Justus-Liebig-University, Gießen
Institute of Agronomy and Plant Breeding II
Department of Biometry and Population Genetics

Comparative genomic prediction in winter wheat

Dissertation for a Doctorate Degree in Agricultural Sciences

- Dr. agr. -

in the Faculty of Agricultural Sciences,
Nutritional Sciences and Environmental Management

Examiners:
Prof. Dr. Matthias Frisch
Prof. Dr. Rod J. Snowdon

Submitted by:
Yohannes Fekadu Difabachew

Gießen, October 2024

Contents

[1] Difabachew YF, Frisch M , Moritz AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Weber S, Okoye UJ and Zenke-Philippi C (2023) Genomic prediction with haplotype blocks in wheat. *Front. Plant Sci.* **14**:1168547

[2] Heilmann PG, Difabachew YF, Frisch M, Moritz AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Zenke-Philippi C (2024). Machine learning for prediction of resistance scores in wheat (Triticum aestivum L.). *Plant Breeding.* In Press.
PG Heilmann and YF Difabachew contributed equally to this article.

Abbreviations

ANOVA	analysis of variance
AFLP	amplified fragment length polymorphism
BGLR	Bayesian general linear regression
BLUP	best linear unbiased prediction
BRR	Bayesian ridge regression
DNA	deoxyribonucleic acid
GBM	gradient boosting machine
GEGVs	genomic estimated genetic values
GVCHAP	genomic prediction and variance component estimation using haplotype blocks
κ	kappa coefficient
LD	linkage disequilibrium
MAS	marker-assisted selection
MLM	mixed linear model
PA	prediction accuracy
QTL	quantitative trait locus
RAND	random amplified polymorphic DNA
REML	restricted maximum likelihood
RF	random forest
RR-BLUP	ridge regression BLUP
RMLA	restricted maximum likelihood ANOVA
RMSE	root mean square error
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
SVM	support vector machine
SVR	support vector regression

Chapter 1

General introduction

1.1 Genomic prediction

Since its introduction in animal breeding by Meuwissen et al. (2001), genomic prediction has revolutionized plant breeding, enabling earlier and more accurate identification of superior genotypes. This transformative approach not only accelerates breeding cycles but also significantly reduces costs (Crossa et al. 2017; Hickey et al. 2017). Genomic prediction has been widely applied across various crops, substantially enhancing the precision of selection for complex traits (Crossa et al. 2017). However, it has been affected by several factors, including trait heritability, marker density, the size of training set, the degree of relatedness between breeding populations, genotype-environment interactions, trait genetic architecture, population structure, and the quality of both phenotypic and genotypic data (Zhang et al. 2019; Crossa et al. 2014).

In crops such as maize and wheat, the integration of genomic prediction has led to substantial improvement in genetic gains (Crossa et al. 2014). Genetic gain has been considered as key performance indicator to measure the progress and success of breeding programs. Although challenges remain, genomic prediction has proven to be an invaluable tool, particularly for crops with a long breeding cycle or complex genetic architecture (Crossa et al. 2017).

Developing proper breeding strategies and implementing appropriate statistical models help to explore the relationship between predictor variables (such as SNPs) and response variables (phenotypes) accurately. Traits, or phenotypic values, are typically

categorized based on the number of genes involved, which affects both their inheritance patterns and complexity. Simple traits, controlled by a single gene or a pair of genes, follow straightforward inheritance patterns and can be effectively modeled using basic statistical methods. For instance, colors in flowering plants can be determined (controlled) by the presence or absence of dominant or recessive alleles in the plants gene (*e.g.*, RR or Rr for red flowers, rr for white flowers) (Hartwell et al. 2018; Lynch and Walsh 1998; Falconer and Mackay 1996).

In contrast, oligogenic traits, controlled by a few major genes with large effects along with others of smaller effect, are characterized by heterogeneous variances between markers and exhibit more complex inheritance patterns compared to simple traits. These traits may also involve gene interactions. For example, in wheat, resistance to stem rust is influenced by several major genes, each contributing to varying levels (Pérez and de los Campos 2014). For such traits, statistical models that account for heterogeneous marker variances and utilize different prior distributions, such as Bayesian approaches (Pérez and de los Campos 2014; Meuwissen et al. 2001) or frequentist models that account for heterogeneous genetic variance components with restricted maximum likelihood estimation of error and partitioning according to analysis of variance (ANOVA) variance components (RMLA) (Hofheinz and Frisch 2014) has been recommended for oligogenic traits.

Polygenic traits, influenced by many genes spread throughout the genome, result in continuous phenotypic variation. Examples include complex agronomic traits such as grain yield or plant height in crops, where multiple genes contribute to determining yield or height (Bernardo 2010; Visscher et al. 2008). Unlike traditional marker-assisted selection (MAS), which focuses on a few major quantitative trait loci (QTLs), genomic prediction has been used multiple loci each with small effect across the genome, enabling accurate prediction of complex traits (Crossa et al. 2017; Jannink et al. 2010; Goddard 2009; Goddard and Hayes 2007). Genomic prediction has been widely implemented using statistical models that accommodate all markers across the genome, such as ridge regression best linear unbiased prediction (RR-BLUP), genomic BLUP (GBLUP), and Bayesian methods (VanRaden 2008; Meuwissen et al. 2001; Bernardo 1994). The RR-BLUP, assumes homogeneous marker variance, are primarily designed to capture additive effects based on bi-allelic markers (VanRaden 2008; Meuwissen et al. 2001). However, Bayesian methods has been used for genomic data with heterogeneous marker variance (Pérez and de los Campos 2014; Meuwissen et al. 2001).

The above mentioned genomic prediction methods play an important role in accelerating breeding cycles by enabling early selection and reducing the need for costly phenotyping in the validation set. The accuracy of genomic prediction for polygenic traits can be varied with different types predictor sets, statistical models, and the genetic architecture of the traits (Weber et al. 2023; Zhang et al. 2019; Desta and Ortiz 2014; Crossa et al. 2011). Prediction accuracy, defined as the correlation between observed values from field trials and predicted values from statistical models, is essential for identifying the most promising individuals. Achieving high prediction accuracy is crucial to the success of genomic prediction.

In this study, various comparisons were taking place to identify the appropriate predictor sets, the most effective statistical prediction models, and the interaction between different predictor sets and prediction methods for accurately identifying the best genotypes (Heilmann et al. 2024; Difabachew et al. 2023). These comparisons were made across both agronomic and resistance traits, focusing on prediction accuracy and Cohen's (κ) (kappa coefficient) based on the validation sets (Fig. 1.1).

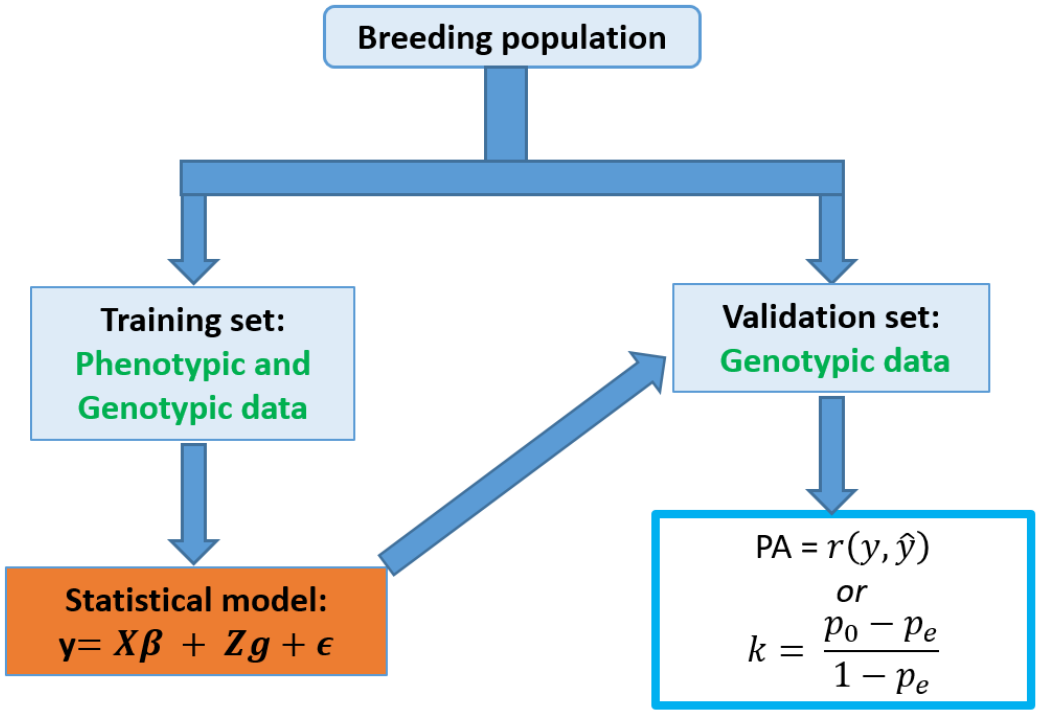

Figure 1.1. Prediction of performance with genetic markers. Individuals in the training set are genotyped and phenotyped indicated with the orange box on the bottom right and comprised 80% of the population. The phenotype is linked to the genotype with a statistical model. The validation set is genotyped and their phenotypes are predicted with the prediction model. The performance of difference prediction approaches were compared in terms of genomic prediction accuracies for metric values and Cohen's (κ) for classification approaches.

1.2 Molecular markers

The rapid development of molecular genetic tools in the 1980s and 1990s enabled the identification of genetic variation at the molecular level, deepening our understanding of the relationship between genetic variants and phenotypes.

Starting in the late 1970s, restriction fragment length polymorphism (RFLP) emerged as a pioneering technique known for its high reproducibility and codominance, despite being labor-intensive and expensive (Jones et al. 1997; Botstein et al. 1980). In the early 1990s, random amplified polymorphic deoxyribonucleic acid (RAPD) provided a simpler, cheaper alternative but had reproducibility and dominance issues (Jones et al. 1997; Rafalski and Tingey 1993). By the mid-1990s, amplified fragment length polymorphism (AFLP) offered high polymorphism and reproducibility, though it was technically complex and expensive (Jones et al. 1997; Vos et al. 1995). By the late 1990s, simple sequence repeats (SSR) gained favor for their high polymorphism and codominance, but remained costly and labor-intensive to develop (Varshney et al. 2005; Jones et al. 1997).

Around early 2000s, single nucleotide polymorphisms (SNPs) became dominant due to their abundance and genome-wide coverage, making them suitable for various genomic applications (Rafalski 2002; Vignal et al. 2002). However, SNPs face challenges like redundancy due to linkage disequilibrium and difficulties detecting rare variants or complex interactions (Myles et al. 2009; Collard and Mackill 2008; Bernardo and Yu 2007). To address these problem, methodologies have been developed, such as grouping adjacent SNPs into a block called haplotype blocks (Gabriel et al. 2002), transforming SNP markers into new features based on neural networks which are called autoencoders (Islam et al. 2023; Goodfellow et al. 2016; Hastie et al. 2009), and selecting most informative SNP markers based on genome wide association study (GWAS) which is called incremental feature selection to implement with machine learning models (Heinrich et al. 2023).

1.2.1 Optimizing SNP information content

SNP markers are commonly used in genomic prediction; however, they capture only additive effects and they are bi-allelic in nature (Villumsen et al. 2009; Barrett et al.

2005). Although, when they are implemented in standard prediction models such as RR-BLUP or GBLUP, inherent redundancy of SNP markers and their interaction effects are usually not captured. This indicates improving SNP information content has been important.

To improve the information content of SNPs, several strategies have been employed, including haplotype blocks (Barrett et al. 2005; Wall and Pritchard 2003; Gabriel et al. 2002), autoencoders (Islam et al. 2023), and incremental feature selection (Heinrich et al. 2023; Li et al. 2018). In this study, we employed these strategies to optimize information contents of SNP markers based on different criteria to improve the prediction accuracy. Different strategies that we utilized are outlined in (Table 1.1).

Table 1.1. Optimizing SNP information content: Various criteria were used to construct different predictor sets, and different statistical methods were applied (Heilmann et al. 2024; Difabachew et al. 2023). Some of these approaches were used for both untransformed and logit-transformed resistance traits.

Predictors type	Criterion 1	Criterion 2	Prediction method
SNP markers	-	-	RR-BLUP, RMLA, BGLR, SVR SVM-c, GBM, GBM-c, RF
LD-AVERAGE	LD-values vs. $t = (0.1, \ldots, 0.9)$	(0,1) tolerance	RR-BLUP, RMLA, GVCHAP
LD-FLANKING	LD-values vs. $t = (0.1, \ldots, 0.9)$	(0,1) tolerance	" , " , "
FIXED-SNP	(5, 10, 20, 50, 100)	-	" , " , "
FIXED-CM	(5, 10, 20) cM	-	" , " , "
HAPLOBLOCKER	(5, 12)	(0.9, 0.95) coverage & (Yes, No) overlap	RR-BLUP,RMLA
Autoencoders	[4000, 1000, 250, 1000, 4000]	-	SVR, RF, GBM
Feature selection	Based on p-values from GWAS	-	RF

1.2.1.1 Haplotype blocks

Haplotype blocks capture local epistasis effects and linkage disequilibrium (LD) patterns (Wall and Pritchard 2003; Gabriel et al. 2002). Haplotype blocks can be constructed using various methods (Table 1.1). Genomic prediction based on LD-based

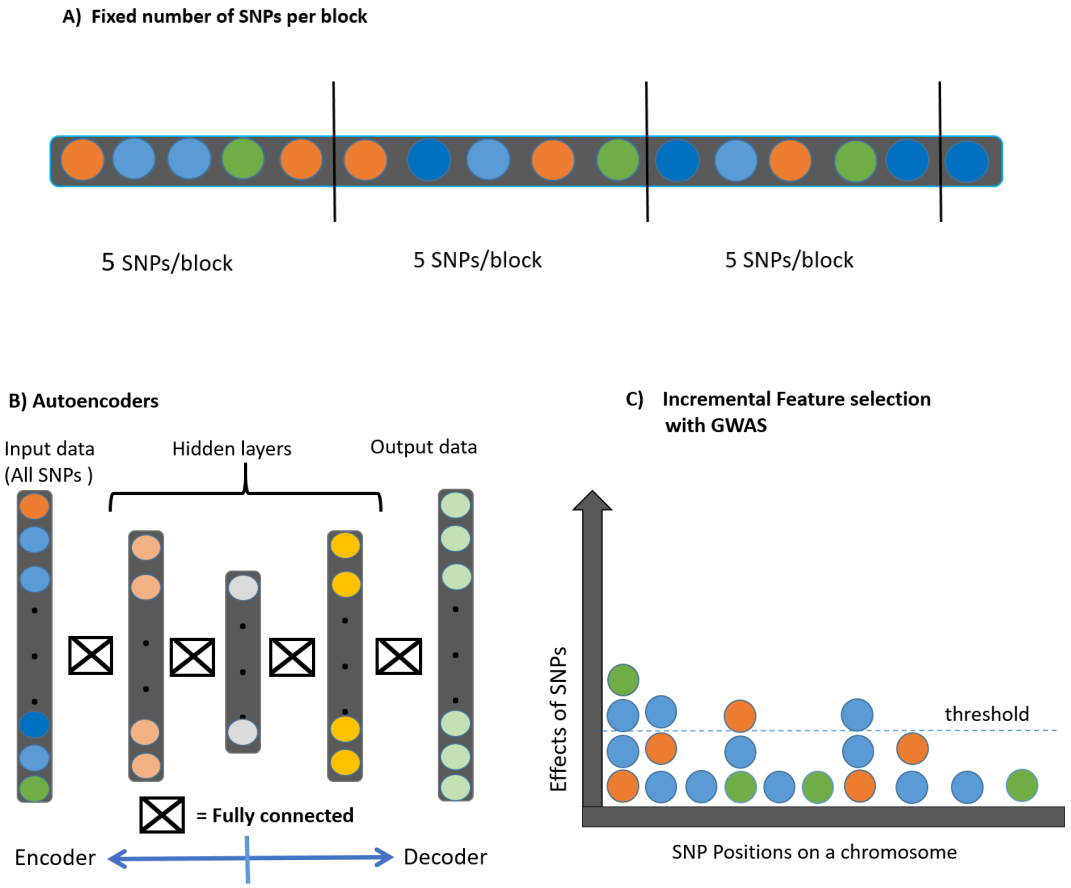

Figure 1.2. In (A), haplotype blocks are formed with fixed SNP markers per block, with each color representing individual SNPs, and five SNPs grouped into each block. In (B), autoencoders are depicted with input and output layers containing similar information with minimal loss, while the middle layer holds a reduced dimensional representation compared to the input layer, without significant loss of information. In (C), incremental feature selection is shown, where SNP markers significantly associated with traits are identified and selected for implementation in the prediction model.

haplotype block is a reasonable approach under the hypothesis that haplotypes are expected to be in stronger linkage LD than SNP markers (Cuyabano et al. 2014). To construct LD-based haplotype blocks, we start by selecting two adjacent loci (SNPs), A and B. We calculate their pairwise LD suggested by Hill and Robertson (1968):

$$r^2 = \frac{D^2}{P(A_1)P(A_2)P(B_1)P(B_2)}, \tag{1.1}$$

where, $D = P(A_1B_1) - P(A_1)P(B_1)$, $P(A_1B_1)$ is the frequency of the haplotype A_1B_1, and $P(A_1)$, $P(A_2)$, $P(B_1)$, and $P(B_2)$ are the allele frequencies at loci A and B. If the r^2 value exceeds a predefined threshold given in (Table 1.1), then loci A and B are grouped into the same haplotype block. Building haplotype blocks by compares the threshold to the average LD of all marker pairs within the block is called LD-AVERAGE. On the other hand, compares the threshold to the LD between the new SNP and the flanking SNP is termed as LD-FLANKING (Difabachew et al. 2023). Then, consider another adjacent SNP markers and calculate the r^2. If this value also exceeds the threshold, include the new SNP to a block. This process continues until the r^2 value between adjacent SNP markers falls below the threshold, marking the boundary of the haplotype block. In both cases (i.e, LD-AVERAGE and LD-FLANKING), if the threshold criteria are fulfilled by all markers in a block then tolerance will be 0. However if a single SNP is added to the block without fulfilling the threshold criteria then tolerance will be 1 (Difabachew et al. 2023; Frisch 2022). Once the block is defined, we move to the next set of adjacent loci and repeat the process across the genome to construct multiple haplotype blocks. SNPs that do not fit into any block are treated as individual blocks.

Another approach to form haplotype blocks is by simply grouping adjacent SNP markers into a block (Fig 1.2 A), referred to as FIXED-SNP (Difabachew et al. 2023). The number of SNP markers per block is chosen arbitrarily, and various block sizes were evaluated based on the criteria in (Table 1.1). Similarly, haplotype blocks can also be constructed by grouping adjacent SNP markers based on specific genetic distances in centiMorgans(cM) along a chromosome segment, known as FIXED-CM (Difabachew et al. 2023). In this method, the length of the chromosome segment is set arbitrarily (Table 1.1). The number of SNPs per block can be different depending on marker density along a chromosome segment (Difabachew et al. 2023).

Haplotype blocks based on shared local allelic sequences, rather than LD between pairwise SNP markers, are referred to as HaploBlocker (Pook et al. 2019). The block

construction process involves several key criteria essential for forming haplotype blocks. Initially, an arbitrary number of SNPs was selected to create local allelic sequences, followed by block identification, merging, filtering, and extension (Pook et al. 2019). This method has been implemented using the HaploBlocker algorithm in the R package (Pook et al. 2019). Other studies (Weber et al. 2023; Da et al. 2022) found that both the traits under investigation and the criteria used during block construction contributed to improving genomic prediction accuracy. In this study different haplotype block building strategies were used and several haplotype blocks were constructed to evaluate their prediction accuracies against SNP markers and among each others.

1.2.1.2 Autoencoders

Autoencoders are unsupervised machine learning techniques based on neural networks, employed for feature extraction to address multicollinearity among non-linear input variables (Kramer 1991). Although inspired by principal component analysis (PCA), autoencoders differ in their ability to capture non-linear relationships, while PCA is limited to linear associations (*cf.* Van Der Maaten et al. 2009). The architecture of autoencoders comprises two main components: an encoder, which compresses input features into a lower-dimensional representation using weight matrices, biases, and activation functions, and a decoder, which reconstructs the original input (Goodfellow et al. 2016; Yasi Wang 2016; Kramer 1991). The intermediate layers are known as hidden layers, and the number of hidden layers are determined by the user. In genomic prediction, SNP markers (input layer) have been transformed to hidden layers, which are then used as input variables for prediction (Islam et al. 2023). In this study, five hidden layers were constructed (Table 1.1 and Fig. 1.2 B), and applied for genomic prediction with the minimum input features based on machine learning algorithms (Heilmann et al. 2024).

1.2.1.3 Incremental feature selection

In addition to the aforementioned techniques in improving SNP information content, simple random sampling was employed and subsets of SNP markers were chosen arbitrary. This allows to determine the minimum number of SNPs required to achieve prediction accuracy comparable to the full SNP marker dataset (Difabachew et al. 2023). However,

more advanced methods for selecting informative SNP subsets, such as incremental feature selection Fig. 1.2 C), which is typically implemented in random forest RF models, have been used after ranking SNP markers based on p-values from GWAS (Heinrich et al. 2023; Li et al. 2018). These selected SNPs are considered highly relevant due to their association with specific phenotypic traits. Applying these subsets of informative SNP markers to the validation set helps to reduce the number of predictor variables. A recent study (Heinrich et al. 2023) demonstrated that prediction accuracies increased when incremental feature selection was used instead of single SNP markers for maize and soy, though no difference was observed for switchgrass. We implemented this approach to assess whether it improves prediction accuracies resistance traits particularly using random forest (Heilmann et al. 2024).

1.2.2 Statistical prediction models

In addition to leveraging various methods for optimizing SNP information content, statistical models are critical for enhancing genomic prediction accuracy. These models range from simple linear frameworks to more complex methods capable of capturing interactions and non-linear effects among predictor variables (Gianola 2013; Meuwissen et al. 2001). In my thesis, I evaluated various statistical models, including linear approaches such as ridge regression best linear unbiased prediction (RR-BLUP) (Endelman 2011; Meuwissen et al. 2001) and Bayesian generalized linear regression (BGLR) for ordinal responses (Pérez and de los Campos 2014). Additionally, I explored kernel and ensemble methods in machine learning algorithms (Vapnik 2013, 1995; Friedman 2001; Breiman 2001), employing different sets of hyperparameter tuning and optimization (Kuhn and Frick 2024; Snoek et al. 2012).

1.2.2.1 Parametric prediction models

Since R.A. Fisher introduced statistical models to agricultural research in 1925, these models have evolved in tandem with advances in data generation across various sectors. A major breakthrough in plant breeding occurred during the Green Revolution in the 1960s, followed by the rise of molecular breeding in the 1980s. With molecular breeding, the ratio of predictor variables (markers) to observations exceeded one, creating challenges for traditional statistical models that require more observations than predictors. In the 1990s, MAS was introduced by Lande and Thompson (1990) to identify markers significantly associated with traits of interest. However, MAS often overestimated its predictive power for complex traits, as non-selected markers could still influence these traits (Meuwissen et al. 2001). While MAS is effective for simple traits controlled by major genes, complex traits like grain yield and disease resistance require more sophisticated statistical approaches capable of incorporating all markers into the model to capture the full complexity of these traits.

The mixed linear model (MLM) was introduced by Henderson (1950) to estimate model parameters that account for both genetic and non-genetic factors. By integrating fixed and random effects, MLM uses as a framework for genomic prediction in animal and plant breeding. This approach facilitates the effective incorporation of genomic information into traditional breeding programs (Endelman and Jannink 2012; Endelman 2011;

Hayes et al. 2009; VanRaden 2008; Habier et al. 2007; Meuwissen et al. 2001; Bernardo 1994). As a result, it enables the estimation of an individual's breeding potential based solely on their genetic material, making the selection process both efficient and powerful for targeting traits of interest (VanRaden 2008; Meuwissen et al. 2001; Bernardo 1994).

The introduction of BLUP enabled the prediction of random effects while accounting for fixed effects and covariance structures (Henderson 1975). In genomic prediction, the MLM is usually expressed as:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Z}g + \epsilon \tag{1.2}$$

where, y is the vector of observed phenotypic values, X is the design matrix of fixed effects, β is the vector of fixed effects, Z is the design matrix allocating phenotype value to individual, g is the vector of random effects, assumed to follow a normal distribution $g \sim N(0, \sigma_g^2\mathrm{G})$, where G is the genetic relationship matrix, ϵ is the vector of residual errors, assumed to follow $\epsilon \sim N(0, \sigma_e^2\mathrm{I})$.

Unlike traditional method that rely on pedigree-based estimates of additive genetic relationships, GBLUP uses the marker-based genomic relationship matrix G, which is more accurately measures the proportion of shared genomic information between individuals (VanRaden 2008; Bernardo 1994). The general formula for estimating genetic values (g) is performed using the mixed model equations (Robinson 1991; Henderson 1975):

$$\begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} \mathrm{X}'\mathrm{X} & \mathrm{X}'\mathrm{Z} \\ \mathrm{Z}'\mathrm{X} & \mathrm{Z}'\mathrm{Z} + \lambda\mathrm{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{X}'\mathrm{y} \\ \mathrm{Z}'\mathrm{y} \end{bmatrix} \tag{1.3}$$

where $\lambda = \sigma_e^2/\sigma_g^2$ is the regularization parameter that controls the shrinkage of genetic values, and I is the identity matrix. Replacing I (Eqn. 1.3) with G changes the prediction method from RR-BLUP to GBLUP (VanRaden 2008).

