G, H, K, L, O, P). In all scenarios, increasing block size resulted in fewer unblocked markers, especially with the fixed window approaches. In all datasets, the *"HaploBlocker"* method produced the fewest haplotypes, considerably fewer than the number of SNPs (Figures 1A, E, I, M). Furthermore, across all datasets and methods, except for blocks based on *"HaploBlocker"*, most of the introduced haplotypes can be classified as rare (Frequency ≤ 0.05) or very rare (Frequency ≤ 0.01) (Figure 1).

Across the four datasets, the examination of the correlations between relationship coefficients derived from SNPs and haplotypes revealed high redundancy between the two marker types in many method/parameter combinations. The methods implemented in *"Haploview"* resulted in relationship coefficients that were highly correlated to those obtained from SNPs, closely approaching a correlation coefficient of 1, in canola, wheat, and soybean (Figures S1A, E, I, M). However, in maize, these methods only produced intermediate correlations (GAB = 0.60, GAM = 0.50, SPI = 0.46) (Figure S1E). In all datasets, relationship coefficients from haplotypes from LD-based haplotype blocks were highly correlated to those obtained from SNPs ($r > 0.75$) with little variation observed across LD thresholds. Only at very low LD thresholds, this correlation was slightly lower, while it was slightly higher for very high thresholds (Figures S1B, F, J, N). Additionally, small fixed window blocks resulted in relationship coefficients similar to those obtained from SNPs, closely approaching a correlation coefficient of 1. However, this similarity eroded drastically with increasing block size (Figures S1C, D, G, H, K, L, O, P). Notably, in Soybean, while the correlation between relationship coefficients from SNPs and haplotypes decreased with increasing block size of the fixed window of adjacent base pairs, it slightly increased again with the largest blocks (*nKB* = 67108.86) (Figure S1P). In canola and soybean, relationship coefficients obtained from *"HaploBlocker"* were highly correlated to those obtained from SNPs (Figures S1A, M). In wheat, this correlation was lower (r = 0.75), and in maize, it was close to zero (r = 0.058), indicating that these blocks capture different information (Figures S1E, I).

TABLE 1 Average size ranges of haplotype blocks constructed by LD, fixed window of adjacent markers and fixed window of adjacent base pairs in the canola, maize, wheat and soybean dataset.

Dataset	Method	minimal average size (kbp)	maximal average size (kbp)
Canola	LD	97.49 (r^2 = 1)	2629.87 (r^2 = 0.01)
	fixed window of adjacent marker	25.46 (*nSNP* = 2)	39801.09 (*nSNP* = 2048)
	fixed window of adjacent base pairs	1.02 kbp	47453.13 kbp
Maize	LD	8.08 (r^2 = 1)	21556.53 (r^2 = 0.01)
	fixed window of adjacent marker	64.31 (*nSNP* = 2)	205312.88 (*nSNP* = 5793)
	fixed window of adjacent base pairs	1.02 kbp	379625.06 kbp
Wheat	LD	106.79 (r^2 = 1)	64954.10 (r^2 = 0.01)
	fixed window of adjacent marker	1544.58 (*nSNP* = 2)	667692.8 (*nSNP* = 1024)
	fixed window of adjacent base pairs	1.02 kbp	1073741.82 kbp
Soybean	LD	138.55 bp (r^2 = 1)	1587.07 (r^2 = 0.01)
	fixed window of adjacent marker	430.27 (*nSNP* = 2)	1526.61 (*nSNP* = 2897)
	fixed window of adjacent base pairs	1.02 kbp	47453.13 kbp

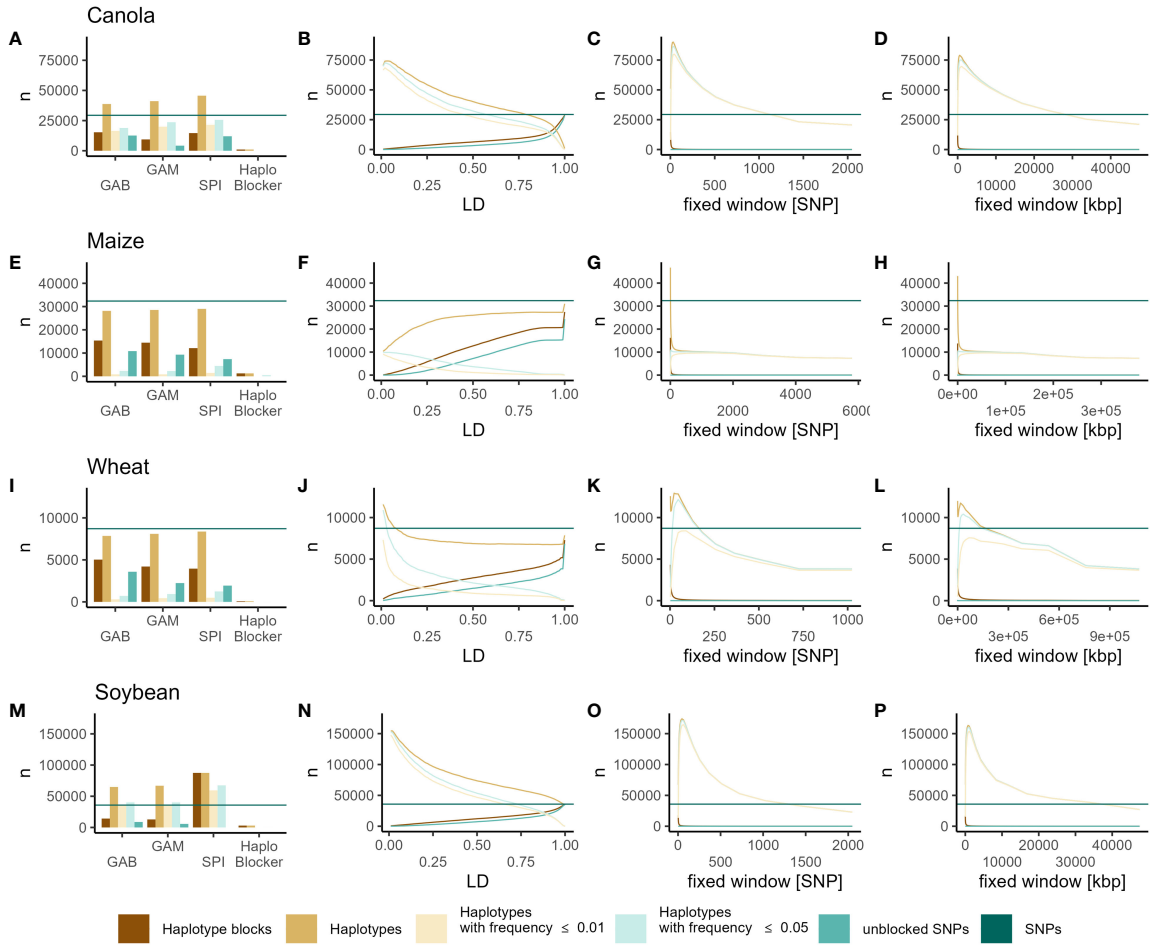

FIGURE 1
Numbers of SNPs (dark green horizontal line), haplotype blocks (dark orange), haplotypes (light orange), haplotypes with frequency ≤ 0.05 (light green), haplotypes with frequency ≤ 0.01 (light yellow) and unblocked SNP markers (green) identified by GAB, GAM, SPI, *"HaploBlocker"* **(A, E, I, M)**, LD **(B, F, J, N)**, fixed window of adjacent markers **(C, G, K, O)** and fixed window of adjacent base pairs **(D, H, I, P)** in canola **(A–D)**, maize **(E–H)**, wheat **(I–L)** and soybean **(M–P)**.

3.2 Genomic prediction

3.2.1 Canola

Within the canola dataset, the prediction accuracy across different models ranged from 0.3 to 0.85, with a strong dependence on the specific trait. Notably, for oil yield, field emergence, glucosinolate content, and lodging, the models considering epistatic effects (EGBLUP and RKHS) consistently outperformed by the other SNP-based models (Figure S2). However, this effect did not consistently translate to haplotype-based predictions. Prediction accuracy showed little variation across LD threshold as well as between LD base, *"Haploview"* or *"HaploBlocker"* methods (Figures 2A, B, S2). On the other hand, the fixed-window approaches exhibited the most variation, with a substantial decrease in prediction accuracy as the block size increased for every trait, while small blocks based on fixed windows resulted in prediction accuracies similar to those based on SNPs or the remaining methods (Figures 2C, D, S2).

Comparing haplotype blocks to SNP-based prediction, the improvement in prediction accuracy ranged from 0.007 to 0.021 for GBLUP, 0.008 to 0.024 for Bayesian LASSO, 0.008 to 0.023 for EGBLUP, and 0.007 to 0.022 for RKHS. These values were based on the haplotyping method that yielded the highest prediction accuracy for each specific trait and model (Tables 2, S1). Interestingly, the use of haplotypes seemed to have the least impact on oil yield (Figure S1; Table S1). Except for flowering time with RKHS, the LD-based methods generally resulted in the most significant improvements. However, no ideal LD threshold or range of thresholds could be identified (Table S1). In the case of flowering time with RKHS, the optimal haplotyping method involved a fixed window of adjacent base pairs measuring 20987.15 kbp.

3.2.2 Maize

Prediction accuracy obtained from the random cross validation ranged from 0.4 to 0.9 and was trait-dependent. Here, little difference between models was observed with SNP-based prediction (Figures 2E, S3). With haplotypes, however, there were considerable differences between Models implemented in a Bayesian framework and frequentist models (Figures 2, S3). With

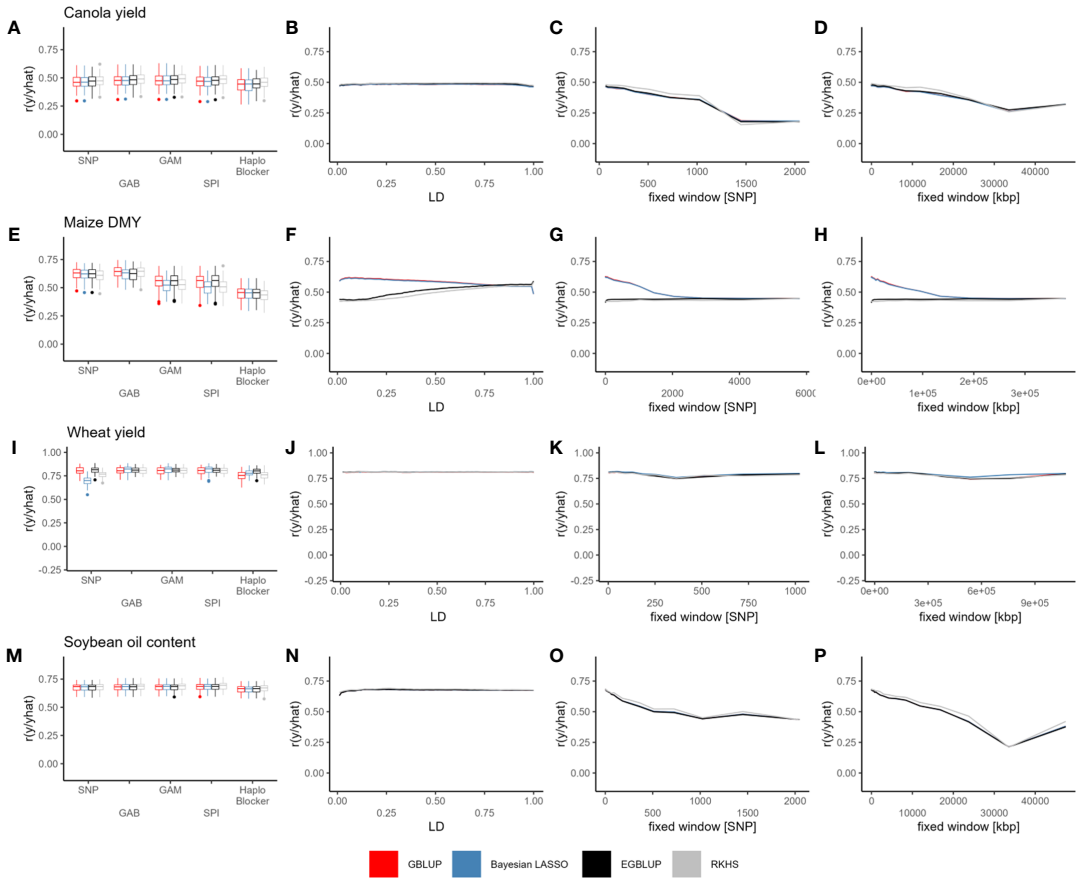

FIGURE 2
Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, GAB, GAM, SPI, *"HaploBlocker"* **(A, E, I, M)**, LD **(B, F, J, N)**, fixed window of adjacent markers **(C, G, K, O)** and fixed window of adjacent base pairs **(D, H, L, P)** based haplotype blocks, in canola seed yield **(A–D)**, maize DMY **(E–H)**, wheat seed yield **(I–L)** and soybean oil content **(M–P)**. Individual points in the line plots represent the mean over all cross validation runs for each haplotype block parameter and model combination.

haplotypes based on LD, prediction accuracy decreased with higher LD thresholds for GBLUP and EGBLUP and increased for Bayesian LASSO and RKHS, respectively (Figures 2F, S4). Here for DMY, DMC and PH, respectively, all models approached a similar prediction accuracy around r^2~0.75 (Figure S3). And for DtTAS and DtSILK all models approached the same prediction accuracy around r^2~0.55 (Figure S3). The same behavior could not be observed with the fixed window haplotypes, where prediction accuracy obtained from GBLUP and EGBLUP decreased drastically with increasing block size. Here, for models implemented in a Bayesian framework the prediction accuracy remained low independent of the block size (Figures 2G, H, S3). Except for DMY, where the GAB method slightly improved prediction accuracy, haplotypes based on the algorithms implemented in "*Haploview*" decreased prediction accuracy in every scenario (Figures 2E, S4). In general, there was no discernable improvement of prediction accuracy by haplotypes compared to SNP-based predictions. In all traits but DMY, the haplotyping method with the highest prediction accuracy even decreased prediction accuracy with Bayesian LASSO and RKHS (Tables 2, S1), whereas for GBLUP and EGBLUP prediction accuracy did not or only slightly increased prediction accuracy compared to SNP based prediction. DMY profited most from haplotypes, whereby GBLUP, Bayesian LASSO and RKHS worked best with the GAB method while for EGBLUP a fixed window of 8 SNPs was ideal (Figure 2; Table 2). Besides DMY, for Bayesian LASSO and RKHS haplotypes worked best with an LD threshold of $r^2 = 1$, however, prediction accuracy was still worse than SNP based prediction (Table S1; Figure S4). For the same traits, Frequentists model worked best with varying fixed window size haplotypes, with a maximal improvement of 0.002 (Table S1). The "*HaploBlocker*" method together with very large fixed window blocks yielded the lowest prediction accuracies across all traits.

The family-wise cross validation generally yielded considerably lower prediction accuracies than its random counterpart (Figure S4; Table S2). The ranking in prediction accuracies obtained from haplotype blocks followed the pattern of the random counterpart, albeit being lower (Figure S4; Table S2). Mentionable, prediction accuracy approached zero for the *HaploBlocker*" method together with very large fixed window blocks.

3.2.3 Wheat

Prediction accuracy in the wheat dataset exhibited much greater variability between traits compared to the other three datasets,

10.3389/fpls.2023.1217589

TABLE 2 Average prediction accuracy of SNP based prediction compared to the best haplotyping method of canola, maize, wheat and soybean for some example traits.

Dataset	Trait	Model	SNP prediction accuracy	Best haplotyping algorithm	Prediction accuracy by best haplotyping algorithm	Improvement by best haplotyping algorithm
Canola	yield	GBLUP	0.464	r^2 = 0.6	0.485	0.021
		Bayesian LASSO	0.462	r^2 = 0.59	0.486	0.024
		EGBLUP	0.471	r^2 = 0.6	0.492	0.021
		RKHS	0.474	r^2 = 0.71	0.496	0.022
	flowering time	GBLUP	0.709	r^2 = 0.12	0.721	0.012
		Bayesian LASSO	0.697	r^2 = 0.15	0.721	0.024
		EGBLUP	0.711	r^2 = 0.14	0.723	0.012
		RKHS	0.704	*nKB* = 2097.15	0.719	0.015
Maize	DMY	GBLUP	0.624	GAB	0.635	0.011
		Bayesian LASSO	0.616	GAB	0.621	0.006
		EGBLUP	0.620	*nSNP* = 8	0.622	0.002
		RKHS	0.608	GAB	0.631	0.023
	DtTAS	GBLUP	0.847	*nSNP* = 4	0.847	0.000
		Bayesian LASSO	0.846	r^2 = 1	0.842	-0.004
		EGBLUP	0.845	*nSNP* = 4	0.846	0.000
		RKHS	0.846	r^2 = 1	0.842	-0.003
Wheat	yield	GBLUP	0.805	r^2 = 0.23	0.813	0.008
		Bayesian LASSO	0.697	*nSNP* = 46	0.818	0.122
		EGBLUP	0.811	r^2 = 0.1	0.815	0.005
		RKHS	0.765	r^2 = 0.1	0.814	0.049
	sedimentation value	GBLUP	0.493	*nKB* = 1073741.82	0.619	0.126
		Bayesian LASSO	0.488	*nKB* = 1073741.82	0.636	0.148
		EGBLUP	0.620	*nKB* = 1073741.82	0.627	0.006
		RKHS	0.610	*nKB* = 1073741.82	0.631	0.021
Soybean	oil content	GBLUP	0.674	r^2 = 0.24	0.682	0.008
		Bayesian LASSO	0.675	r^2 = 0.24	0.683	0.008
		EGBLUP	0.674	r^2 = 0.24	0.682	0.008
		RKHS	0.677	r^2 = 0.26	0.691	0.014
	protein content	GBLUP	0.601	*nSNP* = 4	0.606	0.006
		Bayesian LASSO	0.602	*nSNP* = 4	0.608	0.006
		EGBLUP	0.609	*nSNP* = 4	0.611	0.003
		RKHS	0.609	*nSNP* = 4	0.613	0.003

ranging from -0.4 to 0.9, depending on the specific trait. Interestingly, even with SNP-based predictions, considerable differences in prediction accuracy were observed across (i) models that consider epistasis and those that do not, (ii) frequentist and models implemented in a Bayesian framework, and (iii) combinations of (i) and (ii) (Figures S5–S7). However, when haplotype blocks were utilized, all models achieved at least the average prediction accuracy of the best SNP-based model for 13 out of 15 traits (Figures S5–S7; Table S1). This was achieved by using haplotype blocks constructed with varying methods, including even the largest possible haplotype blocks based on fixed windows (e.g., using whole chromosomes as blocks) (Figures 2, S5–S7; Table S1).

Furthermore, for traits such as yield, biomass yield, NUE, protein yield, sedimentation value, stripe rust, and falling number, the previously worst-performing SNP-based model became the best-performing model when using haplotype blocks (Table S1). Additionally, for traits with very low or even negative prediction accuracy based on SNPs (e.g., plant height, TKW, days till heading, falling number, powdery mildew, and stripe rust), strong improvements were achieved through the use of haplotypes (Figures S5–S7; Table S1). Models implemented in a Bayesian framework seemed to benefit the most from the utilization of haplotypes, with changes in prediction accuracy ranging from -0.039 to 0.170 for GBLUP, from 0.006 to 0.277 for Bayesian LASSO, from -0.003 to 0.085 for EGBLUP, and from 0.025 to 0.291 for RKHS (Tables 2, S1). The most notable improvements were typically seen when prediction accuracy varied considerably between models using SNP data. Only for cases, such as falling number with RKHS, kernel spike^{-1} with EGBLUP, spike m^{-2} with RKHS and EGBLUP, and stripe rust resistance with GBLUP, did the prediction accuracy decrease compared to SNP-based prediction when using haplotype blocks (Table S1).

3.2.4 Soybean

The prediction accuracy in the soybean dataset ranged from 0.5 to 0.8 and exhibited a striking similarity between oil content and protein content. No noticeable differences were observed between models based on SNPs (Figures 2M, S8A, E). Moreover, there was minimal variation in prediction accuracy across different LD thresholds, with only a slight decrease in accuracy observed between r^2 = 0.01 and 0.05 (Figures 2N, S8B, F).

When using fixed windows of adjacent marker blocks, the prediction accuracy experienced a decline with increasing block size for all models (Figures 2O, S8C, G). Similar behavior was observed for fixed windows of adjacent base pairs blocks, except for a marginal increase in prediction accuracy with blocks of size 47453.13 kbp (Figures 2P, S8D, H). However, it is worth noting that the prediction accuracy remained lower than the SNP-based prediction in that case. Overall, the improvements achieved with haplotypes were relatively minor (Tables 2, S1). For oil and protein content, the best haplotype block method and parameter improved the prediction accuracy by only 0.006 and 0.008 with GBLUP, 0.006 and 0.009 with Bayesian LASSO, 0.003 and 0.008 with EGBLUP, and 0.003 and 0.014 with RKHS, respectively, compared to the SNP-based prediction (Tables 2, S1).

Interestingly, within the traits, it was observed that the models worked best with the same haplotype block method: an LD threshold of r^2 = 0.24-0.26 for oil content and a fixed window size of *nSNP* = 4 for protein content.

4 Discussion

Using datasets from four diverse crops and haplotype blocks constructed using a broad range of construction parameters, we show how haplotype blocks change in size and influence the effective number of predictors for genomic prediction. While haplotype blocks sometimes drastically change the number of predictors, genomic prediction accuracy was only marginally affected with no consistent improvement for any method and trait.

Haplotype blocks were built based on LD (r^2), fixed window sizes of adjacent marker or base pairs as well as the three algorithms implemented in the software "*Haploview*" and the method "*Haploblocker*". The r^2 measurement of LD between markers (Hill and Robertson, 1968; Hill, 1981) is highly correlated to D' (VanLiere and Rosenberg, 2008), which is more commonly used in tagSNP methods where it showed superior performance to other measures (Carlson et al., 2004; de Bakker et al., 2005). According to Cuyabano et al. (2014), r^2 and D' show no difference in terms of prediction accuracy in genomic prediction. The high resolution of haplotype blocking methods and construction parameters allowed an examination of a wide range of haplotype block sizes that are normally not considered in genomic prediction. Most studies in this regard only include single or few construction methods or parameters (Lorenz et al., 2010; Ballesta et al., 2019; Maldonado et al., 2019), although our results show that the method of haplotype construction can potentially impact prediction quality. We included haplotype blocks of relatively large sizes, such as a LD threshold of 0.01 and whole chromosome blocks, which may initially seem unrealistic. However, we included these large blocks to account for scenarios in which traits are controlled by large chromosome segments (Voss-Fels et al., 2019), possibly resulting from introgression breeding with suppressed recombination (Hao et al., 2020).

Here, in three datasets the number of haplotypes could be increased substantially compared to the number of SNPs. The number of haplotypes we observed in the four examined datasets was lower than observed in cattle (Cuyabano et al., 2014; Cuyabano et al., 2015; Li et al., 2021; Li et al., 2022) and human (Liang et al., 2020) but similar to previous reports in plants including *Eucalyptus globulus* (Ballesta et al., 2019), maize (Matias et al., 2017) and rice (Matias et al., 2017). These variations may arise from differences in population diversity, marker density, and sequencing technology. The haplotype number detected in maize by Matias et al. (2017) was comparable to that observed in our analysis using around ten times fewer SNP markers, indicating that haplotype number is not (solely) dependent on marker density. However, as expected there is a relationship between the population size and haplotype number, with more (diverse) genotypes causing more haplotypes. The number of haplotypes we detected corresponded to the

population size used for each crop, with wheat having the fewest haplotypes and soybean the most, independent of the method. Nevertheless, an effect of genetic diversity within a species or population cannot be discounted without comparative within-species analyses of alternative populations. Some authors argue that use of haplotype blocks can help to reduce dimensionality (Kim et al., 2019; Pook et al., 2019). However, depending on the methods and parameters for haplotype construction the number of haplotypes was sometimes higher in the examined datasets than the number of SNPs. This may reflect lower marker numbers and different methods compared to Kim et al. (2019). Dimensionality can certainly be decreased if rare haplotypes would be excluded (Hess et al., 2017; Li et al., 2022). The method *"HaploBlocker"* described by Pook et al. (2019) decreased the dimensionality in every examined dataset. In all cases, the major drawback of the large number of additional variants is the very low frequency at which the haplotypes occur. However, low frequency variants are often assumed to be in higher LD with recent causal mutations (Bloom et al., 2019; Wainschtein et al., 2022), implying that their detection and use for predictions could be beneficial. However, caution is needed when considering all haplotypes, especially rare ones. In genomic predictions. effect estimation of rare variants require large populations to be estimated accurately (Meuwissen et al., 2001; Goddard and Hayes, 2007). In large populations, rare variants can be observed at higher frequencies which enables a more accurate estimation of their trait effects. In SNP based prediction markers are commonly excluded if they have a minor allele frequency ≤ 0.05 (Technow et al., 2012; Crossa et al., 2013; Jan et al., 2016; Werner et al., 2018a; Zhang et al., 2018). With large populations, filtering could be shifted from frequencies to allele counts, potentially leading to more reliable effect estimates of rare haplotypes. However, increasing the population could again increases the number of rare new haplotypes. In all four datasets, the number of unblocked SNPs decreased with increasing block size. With LD based haplotype blocks, increasing the LD threshold resulted in an increase of unblocked SNPs.

Genomic prediction was conducted using four models: GBLUP, EGBLUP, Bayesian LASSO, and RKHS regression, with the latter two implemented within a Bayesian framework. GBLUP, being the golds standard of genomic prediction, is a widely employed prediction models in breeding, hence we included it in the analysis. However, GBLUP assumes that all markers or haplotypes contribute to the trait (through relationship), prompting the inclusion of Bayesian LASSO, which allows for marker or haplotype-specific shrinkage of effects towards zero. This is beneficial in scenarios where not all markers or haplotypes have an impact on the trait. Given the assumption that haplotypes capture local epistatic effects (Jiang et al., 2018), EGBLUP and RKHS regression were employed to assess whether considering global epistasis between haplotype blocks could yield a substantial improvement in genomic prediction. Although haplotype blocks are typically fewer in number compared to SNPs, the number of haplotypes used for prediction was often comparable to or even greater than the number of SNPs. Therefore, we selected prediction models capable of handling the challenges posed by the large p small n scenario, opting not to explore machine learning models. Furthermore, the application of machine learning methods would have required extensive hyperparameter optimization, which would have significantly exceeded the computational time required for the four prediction models employed in this study. Lastly, the objective of this study was to compare various haplotype blocking methods and parameters, rather than comparing different prediction models.

Generally, genomic prediction accuracies based on SNPs were similar to those reported in the literature across all datasets. In the canola dataset, accuracies closely matched Jan et al. (2016), with a small improvement likely due to the higher number of markers remaining after filtering. Trait prediction accuracies in canola/rapeseed were mostly consistent with previous reports, with minor variations observed for field emergence, and glucosinolate content (Würschum et al., 2014; Jan et al., 2016; Werner et al., 2018a; Werner et al., 2018b). Also in the maize dataset, SNP-based genomic prediction accuracy roughly matched the original publication (Lehermeier et al., 2014), with expected differences due to varying cross-validation schemes. Maize hybrids exhibited high prediction accuracies as previously reported (Technow et al., 2012; Crossa et al., 2014; Millet et al., 2019). In wheat, prediction accuracies based on SNPs for seed yield and yield components were on a very high level (Table S1) compared to many previously published reports (Lado et al., 2013; Zhao et al., 2013; Crossa et al., 2014; Daetwyler et al., 2014; Crossa et al., 2016; Edwards et al., 2019). Furthermore, prediction accuracies based on SNPs for stripe rust resistance, despite population differences, showed a similar level than observed by Daetwyler et al. (2014). Whereas protein content had a higher prediction accuracy compared to Crossa et al. (2016), sedimentation value was predicted equally well. In soybean, prediction accuracies based solely on SNPs were comparable to levels reported by Jarquin et al. (2016) for oil content and protein content, despite considerable differences in the cross-validation and modeling schemes. The lack of differences in prediction accuracies may be explained by the narrow genetic diversity in soybean breeding material due to genetic bottlenecks (Hyten et al., 2006).

Genomic prediction with LD-based haplotype blocks in canola resulted in the highest accuracy improvements for most model/trait combinations. Variation in prediction accuracy across LD thresholds was minimal. The optimal threshold varied significantly by trait and model, ranging from very low (0.01) to high (0.89). In wheat, LD-based haplotype blocks were superior to the other haplotyping methods for 20 out of 60 model/trait combinations, but accuracy didn't always improve compared to SNP-based prediction. Similar low variation across LD thresholds was observed in soybean. For soybean's oil content, the ideal LD threshold for accuracy estimates across all models was 0.24-0.26. In maize, only the Bayesian LASSO and RKHS models achieved the highest improvements with LD based haplotype blocks with a threshold of 1, effectively removing redundant information. In this scenario, only markers in complete LD were grouped into a block, effectively removing redundant information. This process, is similar to LD pruning, which has been demonstrated to enhance prediction accuracy (Ye et al., 2019). Intriguing patterns were observed with LD-based haplotypes in maize, the two models

implemented in a Bayesian framework (Bayesian LASSO and RKHS) behaved in an opposite direction to the other (frequentist) models, potentially due to different estimation procedures. In contrast to Cuyabano et al. (2014), we generally did not find an ideal LD threshold or even an ideal threshold specific to each dataset and mostly not even an ideal threshold within one trait. The prediction accuracy variation along LD thresholds reported in cattle (Cuyabano et al., 2014; Li et al., 2021; Li et al., 2022) was similar to the variation observed in our analyses. This suggests that any LD threshold is reasonable for genomic prediction due to low variation of prediction accuracy. We propose that even with extreme LD thresholds, reasonably accurate haplotype blocks are constructed, which explains the low variation observed across LD thresholds in all datasets. Additionally, in all datasets, the correlation between relationship coefficients obtained from markers and haplotypes was consistently high, with little variation across LD thresholds. This suggests that relationship representation remains consistent when using LD-based haplotype blocks.

The use of small fixed window blocks led to prediction accuracies comparable to those achieved with individual SNPs. Additionally, in maize, our findings aligned with those of Jiang et al. (2018) in Flint material, showing similar prediction accuracy patterns for frequentist models using small fixed window size haplotype blocks (2-5 markers). Interestingly, in maize, prediction accuracy eroded with the two frequentist models and increasing block size based on fixed windows, whereas for two Bayesian models the prediction accuracy was low across all parameters. Except for the wheat dataset, using excessively large fixed windows to build haplotype blocks considerably reduced prediction accuracy, as observed in previous studies with cattle (Hess et al., 2017). Unrealistically large blocks likely obscure the effects of true QTL within them. Furthermore, these larger blocks are generally more prone to errors in genotyping, and imputation, which accumulate in large blocks and limit prediction accuracy of genomic prediction models utilizing these blocks. These errors can also introduce false rare haplotypes, exacerbating issues related to rare variants. Additionally, as block size increases, haplotypes become more specific to genotypes or subpopulations, resulting in the absence of certain haplotypes in the training set but presence in the validation set. This lack of overlap leads to inaccurate estimation of the effects for those haplotypes, thus decreasing prediction accuracy due to the limited shared haplotypes between the training and validation sets. In the case of wheat, however, using very large blocks, such as whole chromosomes, resulted in considerable improvements in prediction accuracy. Mentionable improvements were observed for traits such as wheat stripe rust resistance, powdery mildew resistance, and kernel spike^{-1}. This improvement can likely be attributed to introgression breeding in wheat, where large chromosome segments are introgressed and preserved due to restricted recombination (Hao et al., 2020). Furthermore, the wheat D-subgenome exhibits large LD haplotype blocks that are important for yield and biomass-related traits (Voss-Fels et al., 2019). However, it should be noted that these improvements were observed in cases where the model performance was initially at a very low level with SNPs. The correlation between relationship coefficients obtained from markers and haplotypes was high for small fixed window blocks but decreased as block size increased. This suggests that crucial relationship information is lost or encoded within large haplotype blocks, which cannot be accessed for accurate prediction. As a result, the prediction accuracy in canola, maize, and soybean is reduced. However, it is important to highlight that large blocks can potentially introduce additional trait information, as demonstrated by their impact in some of the wheat traits.

The widely used algorithms implemented in *"Haploview"* did not exhibit superiority in terms of prediction accuracy compared to other methods. Although the method proposed by Gabriel et al. (2002) showed a slight improvement, particularly in maize DMY, these gains remained modest when compared to SNP-based prediction. In contrast to the findings of Matias et al. (2017), our analysis generally revealed a decrease in prediction accuracy rather than a benefit from haplotypes based on *"Haploview"* in the maize dataset. This discrepancy could be attributed to differences in the plant materials studied. While Matias et al. (2017) examined a diverse collection of tropical maize lines, our analyses focused on European dent material characterized by a relatively strong population structure (Lehermeier et al., 2014). Moreover, the population studied by Matias et al. (2017) was nearly twice the size of our investigation, potentially leading to increased recombination events between loci and reducing the potential size of haplotype blocks. Another contributing factor may be the limited representation of relationship captured by those haplotypes, as evidenced by the intermediate correlation between relationship coefficients obtained from markers and haplotypes. In contrast, canola, wheat, and soybean exhibited a high correlation in this regard. Unlike the findings of Ma et al. (2016) suggest, our study did not observe improved prediction accuracies in soybean using the method proposed by Gabriel et al. (2002). This discrepancy could be attributed to several factors, including differences in the traits under examination, as well as substantial variations in population size and marker density. It is worth noting that the method proposed by Gabriel et al. (2002) shares similarities with the LD-based method described earlier, implying that haplotype blocks formed using this method may already be represented using a specific LD threshold.

The *"HaploBlocker"* method (Pook et al., 2020) has the advantage of constructing subgroup-specific haplotype blocks and was implemented to address this aspect. However, this approach did not improve prediction accuracy and even led to a decrease of prediction accuracy in some cases. In canola and soybean, haplotype blocks from *"HaploBlocker"* effectively captured the genomic relationship represented by SNPs. In wheat, the representation was reasonable, but in maize, it was notably inadequate. Similar to the large fixed windows, haplotypes generated by this method are specific to genotypes or subpopulations. Consequently, haplotypes present in the validation set may not be observed in the training set, resulting in the inability to estimate their effects accurately and leading to decreased prediction accuracy due to the limited number of shared haplotypes between the training and validation sets. Particularly in the maize population, which exhibited strong population structure, the *"HaploBlocker"* method resulted in

comparatively low prediction accuracies. This was pronounced with the family-wise cross validation, where the accuracies were diminished to nearly zero. In this scenario, even when using SNPs, the number of shared alleles or haplotypes between the training and validation sets will be minimized. This effect will be particularly prominent when employing a method that constructs subgroup-specific blocks.

In general, with the exception of wheat, prediction accuracies based on haplotype blocks using GBLUP and EGBLUP followed the correlation observed between relationship coefficients obtained from SNPs and haplotype blocks. This suggests that a portion of the prediction accuracy achieved with haplotypes is derived from reinterpreting the SNP information. However, in the case of wheat, this pattern did not hold true, even when using large fixed window blocks. Furthermore, considerable prediction accuracy differences were observed across models for wheat traits, but these differences were consistently compensated for by utilizing haplotype blocks with varying methods and parameters. This indicates that additional information beyond genetic relatedness contributes to the prediction accuracy when using haplotype blocks. One possible explanation is that haplotype blocks are generally considered to exhibit higher LD with QTL compared to individual markers (Jiang et al., 2018).

Multiple factors contribute to the accuracy of genomic prediction. One crucial factor is the relationship among genotypes, which is overlooked in random cross-validation approaches. In such cases, closely related genotypes may be included in both the training and validation sets, leading to higher prediction accuracies for related individuals (Massman et al., 2013; Hickey et al., 2014; Werner et al., 2020). Consequently, the prediction accuracies obtained from random cross-validation are population-specific and cannot be readily adopted to all breeding populations (Werner et al., 2020). To address this issue, we conducted a family-wise cross-validation in the maize dataset to assess the predictive performance of haplotype blocks for less related individuals. As expected from Werner et al. (2020), we observed a decrease in prediction accuracy compared to random cross-validation. However, the relative ranking of haplotype block methods and parameters remained consistent with that of the random cross-validation, indicating no added benefit from haplotypes in predicting the breeding values of genetically distinct materials.