In RR-BLUP method marker effects are estimated simultaneously, making the accurate estimation of individual marker effects essential (Clark and Werf 2013). However, RR-BLUP is works with biallelic SNP markers, which restricts its direct application to haplotypes. Therefore, procedures such as re-parameterization, transformation, and centering are necessary prerequisites for incorporating haplotype information into the model.

Genomic prediction and variance covariance estimation using haplotypes (GVCHAP) (Prakapenka et al. 2020) is a computing pipeline designed to conduct genomic prediction

using multi-allelic effect (Da 2015). This approach uses the genomic additive relationship matrix and it is derived by transposing, scaling (Hayes and Goddard 2010) the design matrix in (Eqn. 1.2), and it requires adjusting the design matrix for the reference variant by eliminating one column from each haplotype block (Prakapenka et al. 2020). The MLM accounts only for the additive effects of haplotype blocks, as described in Model 4 of Prakapenka et al. (2020). These methods; RR-BLUP and GVCHAP with GBLUP, effectively incorporates both fixed and random factors, which are critical for accurate estimation of genetic values (Prakapenka et al. 2020; Meuwissen et al. 2001). To obtain unbiased variance components, restricted maximum likelihood (REML) is commonly employed in the estimation process (Lynch and Walsh 1998; Gilmour et al. 1995; Robinson 1991).

GBLUP and RR-BLUP, while following different techniques to predict genetic values, both assume constant shrinkage effects and they produce similar results. They have been widely adopted and have become standard procedures for genomic prediction in animal and plant breeding programs, particularly for estimating polygenic traits (Clark and Werf 2013; VanRaden 2008). However, they exhibit limitations when applied to oligogenic traits due to their constant shrinkage assumption (Meuwissen et al. 2001).

To address this issue, Bayesian models improve genomic prediction by integrating prior knowledge and accounting for uncertainties. This approach is particularly advantageous for small sample sizes or when prior information regarding genetic architecture is available (Montesinos López et al. 2022, 2015; Pérez and de los Campos 2014). Bayesian ridge regression (BRR), implemented in the BGLR R package, produces posterior distributions that improve the accuracies of predictions for complex traits (Pérez and de los Campos 2014). These models also allow for heterogeneous marker variance structures, which can be applied to ordinal resistance scores treated as ordered categories (McCullagh 1980). Similarly, RMLA (Hofheinz and Frisch 2014) accommodate heterogeneous marker variances. This helps for accurate estimation of effects of markers.

1.2.2.2 Machine learning algorithms in genomic prediction

In the past two decades, supervised machine learning algorithms have increasingly been recognized as valuable alternatives to linear and Bayesian models for genomic prediction. These algorithms are designed to learn from labeled data, where predictor variables and

output values are used to train a model. Supervised machine learning can be applied to both regression and classification tasks. Free from statistical distributional assumptions, these algorithms effectively handle heterogeneous effects and complex interactions in high-dimensional data, enhancing the accuracy of genomic prediction for both oligogenic and polygenic traits (Alemu et al. 2024; Lourenço et al. 2024; Montesinos López et al. 2022; Merrick et al. 2022; Azodi et al. 2019).

Techniques such as cross-validation are used to assess model performance (Kuhn and Frick 2024; Shewry and Wynn 1987), while hyperparameters control the learning process and fine-tune the model parameters that affect overall performance (Goodfellow et al. 2016; Hastie et al. 2009). Bayesian optimization has been applied to further refine hyperparameters by exploring the hyperparameter space, thereby improving both the accuracy and generalization of machine learning models (Kuhn and Frick 2024; Snoek et al. 2012).

Two common types of supervised machine learning algorithms are kernel and ensemble methods. Kernel methods are implemented to map input data (usually non-linear) into higher-dimensional feature spaces, allowing them to capture complex patterns without assuming linearity (Goodfellow et al. 2016; Hastie et al. 2009). Support vector machines/regression (SVM/SVR) are types of kernel methods, proposed by Drucker et al. (1996) and Cortes and Vapnik (1995). These models aim to find the optimal hyperplane that separates classes (in SVM) or fits the data with minimal error (in SVR), maximizing the margin between the hyperplane and the support vectors (Montesinos López et al. 2022; Hastie et al. 2009). In particular, SVR uses the model:

$$y_i = f(x_i) = \beta + \mathbf{h}(\mathbf{x}_i)^T\mathbf{B} \tag{1.4}$$

where y_i is the quantitative response for the i-th observation, and $f(x_i)$ is the decision function used to predict y_i based on the input x_i. The term $\mathbf{h}(\mathbf{x}_i)^T$ represents the mapping of x_i into a higher-dimensional feature space. β is the bias term (intercept), and $\mathbf{B}$ is the vector of weights for the transformed features.

Gradient boosting machines (GBM) (Friedman 2001) and random forests (RF) (Breiman 2001) are types of ensemble methods. RF constructs multiple decision trees using bootstrap sampling with replacement from the training data and selects a random subset of predictor variables at each tree node for splitting. This approach results in a

low-variance estimator due to averaging across multiple trees, although it may increase bias due to the bias-variance trade-off. For a new observation y_i the RF regression prediction estimated as $\hat{y}_i^B$, the model can be represented as (Hastie et al. 2009):

$$\hat{y}_i^B = \frac{1}{B}\sum_{b=1}^{B} T(\mathbf{x}_i, \Psi_b) \tag{1.5}$$

where the term $T(\mathbf{x}_i, \Psi_b)$ resulted a predicted value for observation i from the b-th tree and $\mathbf{x}_i$ input features. Ψ_b comprises the b-th tree in terms of splitting predictor variables, cut points at each node (where splits are conducted), and terminal node values (predicted values at the leaves of the tree). $\hat{y}_i^B$ is the predicted output for observation i, obtained by averaging over B trees.

GBM enhances the predictive performance in classification or regression trees through sequential model building techniques (Hastie et al. 2009). GBM combines weak learners (typically decision trees) by iteratively minimizing a loss function using gradient descent and each subsequent model is trained on the residuals of the previous one. This process yields a strong predictive model, with each learner progressively improving the overall performance (Lourenço et al. 2024; Hastie et al. 2009; Friedman 2001). In regression, GBM can be represented as follows:

$$\mathbf{y} = f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h(\mathbf{X}; \gamma_m) \tag{1.6}$$

where $\mathbf{y}$ is vector of continuous response values, $\beta_1, ..., \beta_M$ are the coefficients and a weighting factor for each component and the basis functions $h(\mathbf{X}; \gamma_m)$ are weak learners, and comprise the multivariate argument $\mathbf{X}$, characterized by the the set of splitting parameters, $\gamma = (\gamma_1, ..., \gamma_M)$. Prediction is conducted by weighting the ensemble results of all the regression trees (Lourenço et al. 2024; Hastie et al. 2009).

Genomic prediction with supervised machine learning algorithms has, for example, been done for plant breeding, maize (Lourenço et al. 2024; Montesinos-López et al. 2023; Heilmann et al. 2023; Weber et al. 2023; Merrick et al. 2022; Azodi et al. 2019; Zhang et al. 2019; Crossa et al. 2017; Heslot et al. 2014; Ornella et al. 2014), wheat (Ahmadi and Bartholomé 2022; Gonzalez-Camacho et al. 2018; Crossa et al. 2017; Heslot et al. 2014; Ornella et al. 2014, 2012) and in livestock breeding, for cattle and forage research (Mota et al. 2024; Abdollahi-Arpanahi et al. 2020) and pig (Wang et al. 2022) However,

most studies have relied on cross-validation based on single SNP markers to evaluate model performance, often using default model parameters.

In this thesis, we implemented a two-step hyperparameter optimization strategies and initial model tuning for all three supervised machine learning models using five-fold cross-validation, and the performance of the model was evaluated based on the root mean square error (RMSE). Following this, 10 models were trained using a range of hyperparameter combinations using random grid search. To further improve model performance, Bayesian optimization was then applied, to refine the search using 10 additional models, to explore and identify the most promising regions of the hyperparameter space.

1.2.3 Transform response values

Transforming response variables, such as disease score records, offers several advantages in prediction study. Techniques like log or square-root transformations stabilize variances and normalize distributions, improving model robustness and predictive accuracy (Merrick et al. 2022; Montesinos López et al. 2015). These transformations can also linearize relationships between predictors and responses, simplifying model assumptions and enhancing interpretability in plant breeding studies (Bartlett 1947).

A study in rice breeding demonstrated that transforming disease scores effectively enhances the accuracy of predicting resistance to blast disease, thereby facilitating the development of resistant varieties (Xu et al. 2014). Therefore, the choice of response type is pivotal as it dictates the suitability of statistical models and the precision of predictions in genomic prediction studies. In this study, we applied this logit-transformation method for the resistance traits using two frequentist models, such as RR-BLUP and RMLA to evaluate the prediction accuracies and Cohen's κ.

1.2.4 Assessment of prediction accuracy

The prediction performance of various approaches was evaluated using cross-validation, focusing on both continuous and ordinal traits. In each cross-validation run, the dataset was randomly divided into a training set (80%, 289 genotypes) and a validation set (20%, 72 genotypes), ensuring consistency across all predictor sets (Heilmann et al.

2024; Difabachew et al. 2023). Prediction accuracy, $r(y, \hat{y})$, was calculated based on Pearson correlation between the actual phenotypic values (y) and the predicted values ($\hat{y}$), providing a direct measure of how well the predictions aligned with observed outcomes. This method was similarly applied to ordinal traits, such as resistance scores, where logit-transformed predicted values were first reverted to the original scale before computing the correlation (Montesinos López et al. 2022).

In addition to correlation-based accuracy, Cohen's κ was employed to assess classification performance, particularly for identifying genotypes with extreme resistance scores. Genotypes were classified based on their resistance scores as "top" (less than the 10% quantile) or "flop" (greater than the 10% quantile) (Heilmann et al. 2024). Cohen's κ measures the agreement between observed and predicted class assignments, with values ranging from -1 (complete disagreement) to 1 (perfect agreement) (Kuhn and Johnson 2013). A κ value between 0.3 and 0.5 indicates acceptable classification performance, especially in identifying genotypes with critical traits for selection decisions (Gonzalez-Camacho et al. 2018). Machine learning methods; SVM and GBM with SNP markers were used to classify genotypes into "top" and "flop" categories, and Cohen's κ was used to evaluate the results (Kuhn and Johnson 2013). We employed these dual evaluation methods to compare performances of statistical models, predictor sets, and response types to correctly identify best resistance genotypes.

1.2.5 Objectives

Both agronomic traits and resistance to rust disease are great interests for the development of wheat varieties. This study was conducted as part of the Haploselekt project, evaluating 378 elite winter wheat lines tested in 2020 across six locations in Germany (Difabachew et al. 2023). After genotyping and data cleaning, 361 genotypes and 16,667 SNP markers were retained for analysis and to construct different alternative predictor sets (Heilmann et al. 2024; Difabachew et al. 2023). Genomic prediction was performed using different statistical models and sets of predictors to assess the prediction accuracies of resistance traits, including *S. tritici*, *F. graminearum*, *P. triticina* (brown rust), *B. graminis* (mildew), and *P. striiformis* (yellow rust) (Heilmann et al. 2024; Difabachew et al. 2023), along with agronomic traits such as grain yield, protein concentration, starch concentration, hectoliter weight, and plant height (Difabachew et al. 2023).

In breeding programs, enhancing accuracy and efficiency is crucial for both product development and the improvement of breeding populations (Crossa et al. 2017; Jannink et al. 2010; Heffner et al. 2009). Genomic prediction has emerged as a key strategy to achieve these goals by integrating statistical methods with predictor variables and response values (Crossa et al. 2014, 2010; Bernardo and Yu 2007; Meuwissen et al. 2001).

The primary aim of this thesis is to examine the genomic predictions accuracies of various traits and identification of best disease resistance genotypes in winter wheat (*Triticum aestivum* L.). This will be accomplished by evaluating various types of approaches, which integrates different sets of predictors, different prediction models, and response types.

Specifically, this thesis seeks to:

1. Compare the prediction accuracies of different types of predictor sets with the baseline of single SNP markers.
2. Examine the prediction accuracies of various linear models and their interplay with different predictor sets, comparing their performance to the baseline.
3. Investigate the prediction accuracies of machine learning algorithms with linear models for both untransformed and logit-transformed disease resistance traits.
4. Explore the relationship between prediction accuracy and classification metrics for the identification of resistant genotypes.

Chapter 2

Genomic prediction with haplotype blocks in wheat [1]

[1] Difabachew YF, Frisch M , Moritz AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Weber S, Okoye UJ and Zenke-Philippi C (2023) Genomic prediction with haplotype blocks in wheat. *Front. Plant Sci.* **14**:1168547

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 09 May 2023
DOI 10.3389/fpls.2023.1168547

OPEN ACCESS

EDITED BY
Valerio Hoyos-Villegas,
McGill University, Canada

REVIEWED BY
Prabina Kumar Meher,
Indian Council of Agricultural Research, India
Tian Li,
Chinese Academy of Agricultural Sciences, China

*CORRESPONDENCE
Carola Zenke-Philippi
biometry.popgen@uni-giessen.de

RECEIVED 17 February 2023
ACCEPTED 17 April 2023
PUBLISHED 09 May 2023

CITATION
Difabachew YF, Frisch M, Langstroff AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Weber S, Okoye UJ and Zenke-Philippi C (2023) Genomic prediction with haplotype blocks in wheat.
Front. Plant Sci. 14:1168547.
doi: 10.3389/fpls.2023.1168547

Genomic prediction with haplotype blocks in wheat

Yohannes Fekadu Difabachew[1], Matthias Frisch[1], Anna Luise Langstroff[2], Andreas Stahl[3], Benjamin Wittkop[2], Rod J. Snowdon[2], Michael Koch[4], Martin Kirchhoff[5], László Cselényi[6], Markus Wolf[7,8], Jutta Förster[8], Sven Weber[2], Uche Joshua Okoye[1] and Carola Zenke-Philippi[1]*

[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, Gießen, Germany, [2]Institute of Agronomy and Plant Breeding I, Justus Liebig University, Gießen, Germany, [3]Institute for Resistance Research and Stress Tolerance, Julius Kühn Institute, Quedlinburg, Germany, [4]Deutsche Saatveredelung AG, Lippstadt, Germany, [5]Nordsaat Saatzucht GmbH, Langenstein, Germany, [6]Department of Cereal Breeding, W. von Borries-Eckendorf GmbH & Co. KG, Leopoldshöhe, Germany, [7]German Seed Alliance GmbH, Holtsee, Germany, [8]Saaten-Union Biotec GmbH, Leopoldshöhe, Germany

Haplotype blocks might carry additional information compared to single SNPs and have therefore been suggested for use as independent variables in genomic prediction. Studies in different species resulted in more accurate predictions than with single SNPs in some traits but not in others. In addition, it remains unclear how the blocks should be built to obtain the greatest prediction accuracies. Our objective was to compare the results of genomic prediction with different types of haplotype blocks to prediction with single SNPs in 11 traits in winter wheat. We built haplotype blocks from marker data from 361 winter wheat lines based on linkage disequilibrium, fixed SNP numbers, fixed lengths in cM and with the R package HaploBlocker. We used these blocks together with data from single-year field trials in a cross-validation study for predictions with RR-BLUP, an alternative method (RMLA) that allows for heterogeneous marker variances, and GBLUP performed with the software GVCHAP. The greatest prediction accuracies for resistance scores for *B. graminis*, *P. triticina*, and *F. graminearum* were obtained with LD-based haplotype blocks while blocks with fixed marker numbers and fixed lengths in cM resulted in the greatest prediction accuracies for plant height. Prediction accuracies of haplotype blocks built with HaploBlocker were greater than those of the other methods for protein concentration and resistances scores for *S. tritici*, *B. graminis*, and *P. striiformis*. We hypothesize that the trait-dependence is caused by properties of the haplotype blocks that have overlapping and contrasting effects on the prediction accuracy. While they might be able to capture local epistatic effects and to detect ancestral relationships better than single SNPs, prediction accuracy might be reduced by unfavorable characteristics of the design matrices in the models that are due to their multi-allelic nature.

KEYWORDS

genomic prediction, wheat, haplotype blocks, prediction accuracy, cross-validation

Highlights

- Use of haplotype blocks instead of single SNP markers leads to greater accuracy of genomic prediction of quantitative and qualitative traits in wheat.

1 Introduction

Haplotype blocks, most often defined as a set of adjacent markers on a chromosome, were originally proposed as a means of reducing the number of single-nucelotide polymorphisms (SNPs) required to infer the genotype of an individual by the use of tag SNPs (van den Oord and Neale, 2004). This was particularly important when genotyping costs were still very high. More recently, "haplotype stacking", i.e. the combination of favorable haplotype blocks, has been suggested as a promising way for breeders to exploit available genetic variation (Voss-Fels et al., 2019). Moreover, haplotype blocks can identify relationship structures in breeding material and founder lines (Coffman et al., 2020). Functional haplotypes use additive and epistatic marker effects to combine SNPs into haplotype blocks, rather than combining consecutive SNPs. They were shown to identify more candidate regions in a genome-wide association study (GWAS) than single SNPs or other types of haplotype blocks (Liu et al., 2019). Other studies focused on the use of haplotype blocks in genomic prediction. Observed increases in prediction accuracy compared to single SNPs were usually attributed to either local epistasis which is by default captured by haplotype blocks (Jiang et al., 2018; Da et al., 2022) or to the fact that the LD between quantitative trait loci (QTL) and haplotype blocks might be greater than the LD between QTL and single SNPs (Hess et al., 2017). Additionally, it was argued combining SNPs into haplotype blocks can reduce the parameter space in genomic prediction by covering genome stretches that are in linkage disequilibrium (LD) (Cuyabano et al., 2014).

Studies on genomic prediction with haplotype blocks have been conducted for different species, different traits, different types of haplotype blocks and different estimation methods. Investigations have been carried out primarily in animal data sets. In six traits in sheep, prediction accuracies with a GBLUP model which used haplotype blocks were either greater than or similar to the prediction accuracies observed with SNPs only (Araujo et al., 2022). In the three carcass traits liveweight, dressing percentage, and longissimus dorsi muscle weight in beef cattle, haplotype blocks based on either LD or 5, 10, or 20 different SNPs were used in predictions together with either genomic best unbiased prediction (GBLUP) or Bayesian models. It depended on the combination of the trait, the type of haplotype blocks, and the prediction model whether prediction accuracies were greater than, similar to or smaller than the respective reference with single SNPs only (Li et al., 2022). In a Duroc population, the prediction accuracies with GBLUP models that incorporated either haplotype blocks with fixed sizes of 50 to 5000 kilobases per block or haplotype blocks based on the location of genes were up to 7.4% greater than with models that used SNPs only (Bian et al., 2021). In seven traits in humans, increases in prediction accuracies of 1.86 to 8.12% were shown for GBLUP with haplotype blocks with either fixed numbers of SNPs or fixed chromosome distances, or gene-based haplotype blocks (Liang et al., 2020). In Korean cattle, GBLUP with haplotype blocks built from either a fixed number of SNPs, a fixed length in base pairs, or agglomerative hierarchical clustering based on LD showed increased accuracy compared to SNPs for carcass weight and eye muscle area, but found small or no increases in accuracy for backfat thickness (Won et al., 2020). For three traits in three different breeds of dairy cattle, using haplotype blocks of fixed lengths in kb rather than SNPs increased prediction accuracy with different Bayesian methods, with the exception of long (> 500 kb) haplotype blocks. Moreover, increases could only be observed in some combination of traits and breeds but not in others (Hess et al., 2017). In a dairy cattle population, genomic prediction of milk protein, fertility, and mastitis was carried out with LD-based haplotype blocks and either GBLUP or a Bayesian mixture model. An average LD threshold of $D' > 0.45$ increased the prediction accuracies for all three traits. For the other LD thresholds, it depended on the combination of trait and prediction model whether the prediction accuracies for haplotype blocks were greater than those for single SNPs (Cuyabano et al., 2014).

Fewer studies are available for genomic prediction with haplotype blocks in plants. GBLUP with haplotype blocks of 5, 10, 15, or 20 adjacent SNPs resulted in greater prediction accuracies than GBLUP with single SNPs for genomic prediction of yield, test weight, and protein content in a set of wheat lines (Sallam et al., 2020). The use of LD-based haplotype blocks in genomic prediction with different Bayesian models led to greater prediction accuracies compared to single SNPs in *Eucalyptus globulus* (Ballesta et al., 2019). In two data sets with rice genotypes and doubled-haploid maize lines, only a small subset of the traits showed an increase in prediction accuracy with GBLUP based on haplotype blocks with fixed lengths of 2 to 10 SNPs compared to GBLUP with SNP markers (Jiang et al., 2018). Genomic prediction with Bayesian methods was carried out for haplotype blocks built based on LD or with the four-gamete method in rice and maize. The use of haplotype blocks led to greater prediction accuracies in the maize breeding population while in rice, the use of single SNPs was more efficient (Matias et al., 2017).

The simplest ways of constructing haplotypes blocks is to group a fixed number of SNPs or all SNPs within a certain genetic or physical distance on the chromosome into a block. More sophisticated methods employ the LD between SNPs and built haplotype blocks out of those SNPs which are commonly inherited together, shifting the meaning of the block from distance on the chromosome to joint inheritance of SNPs within a block. Some procedures aim to exploit the haplotype diversity across genotypes and result in a haplotype block library that is representative for most of the original SNP data (Zhang et al., 2002; Pook et al., 2019). Other authors built haplotype blocks based on identified genes (Bian et al., 2021) or local genealogy (Edriss et al., 2013).

For wheat, results for genomic prediction with haplotype blocks are available only for blocks with a fixed number of SNPs (Sallam et al.,

2020). Our goal was to compare the accuracy of genomic prediction with haplotype blocks to the standard prediction with single SNPs in 11 traits in winter wheat. In particular, our objectives were to compare (1) different types of block-building methods, (2) different prediction models, and (3) the interaction between both with each other and to a baseline scenario (GBLUP with single SNP markers).

2 Materials and methods

2.1 Field data

378 elite wheat lines were evaluated in a one-year field trial in 2020. We evaluated the resistances against *Septoria tritici*, *Fusarium graminearum*, *Puccinia triticina* (brown rust), *Blumeria graminis* (mildew), and *Puccinia striiformis* (yellow rust). Resistances were scored in observation plots in one replication at one (*S. tritici*, *F. graminearum*), two (*P. triticina*), or three locations (*B. graminis*, *P. striiformis*). In case there was more than one location, the arithmetic mean of the two or three observations was used as the resistance score. In order to improve the readability of the manuscript, we use only the name of the disease instead of the full term for the trait, *e.g.* "*S. tritici*" instead of "*S. tritici* resistance score".

A p-rep design with 54 genotypes in the second replication was conducted at six locations in Germany (Asendorf, Niedersachsen; Böhnshausen, Sachsen-Anhalt; Granskevitz, Mecklenburg-Vorpommern; Groß-Gerau, Hessen; Hovedissen, Nordrhein-Westfalen; Leutewitz, Sachsen) and one location in Poland (Gola) for the quantitative traits grain yield, protein concentration, starch concentration, hectoliter weight, and plant height. Results from Böhnshausen and Groß-Gerau were removed from the analysis due to extreme weather conditions. The remaining field data were analyzed with the mixed linear model

$$g = \mu + l + e + l{:}e + r{:}e + b{:}r{:}e + \epsilon$$

where l is the effect of the line, e is the effect of the environment (location), $l{:}e$ is the genotype-by-environment interaction, $r{:}e$ is the replication-within-environment effect, $b{:}r{:}e$ is the block effect nested within replication and environment, and ϵ is the residual. The genotype was analyzed as a fixed factor, the remaining factors of the model were random. The adjusted entry means were used in further calculations. Protein yield was calculated as the product of yield and protein concentration.

2.2 Genotypic data

All wheat lines were genotyped with the 25k Illumina iSelect SNP array (SGS TraitGenetics, Gatersleben, Germany). All SNP markers with more than two recorded alleles, more than 10% missing values and an expected heterozygosity of < 5% as well as all individuals with more than 10% missing marker information were excluded from the analysis. As a result, 16,667 SNP markers and 361 genotypes remained for further analysis. We used this data set for all further calculations.

2.3 Methods for building haplotype blocks

Haplotype blocks were built with the following methods:

LD-AVERAGE-0, LD-AVERAGE-1, LD-FLANKING-0, LD-FLANKING-1: LD-based haplotype blocks were based on r^2 as a measure of LD (Zhao et al., 2005). r^2 was calculated between all SNPs on each chromosome. Haplotype blocks were then built based on different threshold values t = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9 for r^2. For methods LD-AVERAGE-0 and LD-AVERAGE-1, t was compared to the new average LD between each of the marker pairs within the block if a new SNP was added. For methods LD-FLANKING-0 and LD-FLANKING-1, t was compared to the LD between the new SNP and the SNP flanking the block. In each case, the new SNP was added if the threshold t was exceeded. Blocks were built for tolerance values of zero (all SNPs in the block must meet the criterion; LD-AVERAGE-0 and LD-FLANKING-0) or one (one SNP in the block may fail to meet the criterion; LD-AVERAGE-1 and LD-FLANKING-1). SNPs that could not be assigned to any block were treated as haplotype blocks with just one SNP. An example for how LD-based haplotype blocks were built can be found in Supplementary File S2.

FIXED-SNP: n = 5, 10, 20, 50, or 100 adjacent SNPs were grouped into haplotype blocks. An example for how haplotype blocks were built with method FIXED-SNP can be found in Supplementary File S2.

FIXED-CM: Adjacent SNPs within window sizes of 5, 10, or 20 cM were grouped into haplotype blocks.