Moreover, GBLUP models trained with small haplotype blocks exhibited very similar prediction accuracies to models trained with SNPs. This is expected since haplotype effects can be partially defined as the sum of individual marker effects within their respective block. Another advantage of haplotype effects is their ability to capture local epistasis, as demonstrated by Jiang et al. (2018). However, it is worth noting that purely additive models, especially in prediction methods like GBLUP where marker effects are estimated simultaneously, already implicitly capture local epistasis among markers in complete LD.

The use of haplotypes has been proposed as a means to address the challenges associated with apparent or phantom epistasis (Wood et al., 2014). Apparent or phantom epistasis can occur when two markers are in incomplete LD with QTL, resulting in statistically significant marker interactions in association studies and enhanced prediction accuracies in genomic prediction with models considering epistasis (Wood et al., 2014; de los Campos et al., 2019; Schrauf et al., 2020). This effect may be particularly pronounced in the wheat dataset, which had a significantly lower marker density compared to the other three datasets. Consequently, the use of haplotype blocks sometimes led to considerable improvements in prediction accuracy.

There is a multitude of factors affecting the accurate assembly of haplotype blocks and their respective haplotypes. Especially in complex plant genomes like the allopolyploids canola and wheat, SNP array markers can potentially be non-specific in terms of physical position, representing different homoeologous loci in different individuals (Mason et al., 2017; Makhoul et al., 2020). Furthermore, all methods to build haplotype blocks rely on known marker positions along the genome. These positions are obtained from a reference genome and are not necessarily the same in every population or even genotype. Especially if the reference genome is only distantly related. In such cases, a lack of precision in assembled haplotype blocks and their corresponding haplotypes may limit their potential in genomic prediction. Furthermore, haplotype block borders are not necessarily the same across populations and generations. Even though, Gabriel et al. (2002) showed high harmony of block structure across different human populations, however in plant breeding, with selection favoring positive alleles or haplotypes, this could ultimately change. Especially LD based haplotype blocks may only be useful for very few generations, since initially defined blocks will rapidly be disrupted by recombination or extended due to selection in later generations as the breeding program progresses. Indeed, an important goal of breeding is to accumulate favorable alleles through selection and recombination. This underlines the need for constant updating of both, the haplotype block assignment and the prediction model. Furthermore, besides the two fixed window approaches, all of the methods tested are only capable of identifying a proxy to true chromosomal recombination breakpoints. Even though crossovers tends to aggregate in recombination hotspots (Li and Stephens, 2003; Myers et al., 2006), haplotype blocking methods with limited marker density and population size may not necessarily be able to detect these hotspots. Therefore, there is a need to develop enhanced haplotype blocking pipelines that can effectively capture natural recombination patterns and address challenges associated with polyploidy, structural variations, and chromosomal rearrangements commonly observed in crop plants (Mason et al., 2017; Schiessl et al., 2019). Consequently, ongoing efforts focus on the development of innovative methods to capture local epistatic effects (Pook et al., 2020).

Unfortunately, we could not identify a single optimal haplotype blocking method that suits all datasets. Therefore, it is important to consider haplotype block construction methods and parameters as hyperparameters that require careful optimization, rather than fixed biological parameters. A breeding program that adopts haplotype block-based genomic selection should explore multiple haplotype blocking methods with different parameter settings. In general, the selected method should effectively capture relationships among

individuals. Additionally, it is worth examining blocks of large size, as, in the case of the wheat dataset, larger blocks proved beneficial in improving prediction accuracy. The wheat dataset, which had the lowest marker density, generally showed the greatest improvements. This suggests that haplotype block-based genomic selection could be particularly valuable for breeding programs lacking access to high-density SNP arrays. However, further investigation is required in other datasets with varying SNP densities to validate these findings.

Although we observed only marginal beneficial effects of haplotype blocks in the canola, maize and soybean datasets on genomic prediction, they can still have a beneficial effect when used in other contexts. For example, haplotype blocks can help to identify regions of interest for the identification of candidate genes near significant marker-trait associations, or to compare different genotype groups at such loci (Clark, 2004; Li et al., 2017; Vollrath et al., 2021). Moreover, even if the majority of SNP markers exhibit intermediate minor allele frequency in a population, specific combinations of alleles represented as haplotypes may not be common in a population. Therefore, haplotypes can assist in identifying rare variants that have a potential impact on phenotypic traits. (Bloom et al., 2019; Wainschtein et al., 2022; Wang et al., 2023). Furthermore, especially in highly quantitative traits like yield where markers tend to have very small effects on traits, haplotype blocks can identify positive or negative chromosomal segments. This information can be implemented for cross designs to recombine haplotypes with positive effects (Bernardo and Thompson, 2016; Werner et al., 2018a). This can be considerably easier than selecting for single positive SNPs, as their positive effect can be obscured by deleterious SNPs in proximity that are only rarely separated by recombination in subsequent generations.

5 Conclusion

As anticipated based on numerous previous reports, our study confirms that haplotype blocks have the potential to enhance genomic selection, although the magnitude of improvement is sometimes only marginal. Haplotype blocks can particularly compensate for model differences when there is considerable variation in model performance across different prediction models. The extent of improvement with haplotypes compared to SNP-based predictions seem to be highly dependent on factors such as population, population structure, trait, and model. for a multitude of different traits from different crop species with different genome properties and breeding schemes, we were unable to identify optimal methods or parameters for constructing haplotype blocks in terms of prediction accuracy. Approaches based on LD resulted in improved prediction accuracies across various traits and demonstrated robustness in LD-threshold selection. However, the greatest improvements were observed with haplotype blocks consisting of entire chromosomes. Therefore, we recommend treating haplotype block definition as a tunable hyperparameter when employing genomic selection, taking into account extremely large haplotype blocks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: Please refer to the original publications of the four datasets.

Author contributions

SW and RS designed the study. SW conceived the analysis, MF developed the software for Linkage Disequilibrium (LD) based haplotyping. KV-F and MF supervised the statistical analysis. SW wrote the manuscript. RS and KV-F 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. Informatics infrastructure was provided by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics (de.NBI).

Acknowledgments

The authors thank Benjamin Wittkop, Christian Obermeier, Carola Zenke-Philippi and Lennard Ehrig for discussions on potential applications of haplotype blocks 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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1217589/full#supplementary-material

References

Atanda, S. A., Govindan, V., Singh, R., Robbins, K. R., Crossa, J., and Bentley, A. R. (2022). Sparse testing using genomic prediction improves selection for breeding targets in elite spring wheat. *Theor. Appl. Genet.* 135, 1939–1950. doi: 10.1007/s00122-022-04085-0

Ballesta, P., Maldonado, C., Pérez-Rodríguez, P., and Mora, F. (2019). SNP and haplotype-based genomic selection of quantitative traits in eucalyptus globulus. *Plants* 8, 331. doi: 10.3390/plants8090331

Bandillo, N., Jarquin, D., Song, Q., Nelson, R., Cregan, P., Specht, J., et al. (2015). A population structure and genome-wide association analysis on the USDA soybean germplasm collection. *Plant Genome* 8, plantgenome2015.04.0024. doi: 10.3835/plantgenome2015.04.0024

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* 21, 263–265. doi: 10.1093/bioinformatics/bth457

Bauer, E., Falque, M., Walter, H., Bauland, C., Camisan, C., Campo, L., et al. (2013). Intraspecific variation of recombination rate in maize. *Genome Biol.* 9 (14). doi: 10.1186/gb-2013-14-9-r103

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

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

Bernardo, R., and Thompson, A. M. (2016). Germplasm architecture revealed through chromosomal effects for quantitative traits in maize. *Plant Genome* 9, plantgenome2016.03.0028. doi: 10.3835/plantgenome2016.03.0028

Bloom, J. S., Boocock, J., Treusch, S., Sadhu, M. J., Day, L., Oates-Barker, H., et al. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. *eLife* 8, e49212. doi: 10.7554/eLife.49212

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

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., and Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* 74, 106–120. doi: 10.1086/381000

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genet. Epidemiol.* 27, 321–333. doi: 10.1002/gepi.20025

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

Combs, E., and Bernardo, R. (2013). Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. *Plant Genome* 6, plantgenome2012.11.0030. doi: 10.3835/plantgenome2012.11.0030

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

Crespo-Herrera, L., Howard, R., Piepho, H.-P., Pérez-Rodríguez, P., Montesinos-Lopez, O., Burgueño, J., et al. (2021). Genome-enabled prediction for sparse testing in multi-environmental wheat trials. *Plant Genome* 14, e20151. doi: 10.1002/tpg2.20151

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 Genes|Genomes|Genetics* 3, 1903–1926. doi: 10.1534/g3.113.008227

Crossa, J., Jarquín, D., Franco, J., Pérez-Rodríguez, P., Burgueño, J., Saint-Pierre, C., et al. (2016). Genomic prediction of gene bank wheat landraces. *G3 Genes|Genomes|Genetics* 6, 1819–1834. doi: 10.1534/g3.116.029637

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

Cuyabano, B. C., Su, G., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15, 1171. doi: 10.1186/1471-2164-15-1171

Cuyabano, B. C., Su, G., and Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Selection Evol.* 47, 61. doi: 10.1186/s12711-015-0143-3

Daetwyler, H. D., Bansal, U. K., Bariana, H. S., Hayden, M. J., and Hayes, B. J. (2014). Genomic prediction for rust resistance in diverse wheat landraces. *Theor. Appl. Genet.* 127, 1795–1803. doi: 10.1007/s00122-014-2341-8

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat. Genet.* 29, 229–232. doi: 10.1038/ng1001-229

de Bakker, P. I. W., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat. Genet.* 37, 1217–1223. doi: 10.1038/ng1669

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

Devlin, B., and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29, 311–322. doi: 10.1006/geno.1995.9003

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* 112, 39–47. doi: 10.1038/hdy.2013.13

Edwards, D., and Batley, J. (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.* 8, 2–9. doi: 10.1111/j.1467-7652.2009.00459.x

Edwards, S. M., Buntjer, J. B., Jackson, R., Bentley, A. R., Lage, J., Byrne, E., et al. (2019). The effects of training population design on genomic prediction accuracy in wheat. *Theor. Appl. Genet.* 132, 1943–1952. doi: 10.1007/s00122-019-03327-y

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424

Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. *Genetics* 194, 573–596. doi: 10.1534/genetics.113.151753

Goddard, M. E., and Hayes, B. j. (2007). Genomic selection. *J. Anim. Breed. Genet.* 124, 323–330. doi: 10.1111/j.1439-0388.2007.00702.x

Grant, D., Nelson, R. T., Cannon, S. B., and Shoemaker, R. C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.* 38, D843–D846. doi: 10.1093/nar/gkp798

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

Hao, M., Zhang, L., Ning, S., Huang, L., Yuan, Z., Wu, B., et al. (2020). The resurgence of introgression breeding, as exemplified in wheat improvement. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.00252

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci.* 92, 433–443. doi: 10.3168/jds.2008-1646

Hayes, B. J., Macleod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlian, A. J., Vander Jagt, C. J., et al. (2014)Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. in *10*. In: *World Congress of Genetics Applied to Livestock Production* (Vancouver, Canada) (Accessed November 22, 2022).

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

Hess, M., Druet, T., Hess, A., and Garrick, D. (2017). Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. *Genet. Selection Evol.* 49, 54. doi: 10.1186/s12711-017-0329-y

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat. Genet.* 49, 1297–1303. doi: 10.1038/ng.3920

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. *Crop Sci.* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195

Hill, W. G. (1981). Estimation of effective population size from data on linkage disequilibrium1. *Genet. Res.* 38, 209–216. doi: 10.1017/S0016672300020553

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

Hofheinz, N., and Frisch, M. (2014). Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3 Genes|Genomes|Genetics* 4, 539–546. doi: 10.1534/g3.113.010025

Hyten, D. L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R. L., Costa, J. M., et al. (2006). Impacts of genetic bottlenecks on soybean genome diversity. *Proc. Natl. Acad. Sci.* 103, 16666–16671. doi: 10.1073/pnas.0604379103

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

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic prediction enhanced sparse testing for multi-environment trials. *G3 Genes|Genomes|Genetics* 10, 2725–2739. doi: 10.1534/g3.120.401349

Jarquin, D., Specht, J., and Lorenz, A. (2016). Prospects of genomic prediction in the USDA soybean germplasm collection: historical data creates robust models for enhancing selection of accessions. *G3 Genes|Genomes|Genetics* 6, 2329–2341. doi: 10.1534/g3.116.031443

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.* 29, 217–222. doi: 10.1038/ng1001-217

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

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3 (Bethesda)* 8, 1687–1699. doi: 10.1534/g3.117.300548

Kim, S. A., Brossard, M., Roshandel, D., Paterson, A. D., Bull, S. B., and Yoo, Y. J. (2019). gpart: human genome partitioning and visualization of high-density SNP data by identifying haplotype blocks. *Bioinformatics* 35, 4419–4421. doi: 10.1093/bioinformatics/btz308

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., et al. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3 Genes|Genomes|Genetics* 3, 2105–2114. doi: 10.1534/g3.113.007807

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

Li, T., Ma, X., Li, N., Zhou, L., Liu, Z., Han, H., et al. (2017). Genome-wide association study discovered candidate genes of Verticillium wilt resistance in upland cotton (Gossypium hirsutum L.). *Plant Biotechnol. J.* 15, 1520–1532. doi: 10.1111/pbi.12734

Li, N., and Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* 165, 2213–2233. doi: 10.1093/genetics/165.4.2213

Li, H., Wang, Z., Xu, L., Li, Q., Gao, H., Ma, H., et al. (2022). Genomic prediction of carcass traits using different haplotype block partitioning methods in beef cattle. *Evolutionary Appl.* 15, 2028–2042. doi: 10.1111/eva.13491

Li, H., Zhu, B., Xu, L., Wang, Z., Xu, L., Zhou, P., et al. (2021). Genomic prediction using LD-based haplotypes inferred from high-density chip and imputed sequence variants in Chinese simmental beef cattle. *Front. Genet.* 12. doi: 10.3389/fgene.2021.665382

Liang, Z., Tan, C., Prakapenka, D., Ma, L., and Da, Y. (2020). Haplotype analysis of genomic prediction using structural and functional genomic information for seven human phenotypes. *Front. Genet.* 11. doi: 10.3389/fgene.2020.588907

Liu, F., Schmidt, R. H., Reif, J. C., and Jiang, Y. (2019). Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping. *G3 Genes|Genomes|Genetics* 9, 4115–4126. doi: 10.1534/g3.119.400451

Lorenz, A. J., Hamblin, M. T., and Jannink, J.-L. (2010). Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One* 5, e14079. doi: 10.1371/journal.pone.0014079

Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., et al. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). *Mol. Breed.* 36, 113. doi: 10.1007/s11032-016-0504-9

Makhoul, M., Rambla, C., Voss-Fels, K. P., Hickey, L. T., Snowdon, R. J., and Obermeier, C. (2020). Overcoming polyploidy pitfalls: a user guide for effective SNP conversion into KASP markers in wheat. *Theor. Appl. Genet.* 133, 2413–2430. doi: 10.1007/s00122-020-03608-x

Maldonado, C., Mora, F., Bertagna, F. A. B., Kuki, M. C., and Scapim, C. A. (2019). SNP- and haplotype-based GWAS of flowering-related traits in maize with network-assisted gene prioritization. *Agronomy* 9, 725. doi: 10.3390/agronomy9110725

Mason, A. S., Higgins, E. E., Snowdon, R. J., Batley, J., Stein, A., Werner, C., et al. (2017). A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. *Theor. Appl. Genet.* 130, 621–633. doi: 10.1007/s00122-016-2849-1

Massman, J. M., Gordillo, A., Lorenzana, R. E., and Bernardo, R. (2013). Genomewide predictions from maize single-cross data. *Theor. Appl. Genet.* 126, 13–22. doi: 10.1007/s00122-012-1955-y

Matias, F. I., Galli, G., Correia Granato, I. S., and Fritsche-Neto, R. (2017). Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Sci.* 57, 2951–2958. doi: 10.2135/cropsci2017.01.0022

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

Myers, S., Spencer, C. C. A., Auton, A., Bottolo, L., Freeman, C., Donnelly, P., et al. (2006). The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.* 34, 526–530. doi: 10.1042/BST0340526

Ni, G., Cavero, D., Fangmann, A., Erbe, M., and Simianer, H. (2017). Whole-genome sequence-based genomic prediction in laying chickens with different genomic relationship matrices to account for genetic architecture. *Genet. Selection Evol.* 49, 8. doi: 10.1186/s12711-016-0277-y

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. *G3 Genes|Genomes|Genetics* 8, 2889–2899. doi: 10.1534/g3.118.200311

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

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

Pook, T., Freudenthal, J., Korte, A., and Simianer, H. (2020). Using local convolutional neural networks for genomic prediction. *Front. Genet.* 11. doi: 10.3389/fgene.2020.561497

Pook, T., Schlather, M., de los Campos, G., Mayer, M., Schoen, C. C., and Simianer, H. (2019). HaploBlocker: creation of subgroup-specific haplotype blocks and libraries. *Genetics* 212, 1045–1061. doi: 10.1534/genetics.119.302283

Raymond, B., Bouwman, A. C., Schrooten, C., Houwing-Duistermaat, J., and Veerkamp, R. F. (2018). Utility of whole-genome sequence data for across-breed genomic prediction. *Genet. Sel Evol.* 50, 27. doi: 10.1186/s12711-018-0396-8

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199–204. doi: 10.1038/35075590

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

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183. doi: 10.1038/nature08670

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

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *J. Anim. Sci.* 86, 2447–2454. doi: 10.2527/jas.2007-0010

Soleimani, B., Lehnert, H., Keilwagen, J., Plieske, J., Ordon, F., Naseri Rad, S., et al. (2020). Comparison between core set selection methods using different illumina marker platforms: A case study of assessment of diversity in wheat. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.01040

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and evaluation of soySNP50K, a high-density genotyping array for soybean. *PloS One* 8, e54985. doi: 10.1371/journal.pone.0054985

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

Terraillon, J., Frisch, M., Falke, K. C., Jaiser, H., Spiller, M., Cselényi, L., et al. (2022). Genomic prediction can provide precise estimates of the genotypic value of barley lines evaluated in unreplicated trials. *Front. Plant Sci.* 13. doi: 10.3389/fpls.2022.735256

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

VanLiere, J. M., and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theor. Population Biol.* 74, 130–137. doi: 10.1016/j.tpb.2008.05.006

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

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., et al. (2009). Invited Review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92, 16–24. doi: 10.3168/jds.2008-1514

Villumsen, T. M., and Janss, L. (2009). Bayesian genomic selection: the effect of haplotype length and priors. *BMC Proc.* 3, S11. doi: 10.1186/1753-6561-3-S1-S11

Villumsen, T., Janss, L., and Lund, M. s. (2009). The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126, 3–13. doi: 10.1111/j.1439-0388.2008.00747.x

Vollrath, P., Chawla, H. S., Alnajar, D., Gabur, I., Lee, H., Weber, S., et al. (2021). 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

Voss-Fels, K. P., Stahl, A., Wittkop, B., Lichthardt, C., Nagler, S., Rose, T., et al. (2019). Breeding improves wheat productivity under contrasting agrochemical input levels. *Nat. Plants* 5, 706–714. doi: 10.1038/s41477-019-0445-5

Voss-Fels, K. P., Wei, X., Ross, E. M., Frisch, M., Aitken, K. S., Cooper, M., et al. (2021). Strategies and considerations for implementing genomic selection to improve traits with additive and non-additive genetic architectures in sugarcane breeding. *Theor. Appl. Genet.* 134, 1493–1511. doi: 10.1007/s00122-021-03785-3

Wainschtein, P., Jain, D., Zheng, Z., Cupples, L. A., Shadyab, A. H., McKnight, B., et al. (2022). Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat. Genet.* 54, 263–273. doi: 10.1038/s41588-021-00997-7

Wang, N., Akey, J. M., Zhang, K., Chakraborty, R., and Jin, L. (2002). Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *Am. J. Hum. Genet.* 71, 1227–1234. doi: 10.1086/344398

Wang, F., Moon, W., Letsou, W., Sapkota, Y., Wang, Z., Im, C., et al. (2023). Genome-wide analysis of rare haplotypes associated with breast cancer risk. *Cancer Res.* 83, 332–345. doi: 10.1158/0008-5472.CAN-22-1888

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., et al. (2020). How population structure impacts genomic selection accuracy in cross-validation: implications for practical breeding. *Front. Plant Sci.* 11. doi: 10.3389/fpls.2020.592977

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

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., et al. (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genet. Selection Evol.* 43, 5. doi: 10.1186/1297-9686-43-5

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

Ye, S., Gao, N., Zheng, R., Chen, Z., Teng, J., Yuan, X., et al. (2019). Strategies for obtaining and pruning imputed whole-genome sequence data for genomic prediction. *Front. Genet.* 10. doi: 10.3389/fgene.2019.00673

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., et al. (2011). Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS One* 6, e17595. doi: 10.1371/journal.pone.0017595

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* 42, 355–360. doi: 10.1038/ng.546

Zhang, Q., Sahana, G., Su, G., Guldbrandtsen, B., Lund, M. S., and Calus, M. P. L. (2018). Impact of rare and low-frequency sequence variants on reliability of genomic prediction in dairy cattle. *Genet. Selection Evol.* 50, 62. doi: 10.1186/s12711-018-0432-8

Zhang, A., Wang, H., Beyene, Y., Semagn, K., Liu, Y., Cao, S., et al. (2017). Effect of trait heritability, training population size and marker density on genomic prediction accuracy estimation in 22 bi-parental tropical maize populations. *Front. Plant Sci.* 8. doi: 10.3389/fpls.2017.01916

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

Zimin, A. V., Puiu, D., Hall, R., Kingan, S., Clavijo, B. J., and Salzberg, S. L. (2017). The first near-complete assembly of the hexaploid bread wheat genome, Triticum aestivum. *GigaScience* 6, gix097. doi: 10.1093/gigascience/gix097

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

Weber, S. E., Chawla, H. S., Ehrig, L., Hickey, L. T., Frisch, M., Snowdon, R. J. (2023). *Front. Plant Sci. 14:1221750. doi: 10.3389/fpls.2023.1221750*

frontiers | Frontiers in Plant Science

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

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

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 (Belóet 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 $(0\text{-}2p_i)$ for homozygous allele A, $(1\text{-}2p_i)$ for the heterozygous state, and $(2\text{-}2p_i)$ 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 ($0\text{-}2p_i$) for successful allele calls and ($2\text{-}2p_i$) 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:

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

with

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

where y is the vector of observations, while $\boldsymbol{K}_{ul}$ represents an $n \times n$ 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:

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

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

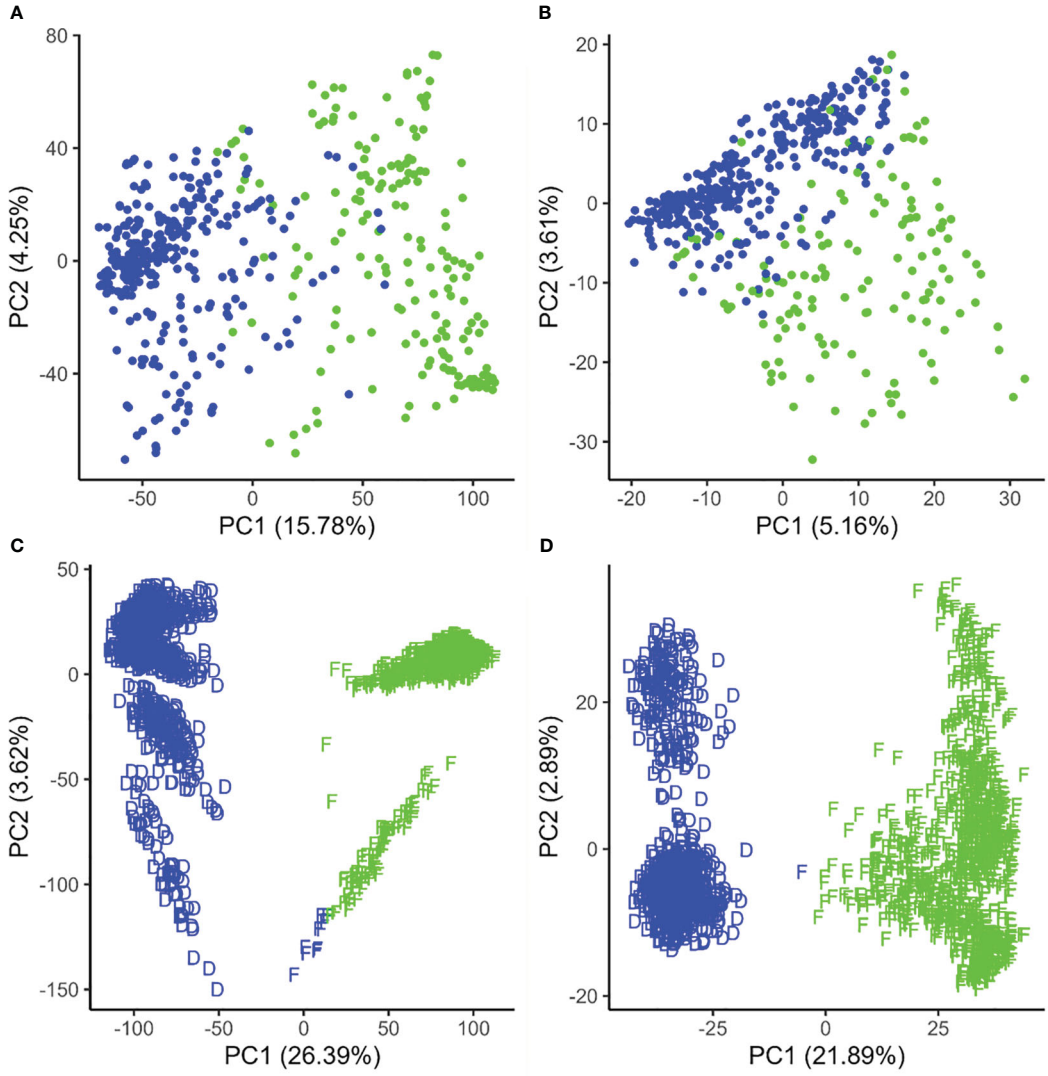

FIGURE 1
Population structure displayed by the first two principal components of the genetic distance in canola **(A, B)** and maize **(C, D)** based on standard single-nucleotide polymorphisms **(A, C)** and failed allele calls **(B, D)**. The colors (red and blue) represent clusters based on k-means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

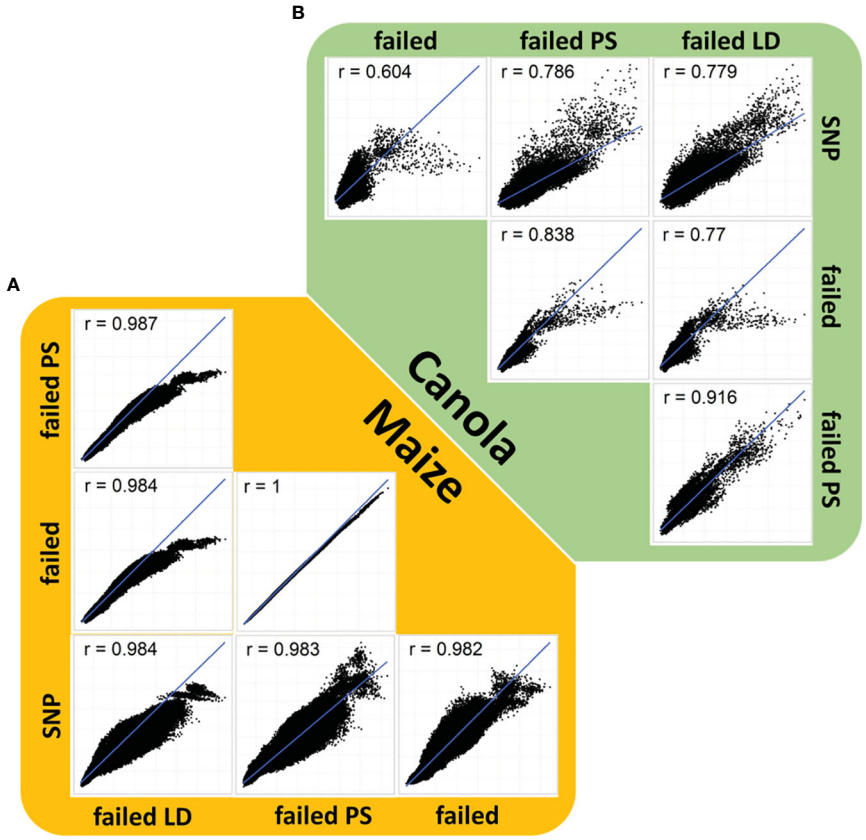

FIGURE 2
Correlation plot of genomic relationship coefficients based on single-nucleotide polymorphisms, failed allele calls (failed), failed allele calls filtered by pool specificity (failed PS), and failed allele calls filtered by LD (failed LD) in **(A)** canola (green) and **(B)** maize (orange).

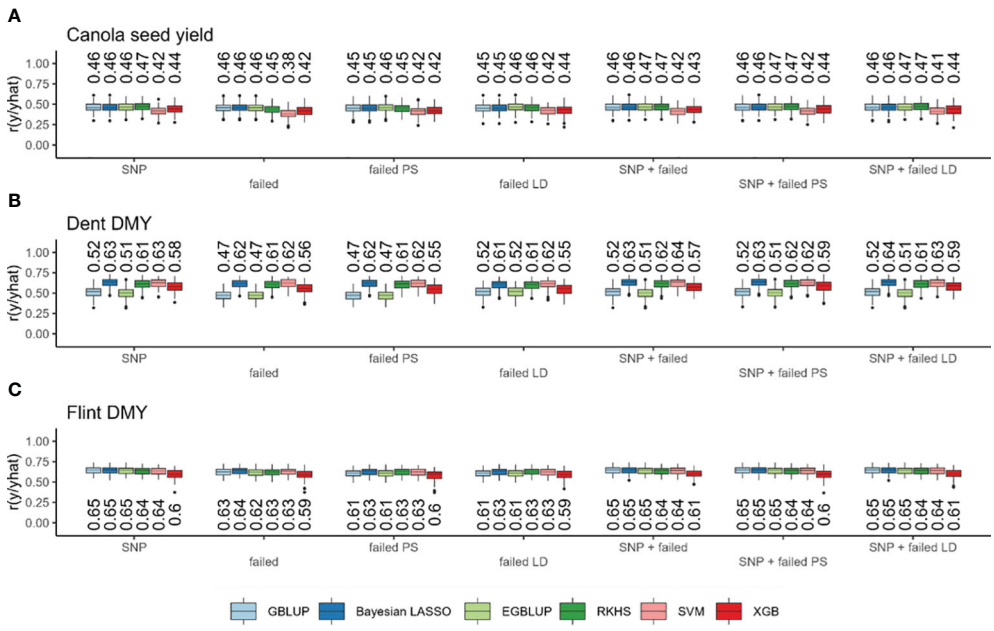

FIGURE 3
Prediction accuracy (*r*) based on standard single-nucleotide polymorphisms (SNPs), failed SNP calls (failed), failed SNP calls filtered by pool specificity (failed PS), and failed SNP calls filtered by LD (failed LD) as well as their combination with GBLUP (light blue), Bayesian Lasso (dark 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.

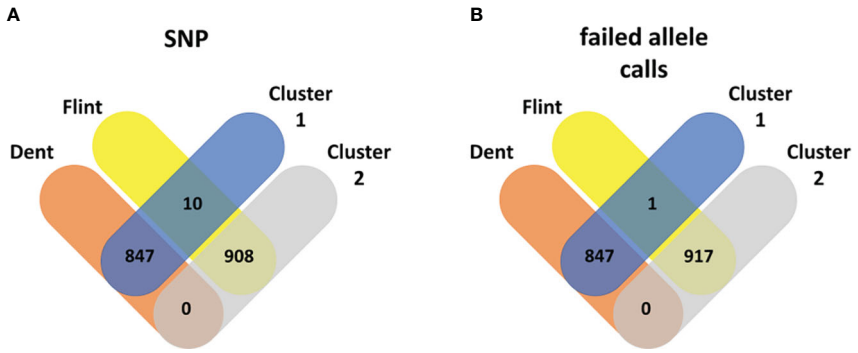

FIGURE 4
Venn diagram: Maize pool assignment to Dent (red) and Flint (yellow) subpools vs. pool assignment based on k-means clustering into cluster 1 (blue) and cluster 2 (gray) based on the genetic distance from standard single-nucleotide polymorphisms **(A)** and failed allele calls **(B)**.

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

4 Genomic Prediction in Brassica napus: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data

Weber, S. E., Roscher-Ehrig, L., Kox, T., Abbadi, A., Stahl, A. and Snowdon, R. J. (2023). under Review in *Genome*

Canadian Science Publishing | Genome

Genomic Prediction in Brassica napus: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data

Journal:	*Genome*
Manuscript ID	gen-2023-0126
Manuscript Type:	Article
Date Submitted by the Author:	30-Nov-2023
Complete List of Authors:	Weber, Sven; Justus Liebig University Giessen, Department of Plant Breeding Roscher-Ehrig, Lennard; Justus Liebig University Giessen, Department of Plant Breeding Kox, Tobias; NPZ innovation GmbH, NPZ Innovation GmbH abbadi, amine; NPZ Innovation GmbH , NPZ Innovation GmbH Stahl, Andreas; Julius Kühn-Institut, Institute for Resistance Research and Stress Tolerance Snowdon, Rod; Justus-Liebig-Universitat Giessen, Department of Plant Breeding
Is the manuscript for consideration in a Special Issue or Collection?:	Oilseed Brassica napus Breeding: From Genomics to Phenomics
Keyword:	Genomic prediction, Imputation, whole-genome sequencing, SNP markers

Genomic Prediction in *Brassica napus*: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data

Sven E. Weber[1]*, Lennard Ehrig[1], Tobias Kox[2], Amine Abbadi[2], Andreas Stahl[3] and Rod J. Snowdon[1]

[1]Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University, Giessen, Germany

[2]NPZ Innovation GmbH, Holtsee, Germany

[3]Julius Kuehn Institute (JKI), Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Quedlinburg, Germany

* Corresponding author:

Sven E. Weber

Sven.E.Weber@agrar.uni-giessen.de

Note: Please find the Figures at the End of the manuscript

Abstract

Advances in sequencing technology allow whole plant genomes to be sequenced with high quality. Combining genotypic and phenotypic data in genomic prediction helps breeders to select crossing partners in partially phenotyped populations. In plant breeding programs, the cost of sequencing entire breeding populations still exceeds available genotyping budgets. Hence, the method for genotyping are still mainly SNP arrays; however, arrays are unable to assess the entire genome- and population-wide diversity. A compromise involves genotyping the entire population using a SNP array and a subset of the population with whole-genome sequencing. Both datasets can then be used to impute markers from whole-genome sequencing onto the entire population. Here, we evaluate whether imputation of whole-genome sequencing data enhances genomic predictions, using data from a nested association mapping population of rapeseed (*Brassica napus*). Employing two cross-validation schemes that mimic scenarios for the prediction of close and distant relatives, we show that imputed marker data does not significantly improve prediction accuracy, likely due to redundancy in relationship estimates and imputation errors. In simulation studies, only small improvements were observed, further corroborating the findings. We conclude that SNP arrays are already equipped with the information that is added by imputation through relationship and linkage disequilibrium.

Keywords

Genomic prediction, Imputation, whole-genome sequencing, SNP markers

1 Introduction

Since the pioneering sequencing and assembly of the first plant genome of *Arabidopsis thaliana* by The Arabidopsis Genome Initiative (2000), the landscape of genomics in plant science has been profoundly transformed. Today, it is possible to assemble, resequence and comprehensively characterize the genome of virtually any plant species. This has also allowed plant science to systematically exploit genomic information and connect it to important phenotypic traits like diseases, grain quality or crop yield.

This revolutionary shift led to the development of a multitude of statistical methods aimed at bridging the gap between phenotypic data and genomic information. Notably, one of the most compelling innovations in plant breeding, adapted from animal breeding, is genomic prediction, which has emerged as the gold standard for identifying genetically superior

accessions within breeding materials. Henderson (1975) was likely one of the first scientists to use relatedness based on pedigree information, along with phenotypic data, for breeding value prediction in a mixed linear model framework. The rapid advancements in genome sequencing technologies over the past three decades have revolutionized the field, enabling the use of genomic data to replace traditional pedigree relationships in statistical prediction models (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008) Today's precision in genome sequencing has reached remarkable levels, empowering researchers to pinpoint millions of polymorphisms across the genome with high quality and confidence. This wealth of genomic information, when combined with phenotypic measurements and utilizing statistical methods, enables plant breeders to predict the genotypic values of individuals who have not undergone phenotypic assessment (Lande and Thompson, 1990; Meuwissen et al., 2001). The statistical methods employed utilize phenotypic and genotypic information from a select group of individuals, often referred to as the "training population," to predict trait performance in individuals for which only genotypic data is collected. Over the years, this approach has given rise to a broad range of mathematical models designed for genomic prediction, including genomic best linear unbiased prediction (GBLUP; Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008), extensions to GBLUP (Jiang and Reif, 2015; Jiang et al., 2018) and models from the “Bayesian alphabet” like Bayesian LASSO (Park and Casella, 2008) or Bayesian ridge regression (Pérez and de los Campos, 2014). Model differences can be attributed to their assumption of variance components, marker effects, marker modes of action and model assumptions.