HAPLOBLOCKER: The R package HaploBlocker (Pook et al., 2019) starts with haplotype blocks built from windows with a fixed number of SNPs. These blocks are then clustered and merged. Blocks are identified, filtered and extended in an iterative procedure (Pook et al., 2019). The result is a library of blocks that are most representative of the data set. Each block can be either present or absent in each genotype but no variants are defined. In the default setting, overlapping blocks are possible. The percentage of the SNP markers that is covered by the blocks in the final haplotype block library is the target coverage (Pook et al., 2019). We used different combinations of a target coverage of 0.90 or 0.95, a starting window size of 5 or 12 SNPs and either overlapping or non-overlapping blocks (Table 1).

Additionally, we investigated subsets of the marker data with n = 500, 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, 11,000, and 12,000 randomly selected SNPs. A different set of n SNPs was used in each cross-validation run.

2.4 Genomic prediction of marker and haplotype block effects

We used ridge regression best linear unbiased prediction (RR-BLUP) of marker and haplotype block effects (Meuwissen et al., 2001), which was technically implemented using a transformation to an animal model (Shen et al., 2013). It was chosen as a baseline scenario since it has proved to be relatively robust in many circumstances (VanRaden, 2008; Clark and Werf, 2013). In order

TABLE 1 Methods for building haplotype blocks with method HAPLOBLOCKER and statistics of the resulting haplotype blocks.

	Version							
	HB1	HB2	HB3	HB4	HB5	HB6	HB7	HB8
Window size	5	5	5	5	12	12	12	12
Target coverage	0.90	0.90	0.95	0.95	0.90	0.90	0.95	0.95
Overlapping blocks	no	yes	no	yes	no	yes	no	yes
Haplotype blocks	5,818	4,725	8,239	7,612	7,594	7,753	7,967	12,499
SNPs per block								
Average	8	24	7	23	13	37	13	37
Maximum	95	308	85	411	132	471	96	471
Distinct variants per block								
Average	1	1	1	1	1	1	1	1
Maximum	1	1	1	1	1	1	1	1

to get robust results for singular design matrices that may occur during the simulation replications we used method 2 of Nazarian and Gezan (2016). The method is available in our software package SelectionTools (http://population-genetics.uni-giessen.de/software0/). For comparison, we used estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (RMLA) (Hofheinz and Frisch, 2014) which allows for heterogeneous marker variances.

Both RR-BLUP and RMLA are based on the assumption of bi-allelic SNPs. Since haplotype blocks are multi-allelic by nature, we had to re-parametrize the marker matrices to allow for the application of both methods, resulting in a design matrix $\mathbf{Z}$ with one column per haplotype block variant (*cf.* Jiang et al. (2018); Hess et al. (2017); Matias et al. (2017); Cuyabano et al. (2014); Villumsen et al. (2009)). An example can be found in Supplementary File S2. Haplotype blocks built with method HAPLOBLOCKER are encoded as either present or absent and do not have variants. The resulting presence-absence matrix for the blocks was treated as a re-parametrized marker matrix in these cases.

We used the software GVCHAP (Prakapenka et al., 2020) for a multi-allelic haplotype model (Da, 2015) which performs genomic best linear unbiased prediction (GBLUP) with a genomic additive relationship matrix $\mathbf{Z}_{GVCHAP}$ based on haplotype blocks. $\mathbf{Z}_{GVCHAP}$ is the design matrix $\mathbf{Z}$ from above, transposed, scaled (Hayes and Goddard, 2010), multiplied by -1 and with one column eliminated for a "reference variant" for each of the haplotype blocks. The mixed linear model that was then used for GBLUP included additive effects for the haplotype blocks only (Model 4 in Prakapenka et al. (2020)). GVCHAP was not used for haplotype blocks built with method HAPLOBLOCKER.

2.5 Assessment of prediction accuracy

Haplotype blocks were built based on the complete data set with 361 genotypes. Cross-validation was then employed in order to assess the prediction accuracy. In each of 1000 cross-validation runs, the data set was randomly divided into a training set with 289 genotypes and a validation set with 72 genotypes. The same splits into training and validation set were used for all the sets of predictors. The prediction accuracy $r(y,\hat{y})$ was calculated as the correlation between the actual phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set.

2.6 Software

We used R version 4.0.3 for all calculations. The adjusted entry means of the genotypes were estimated with ASReml-R 4.1.0.110. Haplotype blocks were built with either the R package SelectionTools version 22.1 or with HaploBlocker version 1.6.06. RR-BLUP and RMLA were calculated with SelectionTools version 22.1. GBLUP was calculated with GVCHAP version 2.1.

3 Results

3.1 Statistics of haplotype blocks built with different methods

For LD-based blocks with an average r^2 of at least 0.1 between all SNPs within a block and zero tolerance (method LD-AVERAGE-0), 210 haplotype blocks with an average number of 79 SNPs and an average of 186 variants were identified (Table 2). The greatest number of SNPs in one block was 671, the greatest number of variants was 360. 177 SNPs remained unassigned so that the total number of haplotype blocks and unassigned SNPs was 387. When the threshold was raised to 0.9, 2,214 haplotype blocks with an average of 3 SNPs and 17 variants were identified. The maximum numbers were 21 SNPs and 19 variants per block. With 9,694 unassigned SNPs, the total number of haplotype blocks and unassigned SNPs was 11,908 (Table 2).

10.3389/fpls.2023.1168547

TABLE 2 Statistics of haplotype blocks (with ≥2 SNPs) built based on linkage disequilibrium (LD).

	LD threshold								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
LD-AVERAGE-0									
Haplotype blocks	210	639	1,160	1,680	2,100	2,333	2,443	2,414	2,214
Unassigned SNPs	177	898	2,022	3,329	4,515	5,673	6,757	7,971	9,694
Total	387	1,537	3,182	5,009	6,615	8,006	9,200	10,385	11,908
SNPs per block									
Average	79	25	13	8	6	5	4	4	3
Maximum	671	260	150	93	82	57	44	31	21
Distinct variants per block									
Average	186	72	35	22	17	15	15	15	17
Maximum	360	354	303	236	140	78	64	36	19
LD-AVERAGE-1									
Haplotype blocks	204	611	1,085	1,583	2,012	2,268	2,403	2,400	2,199
Unassigned SNPs	171	803	1,789	2,928	4,185	5,478	6,609	7,895	9,632
Total	375	1,414	2,874	4,511	6,197	7,746	9,012	10,295	11,831
SNPs per block									
Average	81	26	14	9	6	5	4	4	3
Maximum	671	252	150	95	82	66	44	31	21
Distinct variants per block									
Average	190	75	38	23	18	16	15	15	18
Maximum	360	354	303	264	246	161	64	36	19
LD-FLANKING-0									
Haplotype blocks	2,252	2,611	2,784	2,836	2,828	2,795	2,749	2,603	2,294
Unassigned SNPs	3,324	4,215	4,911	5,525	6,061	6,716	7,431	8,404	10,020
Total	5,576	6,826	7,695	8,361	8,889	9,511	10,180	11,007	12,314
SNPs per block									
Average	6	5	4	4	4	4	3	3	3
Maximum	67	45	32	32	32	28	28	21	18
Distinct variants per block									
Average	15	13	13	13	13	13	13	14	17
Maximum	87	64	45	39	38	36	36	20	15
LD-FLANKING-1									
Haplotype blocks	1,363	1,769	2,032	2,187	2,262	2,314	2,346	2,275	2,056
Unassigned SNPs	1,556	2,250	2,926	3,634	4,152	4,991	5,819	6,933	8,759
Total	2,919	4,019	4,958	5,821	6,414	7,305	8,165	9,208	10,815
SNPs per block									
Average	11	8	7	6	6	5	5	4	4
Maximum	123	76	76	64	56	56	56	34	34

(Continued)

10.3389/fpls.2023.1168547

TABLE 2 Continued

	LD threshold								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Distinct variants per block									
Average	29	21	18	16	16	15	15	16	18
Maximum	242	168	150	109	106	91	99	82	32

For methods LD-AVERAGE-0 and LD-AVERAGE-1, the average r^2 between all SNPs within a haplotype block including the new SNP is compared to the LD threshold. For methods LD-FLANKING-0 and LD-FLANKING-1, r^2 between the new SNP and the SNP flanking a haplotype block is compared to the LD threshold. Either all SNPs within a block must fulfill the criterion (LD-AVERAGE-0, LD-FLANKING-0) or there may be one SNP that does not fulfill the criterion (LD-AVERAGE-1, LD-FLANKING-1

A tolerance level of one (method LD-AVERAGE-1) which allows for one SNP per block that does not fulfill the criterion changed these numbers only slightly (Table 2). For an average r^2 of at least 0.1 between all SNPs within a block, 204 haplotype blocks with an average number of 81 SNPs and an average of 190 variants were found. The maximum numbers were 671 SNPs and 360 variants per block. Since 171 SNPs remained unassigned, the total number of haplotype blocks and unassigned SNPs was 375. When the threshold was raised to 0.9, 2,199 haplotype blocks with an average of 3 SNPs and 18 variants were identified. At most, there were 21 SNPs and 19 variants per block, and 9,632 SNPs remained unassigned (Table 2).

The influence of the tolerance parameter was much greater for methods LD-FLANKING-0 and LD-FLANKING-1 in which the LD of the new SNP with the SNP flanking the block is compared to the threshold value. Also, the blocks for comparable r^2 thresholds were much smaller than those for methods LD-AVERAGE-0 and LD-AVERAGE-1 (Table 2).

With a tolerance of zero (method LD-FLANKING-0) and an r^2 threshold value of 0.1, 2,252 haplotype blocks with an average number of 6 SNPs and an average of 15 variants were identified (Table 2). The maximum values were 67 SNPs and 87 variants per block and 3,324 SNPs remained unassigned, resulting in a total number of 5,576 haplotype blocks and unassigned SNPs. With an r^2 threshold of 0.9, 2,294 haplotype blocks were found. They had 3 SNPs and 17 variants on average and 18 SNPs and 15 variants maximum. Together with 10,020 unassigned SNPs, there was a total of 12,314 haplotype blocks and unassigned SNPs (Table 2).

When one SNP per block was allowed to not exceed the threshold value (method LD-FLANKING-1), 1,363 haplotype blocks with an average number of 11 SNPs and 29 variants were identified for an r^2 threshold of 0.1 (Table 2). The greatest number of SNPs in one block was 123, the greatest number of variants was 242. Since 1,556 SNPs remained unassigned, there were 2,919 haplotype blocks and unassigned SNPs total. When the threshold was raised to 0.9, 2,056 haplotype blocks with an average of 4 SNPs and 18 variants were identified. The maximum numbers were 34 SNPs and 32 variants per block, and 8,759 SNPs remained unassigned so that the total number of haplotype blocks and unassigned SNPs was 10,815 (Table 2).

For a fixed block size of n = 5 SNPs (method FIXED-SNP), 3,339 haplotype blocks were found with an average of 14 variants and a maximum of 47 variants (Table 3). The number of haplotype blocks reduced to 178 for a block size of n = 100 SNPs. The average variant number was 262, the maximum number was 356 (Table 3). For method FIXED-CM, The number of haplotype blocks reduced from 1,400 to 535 when the window size of each block increased from 5 to 20 cM. On average, there were 12 and 31 SNPs and 35 and

TABLE 3 Statistics of haplotype blocks built based on a fixed number of SNPs per haplotype block (FIXED-SNP) or a fixed window size in cM (FIXED-CM).

	FIXED-SNP					FIXED-CM		
	Number of SNPs per block					Window size in cM		
	5	10	20	50	100	5	10	20
Haplotype blocks	3,339	1,676	843	344	178	1,400	897	535
Unassigned SNPs	4	3	2	0	0	0	1	1
Total	3,343	1,679	845	344	178	1,400	898	536
SNPs per block								
Average	5	10	20	48	94	12	19	31
Maximum	5	10	20	50	100	100	165	272
Distinct variants per block								
Average	14	31	71	172	262	35	58	98
Maximum	47	177	311	348	356	248	318	350

Each haplotype block consists of ≥2 SNPs.

98 variants per block, respectively. The maximum numbers were 100 SNPs and 248 variants for blocks with a length of 5 cM and 272 SNPs and 350 variants for blocks with a length of 20 cM (Table 3). For both methods, FIXED-SNP and FIXED-CM, there were almost no unassigned SNPs (Table 3).

Blocks built with method HAPLOBLOCKER (Pook et al., 2019) varied in three parameters: the starting window size (5 or 12 SNPs), the target coverage of the final haplotype block library (0.90 or 0.95), and the possibility for overlapping blocks (yes or no) (Table 1). The number of haplotype blocks was between 4,725 and 12,499 across all investigated versions (Table 1). When overlapping blocks were allowed, the number of haplotype blocks was always smaller than with non-overlapping blocks while the average and maximum numbers of SNPs per block were greater. The number of haplotype blocks was also greater with a greater target coverage and with a starting window size of 12 SNPs when compared to the alternative version with all the other parameters constant. The target coverage did not influence the average and maximum numbers of variants per block. Both numbers were greater for a starting window size of 12 (Table 1).

3.2 Genomic prediction with a reduced set of SNPs

RR-BLUP with reduced SNP numbers achieved the same level of prediction accuracy as the full set of SNPs with at least 3,000 (yield, *B. graminis*), 4,000 (*S. tritici*, *P. striiformis*, *F. graminearum*), 5,000 (protein concentration, protein yield, starch concentration, hectoliter weight, *P. triticina*), or 6,000 (plant height) SNPs, respectively (Supplementary Figure S1).

3.3 RR-BLUP with haplotype blocks

RR-BLUP with the full set of SNPs (baseline) had greater prediction accuracies than all block-based predictions for yield (Figure 1), protein yield, and starch concentration (Supplementary Figures S4, S5). For all other traits, there was at least one type of haplotype block for which the prediction accuracies were greater than the baseline (Supplementary Figures S2, S3, S6–S12). Yield, plant height and hectoliter weight and the resistance scores for *B. graminis*, *P. striiformis*, and *P. triticina* were chosen as illustrative examples for quantitative and qualitative traits, respectively (Figures 1, 2). Results for all investigated traits can be found in the Supplementary Figures S2–S12.

Haplotype blocks with 5, 10, or 20 SNPs (method FIXED-SNP) or with a fixed length of 5, 10, or 20 cM (method FIXED-CM) showed prediction accuracies above baseline for plant height (Figure 1). Haplotype blocks with a fixed number of 5 SNPs and some block types built with method HAPLOBLOCKER resulted in prediction accuracies above baseline for hectoliter weight (Figure 1). In the case of *B. graminis*, LD-based haplotype blocks led to greater prediction accuracies than single SNPs. This was the case for all haplotype blocks built with methods LD-FLANKING-0 and LD-FLANKING-1 and for those blocks built with methods LD-AVERAGE-0 and LD-AVERAGE-1 and an LD threshold of $r^2 > 0.3$. In this trait, haplotype blocks built with method HAPLOBLOCKER based on a starting window size of 5 SNPs also led to greater prediction accuracies compared to the baseline (Figure 2).

In *P. striiformis*, all haplotype blocks types built with method HAPLOBLOCKER resulted in greater prediction accuracies than single SNPs. Prediction accuracies for LD-based haplotype blocks were only greater than the baseline when the haplotype blocks were built with methods LD-FLANKING-0 or LD-FLANKING-1 and a low LD threshold of $r^2 > 0.1$ or 0.2, which was also the case for *P. triticina* (Figure 2).

Some patterns were observed independent of the trait. For LD-based haplotype blocks built with methods LD-AVERAGE-0 and LD-AVERAGE-1, r^2 had to be at least 0.2 or 0.3 to lead to meaningful predictions. Similarly, prediction accuracies declined considerably for haplotype blocks with a fixed SNP number of 50 or sometimes 20 (method FIXED-SNP), and a length of more than 10 or 20 cM (method FIXED-CM). For LD-based haplotype blocks built with methods LD-FLANKING-0 and LD-FLANKING-1, the influence of the threshold value on the prediction accuracies was smaller than for methods LD-AVERAGE-0 and LD-AVERAGE-1 (Figures 1, 2).

Within the haplotype blocks built with method HAPLOBLOCKER, different patterns became apparent. In yield and *P. striiformis*, all prediction accuracies were on the same level (Figures 1, 2). In plant height, the non-overlapping blocks resulted in greater prediction accuracies than the overlapping blocks with the same parameters (Figure 1). In *B. graminis* and *P. triticina*, haplotype blocks based on a starting window size of 5 SNPs showed greater prediction accuracies than haplotype blocks based on a starting window size of 12, regardless of the other parameters (Figure 2).

3.4 Alternative genomic prediction methods

It was dependent on the trait whether ridge regression with homogeneous (RR-BLUP) or heterogeneous (RMLA) marker variances resulted in greater prediction accuracies. In *B. graminis*, RR-BLUP showed greater prediction accuracies than RMLA (Figure 2). RR-BLUP and RMLA resulted in similar prediction accuracies for yield, plant height, and *P. striiformis* (Figures 1, 2). RMLA showed greater prediction accuracies than RR-BLUP in *P. triticina* and hectoliter weight (Figures 1, 2).

Prediction accuracies obtained with GBLUP with GVCHAP were mostly similar to those with RR-BLUP (Figures 1, 2). The only exceptions were the haplotype blocks with a fixed number of 50 or 100 SNPs (method FIXED-SNP) and sometimes fixed block lengths of 10 or 20 cM (method FIXED-CM) where GBLUP with GVCHAP showed greater prediction accuracies than the other estimation methods (Figures 1, 2). These were also the only instances in which the overall ranking of the estimation methods changed. In all other cases, the ranking of the methods remained the same across all types of haplotype blocks (Figures 1, 2).

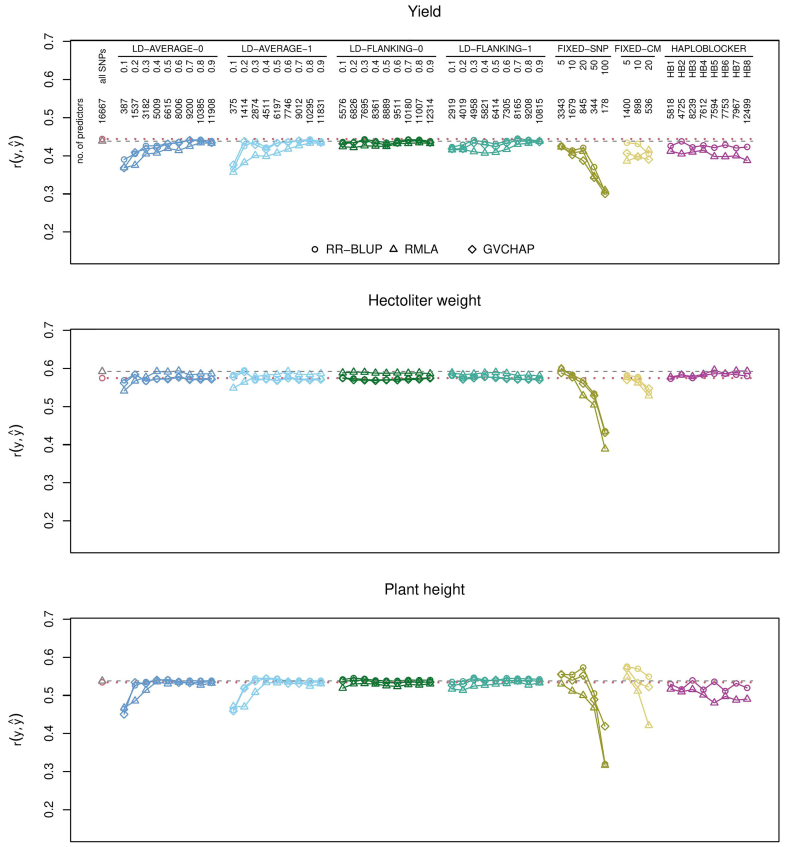

FIGURE 1
Prediction accuracies for genomic prediction of yield, hectoliter weight, and plant height with different types of haplotype blocks and estimation methods. The plots show the medians of the correlations $r(y,\hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (LD-AVERAGE-0, LD-AVERAGE-1, LD-FLANKING-0, LD-FLANKING-1) with different threshold values t=0.1,0.2,...0.9 for r^2, with fixed numbers of SNPs per block (FIXED-SNP), with a fixed block length in cM (FIXED-CM), or with the R package HaploBlocker (HAPLOBLOCKER). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

4 Discussion

4.1 Genomic prediction with haplotype blocks

4.1.1 Number of haplotype blocks

The different procedures for building haplotype blocks resulted in vastly varying number of haplotype blocks and unassigned SNPs that were used for the predictions (Tables 1–3). We determined the minimum number of SNPs required for accurate predictions in order to determine whether changes in prediction accuracy were related to the coverage of the genome with SNPs. Across all traits, RR-BLUP with 3,000 through 6,000 SNPs randomly selected SNPs resulted in prediction accuracies comparable to that of the full set of 16,667 SNPs (baseline) (Supplementary Figure S1). It has to be noted that the number of haplotype blocks does not directly translate into the dimensions of the design matrices in the mixed linear models. In RR-BLUP with m bi-allelic SNP markers, the design matrix $\mathbf{Z}$ for the genotypic effects has m columns. The number of columns of the design matrix $\mathbf{Z}$ in a haplotype model depends on (a) the number of haplotype blocks and unassigned SNPs, and (b) the number of variants at each haplotype block. We found that very often, even though there were fewer haplotype blocks than there were single SNPs in the full marker data set, the dimensions of the resulting design matrix $\mathbf{Z}$ in the reparameterized case were roughly the same because of the high number of variants at some of the haplotype blocks. These dimensions were reduced when monomorphic block-variant combinations were eliminated in the training set. For example, building LD-based blocks with methods LD-FLANKING-0 and an r^2 threshold of 0.1, 2,252 haplotype blocks and 3,324 unassigned SNPs were identified. They translated in a design matrix $\mathbf{Z}$ with 18,330 columns (block/

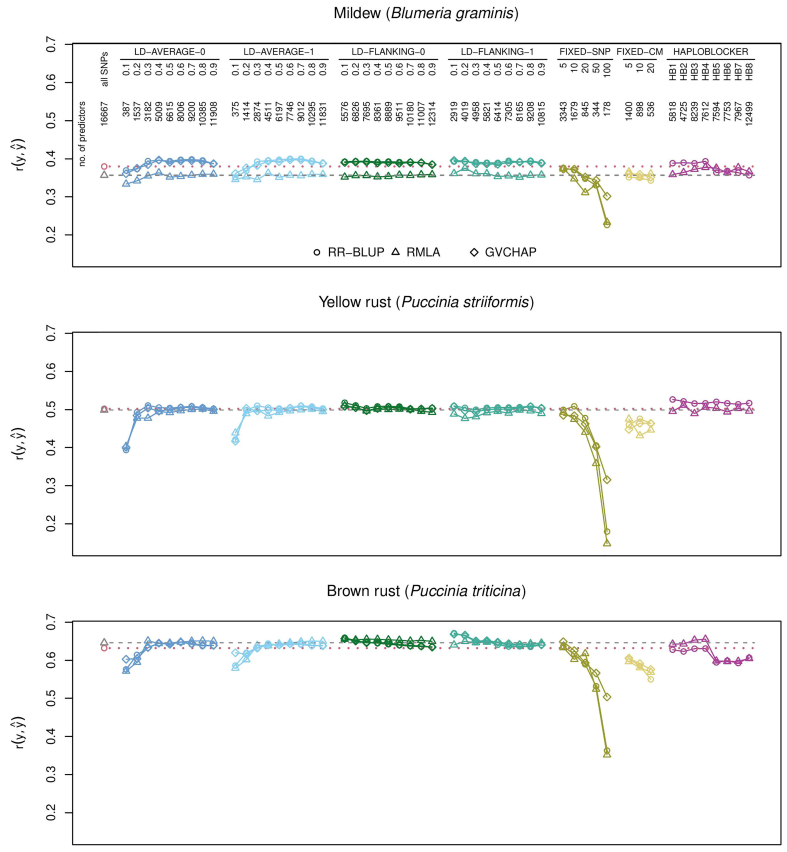

FIGURE 2
Prediction accuracies for genomic prediction of resistance scores for *B. graminis*, *P. striiformis*, and *P. triticina* with different types of haplotype blocks and estimation methods. The plots show the medians of the correlations $r(y,\hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (LD-AVERAGE-0, LD-AVERAGE-1, LD-FLANKING-0, LD-FLANKING-1) with different threshold values t=0.1,0.2,...,0.9 for r^2, with fixed numbers of SNPs per block (FIXED-SNP), with a fixed block length in cM (FIXED-CM), or with the R package HaploBlocker (HAPLOBLOCKER). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

SNP-variant combinations), which were reduced to 13,803 in the first cross-validation run after eliminating those block-variant combinations with an expected heterozygosity of less than 5% in the training set. These observations show that haplotype blocks do not reduce the parameter space, as has been claimed to be the case if they are used in genomic selection (Cuyabano et al., 2014).

In all cases in which the block-building methods resulted in less than 500 haplotype blocks and unassigned SNPs, for example with LD-based blocks built with methods LD-AVERAGE-0 and LD-AVERAGE-0 and low LD thresholds or with method FIXED-SNP and 50 or 100 SNP markers per block, the prediction accuracies were much smaller than in the baseline scenario (Figures 1, 2). Conversely, 536–898 haplotype blocks and unassigned SNPs led to prediction accuracies above baseline for plant height when haplotype blocks were built with methods FIXED-SNP and FIXED-CM (Figure 1), an increase which could not be achieved by simply reducing the number of SNPs in the predictions (Supplementary Figure S1). These findings confirm that haplotype blocks do more than just eliminate noise in the form of redundant information from the SNP data since that reduction should also be achievable by a simple reduction in the number of SNPs used for the predictions.