Genotypic information utilized for genomic prediction normally comprises biallelic single nucleotide polymorphisms (SNPs) that are very abundant in crop genomes (Rafalski, 2002; Frazer et al., 2007; Ganal et al., 2009). In large-scale breeding operations in major crops, breeding populations are frequently genotyped with SNP arrays, which provide a simple, cost-effective and highly reproducible high-throughput technology to assay genotypes on a population scale with a fixed set of genomic markers that were selected during the array development. SNP arrays vary in marker number and can range from a few thousand to more than a million SNP markers. Previous studies of genomic prediction in canola and rapeseed (*Brassica napus*) utilized SNP arrays with 10,000 to 30,000 usable SNPs (Werner et al., 2018a, 2020; Knoch et al., 2021; Weber et al., 2023b), which represents only a fraction of all genomic polymorphisms present in breeding populations and the broader gene pool. With strongly

limited marker numbers, and depending on the population structure and linkage disequilibrium (LD) between genomic loci, genomic prediction accuracy could potentially be limited, because not all quantitative trait loci (QTL) for highly complex traits can be expected to be sufficiently linked with markers from the array. Furthermore, with a limited marker density we may be unable to identify all variants and allelic combinations of genes that contribute to a particular trait, since most genes carry multiple sequence polymorphisms.

Moreover, accurate genomic prediction is often obtained by considering relatives (VanRaden, 2008; Hayes et al., 2009) and accuracy generally drops with increasingly unrelated validation individuals (Wolc et al., 2011; Habier et al., 2013). This implies that SNPs from arrays are not necessarily in LD with all causal QTL and the prediction accuracy is partly driven by capturing genetic relationships among individuals. Hence, one strategy to improve prediction accuracy is to increase marker density in order to capture more QTL, either directly or as a result of LD. With the advance of whole-genome sequencing technologies, increasingly large and dense marker datasets can today be generated for most major crops using whole-genome sequencing (Edwards and Batley, 2010; Yu et al., 2011; Edwards et al., 2013). Furthermore, ongoing technological advancements and growing demand continue to reduce the cost of sequencing individual genomes. Nevertheless, it is still financially unviable to comprehensively sequence entire breeding populations or offspring from controlled crosses, even with the most state-of-the-art whole-genome sequencing technology at our disposal. The challenge is particularly pronounced when dealing with polyploid crop species like *Brassica napus* (oilseed rape/canola), where a high degree of genome duplication and high similarity between the two subgenomes demand special care and extensive sequencing coverage for accurate alignment of genomic variations to their precise locations on the genome (Makhoul et al., 2020).

Alongside the advances in sequencing technology, several pipelines have been developed to impute genomic information of individuals in a population that have been genotyped with different sets of markers, provided there is some marker overlap between individuals. Hence, a potential alternative to whole-genome sequencing of a complete population is to assay a subset of the breeding population with whole-genome sequencing and the entire breeding population with a SNP array. Subsequently, these two datasets can be used to impute markers from whole-genome sequencing onto the complete breeding population based on the subset of whole-genome sequenced genotypes. There are numerous methods for imputation which rely on allele frequencies, LD, haplotypes and flanking marker information, for example

"BEAGLE" (Browning and Browning, 2007; Browning et al., 2018), "SHAPEIT" (Delaneau et al., 2012) and "IMPUTE2" (Howie et al., 2011).

Thus, combining imputation for whole-genome sequencing data with genomic prediction could be a potential solution in the trade-off between sequencing costs and genomic prediction accuracy. Several studies in animal breeding achieved promising results, demonstrating that prediction accuracy in genomic prediction was comparable to real whole-genome sequencing data when this data was imputed from a smaller marker set (Zhang and Druet, 2010; Berry and Kearney, 2011; Cleveland and Hickey, 2013; Tsai et al., 2017; Song et al., 2019; Mancin et al., 2021; Kriaridou et al., 2023). In plant breeding however, studies mostly focused on imputations with regard to imputing missing marker calls (Crossa et al., 2013; Rutkoski et al., 2013; Wang et al., 2016; Edriss et al., 2017; Munyengwa et al., 2021). Only a few studies focus on imputation of unobserved markers from low to high density SNP arrays (Hickey et al., 2012), but the aspect of imputation of whole genome sequencing marker data has not been extensively studied, although simulations indicate beneficial effects especially with regard to genomic prediction accuracy and return on investment (Hickey et al., 2015; Gorjanc et al., 2017a, 2017b).

Here, we evaluate the potential benefits of imputing whole-genome sequencing marker data to enhance the accuracy of genomic predictions in *B. napus*. We leverage data from a *B. napus* Nested Association Mapping (BnNAM) population comprising 46 founder lines. The founders in this population were initially sequenced using the Illumina HiSeq 2500 platform (Snowdon et al., 2015), which closely mirrors the initial phase of a breeding program, providing comprehensive information on all founders. Our primary objective is to determine whether the incorporation of imputed whole-genome sequencing markers can lead to improved genomic prediction accuracy. To achieve this, we employ various cross-validation schemes that simulate predictions for closely or distantly related individuals, while also considering different population sizes. As a reference point, we compare the genomic prediction performance based on imputed data with that of the standard SNP array data.

2 Material and methods

2.1 Dataset

In the presented study, we utilized phenotypic and genotypic data described in Snowdon et al. (2015) and Werner et al. (2018a). In brief, the plant material utilized is a portion of the

*Bn*NAM population initially described in Snowdon et al. (2015). The entire population derives from a panel of 60 genetically diverse *B. napus* founder lines, which were all crossed to the same parental line (a winter oilseed rape inbred line) to generate 60 half-sibling families of double haploid (DH) or recombinant inbred lines (RILs). From these, a subset of the offspring including 17 selected DH families (420 DH lines) and 29 RIL families (520 RILs) were subsequently crossed with an elite male-sterile parent to generate F1 test hybrids that were subsequently evaluated in field trials. All test crosses underwent phenotypic assessment in multi-environment trials, as described by Werner et al. (2020) on an entry mean basis. Traits include seed yield, begin of flowering, plant length, oil content, protein content and glucosinolate content.

All the DH lines and RILs were genotyped using the Brassica 60k SNP array (Clarke et al., 2016; Mason et al., 2017). To ensure data quality, we filtered out markers with non-unique positions on the *B. napus* reference genome "Darmor-bzh" v4.2 (Chalhoub et al., 2014). Specifically, markers were excluded if their 50 bp SNP probe sequence could not be precisely matched to a unique position on the reference sequence without any mismatches (E-value ≤ 7.59E−17) based on an BLASTn analysis (Madden, 2003).

Additionally, all NAM family founders were sequenced using the Illumina HiSeq 2500 platform (Illumina Inc., San Diego, CA, USA) with approximately 12–15x genome coverage. Reads were aligned to the "Darmor-bzh" v4.2 reference genome (Chalhoub et al., 2014) and variants were identified as described by Schmutzer et al. (2015), considering only biallelic SNPs with a unique position on the reference genome.

There were 10,788 SNPs common to both the filtered SNP array data and the whole-genome sequencing SNP data. Subsequently, the array genotyping data from DH lines and RILs were combined with the SNPs obtained from whole-genome sequencing of the founders. Further refining the dataset, we excluded SNPs with a missing rate greater than 10% and a minor allele frequency less than 0.05. This yielded 403,080 high-quality SNPs (19,846 from the SNP array and 383,234 from whole-genome sequencing) for our subsequent analyses.

2.2 Imputation

Utilizing the founders with SNPs from sequencing as reference panel, we imputed all available SNPs onto all individual lines from the NAM population using the software “BEAGLE” v5.4 (Browning and Browning, 2007; Browning et al., 2018). We choose “BEAGLE” as imputation

tool as it has been used in other crops to improve prediction accuracy of genomic prediction when imputing whole-genome sequencing data (e.g. Berry and Kearney, 2011; Song et al., 2019; Munyengwa et al., 2021) and has been widely applied in rapeseed studies (Werner et al., 2018a, 2020; Knoch et al., 2021; Weber et al., 2023b, 2023a).

2.3 Haplotype block analysis

We created haplotype blocks by dividing the genome into equal-sized blocks of 10 Kbp along the chromosomes. Then, we evaluated the number of SNPs from the SNP array within each block.

To better understand how effectively each haplotype block was represented by SNP array SNPs, we counted the number of haplotypes present in each block in the population (based on all available markers in the block). We then calculated the minimum number of biallelic SNPs needed to represent each haplotype within a block. A sequence of n biallelic SNPs can represent 2^n haplotypes. For instance, if a 10 Kbp haplotype block had ten haplotypes, regardless of the actual SNP count within that block, at least four SNPs (2^4 = 16) would be needed to represent each haplotype block.

On a genome-wide basis, we plotted the number of SNPs required to represent each haplotype block, along with the SNP count from the SNP array within that block. This was achieved using a loess curve, with a smoothing parameter (alpha) set to 0.1 with the r function "loess" (R Core Team, 2021). The purpose of this analysis was to evaluate how effectively SNPs on the SNP array could potentially represent the haplotypes present in the population.

2.4 Linkage disequilibrium

An important factor contributing to genomic prediction is the LD between QTL and markers. As a proxy for LD between QTL and SNPs, we calculated the LD between SNPs using only SNPs from the array or imputed whole-genome sequencing SNP data. This was done using the software TASSEL (Bradbury et al., 2007).

2.5 Genomic Prediction

For genomic prediction we utilized the GBLUP model (Bernardo, 1994; Meuwissen et al., 2001; VanRaden, 2008), which employs the following underlying mixed linear model:

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

here, $\boldsymbol{y}$ is the vector of observations for a trait under consideration (i.e. adjusted entry means) and $\boldsymbol{\beta}$ is the vector of fixed effects. This comprises a fixed effect to model the intercept and an effect corresponding to the RILs and DH lines, as they were cultivated in different years (Werner et al., 2020), along with the design matrix of the fixed effects $\boldsymbol{X}$. $\boldsymbol{a}$ is a vector of random additive SNP effects with their associated design matrix $\boldsymbol{Z}_a$.

It is assumed that $\boldsymbol{a} \sim N(0, \boldsymbol{G}_a \sigma_a^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_a^2 and σ_e^2 are additive genetic variance and error variance, respectively. $\boldsymbol{G}_a$ is the additive relationship matrices and $\boldsymbol{I}$ is a n x n identity matrix. The additive genomic relationship matrix was derived following the method proposed by VanRaden (2008):

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

The entries of $\boldsymbol{Z}$ are represented by (0-2pi) for homozygous allele A, (1-2pi) for the heterozygous state and (2-2pi) for homozygous allele B, with pi being the allele frequency of the B allele. $\boldsymbol{Z}$ was either based on the high-quality SNPs from the SNP array or the imputed whole-genome sequencing data (including SNP array SNPs).

As incomplete LD between markers and QTL can lead to apparent or phantom epistasis (Wood et al., 2014; de los Campos et al., 2019; Schrauf et al., 2020), we investigated if a model considering epistasis yields higher prediction accuracy based on SNPs from the SNP-array and if this difference can be leveraged using imputed whole-genome sequencing data. For this purpose, we utilized an extended GBLUP model to account for additive by additive epistasis (EGBLUP), following Jiang and Reif (2015). The model describing the EGBLUP is:

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

here, $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{Z}_a$ and $\boldsymbol{e}$ are the same as in the GBLUP model. Additionally, $\boldsymbol{i}$ is a vector of random epistatic effects with the belonging design matrix $\boldsymbol{Z}_i$.

As in the GBLUP 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 $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$. Here σ_a^2, σ_{aa}^2 and σ_e^2 are additive genetic variance, epistatic genetic variance and error variance, respectively. $\boldsymbol{G}_{aa}$ is the epistatic relationship matrix which is calculated according to Henderson (1985) and Jiang and Reif (2015) as:

$$\boldsymbol{G}_{aa} = \boldsymbol{G}_a \# \boldsymbol{G}_a$$

where # denotes the Hadamard product operation.

The mixed linear mixed models described here (GBLUP and EGBLUP) were solved using the package "sommer" (Covarrubias-Pazaran, 2016, 2018) in R (R Core Team, 2021).

2.6 Evaluation of prediction accuracy

Two methods, namely random cross-validation and family-wise cross-validation, were used to assess how well genomic prediction could predict the phenotype mimicking scenarios for prediction of close and distantly related individuals.

In family-wise cross-validation, the dataset was divided based on the family assignments within the NAM population. The model was trained on the genotypic and phenotypic data of 45 families (training set). The trained model was then used to predict the remaining family (validation set). Each family served once as validation set. To evaluate prediction accuracy, the Pearson correlation coefficient (r) was calculated between the observed and predicted phenotypic values of the validation set.

In random cross-validation, 920 genotypes were randomly selected to create the training set, while the remaining 20 genotypes formed the validation set. The models were trained using the phenotypic and genotypic data of the training set, and then applied to predict the phenotype of the validation set. This process was repeated 200 times with random assignment of genotypes to training and validation sets. The validation set size was chosen to mirror the average family size of the NAM families. This ensured a fair comparison between the two cross-validation methods. Prediction accuracy was measured using the Pearson correlation coefficient, as described above.

Additionally, to explore the impact of varying training population sizes in the random cross-validation scenario, we randomly chose 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, and 90% of the genotypes from the previously mentioned training sets to train the model. The validation sets and the assessment of prediction accuracy remained as described above.

2.7 Genomic relationship

To assess the resemblance between relationship coefficients based on SNPs from the array and from imputed whole-genome sequencing data, we used the relationship coefficients (inbreeding values) obtained from the relationship matrix calculated following VanRaden

(2008) and calculated the Pearson correlation between relationship coefficients from SNPs of the SNP array and from imputed whole-genome sequencing data.

2.8 Simulation

To evaluate further scenarios where imputed whole-genome sequencing data outperforms smaller SNP datasets in genomic prediction, a simulation study was carried out. The simulation was based on the imputed whole-genome sequencing genotypic data described in the "Dataset" section, assuming no imputation errors. Two scenarios were considered. In the first scenario, SNPs were randomly sampled from all SNPs to act as QTL. In the second scenario, we simulated an extreme case where QTL were exclusively present in haplotype blocks from the section "Haplotype block analysis", which were not tagged by SNPs from the SNP array. For both scenarios, we sampled 100, 1,000, or 10,000 SNPs as simulated QTL. QTL effects were sampled from a normal distribution with a mean of 0 and a variance of 1. The phenotype was then determined by adding a residual term to the total additive value of each individual. Residuals were also sampled from a normal distribution with a mean of 0 and a variance V_e. V_e was derived as $\frac{V_g}{H^2} - V_g$, where V_g represents the total genetic variance (variance of the simulated genotypic values) and H^2 is the heritability calculated as $H^2 = \frac{V_g}{V_g + V_e}$. Respective heritabilities of 0.4, 0.6 and 0.8 were simulated for each set of simulated QTL. Each combination of scenario, number of QTL and heritability was repeated 100 times. In each simulation, genomic prediction was conducted with both prediction models, using SNPs from the SNP array or imputed whole-genome sequencing SNPs, respectively. Prediction accuracy was assessed using five-fold cross-validation and accuracy was measured as the Pearson correlation between true and predicted phenotypic values of the validation sets.

3 Results

3.1 Linkage disequilibrium and marker density

When considering SNPs solely from the SNP array, the average marker density was 17.56 SNP Mbp^{-1}, and this increased substantially to 356.71 SNP Mbp^{-1} with the addition of SNPs from whole-genome sequencing. The mean distance between neighboring SNPs was 61.89 kbp for SNPs from the array and 3.01 Kbp for SNPs from imputed whole-genome sequencing data.

Similar trends were observed in LD. Utilizing only array SNPs, the average LD between neighboring SNPs was 0.47, whereas it was 0.53 for imputed whole-genome sequencing SNPs.

The increase in LD was particularly notable on the A-subgenome, where the average LD rose from 0.34 to 0.50, compared to a change from 0.56 to 0.55 on the C-subgenome (Figure 1).

3.2 Haplotype block analysis

Segmenting the genome into haplotype blocks yielded blocks containing an average of 59.4 haplotypes across the whole genome. Furthermore, the analysis unveiled that, out of 25,074 haplotype blocks, only 10,936 were tagged with markers from the SNP array (Figure 2). Closer examination of the haplotypes within each block revealed that in nearly all cases where haplotype blocks were tagged by SNP array SNPs, there was an insufficient number of array SNPs to adequately represent each individual haplotype (Figure 2).

3.3 Genomic relationship

The investigation of relationship coefficients (inbreeding values) revealed a substantial resemblance between the respective coefficients derived from the SNP array data and the imputed whole-genome sequencing data. The correlation coefficient between these relationship coefficients was found to be 0.96, indicating a strong concordance in the assessment of genetic relatedness.

3.4 Genomic Prediction

3.4.1 Random cross-validation

The average prediction accuracies ranged from 0.75 to 0.99 when utilizing array SNPs, and with imputed whole-genome sequencing data it also ranged from 0.75 to 0.99 (Figure 3). The highest prediction accuracies were observed for days to flowering, which had a nearly perfect prediction accuracy. On the other hand, the lowest values were observed for seed yield, irrespective of the marker data used (Figure 3). Generally, only minor differences were observed between predictions using the two marker sets and/or different prediction models (Figure 3).

As expected, a reduction in the training set size resulted in a decline in prediction accuracy for both marker sets (Figure 4). Intriguingly, this decrease followed a similar pattern in both sets of markers. The trait days to flowering exhibited no mentionable change in prediction accuracy with decreasing training set size, independent of the reduction rate (Figure 4). In predictions for glucosinolate content, prediction accuracy showed only marginal alterations when the training set size was reduced by up to 50%. Beyond this point, a nearly linear decrease in prediction accuracy occurred with further reductions in training set size (Figure 4).

For the remaining traits—seed yield, plant length, oil content, and protein content—prediction accuracy, while varying in absolute levels, exhibited a comparable decreasing pattern. Notably, a more pronounced decrease was observed between reductions of 80% to 90%, in contrast to the reductions ranging from 10% to 80% (Figure 4).

3.4.2 Family-wise cross-validation

Prediction accuracies from family-wise cross-validation were generally inferior to those from random cross-validations (Figure 3 and Figure S1). The average prediction accuracy spanned from 0.16 to 0.48 with array SNPs, while with imputed whole-genome sequencing data the accuracy showed a similar range from 0.14 to 0.47 (Figure S1). Only the traits with the highest prediction accuracy was slightly different for the two marker sets—oil content for SNP array SNPs and days to flowering for imputed whole-genome sequencing SNPs. However, the differences in prediction accuracy between days to flowering and oil content were generally negligible (Figure S1). Much like in random cross-validation, the differences between the two marker sets were insignificant, and the model differences were negligible (Figure S1).

3.5 Simulation

In both simulation scenarios—whether involving a random distribution of QTL or exclusive placement of QTL in regions not tagged by array SNPs—the prediction accuracy closely followed the simulated heritability (Figure S2 and S3). This pattern persisted regardless of the simulated number of QTL and the types of markers employed. Generally, differences between the two marker datasets were minimal (Figure S2 and S3). However, small increases were generally observed with the use of imputed whole-genome sequencing data. As heritability increased, and regardless of the simulated number of QTL, the advantage of imputed whole-genome sequencing became more pronounced. Nevertheless, the absolute differences remained quite low (Figure S2 and S3). In the extreme scenario where QTL were solely present in regions without SNPs from the SNP array, the advantage of imputed whole-genome sequencing data was slightly more prominent compared to the scenario with random QTL distribution. Nevertheless, these differences remained negligible (Figure S3).

4 Discussion

Utilizing published data from a population used for hybrid breeding, genomic resequencing and genetic analysis in winter oilseed rape, we show that imputed SNP data form whole-genome sequencing does not necessarily improve genomic prediction. While the marker

number is drastically changed, genomic prediction accuracy was only marginally affected regardless of training population size.

Markers were imputed using the founders of a NAM population, for which SNP markers from whole-genome sequencing of the founder lines were available to impute the entire NAM population. For imputation, the widely applied software “BEAGLE” (Browning and Browning, 2007; Browning et al., 2018) was used as it finds wide application in rapeseed (Werner et al., 2018a, 2020; Knoch et al., 2021; Weber et al., 2023b, 2023a). In previous studies, imputation using “BEAGLE” has been demonstrated to produced high prediction accuracy in animal and plant breeding (e.g. Berry and Kearney, 2011; Song et al., 2019; Munyengwa et al., 2021).

Generally, genomic prediction accuracy based on all marker sets roughly corresponded to the results in the original publication of Werner et al. (2020), independent of the utilized marker data. Furthermore, except for some differences in flowering time and glucosinolate content, the ranking in prediction accuracies across the examined traits closely resembled those reported in previous studies (Würschum et al., 2014; Jan et al., 2016; Werner et al., 2018a, 2018b; Knoch et al., 2021). As expected, reducing the size of the training set resulted in a clear decrease in prediction accuracy (Heffner et al., 2011; Habier et al., 2013; Norman et al., 2018; Fernández-González et al., 2023; Wu et al., 2023). This decrease could potentially be compensated by training set optimization (Akdemir and Isidro-Sánchez, 2019; Isidro y Sánchez and Akdemir, 2021; Fernández-González et al., 2023).

Genomic prediction accuracy varied only marginally between the two prediction models. In addition to the GBLUP, we evaluated the performance of EGBLUP to address additive-by-additive epistasis (Henderson, 1985; Jiang and Reif, 2015). This was done to i) incorporate epistasis and ii) account for apparent or phantom epistasis. The latter phenomenon arises when markers and QTL exhibit incomplete LD with neighboring QTL, potentially leading to significant marker interactions as well as improved prediction accuracies in models that consider epistasis (Wood et al., 2014; de los Campos et al., 2019; Schrauf et al., 2020). As the marker density was significantly lower with the SNP array compared to whole-genome sequencing, we expected considerable phantom or apparent epistasis. While we cannot completely rule out epistasis, as both marker datasets are only a sample of all potential genomic variations, in the population examined here the EGBLUP did not exhibit improved prediction accuracies and is not beneficial compared to the simple GBLUP.

Unfortunately, opposing to the reports in literature (Berry and Kearney, 2011; Cleveland and Hickey, 2013; Gorjanc et al., 2017a, 2017b; Tsai et al., 2017; Song et al., 2019; Mancin et al., 2021; Munyengwa et al., 2021; Kriaridou et al., 2023), which report beneficial effects of imputation to a higher marker density on genetic gain, prediction accuracy or return on investment, we could not find mentionable improvements in genomic prediction for closely and distantly related individuals in the *B. napus* population examined here. In contrast to research in animal breeding, which predominantly centers on heterozygous individuals (e.g. Zhang and Druet, 2010; Berry and Kearney, 2011; Cleveland and Hickey, 2013; Tsai et al., 2017; Song et al., 2019; Mancin et al., 2021; Kriaridou et al., 2023), the present study, as well as most of plant breeding studies, focus on inbred lines—essentially homozygous genotypes. This could potentially influence imputation and genomic prediction.

An important factor contributing to genomic prediction accuracy is the number of markers and the LD between them and QTL (Habier et al., 2013). In this case, the marker density was substantially increased by imputation. Furthermore, the level of LD among neighboring markers, as approximation to LD between markers and QTL, could also be increased, especially on the A-subgenome. This should theoretically favor prediction accuracy. However, it did not translate into higher genomic prediction accuracies.

In *B. napus,* the A-subgenome experiences a more rapid decay in LD (Qian et al., 2014; Zhou et al., 2017; Werner et al., 2018a; Jan et al., 2019; Wu et al., 2019) and exhibits higher genetic diversity (Lu et al., 2019) compared to the C-subgenome. This could provide an explanation for the absence of enhanced prediction accuracy, suggesting that the SNP array may either already adequately capture the diversity or fail to sufficiently represent it. The latter can result in erroneous imputations and, consequently, a lack of improvement in prediction accuracy.

The C-subgenome on the other hand is considerably larger than the A-subgenome (Chalhoub et al., 2014), however it exhibits a considerably slower decay in LD compared to the A-subgenome (Qian et al. 2014). Consequently, the *B. napus* C-subgenome also contains a number of very large LD blocks (Qian et al., 2014; Zhou et al., 2017; Werner et al., 2018a; Jan et al., 2019; Wu et al., 2019) which are likely already effectively tagged with SNPs from the array.

The LD pattern could hence be a reason for the lack of improvement with imputed whole-genome sequencing markers, as the SNPs present on the SNP array may already tag QTL

sufficiently. If this is the case, no increase in prediction accuracy is to be expected with imputed SNPs from whole-genome sequencing. Interestingly, the density of markers can often be lowered without a substantial reduction in prediction accuracy (de Roos et al., 2009; Kriaridou et al., 2023). Furthermore, the advantageous impact of imputation on prediction accuracy is frequently observed when employing low-density marker sets, often in the range of a few hundred markers, to impute data onto high-density marker sets (Berry and Kearney, 2011; Hickey et al., 2015). On the other hand, as one transitions from relatively high marker density to even higher densities, the magnitude of improvement diminishes (Tsai et al., 2017; Song et al., 2019).

Relationship acts as another important contributor to genomic prediction accuracy (Habier et al., 2013). Notably, the estimates of relationships exhibit a high of correlation between the two marker types. This correlation likely explains the absence of discernible differences in prediction accuracies between the two marker types, serving as an additional indication that both sets of markers capture, to some extent, the same information. This alignment is reasonable, given that imputation tools draw upon information related to linkage disequilibrium and relationship for imputation (Browning and Browning, 2007; Howie et al., 2011; Delaneau et al., 2012; Browning et al., 2018).

The imputation software "BEAGLE" relies partially on the marker positions on the reference genome. Here positions were predicted using the *B. napus* reference genome Darmor-bzh v4.2 (Chalhoub et al., 2014), which was the most up-to-date reference when the SNP positions were originally determined. Although a re-positioning on more recent *B. napus* reference genomes may change some of the SNP positions, we do not expect this to have a significant impact on the results, as the whole-genome sequence data from Schmutzer et al. (2015) that was used for the imputation was also mapped to Darmor bzh v4.2. More importantly, in complex plant genomes such as that of allopolyploid *B. napus*, array SNPs may lack specificity in terms of their physical position, potentially representing different homoeologous loci across subgenomes (Mason et al., 2017; Makhoul et al., 2020). Despite our rigorous filtering to ensure confidence in SNP positions on the reference genome, these positions may differ significantly in the studied population or even among specific individuals as a result of genome structural variants (Chawla et al. 2021). These discrepancies have the potential to limit imputation accuracy and restrict the value of using imputed SNPs for genomic predictions.

Furthermore, genotyping and imputation errors have the potential to introduce erroneous haplotypes that do not represent the actual population, consequently negatively affecting prediction accuracy (Weber et al., 2023b). In the examination of haplotype blocks, the mean number of haplotypes per block was 59.4. In the BnNAM population, and under the assumption of minimal recombination within the 10 Kbp block window, we would not expect a higher number of haplotypes than those present in the homozygous founder individuals (i.e. 46). An excess of haplotypes beyond this threshold may be an indicator for genotyping or imputation errors, leading to the generation of spurious rare haplotypes and consequently inaccurate effect estimations. Moreover, accurately estimating the effects of rare variants necessitates large populations for robust effect estimation (Meuwissen et al., 2001; Goddard et al., 2011). Hence, reducing genotyping and imputation errors might potentially minimize the occurrence of falsely imputed genotypes and haplotypes, thereby contributing to an enhancement in genomic prediction accuracy. Previous studies have indicated that incorporating close relatives and expanding the reference panel can improve imputation accuracy (Calus et al., 2014; Roshyara and Scholz, 2015; Ventura et al., 2016). Therefore, a viable strategy for enhancing imputation involves sequencing not only founders but also representative lines from each NAM family.

Another limitation in the current study is the restriction of markers to biallelic SNPs, which are very common in crop genomes (Rafalski, 2002; Ganal et al., 2009). Although SNPs are prevalent, they may not comprehensively elucidate the entire spectrum of genetic variation underlying complex traits, thereby giving rise to the issue of "missing heritability" (Manolio et al., 2009; Forer et al., 2010). Advancements in sequencing technology offer an opportunity to address this limitation by incorporating alternative markers, such as genome structural variants (Manolio et al., 2009; Génin, 2020; Theunissen et al., 2020; Chawla et al., 2021; Zhou et al., 2022). Unfortunately, the assessment of these variations necessitates high sequencing coverage, if possible with long reads that span gene-level size structural variants, and their identification using SNP presence-absence data detected by raw fluorescence levels from SNP arrays is inherently limited (Gabur et al., 2018; Weber et al., 2023a). Consequently, combining structural variations with SNP arrays through imputation could be an appealing approach to improve genomic prediction accuracy.

A promising technology after the identification of all kinds of genomic markers is genotyping by sequencing (GBS), offering a cost-effective alternative to SNP arrays for profiling breeding

populations (Poland and Rife, 2012; Kim et al., 2016; Chung et al., 2017). This technology enables identification of a broader spectrum of alleles and sequence variants, including deletions (Poland and Rife, 2012).

Using imputed whole-genome sequencing marker data for marker pre-selection in genomic prediction has demonstrated enhanced prediction accuracy (Song et al., 2019). The integration of this approach with GBS data holds the potential to further improve genomic selection. By specifically genotyping only relevant markers and optimizing the sequencing strategy of a breeding program, this combined approach offers a promising avenue for improving the precision of genomic selection, however further research is needed to test this hypothesis.

SNP arrays represent the method of choice for sequencing large populations in many plant and crop studies, with arrays available for a diverse range of crops (Ganal et al., 2011; Song et al., 2013; Unterseer et al., 2014; Tian et al., 2015; Bayer et al., 2017; Mason et al., 2017; You et al., 2018; Sun et al., 2020, 660). Most arrays undergo rigorous SNP selection, validation and continuous updating across different populations and over time (Boichard et al., 2012; Bayer et al., 2017). As we showed in this example from *B. napus* breeding population, arrays SNPs are a highly reliable source of information for canola/oilseed rape breeding and should be the primary choice when genotype information is required.

5 Conclusion

In contrast to earlier studies in other species, our study utilizing data from a large *B. napus* NAM population shows that imputation of SNP data from whole-genome sequencing does not improve prediction accuracy in genomic prediction in comparison to SNP data from a medium-density SNP array. This is likely caused by the fact that all necessary information is already present in the SNP array data, through LD or large resemblance between relationship coefficients from the array and sequencing datasets. Furthermore, the prediction accuracy can be limited by potential imputation errors. We conclude that imputation is not necessary or beneficial to obtain acceptable genomic prediction accuracies in canola/rapeseed breeding, and that SNP arrays represent a reliable high-throughput genotyping tool with high suitability for *B. napus* breeding.

6 Conflict of Interest

Amine Abbadi and Tobias Kox are employed at NPZ innovation GmbH. The remaining 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.

7 Author Contributions

SEW and RJS designed the study. SEW conceived the analysis, LE helped in the modeling for genomic prediction. SEW wrote the manuscript. RJS, AA, TK, LE and AS revised the manuscript. All authors read and approved the final manuscript.

8 Funding

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

9 Data availability

The whole-genome sequencing marker data is available in Schmutzer et al. (2015), while the remaining data is available upon request from Werner et al. (2020).

10 References

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci Rep* 9, 1446. doi: 10.1038/s41598-018-38081-6.

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. *Frontiers in Plant Science* 8. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2017.01792 [Accessed January 4, 2023].

Bernardo, R. (1994). Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Science* 34, cropsci1994.0011183X003400010003x. doi: 10.2135/cropsci1994.0011183X003400010003x.

Berry, D. P., and Kearney, J. F. (2011). Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal* 5, 1162–1169. doi: 10.1017/S1751731111000309.

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.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi: 10.1093/bioinformatics/btm308.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015.

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. *The American Journal of Human Genetics* 81, 1084–1097. doi: 10.1086/521987.

Calus, M. P. L., Bouwman, A. C., Hickey, J. M., Veerkamp, R. F., and Mulder, H. A. (2014). Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *animal* 8, 1743–1753. doi: 10.1017/S1751731114001803.

Chalhoub, B., Denoeud, F., Liu, S., Parkin, I. A. P., Tang, H., Wang, X., et al. (2014). Early allopolyploid evolution in the post-Neolithic Brassica napus oilseed genome. *Science* 345, 950–953. doi: 10.1126/science.1253435.

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 Biotechnology Journal* 19, 240–250. doi: 10.1111/pbi.13456.

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.

Cleveland, M. A., and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. *Journal of Animal Science* 91, 3583–3592. doi: 10.2527/jas.2013-6270.

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., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. *G3 Genes|Genomes|Genetics* 3, 1903–1926. doi: 10.1534/g3.113.008227.

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.

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.

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.

Edriss, V., Gao, Y., Zhang, X., Jumbo, M. B., Makumbi, D., Olsen, M. S., et al. (2017). Genomic Prediction in a Large African Maize Population. *Crop Science* 57, 2361–2371. doi: 10.2135/cropsci2016.08.0715.

Edwards, D., and Batley, J. (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnology Journal* 8, 2–9. doi: 10.1111/j.1467-7652.2009.00459.x.

Edwards, D., Batley, J., and Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126, 1–11. doi: 10.1007/s00122-012-1964-x.

Fernández-González, J., Akdemir, D., and Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theor Appl Genet* 136, 30. doi: 10.1007/s00122-023-04265-6.

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 Bioinformatics* 11, 318. doi: 10.1186/1471-2105-11-318.

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.

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 Biotechnology Journal* 16, 2102–2112. doi: 10.1111/pbi.12942.

Ganal, M. W., Altmann, T., and Röder, M. S. (2009). SNP identification in crop plants. *Current Opinion in Plant Biology* 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.

Goddard, M. e., Hayes, B. j., and Meuwissen, T. h. e. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics* 128, 409–421. doi: 10.1111/j.1439-0388.2011.00964.x.

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolin, R., Gaynor, R. C., and Hickey, J. M. (2017a). Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation. *Crop Science* 57, 216–228. doi: 10.2135/cropsci2016.06.0526.

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., and Hickey, J. M. (2017b). Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. *Crop Science* 57, 1404–1420. doi: 10.2135/cropsci2016.08.0675.

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.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92, 433–443. doi: 10.3168/jds.2008-1646.

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., and Sorrells, M. E. (2011). Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. *Crop Science* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253.

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.

Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. (2012). Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Science* 52, 654–663. doi: 10.2135/cropsci2011.07.0358.

Hickey, J. M., Gorjanc, G., Varshney, R. K., and Nettelblad, C. (2015). Imputation of Single Nucleotide Polymorphism Genotypes in Biparental, Backcross, and Topcross Populations with a Hidden Markov Model. *Crop Science* 55, 1934–1946. doi: 10.2135/cropsci2014.09.0648.

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.

Isidro y Sánchez, J., and Akdemir, D. (2021). Training Set Optimization for Sparse Phenotyping in Genomic Selection: A Conceptual Overview. *Frontiers in Plant Science* 12. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2021.715910 [Accessed November 15, 2023].

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.

Jan, H. U., Guan, M., Yao, M., Liu, W., Wei, D., Abbadi, A., et al. (2019). Genome-wide haplotype analysis improves trait predictions in Brassica napus hybrids. *Plant Science* 283, 157–164. doi: 10.1016/j.plantsci.2019.02.007.

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

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. *G3 (Bethesda)* 8, 1687–1699. doi: 10.1534/g3.117.300548.

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 Science* 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., Fraslin, C., Gorjanc, G., Looseley, M. E., Johnston, I. A., et al. (2023). Evaluation of low-density SNP panels and imputation for cost-effective genomic selection in four aquaculture species. *Frontiers in Genetics* 14. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1194266 [Accessed November 9, 2023].

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.

Lu, K., Wei, L., Li, X., Wang, Y., Wu, J., Liu, M., et al. (2019). Whole-genome resequencing reveals Brassica napus origin and genetic loci involved in its improvement. *Nat Commun* 10, 1154. doi: 10.1038/s41467-019-09134-9.