Some authors claim that the additional information carried by haplotype blocks is mainly local epistatis (Jiang et al., 2018; Da et al., 2022). Other hypotheses include that ancestral relationships might be detected better by haplotype blocks and that the LD between causal mutations and haplotype variants might be greater than for single SNP markers (Hess et al., 2017). It is possible that this additional information content is outweighed by the properties of the design matrices because the large number of columns that belong to the variants of a single haplotype block might introduce multicollinearity and estimation errors (Matias et al., 2017). A

further decrease might be caused by the substantial number of haplotype block variants that is removed when filtering out monomorphic loci.

4.1.2 Block-building methods

It depended on the trait whether adding SNPs based on the average LD in the block (methods LD-AVERAGE-0 and LD-AVERAGE-1) or based on their LD with the flanking SNP of a block (methods LD-FLANKING-0 and LD-FLANKING-1) resulted in greater prediction accuracies (Figures 1, 2). Similarly, allowing for no (tolerance zero) or one (tolerance one) SNP in each block that does not fulfill the LD threshold made in a difference in some cases while in others it did not. For example, there was no difference between the prediction accuracies of haplotype blocks built with methods LD-AVERAGE-0, LD-AVERAGE-1, LD-FLANKING-0, and LD-FLANKING-1 in the prediction of plant height (Figure 1). In *B. graminis* and *P. striiformis*, prediction accuracies for haplotype blocks built with methods LD-AVERAGE-0 and LD-AVERAGE-1 increased with an increase in the LD threshold while they decreased for haplotype blocks built with methods LD-FLANKING-0 and LD-FLANKING-1. In *P. triticina*, haplotype blocks built with method LD-FLANKING-1 resulted in much greater prediction accuracies than single SNPs while this increase was not observed for haplotype blocks built with the other methods (Figure 2). Overall, thresholds for r^2 of 0.4 for haplotype blocks built with methods LD-FLANKING-0 and LD-FLANKING-1 and 0.6 for haplotype blocks built with methods LD-AVERAGE-0 and LD-AVERAGE-1 led to prediction accuracies comparable to that of the baseline, indicating that the information content was similar (Figures 1; 2, Supplementary Figure S1). These haplotype blocks did not lead to greater prediction accuracies than single SNPs in any of the cases.

Prediction with haplotype blocks with a fixed number of SNPs (method FIXED-SNP) resulted in a decrease in prediction accuracies in most traits (Figures 1, 2). The most notable exception was plant height which showed an increase in prediction accuracy for RR-BLUP (Figure 1). For the other traits, predictions of *P. striiformis*, *P. triticina*, and hectoliter weight could be increased compared to the baseline with either one or several types of haplotypes blocks built with method FIXED-SNP (Figures 1, 2). The same was observed for haplotypes blocks built with method FIXED-CM: The prediction accuracies for plant height were substantially greater (Figure 1). In the other traits, prediction accuracies for haplotype blocks built with methods FIXED-SNP or FIXED-CM were smaller than those for haplotype blocks built based on LD (Figures 1, 2). A possible reason for this finding could be that blocks with a fixed number of SNPs or fixed window size combine SNPs arbitrarily while LD-based blocks take into account information from the data set regarding the recombination frequencies. This is reflected in the finding that a relatively high number of SNPs remains unassigned with the LD-based block-building methods (Table 2). It can therefore be expected that LD-based blocks should capture more or less the same information about QTL for the trait even with low LD thresholds while marker-trait associations might be broken for blocks with a fixed number of SNPs or fixed length in cM.

Prediction accuracies increased slightly for *B. graminis*, *P. striiformis*, and *P. triticina* for haplotype blocks built with method HAPLOBLOCKER. For plant height, prediction accuracies for non-overlapping haplotype blocks were always greater than their counterparts with overlapping blocks, even though none of the versions led to prediction accuracies greater than RR-BLUP with the full set of SNPs. For *B. graminis* and *P. triticina*, the greatest differences within the haplotype blocks built with method HAPLOBLOCKER were between the starting window sizes 5 and 12. Prediction accuracies were comparable to that of single SNP markers for window size 5 but much smaller for window size 12 (Figure 2).

Trait-dependence of prediction accuracies with different block-building methods was also observed in pigs (Bian et al., 2021), humans (Liang et al., 2020), cattle (Cuyabano et al., 2014; Won et al., 2020), eucalyptus (Ballesta et al., 2019), wheat (Sallam et al., 2020) and rice and maize (Matias et al., 2017). Other authors studied the optimal haplotype block length required for estimation of the genomic relationship matrix and also arrived at the conclusion that it depends on the trait which block length is best (Ferdosi et al., 2016). Apparently, there is no single method that can generally be recommended for building meaningful blocks. If, as proposed by some authors (Jiang et al., 2018; Da et al., 2022), greater prediction accuracies are mostly due to local epistasis that is captured by the haplotype blocks, these findings raise the question if the optimal choice of haplotype blocks depends on the exact structure of "local" epistatis exhibited for the trait.

It is also possible that the genetic architecture of the trait influenced whether haplotype blocks led to greater prediction accuracies than single markers. In our data set, we observed greater prediction accuracies for haplotype blocks than for single SNPs in the prediction of oligogenic traits like the resistance scores for *B. graminis*, *P. striiformis*, and *P. triticina*. In contrast, there was no obvious and consistent advantage of using haplotype blocks for the prediction of polygenic traits like yield, hectoliter weight, and plant height (Figures 1, 2).

Haplotype blocks divide the chromosome into segments and effects are then assigned to these chromosome segments rather than distributed over many markers. This approach might be beneficial for the prediction of oligogenic traits because it reduces the noise caused by a great number of markers that are not linked to genes that are causal for the trait. Additionally, haplotype blocks with a positive effect can then be used to combine favorable chromosome stretches *via* haplotype stacking (Voss-Fels et al., 2019). For highly polygenic traits, grouping markers into chromosome segments and assigning effects to segments rather than to single markers is not expected to lead to greater prediction accuracies because is it precisely the distribution of effects over many markers that corresponds to their polygenic nature. The effect of the haplotype block would then be a "net effect" that is roughly equal to the sum of the effects of the single markers in this block, not adding any additional or removing redundant information.

4.2 Investigating alternative genomic prediction methods

4.2.1 Assumption of heterogeneous marker variances

Using a the RMLA model with heterogeneous marker variances (Hofheinz and Frisch, 2014) instead of the baseline (RR-BLUP with homogeneous marker variances) with all available SNP markers resulted in increases of prediction accuracies for *P. triticina* and hectoliter weight, decreases in prediction accuracies for *B. graminis* and equal prediction accuracies for all other traits (Figures 1, 2). We had hypothesized that RMLA might be beneficial particularly for resistance traits which tend to be oligogenic rather than polygenic and might therefore benefit from the modeling of marker effects with heterogeneous variances. However, the possible advantage in the estimation of more accurate marker effects (Hofheinz and Frisch, 2014) did not translate into greater prediction accuracies for most of the traits we investigated. In most cases, prediction accuracies for genomic prediction with RMLA were either smaller than the corresponding version with RR-BLUP or the same. The overall tendencies (decreases or increases within a particular block-building method) were roughly the same as for RR-BLUP but the deviations from RR-BLUP were greater than those with GVCHAP. Greater r^2 thresholds for the LD-based haplotype blocks were required for RMLA to obtain the same prediction accuracies for plant height as RR-BLUP (Figure 1). It depended on the trait whether the influence of the estimation method, as shown for eucalyptus (Ballesta et al., 2019), or the influence of the block building method on the prediction accuracies was greater (Figures 1, 2). We cannot make a general recommendation for the use of either RR-BLUP or RMLA for the investigated traits.

4.2.2 Multi-allelic GBLUP with GVCHAP

The main difference between RR-BLUP with SelectionTools and GBLUP with GVCHAP is the construction of the genomic relationship matrix G. In SelectionTools, $\mathbf{G}=\mathbf{ZZ}'$ with $\mathbf{Z}$ the re-parametrized design matrix. $\mathbf{Z}$ is subjected to several transformations, including centering with the allele frequencies, to arrive a realized relationship $\mathbf{G}$ in GVCHAP (Da, 2015). Results for RR-BLUP and GVCHAP were very similar with the exception of the long haplotype blocks built with a fixed number of $n = 50$ or 100 SNPs or with a fixed length of 10 or 20 cM (Figures 1, 2). In these special cases, GVCHAP showed greater prediction accuracies than RR-BLUP even though the prediction accuracies were smaller than for RR-BLUP with 16,667 SNPs. The genomic relationship that is captured by both methods is apparently mostly the same and a difference arises only when the blocks become relatively long (Figures 1, 2, Table 2). These instances were also the only ones in which the ranking of the prediction methods (RR-BLUP, RMLA, GBLUP with GVCHAP) changed. In all other cases, their ranking remained the same over all types of haplotype blocks used for the predictions (Figures 1, 2).

4.3 Conclusions

Prediction accuracies for most traits in our data set were greater when haplotype blocks were used instead of single SNP markers in genomic prediction. The ranking of the block-building methods was trait-dependent, with some methods leading to greater prediction accuracies than single SNPs in some traits and to smaller prediction accuracies in others. The greatest prediction accuracies for resistance scores for *B. graminis*, *P. triticina*, and *F. graminearum* were obtained with LD-based haplotype blocks while blocks with fixed marker numbers and fixed lengths in cM resulted in the greatest prediction accuracies for plant height. Prediction accuracies of haplotype blocks built with the R package HaploBlocker (Pook et al., 2019) were greater than those of the other methods for protein concentration and resistances scores for *S. tritici*, *B. graminis*, and *P. striiformis*. For the resistance scores for *B. graminis*, prediction accuracies were greater than for standard RR-BLUP if marker variances were assumed to be heterogeneous (Hofheinz and Frisch, 2014). Results for multi-allelic prediction with software GVCHAP (Prakapenka et al., 2020) were similar to those from RR-BLUP in most cases. The dependence of prediction accuracies on trait and estimation method was also observed in other studies (Ballesta et al., 2019; Liang et al., 2020; Sallam et al., 2020; Won et al., 2020; Bian et al., 2021). It is important to note that all these studies used different species, blocking methods, and marker effect estimation procedures and do not allow for direct numerical comparison of the results. Nevertheless, their findings support our conclusions that (1) haplotype blocks have the potential to increase the accuracy of genomic prediction in winter wheat, and (2) the choice of the best block-building method is trait-dependent. It is likely that the trait-dependence is caused by properties of the haplotype blocks that have overlapping and contrasting effects on the prediction accuracy. While they might be able to capture local epistatic effects and to detect ancestral relationships better than single SNPs, prediction accuracy might be reduced by unfavorable characteristics of the design matrices in the models that are due to their multi-allelic nature. Additionally, haplotype blocks might be better suited for the prediction of oligogenic than polygenic traits. In oligogenic traits like resistances, they might improve the correct assignment of effects to the underlying genes, while in polygenic traits, the precision of marker effect estimates cannot be improved. Our results suggest that building haplotype blocks allows efficient haplotype stacking for oligogenic resistances in wheat.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/czp-jlu/haploblocks, czp-jlu/haploblocks.

Author contributions

MF, RS, and AS conceived the study. MKo, MKi, LC, MW, and JF collected the field data and genotypic data. AL, AS, BW, and MF evaluated the field data. YD, UO, and CZ-P carried out the genomic prediction. YD and CZ-P wrote the manuscript. SW contributed to writing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The project was funded by the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany *via* the Federal Office for Agriculture and Food (BLE) under the innovation support programme (FKZ 2818403A18).

Acknowledgments

We are grateful to Eva Herzog for her feedback on the figures. We would like to thank the reviewers for their suggestions to improve the manuscript.

Conflict of interest

MKo is employed by Deutsche Saatveredelung AG. MKi is employed by Nordsaat Saatzucht GmbH. LC is employed by W. von Borries-Eckendorf GmbH & Co. KG. MW was employed by German Seed Alliance GmbH and is employed by Saaten-Union Biotec GmbH. JF is employed by Saaten-Union Biotec 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.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2023.1168547/full#supplementary-material

References

Araujo, A. C., Carneiro, P. L. S., Oliveira, H. R., Lewis, R. M., and Brito, L. F. (2022). Snp- and haplotype-based single-step genomic predictions for body weight, wool, and reproductive traits in north American rambouillet sheep. *J. Anim. Breed. Genet.* 00, 1–19. doi: 10.1111/jbg.12748

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

Bian, C., Prakapenka, D., Tan, C., Yang, R., Zhu, D., Guo, X., et al. (2021). Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in duroc pigs. *Genet. Selection Evol.* 53, 1–19. doi: 10.1186/s12711-021-00661-y

Clark, S. A., and van der Werf, J. (2013). "Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values," in *Genome-wide association studies and genomic prediction* (Berlin: Springer), 321–330. doi: 10.1007/978-1-62703-447-0\s\do5(1)3

Coffman, S. M., Hufford, M. B., Andorf, C. M., and Lübberstedt, T. (2020). Haplotype structure in commercial maize breeding programs in relation to key founder lines. *Theor. Appl. Genet.* 133, 547–561. doi: 10.1007/s00122-019-03486-y

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, 1–11. doi: 10.1186/1471-2164-15-1171

Da, Y. (2015). Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genet.* 16, 1–12. doi: 10.1186/s12863-015-0301-1

Da, Y., Liang, Z., and Prakapenka, D. (2022). Multifactorial methods integrating haplotype and epistasis effects for genomic estimation and prediction of quantitative traits. *Front. Genet.* 13. doi: 10.3389/fgene.2022.922369

Edriss, V., Fernando, R. L., Su, G., Lund, M. S., and Guldbrandtsen, B. (2013). The effect of using genealogy-based haplotypes for genomic prediction. *Genet. Selection Evol.* 45, 338–348. doi: 10.1186/1297-9686-45-5

Ferdosi, M. H., Henshall, J., and Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. *Genet. Selection Evol.* 48, 75. doi: 10.1186/s12711-016-0253-6

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* 53, 876–883. doi: 10.1139/G10-076

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, 1–14. doi: 10.1186/s12711-017-0329-y

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 Genet.* 4, 539–546. doi: 10.1534/g3.113.010025

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

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

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 Genet.* 9, 4115–4126. doi: 10.1534/g3.119.400451

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

Nazarian, A., and Gezan, S. A. (2016). GenoMatrix: a software package for pedigree-based and genomic prediction analyses on complex traits. *J. Heredity* 107, 372–379. doi: 10.1093/jhered/esw020

Pook, T., Schlather, M., de Los Campos, G., Mayer, M., Schön, 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

Prakapenka, D., Wang, C., Liang, Z., Bian, C., Tan, C., and Da, Y. (2020). GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00282

Sallam, A. H., Conley, E., Prakapenka, D., Da, Y., and Anderson, J. A. (2020). Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3: Genes Genomes Genet.* 10, 2265–2273. doi: 10.1534/g3.120.401165

Shen, X., Alam, M., Fikse, F., and Rönnegård, L. (2013). A novel generalized ridge regression method for quantitative genetics. *Genetics* 193, 1255–1268. doi: 10.1534/genetics.112.146720

van den Oord, E., and Neale, B. (2004). Will haplotype maps be useful for finding genes? *Mol. Psychiatry* 9, 227–236. doi: 10.1038/sj.mp.4001449

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

Villumsen, T. M., 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

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

Won, S., Park, J.-E., Son, J.-H., Lee, S.-H., Park, B. H., Park, M., et al. (2020). Genomic prediction accuracy using haplotypes defined by size and hierarchical clustering based on linkage disequilibrium. *Front. Genet.* 11. doi: 10.3389/fgene.2020.00134

Zhang, K., Deng, M., Chen, T., Waterman, M. S., and Sun, F. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Natl. Acad. Sci.* 99, 7335–7339. doi: 10.1073/pnas.102186799

Zhao, H., Nettleton, D., Soller, M., and Dekkers, J. (2005). Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. *Genet. Res.* 86, 77–87. doi: 10.1017/S001667230500769X

Chapter 3

Machine learning for prediction of resistance scores in wheat (*Tricicum aestivum L.*) [1]

[1] Heilmann PG, Difabachew YF, Frisch M, Moritz AL, Stahl A, Wittkop B, Snowdon RJ, Koch M, Kirchhoff M, Cselényi L, Wolf M, Förster J, Zenke-Philippi C (2024). Machine learning for prediction of resistance scores in wheat (Triticum aestivum L.). *Plant Breeding*. In Press.
PG Heilmann and YF Difabachew contributed equally to this article.

Plant Breeding

WILEY

Plant Breeding

ORIGINAL ARTICLE OPEN ACCESS

Machine Learning for Prediction of Resistance Scores in Wheat (*Triticum aestivum* L.)

Philipp Georg Heilmann[1] | Yohannes Fekadu Difabachew[1] | Matthias Frisch[1] | Anna Luise Moritz[2] | Andreas Stahl[3] | Benjamin Wittkop[2] | Rod J. Snowdon[2] | Michael Koch[4] | Martin Kirchhoff[5] | László Cselényi[6] | Markus Wolf[7,8] | Jutta Förster[8] | Carola Zenke-Philippi[1]

[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, Gießen, Germany | [2]Institute of Agronomy and Plant Breeding I, Justus Liebig University, Gießen, Germany | [3]Institute for Resistance Research and Stress Tolerance, Julius Kühn Institute, Quedlinburg, Germany | [4]Deutsche Saatveredelung AG, Lippstadt, Germany | [5]Nordsaat Saatzucht GmbH, Langenstein, Germany | [6]W. von Borries-Eckendorf GmbH & Co. KG, Leopoldshöhe, Germany | [7]German Seed Alliance GmbH, Holtsee, Germany | [8]Saaten-Union Biotec GmbH, Leopoldshöhe, Germany

Correspondence: Carola Zenke-Philippi (biometry.popgen@uni-giessen.de)

Received: 7 February 2024 | **Revised:** 12 July 2024 | **Accepted:** 10 October 2024

Funding: This research was supported by the German Federal Ministry of Food and Agriculture, Grant/Award number: FKZ 2818403A18.

Keywords: cross-validation | genomic prediction | machine learning | wheat

ABSTRACT

Machine learning methods were shown to improve the prediction accuracies of genomic prediction of resistance scores compared to methods like RR-BLUP, which were originally designed for metric rather than ordinal response values. We conducted a cross-validation study with 361 wheat genotypes evaluated for five fungal diseases. Our objective was to compare the prediction accuracy and the ability to identify the most resistant genotypes of 19 genomic prediction approaches. Each approach consisted of a different combination of prediction method (RR-BLUP, an alternative method with heterogeneous marker variances, Bayesian generalized linear regression with an ordinal response, support vector machine, gradient boosting machine and random forest), predictor (single SNP markers, LD-based haplotype blocks, 250 variables generated with an autoencoder and SNPs identified with incremental feature selection) and response value (untransformed and logit-transformed resistance scores). In our dataset, RR-BLUP was consistently among the methods with the largest prediction accuracies and the best abilities to identify resistant genotypes in four of five investigated traits. However, in *P. triticina*, using gradient boosting machine and random forest instead of RR-BLUP increased the prediction accuracy from 0.64 to 0.71, indicating that machine learning methods may have an advantage over linear models in genomic prediction. We also found that even though there was a positive correlation between the prediction accuracy and Cohen's κ, a measure to judge how well the most resistant genotypes can be identified, the correlation is not perfect and a large value for the prediction accuracy does not necessarily translate into an equally large κ value.

1 | Introduction

In the last two decades, genomic prediction (Meuwissen, Hayes, and Goddard 2001), which aims at predicting the phenotypic value of an individual from its genotypic data, has increasingly replaced phenotypic selection. The advantage is that only a part of all the genotypes in the breeding population have to be phenotyped or, even better, that phenotypic data that are already

Abbreviations: BGLR, Bayesian generalized linear regression; GBLUP, genomic best linear unbiased prediction; GBM, gradient boosting machine; GWAS, genome-wide association study; LD, linkage disequilibrium; RF, random forest; RMLA, estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components; RMSE, square root of the mean square error; RR-BLUP, ridge-regression best linear unbiased prediction; SNP, single nucleotide polymorphism; SVM, support vector machine; SVR, support vector machine regression.

Philipp Georg Heilmann and Yohannes Fekadu Difabachew contributed equally to this work.

available can be used for the predictions. For the remaining genotypes, only marker data are needed. This is especially beneficial for the evaluation of resistance traits, which is time consuming and expensive. The genomic prediction approach has three components: (1) the form of the genotypic data that are used as predictors, (2) the type of the response values (metric values, percentages and values on an ordinal or nominal scale) and (3) the statistical model that links predictor and response. All of these components have an influence on the prediction accurary, defined as Pearson's correlation between the observed and predicted phenotypic values.

The last component, the statistical model, is the one that receives the most attention in studies on genomic prediction. Ridge regression best linear unbiased prediction (RR-BLUP) and Bayesian methods are among the standard methods for genomic prediction (Wang et al. 2018). While RR-BLUP assumes homogeneous marker variances, most Bayesian methods (Meuwissen, Hayes, and Goddard 2001) as well as another method called "estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components" (RMLA) (Hofheinz and Frisch 2014) allow for heterogeneous marker variances. This might be a better fit for the oligogenic nature of resistance traits because the effects of some markers may be large while those of most others may be close to zero (Hofheinz and Frisch 2014). RR-BLUP, methods from the Bayesian alphabet, and RMLA were developed for the prediction of metric response values with single SNP markers. A possible alternative for the prediction of ordinal response values is Bayesian generalized linear regression as implemented in the R package BGLR (Pérez and de los Campos 2014). More recently, machine learning methods such as support vector machine (SVM), gradient boosting machine (GBM) and random forest (RF) have been used for genomic prediction of resistance scores (Azodi et al. 2019; John et al. 2022; Jones et al. 2023; Ornella et al. 2012; Ornella et al. 2014; Tomar et al. 2021). Machine learning methods are non-parametric and can be applied to metric or ordinal response values without any assumption on the underlying distribution. They also allow to reframe the prediction problem as a classification in which not the observed or predicted resistance scores of a genotype are used as response values but rather its assignment to the "top" or "flop" class (González-Camacho et al. 2018).

Other attempts to improve the prediction accuracy of genomic prediction address the predictors of the model by using haplotype blocks (Difabachew et al. 2023; Weber et al. 2023) or autoencoder features (Islam et al. 2023) as predictors instead of single SNPs or by using subsets of SNPs determined with feature selection (Heinrich et al. 2023; Li et al. 2018). Haplotype blocks group adjacent SNPs on the chromosomes together based on different criteria such as linkage disequilibrium (LD), a fixed number of markers, a fixed physical or genetic distance on the chromosome, or algorithms that aim to create haplotype block libraries that are as representative of the whole set of markers as possible (Pook et al. 2019). When haplotype blocks are used as input variables in RR-BLUP, they are able to capture local epistatic effects (Jiang, Schmidt, and Reif 2018). Autoencoder features are extracted from the encoding layer of an autoencoder. Autoencoders are unsupervised neural networks, in which the input variables are also the targets of the model output (Goodfellow, Bengio, and Courville 2016). In between the input and the output layers is at least one hidden layer with fewer nodes than input variables. These layers function as a bottleneck where the input variables are mapped to a lower dimensional representation (encoding). The number of dimensions can be selected by the user. The model then reconstructs the original input variables from this representation (decoding) in the output layer. To minimize the reconstruction loss, the model learns to preserve as much information of the original variables in the hidden layer as possible (Kramer 1991). Once the "optimal" encoding model is found, the encoded data are used as input variables in a genomic prediction model. In a study in rice, this reduction in dimensionality preserved most of the prediction accuracy while it reduced the computation time considerably (Islam et al. 2023). For feature selection, a genome-wide association study is performed to identify markers that are associated with the trait. The optimum number of markers to be used in the prediction is then determined by cross-validation and the final model is fit accordingly. The results on whether feature selection increases the prediction accuracy compared to the full set of SNPs are contradictory (Heinrich et al. 2023; Li et al. 2018).

Apart from ignoring that the response values are not normally distributed and using RR-BLUP or other methods for metric data anyway, researchers have the option to transform the response so that it better fits the normality assumption. The goal here is to avoid potentially biased results when methods that were originally intended for use with normally distributed data are applied to data on an ordinal scale (Montesinos López et al. 2015). Additionally, when marker effects are estimated with methods like RR-BLUP with an additive model, the additivity of effects can lead to genomic estimates of the genotypic value (GEGVs) that are outside of the original scale, that is, smaller than 0 or larger than 9 on a 0–9 scale. The GEGVs then have no direct translation into meaningful resistance scores. The logit transformation addresses both of these issues. It is intended to achieve a normal distribution of the data (Lesaffre, Rizopoulos, and Tsonaka 2007) and shrinks the score values at both ends of the scale so that GEGVs below or above the limits of the scale are avoided.

We designed this study in order to evaluate the potential of machine learning methods for genomic prediction not only for single SNP markers but also for alternative input features, precisely haplotype blocks and autoencoder features and for subsets of SNP markers determined with feature selection. In order to compare these newer methods with established approaches, we also included Bayesian generalized linear regression and the use of logit-transformed response values. In particular, our objectives were to compare (1) the prediction accuracy of different prediction approaches, including machine-learning methods, and (2) the ability of these approaches to identify the genotypes with the smallest resistance scores with a reference scenario (RR-BLUP with single SNP markers) for the prediction of resistance to five different fungal diseases in a panel of 361 German elite winter wheat lines.

2 | Materials and Methods

2.1 | Phenotypic Data

We evaluated the resistances against *Puccinia triticina* (brown rust), *Fusarium graminearum*, *Septoria tritici*, *Blumeria graminis* (mildew) and *Puccinia striiformis* (yellow rust) of 378

elite wheat lines at three locations in Germany (Böhnshausen, Sachsen-Anhalt; Hovedissen, Nordrhein-Westfalen; Leutewitz, Sachsen) in 2020. Resistances were scored on a 1–9 scale in observation plots in one replication at one (*S. tritici*), two (*F. graminearum*, *P. triticina*), or three locations (*B. graminis*, *P. striiformis*). In case there was more than one location, the arithmetic mean of the two or three observations was used as the resistance score. In order to improve the readability of the manuscript, we use only the name of the disease instead of the full term for the trait, for example, "*S. tritici*" instead of "*S. tritici* resistance score."

2.2 | Genotypic Data

All wheat lines were genotyped with the 25k Illumina iSelect SNP array (SGS TraitGenetics, Gatersleben, Germany). All SNP markers with more than two recorded alleles, more than 10% missing values and an expected heterozygosity of $< 5\%$ as well as all individuals with more than 10% missing marker information were excluded from the analysis. As a result, 16,667 SNP markers and 361 genotypes remained for further analysis. Missing marker data were imputed with BEAGLE (Browning, Zhou, and Browning 2018). We used this dataset for all further calculations. There was no population structure in the dataset (Figure S1).

2.3 | Genomic Prediction Methods

We used genomic prediction based on linear models and machine learning algorithms to evaluate genomic prediction accuracy and efficiency for resistance traits. We used RR-BLUP (Meuwissen, Hayes, and Goddard 2001), RMLA (Hofheinz and Frisch 2014) and Bayesian generalized linear regression (BGLR) with an ordinal response (Pérez and de los Campos 2014). RR-BLUP was technically implemented using a transformation to an animal model (Shen et al. 2013). In order to obtain more robust results in case singular design matrices occur during the cross-validations, we used method 2 of Nazarian and Gezan 2016. The method is available in our software package SelectionTools: https://www.uni-giessen.de/de/fbz/fb09/institute/pflbz2/population-genetics/software. RR-BLUP is considered a standard genomic prediction method in plant and animal breeding programs as it provides stable prediction results (Clark and van der Werf 2013; VanRaden 2008) and is therefore, together with single SNP markers as predictors and resistance scores as response values, treated as the reference in this study.

We also used three supervised machine learning algorithms: support vector regression (SVR)/SVM, GBM and RF. Hyperparameter optimization was performed for all algorithms. SVR is a special case of SVM that is used for metric response values (Drucker et al. 1996). We used a radial basis function as the kernel and tuned the `cost`, the error margin (`margin`) and the influence reach of the individual data points (`sigma`). GBM and RF are both based on ensembles of decision trees (Breiman 2001; Friedman 2001). For GBMs, decision trees are trained in a consecutive order, each tree based on the previous one. For RFs, multiple trees are trained in parallel, each based on a different subset of the training data. The final prediction of the RF model is the average of the predictions of all trees. For both algorithms, we tuned the number of trees used by the model (`ntrees`), the random column sampling rate (`mtry`) and the minimum data points required for a split (`min_n`). We manually set a learning rate of 0.001 for GBM. Default settings were used for all other hyperparameters.

As an alternative, we treated the prediction of resistance scores as a classification task. We used SVM and GBM to predict whether a line was resistant, that is, had a resistance score y smaller than or equal to the 10% quantile Q_{10}, or not. For classification, we used a linear kernel for the SVM and only tuned the `cost` and `margin`. Learning rate for GBM was increased to 0.01 and `min_n` was manually set to 1.

We used a two-step procedure to optimize the hyperparameters for SVR, RF and GBM. The procedure was the same for all algorithms, only the hyperparameters changed (Table 1). We used a 5-fold cross-validation based on the training set to evaluate the hyperparameters. The metric used for evaluation was the square root of the mean square error (RMSE). First, we trained 10 models with hyperparameter combinations based on a maximum entropy grid (Kuhn and Frick 2024; Shewry and Wynn 1987). The essential idea of the maximum entropy grid is to sample points (i.e., combinations of hyperparameters) that cover the hyperparameter space as well as possible, which ensures that the grid search explores a broad range of hyperparameter combinations. Since the points are sampled, they vary between replications. The range of the hyperparameters is shown in Table 1. We used the results of the grid search to initialise an iterative Bayesian optimization, training 10 more models (Snoek, Larochelle, and Adams 2012). Based on the error distribution of the initial maximum entropy grid points, a Bayesian optimization approach can sample and test new combinations from the most promising

TABLE 1 | Overview of hyperparameter ranges considered during tuning.

Hyperparameter	Regression	Classification
RF		
ntrees	(200, 1000)	—
mtry	(0.01, 0.33)	—
min_n	(1, 20)	—
GBM		
ntrees	(50, 500)	(500, 2000)
mtry	(0.01, 0.2)	(0.01, 0.8)
min_n	2, 40)	—
SVR/SVM		
cost	(−10, 5)	(−10, 5)
margin	(0, 0.2)	(0, 0.2)
sigma	(−10, 0)	—

Note: Names of the listed hyperparameters correspond to the argument names used in the software.

regions of the hyperparameter space more quickly. The hyperparameter combination of the model with the smallest RMSE was used to train the final model. The optimization of the hyperparameters for classification was performed analogously, except that some of the parameter ranges in the grid were changed and Cohen's κ was used as the evaluation metric.

2.4 | Feature Engineering

In addition to the complete set of SNP markers, we used three alternative sets of predictors. For the first set, we constructed haplotype blocks based on linkage disequilibrium (LD), which can be measured by r^2 (Zhao et al. 2005). Pairwise LD values were calculated for all SNP markers on each chromosome. SNP markers were added to the left or to the right of a haplotype block as long as the average r^2 between all pairs of SNPs within a block was greater than $t = 0.7$. In order to be able to apply RR-BLUP and RMLA to multi-allelic haplotype block data, the design matrix $\mathbf{Z}$ was re-parameterized (Difabachew et al. 2023).

For the second alternative set of predictors, we extracted the outputs of the encoding layer of an autoencoder. Our autoencoder consisted of five fully connected hidden layers. The layers consisted of [4000,1000, 250,1000, 4000] nodes. The input and output layers consisted of as many nodes as there were predictor variables. The output of the centre layer, consisting of 250 nodes, was treated as the encoding and extracted after model training. We used a rectified linear unit activation function in the hidden layers and applied batch normalization to the outputs of all hidden layers except for the encoding layer. Our data consisted only of homozygous inbred lines with no heterozygous markers present after filtering. Therefore, the markers could be encoded in a binary format, represented by 0 and 1. This allowed for the use of a sigmoid activation function in the output layer. We used binary crossentropy as the loss function and Adam as the optimizer (Kingma and Ba 2015) and trained the autoencoder for 100 epochs.

The third alternative set of predictors was determined by feature selection with a RF model based on GWAS (Heinrich et al. 2023). Analogous to the grid search in the hyperparameter optimization, we conducted a 5-fold cross-valiation on the training set. First, a GWAS was conducted and the markers were ranked according to their p values. Next, RF models were trained, starting with only the most important markers and then incrementing the number of markers in a iterative procedure in steps of 50 from 100 to 1000, of 100 from 1001 to 5000 and of 1000 beyond 5000 markers. The number of markers that resulted in the largest prediction accuracy was determined as the optimum number and the marker set was then used to train another RF model on the complete training set in order to predict the phenotypic values in the validation set. Default settings were used in all RF models. The distribution of the number of SNPs selected by the feature selection procedure is shown in Figure S2.

2.5 | Response Values

For the regression approaches, we used either the resistance scores y or the logit-transformed resistance scores $y^* = \text{logit}\left(\frac{y}{10}\right) = \ln\left(\frac{\frac{1}{1-\frac{y}{10}}}{}\right)$ (Lesaffre, Rizopoulos, and Tsonaka 2007) as response values. The division by 10 was necessary because the logit transformation can only be applied to values in the interval (0,1). For the classification methods, the observations y were transformed into two classes: Individuals in the "top" class had a y below or equal to the 10% quantile Q_{10}, and individuals in the "flop" class had a y above the 10% quantile Q_{10}.

2.6 | Prediction Approaches

We define a prediction "approach" as the combination of prediction method, predictor and response values. The name of each approach consists of three elements, divided by a hyphen. The first element is the prediction method: ridge regression BLUP (`RR-BLUP-...`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-...`), Bayesian generalized linear regression with an ordinal response (`BGLR-...`), support vector regression (`SVR-...`), support vector machine (`SVM-...`), gradient boosting machine (`GBM-...`) and random forest (`RF-...`). The predictors can either be SNPs, indicated by `...-SNP-...` as the second element of the approaches, haplotype blocks, indicated by `...-HAP-...`, the autoencoder output, indicated by `...-AEN-...`, or a set of SNP markers determined by feature selection (`...-FS-...`). The last element of each approach is the type of the response value: The use of untransformed values y is indicated by `...-...-0` in the name of the approach, the use of logit-transformed values y^* is indicated by `...-...-1`. Classified values are denoted by `...-...-c`. For example, the approach with the name `SVR-AEN-0` means that a support vector regression was applied on the autoencoder data with the untransformed resistance scores as the response values.

2.7 | Evaluation of the Prediction Approaches

Each prediction approach was evaluated in 200 cross-validation runs. In each of the 200 runs, the dataset was randomly divided into a training set with 289 genotypes (80%) and a validation set with 72 genotypes (20%). The same splits into training and validation set were used for all sets of predictors and algorithms. When predicting ordinal values, the prediction accuracy $r(y, \hat{y})$ was calculated as the correlation between the actual phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set. The predicted logit-transformed resistance scores $\hat{y}^*$ were transformed back to $\hat{y}$ and the prediction accuracy was then calculated as $r(y, \hat{y})$.

Cohen's κ (Cohen 1960; Fielding and Bell 1997) as a measure for the agreement between observed and predicted class can be calculated from the confusion matrix for the class assignment (Table 2) as $\kappa = \frac{P_o - P_e}{1 - P_e}$ with $P_0 = \frac{\text{tp}+\text{tn}}{n}$ (the proportion of agreement between the observed and predicted values) and $P_e = \frac{\text{tp}+\text{fn}}{n} \times \frac{\text{tp}+\text{fp}}{n} + \frac{\text{fp}+\text{tn}}{n} \times \frac{\text{fn}+\text{tn}}{n}$ (the expected agreement by random chance) (Montesinos López, Montesinos López, and Crossa 2022). The values for κ range from -1 to 1 where $\kappa = 1$ for perfect agreement and $\kappa \leq 0$ for agreement only by random

chance (González-Camacho et al. 2018). The assignment of the observed values y to the "top" or "flop" class was based on the 10% quantile Q_{10}. Individuals in the "top" class had a y smaller than or equal to Q_{10}, and individuals in the "flop" class had a y greater than Q_{10}. This assignment led to different numbers of individuals in the "top" and "flop" classes for the different diseases (Table 3). To account for the different numbers n_{top}, an individual was assigned to the "top" class of the predictions $\hat{y}$ if its predicted value $\hat{y}$ was among the n_{top} individuals with the smallest $\hat{y}$ values for this disease and to the "flop" class otherwise. For the classification approaches `SVM-SNP-c` and `GBM-SNP-c`, the observed values in the confusion matrix resulted from the assignment of the genotypes to the "top" and "flop" classes by the algorithm. A "good" prediction can mean that (a) the prediction accuracy is high and (b) a prediction approach is able to correctly identify the genotypes with extreme resistance scores, that is, the ones that are most interesting for selection decisions, which would be reflected in a κ value of at least 0.3 to 0.5 (Kuhn and Johnson 2013).

The efficiency of an algorithm was evaluated as the mean of the computation time required for one cross-validation run.

2.8 | Software and Hardware

We used R 4.2.2 (R Core Team 2022) for all calculations except the autoencoders, which where calculated using Python 3.10 (Van Rossum and Drake 2009). The adjusted entry means of the genotypes were estimated using "ASReml-R 4.1.0.110" (Butler et al. 2017). Haplotype blocks were built and RR-BLUP and RMLA were calculated using the R package "SelectionTools 22.1." BGLR was calculated using "BGLR" version 1.1.0 (Pérez and de los Campos 2014). SVR was calculated using the package "kernlab 0.9-30" (Karatzoglou, Smola, and Hornik 2022). For RF, we used "ranger 0.16.0" (Wright and Ziegler 2017). GBMs were trained using "lightgbm 3.3.5" (Shi et al. 2023). Maximum entropy grids were constructed using "dials 1.2.0" (Kuhn and Frick 2024) and Bayesian optimization was based on "tune 1.2.1" (Kuhn 2024). We used "parsnip 1.2.1" (Kuhn and Vaughan 2024) and "tidymodels 1.2.0" (Kuhn and Wickham 2020) as wrapper packages to access all the machine learning-related packages. Autoencoders were built using "tensorflow 2.10.0" (Abadi et al. 2015). Missing marker data were imputed with "BEAGLE 5.4" (Browning, Zhou, and Browning 2018). "plink 1.90b6.12" (Chang et al. 2015; Purcell and Chang 2018) was used for recoding the data into VCF format and conducting the GWAS for incremental feature selection.

All calculations were performed on four Intel Xeon Platinum processors 8276 (28 × 2.20 GHz) with 1 TB DDR4 RAM each and 112 kernels in total. For the ML methods, a maximum of 50 kernels was used at the same time. Due to technical limitations on the package side, it was not possible to run one iteration of SVR or SVM on multiple threads. To keep comparability between machine learning algorithms, we ran 50 instances of SVR at the same time and divided the runtime by 50. This way, 50 cores could be used for training.

TABLE 2 | Confusion matrix for a classification problem with two classes.

		Predicted values		
		Top	**Flop**	**Σ**
Observed values	Top	`tp`	`fn`	`tp + fn`
	Flop	`fp`	`tn`	`fp + tn`
	Σ	`tp + fp`	`fn + tn`	n

Abbreviations: `fn`, number of false negatives; `fp`, number of false positives; n, total number of individuals; `tn`, number of true negatives; `tp`, number of true positives.

3 | Results

3.1 | Phenotypic Values

The observed resistance scores covered only a part of the available range from 1 to 9. *F. graminearum* had the smallest range with observed scores between 3 and 7. The proportion of individuals assigned to the "top" class ranged from 12% in *S. tritici* to 49% in *P. striiformis* (Table 3). An illustration of the distribution of the phenotypic data in one particular validation set can be found in the Supporting Information (Figure S3).

3.2 | Prediction Accuracy of Different Prediction Approaches

All results presented in this section are shown in Figure 1. The overall level of the prediction accuracy was determined by the trait. The reference prediction approach `RR-BLUP-SNP-0`,

TABLE 3 | Summary statistics for the five resistance scores.

Trait	Min	Q_{10}	$Z = Q_{50}$	Q_{90}	Max	n_{top} ($y \leq Q_{10}$)	n_{flop} ($y > Q_{10}$)
S. tritici	1.00	1.00	2.00	3.00	6.00	42	319
B. graminis	1.00	1.50	2.00	3.50	5.50	119	242
P. triticina	1.00	1.00	2.00	3.75	7.50	66	295
P. striiformis	1.00	1.00	1.33	3.50	6.50	177	184
F. graminearum	3.00	4.00	4.50	5.50	7.00	86	275

Note: The last two columns show how many of the $n = 361$ individuals have a phenotypic value y below/equal to or above the 10% quantile.

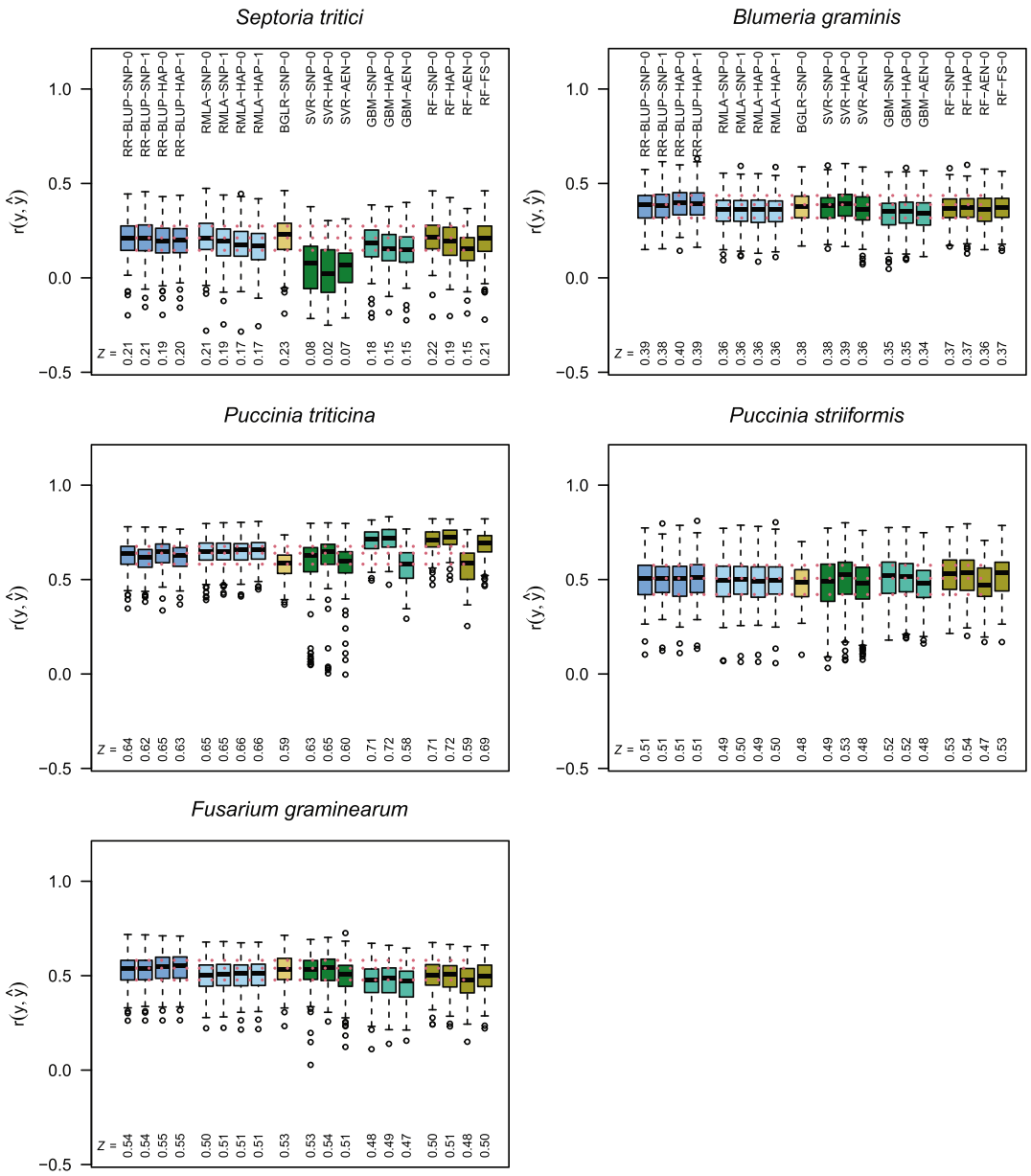

FIGURE 1 | Prediction accuracies for genomic prediction of resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis* and *F. graminearum* with different prediction approaches. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 200 cross-validation runs. Predictions were made with methods ridge regression BLUP (`RR-BLUP-…`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-…`), Bayesian generalized linear regression (`BGLR-…`), support vector regression (`SVR-…`), gradient boosting machine (`GBM-…`) and random forest (`RF-…`). Predictors were either the full set of 16,667 SNP markers (`…-SNP-…`), haplotype blocks based on linkage disquilibrium (`…-HAP-…`), 250 autoencoder features (`…-AEN-…`), or SNP markers identified by feature selection (`…-FS…-`). The response values were either the untransformed resistance scores (`…-...-0`) or the logit-transformed resistance scores (`…-...-1`). Red dotted lines: quartiles from RR-BLUP with 16,667 SNPs (reference). *Z*: median of the correlations $r(y, \hat{y})$ in the 200 cross-validation runs.

RR-BLUP with SNP markers as predictors and the untransformed resistance scores as the response, resulted in medians of $r(y, \hat{y})$ from 0.21 in *S. tritici* to 0.64 in *P. triticina*.

In *S. tritici*, medians of the prediction accuracy ranged from 0.19 to 0.21 in the RR-BLUP approaches and from 0.17 to 0.21 in the RMLA approaches, with the smaller values in the approaches

that used haplotype blocks as predictors. BGLR-SNP-0 had a median of 0.23, the largest value that was observed in this trait. Medians for SVR-SNP-0 and SVR-HAP-0 were 0.08 and 0.02, respectively, while the SVR approach with autoencoder features as predictors, SVR-AEN-0, had a median of 0.07. Medians for the GBM approaches ranged from 0.15 for GBM-HAP-0 to 0.18 for GBM-SNP-0. Medians for the random forest approaches were between 0.15 when autoencoder features were used as predictors (RF-AEN-0) and 0.22 when single SNPs were used instead (RF-SNP-0).

In *B. graminis*, all medians of the correlations $r(y,\hat{y})$ were between 0.34 and 0.40. The largest median, 0.40, was observed with approach RR-BLUP-HAP-0, and the smallest values of 0.34 and 0.35 with the GBM approaches. The medians of the other approaches were in between.

The largest prediction accuracies of all traits were observed in *P. triticina*. The reference approach RR-BLUP-SNP-0 had a median of 0.64, with medians of the other RR-BLUP approaches ranging from 0.62 to 0.65. Medians of the RMLA approaches were 0.65 with untransformed and 0.66 with logit-transformed response values. The median of BGLR-SNP-0 was 0.59. The medians of the SVR approaches ranged from 0.60 for autoencoder features as predictors (SVR-AEN-0) to 0.65 for haplotype blocks (SVR-HAP-0). Medians of the GBM and RF approaches were similar: 0.71 for SNPs as predictors (GBM-SNP-0 and RF-SNP-0), 0.72 for haplotype blocks (GBM-HAP-0 and RF-HAP-0) and 0.58 and 0.59 for autoencoder features (GBM-AEN-0 and RF-AEN-0, respectively). The random forest approach with incremental feature selection (RF-FS-0) was in between with a median of 0.69.

All medians of the prediction accuracies in *P. striiformis* were in the range between 0.47 (for appraoch RF-AEN-0) and 0.53 (approaches SVR-HAP-0, RF-SNP-0 and RF-FS-0). The median of the reference, RR-BLUP-SNP-0, was 0.51 in this case.

In *F. graminearum*, the reference approach RR-BLUP-SNP-0 resulted in a median of the prediction accuracies of 0.54, as did the corresponding approach with haplotype blocks. When logit-transformed response values were used instead, the medians of the prediction accuracies increased to 0.55. RMLA approaches resulted in medians of 0.50 with single SNPs and untransformed response values (RMLA-SNP-0) and 0.51 otherwise. The median of approach BGRL-SNP-0 was 0.53. Among the machine learning methods, the SVR approaches had the largest medians with 0.54 for SVR-HAP-0 and 0.53 for SVR-SNP-0. The smallest medians were observed in the GBM and RF approaches with values of 0.48 for GBM-SNP-0 and RF-AEN-0 and 0.47 for GBM-AEN-0. The remaining RF approaches resulted in medians of 0.50 or 0.51.

3.3 | Identification of the Most Resistant Genotypes

Figure 2 visualizes the results presented in this section. When Cohen's κ was used to evaluate the approaches for how well they were able to identify the most resistant genotypes, the overall level of the κ values was again dependent on the trait.

In *S. tritici*, the reference approach RR-BLUP-SNP-0 had a median of 0.11, as did the corresponding approach with haplotype blocks, RR-BLUP-HAP-0. Using logit-transformed response values led to medians of 0.13 in the RR-BLUP approaches. A similar pattern could be observed in the RMLA approaches, with medians of 0.07 for RMLA-SNP-0 and 0.08 for RMLA-HAP-0 and 0.10 and 0.11 for RMLA-SNP-1 and RMLA-HAP-1, respectively. BGLR-SNP-0 had a median of 0.10. The medians were 0.01 for SVR-SNP-0 and 0.13 and SVR-HAP-0. The use of autoencoder features led to a median of 0.04 in SVR-AEN-0 and the classification approach SVM-SNP-C resulted in a median of 0.02. In the GBM approaches based on regression, the medians ranged between 0.09 for GBM-HAP-0 and 0.13 for GBM-AEN-0. The classification approach GBM-SNP-C had a median of -0.03. The medians of the RF approaches were 0.11 for RF-SNP-0, RF-HAP-0 and RF-FS-0 and 0.10 for RF-AEN-0.

In *B. graminis*, all medians were between 0.21 and 0.23, with the exception of the classification approaches SVM-SNP-C and GBM-SNP-C with median of 0.05 and 0.11, respectively.

In *P. triticina*, the reference RR-BLUP-SNP-0 had a median of 0.16. The medians of the remaining RR-BLUP approaches ranged from 0.15 to 0.18 and showed more variation than the reference. The RMLA approaches resulted in medians of 0.22 for RMLA-SNP-1 and 0.20 for the others. BGRL-SNP-0 had a median of 0.16. The medians of the SVR approaches were either 0.15 or 0.16, with a smaller median of 0.04 for the classification approach SVM-SNP-C. GBM-SNP-0 and GBM-HAP-0 resulted in medians of 0.25. Smaller medians of 0.16 and 0.11 were observed for GBM-AEN-0 and GBM-SNP-C. The pattern for the random forest approaches was similar, with medians of 0.25, 0.28 and 0.15 for approaches RF-SNP-0, RF-HAP-0 and RF-AEN-0, respectively. Approach RF-FS-0 was in between with a median of 0.20.