Madden, T. (2003). "The BLAST Sequence Analysis Tool," in *The NCBI Handbook [Internet]* (National Center for Biotechnology Information (US)). Available at: https://www.ncbi.nlm.nih.gov/books/NBK21097/ [Accessed November 10, 2023].

Makhoul, M., Rambla, C., Voss-Fels, K. P., Hickey, L. T., Snowdon, R. J., and Obermeier, C. (2020). Overcoming polyploidy pitfalls: a user guide for effective SNP conversion into KASP markers in wheat. *Theor Appl Genet* 133, 2413–2430. doi: 10.1007/s00122-020-03608-x.

Mancin, E., Sosa-Madrid, B. S., Blasco, A., and Ibáñez-Escriche, N. (2021). Genotype Imputation to Improve the Cost-Efficiency of Genomic Selection in Rabbits. *Animals* 11, 803. doi: 10.3390/ani11030803.

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.

690 Mason, A. S., Higgins, E. E., Snowdon, R. J., Batley, J., Stein, A., Werner, C., et al. (2017). A
691 user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. *Theor Appl*
692 *Genet* 130, 621–633. doi: 10.1007/s00122-016-2849-1.

693 Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic
694 Value Using Genome-Wide Dense Marker Maps. *Genetics* 157, 1819–1829. doi:
695 10.1093/genetics/157.4.1819.

696 Munyengwa, N., Le Guen, V., Bille, H. N., Souza, L. M., Clément-Demange, A., Mournet, P., et
697 al. (2021). Optimizing imputation of marker data from genotyping-by-sequencing
698 (GBS) for genomic selection in non-model species: Rubber tree (Hevea brasiliensis) as
699 a case study. *Genomics* 113, 655–668. doi: 10.1016/j.ygeno.2021.01.012.

700 Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in
701 Wheat: Effect of Marker Density, Population Size and Population Structure on
702 Prediction Accuracy. *G3 Genes|Genomes|Genetics* 8, 2889–2899. doi:
703 10.1534/g3.118.200311.

704 Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical*
705 *Association* 103, 681–686. doi: 10.1198/016214508000000337.

706 Pérez, P., and de los Campos, G. (2014). Genome-Wide Regression and Prediction with the
707 BGLR Statistical Package. *Genetics* 198, 483–495. doi: 10.1534/genetics.114.164442.

708 Poland, J. A., and Rife, T. W. (2012). Genotyping-by-Sequencing for Plant Breeding and
709 Genetics. *The Plant Genome* 5. doi: 10.3835/plantgenome2012.05.0005.

710 Qian, L., Qian, W., and Snowdon, R. J. (2014). Sub-genomic selection patterns as a signature
711 of breeding in the allopolyploid Brassica napus genome. *BMC Genomics* 15, 1170. doi:
712 10.1186/1471-2164-15-1170.

713 R Core Team, R. C. T. (2021). R: A language and environment for statistical computing.
714 Vienna, Austria: R Foundation for Statistical Computing Available at: https://www.R-
715 project.org/.

716 Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. *Current*
717 *Opinion in Plant Biology* 5, 94–100. doi: 10.1016/S1369-5266(02)00240-6.

718 Roshyara, N. R., and Scholz, M. (2015). Impact of genetic similarity on imputation accuracy.
719 *BMC Genet* 16, 90. doi: 10.1186/s12863-015-0248-2.

720 Rutkoski, J. E., Poland, J., Jannink, J.-L., and Sorrells, M. E. (2013). Imputation of Unordered
721 Markers and the Impact on Genomic Selection Accuracy. *G3*
722 *Genes|Genomes|Genetics* 3, 427–439. doi: 10.1534/g3.112.005363.

723 Schmutzer, T., Samans, B., Dyrszka, E., Ulpinnis, C., Weise, S., Stengel, D., et al. (2015).
724 Species-wide genome sequence and nucleotide polymorphisms from the model
725 allopolyploid plant Brassica napus. *Sci Data* 2, 150072. doi: 10.1038/sdata.2015.72.

726 Schrauf, M. F., Martini, J. W. R., Simianer, H., de los Campos, G., Cantet, R., Freudenthal, J.,
727 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.

Snowdon, R. J., Abbadi, A., Kox, T., Schmutzer, T., and Leckband, G. (2015). Heterotic Haplotype Capture: precision breeding for hybrid performance. *Trends in Plant Science* 20, 410–413. doi: 10.1016/j.tplants.2015.04.013.

Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. *Genetics Selection Evolution* 51, 58. doi: 10.1186/s12711-019-0500-8.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLOS ONE* 8, e54985. doi: 10.1371/journal.pone.0054985.

Sun, C., Dong, Z., Zhao, L., Ren, Y., Zhang, N., and Chen, F. (2020). The Wheat 660K SNP array demonstrates great potential for marker-assisted selection in polyploid wheat. *Plant Biotechnology Journal* 18, 1354–1360. doi: 10.1111/pbi.13361.

The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796–815. doi: 10.1038/35048692.

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. *Frontiers in Neuroscience* 14. Available at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00047 [Accessed February 16, 2023].

Tian, H.-L., Wang, F.-G., Zhao, J.-R., Yi, H.-M., Wang, L., Wang, R., et al. (2015). Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. *Mol Breeding* 35, 136. doi: 10.1007/s11032-015-0335-0.

Tsai, H.-Y., Matika, O., Edwards, S. M., Antolín–Sánchez, R., Hamilton, A., Guy, D. R., et al. (2017). Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. *G3 (Bethesda)* 7, 1377–1383. doi: 10.1534/g3.117.040717.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15, 823. doi: 10.1186/1471-2164-15-823.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91, 4414–4423. doi: 10.3168/jds.2007-0980.

Ventura, R. V., Miller, S. P., Dodds, K. G., Auvray, B., Lee, M., Bixley, M., et al. (2016). Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. *Genetics Selection Evolution* 48, 71. doi: 10.1186/s12711-016-0244-7.

Wang, Y., Lin, G., Li, C., and Stothard, P. (2016). Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle. *Springer Science Reviews* 4, 79–98. doi: 10.1007/s40362-017-0041-x.

Weber, S. E., Chawla, H. S., Ehrig, L., Hickey, L. T., Frisch, M., and Snowdon, R. J. (2023a). Accurate prediction of quantitative traits with failed SNP calls in canola and maize. *Frontiers in Plant Science* 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1221750 [Accessed November 10, 2023].

Weber, S. E., Frisch, M., Snowdon, R. J., and Voss-Fels, K. P. (2023b). Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Frontiers in Plant Science* 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1217589 [Accessed November 10, 2023].

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., et al. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science* 11. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.592977 [Accessed December 22, 2022].

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. *The Plant Genome* 11, 170084. doi: 10.3835/plantgenome2017.09.0084.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., et al. (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* 43, 5. doi: 10.1186/1297-9686-43-5.

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.

Wu, D., Liang, Z., Yan, T., Xu, Y., Xuan, L., Tang, J., et al. (2019). Whole-Genome Resequencing of a Worldwide Collection of Rapeseed Accessions Reveals the Genetic Basis of Ecotype Divergence. *Molecular Plant* 12, 30–43. doi: 10.1016/j.molp.2018.11.007.

Wu, P.-Y., Ou, J.-H., and Liao, C.-T. (2023). Sample size determination for training set optimization in genomic prediction. *Theor Appl Genet* 136, 57. doi: 10.1007/s00122-023-04254-9.

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

You, Q., Yang, X., Peng, Z., Xu, L., and Wang, J. (2018). Development and Applications of a High Throughput Genotyping Tool for Polyploid Crops: Single Nucleotide Polymorphism (SNP) Array. *Frontiers in Plant Science* 9. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2018.00104 [Accessed August 24, 2022].

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., et al. (2011). Gains in QTL Detection Using an Ultra-High Density SNP Map Based on Population Sequencing Relative to Traditional RFLP/SSR Markers. *PLOS ONE* 6, e17595. doi: 10.1371/journal.pone.0017595.

Zhang, Z., and Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science* 93, 5487–5494. doi: 10.3168/jds.2010-3501.

Zhou, Q., Zhou, C., Zheng, W., Mason, A. S., Fan, S., Wu, C., et al. (2017). Genome-Wide SNP Markers Based on SLAF-Seq Uncover Breeding Traces in Rapeseed (Brassica napus L.). *Frontiers in Plant Science* 8. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2017.00648 [Accessed November 15, 2023].

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.

11 Figures

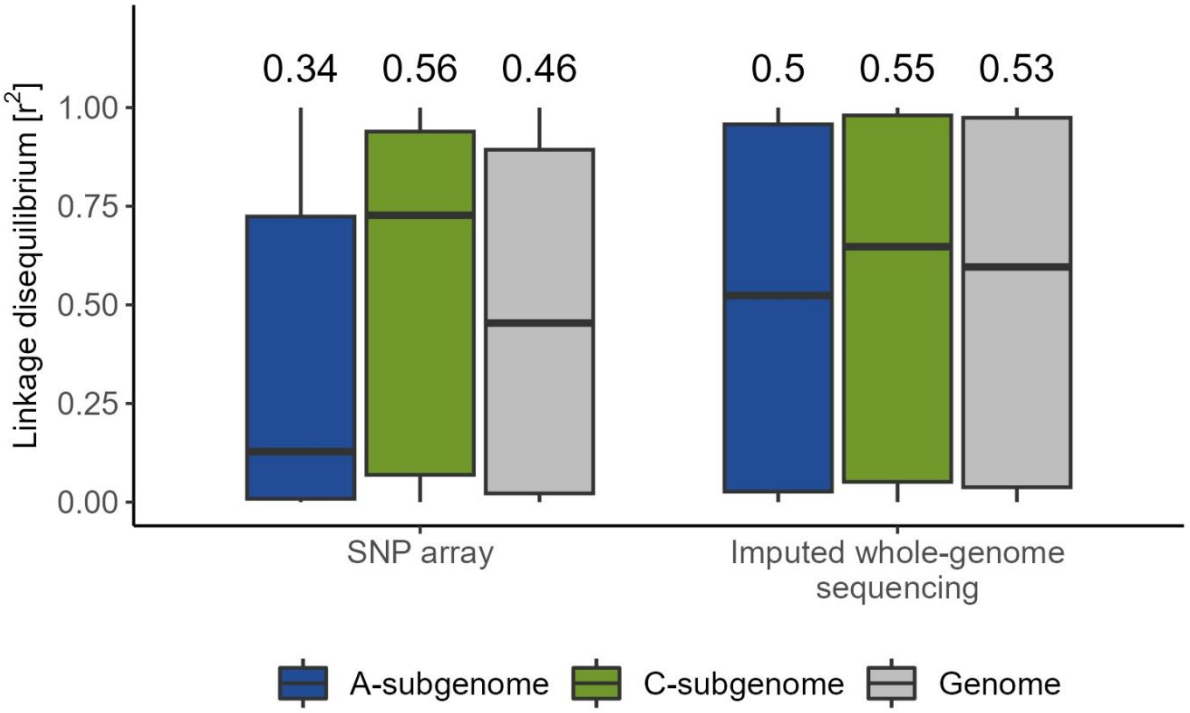

Figure 1 Average linkage disequilibrium of neighboring markers based on SNPs from the SNP array and imputed whole-genome sequencing SNP data on the genome (grey), A-subgenome (blue) and the C-subgenome (green).

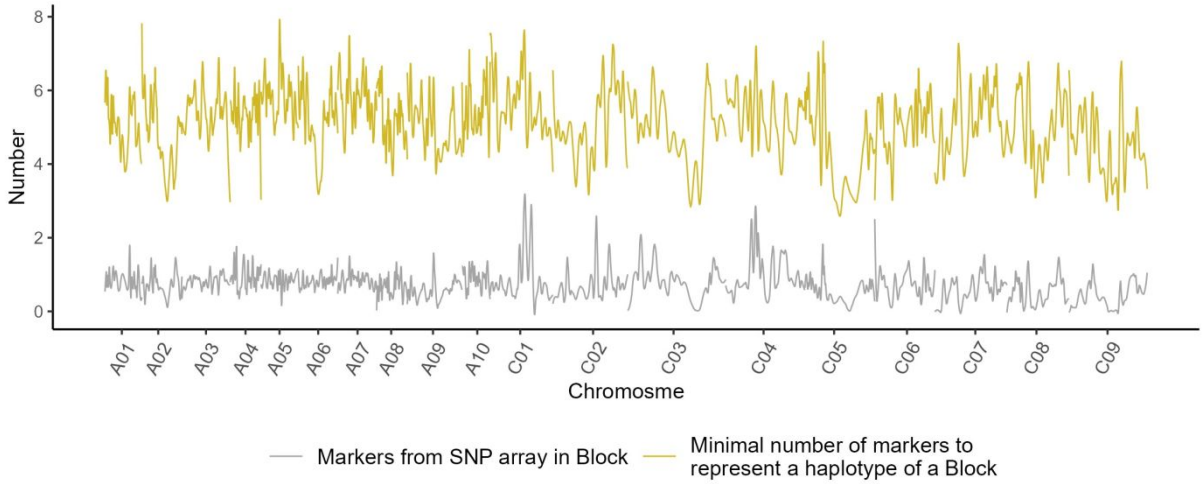

Figure 2 Minimal number of biallelic SNPs required to represent haplotypes of 10 Kbp haplotype blocks (yellow) and number of SNPs from the SNP array within the block (grey) displayed as loess curves across the whole genome.

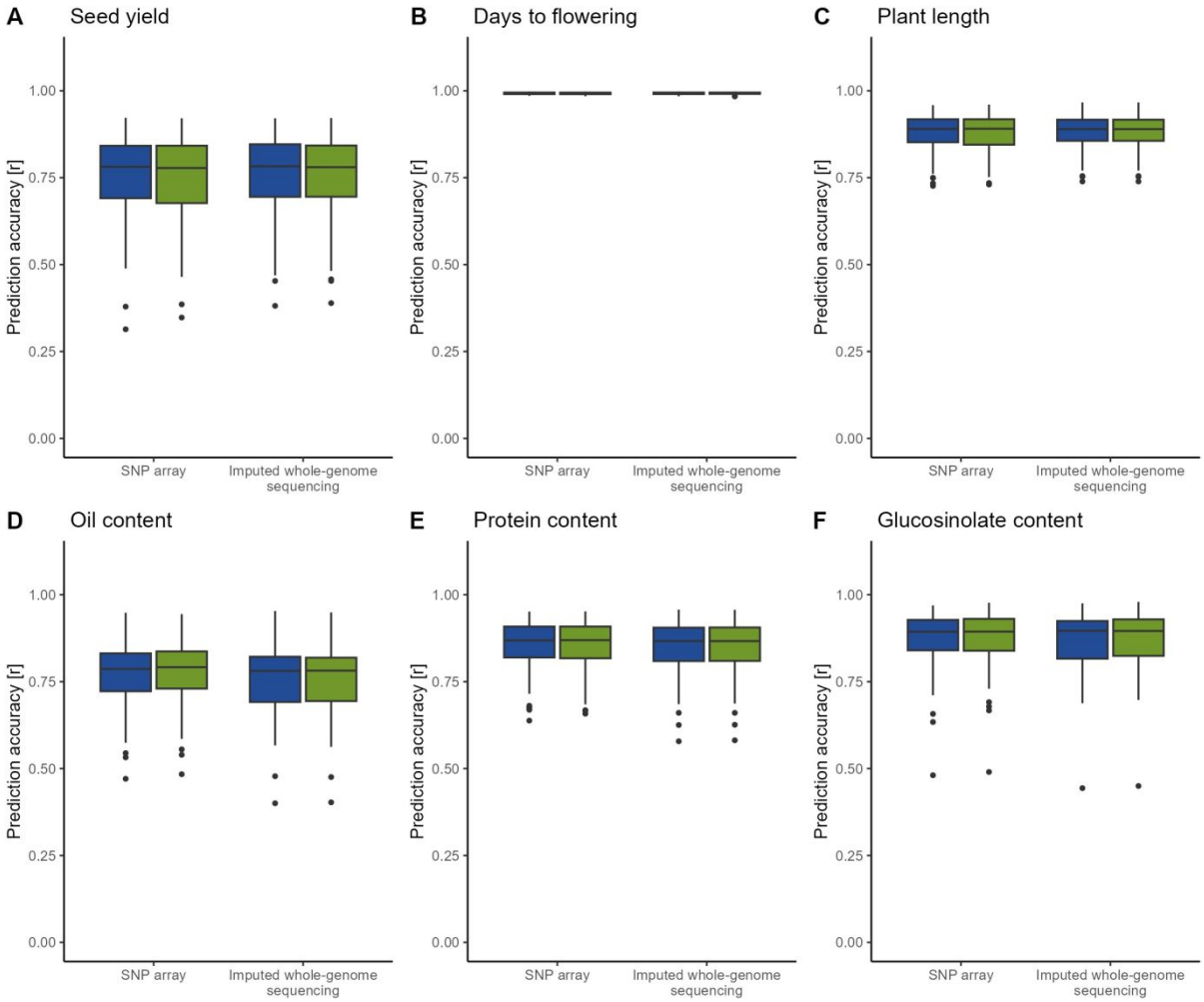

Figure 3 Prediction accuracy (r) with random cross-validation based on SNPs from the SNP array and imputed whole-genome sequencing SNP data using the GBLUP (blue) and EGBLUP (green). In *brassica napus* traits seed yield (A), days to flowering (B), plant length (C), oil content (D), protein content (E) and glucosinolate content.

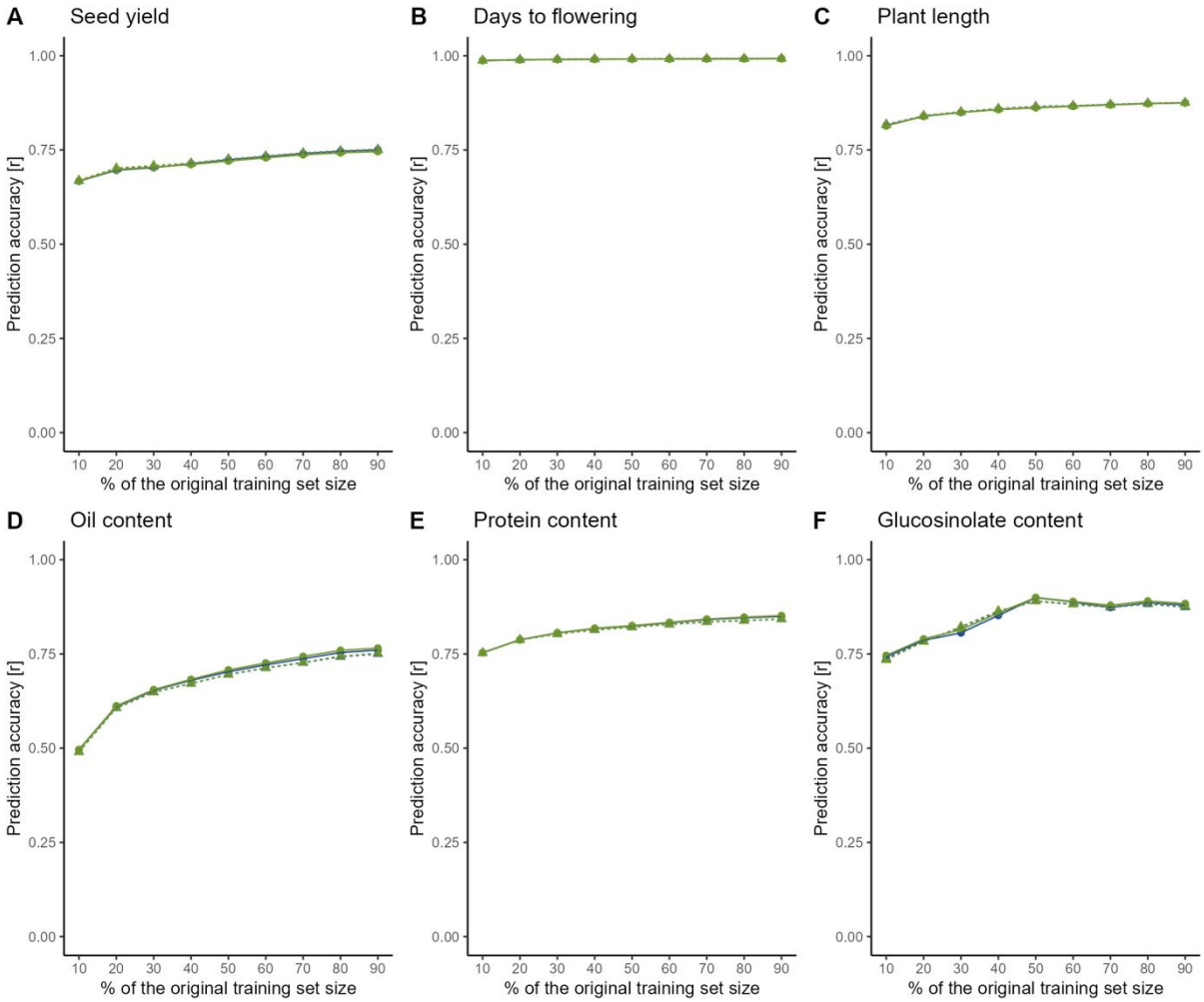

Figure 4 Prediction accuracy (r) with random cross-validation and reduction of the training set size based on SNPs from the SNP array (solid lines and triangles) and imputed whole-genome sequencing SNP data (dotted lines and circles) using the GBLUP (blue) and EGBLUP (green). In *brassica napus* traits seed yield (A), days to flowering (B), plant length (C), oil content (D), protein content (E) and glucosinolate content.

5 General Discussion

In numerous reports (Bernardo, 1994; Meuwissen et al., 2001; Hickey et al., 2014; Lehermeier et al., 2014; Gaynor et al., 2017; Werner et al., 2018), genomic prediction has been demonstrated to be a promising and effective tool within the selection process of plant breeding. This method relies on the acquisition of phenotypic information from a subset of the breeding population, complete genotypic information for the entire breeding population, and the application of a robust statistical framework for extrapolating phenotype implications to the broader breeding population.

Since the inception of genotyping technologies, various genotyping and marker systems have been employed in plant breeding. Early studies focused on utilizing RFLP markers, amplified fragment length polymorphism (AFLP) markers, or simple sequence repeat (SSR) markers (Melchinger, 1993; Bernardo, 1994; Bohn et al., 1999; Bernardo et al., 2000; Jordan et al., 2003) for genomic prediction. However, with the advent of sequencing technologies and, notably, the development of high-throughput SNP arrays (LaFramboise, 2009), SNPs have become the predominant marker system in various studies.The predominant statistical prediction models were established several decades ago (Henderson, 1975; Bernardo, 1994) and have since undergone extensive development, evaluation and successful application in genomic prediction (Cooper et al., 2014; Crossa et al., 2014; García-Ruiz et al., 2016). In particular, the foundation of these modeling methods can be traced back to Henderson (1975), who is considered one of the pioneering scientists to incorporate relatedness based on pedigree information, alongside phenotypic data, for breeding value prediction within a mixed linear model framework. Subsequently, genomic prediction has found widespread application in numerous breeding programs.

Maize stands as a prominent example for genomic prediction with remarkable efficacy (Schrag et al., 2007; Crossa et al., 2014; Lehermeier et al., 2014). This has helped breeders to discern superior accessions without necessitating overly extensive phenotyping efforts (Cooper et al., 2014). Meanwhile, genomic prediction has been employed in most major crops, including the present study's focus on rapeseed/canola, as well as wheat and soybean (Crossa et al., 2014; Lehermeier et al., 2014; Ma et al., 2016; Werner et al., 2018; Knoch et al., 2021). Our investigation validates the promising outcomes in all of these crops. Irrespective of the

statistical model employed, our results demonstrate reasonable prediction accuracies for many traits across the examined species, underscoring the utility of genomic prediction in aiding breeders to enhance their populations in important agronomic traits with reduced phenotyping costs and higher selection accuracy.

Genomic prediction enables the prediction of unphenotyped genotypes within a breeding population, facilitating sparse testing by lowering the requirement for phenotyping a large number of genotypes (Jarquin et al., 2020) and enabling offspring prediction across multiple breeding cycles to select genotypes or families for intensive testing (Auinger et al., 2016).

The statistical models employed for genomic prediction exhibit remarkable flexibility across traits with varying genetic architectures, ranging from qualitative to quantitative traits. Although these models were initially introduced to address quantitative traits, their utility has been demonstrated to extend effectively to traits with a more qualitative genetic architecture. (Bernardo and Yu, 2007).

5.1 Does the marker type matter?

Despite the high abundance of SNPs in eukaryotic genomes (Rafalski, 2002; Frazer et al., 2007; Ganal et al., 2011), they sometimes fail to comprehensively capture the entire spectrum of genetic variation underlying complex traits. This phenomenon has given rise to the problem of "missing heritability" (Manolio et al., 2009), where not all variations in genotypic values can be attributed to SNP markers.

To address this issue and to increase the information content of SNP arrays, the first approach applied in this thesis, given the availability of reliable SNP arrays, is to identify additional variants beyond SNPs. An attractive avenue examined in Chapter 2 is to leverage the block-like inheritance pattern of chromosomal regions, which are interrupted by recombination hotspots (Daly et al., 2001; Jeffreys et al., 2001; Reich et al., 2001; Gabriel et al., 2002). In this work it was hypothesized that the identification of haplotype blocks using SNP array markers could potentially address the limitations associated with missing heritability. This approach captures local epistasis (Jiang et al., 2018) and eliminates redundant information from markers in high LD. However, as detailed in Chapter 2, regardless of the dataset, method, or parameters used to identify haplotype blocks, the improvement in genomic prediction is often

marginal. In some cases, haplotype blocks may even significantly decrease prediction accuracy, especially in scenarios with large blocks based on fixed windows. In accordance with results shown by Difabachew et al. (2023), the study in Chapter 2 was unable to identify generally ideal methods or parameters to construct haplotype blocks. Interestingly, the patterns in prediction accuracy for genomic prediction accuracy in wheat were surprisingly similar between Chapter 2 and Difabachew et al. (2023) with frequentist models, showing increasing prediction accuracy as LD thresholds increased and decreasing accuracy as LD blocks increased in size. Whereas Difabachew et al. (2023) concluded that haplotype blocks are especially beneficial for disease resistance, the study presented in Chapter 2 fail to reach this conclusion for the examined wheat disease resistance trait. In general, it can be concluded that haplotype blocks can improve genomic prediction, but the extent is only marginal and likely insignificant. This is in accordance with the findings of Jiang et al. (2018), but contrasts with frequent reports in the literature showing improved prediction or even "optimal" haplotype block construction methods or parameters (Cuyabano et al., 2014; Ballesta et al., 2019). This lack of improvement can potentially be attributed to the introduction of rare alleles due to genotyping, phasing and imputation errors, thereby limiting the validity of haplotypes within a block.

Recent advancements in sequencing technology provide an opportunity to address the challenge of missing heritability by incorporating various markers, including genome structural variants such as deletions, insertions and copy number variations (Manolio et al., 2009; Génin, 2020; Theunissen et al., 2020; Chawla et al., 2021; Zhou et al., 2022). These kinds of markers have been associated with important agricultural traits in many crops, for example disease resistance and flowering time in canola/rapeseed, disease resistance and boron toxicity tolerance in barley, pathogen response and aluminum tolerance in maize, or plant height and heading date in wheat (Sutton et al., 2007; Beló et al., 2010; Li et al., 2012; Maron et al., 2013; Muñoz-Amatriaín et al., 2013; Nishida et al., 2013; Gabur et al., 2018, 2020; Vollrath et al., 2021b, 2021a).

In genomic prediction, the impact of incorporating structural variants on prediction accuracy varies across species. While some studies report improved accuracy in cattle and rapeseed (Hay et al., 2018; Chen et al., 2021; Knoch et al., 2021), others (Lyra et al., 2019), including our study in Chapter 3, failed to observe significant improvements. Chapter 3 deviates from the

method of calling structural variations from light intensity scores, as done in previous studies (Lyra et al., 2019; Knoch et al., 2021), and instead employs a strategy similar to "Single Nucleotide absence Polymorphisms" (Gabur et al., 2018). The findings indicate that failed allele calls, possibly stemming from structural variations (as demonstrated in rapeseed by Gabur et al., 2018), can result in prediction accuracies comparable to those achieved with SNP markers. This underscores the importance of failed allele calls and structural variants in the context of agricultural traits. However, combining these markers with SNPs does not enhance prediction accuracy compared to SNPs alone. One could hypothesize that structural variants, like SNPs, act as genetic markers, and SNPs alone are likely sufficient to tag most important QTL and the genetic background through relationships for quantitative traits. Hence, structural variations may to some extent only add redundant information, as indicated by the highly similar relationship estimates revealed in Chapter 3.

Generally, one limitation of calling other variants from SNP arrays, as discussed for haplotype blocks and failed allele calls, is the introduction of only "new alleles" compared to the two SNP alleles. While this captures allelic diversity more comprehensively, it does not increase marker density *per se*, as no regions untagged by the SNP array are introduced. Therefore, it may be advantageous to identify variants beyond those on the SNP array. Here, whole-genome sequencing has demonstrated the ability to generate large and dense marker datasets in various crops (Edwards and Batley, 2010; Yu et al., 2011; Edwards et al., 2013). However, in large-scale commercial breeding operations, this option remains financially unviable.

To address the balance between capturing genomic diversity and maintaining financial feasibility, the imputation of marker data from whole-genome sequencing was introduced, as discussed in Chapter 4. Numerous studies have documented the positive effects of imputation from smaller to larger marker sets in animal breeding (Zhang and Druet, 2010; Berry and Kearney, 2011; Cleveland and Hickey, 2013; Tsai et al., 2017; Song et al., 2019; Kriaridou et al., 2020; Mancin et al., 2021). This was also extended to applications in plant breeding (Hickey et al., 2012) and simulations in plant breeding (Hickey et al., 2015; Gorjanc et al., 2017a, 2017b).

However, our attempts to replicate these findings in rapeseed proved unsuccessful when imputing from a relatively high-density SNP dataset (19,846 markers) to whole-genome sequencing marker data containing 403,080 high-quality SNPs. In Chapter 4 this is attributed

to the fact that markers from the SNP array already effectively capture genome-wide QTL effects. Additionally, imputation errors, exacerbated by the inherent complexity of the rapeseed genome (Hurgobin et al., 2018; Lee et al., 2020; Sourdille and Jenczewski, 2021), may have contributed to the lack of success.

Across the three studies described in Chapters 2, 3, and 4, relationships emerged as a potential explanation for the lack of improvement in prediction accuracy. Relationships based on failed allele calls, imputed whole-genome sequencing marker data, and often also those based on haplotype blocks, exhibited a high resemblance to relationships based on SNPs, as indicated by the correlation between relationship estimates. Indeed, the relationship between the training and prediction (or validation) population is an important driver of prediction accuracy (Habier et al., 2013). High prediction accuracies are frequently achieved when the training and validation/prediction population are closely related (VanRaden, 2008; Hayes et al., 2009), whereas accuracy tends to decline with increasingly unrelated validation/prediction individuals (Wolc et al., 2011; Habier et al., 2013). Therefore, it is advisable, before conducting genomic prediction and irrespective of the sequencing technology employed, to assess the added information by evaluating the correlation between relationship estimates of some test individuals based on existing information (e.g. SNPs or pedigree information) and that of the sequencing or marker technology under consideration. This assessment aims to ascertain whether genuinely new information, particularly in terms of relationship, is introduced.

So far, the discussion has primarily centered on genetic markers within the realm of genomic prediction. While genetic information remains a crucial focus in this context, scientists have also leveraged epigenetic patterns to predict genotypic values. Notably, RNA and DNA methylation data have proven to result in genomic prediction accuracies competitive to SNPs (Hu et al., 2015; Westhues et al., 2017; Schrag et al., 2018; Li et al., 2019). However, the applicability of these markers in breeding is still under exploration, given their dependence on environmental conditions and a limited degree of heritability. Nevertheless, certain methylation patterns could be shown to be stable across generations (Becker et al., 2011; Schmitz et al., 2011; Hagmann et al., 2015; Hofmeister et al., 2017), indicating their heritability and potential utility in breeding (Crisp et al., 2022).

A novel avenue in predictive breeding is phenomic prediction, which predominantly relies on near-infrared spectroscopy patterns of seeds or leaves, yielding promising prediction accuracies (Rincent et al., 2018). The study indicates that utilizing near-infrared spectroscopy of seeds makes phenomic prediction a cost-effective alternative to genome-wide SNPs. Additionally, combining this profile with SNPs has the potential to capture environmental effects.

5.2 Factors influencing genomic prediction accuracy

Several factors significantly influence genomic prediction accuracy, with some tied to the marker data used and others associated with the characteristics of the training population. The impact of training set size on prediction accuracy is evident in Chapter 4, highlighting an asymptotic increase in accuracy as the training set size grows. This phenomenon was also observed in existing literature (Heffner et al., 2011a; Habier et al., 2013; Norman et al., 2018; Fernández-González et al., 2023; Wu et al., 2023). Maintaining a reasonable training set size is crucial for reliably predicting marker effects. However, training set optimization allows to decrease the training set size without compromising genomic prediction accuracy to some degree (Akdemir and Isidro-Sánchez, 2019; Fernández-González et al., 2023).

Prediction accuracies are notably higher when predicting close relatives (VanRaden, 2008; Hayes et al., 2009), emphasizing the importance of accurately estimating relationships. Our study revealed that failed allele calls, imputed whole-genome sequencing marker data, and certain haplotype blocking procedures capture relationship coefficients highly correlated to those of SNPs. This suggests that SNPs from an array may already offer a representative sample of the genome for relationship estimation. Accurate pedigree information also serves as a reliable source of relationship data, yielding competitive genomic prediction accuracy (Schrag et al., 2018; Calleja-Rodriguez et al., 2020). However, genomic data holds a theoretical advantage because it can be implemented without necessitating knowledge of pedigree information. This allow the modeling of Mendelian sampling and eliminates the need to make assumptions about relationships between founders or relatives.

Population structure, as illustrated by Werner et al. (2020), is another important factor influencing prediction accuracy. The results in Chapters 2, 3 and 4 support this observation, revealing diminished prediction accuracy in cross-validations across subpopulations or

families when compared to a random cross-validation scheme. Notably, the investigation described in Chapter 3 unveiled a noteworthy finding: Failed allele calls proved highly valuable in delineating the population structure within maize. Surprisingly, these failed allele calls facilitated a more accurate identification of Flint and Dent heterotic pools than the pool differentiation via SNPs. This underscores their utility in capturing nuances of the population structure. The interplay between relationship and population structure suggests that insights regarding one may extend to the other.

LD between markers and QTL is a persistent contributor to prediction accuracy across generations (Habier et al., 2013). In Chapter 4, average intrachromosomal pairwise marker LD was utilized as a proxy for LD between markers and QTL, revealing a considerable increase with imputed whole-genome sequencing marker data. However, this did not translate into improved prediction accuracy. Additionally, efforts to filter failed allele calls using LD did not enhance accuracy, as LD with SNPs potentially already captured the information present in the failed allele calls. This indicates that QTL that are captured by failed allele calls or imputed whole-genome sequencing marker data are already captured by SNPs. Explicitly, incorporating the LD structure through LD haplotype blocks did not yield general improvements in genomic prediction. This suggests that information on LD is already captured in prediction models utilizing SNPs. This observation is particularly true for rrBLUP, where marker effects are simultaneously predicted.

Cosegregation, where two loci originate from the same parental genotype, is identified as another contributor to prediction accuracy (Habier et al., 2007, 2013). However, across Chapters 2, 3, and 4, we were unable to draw implications about cosegregation due to limitations in the population design.

Irrespective of the contributing factors, the heritability of a given trait establishes a theoretical upper limit for genomic prediction accuracy, assuming correct heritability calculations. This principle is further substantiated by the simulation studies undertaken in Chapters 3 and 4. In these chapters, it becomes evident that anticipating high prediction accuracies is unrealistic in instances where heritability estimates are low.

Disentangling all contributing factors to prediction accuracy is challenging, and it is often only feasible through simulations rather than empirical breeding datasets. A comprehensive overview of this topic, including detailed simulations, is provided by Habier et al. (2013).

5.3 The challenges associated with plant genomes

Many plants investigated in the field of crop science have often undergone multiple whole-genome duplication (Almeida-Silva and Van de Peer, 2023) or polyploidization events (Adams and Wendel, 2005) throughout their evolutionary history. Polyploidy is categorized into autopolyploidy, where a plant possesses more than one pair of homologous chromosomes, and allopolyploidy, wherein interspecific hybridization between distinct, often closely related plant species results in a plant carrying two sets of homeologous chromosomes, one from each species (Adams and Wendel, 2005). Here, rapeseed was examined in Chapters 2 to 4 and wheat in Chapter 1, both of which are allopolyploid, carrying two and three chromosome sets respectively. In canola, the two subgenomes are very closely related, with nearly identical gene order and very similar gene sequences. This presents a challenge in SNP arrays, because some SNP probes may map to different homoeologous regions in the genome (Makhoul et al., 2020). Moreover, homoeologous exchanges among closely related chromosomes can cause a positional shift of an SNP locus (Zhang et al., 2020). This challenge is exacerbated by the reliance on reference genomes for determining chromosomal positions, which may not accurately represent the true positions in the assessed population and may vary among individuals within the population. Hence, caution is advised when assigning markers to physical positions (Makhoul et al., 2020). While this phenomenon is particularly prominent in allopolyploids, it is not strictly confined to polyploids, as observed as chromosomal translocations in diploids such as maize (Sheridan and Auger, 2006) and soybean (Wang et al., 2021).

The limitations concerning chromosomal position of markers impedes their use in pipelines where a fixed chromosomal position is vital for meaningful outcomes. In both haplotype blocks and the imputation of whole-genome sequencing data, the utilization of the physical position of SNP markers on a given reference genome was necessary. Incorrect or variable positions in the examined population, even among a few individuals, can result in incorrect LD estimates between markers and imputation errors. Assumed neighboring markers may be

millions of base pairs apart or on an entirely different chromosome. In the case of haplotype blocks based on LD, this can lead to incorrect haplotype block borders and incorrect haplotypes in all blocking methods. Furthermore, the imputation of whole-genome sequencing marker data may also be erroneous. This, in turn, can lead to the inference of numerous false and rare haplotypes, complicating the accurate estimation of their effects, especially as the robust estimation of effects for rare variants demands large population sizes (Meuwissen et al., 2001; Goddard et al., 2011). The observed frequencies and the numbers of haplotypes per block reported in the examples in Chapters 2 and 4 indicate that most haplotypes are rare and potentially spurious. Specifically, the investigation of the nested association mapping population in Chapter 4 suggests that the prevalence of these rare haplotypes may stem from imputation errors. This not only renders effect estimation challenging, but may also lead to inaccuracies, as it involves estimating effects for haplotypes that may not genuinely exist in the population.

In summary, caution is advised when using pipelines that assume a known physical position of a marker, necessitating a high-quality reference genome that is ideally closely related to the examined population, in order to to ensure positional overlap between the population and the reference genome. One potential approach to address marker position issues is the utilization of multiple reference genomes in pangenomes (Edwards and Batley, 2022). However, it remains to be demonstrated if this approach will enhance genomic prediction, as both examined methods which utilize physical positions in genomic prediction generally did not significantly improve genomic prediction, except under conditions of very poor model performance (as shown for wheat in Chapter 2).

Adding another layer of complexity to plant genomes is the abundance of structural variants, including deletions, insertions, translocations, inversions and copy number variants (Feschotte et al., 2002; Yuan et al., 2021). Identifying these variants is inherently challenging and typically requires extensive long-read sequencing (Chawla et al., 2021). Nevertheless, as outlined in the review by Gabur et al. (2018), numerous studies have extensively demonstrated the association of structural variants with critical agricultural traits. In Chapter 3, this connection was validated through failed allele calls from SNP arrays.

5.4 SNP arrays are a reliable genotype resource

SNP arrays have become widely accessible for numerous crops, with availability for various species (Ganal et al., 2011; Song et al., 2013; Unterseer et al., 2014; Tian et al., 2015; Bayer et al., 2017; Mason et al., 2017). For instance, as outlined in the introduction, the genotyping service provider which generated the data used in this thesis provides array genotyping services for a wide range of crops. Furthermore, the major SNP array manufacturers offer the possibility for anyone to obtain a custom SNP array tailored to his or her specific requirements and populations.

In general, these arrays present a straightforward and efficient high-throughput technology for assessing genotypes within large breeding populations, utilizing a fixed set of genomic markers carefully selected during array development (Dalton-Morgan et al., 2014; Mason et al., 2017). The widespread adoption of SNP markers extends across both animal and plant breeding disciplines. Even post-SNP array design, these arrays undergo thorough reevaluation and optimization (Boichard et al., 2012; Bayer et al., 2017). Coupled with stringent filtering and quality control processes preceding genomic prediction, as described in in Chapters 2 to 4, they stand as a dependable source of genotypic information that is challenging to surpass without additional non-redundant genotypic data.

5.5 Genomic prediction: Does the prediction model matter?

Chapters 2 to 4 explored various prediction models for genomic prediction. The models included GBLUP (Bernardo, 1994) and an extension of GBLUP to account for additive-by-additive epistasis (EGBLUP) (Henderson, 1985; Jiang and Reif, 2015), along with Bayesian LASSO (Park and Casella, 2008), RKHS (de los Campos et al., 2009), support vector machines (Chang and Lin, 2011) and extreme gradient boosting (Friedman, 2001). It is important to note that this selection represents only a fraction of the possible prediction models. For instance, within the Bayesian framework alone, the "BGLR" package (Pérez and de los Campos, 2014) encompasses 6 models.

In Chapter 2 and 3 within the maize population, model differences were observed between frequentist approaches (e.g. GBLUP, EGBLUP) and models within a Bayesian framework (e.g. Bayesian LASSO, RKHS). However, in most of the remaining traits and datasets, no great

differences between prediction models were identified. This aligns with the broader body of research on genomic prediction models, where major discrepancies are generally not observed although small differences exist between models (Crossa et al., 2013, 2014; Zhao et al., 2013; Jiang and Reif, 2015; Jiang et al., 2018; Werner et al., 2018; Knoch et al., 2021).

In Chapters 2, 3, and 4, no substantial differences were observed between models accounting for epistasis (e.g. RKHS and EGBLUP) and those that do not (e.g. GBLUP and Bayesian LASSO). While this might suggest the absence of nonlinear marker effects or epistasis, such a conclusion oversimplifies the relationship between genotype and phenotype. Examining the average effect of an allelic substitution reveals that only in the absence of dominance or epistasis does the average effect arise solely from additivity (Lynch and Walsh, 1998; Falconer and Mackay, 2009). In any other scenario, the average effect cannot be disentangled from the influence of dominance or epistasis. Therefore, even a simple GBLUP or rrBLUP captures a broad spectrum of genetic variance, including dominance and epistasis effects. This could potentially explain the lack of improvements using models other than GBLUP.

Typically, models like GBLUP or rrBLUP are preferred for traits with a highly quantitative structure, assuming equal contribution of each locus to genetic variance. However, as one transitions to traits with a more qualitative architecture, or with large-effect QTL, models accommodating this kind of variability might be superior. These include Bayesian LASSO, where not all markers receive the same variance, or GBLUP, which includes a fixed factor for known QTL (Spindel et al., 2016; Werner et al., 2018). Nevertheless, in the case of flowering time in rapeseed, a trait known to be influenced by several major QTL (Schiessl, 2020), no beneficial effect of Bayesian LASSO was observed in Chapters 2 and 3. This is consistent with findings by Werner et al. (2018) with the method “RR-BLUP + de novo GWAS”.

Despite the assumption that machine learning models are capturing nonlinear relationships between markers and phenotypes (Montesinos López et al., 2022), no superiority of machine learning-based genomic prediction was observed in Chapter 3. This observation corroborates findings from previous studies, emphasizing that the performance of such models may not consistently surpass traditional approaches (González-Recio et al., 2010; Long et al., 2010; Gianola et al., 2011; Heslot et al., 2012; Montesinos-López et al., 2018; Azodi et al., 2019; Perez et al., 2022; Chen et al., 2023).

Moreover, it is argued that machine learning performs well in scenarios with large datasets (i.e., n>p problems). This may not be fully applicable in the context of most genomic prediction studies, including those discussed here, where the sample size is still comparable to or smaller than the number of predictors. In such cases, models that can fit flexible nonlinear functions may run the risk of overfitting (Montesinos López et al., 2022). In comparison, GBLUP might be less prone to overfitting, because it represents an inflexible form of penalized linear regression. Nevertheless, the potential for machine learning may become more evident as we approach scenarios with larger training populations (n>p) or, ideally, considerably larger training populations.

5.6 How genomic prediction helps in a breeding

The extensive discussion and comparison of genomic prediction across Chapters 2 to 4 underline its potential utility and impact within a breeding program. The overarching objective of a breeding program is to genetically optimize traits of interest, with a primary focus on enhancing yields, encompassing seed yield, biomass and end-product yields such as oil, protein, or sugar yield in major crops. This success is commonly denoted as the genetic gain in a target trait, and its quantification is expressed through the response to selection, as formulated by Lush (1937):

$$R = ih\sigma_g$$

where R is the response to selection, i is the selection intensity, i.e. how much the selected fraction is better than the base population expressed in standard deviations, h is the square root of the heritability and σ_g is the square root of the genotypic variance. Some authors divide the response to selection by time or breeding cycle duration to underscore the importance of the time aspect in breeding. However, such a division oversimplifies the breeding process, implying a reduction to a single selection step per breeding cycle. Further, as eloquently expressed by Frisch (2023), "All successful line breeding programs use multi-stage selection".

Genomic prediction does not directly contribute parameters to the calculation of the response to selection. Nevertheless, it empowers breeders to estimate genotypic values of genotypes without phenotype, thereby expanding the base population and potentially enabling the

selection of superior genotypes for ongoing breeding programs. The extension of the population introduces genetic diversity for the trait, potentially impacting genetic variance, with the accuracy of the prediction model influencing the achievable response to selection. Moreover, having knowledge of the loci influencing a trait through a genomic prediction model enables more precise and informed selection.

Recognizing the time aspect in breeding, genomic prediction proves instrumental in increasing the response to selection per unit time. By partially substituting extensive multi-year field trials, it minimizes the time component, facilitating the identification of promising crossing partners in early stages or even without testing, ultimately shortening the breeding cycle (Heffner et al., 2009, 2011b).

Studies in both animal and plant breeding underscore the benefits of genomic prediction. Several simulation studies (Bernardo and Yu, 2007; Bernardo, 2010; Gaynor et al., 2017; Voss-Fels et al., 2021; Wientjes et al., 2022) demonstrate that genomic prediction can outperform classical phenotypic selection for various traits. Even if genomic prediction does not significantly improve response to selection, it can nevertheless achieve comparable gains with lower costs (Beyene et al., 2019). This approach is already being applied in breeding programs (Cooper et al., 2014; Crossa et al., 2014; Gaffney et al., 2015) with notable success. A notable example with extremely high success was the development of Aquamax© maize/corn varieties for the US cornbelt, assisted by genomic prediction (Cooper et al., 2014). Furthermore, after a decade of application in US cattle breeding, genomic prediction has been shown to also improve traits with very low heritability (García-Ruiz et al., 2016).

While genomic prediction proves advantageous in the short run for selecting beneficial crossing parents and developing superior varieties, in the long run it can pose challenges related to allelic diversity. The intense selection processes facilitated by genomic prediction can lead to the loss of alleles, narrowing down the genetic variance and thereby limiting long time genetic gain (Jannink, 2010; Rutkoski et al., 2015; Gaynor et al., 2017). Thus, a breeding program employing genomic selection must also consider allelic diversity. The "optimal contribution selection" strategy, widely employed in animal breeding, attempts to balance the loss of genetic variation against long-term genetic gain (Meuwissen, 1997). Although originally proposed for animal breeding with pedigree-based relationship coefficients, its

implementation in plant breeding, with complex pedigrees, is challenging (Shaw et al., 2014). However, the introduction of genomic information in optimal contribution selection allows replacement of pedigree relationships by realized relationships determined with genotypic information. Genomic optimal contribution selection has proven more efficient in controlling inbreeding compared to pedigree-based approaches (Sonesson et al., 2012). In summary, a breeder must balance allelic diversity and genetic gain, recognizing that the depletion of pertinent allelic diversity typically occurs after several generations, often spanning many years.

However, genomic prediction is not the only key to breeding success, and breeders still need to conduct accurate multi-environmental trials to generate sufficient and updated training data for genomic prediction. As the breeding population progresses through cycles, the training population becomes increasingly unrelated to the original prediction individuals, potentially limiting prediction accuracy (Auinger et al., 2016). Moreover, in these trials, it is essential to emphasize the critical role of precise trial execution to avoid constraining heritability due to high residual variance. Heritability, being a direct factor in the calculation of the response to selection, underscores the significance of meticulous trial procedures.

6 Conclusions

This thesis provides a further example for the potential of genomic prediction across diverse crops, demonstrating the capability of this technique to enhance genetic gains within plant breeding programs. The findings presented here underline the robustness of genomic prediction, showcasing high prediction accuracies irrespective of the statistical model that is employed. Notably, even the simple GBLUP model demonstrates prediction accuracies that are competitive to those of more complex machine learning models, thereby rendering genomic prediction feasible for virtually any breeding program, contingent upon the availability of comprehensive phenotypic and genotypic datasets. Additionally, the widespread integration of high-throughput array genotyping frameworks for most major plant species establishes genotyping as an accessible tool across virtually any crop.

The SNP arrays employed in the practical examples described in this thesis emerged as dependable markers for accurate genomic predictions across the four crops investigated. However, the exploration of alternative marker types called from arrays failed to yield

significant enhancements in prediction accuracy. This limitation can be attributed to the redundancy in estimating relationships, resulting in a lack of novel information added by those markers.

Although the application of haplotype blocks did not improve genomic prediction accuracy in general, sometimes they exhibited positive effects on prediction accuracy in scenarios with comparably underperforming models. However, to exploit these beneficial effects it is necessary to meticulously tune haplotype construction parameters, treating them as crucial hyperparameters. Further investigation revealed that failed allele calls from SNP arrays are linked to quantitative traits, likely due to deletions spanning the SNP loci. While the inclusion of this information did not lead to heightened prediction accuracies, its utility in exploring population structure at no additional costs advocates for its integration into SNP array data analysis pipelines. Similarly, the imputation of whole-genome sequencing data failed to surpass the predictive performance achieved with SNP arrays. The lack of improvement across different levels of marker resolution can be explained through the redundant relationship information explored in SNPs and the examined methods to increase the information content of SNP arrays.

In summary, while alternative marker types showed limited impact on prediction accuracy, their nuanced application and the incorporation of additional information can contribute to a more comprehensive understanding of genomic landscapes in crop plants with negligible additional costs.

7 Summary

Genomic prediction is a promising tool for improving genetic gains in various crops, serving as a valuable tool for plant breeders. SNP arrays are the preferred genotyping tool for breeders of most major crops, however the limited predefined marker number associated with SNP arrays has the potential to impede achievable prediction accuracy in genomic prediction.

The objective of this study was to evaluate cost-effective methods for maximizing the information content of SNP arrays. Three methods were explored and their information content was assessed using prediction accuracies from six genomic prediction models across diverse crops and agronomic traits. Independently of the method used to increase the

information content of SNP arrays, the applied genomic prediction models consistently demonstrated similar performance in terms of prediction accuracy within traits, making them equally suitable for genomic prediction across a variety of crops and traits.

The first method to maximize the information content of SNP arrays involved constructing haplotype blocks with various methods and parameters and utilizing their haplotypes for genomic prediction. Analyzing data from rapeseed, maize, wheat and soybean in genomic prediction models revealed only marginal improvements in genomic prediction accuracy across most traits. Notably, haplotype blocks demonstrated effectiveness in compensating for poorly performing models in scenarios with highly variable prediction accuracies across prediction models. Nevertheless, the absence of a consistent ideal method or parameter for constructing haplotype blocks makes them a hyperparameter requiring careful tuning.

Furthermore, failed allele calls from SNP arrays were examined for their information content in genomic prediction of agronomic traits in maize and rapeseed. Two statistical pipelines were developed and tested to filter non-random failed allele calls from random technical errors. Surprisingly, failed allele calls, potentially originating from genome structural variants, exhibited prediction accuracies comparable to genome-wide SNP datasets. However, the combination of SNPs and failed allele calls did not enhance genomic prediction.

As an alternative to whole-genome sequencing marker data, imputation of whole-genome sequencing marker data from SNP arrays was explored. While there was a considerable improvement in LD and marker density, no increase in prediction accuracy was observed. This can likely be attributed to erroneous haplotypes and marker calls resulting from imputation errors. A suitable hypothesis to explain this observation is that these errors are introduced by the high complexity and redundancy of crop plant genomes.

Across all three methods, relationships emerged as an explanation for the lack of improvement in genomic prediction accuracy. Relationship estimates exhibited a high correlation between those obtained from SNP array data and methods to increase the information content of SNP arrays, contributing predominantly redundant information. Moreover, it can be assumed that markers on arrays generally exhibit sufficient LD with adjacent QTL.

In conclusion, SNP arrays were proven to be a reliable genotyping technology, offering a representative sample of the genome for estimating relationships. Furthermore, this study reaffirms the potential of genomic prediction as a breeding tool to improve genetic gain in several crops.

8 Zusammenfassung

Die genomische Vorhersage ist ein vielversprechendes Werkzeug zur Verbesserung des Zuchtfortschritts in vielen Nutzpflanzen und ist deshalb ein wertvolles Instrument für Pflanzenzüchter. Der SNP-Array ist die bevorzugte Methode der Genotypisierung für die meisten Züchter; jedoch hat die begrenzte und vordefinierte Anzahl von Markern von SNP-Arrays das Potenzial, die erreichbare Vorhersagegenauigkeit bei der genomischen Vorhersage zu limitieren. Deshalb war das Ziel dieser Studie, kostengünstige Methoden zur Maximierung des Informationsgehalts von SNP-Arrays für die genomische Vorhersage zu evaluieren. Hier wurden drei Methoden untersucht, und ihr Informationsgehalt wurde anhand der Vorhersagegenauigkeit von sechs genomischen Vorhersagemodellen über verschiedene Kulturen und agronomische Merkmale hinweg bewertet. Unabhängig von der Methode zur Steigerung des Informationsgehalts von SNP-Arrays zeigten die angewendeten genomischen Vorhersagemodelle konsistent ähnliche Leistungen in Bezug auf Vorhersagegenauigkeit innerhalb von Merkmalen, was sie gleichermaßen für die genomische Vorhersage in verschiedenen Kulturen und Merkmalen geeignet macht.

Die erste Methode zur Maximierung des Informationsgehalts von SNP-Arrays bestand darin, Haplotypenblöcke mit verschiedenen Methoden und Parametern zu erstellen und ihre Haplotypen für die genomische Vorhersage zu nutzen. Die Analyse von Daten aus Raps, Mais, Weizen und Sojabohnen mit genomischen Vorhersagemodellen zeigte nur marginale Verbesserungen in der genomischen Vorhersagegenauigkeit über die meisten Merkmale hinweg. Insbesondere zeigten Haplotypenblöcke ihre Wirksamkeit, um Modellunterschiede in Szenarien mit stark variablen Vorhersagegenauigkeiten zwischen Vorhersagemodellen auszugleichen. Dennoch machen die fehlenden ideale Methoden oder Parameter für die Konstruktion von Haplotypenblöcken sie zu einem Hyperparameter, der sorgfältig abgestimmt werden muss.

Darüber hinaus wurden „failed allele calls“ von SNP-Arrays (Loci, an denen keins der beiden Allele gefunden werden konnte) auf ihren Informationsgehalt in der genomischen Vorhersage für agronomische Merkmale in Mais und Raps untersucht. Überraschenderweise zeigten „failed allele calls“, die möglicherweise auf strukturelle Genomvariationen zurückzuführen sind, Vorhersagegenauigkeiten, die mit denjenigen von genomweiten SNPs vergleichbar waren. Zusätzlich wurden zwei statistische Pipelines entwickelt und getestet, um systematische „failed allele calls“ von zufälligen technischen Fehlern zu filtern, wobei Vorhersagegenauigkeiten ähnlich zu denen von SNPs beobachtet wurden. Die Kombination von SNPs und fehlgeschlagenen Allel-Calls verbesserte jedoch nicht die genomische Vorhersage.

Als Alternative zu Markerdaten aus der Sequenzierung des gesamten Genoms wurde die Imputation von diesen basierend auf genomweiten SNP-Markern vom Array untersucht. Dies ist wahrscheinlich auf fehlerhafte Haplotypen und Marker-Calls aufgrund von Imputations- und Genotypisierungsfehlern zurückzuführen. Es kann vermutet werden, dass diese Fehler auf die hohe Komplexität und Redundanz von Kulturpflanzengenomen zurückzuführen sind.

Über alle drei Methoden hinweg ergaben sich die Bestimmung der Verwandtschaft als Erklärung für das Fehlen einer Verbesserung der genomischen Vorhersagegenauigkeit. Verwandschaftskoeffizienten zeigten eine hohe Korrelation zwischen denen, die aus den Daten von SNP-Arrays und Methoden zur Steigerung des Informationsgehalts von SNP-Arrays erhalten wurden, und trugen deshalb hauptsächlich redundante Informationen bei. Darüber kann man davon ausgehen, dass Marker auf Arrays in der Regel in ausreichende LD mit benachbarten QTL sind.

Zusammenfassend erweisen sich SNP-Arrays als zuverlässige Genotypisierungstechnologie, die eine repräsentative Stichprobe des Genoms für die Schätzung von Verwandtschaft bietet. Darüber hinaus bestätigt diese Studie das Potenzial der genomischen Vorhersage als Züchtungsinstrument zur Verbesserung genetischer Fortschritte in mehreren Nutzpflanzen.

9 References

Adam-Blondon, A.-F., Martínez-Zapater, J. M., and Kole, C. eds. (2011). *Genetics, genomics and breeding of grapes*. Enfield, NH : Boca Raton, FL: Science Publishers ; Marketed and distributed by CRC Press.

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology* 8, 135–141. doi: 10.1016/j.pbi.2005.01.001.

Adhikari, K. N., Khazaei, H., Ghaouti, L., Maalouf, F., Vandenberg, A., Link, W., et al. (2021). Conventional and Molecular Breeding Tools for Accelerating Genetic Gain in Faba Bean (Vicia Faba L.). *Frontiers in Plant Science* 12. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2021.744259 [Accessed December 8, 2023].

Akdemir, D., and Isidro-Sánchez, J. (2019). Design of training populations for selective phenotyping in genomic prediction. *Sci Rep* 9, 1446. doi: 10.1038/s41598-018-38081-6.

Almeida-Silva, F., and Van de Peer, Y. (2023). Whole-genome Duplications and the Long-term Evolution of Gene Regulatory Networks in Angiosperms. *Molecular Biology and Evolution* 40, msad141. doi: 10.1093/molbev/msad141.

Auinger, H.-J., Schönleben, M., Lehermeier, C., Schmidt, M., Korzun, V., Geiger, H. H., et al. (2016). Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). *Theor Appl Genet* 129, 2043–2053. doi: 10.1007/s00122-016-2756-5.

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.

Ballesta, P., Maldonado, C., Pérez-Rodríguez, P., and Mora, F. (2019). SNP and Haplotype-Based Genomic Selection of Quantitative Traits in Eucalyptus globulus. *Plants* 8, 331. doi: 10.3390/plants8090331.

Barrie Wetherill, G., Duncombe, P., Kenward, M., Köllerström, J., Paul, S. R., and Vowden, B. J. (1986). "Multicollinearity," in *Regression Analysis with Applications* Monographs on Statistics and Applied Probability., eds. G. B. Wetherill, P. Duncombe, M. Kenward, J. Köllerström, S. R. Paul, and B. J. Vowden (Dordrecht: Springer Netherlands), 82–107. doi: 10.1007/978-94-009-4105-2_4.

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. *Frontiers in Plant Science* 8. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2017.01792 [Accessed January 4, 2023].

Becker, C., Hagmann, J., Müller, J., Koenig, D., Stegle, O., Borgwardt, K., et al. (2011). Spontaneous epigenetic variation in the Arabidopsis thaliana methylome. *Nature* 480, 245–249. doi: 10.1038/nature10555.

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.

Bernardo, R. (1994). Prediction of Maize Single-Cross Performance Using RFLPs and Information from Related Hybrids. *Crop Science* 34, cropsci1994.0011183X003400010003x. doi: 10.2135/cropsci1994.0011183X003400010003x.

Bernardo, R. (1997). RFLP markers and predicted testcross performance of maize sister inbreds. *Theor Appl Genet* 95, 655–659. doi: 10.1007/s001220050608.

Bernardo, R. (2009). Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. *Crop Science* 49, 419–425. doi: 10.2135/cropsci2008.08.0452.

Bernardo, R. (2010). Genomewide Selection with Minimal Crossing in Self-Pollinated Crops. *Crop Science* 50, 624–627. doi: 10.2135/cropsci2009.05.0250.

Bernardo, R. (2014). *Essentials of plant breeding*. Woodbury, Minnesota: Stemma Press.

Bernardo, R. N. (2020). *Breeding for quantitative traits in plants*. Third edition. Woodbury, Minnesota: Stemma Press.

Bernardo, R., Romero-Severson, J., Ziegle, J., Hauser, J., Joe, L., Hookstra, G., et al. (2000). Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. *Theor Appl Genet* 100, 552–556. doi: 10.1007/s001220050072.

Bernardo, R., and Yu, J. (2007). Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Science* 47, 1082–1090. doi: 10.2135/cropsci2006.11.0690.

Berry, D. P., and Kearney, J. F. (2011). Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *Animal* 5, 1162–1169. doi: 10.1017/S1751731111000309.

Beyene, Y., Gowda, M., Olsen, M., Robbins, K. R., Pérez-Rodríguez, P., Alvarado, G., et al. (2019). Empirical Comparison of Tropical Maize Hybrids Selected Through Genomic and Phenotypic Selections. *Frontiers in Plant Science* 10. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01502 [Accessed December 14, 2023].

Bohn, M., Utz, H. F., and Melchinger, A. E. (1999). Genetic Similarities among Winter Wheat Cultivars Determined on the Basis of RFLPs, AFLPs, and SSRs and Their Use for Predicting Progeny Variance. *Crop Science* 39, cropsci1999.0011183X003900010035x. doi: 10.2135/cropsci1999.0011183X003900010035x.

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.

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics* 103, 338–348. doi: 10.1016/j.ajhg.2018.07.015.

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. *The American Journal of Human Genetics* 81, 1084–1097. doi: 10.1086/521987.

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., et al. (2011). Targeted Amplicon Sequencing (TAS): A Scalable Next-Gen Approach to Multilocus, Multitaxa Phylogenetics. *Genome Biol Evol* 3, 1312–1323. doi: 10.1093/gbe/evr106.

Calleja-Rodriguez, A., Pan, J., Funda, T., Chen, Z., Baison, J., Isik, F., et al. (2020). Evaluation of the efficiency of genomic versus pedigree predictions for growth and wood quality traits in Scots pine. *BMC Genomics* 21, 796. doi: 10.1186/s12864-020-07188-4.

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 Biotechnology Journal* 19, 240–250. doi: 10.1111/pbi.13456.

Chen, C., Powell, O., Dinglasan, E., Ross, E. M., Yadav, S., Wei, X., et al. (2023). Genomic prediction with machine learning in sugarcane, a complex highly polyploid clonally propagated crop with substantial non-additive variation for key traits. *The Plant Genome* n/a, e20390. doi: 10.1002/tpg2.20390.

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.

Cleveland, M. A., and Hickey, J. M. (2013). Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation1. *Journal of Animal Science* 91, 3583–3592. doi: 10.2527/jas.2013-6270.

Cochran, W. G., and Cox, G. M. (1992). *Experimental designs*. 2nd ed. New York: Wiley.

Cockerham, C. C. (1980). Random and fixed effects in plant genetics. *Theoret. Appl. Genetics* 56, 119–131. doi: 10.1007/BF00265082.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., and Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142, 169–196. doi: 10.1007/s10681-005-1681-5.

Collard, B. C. Y., and Mackill, D. J. (2007). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363, 557–572. doi: 10.1098/rstb.2007.2170.

Cooper, M., Gho, C., Leafgren, R., Tang, T., and Messina, C. (2014). Breeding drought-tolerant maize hybrids for the US corn-belt: discovery to product. *Journal of Experimental Botany* 65, 6191–6204. doi: 10.1093/jxb/eru064.

Crisp, P. A., Bhatnagar-Mathur, P., Hundleby, P., Godwin, I. D., Waterhouse, P. M., and Hickey, L. T. (2022). Beyond the gene: epigenetic and cis-regulatory targets offer new breeding potential for the future. *Current Opinion in Biotechnology* 73, 88–94. doi: 10.1016/j.copbio.2021.07.008.

Crossa, J., Beyene, Y., Kassa, S., Pérez, P., Hickey, J. M., Chen, C., et al. (2013). Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing. *G3 Genes|Genomes|Genetics* 3, 1903–1926. doi: 10.1534/g3.113.008227.

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 in Plant Science* 22, 961–975. doi: 10.1016/j.tplants.2017.08.011.

Cuyabano, B. C., Su, G., and Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics* 15, 1171. doi: 10.1186/1471-2164-15-1171.

Dalton-Morgan, J., Hayward, A., Alamery, S., Tollenaere, R., Mason, A. S., Campbell, E., et al. (2014). A high-throughput SNP array in the amphidiploid species Brassica napus shows diversity in resistance genes. *Funct Integr Genomics* 14, 643–655. doi: 10.1007/s10142-014-0391-2.

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., and Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nat Genet* 29, 229–232. doi: 10.1038/ng1001-229.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12, 499–510. doi: 10.1038/nrg3012.

Davis, G. P., and DeNise, S. K. (1998). The impact of genetic markers on selection. *Journal of Animal Science* 76, 2331–2339. doi: 10.2527/1998.7692331x.

de los Campos, G., Gianola, D., and Rosa, G. J. M. (2009). Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation1. *Journal of Animal Science* 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.

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.

Difabachew, Y. F., Frisch, M., Langstroff, A. L., Stahl, A., Wittkop, B., Snowdon, R. J., et al. (2023). Genomic prediction with haplotype blocks in wheat. *Frontiers in Plant Science* 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1168547 [Accessed June 26, 2023].

Druet, T., Macleod, I. M., and Hayes, B. J. (2014). Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity* 112, 39–47. doi: 10.1038/hdy.2013.13.

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

Edwards, D., and Batley, J. (2010). Plant genome sequencing: applications for crop improvement. *Plant Biotechnology Journal* 8, 2–9. doi: 10.1111/j.1467-7652.2009.00459.x.

Edwards, D., and Batley, J. (2022). Graph pangenomes find missing heritability. *Nat Genet* 54, 919–920. doi: 10.1038/s41588-022-01099-8.

Edwards, D., Batley, J., and Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126, 1–11. doi: 10.1007/s00122-012-1964-x.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11, 446–450. doi: 10.1038/nrg2809.

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 Physiology* 156, 1679–1690. doi: 10.1104/pp.111.174748.

Endelman, J. B., and Jannink, J.-L. (2012). Shrinkage Estimation of the Realized Relationship Matrix. *G3 Genes|Genomes|Genetics* 2, 1405–1413. doi: 10.1534/g3.112.004259.

Falconer, D. S., and Mackay, T. (2009). *Introduction to quantitative genetics*. 4. ed., [16. print.]. Harlow: Pearson, Prentice Hall.

Fernández-González, J., Akdemir, D., and Isidro y Sánchez, J. (2023). A comparison of methods for training population optimization in genomic selection. *Theor Appl Genet* 136, 30. doi: 10.1007/s00122-023-04265-6.

Feschotte, C., Jiang, N., and Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 3, 329–341. doi: 10.1038/nrg793.

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.

Frisch, M. (2023). A closer look at the breeder´s equation. AG Seminar Department of Biometry and Population Genetics. 09. October 2023

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

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

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., et al. (2002). The Structure of Haplotype Blocks in the Human Genome. *Science* 296, 2225–2229. doi: 10.1126/science.1069424.

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 Biotechnology Journal* 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., 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.

García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., and Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences* 113, E3995–E4004. doi: 10.1073/pnas.1519061113.

Gaynor, R. C., Gorjanc, G., Bentley, A. R., Ober, E. S., Howell, P., Jackson, R., et al. (2017). A Two-Part Strategy for Using Genomic Selection to Develop Inbred Lines. *Crop Science* 57, 2372–2386. doi: 10.2135/cropsci2016.09.0742.

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. *Journal of Animal Breeding and Genetics* 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. *Genetics Research* 92, 227–237. doi: 10.1017/S0016672310000261.

Gorjanc, G., Battagin, M., Dumasy, J.-F., Antolin, R., Gaynor, R. C., and Hickey, J. M. (2017a). Prospects for Cost-Effective Genomic Selection via Accurate Within-Family Imputation. *Crop Science* 57, 216–228. doi: 10.2135/cropsci2016.06.0526.

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R. C., Antolin, R., and Hickey, J. M. (2017b). Potential of Low-Coverage Genotyping-by-Sequencing and Imputation for Cost-Effective Genomic Selection in Biparental Segregating Populations. *Crop Science* 57, 1404–1420. doi: 10.2135/cropsci2016.08.0675.

Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177, 2389–2397. doi: 10.1534/genetics.107.081190.

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.

Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R. C., Wang, G., et al. (2015). Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLOS Genetics* 11, e1004920. doi: 10.1371/journal.pgen.1004920.

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.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* 92, 433–443. doi: 10.3168/jds.2008-1646.

Hayes, B. J., Macleod, I. M., Daetwyler, H. D., Bowman, P. J., Chamberlian, A. J., Vander Jagt, C. J., et al. (2014). Genomic prediction from whole genome sequence in livestock: the 1000 Bull Genomes Project. in *10. World Congress of Genetics Applied to Livestock Production* (Vancouver, Canada), np. Available at: https://hal.archives-ouvertes.fr/hal-01193911 [Accessed November 22, 2022].

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., and Sorrells, M. E. (2011a). Genomic Selection Accuracy for Grain Quality Traits in Biparental Wheat Populations. *Crop Science* 51, 2597–2606. doi: 10.2135/cropsci2011.05.0253.

Heffner, E. L., Jannink, J.-L., and Sorrells, M. E. (2011b). Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. *The Plant Genome* 4. doi: 10.3835/plantgenome2010.12.0029.

Heffner, E. L., Sorrells, M. E., and Jannink, J.-L. (2009). Genomic Selection for Crop Improvement. *Crop Science* 49, 1–12. doi: 10.2135/cropsci2008.08.0512.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Ann Math Stat* 21, 309–310.

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.

Herrmann, A. (2013). Biogas Production from Maize: Current State, Challenges and Prospects. 2. Agronomic and Environmental Aspects. *Bioenerg. Res.* 6, 372–387. doi: 10.1007/s12155-012-9227-x.

Heslot, N., Yang, H.-P., Sorrells, M. E., and Jannink, J.-L. (2012). Genomic Selection in Plant Breeding: A Comparison of Models. *Crop Science* 52, 146–160. doi: 10.2135/cropsci2011.06.0297.

Hickey, J. M., Chiurugwi, T., Mackay, I., and Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nat Genet* 49, 1297–1303. doi: 10.1038/ng.3920.

Hickey, J. M., Crossa, J., Babu, R., and de los Campos, G. (2012). Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Science* 52, 654–663. doi: 10.2135/cropsci2011.07.0358.

Hickey, J. M., Dreisigacker, S., Crossa, J., Hearne, S., Babu, R., Prasanna, B. M., et al. (2014). Evaluation of Genomic Selection Training Population Designs and Genotyping Strategies in Plant Breeding Programs Using Simulation. *Crop Science* 54, 1476–1488. doi: 10.2135/cropsci2013.03.0195.

Hickey, J. M., Gorjanc, G., Varshney, R. K., and Nettelblad, C. (2015). Imputation of Single Nucleotide Polymorphism Genotypes in Biparental, Backcross, and Topcross Populations with a Hidden Markov Model. *Crop Science* 55, 1934–1946. doi: 10.2135/cropsci2014.09.0648.

Hofmeister, B. T., Lee, K., Rohr, N. A., Hall, D. W., and Schmitz, R. J. (2017). Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biology* 18, 155. doi: 10.1186/s13059-017-1288-x.

Holland, J. B., Nyquist, W. E., and Cervantes-Martínez, C. T. (2002). "Estimating and Interpreting Heritability for Plant Breeding: An Update," in *Plant Breeding Reviews* (John Wiley & Sons, Ltd), 9–112. doi: 10.1002/9780470650202.ch2.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics* 5, e1000529. doi: 10.1371/journal.pgen.1000529.

Hu, Y., Morota, G., Rosa, G. J. M., and Gianola, D. (2015). Prediction of Plant Height in Arabidopsis thaliana Using DNA Methylation Data. *Genetics* 201, 779–793. doi: 10.1534/genetics.115.177204.