In *P. striiformis*, the range of the κ values was smaller than for the other traits. The RR-BLUP approaches resulted in medians of 0.30 (for RR-BLUP-SNP-0) to 0.33 (for RR-BLUP-HAP-1). All RMLA approaches had medians of 0.28. The median of BGLR-SNP-0 was 0.31. SVR-HAP-0 had a median of 0.31, compared to medians of 0.28 and 0.25 in the other SVR/SVM approaches. The medians of the GBM approaches were between 0.25 and 0.30. The medians of RF-SNP-0 and RF-HAP-0 were 0.28 and 0.30, respectively, compared to medians of 0.25 for RF-AEN-0 and 0.28 for RF-FS-0.

The overall level of the κ values was highest in *F. graminearum*. All RR-BLUP and RMLA approaches as well as BGLR-SNP-0 had medians of 0.36. The largest median for this trait, 0.38, was observed for approach SVR-SNP-0. The medians of SVR-HAP-0, SVR-AEN-0 and SVM-SNP-C were 0.36, 0.31 and 0.30, respectively. The medians of the GBM regression approaches ranged from 0.26 to 0.31 and the median of the classification approach GBM-SNP-C was 0.24. The medians of the RF approaches ranged from 0.28 to 0.31.

3.4 | Correlation Between r and κ

Figure 3 visualizes the relationship between the prediction accuracy $r(y,\hat{y})$ and Cohen's κ. The means of both measures

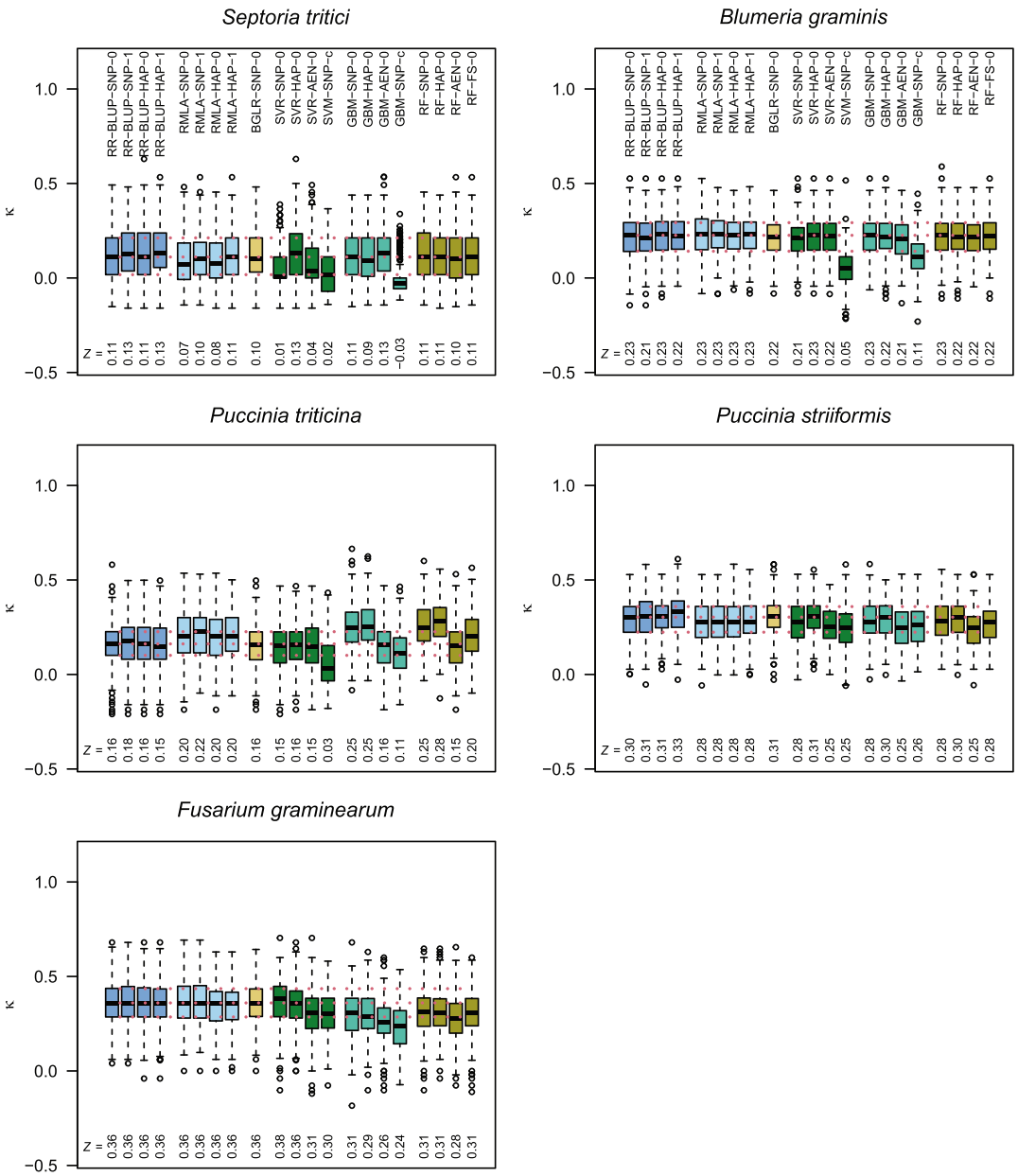

FIGURE 2 | Cohen's κ for genomic prediction of resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis*, and *F. graminearum* with different prediction approaches. The boxplots show the κ values for the agreement between the assignment to the "top" class (y or $\hat{y}$ equal to or below the 10% quantile Q_{10}) and the "flop" class (y or $\hat{y}$ greater than the 10% quantile Q_{10}) in the validation set for 200 cross-validation runs. Predictions were made with methods ridge regression BLUP (`RR-BLUP-…`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-…`), Bayesian generalized linear regression (`BGLR-…`), support vector regression (`SVR-…`), support vector machine (`SVM-…`), gradient boosting machine (`GBM-…`), and random forest (`RF-…`). Predictors were either the full set of 16,667 SNP markers (`…-SNP-…`), haplotype blocks based on linkage disquilibrium (`…-HAP-…`), 250 autoencoder features (`…-AEN-…`), or SNP markers identified by feature selection (`…-FS…-`). The response values were either the untransformed resistance scores (`…-…-0`), the logit-transformed resistance scores (`…-…-1`), or classifications based on the 10% quantile Q_{10} (`…-…-c`). Red dotted lines: quartiles from RR-BLUP with 16,667 SNPs (reference). *Z*: median of the κ values in the 200 cross-validation runs.

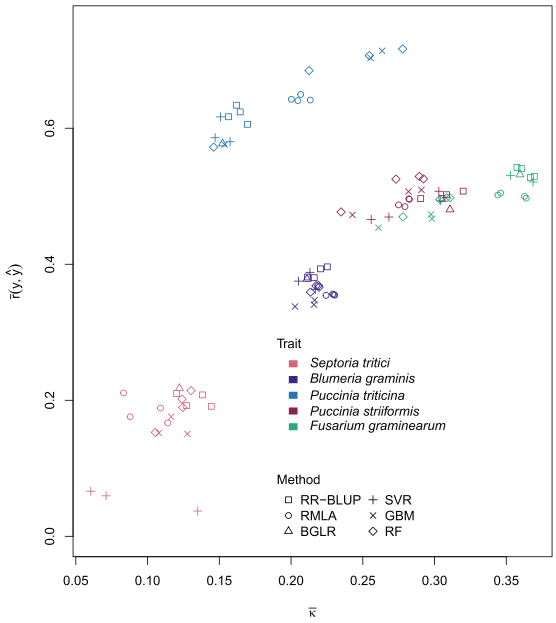

FIGURE 3 | Mean values of correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set and of Cohen's κ for 200 cross-validation runs. Displayed are the values for the resistance scores for *S. tritici*, *B. graminis*, *P. triticina*, *P. striiformis* and *F. graminearum* for different prediction approaches. Predictions were made with methods ridge regression BLUP (`RR-BLUP`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA`), Bayesian generalized linear regression (`BGLR`), support vector regression (`SVR`), gradient boosting machine (`GBM`) and random forest (`RF`). Predictors were either the full set of 16,667 SNP markers, haplotype blocks based on linkage disquilibrium, 250 autoencoder features, or SNP markers identified by feature selection. The phenotypic values used as response values were either the untransformed resistance scores or the logit-transformed resistance scores. Different predictors and response values are not visualized.

showed a correlation across the traits. *P. triticina*, which had the largest mean prediction accuracies of around 0.60, had mean κ values of 0.12 to 0.28. Larger mean κ values of between 0.23 and 0.37 were observed in *P. striiformis* and *F. graminearum* together with smaller mean prediction accuracies of around 0.50. Within the traits, a linear relationship between $r(y, \hat{y})$ and κ could be observed in *P. triticina* and, to a smaller extent, in *P. striiformis* and *F. graminearum*, but not in *S. tritici* and *B. graminis*.

3.5 | Computation Times

The computation times of all approaches can be found in Table 4. RR-BLUP with all 16,667 single SNP markers as predictors was the fastest method with a computation time of 0.68 second per individual cross-validation run on average. However, SVR with autoencoder features was faster when considering the averaged runtime of the parallelization. RMLA and BGLR had a computation time that was about twice as long for the same set of predictors. The computation times of the machine learning methods with single SNP markers were longer with 1.65 min for SVR (averaged), 5.02 min for GBM regression and 3.78 min for RF. When haplotype blocks instead of single SNPs were used, computation times of both RR-BLUP and RMLA increased compared to single SNP markers. An increase of the computation time was also observed for GBM regression, RF and SVR. The use of autoencoder features as predictors in the machine learning methods reduced their computation times to around 1 min or less. The computation time of SVM (averaged) was about 2.87 min and longer compared to that of SVR with single SNP markers. In contrast, the GBM classification took more than twice as long per run as the corresponding regression approach. Computation time was much longer for SVR and SVM compared to other approaches when considering the runtime of the individual cross-validation runs. SVR took 82 min with SNPs and 132 min with haplotype blocks. The autoencoder-based approach (`SVR-AEN-0`) was closer to the other machine learning approaches with 6.36 minutes of individual computation time. SVM took slightly longer than the SNP-based approach `SVR-SNP-0`.

4 | Discussion

4.1 | Prediction Accuracy of Different Prediction Approaches

4.1.1 | Trait

The overall level of the prediction accuracy was determined by the trait (Figure 1). Prediction accuracies between the different approaches varied less for *B. graminis*, *P. striiformis* and *F. graminearum* and more for *S. tritici*, the trait with the smallest overall prediction accuracies with medians around 0.20, and *P. triticina*, the trait with the largest overall prediction accuracies with medians around 0.60 or greater (Figure 1).

The wheat lines in this study are either registered elite varieties or genotypes that are already close to registration. They have therefore been bred for resistance against a variety of pathogens which is reflected in the distribution of the phenotypic values: The observations only cover part of the available scale from 1 to 9 and the larger values, indicating less resistance, are relatively rare (Table 3 and Figure S3). Small prediction accuracies could therefore be at least partially due to the low variation in the response values. In order to obtain reliable results for the genomic predictions, other authors suggest a training set of diverse lines which is continually updated with new breeding material and which can be phenotyped once per season (Juliana et al. 2017).

4.1.2 | Prediction Method

Predictions made with RMLA resulted in similar prediction accuracies as predictions made with RR-BLUP in most cases (Figure 1), even though the genetic architecture of resistance traits is made up of major and minor genes and should, in theory, be captured better by a prediction model like RMLA that allows for heterogeneous marker variances (Hofheinz and Frisch 2014).

TABLE 4 | Computation times in minutes for the different prediction approaches.

Prediction method	Computation time (in minutes)
RR-BLUP-SNP-0	0.68
RR-BLUP-SNP-1	0.68
RR-BLUP-HAP-0	4.97
RR-BLUP-HAP-1	4.74
RMLA-SNP-0	1.45
RMLA-SNP-1	1.45
RMLA-HAP-0	2.10
RMLA-HAP-1	2.10
BGLR-SNP-0	1.33
SVR-SNP-0	1.65 (82.51)
SVR-HAP-0	2.65 (132.53)
SVR-AEN-0	0.13 (6.36)
SVM-SNP-c	2.87 (143.44)
GBM-SNP-0	5.02
GBM-HAP-0	7.2
GBM-AEN-0	0.95
GBM-SNP-c	12.8
RF-SNP-0	3.78
RF-HAP-0	4.89
RF-AEN-0	0.75
RF-FS-0	2.14

Note: The table contains the average values of 200 cross-validation runs for all five traits. For SVR/SVM, due to parallelization, we provide the averaged time per run for 200 cross-validation runs and the time required for a single run in brackets ( ). Predictions were made with methods ridge regression BLUP (RR-BLUP-…), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (RMLA-…), Bayesian generalized linear regression (BGLR-…), support vector regression (SVR-…), support vector machine (SVM-…), gradient boosting machine (GBM-…) and random forest (RF-…). Predictors were either the full set of 16,667 SNP markers (…-SNP-…), haplotype blocks based on linkage disquilibrium (…-HAP-…), 250 autoencoder features (…-AEN-…), or SNP markers identified by feature selection (…-FS…-). The response values were either the untransformed resistance scores (…-…-0), the logit-transformed resistance scores (…-…-1), or classifications based on the 10% quantile Q_{10} (…-…-c).

Other studies on genomic prediction of rust in wheat found that Bayesian methods, which also allow for heterogeneous marker variances, are not necessarily superior to RR-BLUP or genomic BLUP (GBLUP) for the prediction of resistance scores in empirical datasets (Tehseen et al. 2021; Mahmood et al. 2022) even though simulation studies predict that they should be (Meher, Rustgi, and Kumar 2022). A study on both empirical and simulated datasets found the same discrepancy between properties of the methods that should result in better prediction accuracies in theory—and do in simulated datasets—and the actual performance in real-life data (John et al. 2022).

Bayesian generalized linear regression with ordinal response values (approach BGLR-SNP-0) also led to correlations $r(y,\hat{y})$ that were mostly similar to those of RR-BLUP-SNP-0, except for *P. triticina*, in which the values were smaller (Figure 1). This was true regardless of the distribution of the phenotypic values in the validation set. The use of a method specifically designed for ordinal response values therefore did not result in greater prediction accuracies than the use of methods designed for metric response values.

For the machine learning approaches, we did not observe larger prediction accuracies than for the reference approach except for GBM-SNP-0, GBM-HAP-0, RF-SNP-0 and RF-HAP-0 in *P. triticina*. Since we showed in another study that haplotype blocks also led to larger prediction accuracies in this trait compared to single SNPs (Difabachew et al. 2023), we hypothesize that local epistatic effects that can be incorporated by haplotype blocks and machine learning methods, but not by RR-BLUP with single SNPs, may play a role here (Jiang, Schmidt, and Reif 2018; Momen et al. 2018). The prediction accuracies for SVR with single SNP markers (SVR-SNP-0) were generally in the range of those for the corresponding RR-BLUP approach (RR-BLUP-SNP-0), with a difference in the medians of 0.02 at most, except for *S. tritici*. Predictions made with method RF mostly had medians that were 0.01 to 0.04 points greater than those for the corresponding GBM approaches (Figure 1). Only for *P. triticina*, the medians were similar for GBM and RF. Our results partially confirm and partially contradict the results of others. For example, RF resulted in larger prediction accuracies compared to RR-BLUP in the prediction of *P. striiformis* (Tomar et al. 2021) and *F. graminearum* (Rutkoski et al. 2012). In a recent simulation study on genomic prediction with machine learning methods, SVM, RF and GBM showed larger prediction accuracies in a dataset with clear population structure but not in a dataset in which population structure was absent (Jones et al. 2023). The latter corresponds to our dataset (Figure S1), possibly explaining the equal performance of linear and machine learning genomic prediction approaches in four of the five traits in our study. In an extensive study spanning six crops with mostly quantitative traits that compared the prediction accuracy of RR-BLUP, Bayes A and B, Bayesian LASSO, Baysian ridge regression, SVR with linear and nonlinear kernels, gradient tree boosting, artificial neural networks and convolutional neural networks, the results were similar to ours: No single genomic prediction method performed best in all crop/trait combinations, and RR-BLUP was close to the method with the largest prediction accuracy in most cases (Azodi et al. 2019). The same result was found in another study on a simulated animal dataset and three real-life datasets for maize (Lourenço et al. 2024). Our study confirms these findings for resistance traits in wheat.

4.1.3 | Predictor

Replacing single SNP markers with haplotype blocks led to mostly similar prediction accuracies for the corresponding methods, with only small decreases or increases (Figure 1). It has to be noted that there are other possibilities for defining haplotype blocks. In this study, haplotype blocks were built

based on an LD threshold of $r^2 > 0.7$. Other thresholds as well as other methods like building blocks based on a fixed number of markers, fixed window sizes in cM or kilobases on the chromosome, or haplotype block libraries created with the R package HaploBlocker (Pook et al. 2019) are alternative options which have already been investigated in greater detail for this dataset (Difabachew et al. 2023) and others (Weber et al. 2023) and have been shown to increase prediction accuracy in some but not in all cases.

Using autoencoder features as predictors in the machine learning methods resulted in medians of the prediction accuracies that were either similar to or smaller than those of the other approaches, regardless of the method they were used in (Figure 1). Their use led to a reduction in the computation time compared to other predictors for the machine learning methods (Table 4). However, since the computation of the autoencoder features also needs time and the prediction accuracy is generally decreased compared to other predictors, their use as inputs for the machine learning methods was not advantageous in our dataset. More complex studies (Islam et al. 2023) demonstrate the feasibility of preserving prediction accuracy with a reduced set of autoencoder features. We found larger prediction accuracies for GBM and RF than for RR-BLUP with single SNP markers in *P. triticina*, albeit with longer computation times (Table 4). Further research is required to find an easily applicable way to use the autoencoder while maintaining the prediction accuracy and thus save a lot of computation time.

When sets of markers determined by feature selection were used as predictors in a random forest prediction approach (`RF-FS-0`), prediction accuracies were similar to those obtained with the full set of SNPs in nearly all cases (Figure 1), even though the distributions of the numbers of selected SNPs were different between the traits (Figure S2). The findings from other authors in this respect are contradictory: Some found substantial increases with incremental feature selection compared to using the full set of SNPs (Heinrich et al. 2023) while the results of others are similar to ours (Li et al. 2018). We conclude that while feature selection can be beneficial in some cases, further research is needed to determine under which circumstances exactly it can improve the prediction accuracy.

4.1.4 | Response Values

When logit-transformed resistance scores (approaches `RR-BLUP-SNP-1`, `RR-BLUP-HAP-1`, `RMLA-SNP-1` and `RMLA-HAP-1`) were used as response variables instead of the untransformed resistance scores (approaches `RR-BLUP-SNP-0`, `RR-BLUP-HAP-0`, `RMLA-SNP-0` and `RMLA-HAP-0`), differences between the prediction accuracies were small with a maximum of 0.02 points in the medians of the prediction accuracies of the corresponding approaches (Figure 1). We conclude that the logit transformation could successfully address the problem of GEGVs outside the interpretable range and yields predictions with a similar accuracy to those obtained with untransformed data in our dataset. However, it did not improve the predictions by a change in the distribution of the response values. These findings are supported by a study on *P. striiformis* infection in wheat in which the use of logarithmic, boxcox and square root transformations on the observed data did not result in consistent increases in the prediction accuracies obtained with RR-BLUP (Merrick et al. 2022).

4.2 | Identification of the Most Resistant Genotypes

Overall, κ should have a value between 0.3 and 0.5 for acceptable agreement between the classes (Kuhn and Johnson 2013), indicating that an approach is able to identify the most resistant genotypes. We found values in this range only for *F. graminearum*. In the other traits, the κ values were usually smaller.

The patterns for the comparisons between the κ values of the regression approaches in terms of the prediction methods, predictors and response values were the same as for the prediction accuracy (Figure 2). The use of alternative prediction methods, predictors and logit-transformed response values led to medians of the κ values that were either smaller than or similar to the reference approach `RR-BLUP-SNP-0`. The only exception was *P. triticina*, with an increase for GBM and RF from a median of the κ values of 0.16 for `RR-BLUP-SNP-0` to 0.25 for `GBM-SNP-0`, `GBM-HAP-0` and `RF-SNP-0` and 0.28 for `RF-HAP-0`. Autoencoder features as predictors led to smaller κ values in most cases in comparison to `RR-BLUP-SNP-0` (Figure 2). We could not confirm the superiority of SVM for the identification of superior genotypes that was found in 16 wheat datasets (Ornella et al. 2014).

In most studies on genomic prediction, only the prediction accuracy $r(y,\hat{y})$ is reported. However, while a large value for the prediction accuracy indicates that the predictions are accurate on average, this is different from the correct identification of the most resistant genotypes, which are the ones that are interesting for selection. Ideally, a prediction approach would yield both large κ values as well as have a large prediction accuracy. We found a positive correlation between the means of the prediction accuracy $r(y,\hat{y})$ and the means of κ across the traits (Figure 3). Apart from the smaller range of the κ values, these findings are mostly similar to those for rust resistance in wheat (Ornella et al. 2014; González-Camacho et al. 2018). However, both measures must be considered together when the suitability of a method identify superior genotypes is evaluated: In *P. triticina*, the prediction accuracies were largest for all traits, with mean values around 0.6, while the means of the κ values were between 0.12 and 0.28. In contrast, the mean prediction accuracies in *F. graminearum* were around 0.5, but the means of the κ values were all greater than 0.25 (Figure 3). Our findings show that even if κ and $r(y,\hat{y})$ are positively correlated, a large prediction accuracy does not automatically translate into a κ value that is sufficient for the selection of superior genotypes.

4.3 | Summary

A good genomic prediction model is supposed to extract the relevant information from the genotypic data while simultaneously dealing with the noise which comes from other factors. Linear models like RR-BLUP make simplifying assumptions in this situation, particularly when they include only additive effects, like

in our study. The questions then become if there are additional patterns in the genotypic data that cannot be captured by linear models and if machine learning methods are able to find these patterns. In our dataset, RR-BLUP was consistently among the methods with the largest prediction accuracies and the best abilities to identify resistant genotypes in four of the five investigated traits. Compared to machine learning methods, RR-BLUP is implemented in most genomic prediction software. It is easy to apply without the need for hyperparameter tuning and consequently very fast. Additionally, the resulting marker effects are easy to interpret and understand. However, we found substantial increases in the prediction accuracies and κ values compared to the reference approach `RR-BLUP-SNP-0` in *P. triticina*, indicating that investing the additional effort to fine-tune such a method may be worth it. We also found that even though there was a positive correlation between the prediction accuracy and Cohen's κ, a measure to judge how well the most resistant genotypes can be identified, the correlation is not perfect and a large value for the prediction accuracy does not necessarily translate into an equally large κ value. This shows that the prediction accuracy should not be the only measure that is used to select a "good" genomic prediction method.

Author Contributions

Matthias Frisch, Rod Snowdon and Andreas Stahl conceived the study. Michael Koch, Martin Kirchhoff, László Cselényi, Markus Wolf and Jutta Förster collected the field data and genotypic data. Anna Moritz, Andreas Stahl, Benjamin Wittkop and Matthias Frisch evaluated the field data. Yohannes Difabachew carried out the genomic predictions with RR-BLUP, RMLA and BGLR. Philipp Heilmann carried out the genomic predictions with SVR/SVM, GBM and RF. Philipp Heilmann, Yohannes Difabachew and Carola Zenke-Philippi wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgments

The project was funded by the Federal Ministry of Food and Agriculture (BMEL) based on a decision of the Parliament of the Federal Republic of Germany via the Federal Office for Agriculture and Food (BLE) under the innovation support programme (FKZ 2818403A18). Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

Michael Koch is employed by Deutsche Saatveredelung AG. Martin Kirchhoff was employed by Nordsaat Saatzucht GmbH and is employed by Nordzucker AG. László Cselényi is employed by W. von Borries-Eckendorf GmbH & Co. KG. Markus Wolf is employed by German Seed Alliance GmbH. Jutta Förster is employed by Saaten-Union Biotec 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.

Data Availability Statement

The genotypic and phenotypic data as well as the scripts used for this study can be downloaded from https://github.com/czp-jlu/resistance.

References

Abadi, M., A. Agarwal, P. Barham, et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems." https://www.tensorflow.org/. Software available from tensorflow.org.

Azodi, C. B., E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. 2019. "Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits." *G3: Genes, Genomes, Genetics* 9, no. 11: 3691–3702.

Breiman, L. 2001. "Random forests." *Machine Learning* 45: 5–32.

Browning, B. L., Y. Zhou, and S. R. Browning. 2018. "A One-Penny Imputed Genome From Next Generation Reference Panels." *American Journal of Human Genetics* 103: 338–348.

Butler, D. G., B. R. Cullis, A. R. Gilmour, B. G. Gogel, and R. Thompson. 2017. *ASReml-R Reference Manual Version 4*. Hemel Hempstead, HP1 1ES, UK: VSN International Ltd. https://asreml.kb.vsni.co.uk/knowledge-base/asreml_r_documentation/.

Chang, C. C., C. C. Chow, LCAM Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4, no. 1: s13742–015.

Clark, S. A., and J. van der Werf. 2013. "Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values." edited by C. Gondro, J. van der Werf, and B. Hayes, *Genome-Wide Association Studies and Genomic Prediction*. Totowa, NJ: Humana Press, pp. 321–330.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20, no. 1: 37–46.