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 Biotechnology Journal* 16, 1265–1274. doi: 10.1111/pbi.12867.

Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001). Diversity Arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29, e25. doi: 10.1093/nar/29.4.e25.

Jannink, J.-L. (2010). Dynamics of long-term genomic selection. *Genetics Selection Evolution* 42, 35. doi: 10.1186/1297-9686-42-35.

Jannink, J.-L., Lorenz, A. J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics* 9, 166–177. doi: 10.1093/bfgp/elq001.

Jarquin, D., Howard, R., Crossa, J., Beyene, Y., Gowda, M., Martini, J. W. R., et al. (2020). Genomic Prediction Enhanced Sparse Testing for Multi-environment Trials. *G3 Genes|Genomes|Genetics* 10, 2725–2739. doi: 10.1534/g3.120.401349.

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29, 217–222. doi: 10.1038/ng1001-217.

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

Jiang, Y., Schmidt, R. H., and Reif, J. C. (2018). Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. *G3 (Bethesda)* 8, 1687–1699. doi: 10.1534/g3.117.300548.

Jighly, A., Thayalakumaran, T., O'Leary, G. J., Kant, S., Panozzo, J., Aggarwal, R., et al. (2023). Using genomic prediction with crop growth models enables the prediction of associated traits in wheat. *Journal of Experimental Botany* 74, 1389–1402. doi: 10.1093/jxb/erac393.

Jordan, D., Tao, Y., Godwin, I., Henzell, R., Cooper, M., and McIntyre, C. (2003). Prediction of hybrid performance in grain sorghum using RFLP markers. *Theor Appl Genet* 106, 559–567. doi: 10.1007/s00122-002-1144-5.

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. *Frontiers in Genetics* 11. Available at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00124 [Accessed February 9, 2023].

Kumar, L. S. (1999). DNA markers in plant improvement: An overview. *Biotechnology Advances* 17, 143–182. doi: 10.1016/S0734-9750(98)00018-4.

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. *Frontiers in Plant Science* 11. Available at: https://www.frontiersin.org/article/10.3389/fpls.2020.00496 [Accessed June 28, 2022].

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.

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 Phytologist* 196, 282–291. doi: 10.1111/j.1469-8137.2012.04243.x.

Li, Z., Gao, N., Martini, J. W. R., and Simianer, H. (2019). Integrating Gene Expression Data Into Genomic Prediction. *Front. Genet.* 10, 430679. doi: 10.3389/fgene.2019.00126.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2013). *SAS for mixed models*. 2. ed., 7. print. Cary, NC: SAS Institute.

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. *Genetics Research* 92, 209–225. doi: 10.1017/S0016672310000157.

Lush, J. L. (1937). *Animal Breeding Plans*. Ames, Iowa: Iowa State Press.

Lynch, M., and Walsh, B. (1998). *Genetics and analysis of quantitative traits*. Sunderland, Mass: Sinauer Assoc.

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.

Ma, Y., Reif, J. C., Jiang, Y., Wen, Z., Wang, D., Liu, Z., et al. (2016). Potential of marker selection to increase prediction accuracy of genomic selection in soybean (Glycine max L.). *Mol Breeding* 36, 113. doi: 10.1007/s11032-016-0504-9.

Makhoul, M., Rambla, C., Voss-Fels, K. P., Hickey, L. T., Snowdon, R. J., and Obermeier, C. (2020). Overcoming polyploidy pitfalls: a user guide for effective SNP conversion into KASP markers in wheat. *Theor Appl Genet* 133, 2413–2430. doi: 10.1007/s00122-020-03608-x.

Mancin, E., Sosa-Madrid, B. S., Blasco, A., and Ibáñez-Escriche, N. (2021). Genotype Imputation to Improve the Cost-Efficiency of Genomic Selection in Rabbits. *Animals* 11, 803. doi: 10.3390/ani11030803.

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. *Proceedings of the National Academy of Sciences* 110, 5241–5246. doi: 10.1073/pnas.1220766110.

Mason, A. S., Higgins, E. E., Snowdon, R. J., Batley, J., Stein, A., Werner, C., et al. (2017). A user guide to the Brassica 60K Illumina Infinium™ SNP genotyping array. *Theor Appl Genet* 130, 621–633. doi: 10.1007/s00122-016-2849-1.

Melchinger, A. e. (1993). "Use of RFLP Markers for Analysis of Genetic Relationships Among Breeding Materials and Prediction of Hybrid Performance," in *International Crop Science I* (John Wiley & Sons, Ltd), 621–628. doi: 10.2135/1993.internationalcropscience.c98.

Meuwissen, T. H. E. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *Journal of Animal Science* 75, 934–940. doi: 10.2527/1997.754934x.

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.

Montesinos López, O. A., Montesinos López, A., and Crossa, J. (2022). *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer International Publishing 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.

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 Breeding* 31, 27–37. doi: 10.1007/s11032-012-9765-0.

Norman, A., Taylor, J., Edwards, J., and Kuchel, H. (2018). Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy. *G3 Genes|Genomes|Genetics* 8, 2889–2899. doi: 10.1534/g3.118.200311.

Nyquist, W. E., and Baker, R. J. (1991). Estimation of heritability and prediction of selection response in plant populations. *Critical Reviews in Plant Sciences* 10, 235–322. doi: 10.1080/07352689109382313.

Park, T., and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 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.

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica* 161, 209–228. doi: 10.1007/s10681-007-9449-8.

Piepho, H.-P., and Möhring, J. (2007). Computing Heritability and Selection Response From Unbalanced Plant Breeding Trials. *Genetics* 177, 1881–1888. doi: 10.1534/genetics.107.074229.

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

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001). Linkage disequilibrium in the human genome. *Nature* 411, 199–204. doi: 10.1038/35075590.

Rincent, R., Charpentier, J.-P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., et al. (2018). Phenomic Selection Is a Low-Cost and High-Throughput Method Based on Indirect Predictions: Proof of Concept on Wheat and Poplar. *G3 Genes|Genomes|Genetics* 8, 3961–3972. doi: 10.1534/g3.118.200760.

Rutkoski, J., Singh, R. p., Huerta-Espino, J., Bhavani, S., Poland, J., Jannink, J. l., et al. (2015). Genetic Gain from Phenotypic and Genomic Selection for Quantitative Resistance to Stem Rust of Wheat. *The Plant Genome* 8, plantgenome2014.10.0074. doi: 10.3835/plantgenome2014.10.0074.

Schiessl, S. (2020). Regulation and Subfunctionalization of Flowering Time Genes in the Allotetraploid Oil Crop Brassica napus. *Frontiers in Plant Science* 11. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.605155 [Accessed December 13, 2023].

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. *The Crop Journal* 7, 127–140. doi: 10.1016/j.cj.2018.07.006.

Schmidt, P., Hartung, J., Bennewitz, J., and Piepho, H.-P. (2019). Heritability in Plant Breeding on a Genotype-Difference Basis. *Genetics* 212, 991–1008. doi: 10.1534/genetics.119.302134.

Schmitz, R. J., Schultz, M. D., Lewsey, M. G., O'Malley, R. C., Urich, M. A., Libiger, O., et al. (2011). Transgenerational Epigenetic Instability Is a Source of Novel Methylation Variants. *Science* 334, 369–373. doi: 10.1126/science.1212959.

Schrag, T. A., Maurer, H. P., Melchinger, A. E., Piepho, H.-P., Peleman, J., and Frisch, M. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theor Appl Genet* 114, 1345–1355. doi: 10.1007/s00122-007-0521-5.

Schrag, T. A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., et al. (2018). Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. *Genetics* 208, 1373–1385. doi: 10.1534/genetics.117.300374.

Shaw, P. D., Graham, M., Kennedy, J., Milne, I., and Marshall, D. F. (2014). Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics* 15, 259. doi: 10.1186/1471-2105-15-259.

Shelar, M. N., Matsagar, V. K., Patil, V. S., and Barahate, S. D. (2023). Net energy analysis of sugarcane based ethanol production. *Cleaner Energy Systems* 4, 100059. doi: 10.1016/j.cles.2023.100059.

Sheridan, W. F., and Auger, D. L. (2006). Construction and Uses of New Compound B-A-A Maize Chromosome Translocations. *Genetics* 174, 1755–1765. doi: 10.1534/genetics.106.065540.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. E. (2008). Genomic selection using different marker types and densities. *Journal of Animal Science* 86, 2447–2454. doi: 10.2527/jas.2007-0010.

Soller, M., and Beckmann, J. S. (1983). Genetic polymorphism in varietal identification and genetic improvement. *Theoret. Appl. Genetics* 67, 25–33. doi: 10.1007/BF00303917.

Sonesson, A. K., Woolliams, J. A., and Meuwissen, T. H. (2012). Genomic selection requires genomic control of inbreeding. *Genetics Selection Evolution* 44, 27. doi: 10.1186/1297-9686-44-27.

Song, H., Ye, S., Jiang, Y., Zhang, Z., Zhang, Q., and Ding, X. (2019). Using imputation-based whole-genome sequencing data to improve the accuracy of genomic prediction for combined populations in pigs. *Genetics Selection Evolution* 51, 58. doi: 10.1186/s12711-019-0500-8.

Song, Q., Hyten, D. L., Jia, G., Quigley, C. V., Fickus, E. W., Nelson, R. L., et al. (2013). Development and Evaluation of SoySNP50K, a High-Density Genotyping Array for Soybean. *PLOS ONE* 8, e54985. doi: 10.1371/journal.pone.0054985.

Sourdille, P., and Jenczewski, E. (2021). Homoeologous exchanges in allopolyploids: how Brassica napus established self-control! *New Phytologist* 229, 3041–3043. doi: 10.1111/nph.17222.

Spindel, J. E., Begum, H., Akdemir, D., Collard, B., Redoña, E., Jannink, J.-L., et al. (2016). Genome-wide prediction models that incorporate de novo GWAS are a powerful new tool for tropical rice improvement. *Heredity* 116, 395–408. doi: 10.1038/hdy.2015.113.

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.

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. *Frontiers in Neuroscience* 14. Available at: https://www.frontiersin.org/articles/10.3389/fnins.2020.00047 [Accessed February 16, 2023].

Tian, H.-L., Wang, F.-G., Zhao, J.-R., Yi, H.-M., Wang, L., Wang, R., et al. (2015). Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. *Mol Breeding* 35, 136. doi: 10.1007/s11032-015-0335-0.

Tsai, H.-Y., Matika, O., Edwards, S. M., Antolín–Sánchez, R., Hamilton, A., Guy, D. R., et al. (2017). Genotype Imputation To Improve the Cost-Efficiency of Genomic Selection in Farmed Atlantic Salmon. *G3 (Bethesda)* 7, 1377–1383. doi: 10.1534/g3.117.040717.

Unterseer, S., Bauer, E., Haberer, G., Seidel, M., Knaak, C., Ouzunova, M., et al. (2014). A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array. *BMC Genomics* 15, 823. doi: 10.1186/1471-2164-15-823.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91, 4414–4423. doi: 10.3168/jds.2007-0980.

Visković, J., Zheljazkov, V. D., Sikora, V., Noller, J., Latković, D., Ocamb, C. M., et al. (2023). Industrial Hemp (Cannabis sativa L.) Agronomy and Utilization: A Review. *Agronomy* 13, 931. doi: 10.3390/agronomy13030931.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., et al. (2006). Assumption-Free Estimation of Heritability from Genome-Wide Identity-by-Descent Sharing between Full Siblings. *PLOS Genetics* 2, e41. doi: 10.1371/journal.pgen.0020041.

Vitezica, Z. G., Legarra, A., Toro, M. A., and Varona, L. (2017). Orthogonal Estimates of Variances for Additive, Dominance, and Epistatic Effects in Populations. *Genetics* 206, 1297–1307. doi: 10.1534/genetics.116.199406.

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. *Frontiers in Plant Science* 12. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2021.749491 [Accessed January 5, 2023].

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.

Voss-Fels, K. P., Stahl, A., and Hickey, L. T. (2019). Q&A: modern crop breeding for future food security. *BMC Biology* 17, 18. doi: 10.1186/s12915-019-0638-4.

Voss-Fels, K. P., Wei, X., Ross, E. M., Frisch, M., Aitken, K. S., Cooper, M., et al. (2021). Strategies and considerations for implementing genomic selection to improve traits with additive and non-additive genetic architectures in sugarcane breeding. *Theor Appl Genet* 134, 1493–1511. doi: 10.1007/s00122-021-03785-3.

Wang, W., Chen, L., Wang, X., Duan, J., Flynn, R. D., Wang, Y., et al. (2021). A transposon-mediated reciprocal translocation promotes environmental adaptation but compromises domesticability of wild soybeans. *New Phytologist* 232, 1765–1777. doi: 10.1111/nph.17671.

Weber, S. E., Chawla, H. S., Ehrig, L., Hickey, L. T., Frisch, M., and Snowdon, R. J. (2023a). Accurate prediction of quantitative traits with failed SNP calls in canola and maize. *Frontiers in Plant Science* 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1221750 [Accessed November 10, 2023].

Weber, S. E., Frisch, M., Snowdon, R. J., and Voss-Fels, K. P. (2023b). Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Frontiers in Plant Science* 14. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1217589 [Accessed November 10, 2023].

Weber, S. E., Roscher-Ehrig, L., Kox, T., Abbadi, A., Stahl, A. and Snowdon, R. J. (2023). Genomic Prediction in Brassica napus: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data. under Review in Genome

Werner, C. R., Gaynor, R. C., Gorjanc, G., Hickey, J. M., Kox, T., Abbadi, A., et al. (2020). How Population Structure Impacts Genomic Selection Accuracy in Cross-Validation: Implications for Practical Breeding. *Frontiers in Plant Science* 11. Available at: https://www.frontiersin.org/articles/10.3389/fpls.2020.592977 [Accessed December 22, 2022].

Werner, C. R., Qian, L., Voss-Fels, K. P., Abbadi, A., Leckband, G., Frisch, M., et al. (2018). 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.

Westhues, M., Schrag, T. A., Heuer, C., Thaller, G., Utz, H. F., Schipprack, W., et al. (2017). Omics-based hybrid prediction in maize. *Theor Appl Genet* 130, 1927–1939. doi: 10.1007/s00122-017-2934-0.

Whittaker, J. C., Thompson, R., and Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetics Research* 75, 249–252. doi: 10.1017/S0016672399004462.

Wientjes, Y. C. J., Bijma, P., Calus, M. P. L., Zwaan, B. J., Vitezica, Z. G., and van den Heuvel, J. (2022). The long-term effects of genomic selection: 1. Response to selection, additive genetic variance, and genetic architecture. *Genetics Selection Evolution* 54, 19. doi: 10.1186/s12711-022-00709-7.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., et al. (2011). Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. *Genetics Selection Evolution* 43, 5. doi: 10.1186/1297-9686-43-5.

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.

Wu, P.-Y., Ou, J.-H., and Liao, C.-T. (2023). Sample size determination for training set optimization in genomic prediction. *Theor Appl Genet* 136, 57. doi: 10.1007/s00122-023-04254-9.

Xu, Y., and Crouch, J. H. (2008). Marker-Assisted Selection in Plant Breeding: From Publications to Practice. *Crop Science* 48, 391–407. doi: 10.2135/cropsci2007.04.0191.

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.

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., et al. (2011). Gains in QTL Detection Using an Ultra-High Density SNP Map Based on Population Sequencing Relative to Traditional RFLP/SSR Markers. *PLOS ONE* 6, e17595. doi: 10.1371/journal.pone.0017595.

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, Z., and Druet, T. (2010). Marker imputation with low-density marker panels in Dutch Holstein cattle. *Journal of Dairy Science* 93, 5487–5494. doi: 10.3168/jds.2010-3501.

Zhang, Z., Gou, X., Xun, H., Bian, Y., Ma, X., Li, J., et al. (2020). Homoeologous exchanges occur through intragenic recombination generating novel transcripts and proteins in wheat and other polyploids. *Proceedings of the National Academy of Sciences* 117, 14561–14571. doi: 10.1073/pnas.2003505117.

Zhao, Y., Zeng, J., Fernando, R., and Reif, J. C. (2013). Genomic Prediction of Hybrid Wheat Performance. *Crop Science* 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.

Appendix

Appendix I: Supplementary material from

Weber, S. E., Frisch, M., Snowdon, R. J. and Voss-Fels, K.-P. (2023). Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. Front. Plant Sci. 14:1217589. doi: 10.3389/fpls.2023.1217589

frontiers

Supplementary Material

Haplotype blocks for genomic prediction: A comparative evaluation in multiple crop datasets

Sven E. Weber[1*], Matthias Frisch[2], Rod J. Snowdon[1], Kai P. Voss-Fels[3]

*** Correspondence:**

Sven E. Weber
Sven.E.Weber@agrar.uni-giessen.de

1.1 Supplementary Figures

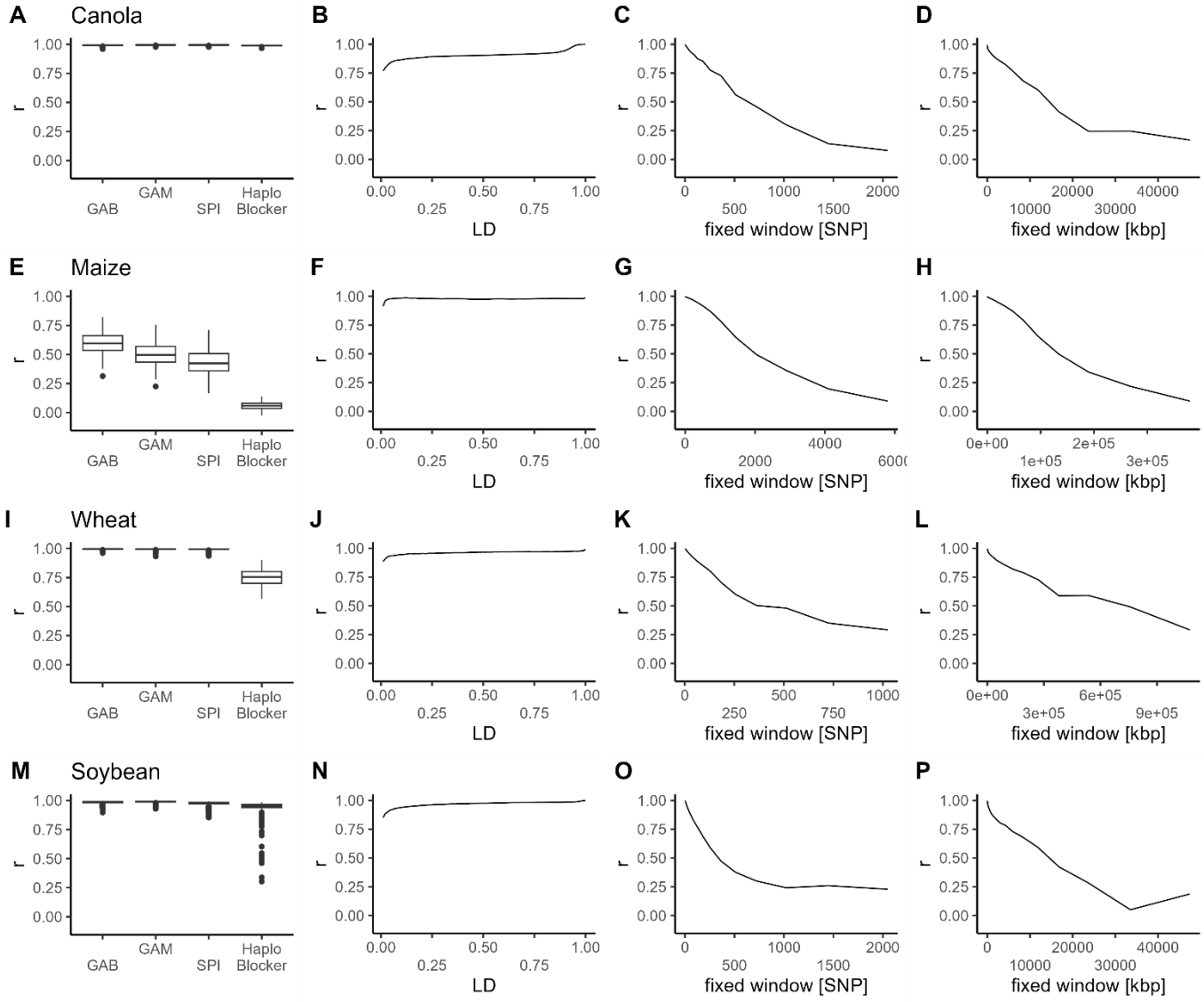

Figure S1 Mean Correlation (r) between SNP and haplotype based genomic relationship coefficients identified by the methods implemented in "Haploview" and "HaploBlocker" (**A**, **E**, **I**, **M**), LD (**B**, **F**, **J**, **N**), fixed window of adjacent base pairs (**C**, **G**, **K**, **O**) and fixed window of adjacent markers (**D**, **H**, **L**, **P**) based haplotype blocks, in canola (**A**, **B**, **C**, **D**), maize (**E**, **F**, **G**, **H**), wheat (**I**, **J**, **K**, **L**) and soybean (**M**, **N**, **O**, **P**)

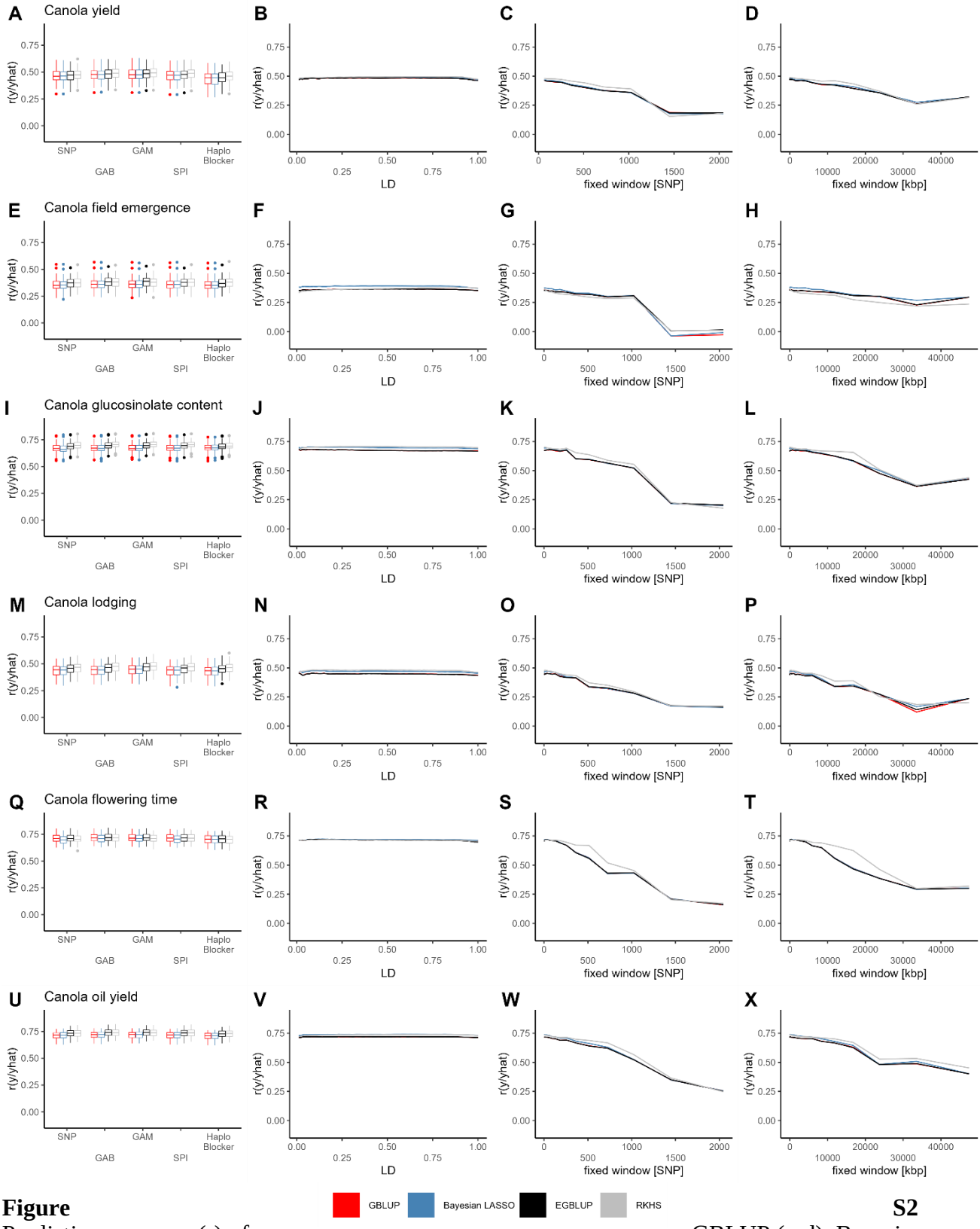

Figure S2
Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs,"Haploview" and "HaploBlocker" (**A**, **E**, **I**, **M**, **Q**, **U**), LD (**B**, **F**, **J**, **N**, **R**, **V**), fixed window of adjacent base pairs (**C**, **G**, **K**, **O**, **S**, **W**) and fixed window of adjacent markers (**D**, **H**, **L**, **P**, **T**, **X**) based haplotype blocks, in canola: seed yield (**A**, **B**, **C**, **D**), field emergence (**E**, **F**, **G**, **H**), glucosinolate content (**I**, **J**, **K** ,**L**),lodging (**M**, **N**, **O**, **P**), flowering time (**Q**, **R**, **S**, **T**), oil yield (**U**, **V**, **W**, **X**). Individual points in the lines represent the mean over all cross validation runs for each haplotype block parameter and model combination

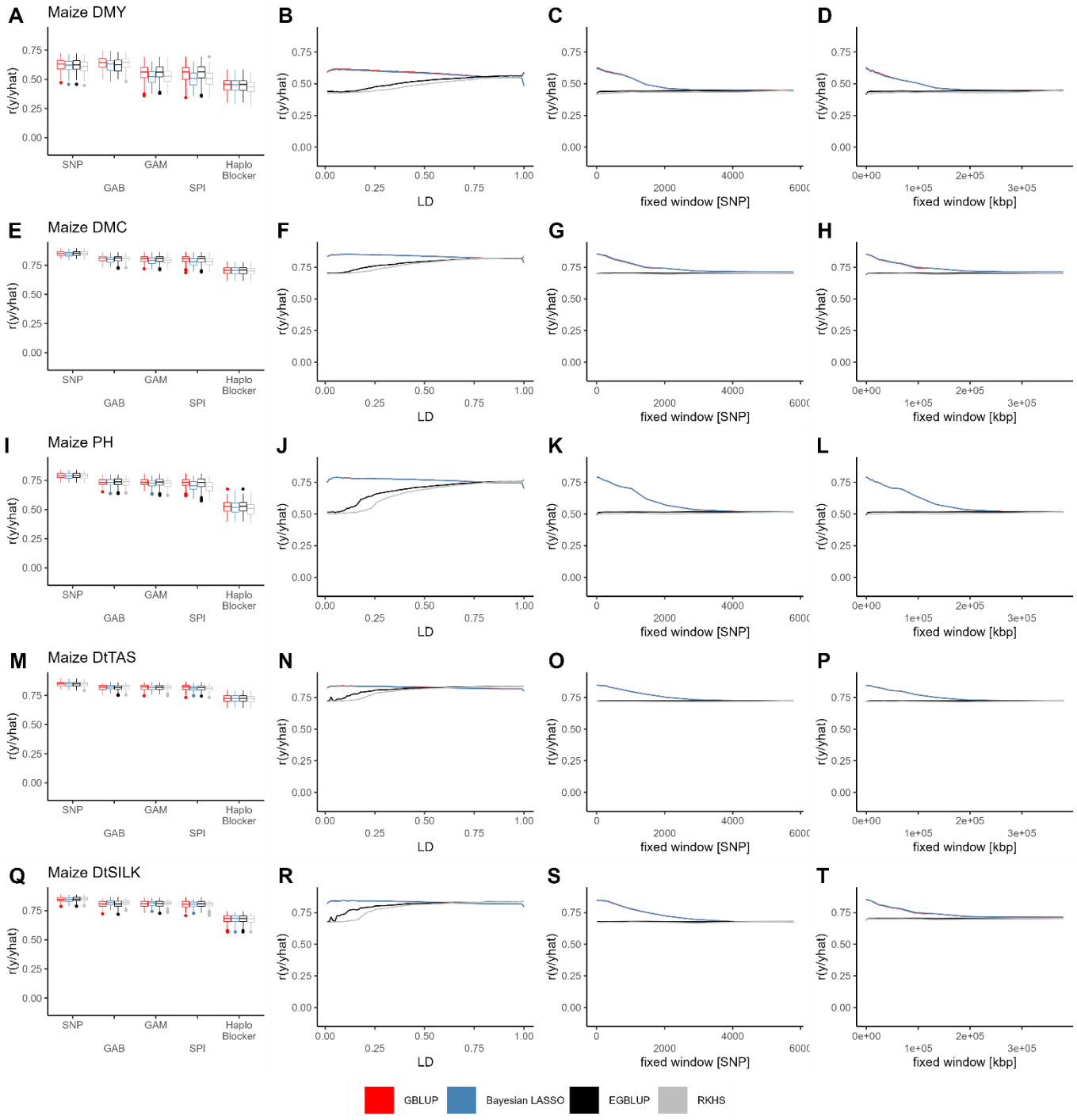

Figure S3 Prediction accuracy (r) (random cross validation) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E, I, M, Q**), LD (**B, F, J, N, R**), fixed window of adjacent base pairs (**C, G, K, O, S**) and fixed window of adjacent markers (**D, H, L, P, T**) based haplotype blocks, in maize: DMY (**A, B, C, D**), DMC (**E, F, G, H**), PH (**I, J, K, L**), DtTAS (**M, N, O, P**), DtSILK (**Q, R, S, T**). Individual points in the lines represent the mean over all cross validation runs for each haplotype block parameter and model combination

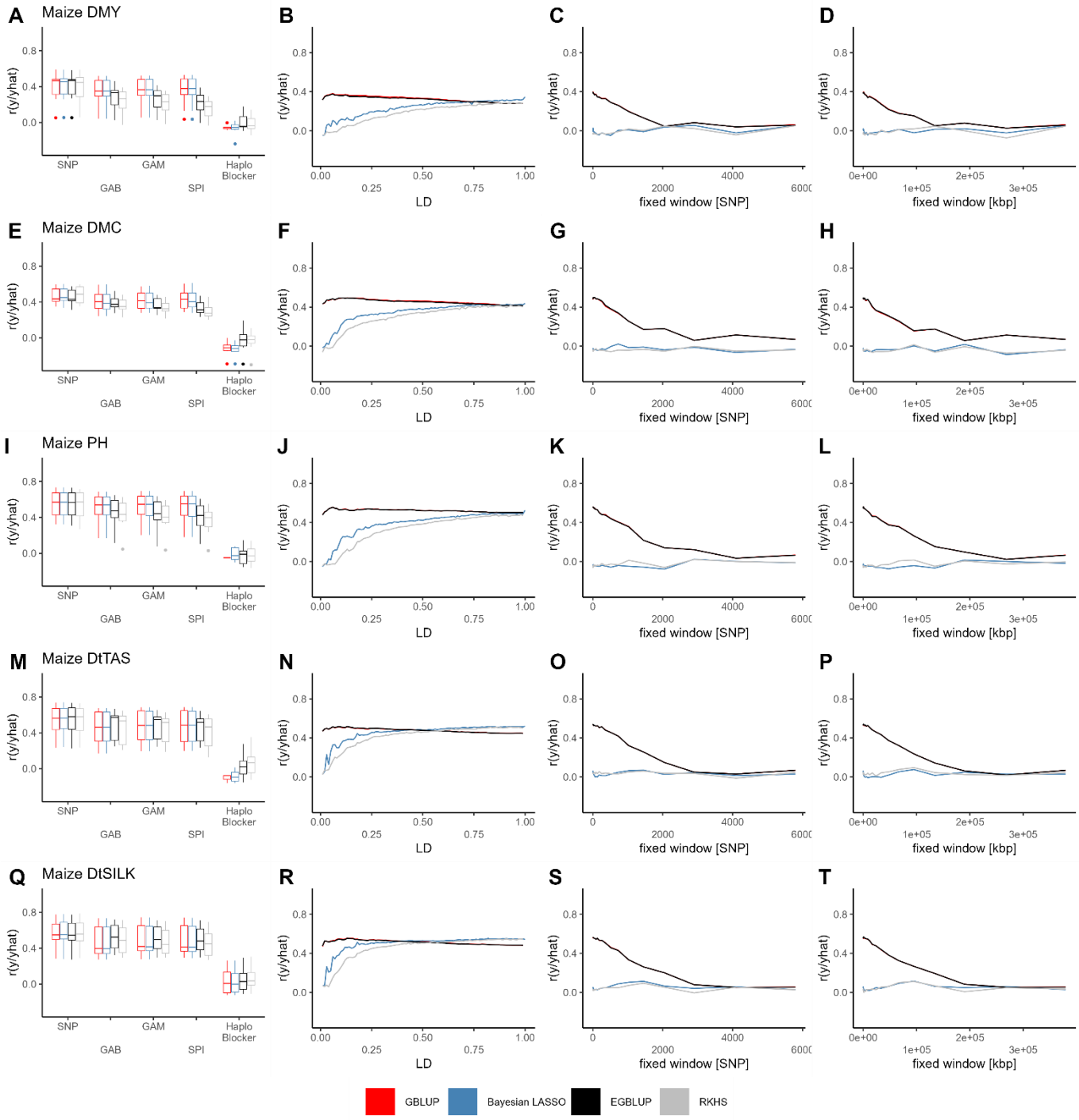

Figure S4 Prediction accuracy (r) (family-wise cross validation) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E, I, M, Q**), LD (**B, F, J, N, R**), fixed window of adjacent base pairs (**C, G, K, O, S**) and fixed window of adjacent markers (**D, H, L, P, T**) based haplotype blocks, in maize: DMY (**A, B, C, D**), DMC (**E, F, G, H**), PH (**I, J, K, L**), DtTAS (**M, N, O, P**), DtSILK (**Q, R, S, T**). Individual points in the lines represent the mean over all cross validation runs for each haplotype block parameter and model combination

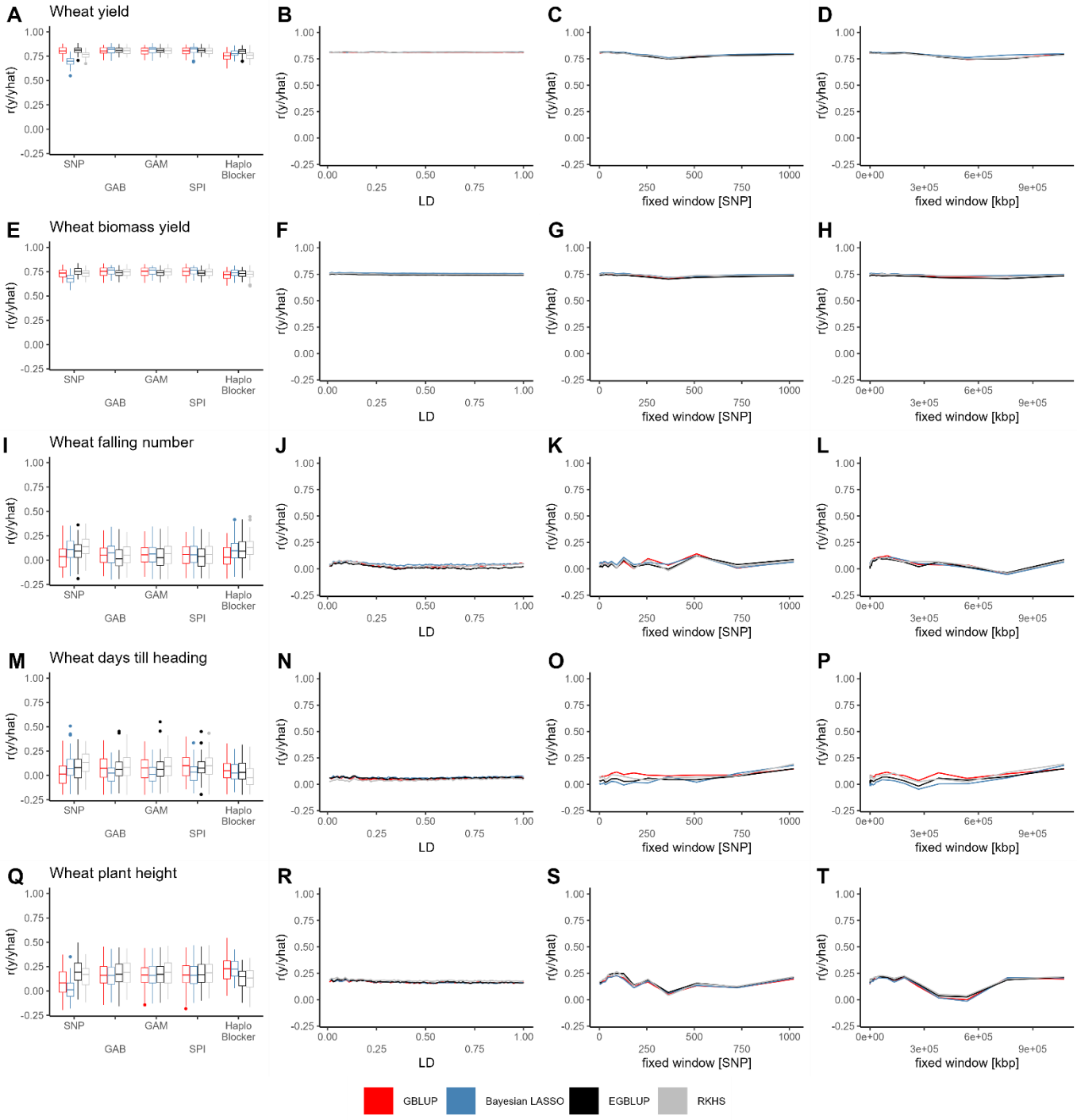

Figure S5 Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E, I, M, Q**), LD (**B, F, J, N, R**), fixed window of adjacent base pairs (**C, G, K, O, S**) and fixed window of adjacent markers (**D, H, L, P, T**) based haplotype blocks, in wheat: seed yield (**A, B, C, D**), biomass yield (**E, F, G, H**), falling number (**I, J, K, L**), days till heading (**M, N, O, P**), plant height (**Q, R, S, T**). Individual points in the lines represent the mean over all cross validation runs for each haplotype block parameter and model combination