Difabachew, Y. F., M. Frisch, A. L. Langstroff, et al. 2023. "Genomic Prediction With Haplotype Blocks in Wheat." *Frontiers in Plant Science* 14: 1168547.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. 1996. "Support Vector Regression Machines." edited by M. C. Mozer, M. Jordan, and T. Petsche, *Advances in Neural Information Processing Systems*, Vol. 9. MIT Press, pp. 155–161. https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.

Fielding, A. H., and J. F. Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation* 24, no. 1: 38–49.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29: 1189–1232.

González-Camacho, J. M., L. Ornella, P. Pérez-Rodríguez, D. Gianola, S. Dreisigacker, and J. Crossa. 2018. "Applications of Machine Learning Methods to Genomic Selection in Breeding Wheat for Rust Resistance." *Plant Genome* 11, no. 2: 170104.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Heinrich, F., T. M. Lange, M. Kircher, F. Ramzan, A. O. Schmitt, and M. Gültas. 2023. "Exploring the Potential of Incremental Feature Selection to Improve Genomic Prediction Accuracy." *Genetics Selection Evolution* 55, no. 1: 78.

Hofheinz, N., and M. Frisch. 2014. "Heteroscedastic Ridge Regression Approaches for Genome-Wide Prediction With a Focus on Computational Efficiency and Accurate Effect Estimation." *G3: Genes, Genomes, Genetics* 4, no. 3: 539–546.

Islam, T., C. Kim, H. Iwata, H. Shimono, and A. Kimura. 2023. "DeepCGP: A Deep Learning Method to Compress Genome-Wide Polymorphisms for Predicting Phenotype of Rice." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20, no. 3: 2078–2088.

Jiang, Y., R. H. Schmidt, and J. C. Reif. 2018. "Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers." *G3: Genes, Genomes, Genetics* 8, no. 5: 1687–1699.

John, M., F. Haselbeck, R. Dass, et al. 2022. "A Comparison of Classical and Machine Learning-Based Phenotype Prediction Methods on Simulated Data and Three Plant Species." *Frontiers in Plant Science* 13: 932512.

Jones, D., R. Fornarelli, M. Derbyshire, M. Gibberd, K. Barker, and J. Hane. 2023. "The Pursuit of Genetic Gain in Agricultural Crops Through the Application of Machine-Learning to Genomic Prediction." *Frontiers in Genetics* 14: 1186782.

Juliana, P., R. P. Singh, P. K. Singh, et al. 2017. "Genomic and Pedigree-Based Prediction for Leaf, Stem, and Stripe Rust Resistance in Wheat." *Theoretical and Applied Genetics* 130: 1415–1430.

Karatzoglou, A., A. Smola, and K. Hornik. 2022. "kernlab: Kernel-Based Machine Learning Lab." https://CRAN.R-project.org/package%3Dkernlab. R package version 0.9-30.

Kingma, D. P., and J. Ba. 2015. "Adam: A Method for Stochastic Optimization." In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* edited by Y. Bengio, and Y. LeCun. https://arxiv.org/abs/1412.6980.

Kramer, M. A. 1991. "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks." *AIChE Journal* 37, no. 2: 233–243.

Kuhn, M. 2024. "tune: Tidy Tuning Tools." https://CRAN.R-project.org/package%3Dtune. R package version 1.2.1.

Kuhn, M., and H. Frick. 2024. "dials: Tools for Creating Tuning Parameter Values." https://CRAN.R-project.org/package%3Ddials. R package version 1.2.1.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. New York, NY: Springer.

Kuhn, M., and D. Vaughan. 2024. "parsnip: A Common API to Modeling and Analysis Functions." https://CRAN.R-project.org/package%3Dparsnip. R package version 1.2.1.

Kuhn, M., and H. Wickham. 2020. "tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles." https://www.tidymodels.org.

Lesaffre, E., D. Rizopoulos, and R. Tsonaka. 2007. "The Logistic Transform for Bounded Outcome Scores." *Biostatistics* 8, no. 1: 72–85.

Li, B., N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li. 2018. "Genomic Prediction of Breeding Values Using a Subset of SNPs Identified by Three Machine Learning Methods." *Frontiers in Genetics* 9: 237.

Lourenço, V. M., J. O. Ogutu, R. A. P. Rodrigues, A. Posekany, and H.-P. Piepho. 2024. "Genomic Prediction Using Machine Learning: A Comparison of the Performance of Regularized Regression, Ensemble, Instance-Based and Deep Learning Methods on Synthetic and Empirical Data." *BMC Genomics* 25, no. 1: 152.

Mahmood, Z., M. Ali, J. I. Mirza, et al. 2022. "Genome-Wide Association and Genomic Prediction for Stripe Rust Resistance in Synthetic-Derived Wheats." *Frontiers in Plant Science* 13: 788593.

Meher, P. K., S. Rustgi, and A. Kumar. 2022. "Performance of Bayesian and BLUP Alphabets for Genomic Prediction: Analysis, Comparison and Results." *Heredity* 128, no. 6: 519–530.

Merrick, L. F., D. N. Lozada, X. Chen, and A. H. Carter. 2022. "Classification and Regression Models for Genomic Selection of Skewed Phenotypes: A Case for Disease Resistance in Winter Wheat (Triticum aestivum L.)."*Frontiers in Genetics* 13: 835781.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. "Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps." *Genetics* 157: 1819–1829.

Momen, M., A. A. Mehrgardi, A. Sheikhi, et al. 2018. "Predictive ability of Genome-Assisted Statistical Models Under Various Forms of Gene Action." *Scientific Reports* 8: 12309.

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

Montesinos López, O. A., A. Montesinos López, P. Pérez-Rodríguez, G. de los Campos, K. Eskridge, and J. Crossa. 2015. "Threshold Models for Genome-Enabled Prediction or Ordinal Categorical Traits in Plant Breeding." *G3: Genes, Genomes, Genetics* 5: 291–300.

Nazarian, A., and S. A. Gezan. 2016. "GenoMatrix: A Software Package for Pedigree-Based and Genomic Prediction Analyses on Complex Traits." *Journal of Heredity* 107, no. 4: 372–379.

Ornella, L., P. Pérez, E. Tapia, et al. 2014. "Genomic-Enabled Prediction With Classification Algorithms." *Heredity* 112: 616–626.

Ornella, L., S. Singh, P. Perez, et al. 2012. "Genomic Prediction of Genetic Values for Resistance to Wheat Rusts." *The Plant Genome* 5: 136–148.

Pérez, P., and G. de los Campos. 2014. "Genome-Wide Regression and Prediction With the BGLR Statistical Package." *Genetics* 198, no. 2: 483–495.

Pook, T., M. Schlather, G. de Los Campos, M. Mayer, C. C. Schön, and H. Simianer. 2019. "HaploBlocker: Creation of Subgroup-Specific Haplotype Blocks and Libraries." *Genetics* 212, no. 4: 1045–1061.

Purcell, S., and C. Chang. 2018. "Plink v1.90b6.12." https://www.cog-genomics.org/plink/1.9/.

R Core Team. 2022. "R: A Language and Environment for Statistical Computing." Vienna, Austria. https://www.R-project.org.

Rutkoski, J., J. Benson, Y. Jia, G. Brown-Guedira, J.-L. Jannink, and M. Sorrells. 2012. "Evaluation of Genomic Prediction Methods for Fusarium Head Blight Resistance in Wheat." *Plant Genome* 5: 51–61.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård. 2013. "A Novel Generalized Ridge Regression Method for Quantitative Genetics." *Genetics* 193, no. 4: 1255–1268.

Shewry, M. C., and H. P. Wynn. 1987. "Maximum Entropy Sampling." *Journal of Applied Statistics* 14, no. 2: 165–170.

Shi, Y., G. Ke, D. Soukhavong, et al. 2023. "lightgbm: Light Gradient Boosting Machine." https://CRAN.R-project.org/package%3Dlightgbm. R package version 3.3.5.

Snoek, J., H. Larochelle, and R. P. Adams. 2012. "Practical Bayesian Optimization of Machine Learning Algorithms." edited by F. Pereira, C. J. Burges, L. Bottou, and K. Q. Weinberger, *Advances in Neural Information Processing Systems*, Vol. 25. Curran Associates, Inc., pp. 2951–2959.

Tehseen, M. M., Z. Kehel, C. P. Sansaloni, et al. 2021. "Comparison of Genomic Prediction Methods for Yellow, Stem, and Leaf Rust Resistance in Wheat Landraces From Afghanistan." *Plants* 10: 558.

Tomar, V., G. S. Dhillon, D. Singh, et al. 2021. "Evaluations of Genomic Prediction and Identification of New Loci for Resistance to Stripe Rust Disease in Wheat (Triticum aestivum L.)."*Frontiers in Genetics* 12: 710485.

Van Rossum, G., and F. L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. https://api.semanticscholar.org/CorpusID:61259041.

VanRaden, P. M. 2008. "Efficient Methods to Compute Genomic Predictions." *Journal of Dairy Science* 91, no. 11: 4414–4423.

Wang, X., Y. Xu, Z. Hu, and C. Xu. 2018. "Genomic Selection Methods for Crop Improvement: Current Status and Prospects." *Crop Journal* 6, no. 4: 330–340.

Weber, S. E., M. Frisch, R. J. Snowdon, and K. P. Voss-Fels. 2023. "Haplotype Blocks for Genomic Prediction: A Comparative Evaluation in Multiple Crop Datasets." *Frontiers in Plant Science* 14: 1217589.

Wright, M. N., and A. Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77, no. 1: 1–17.

Zhao, H., D. Nettleton, M. Soller, and J. C. M. Dekkers. 2005. "Evaluation of Linkage Disequilibrium Measures Between Multi-Allelic Markers as Predictors of Linkage Disequilibrium Between Markers and QTL." *Genetics Research* 86, no. 1: 77–87.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Chapter 4

General discussion

4.1 SNP markers vs. alternative predictor sets in genomic prediction

In addition to single SNP markers, genomic prediction can be performed using alternative sets of predictors, such as haplotype blocks (Lin et al. 2024; Weber et al. 2023; Da et al. 2022; Li et al. 2021; Sallam et al. 2020; Cuyabano et al. 2014; Villumsen et al. 2009), autoencoders (Islam et al. 2023), or incremental feature selection (Heinrich et al. 2023; Li et al. 2018). Utilizing these alternative predictors enhances genomic prediction accuracy, and facilitates accurate identification of superior genotypes. Haplotype blocks are able to capture ancestral relationship, local epistasis effects and patterns of LD (Shipilina et al. 2023; Da et al. 2022; Jiang et al. 2018; Barrett et al. 2005; Zhao et al. 2003; Gabriel et al. 2002; Zhang et al. 2002), leading to improved prediction accuracy. Meanwhile, genomic prediction based on autoencoders (Islam et al. 2023) or incremental feature selection (Heinrich et al. 2023; Li et al. 2018) helps to reduce data dimensionality and noise, concentrating predictive models on the most informative genetic features, thereby enhance model performance and prediction accuracy.

4.1.1 Methods of haplotype block building

Haplotype blocks were built with five different approaches using 16,667 SNP markers, each approach comprises different criteria (Table 1.1). For instance, the LD-based methods (LD-AVERAGE and LD-FLANKING) with thresholds ranging from 0.1 to 0.9

and tolerance values of zero or one, comprises 36 different haplotype blocks and each were evaluated for prediction accuracy (Difabachew et al. 2023). The LD threshold determined which SNP markers must be added into the block and the tolerance values depends on the violation of threshold criteria. Genomic prediction accuracy using haplotype blocks is influenced by several factors, such as traits, block-building methods, block size, LD threshold, and the number of assigned and unassigned SNPs (Difabachew et al. 2023). For LD-based methods, both LD-AVERAGE and LD-FLANKING exhibited varying impacts on prediction accuracy based on the trait (Difabachew et al. 2023). The resistance traits *B. graminis* and *P. striiformis* prediction accuracies increased with the LD threshold for haplotype block with LD-AVERAGE, whereas it decreased with LD-FLANKING (Difabachew et al. 2023). LD-FLANKING with a tolerance of one yielded superior accuracy for *P. triticina* compared to single SNPs (Difabachew et al. 2023). In general, as LD threshold values increased, regardless of tolerance, prediction accuracies tended to improve for most traits when using the LD-AVERAGE method; however, the extent of improvement varied by trait (Difabachew et al. 2023). Other studies have indicated that the selection of LD thresholds for constructing haplotype blocks significantly impacts both the number of predictors and prediction accuracy (Da et al. 2022; Li et al. 2021; Cuyabano et al. 2014). Furthermore, population characteristics, such as effective population size and LD patterns, also influence the performance of haplotype-based predictions (Shipilina et al. 2023; Weber et al. 2023).

When evaluating haplotype block-building methods based on a fixed number of SNPs (FIXED-SNP) or fixed chromosome length (FIXED-CM), we found that these methods generally produced lower prediction accuracies compared to single SNPs and LD-based blocks for most traits, except for plant height (Difabachew et al. 2023). This may be due to the adaptive nature of LD-based methods, which capture recombination frequencies and better reflect the genetic architecture of various traits. In contrast, uniform haplotype blocks constructed using FIXED-SNP or FIXED-CM may not align well with recombination hotspots. Other studies (Bian et al. 2021; Sallam et al. 2020; Matias et al. 2017) have demonstrated that the size of SNPs per block or the length of chromosome segments affects prediction accuracy, consistent with our findings that prediction accuracies decreased as the number of SNPs per block or chromosome segment length increased. Consequently, these fixed approaches often fail to capture recombination as effectively as LD-based methods, which rely on data-driven recombination frequencies to define haplotype blocks (Cuyabano et al. 2014; Wall and Pritchard 2003; Gabriel et al.

2002). The improvement in prediction accuracy for plant height with FIXED-SNP and FIXED-CM suggests that certain traits may have genetic architectures better suited to uniform blocks. Additionally, plant height may harbor key loci that benefit from the stability of these fixed methods, though this effect was not observed for other traits in our study.

Haplotype block building with the HaploBlocker algorithms, initially clusters SNP markers and merges them into blocks based on local allelic sequence similarity and other criteria (Table 1.1). Unlike LD-based approaches, HaploBlocker uses linkage rather than LD to define these blocks (Pook et al. 2019). In our study, prediction accuracies increased for resistance traits such as *B. graminis*, *P. striiformis*, and *P. triticina* when HaploBlocker was employed compared to LD-based approaches (Difabachew et al. 2023). In addition, non-overlapping blocking strategies led to improved prediction accuracies for plant height compared to overlapping ones. For *B. graminis* and *P. triticina*, the prediction accuracy was closely linked to block-building strategies, such as initial clustering SNP markers, starting with small number SNP count improves prediction accuracy as compared to larger number of SNP count (Difabachew et al. 2023). These findings are consistent with other studies, which have also demonstrated that HaploBlocker's performance depends on the specific block-building strategies used (Lin et al. 2024; Weber et al. 2023; Da et al. 2022; Pook et al. 2019). Overall, selecting proper haplotype block construction strategies, including number SNP markers in the clustering stages, block overlap or non-overlap, and genomic coverage, are crucial for improving prediction accuracy.

Overall, our findings indicate that apart from traits, genomic prediction accuracy is also influenced by the choice of haplotype block-building methods and the criteria used to built the haplotype blocks. LD-based approaches generally improved prediction accuracy for most traits, particularly resistance scores, due to their ability to capture recombination frequencies. However, for traits like plant height, fixed methods such as FIXED-SNP and FIXED-CM outperformed LD-based methods, likely due to the of uniform block size fits well. Haplotype blocks with HaploBlocker algorithms improved prediction accuracy for certain traits, but its performance varied depending on the block-building strategies, such as initial clustering SNP markers, overlapping versus non-overlapping blocks (Difabachew et al. 2023). These results highlight the importance of optimizing block-building strategies for different traits to achieve the best prediction outcomes.

4.1.2 Feature engineering with Autoencoders

The use of autoencoders as predictors significantly reduces the number of predictors (SNP markers) and is applied exclusively in machine learning algorithms. Autoencoders function by compressing high-dimensional input data into a lower-dimensional latent space and then reconstruct the output features from this compressed representation (Goodfellow et al. 2016; LeCun et al. 2015; Kramer 1991). Studies such as (Wang et al. 2024; Jurado-Ruiz et al. 2023; Islam et al. 2023) have demonstrated the feasibility of using reduced sets of autoencoder-derived features while maintaining prediction accuracy.

In this study, the full SNP markers were compressed from 16,667 to 250 features, with the number of features in each layer selected arbitrarily (Table 1.1). This compression of predictor variables resulted in enhanced computational efficiency (Heilmann et al. 2024). However, across all resistance traits investigated, the median of prediction accuracies and κ derived from autoencoder-based features were comparable to or lower than from full SNP markers or from other predictor sets, regardless of the prediction method used (Heilmann et al. 2024). This suggests that the autoencoder's feature compression process may sometimes result in the loss of essential genetic information, indicating a need for further optimization of autoencoder architectures and training methods (Li et al. 2018). Despite this, the ability of autoencoders to capture essential features for trait variation makes them useful for managing high-dimensional SNP datasets by reducing their size while retaining critical genetic information (Islam et al. 2023; Li et al. 2018).

4.1.3 Incremental feature selection

Previous studies (Jeong et al. 2020; Bermingham et al. 2015) have reported that feature selection approaches, such as ranking SNPs based on GWAS analysis, can lead to increased prediction accuracy. Results from incremental feature selection with GWAS implemented in RF models have demonstrated that using only informative features can improve prediction accuracy and reduce computational time under certain conditions (Heinrich et al. 2023). This approach has shown improved prediction accuracies in maize and soy, though not in switchgrass (Heinrich et al. 2023).

In our study, despite these efforts, feature selection did not significantly enhance prediction accuracies for resistance traits compared to the full set of SNP markers or other

predictor sets, such as haplotype blocks (Heilmann et al. 2024). GWAS identifies SNPs significantly associated with traits, allowing researchers to focus on a subset of markers most likely to influence the traits of interest (Heinrich et al. 2023; Li et al. 2018). This method can improve model interpretability and potentially enhance prediction accuracy by eliminating noise from irrelevant markers (Heinrich et al. 2023; Li et al. 2018).

In conclusion, the comparative analysis of different predictor variables in genomic prediction, such as haplotype blocks, autoencoders, and feature selection reveals diverse outcomes across traits and methodologies. Haplotype blocks, particularly those built using LD-based methods, have shown promise in capturing local genetic variation but have not consistently outperformed traditional models like RR-BLUP. Similarly, feature selection has demonstrated potential in improving prediction accuracy for specific traits, as shown in some studies (Heinrich et al. 2023; Li et al. 2018), but it often performs comparably to full SNP datasets depending on the dataset and traits analyzed. Overall, while each method offers certain advantages, the choice of predictor variables should be tailored to the trait genetic architecture, with careful consideration of the trade-offs between model complexity, computational efficiency, and prediction accuracy.

4.2 Statistical prediction models

In this study, various statistical models were employed for genomic prediction of different types traits (Heilmann et al. 2024; Difabachew et al. 2023). Linear models, such as RR-BLUP (Meuwissen et al. 2001) and GBLUP (Bernardo 1994), assume homogeneous marker variances. The RMLA model (Hofheinz and Frisch 2014) extends RR-BLUP to account for heterogeneous predictor variances, while GVCHAP (Prakapenka et al. 2020; Sallam et al. 2020) builds on GBLUP with haplotype blocks. Additionally, Bayesian models, implemented in the R package "BGLR", were used to manage ordinal resistance scores (Pérez and de los Campos 2014; Gianola 2013).

Machine learning algorithms, including kernel-based methods like SVR and SVM, use hyperplanes as decision boundaries and kernel functions to map input features into high-dimensional spaces for better separation of data points (Drucker et al. 1996). Ensemble methods, such as GBM, where each subsequent model is trained based on the residuals of the previous ones (Friedman 2001), and RF, in which each tree is trained

independently and the results are averaged (Breiman 2001), have also been explored for genomic prediction due to their ability to capture complex patterns and non-linear interactions among genetic markers.

4.2.1 Linear models: Homogeneous vs. heterogeneous variance

Genomic prediction using the RMLA model resulted in increased prediction accuracy compared to the baseline (RR-BLUP with full SNP markers) for certain traits, such as hectoliter weight and *P. triticina*, due to its ability to account for heterogeneous marker variance, particularly for oligogenic traits (Difabachew et al. 2023). However, for most other traits, RMLA showed either similar prediction accuracies or lower performance compared to the baseline model (Heilmann et al. 2024; Difabachew et al. 2023). Other studies on rust resistance in wheat (Tehseen et al. 2021; Tomar et al. 2021), using Bayesian methods ,which is not included in this work, also account for heterogeneous marker variance, reported results that were not superior to other conventional methods such as GBLUP and RR-BLUP. While simulation studies have shown promising results for improving prediction accuracy in traits characterized by a mixture of markers with large effects and others with minimal effects (Meher et al. 2022), practical studies have reported contrasting outcomes (John et al. 2022). Substituting SNP markers with haplotype blocks, constructed using various approaches and employed as predictors in both RR-BLUP and RMLA methods, did not result in higher prediction accuracy compared to RR-BLUP using full SNP markers (Difabachew et al. 2023). Both genomic prediction methods exhibited similar performance patterns across haplotype blocks (Difabachew et al. 2023). This suggests that haplotype block-based approaches, while potentially useful under certain condition, may not always provide significant advantages over single SNP markers in terms of prediction accuracy.

Similarly, GVCHAP, which operates exclusively with haplotype blocks within the GBLUP framework, resulted higher prediction accuracies for longer haplotype blocks, such as when a fixed number of SNPs per block (*e.g.*, 50 or 100 SNPs) or fixed chromosome lengths (*e.g.*, 10 or 20 cM) were used (Difabachew et al. 2023). However, it did not consistently outperform the standard genomic prediction method, RR-BLUP with single SNP markers (Difabachew et al. 2023). Previous studies have highlighted that, while GVCHAP can capture more complex genetic relationships, its performance is often

trait-dependent and varies according to the specific genetic architecture (Da et al. 2022; Prakapenka et al. 2020). These findings suggest that the genetic architecture of the trait may be more critical in determining prediction accuracy than the choice of prediction model.

Genomic prediction using linear models, such as RR-BLUP, RMLA, and GVCHAP with various haplotype blocks constructed based on LD, fixed SNP markers, and fixed chromosome length yielded similar patterns of prediction accuracies for most of the investigated traits (Difabachew et al. 2023). In general, among these frequentist prediction models, RR-BLUP proved to be consistent across multiple traits, confirming its broad applicability and effectiveness in genomic prediction.

Bayesian approaches implemented via the BGLR R package are designed to capture complex genetic architectures by accounting for heterogeneous marker variances across various trait types (continues, binary, count or ordinal) (Pérez and de los Campos 2014). In this study, ordinal Bayesian regression using the BGLR R package with full SNP markers was applied; however, the results for both prediction accuracies and κ were mostly lower than that of the baseline RR-BLUP with full SNP markers, except for the resistance trait *S.tritici* (Heilmann et al. 2024). Other research on resistance traits (Merrick et al. 2022; Montesinos López et al. 2022) has also shown only marginal improvements in prediction accuracy with Bayesian methods compared to the standard GBLUP. While Bayesian models can incorporate prior information and handle complex genetic architectures, their practical performance often remains similar to or even lower than that of GBLUP (John et al. 2022), which aligns with the results we found.

4.2.2 Linear models vs. machine learning algorithms

Machine learning methods, such as SVR, GBM, and RF, when applied to various predictor sets, did not consistently outperform the baseline (RR-BLUP with full SNP) for most traits investigated (Heilmann et al. 2024). However, when ensemble methods (GBM and RF) were employed using full SNP markers and haplotype blocks constructed based on LD, genomic prediction accuracy increased by up to 6% for resistance trait *P. triticina* (Heilmann et al. 2024). This improvement in accuracy is comparable to results from both GVCHAP and RR-BLUP, particularly when haplotype blocks were constructed

using a flanking technique at low LD threshold values (Difabachew et al. 2023). Additionally, RF generally performed slightly better than GBM, with marginally higher median prediction accuracies across various traits, though these differences were not statistically significant (Heilmann et al. 2024). The RF often results greater prediction accuracy compared to other genomic prediction methods, for datasets with a clear population structure, particularly in studies on rust diseases and Fusarium head blight (Tomar et al. 2021; Rutkoski et al. 2012). Despite these promising results for machine learning methods in certain scenarios, our study, along with others (Alemu et al. 2024; Lourenço et al. 2024; Tehseen et al. 2021; Tomar et al. 2021; Azodi et al. 2019), concluded that these methods did not consistently outperform conventional linear models such as RR-BLUP across all traits (Heilmann et al. 2024). This outcome aligns with broader benchmark research across different species, which suggests that no single genomic prediction method excels universally (Alemu et al. 2024; Lourenço et al. 2024; Azodi et al. 2019).

Selecting the appropriate model is crucial in genomic prediction, as no single model performs optimally for all traits and data types. Linear models like RR-BLUP and GBLUP, which assume homogeneous marker variances and primarily capture additive effects, provide consistent accuracy across different predictor sets but struggle with complex genetic architectures. The GVCHAP model, designed for haplotype blocks, enhances prediction accuracy with longer blocks but yields similar results to RR-BLUP for shorter ones. The RMLA extends RR-BLUP to accommodate heterogeneous marker variances, improving predictions for oligogenic traits. Conversely, Bayesian models implemented via the BGLR R package generally do not enhance prediction accuracy compared to RR-BLUP. While machine learning models such as SVR, GBM, and RF can manage complex, non-linear interactions and show slight improvements for some traits, they do not consistently outperform linear models. This study underscores the necessity of tailoring model selection to specific traits and datasets, reinforcing that no single model is best for all scenarios.