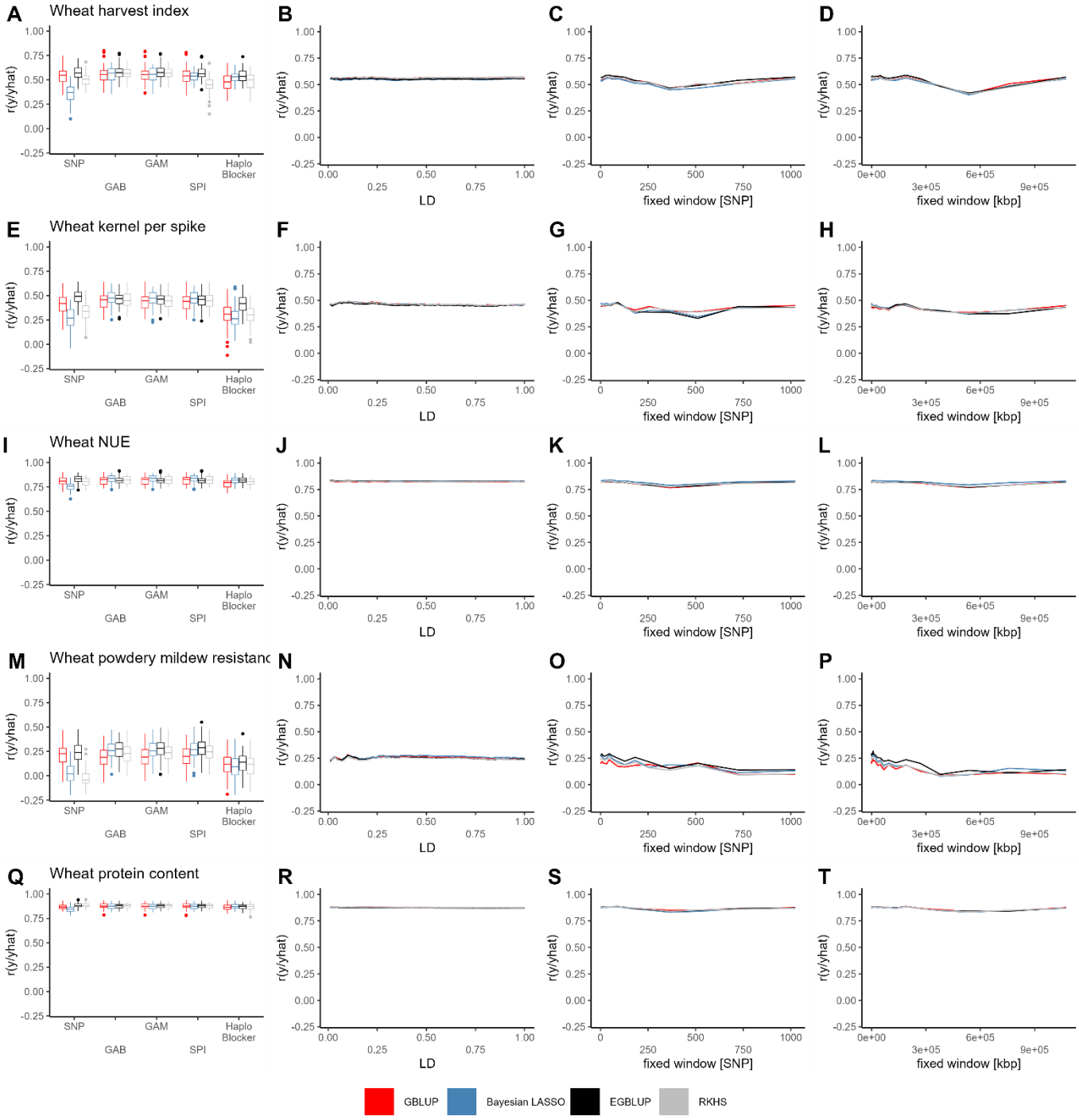

Figure S6 Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E, I, M, Q**), LD (**B, F, J, N, R**), fixed window of adjacent base pairs (**C, G, K, O, S**) and fixed window of adjacent markers (**D, H, L, P, T**) based haplotype blocks, in wheat: harvest index (**A, B, C, D**), kernel spike^{-1} (**E, F, G, H**), NUE (**I, J, K, L**), powdery mildew resistance (**M, N, O, P**), protein content (**Q, R, S, T**). Individual points in the lines represent the mean over all cross validation runs for each haplotype block parameter and model combination

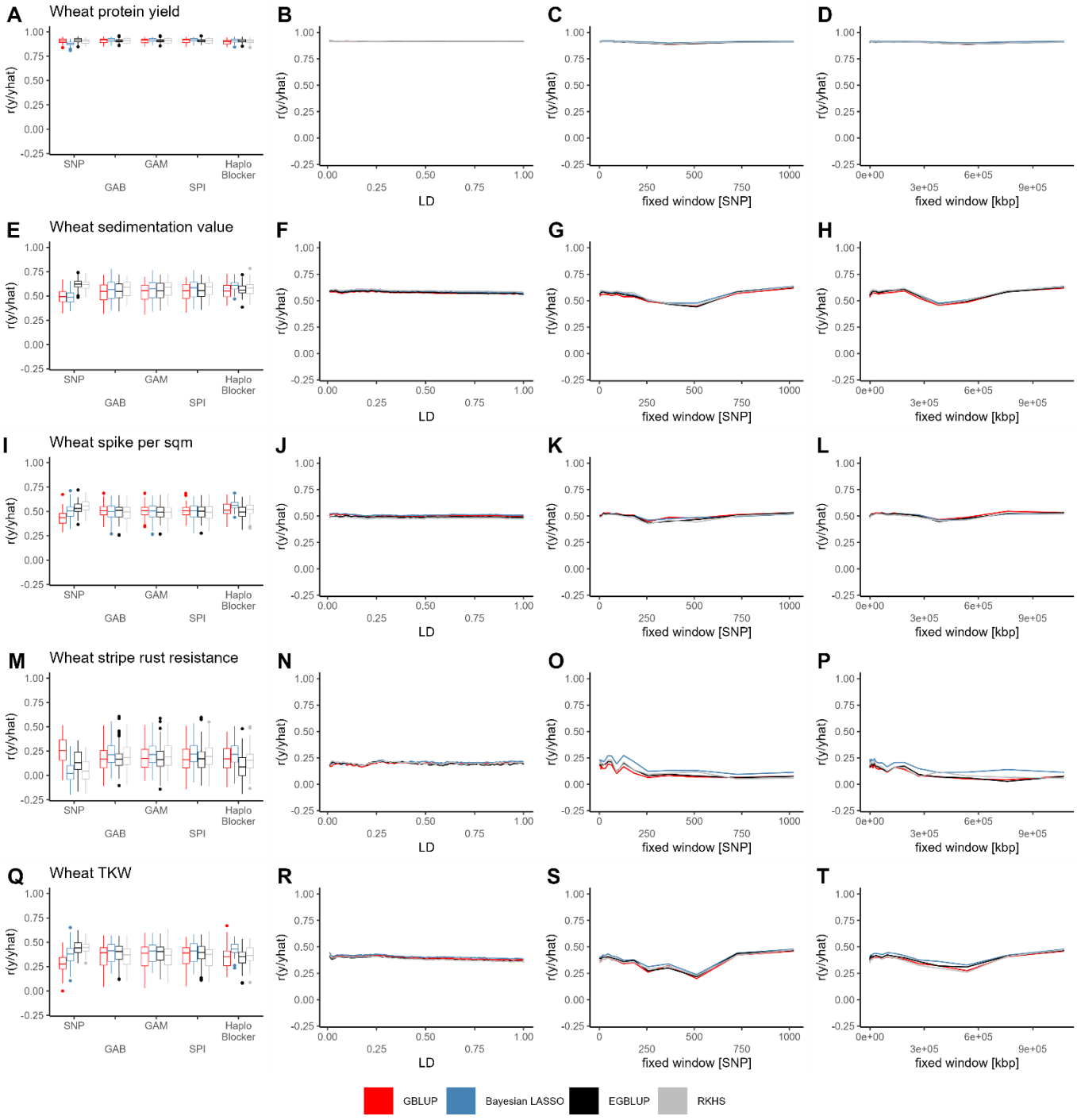

Figure S6 Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E, I, M, Q**), LD (**B, F, J, N, R**), fixed window of adjacent base pairs (**C, G, K, O, S**) and fixed window of adjacent markers (**D, H, L, P, T**) based haplotype blocks, in wheat: protein yield (**A, B, C, D**), sedimentation value (**E, F, G, H**), spike m^{-2} (**I, J, K, L**), stripe rust resistance (**M, N, O, P**), TKW (**Q, R, S, T**). Individual points in the lines represent the mean over all cross validation run for each haplotype block parameter and model combination

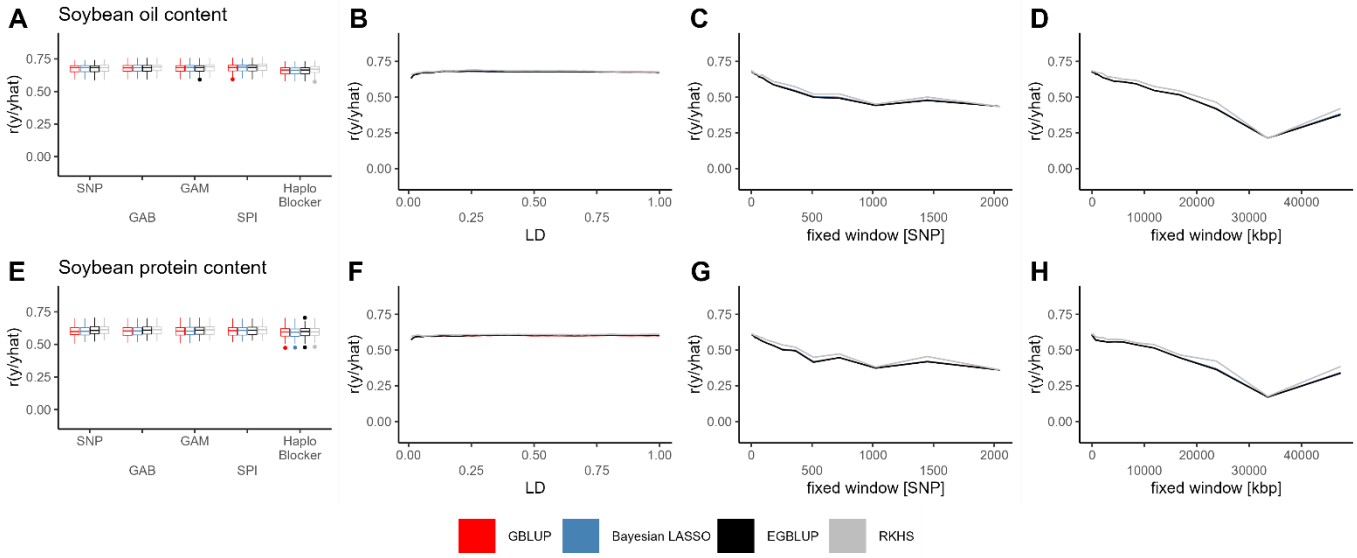

Figure S7 Prediction accuracy (r) of GBLUP (red), Bayesian LASSO (blue), EGBLUP (black) and RKHS (grey) with SNPs, *"Haploview"* and *"HaploBlocker"* (**A, E**), LD (**B, F**), fixed window of adjacent base pairs (**C, G**) and fixed window of adjacent markers (**D, H**) based haplotype blocks, in soybean: oil content (**A, B, C, D**), protein content (**E, F, G, H**). Individual points in the lines represent the mean over all cross validation run for each haplotype block parameter and model combination

Table S1 Average prediction accuracy of SNP based prediction compared to the best haplotyping method of Canola, Maize, Wheat and Soybean for all traits

Dataset	Trait	Model	SNP prediction accuracy	Best haplotyping algorithm	Prediction accuracy by best haplotyping algorithm	Improvement by best haplotyping algorithm
Canola	yield	GBLUP	0.464	LD = 0.6	0.485	0.021
		Bayesian LASSO	0.462	LD = 0.59	0.486	0.024
		EGBLUP	0.471	LD = 0.6	0.492	0.021
		RKHS	0.474	LD = 0.71	0.496	0.022
	field emergence	GBLUP	0.354	LD = 0.43	0.367	0.013
		Bayesian LASSO	0.352	LD = 0.43	0.367	0.014
		EGBLUP	0.371	LD = 0.52	0.393	0.022
		RKHS	0.373	LD = 0.89	0.379	0.006
	glucosinolate content	GBLUP	0.669	LD = 0.08	0.681	0.013
		Bayesian LASSO	0.665	LD = 0.01	0.683	0.018
		EGBLUP	0.687	LD = 0.08	0.704	0.017
		RKHS	0.696	LD = 0.24	0.707	0.011
	lodging	GBLUP	0.436	LD = 0.01	0.456	0.02
		Bayesian LASSO	0.434	LD = 0.09	0.454	0.02
		EGBLUP	0.453	LD = 0.09	0.476	0.023
		RKHS	0.462	LD = 0.16	0.482	0.02
	flowering time	GBLUP	0.709	LD = 0.12	0.721	0.012
		Bayesian LASSO	0.697	LD = 0.15	0.721	0.024
		EGBLUP	0.711	LD = 0.14	0.723	0.012
		RKHS	0.704	nKB = 2097.15	0.719	0.015
	oil yield	GBLUP	0.713	LD = 0.23	0.72	0.007
		Bayesian LASSO	0.713	LD = 0.23	0.721	0.008
		EGBLUP	0.732	LD = 0.59	0.74	0.008
		RKHS	0.731	LD = 0.82	0.738	0.007
Maize	DMY	GBLUP	0.624	GAB	0.635	0.011
		Bayesian LASSO	0.616	GAB	0.621	0.006
		EGBLUP	0.62	nSNP = 8	0.622	0.002
		RKHS	0.608	GAB	0.631	0.023
	DMC	GBLUP	0.852	nSNP = 32	0.854	0.002
		Bayesian LASSO	0.845	LD = 1	0.835	-0.011
		EGBLUP	0.852	nSNP = 32	0.854	0.002
		RKHS	0.851	LD = 1	0.84	-0.011
	PH	GBLUP	0.791	nKB = 131.07	0.793	0.002
		Bayesian LASSO	0.786	LD = 1	0.771	-0.015
		EGBLUP	0.791	nKB = 131.07	0.793	0.002
		RKHS	0.786	LD = 1	0.77	-0.016
	DtTAS	GBLUP	0.847	nSNP = 4	0.847	0.000
		Bayesian LASSO	0.846	LD = 1	0.842	-0.004
		EGBLUP	0.845	nSNP = 4	0.846	0.000

Table S1 Average prediction accuracy of SNP based prediction compared to the best haplotyping method of Canola, Maize, Wheat and Soybean for all traits

Dataset	Trait	Model	SNP prediction accuracy	Best haplotyping algorithm	Prediction accuracy by best haplotyping algorithm	Improvement by best haplotyping algorithm
	DtSILK	RKHS	0.846	LD = 1	0.842	-0.003
		GBLUP	0.847	nSNP = 16	0.848	0.002
		Bayesian LASSO	0.845	LD = 1	0.841	-0.004
		EGBLUP	0.848	nSNP = 16	0.849	0.001
		RKHS	0.849	LD = 1	0.843	-0.006
Wheat	yield	GBLUP	0.805	LD = 0.23	0.813	0.008
		Bayesian LASSO	0.697	nSNP = 46	0.818	0.122
		EGBLUP	0.811	LD = 0.1	0.815	0.005
		RKHS	0.765	LD = 0.1	0.814	0.049
	biomass yield	GBLUP	0.732	LD = 0.02	0.759	0.027
		Bayesian LASSO	0.678	LD = 0.02	0.766	0.089
		EGBLUP	0.752	LD = 0.06	0.752	0.001
		RKHS	0.729	LD = 0.02	0.761	0.032
	falling number	GBLUP	0.009	nSNP = 512	0.143	0.134
		Bayesian LASSO	0.102	nSNP = 512	0.124	0.022
		EGBLUP	0.096	nSNP = 512	0.123	0.027
		RKHS	0.141	HaploBlocker	0.123	-0.018
	days till heading	GBLUP	-0.025	nKB = 1073741.82	0.145	0.17
		Bayesian LASSO	0.071	nKB = 1073741.82	0.18	0.109
		EGBLUP	0.077	nKB = 1073741.82	0.147	0.07
		RKHS	0.129	nKB = 1073741.82	0.192	0.063
	plant height	GBLUP	0.092	nSNP = 91	0.226	0.134
		Bayesian LASSO	0.021	nSNP = 91	0.23	0.208
		EGBLUP	0.188	nSNP = 91	0.249	0.06
		RKHS	0.15	nSNP = 91	0.268	0.119
	harvest index	GBLUP	0.538	nSNP = 46	0.578	0.04
		Bayesian LASSO	0.364	nKB = 47453.13	0.563	0.199
		EGBLUP	0.571	nKB = 189812.53	0.589	0.019
		RKHS	0.503	nSNP = 46	0.578	0.075
	kernel per spike	GBLUP	0.414	LD = 0.08	0.488	0.074
		Bayesian LASSO	0.266	LD = 0.1	0.488	0.222
		EGBLUP	0.482	LD = 0.08	0.48	-0.003
		RKHS	0.331	LD = 0.1	0.493	0.162
	NUE	GBLUP	0.812	LD = 0.01	0.83	0.018
		Bayesian LASSO	0.755	nSNP = 46	0.841	0.085
		EGBLUP	0.83	LD = 0.01	0.836	0.006
		RKHS	0.802	LD = 0.01	0.832	0.03
	powdery mildew	GBLUP	0.214	LD = 0.1	0.283	0.069
		Bayesian LASSO	0.009	nKB = 8388.61	0.286	0.277

Table S1 Average prediction accuracy of SNP based prediction compared to the best haplotyping method of Canola, Maize, Wheat and Soybean for all traits

Dataset	Trait	Model	SNP prediction accuracy	Best haplotyping algorithm	Prediction accuracy by best haplotyping algorithm	Improvement by best haplotyping algorithm
	resistance	EGBLUP	0.243	nKB = 8388.61	0.316	0.073
		RKHS	-0.029	nKB = 8388.61	0.262	0.291
	protein content	GBLUP	0.867	nKB = 189812.53	0.888	0.021
		Bayesian LASSO	0.843	nKB = 189812.53	0.885	0.042
		EGBLUP	0.881	nSNP = 91	0.888	0.006
		RKHS	0.884	nSNP = 91	0.887	0.003
	protein yield	GBLUP	0.907	LD = 0.01	0.915	0.008
		Bayesian LASSO	0.88	nSNP = 46	0.921	0.041
		EGBLUP	0.915	LD = 0.01	0.918	0.003
		RKHS	0.901	LD = 0.01	0.916	0.014
	sedimentation value	GBLUP	0.493	nKB = 1073741.82	0.619	0.126
		Bayesian LASSO	0.488	nKB = 1073741.83	0.636	0.148
		EGBLUP	0.62	nKB = 1073741.84	0.627	0.006
		RKHS	0.61	nKB = 1073741.85	0.631	0.021
	spike per sqm	GBLUP	0.432	nKB = 759250.13	0.546	0.114
		Bayesian LASSO	0.503	HaploBlocker	0.564	0.061
		EGBLUP	0.536	nKB = 1073741.82	0.532	-0.003
		RKHS	0.556	nKB = 759250.13	0.531	-0.025
	stripe rust resistance	GBLUP	0.254	LD = 0.28	0.215	-0.039
		Bayesian LASSO	0.011	nSNP = 128	0.274	0.263
		EGBLUP	0.134	LD = 0.27	0.219	0.085
		RKHS	0.047	nSNP = 46	0.229	0.183
	TKW	GBLUP	0.281	nKB = 1073741.82	0.461	0.179
		Bayesian LASSO	0.378	nKB = 1073741.83	0.478	0.1
		EGBLUP	0.442	nKB = 1073741.84	0.467	0.025
		RKHS	0.446	nKB = 1073741.85	0.473	0.028
Soybean	oil content	GBLUP	0.674	LD = 0.24	0.682	0.008
		Bayesian LASSO	0.675	LD = 0.24	0.683	0.008
		EGBLUP	0.674	LD = 0.24	0.682	0.008
		RKHS	0.677	LD = 0.26	0.691	0.014
	protein content	GBLUP	0.601	nSNP = 4	0.606	0.006
		Bayesian LASSO	0.602	nSNP = 4	0.608	0.006
		EGBLUP	0.609	nSNP = 4	0.611	0.003
		RKHS	0.609	nSNP = 4	0.613	0.003

Table S2 Average prediction accuracy of SNP based prediction compared to the best haplotyping method of Maize traits in the family-wise cross validation

Dataset	Trait	Model	SNP prediction accuracy	Best haplotyping algorithm	Prediction accuracy by best haplotyping algorithm	Improvement by best haplotyping algorithm
Maize	DMY	GBLUP	0.394	nSNP = 4	0.397	0.003
		Bayesian LASSO	0.396	LD = 1	0.348	-0.049
		EGBLUP	0.392	nSNP = 4	0.393	0.002
		RKHS	0.39	LD = 0.94	0.29	-0.101
	DMC	GBLUP	0.465	nSNP = 46	0.5	0.035
		Bayesian LASSO	0.458	LD = 1	0.437	-0.021
		EGBLUP	0.47	nSNP = 46	0.5	0.03
		RKHS	0.481	LD = 0.93	0.421	-0.06
	PH	GBLUP	0.548	nKB = 92.68	0.561	0.012
		Bayesian LASSO	0.544	LD = 1	0.526	-0.018
		EGBLUP	0.548	nKB = 92.68	0.561	0.012
		RKHS	0.539	LD = 0.98	0.48	-0.059
	DtTAS	GBLUP	0.533	nKB = 131.07	0.538	0.005
		Bayesian LASSO	0.537	LD = 1	0.518	-0.019
		EGBLUP	0.538	nKB = 131.07	0.541	0.003
		RKHS	0.536	LD = 0.93	0.512	-0.024
	DtSILK	GBLUP	0.553	nKB = 524.29	0.566	0.014
		Bayesian LASSO	0.551	LD = 0.82	0.554	0.002
		EGBLUP	0.559	nKB = 262.14	0.568	0.009
		RKHS	0.554	LD = 0.97	0.549	-0.006

Appendix II: Supplementary material from

Weber, S. E., Chawla, H. S., Ehrig, L., Hickey, L. T., Frisch, M. and Snowdon, R. J. (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

frontiers

Supplementary Material

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], Rod J. Snowdon[1]

*** Correspondence: Sven E. Weber**: Sven.E.Weber@agrar.uni-giessen.de

Supplementary Figures

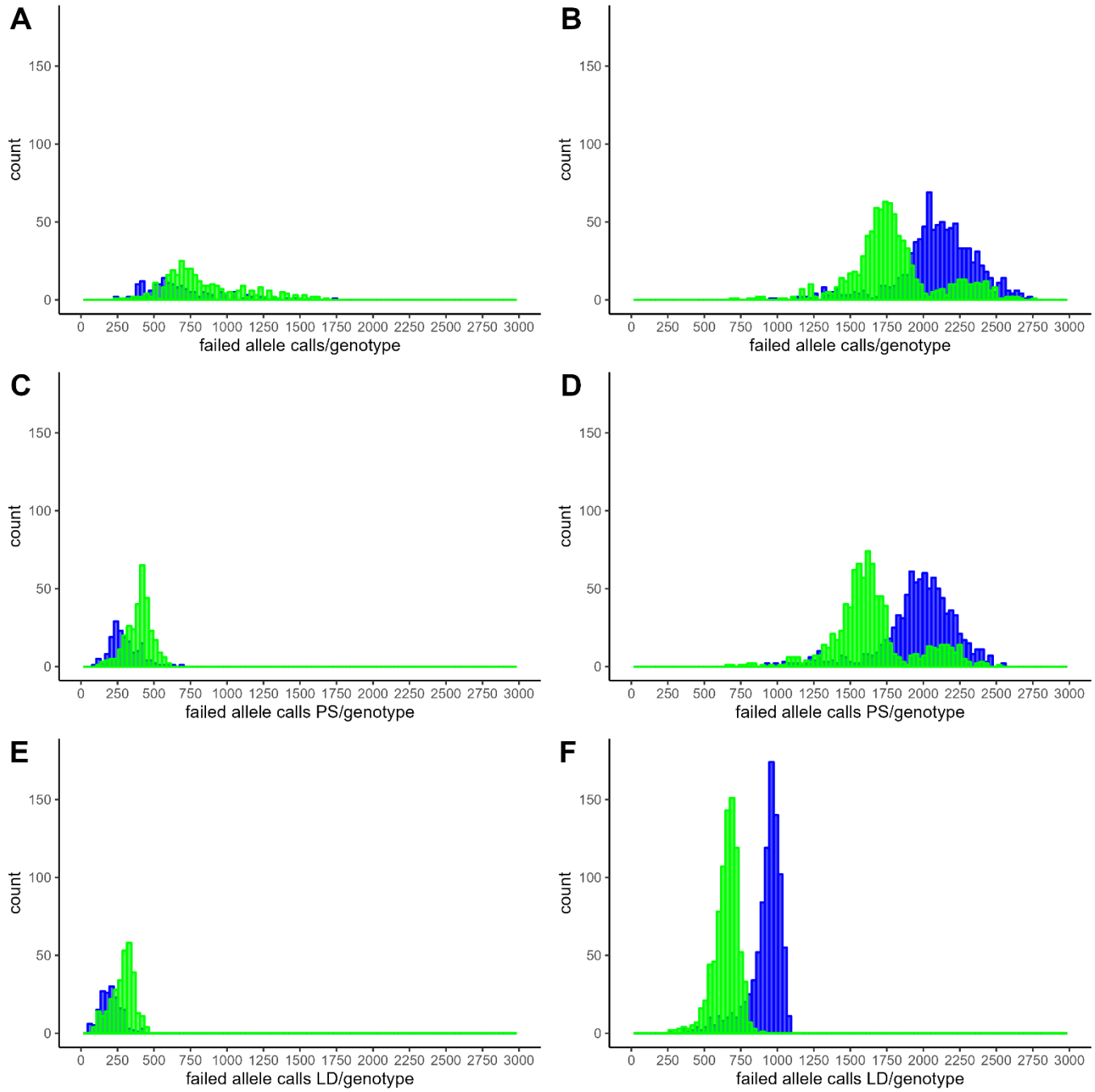

Figure S1 Histograms of number of failed allele calls per genotype (**A**, **B**) failed allele calls filtered by pool specificity (**C**, **D**) and failed allele calls filtered by LD (**E**, **F**) in canola (**A**, **C**, **E**) and maize (**B**, **D**, **F**). In canola, the color blue represents pool A and lime pool B. In maize, blue represents the flint pool and lime the dent pool

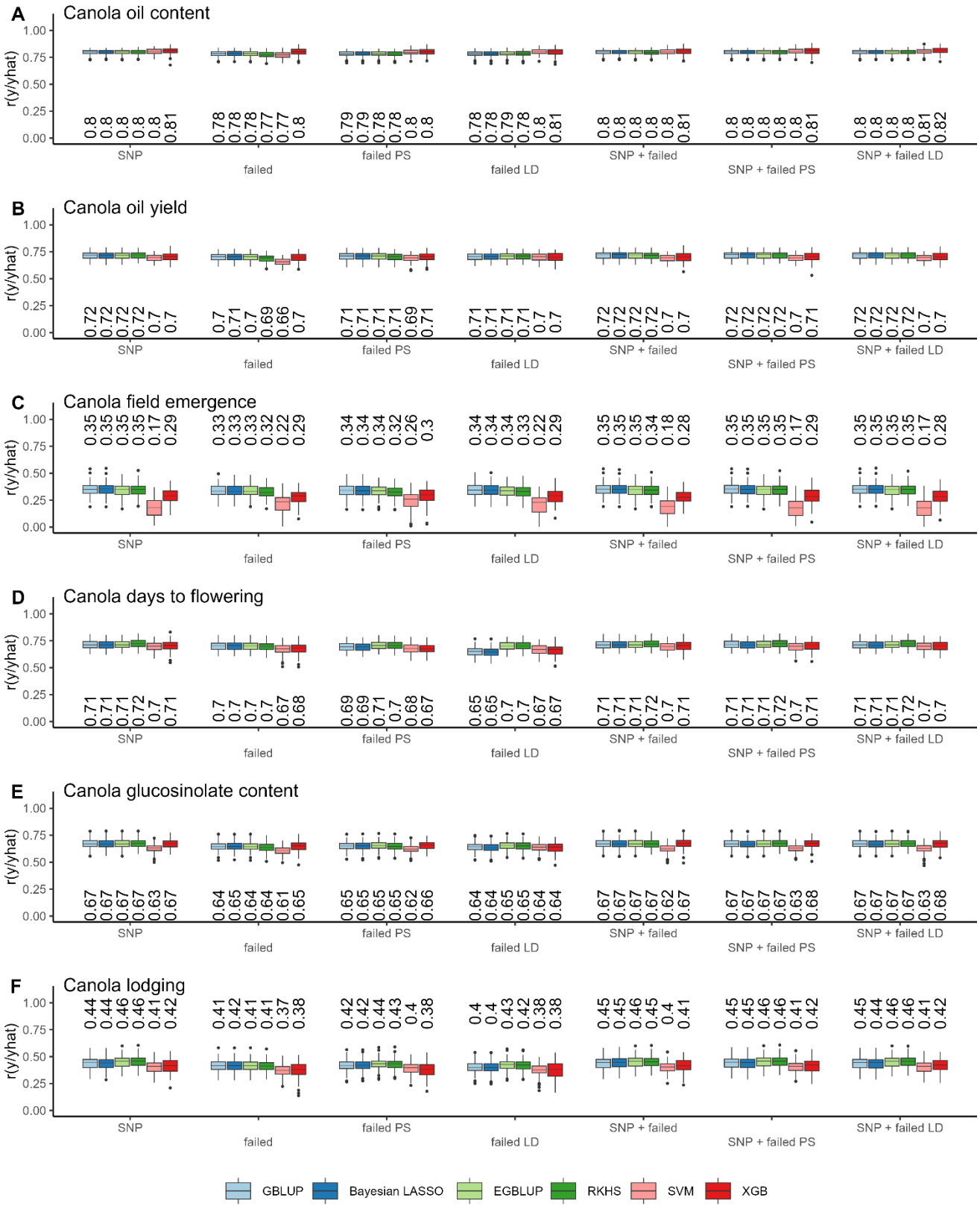

Figure S2 Prediction accuracy (r) based on standard SNPs, failed SNP calls (failed), failed SNP calls filtered by pool specificity (failed PS) and failed SNP calls filtered by LD (failed LD) as well as their combination with GBLUP (light blue), Bayesian Lasso (dark blue), EGBLUP (light green), RKHS (dark green), SVM (pink) and XGB (red). In canola traits: oil content (**A**), oil yield (**B**), field emergence (**C**), days to flowering (**D**), glucosinolate content (**E**) and lodging (**F**) . Values above boxplots represent median values across all cross validation runs

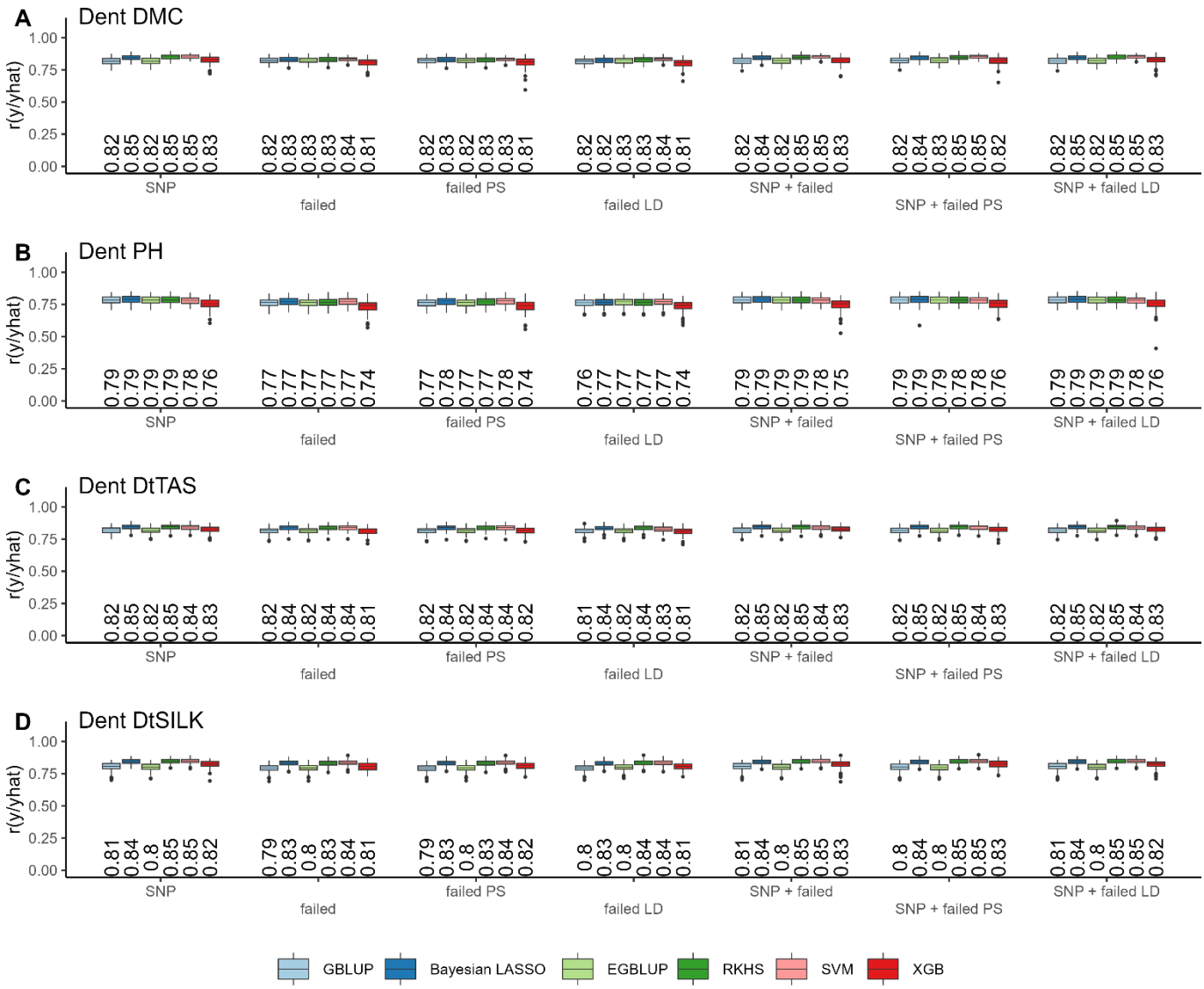

Figure S3 Prediction accuracy (r) based on standard SNPs, failed SNP calls (failed), failed SNP calls filtered by pool specificity (failed PS) and failed SNP calls filtered by LD (failed LD) as well as their combination with GBLUP (light blue), Bayesian Lasso (dark blue), EGBLUP (light green), RKHS (dark green), SVM (pink) and XGB (red). In maize dent traits: DMY (**A**), DMC (**B**), DtTAS (**C**) and DtSILK (**E**). Values above boxplots represent median values across all cross validation runs

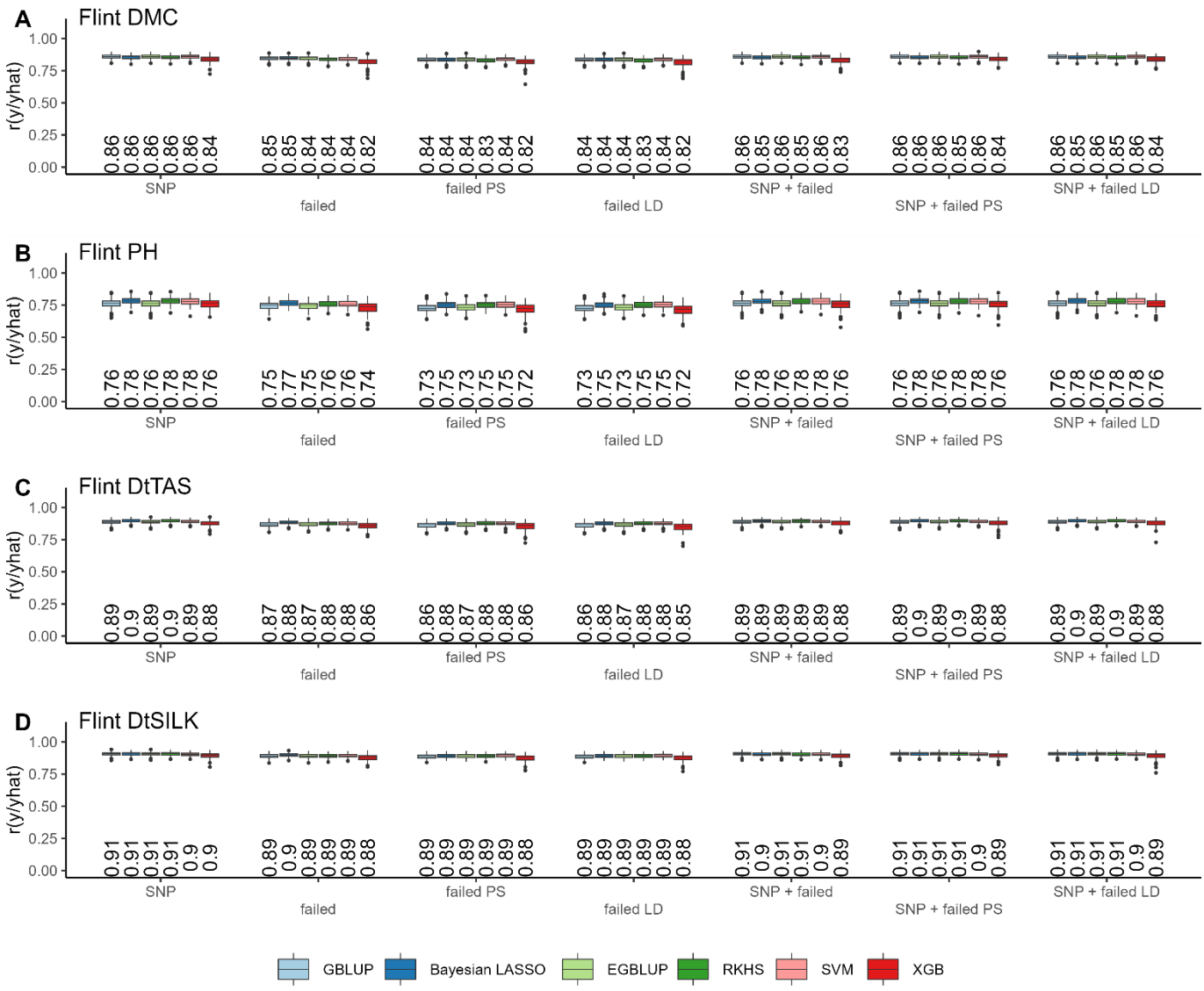

Figure S4 Prediction accuracy (r) based on SNPs, calls failed SNP calls (failed), failed SNP calls filtered by pool specificity (failed PS) and failed SNP calls filtered by LD (failed LD) as well as their combination with GBLUP (light blue), Bayesian Lasso (dark blue), EGBLUP (light green), RKHS (dark green), SVM (pink) and XGB (red). In maize flint traits: DMY (**A**), DMC (**B**), DtTAS (**C**) and DtSILK (**E**). Values above boxplots represent median values across all cross validation runs

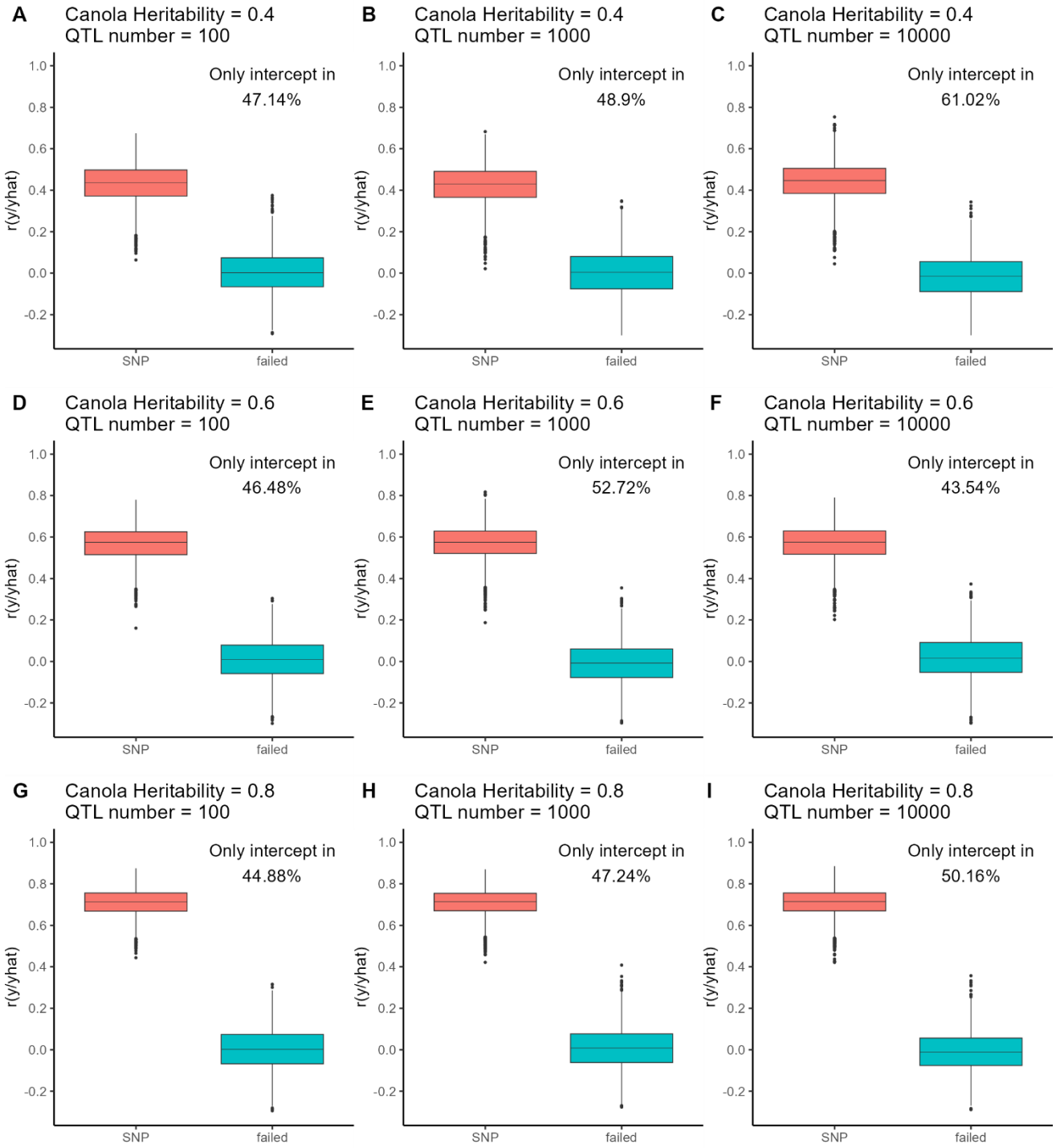

Figure S5 Results of simulated traits based on the genotypic data of the canola dataset. Prediction accuracy (r) across all simulation and cross-validation runs based on SNPs and a randomly sampled failed SNP calls (failed) with the GBLUP model. **A**, **B**, **C** show traits with a simulated heritability of 0.4. **D**, **E**, **F** show traits with simulated heritability of 0.6. While **G**, **H**, **I** display traits with heritability of 0.8. The number of QTL was 100 for **A**, **D**, **G**, 1,000 for **B**, **E**, **H** and 10,000 for **C**, **F**, **I**.

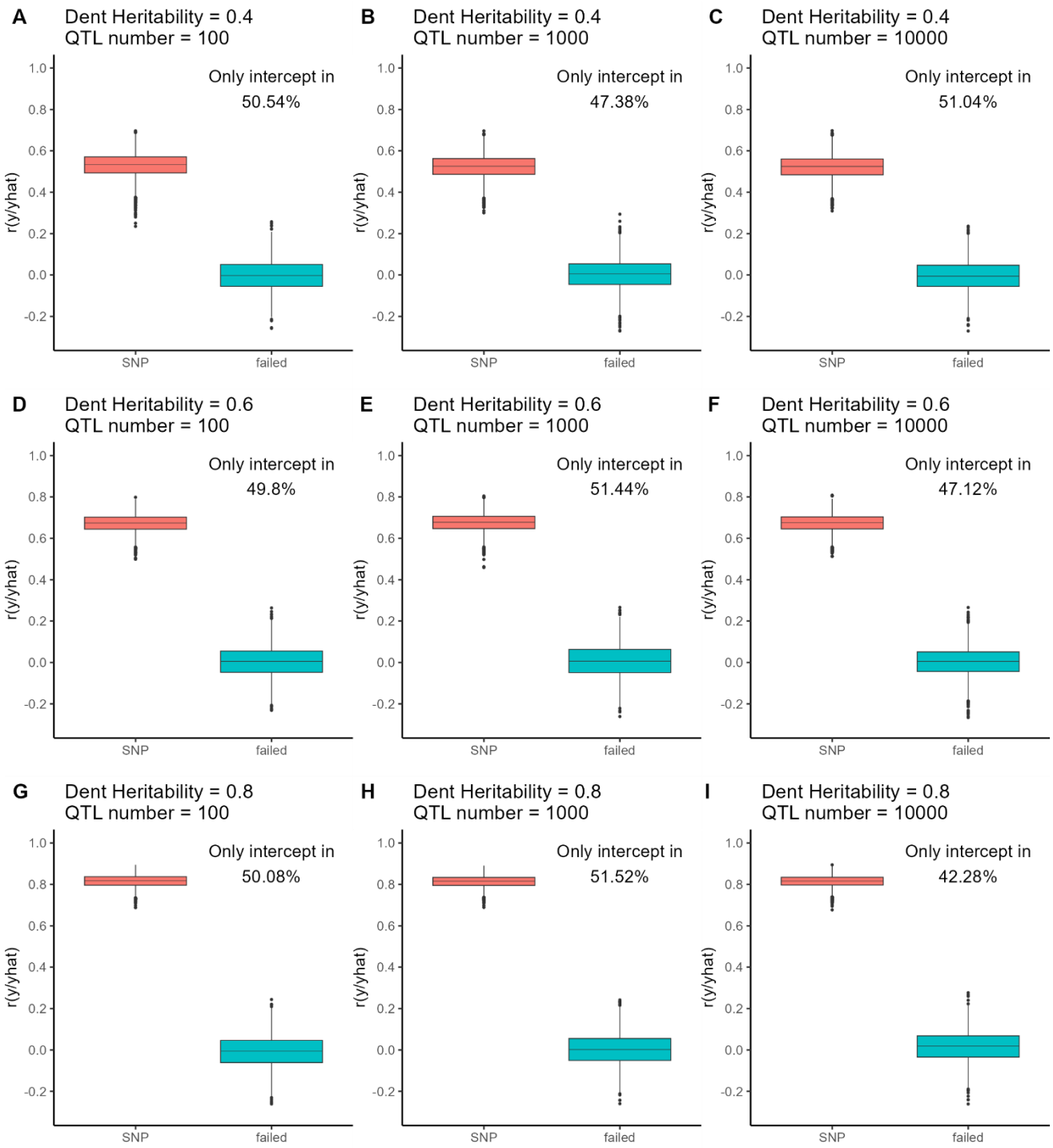

Figure S6 Results of simulated traits based on the genotypic data of the maize dent dataset. Prediction accuracy (r) across all simulation and cross-validation runs based on SNPs and a randomly sampled failed SNP calls (failed) with the GBLUP model. **A**, **B**, **C** show traits with a simulated heritability of 0.4. **D**, **E**, **F** show traits with simulated heritability of 0.6. While **G**, **H**, **I** display traits with heritability of 0.8. The number of QTL was 100 for **A**, **D**, **G**, 1,000 for **B**, **E**, **H** and 10,000 for **C**, **F**, **I**.

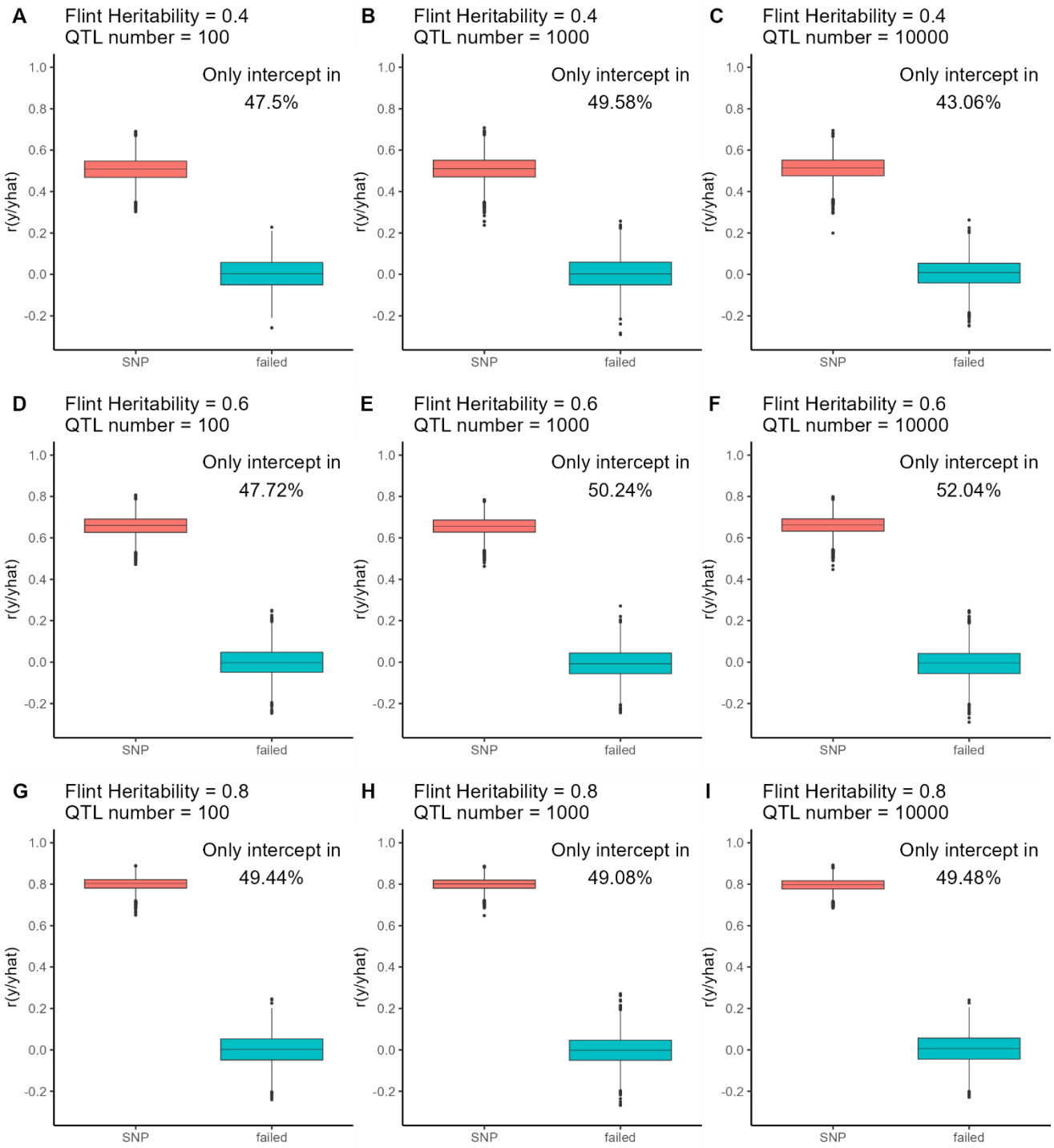

Figure S7 Results of simulated traits based on the genotypic data of the maize flint dataset. Prediction accuracy (r) across all simulation and cross-validation runs based on SNPs and a randomly sampled failed SNP calls (failed) with the GBLUP model. **A**, **B**, **C** show traits with a simulated heritability of 0.4. **D**, **E**, **F** show traits with simulated heritability of 0.6. While **G**, **H**, **I** display traits with heritability of 0.8. The number of QTL was 100 for **A**, **D**, **G**, 1,000 for **B**, **E**, **H** and 10,000 for **C**, **F**, **I**.

Appendix III: Supplementary material from

Weber, S. E., Roscher-Ehrig, L., Kox, T., Abbadi, A., Stahl, A. and Snowdon, R. J. (2023). Genomic Prediction in Brassica napus: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data. under Review in Genome

Genomic Prediction in *Brassica napus*: Evaluating the Benefit of Imputed Whole-Genome Sequencing Data

Sven E. Weber[1]*, Lennard Ehrig[1], Tobias Kox[2],Amine Abbadi[2], Andreas Stahl[3] and Rod J. Snowdon[1]

[1]Department of Plant Breeding, IFZ Research Centre for Biosystems, Land Use and Nutrition, Justus Liebig University, Giessen, Germany

[2]NPZ Innovation GmbH, Holtsee, Germany

[3]Julius Kuehn Institute (JKI), Federal Research Centre for Cultivated Plants, Institute for Resistance Research and Stress Tolerance, Quedlinburg, Germany

* Corresponding author:

Sven E. Weber

Sven.E.Weber@agrar.uni-giessen.de

Supplementary Figures

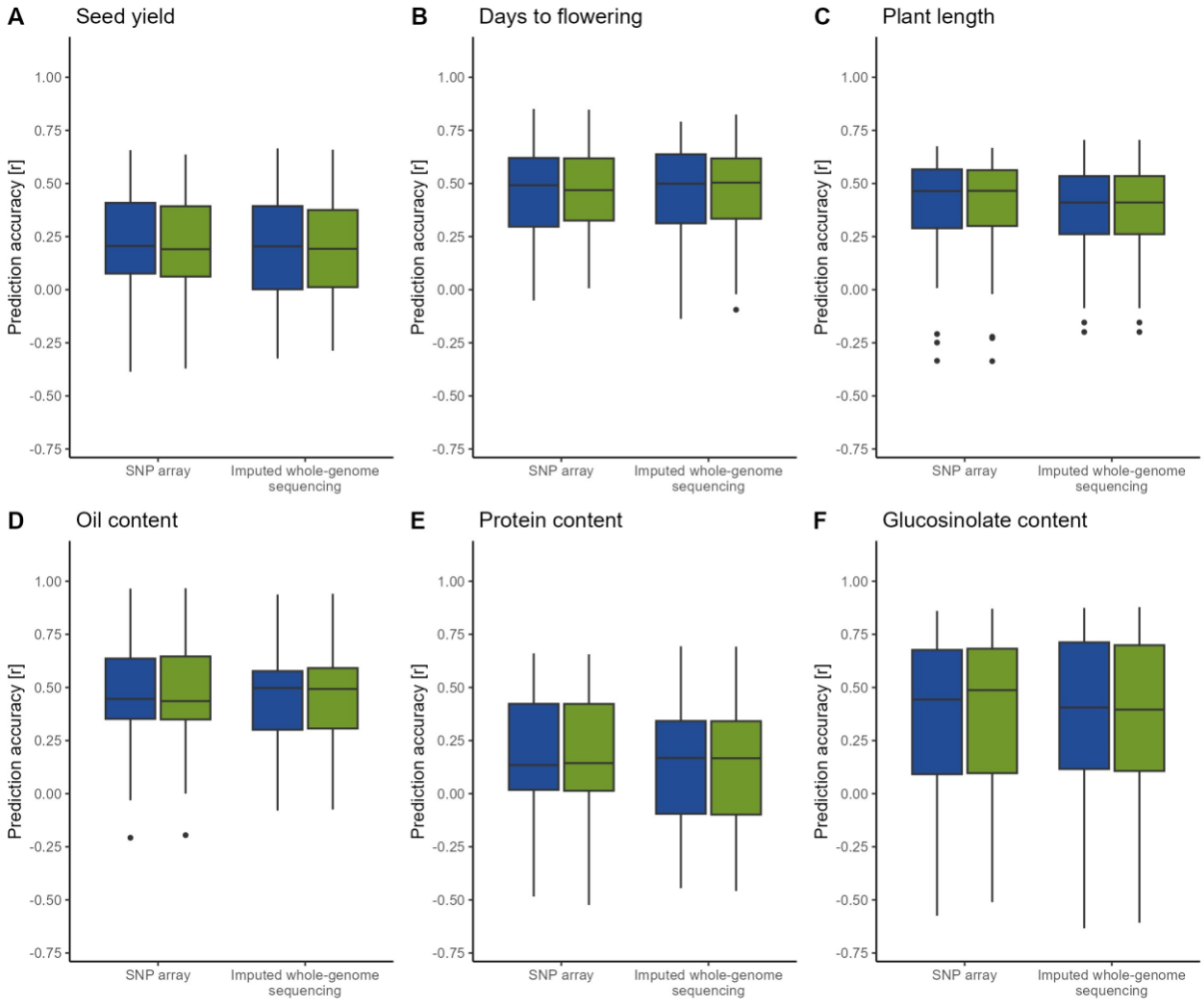

Figure S1 Prediction accuracy (r) with family-wise cross-validation based on SNP from the SNP array and imputed whole-genome sequencing marker data using the GBLUP (blue) and EGBLUP (green). In *brassica napus* traits seed yield (A), days to flowering (B), plant length (C), oil content (D), protein content (E) and glucosinolate content.

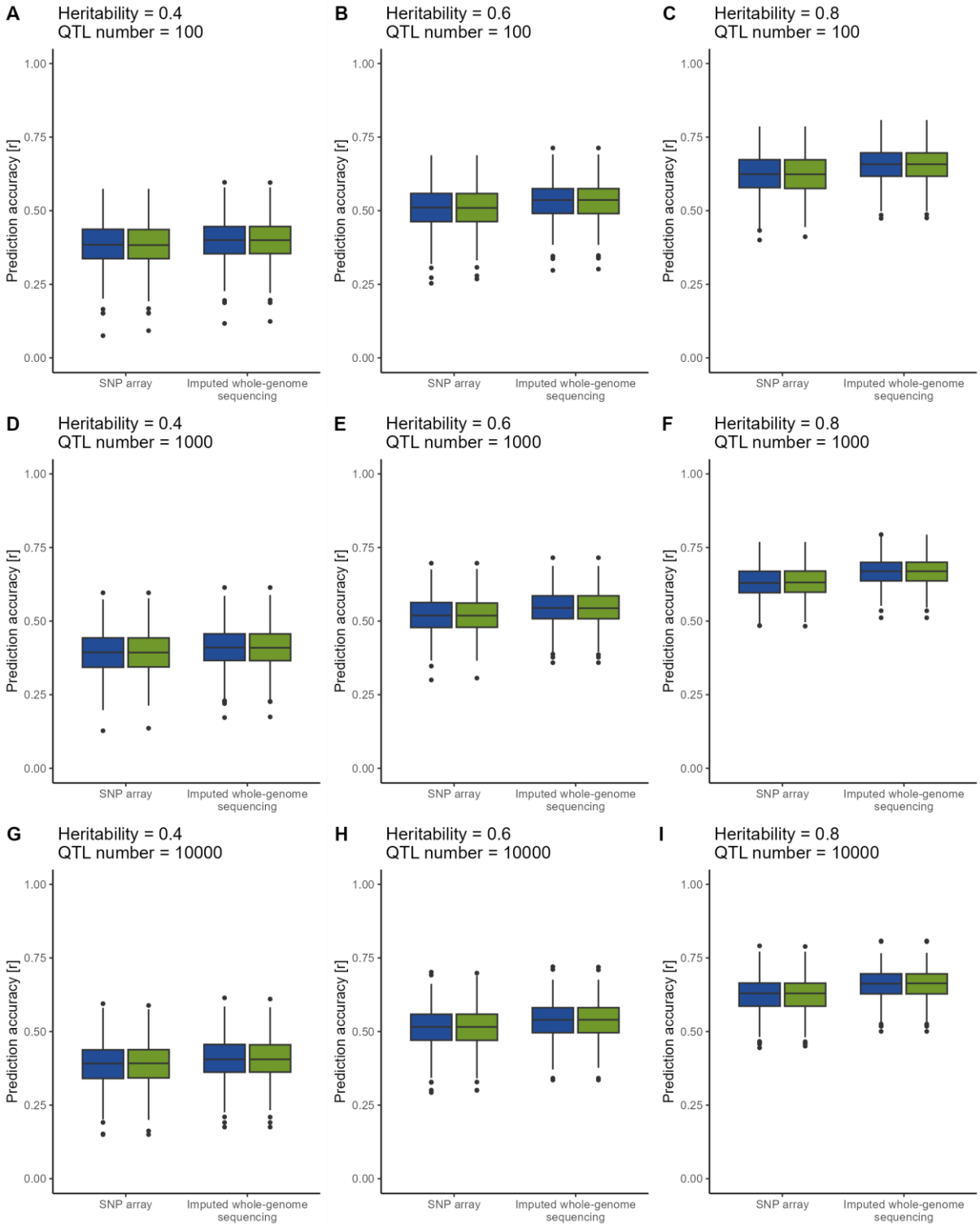

Figure S2 Prediction accuracy (r) with based on fivefold cross-validation based on SNPs from the SNP array and imputed whole-genome sequencing SNP data using the GBLUP (blue) and EGBLUP (green). In simulations with random placement of markers across the genome with heritabilities of 0.4 (A, D, G) 0.6 (B, E, H) and 0.8 (C, F, I) and 100 (A, B, C), 1000 (D, E, F) and 10,000 (G, H, I) QTL.

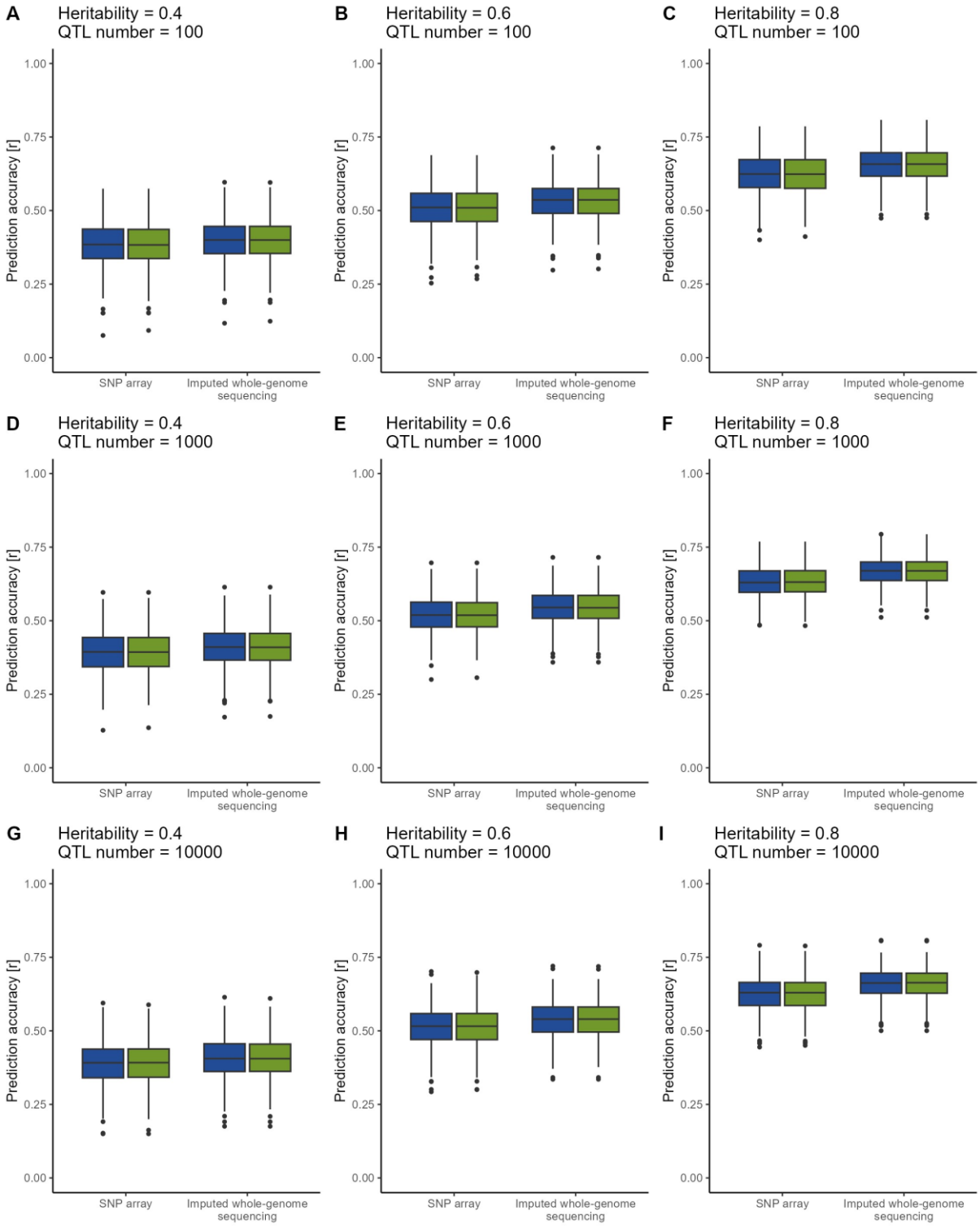

Figure S3 Prediction accuracy (r) with based on fivefold cross-validation based on SNPs from the SNP array and imputed whole-genome sequencing SNP data using the GBLUP (blue) and EGBLUP (green). In simulations with placement of markers in regions not tagged by markers from the SNP array across the genome with heritabilities of 0.4 (A, D, G) 0.6 (B, E, H) and 0.8 (C, F, I) and 100 (A, B, C), 1000 (D, E, F) and 10,000 (G, H, I) QTL.

Declaration of Academic Integrity

„Ich erkläre: Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.“

Gießen 22.12.2023 **Sven Ernst Weber**

Acknowledgments

Die Liste der Personen, die mich während meiner Promotion begleitet haben, ist sehr lang. Viele Menschen haben mich auf meinem Weg unterstützt. Leider ist diese Liste zu lang, um jeden hier namentlich zu erwähnen. Dennoch vielen Dank an alle – danke für die Motivation, die Hilfe und den Spaß, den ich seit meinem Studium und besonders während meiner Promotion erleben durfte.

Am meisten bin ich natürlich meinem Doktorvater Professor Rod Snowdon dankbar. Danke, Rod, dass du damals einem Studenten aus der Pflanzenernährung eine Chance gegeben hast und mir meine Stelle angeboten hast. Ich schätze wirklich, wie du mich in den letzten Jahren gefördert hast – durch deine immer offene Tür, deinen Enthusiasmus für meine Erkenntnisse und das Behandeln als gleichwertigen Wissenschaftler. Durch die Arbeit in deinem Institut konnte ich die Welt sehen und Menschen kennenlernen, von denen ich viel lernen konnte. Danke, dass du mir die Freiheit gibst, mich entsprechend meiner Interessen frei zu entwickeln. Ich bin wirklich froh, bei dir gelandet zu sein.

Des Weiteren möchte ich Professor Matthias Frisch, meinem Zweitbetreuer, danken. Die Gespräche nachmittags mit dir waren immer mehr als erfrischend. Es war mir immer eine Freude, mit dir zu reden. Besonders außergewöhnlich fand ich, dass während wir uns mit den anderen „Jungs“ getroffen haben und eigentlich nichts Wissenschaftliches besprochen haben, du es geschafft hast, mir mit nur ein oder zwei Nebensätzen etwas beizubringen. Danke. Ich habe unsere Gespräche über Heritabilität, Response to Selection, Züchtung und New Kids sehr genossen!

Auch Andreas Stahl darf nicht vergessen werden, der mir auf meinem wissenschaftlichen Werdegang stets zu Rate stand und mich darüber hinaus sehr gefördert hat. Gleichermaßen möchte ich mich bei Benjamin Wittkop bedanken für die vielen Gespräche über Züchtung und Landwirtschaft. Christian Obermeier möchte ich danken für viele Ratschläge über Genetik und Krankheitsresitenz.

Die Liste der Arbeitskollegen in diesem Institut ist sehr lang. Dennoch möchte ich einige Weggefährten besonders erwähnen. Harmeet, du warst und bist mir immer eine große Hilfe gewesen in allen Dingen, die mit Genomik und Bioinformatik zu tun haben. Danke für deine Geduld, die Hilfe und nicht zuletzt für das Korrekturlesen meiner Arbeit. Lennard, als mein

bester Freund und Arbeitskollege, weiß ich dich sehr zu schätzen. Danke für den ständigen Austausch, das gegenseitige Helfen, danke für dein immer offenes Ohr für meine Probleme. Danke, dass du dir immer angehört hast, wer keine Ahnung von Züchtung hat. Danke für die Zeit, die wir zusammen in Australien, Paris und den USA hatten. Ich hoffe wirklich, wir werden weiter die Möglichkeit haben, so zusammen zu arbeiten. Danke, Göle, für deine Verlässlichkeit und dein offenes Ohr. Danke, Philipp, für den ständigen Spaß in Gesprächen, danke für das Fachsimpeln über Machine Learning und alles andere. Danke, Rica, danke Paul, danke Kevin, danke Lukas, danke Erick, danke Samson, danke Manar, danke Stjepan, danke Luisa, danke Luisa, danke Nata. Ihr habt die Zeit hier wirklich besser gemacht.

Besonders möchte ich auch der Basis des Instituts danken. Danke an jeden TA, ihr habt mich alle in meiner Zeit im Labor so unterstützt und mir die Zeit hier leichter gemacht! Besonders möchte ich aber Stavros danken. Danke für deine pausenlose Unterstützung im Labor, das Erklären, wie Sachen laufen, und wo Sachen zu finden sind. Ich weiß, dich nervt das, aber nicht ohne Grund bist du die erste Anlaufstelle für jede Frage nach Hilfe im Labor! Vielen Dank gilt auch Regina, danke dafür, dass du mich so stark beim Extrahieren von DNA unterstützt hast. Du weißt, wir hatten damals eine sehr stressige Zeit. Danke für deine Hilfe!

Auch den eigentlichen Institutschefinnen möchte ich herzlich danken. Danke an das mittlerweile sehr große Sekretariat. Besonderen Dank gilt Frau Schomber, sie haben mir wirklich geholfen, die Untiefen des Verwaltungsapparats zu umschiffen. Danke für jedes „Das machen wir jetzt folgendermaßen…“.

Der größte Dank gebührt allerdings meiner Familie. Danke an meine Eltern Gerold und Edith für eure grenzenlose Unterstützung, für das Unterstützen meiner Reisen, meines Studiums. Einfach danke für alles – ohne euch wäre ich nicht hier! Danke meinem großen Bruder Christian für die Hilfe auf dem akademischen Weg, für jedes Korrekturlesen (nicht zuletzt auch dieser Arbeit). Danke auch an meinen kleinen Bruder Tobias, der mir immer bei der Arbeit zuhause den Rücken freihält. Danke an meine Schwägerin Tina, danke an meine beiden Nichten Emma und Anni!