4.3 Untransformed vs. transformed response values

The transformation of skewed phenotypic data, such as resistance traits, has been used to stabilize variance and fit the data into a normal distribution (Bartlett 1947). This ap-

proach is essential for ensuring that predicted values remain within a valid scale (Lesaffre et al. 2007). In this study, we compared untransformed and logit-transformed resistance scores using full SNP markers and haplotype blocks. However, no significant improvement in prediction accuracy was observed as a result of transforming phenotypic data, regardless of the prediction methods used, including RR-BLUP and RMLA (Heilmann et al. 2024).

Our results showed that the logit transformation effectively maintained GEGVs within an interpretable range Lesaffre et al. (2007), but the differences in prediction accuracy between the two methods were minimal (Heilmann et al. 2024). This aligns with previous findings by (Merrick et al. 2022; Montesinos López et al. 2022, 2015; Ornella et al. 2014, 2012), where transformations, including logarithmic, failed to consistently improve prediction accuracy for resistance traits. These studies suggested that while transformations may be beneficial for certain skewed traits, their applicability is dependent on the specific characteristics of the data and the genetic architecture of the trait under study.

4.4 Assessment of most resistance genotype: Prediction accuracy vs. Cohen's κ

Prediction accuracy is a key metric for assessing the applicability of genomic prediction for quantitative response values within a regression framework. Evaluation of prediction accuracy is generally depends on correlation between observed and predicted values. An increase in prediction accuracy may indicate that predictions are close to observed values, but it does not necessarily guarantee the identification of most resistant genotypes, which is crucial for breeding program selection. The inclusion of additional metrics, such as Cohen's κ measures the agreement between observed and predicted values on grouping disease resistance or non resistance traits, particularly useful for identifying the best genotypes (Merrick et al. 2022; Montesinos López et al. 2022). It has also been possible to reframe quantitative responses and predicted values using different quantiles, and the results have been presented in a classification methods (Gonzalez-Camacho et al. 2018).

Our comparative analysis of these two metrics (prediction accuracy and Cohen's κ), across various combinations of prediction methods, predictors, and response values,

revealed largely consistent patterns for the investigated resistance traits (Heilmann et al. 2024). Both prediction accuracy and Cohen's κ indicated low mean and median values for the resistance trait against *S. tritici* (Heilmann et al. 2024). On the other hand, prediction accuracy for *P. triticina* exhibited highest mean and median values, while the highest Cohen's κ is recorded for *F. graminearum* (Heilmann et al. 2024). Although both prediction accuracy and Cohen's κ exhibited similar patterns, only *F. graminearum* was found to be within the acceptable agreement range of 0.3 to 0.5 (Kuhn and Johnson 2013).

Ensemble machine learning methods, such as GBM and RF, implemented for full SNP markers and haplotype blocks constructed using the LD method, improved the median Cohen's κ for *P.triticina* by 9% and 11%, respectively, compared to the baseline (Heilmann et al. 2024). For all other combinations of prediction methods, predictors, and response values, the Cohen's κ values were either similar to or below the baseline RR-BLUP. While other studies in wheat (Ornella et al. 2014) were suggested that, SVM performed best for genotype identification, our results did not support this finding.

For robust and reliable selection, both a high Cohen's κ value and high prediction accuracy are ideally required. However, the relationship between these two metrics is not always perfect, as observed in other studies (Gonzalez-Camacho et al. 2018; Ornella et al. 2014). These findings underscore the importance of evaluating genomic prediction models using both prediction accuracy and Cohen's κ values. Focusing solely on one metric could lead to incomplete conclusions about a model's capacity to identify superior genotypes. Thus, breeding programs should adopt a dual-criteria approach, combining prediction accuracy with classification performance metrics like Cohen's κ. This strategy ensures that the methods employed are effective not only at predicting overall average performance but also at identifying the genotypes most resistant to diseases.

4.5 Conclusions

The conclusions drawn from this comparative genomic prediction study highlight the necessity of tailoring approaches to the genetic architecture of specific traits. Haplotype blocks, especially those constructed using LD-based methods, can improve prediction

accuracy by capturing local genetic variation. However, they do not universally outperform SNP-based models like RR-BLUP. In certain cases, such as predicting *P. triticina* resistance, machine learning models like random forest and gradient boosting machines offered better performance metrics, such as Cohen's κ, despite similar prediction accuracies to RR-BLUP. This indicates that machine learning models can uncover complex trait relationships but should be selected based on trait complexity. Autoencoders and feature selection enhance computational efficiency and reduce noise but vary in effectiveness depending on the trait. While prediction accuracy is an important metric for genomic prediction, it reflects average performance of the genotypes rather than individual potential. Metrics like Cohen's κ provide a clearer picture of model performance based on proper classification. Therefore, selecting the appropriate models and predictors requires balancing model complexity, computational cost, and the genetic nature of traits. Ultimately, this thesis underscores the importance of a trait-specific approach, affirming that no single model is universally optimal across all genomic prediction scenarios.

Chapter 5

Summary

This thesis explores genomic prediction methods in winter wheat, focusing on three main components: (1) different sets of predictors, (2) various statistical prediction models, including machine learning algorithms, and (3) response values. The study aims to compare the prediction accuracies of different haplotype block-building methods and linear models across diverse traits, providing insights into the strengths and limitations of various approaches to genomic prediction. Furthermore, it evaluates the prediction accuracies and Cohen's κ for different combinations of predictors and models, comparing both untransformed and logit-transformed resistance scores with the standard RR-BLUP.

The sets of predictors utilized in this study include SNP markers, haplotype blocks, autoencoders, and subsets of SNPs selected through feature selection based on GWAS. Incorporating haplotype blocks facilitates the capture of local epistatic interactions and ancestral relationships, potentially enhancing prediction accuracy for certain traits compared to individual SNP markers. However, the effectiveness of haplotype blocks varies significantly by trait. Improved results are often observed for oligogenic traits, which exhibit less complex genetic architectures, such as resistance traits. In contrast, traits like yield and plant height, which involve more complex genetic architectures, show that haplotype blocks are less effective than single SNP markers. For polygenetic traits, SNP markers remain effective and efficient in these contexts. Other predictor sets, such as autoencoders and feature selection offer advantages in dimension reduction and computational efficiency.

Various statistical models, ranging from traditional linear models like GBLUP to more flexible machine learning algorithms such as SVR, RF, and GBM, are employed

to predict complex traits. Linear models, particularly RR-BLUP and GVCHAP implemented with GBLUP, consistently deliver accurate predictions for traits governed by multiple SNP markers spread across the genome. However, their performance diminishes when predicting traits characterized by complex non-linear interactions, where machine learning algorithms tend to excel. In particular, models such as RF and GBM demonstrated improvements in the prediction of traits like resistance to certain fungal diseases, although they do not consistently surpass RR-BLUP method across all traits.

The study also evaluates the prediction accuracies and Cohen's κ for Bayesian ordinal regression, implemented via the BGLR R package, which theoretically offers greater flexibility by incorporating prior distributions. While methods involving heterogeneous marker variances are hypothesized to improve prediction accuracies for oligogenic traits. However, empirical results indicate that prediction methods such as RMLA and Bayesian ordinal regression do not consistently outperform the RR-BLUP method.

A key finding of this research is that no single model or predictor set performs best for all traits in genomic prediction. The choice of model and predictors must be tailored to the specific genetic architecture of the trait being predicted. Traditional linear models, such as GBLUP and RR-BLUP, perform well with polygenic traits, while machine learning algorithms are more effective for oligogenic traits. Logit-transformation is particularly beneficial for maintaining predicted resistance scores, ensuring that the GEGVs stay within an interpretable range. Furthermore, dual evaluation approaches enhance both the reliability and acceptability of methods for identifying resistance genotypes in breeding programs.

In conclusion, this thesis highlights the importance of a tailored, trait-specific approach to genomic prediction. By evaluating different predictor sets, models, and traits, it underscores the necessity of choosing models based on the characteristics of the data and the genetic architecture of the trait, reinforcing the idea that no single approach is universally superior in the realm of genomic prediction.

Chapter 6

Zusammenfassung

Diese Arbeit untersucht genomische Vorhersagemethoden bei Winterweizen und konzentriert sich dabei auf drei Hauptkomponenten: (1) verschiedene Sätze von Prädiktoren, (2) verschiedene statistische Vorhersagemodelle, einschließlich Algorithmen für maschinelles Lernen, und (3) Antwortwerte. Die Studie zielt darauf ab, die Vorhersagegenauigkeit verschiedener Haplotyp-Blockbildungsmethoden und linearer Modelle für verschiedene Merkmale zu vergleichen und Einblicke in die Stärken und Grenzen verschiedener Ansätze zur genomischen Vorhersage zu geben. Darüber hinaus werden die Vorhersagegenauigkeiten und Cohen's κ für verschiedene Kombinationen von Prädiktoren und Modellen, wobei sowohl untransformierte als auch logit-transformierte Resistenzwerte mit dem Standard-RR-BLUP verglichen werden.

Zu den in dieser Studie verwendeten Prädiktoren gehören SNP-Marker, Haplotyp-Blöcke, Autocoder und Untergruppen von SNPs, die durch Merkmalsauswahl auf der Grundlage von GWAS ausgewählt wurden. Die Einbeziehung von Haplotyp-Blöcken erleichtert die Erfassung lokaler epistatischer Interaktionen und angestammter Beziehungen, wodurch die Vorhersagegenauigkeit für bestimmte Merkmale im Vergleich zu einzelnen SNP-Markern verbessert werden kann. Die Wirksamkeit von Haplotyp-Blöcken ist jedoch je nach Merkmal sehr unterschiedlich. Bessere Ergebnisse werden häufig bei oligogenen Merkmalen beobachtet, die eine weniger komplexe genetische Architektur aufweisen, wie $z.B.$ Resistenzmerkmale. Im Gegensatz dazu zeigen Merkmale wie Ertrag und Pflanzenhöhe, die eine komplexere genetische Architektur aufweisen, dass Haplotyp-Blöcke weniger effektiv sind als einzelne SNP-Marker. Bei polygenetischen Merkmalen bleiben SNP-Marker in diesen Zusammenhängen effektiv und effizient. Andere Prädiktorensätze bieten Vorteile in Bezug auf Recheneffizienz und Dimensionalitätsreduktion.

Verschiedene statistische Modelle, die von traditionellen linearen Modellen wie GBLUP bis hin zu flexibleren Algorithmen des maschinellen Lernens wie SVR, RF und GBM reichen, werden zur Vorhersage komplexer Merkmale eingesetzt. Lineare Modelle, insbesondere RR-BLUP und GVCHAP, die mit GBLUP implementiert wurden, liefern durchweg genaue Vorhersagen für Merkmale, die durch mehrere über das Genom verteilte SNP-Marker bestimmt werden. Ihre Leistung nimmt jedoch ab, wenn Merkmale vorhergesagt werden, die durch komplexe nichtlineare Interaktionen gekennzeichnet sind, bei denen Algorithmen des maschinellen Lernens tendenziell besser abschneiden. Insbesondere Modelle wie RF und GBM zeigten Verbesserungen bei der Vorhersage von Merkmalen wie Resistenz gegen bestimmte Pilzkrankheiten, obwohl sie die RR-BLUP-Methode nicht durchgängig bei allen Merkmalen übertreffen.

In der Studie werden auch die Vorhersagegenauigkeit und Cohens κ für die Bayes'sche ordinale Regression bewertet, die mit dem R-Paket BGLR implementiert wurde, das theoretisch eine größere Flexibilität durch die Einbeziehung von Vorverteilungen bietet. Es wird angenommen, dass Methoden, die heterogene Markervarianzen einbeziehen, die Vorhersagegenauigkeit für oligogene Merkmale verbessern. Empirische Ergebnisse zeigen jedoch, dass Vorhersagemethoden wie RMLA und Bayes'sche ordinale Regression die RR-BLUP-Methode nicht durchgängig übertreffen.

Ein zentrales Ergebnis dieser Forschung ist, dass kein einzelnes Modell oder Prädiktorset bei der genomischen Vorhersage für alle Merkmale am besten abschneidet. Die Wahl des Modells und der Prädiktoren muss auf die spezifische genetische Architektur des vorauszusagenden Merkmals zugeschnitten sein. Herkömmliche lineare Modelle wie GBLUP und RR-BLUP eignen sich gut für polygene Merkmale, während Algorithmen des maschinellen Lernens bei oligogenen Merkmalen effektiver sind. Die Logit-Transformation ist besonders vorteilhaft für die Beibehaltung der vorhergesagten Resistenzwerte und stellt sicher, dass die GEGVs innerhalb eines interpretierbaren Bereichs bleiben. Darüber hinaus verbessern duale Bewertungsansätze sowohl die Zuverlässigkeit als auch die Akzeptanz von Methoden zur Identifizierung von Resistenzgenotypen in Zuchtprogrammen.

Abschließend unterstreicht diese Arbeit die Bedeutung eines maßgeschneiderten, merkmalspezifischen Ansatzes für die genomische Vorhersage. Durch die Bewertung

verschiedener Prädiktorensätze, Modelle und Merkmale wird die Notwendigkeit unterstrichen, Modelle auf der Grundlage der Merkmale der Daten und der genetischen Architektur des Merkmals auszuwählen, was die Idee unterstreicht, dass kein einzelner Ansatz im Bereich der genomischen Vorhersage universell überlegen ist.

Chapter 7

Literature

R. Abdollahi-Arpanahi, D. Gianola, and F. Peñagaricano. Deep learning versus parametric and ensemble methods for genomic prediction of complex phenotypes. *Genetics Selection Evolution*, 52:1–15, 2020.

N. Ahmadi and J. Bartholomé. *Genomic Prediction of Complex Traits*. Springer, 2022.

A. Alemu, J. Åstrand, O. A. Montesinos-Lopez, J. I. y Sanchez, J. Fernandez-Gonzalez, W. Tadesse, R. R. Vetukuri, A. S. Carlsson, A. Ceplitis, J. Crossa, et al. Genomic selection in plant breeding: Key factors shaping two decades of progress. *Molecular Plant*, 2024.

C. B. Azodi, E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. Benchmarking parametric and machine learning models for genomic prediction of complex traits. *G3: Genes, Genomes, Genetics*, 9(11):3691–3702, 11 2019. doi: 10.1534/g3.119.400498.

J. C. Barrett, B. Fry, J. Maller, and M. J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

M. S. Bartlett. The use of transformations. *Biometrics*, 3(1):39–52, 1947. doi: 10.2307/3001536.

M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, et al. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5(1):10312, 2015.

R. Bernardo. Prediction of maize single-cross performance using rflps and information from related hybrids. *Crop Science*, 34(1):20–25, 1994.

R. Bernardo. *Breeding for Quantitative Traits in Plants*. Stemma Press, 2010.

R. Bernardo and J. Yu. Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47(3):1082–1090, 2007.

C. Bian, D. Prakapenka, C. Tan, R. Yang, D. Zhu, X. Guo, D. Liu, G. Cai, Y. Li, Z. Liang, et al. Haplotype genomic prediction of phenotypic values based on chromosome distance and gene boundaries using low-coverage sequencing in Duroc pigs. *Genetics Selection Evolution*, 53(1):1–19, 2021. doi: 10.1186/s12711-021-00661-y.

D. Botstein, R. L. White, M. Skolnick, and R. W. Davis. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics*, 32(3):314, 1980.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 10 2001. ISSN 08856125. doi: 10.1023/A:1010933404324.

S. A. Clark and J. v. d. Werf. Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. pages 321–330, 2013. doi: 10.1007/978-1-62703-447-0_13.

B. C. Collard and D. J. Mackill. Marker-assisted selection: An approach for precision plant breeding in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491):557–572, 2008.

C. Cortes and V. Vapnik. Support-vector network-. machine learning 20: 273–297. *Portfolio Selection, Journal of Global Optimization*, 43(2-3), 1995.

J. Crossa, G. de los Campos, P. Pérez-Rodríguez, D. Gianola, J. Burgueño, J. L. Araus, D. Makumbi, R. P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger, and H.-J. Braun. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, 186(2):713–724, 2010. doi: 10.1534/genetics.110.118521.

J. Crossa, P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho. Genomic selection and prediction in plant breeding. *Journal of Crop Improvement*, 25 (3):239–261, 2011.

J. Crossa, P. Perez, J. Hickey, J. Burgueno, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, et al. Genomic prediction in cimmyt maize and wheat breeding programs. *Heredity*, 112(1):48–60, 2014.

J. Crossa, P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. De Los Campos, J. Burgueño, J. M. González-Camacho, S. Pérez-Elizalde, Y. Beyene, et al. Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, 22(11):961–975, 2017.

B. C. Cuyabano, G. Su, and M. S. Lund. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics*, 15(1):1–11, 2014. doi: 10.1186/1471-2164-15-1171. URL `https://europepmc.org/articles/PMC4367958`.

Y. Da. Multi-allelic haplotype model based on genetic partition for genomic prediction and variance component estimation using SNP markers. *BMC Genetics*, 16(1):1–12, 2015. doi: 10.1186/s12863-015-0301-1. URL `https://europepmc.org/articles/PMC4683770`.

Y. Da, Z. Liang, and D. Prakapenka. Multifactorial methods integrating haplotype and epistasis effects for genomic estimation and prediction of quantitative traits. *Frontiers in Genetics*, 13, 2022. doi: 10.3389/fgene.2022.922369.

Z. A. Desta and R. Ortiz. Genomic selection: Genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9):592–601, 2014.

Y. F. Difabachew, M. Frisch, A. L. Langstroff, A. Stahl, B. Wittkop, R. J. Snowdon, M. Koch, M. Kirchhoff, L. Cselényi, M. Wolf, et al. Genomic prediction with haplotype blocks in wheat. *Frontiers in Plant Science*, 14:1168547, 2023.

H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. 9, 1996. URL `https://proceedings.neurips.cc`.

J. B. Endelman. Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome*, 4(3):250–255, 2011. doi: 10.3835/plantgenome2011.08.0024.

J. B. Endelman and J.-L. Jannink. Shrinkage estimation of the realized relationship matrix. *G3: Genes, Genomes, Genetics*, 2(11):1405–1413, 2012. doi: 10.1534/g3.112.004259.

D. Falconer and T. Mackay. *Introduction to Quantitative Genetics*. Longman, 1996.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 10 2001. ISSN 0090-5364. doi: 10.1214/aos/1013203451.

M. Frisch. *SelectionTools Reference Manual Version 22.1*. Justus-Liebig-University, Heinrich-Buff-Ring 26 35392 Gießen, 2022. URL `https://population-genetics.uni-giessen.de/ software/`.

S. B. Gabriel, S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

D. Gianola. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, 194(3):573–596, 2013.

A. Gilmour, R. Thompson, and B. Cullis. Average information reml: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, 51: 1440–1450, 1995.

M. Goddard. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, 136:245–257, 2009.

M. Goddard and B. Hayes. Genomic selection. *Journal of Animal Breeding and Genetics*, 124:323–330, 2007.

J. M. Gonzalez-Camacho, G. de los Campos, P. Perez, D. Gianola, J. E. Cairns, G. Mahuku, and J. Crossa. Genomic prediction of drought stress response in wheat breeding populations. *Plant Science*, 276:235–242, 2018.

I. Goodfellow, Y. Bengio, and A. Courville. Deep learning. 2016. URL `http://www.deeplearningbook.org`.

D. Habier, R. L. Fernando, and J. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

L. H. Hartwell, L. Hood, M. L. Goldberg, A. E. Reynolds, L. M. Silver, and R. C. Veres. *Genetics: From Genes to Genomes*. McGraw-Hill Education, 5th edition, 2018.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

B. Hayes and M. Goddard. Genome-wide association and genomic selection in animal breeding. *Genome*, 53:876–883, 2010. doi: 10.1139/G10-065.

B. J. Hayes, P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92(2):433–443, 2009.

E. Heffner, M. Sorrells, and J. Jannink. Genomic selection for crop improvement. *Crop Science*, 49(1):1–12, 2009.

P. G. Heilmann, M. Frisch, A. Abbadi, T. Kox, and E. Herzog. Stacked ensembles on basis of parentage information can predict hybrid performance with an accuracy comparable to marker-based gblup. *Frontiers in Plant Science*, 14:1178902, 2023.

P. G. Heilmann, Y. F. Difabachew, M. Frisch, A. L. Moritz, A. Stahl, B. Wittkop, R. J. Snowdon, M. Koch, M. Kirchhoff, L. Cselényi, M. Wolf, J. Förster, and C. Zenke-Philippi. Machine learning for prediction of resistance scores in wheat (triticum aestivum l.). *Plant Breeding*, 2024. In press.

F. Heinrich, T. M. Lange, M. Kircher, F. Ramzan, A. O. Schmitt, and M. Gültas. Exploring the potential of incremental feature selection to improve genomic prediction accuracy. *Genetics Selection Evolution*, 55(1):78, 2023. doi: 10.1186/s12711-023-00853-8.

C. Henderson. Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21: 309–310, 1950.

C. R. Henderson. Best linear unbiased prediction. *Journal of the American Statistical Association*, 70:423–432, 1975.

N. Heslot, H. Yang, M. E. Sorrells, and J.-L. Jannink. Genomic selection in plant breeding: A comparison of models. *Crop Science*, 54(1):148–160, 2014.

J. M. Hickey, T. Chiurugwi, I. Mackay, and W. Powell. Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature genetics*, 49(9):1297–1303, 2017.

W. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and applied genetics*, 38:226–231, 1968.

N. Hofheinz and M. Frisch. Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. *G3: Genes, Genomes, Genetics*, 4(3):539–546, 2014. doi: 10.1534/g3.113.010025.

T. Islam, C. Kim, H. Iwata, H. Shimono, and A. Kimura. DeepCGP: A deep learning method to compress genome-wide polymorphisms for predicting phenotype of rice. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(03):2078–2088, 2023. doi: 10.1109/TCBB.2022.3231466.

J.-L. Jannink, A. J. Lorenz, and H. Iwata. Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, 9(2):166–177, 2010.

S. Jeong, J.-Y. Kim, and N. Kim. Gmstool: Gwas-based marker selection tool for genomic prediction from genomic data. *Scientific reports*, 10(1):19653, 2020.

Y. Jiang, R. H. Schmidt, and J. C. Reif. Haplotype-based genome-wide prediction models exploit local epistatic interactions among markers. *G3: Genes, Genomes, Genetics*, 8 (5):1687–1699, 2018. doi: 10.1534/g3.117.300548.

M. John, F. Haselbeck, R. Dass, C. Malisi, P. Ricca, C. Dreischer, S. J. Schultheiss, and D. G. Grimm. A comparison of classical and machine learning-based phenotype prediction methods on simulated data and three plant species. *Frontiers in Plant Science*, 13, 2022. doi: 10.3389/fpls.2022.932512.

C. Jones, K. Edwards, S. Castaglione, M. Winfield, F. Sala, C. Van de Wiel, G. Bredemeijer, B. Vosman, M. Matthes, A. Daly, et al. Reproducibility testing of rapd, aflp and ssr markers in plants by a network of european laboratories. *Molecular breeding*, 3:381–390, 1997.

F. Jurado-Ruiz, D. Rousseau, J. A. Botía, and M. J. Aranzana. Genodrawing: an autoencoder framework for image prediction from snp markers. *Plant Phenomics*, 5: 0113, 2023.

M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37(2):233–243, 1991. doi: 10.1002/aic.690370209.

M. Kuhn and H. Frick. dials: Tools for creating tuning parameter values. *R package version*, 1(0), 2024.

M. Kuhn and K. Johnson. *Applied Predictive Modeling*. Springer, New York, NY, 2013. ISBN 978-1-4614-6848-6. doi: 10.1007/978-1-4614-6849-3.

R. Lande and R. Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756, 1990.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

E. Lesaffre, D. Rizopoulos, and R. Tsonaka. The logistic transform for bounded outcome scores. *Statistical Methods in Medical Research*, 16(3):209–230, 2007.

B. Li, N. Zhang, Y.-G. Wang, A. W. George, A. Reverter, and Y. Li. Genomic prediction of breeding values using a subset of snps identified by three machine learning methods. *Frontiers in genetics*, 9:237, 2018.

H. Li, X. Han, J. Xu, and X. Jin. The utility of haplotype-based genomic prediction models for the assessment of complex traits in crops. *G3: Genes, Genomes, Genetics*, 11(4):1–11, 2021. doi: 10.1093/g3journal/jkab085.

Y.-C. Lin, M. Mayer, D. Valle Torres, T. Pook, A. C. Hölker, T. Presterl, M. Ouzunova, and C.-C. Schön. Genomic prediction within and across maize landrace derived populations using haplotypes. *Frontiers in Plant Science*, 15:1351466, 2024.

V. M. Lourenço, J. O. Ogutu, R. A. Rodrigues, A. Posekany, and H.-P. Piepho. Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. *BMC Genomics*, 25(1):152, 2024. doi: 10.1186/s12864-023-09933-x.

M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Sunderland, MA, 1998.

F. I. Matias, G. Galli, I. S. Correia Granato, and R. Fritsche-Neto. Genomic prediction of autogamous and allogamous plants by SNPs and haplotypes. *Crop Science*, 57(6): 2951–2958, 2017. doi: 10.2135/cropsci2017.01.0022.

P. McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

P. Meher, S. Rustgi, and A. Kumar. Performance of Bayesian and BLUP alphabets for genomic prediction: Analysis, comparison and results. *Heredity*, 128, 2022. doi: 10.1038/s41437-022-00539-9.

L. F. Merrick, D. N. Lozada, X. Chen, and A. H. Carter. Classification and regression models for genomic selection of skewed phenotypes: a case for disease resistance in winter wheat (triticum aestivum l.). *Frontiers in Genetics*, 13:835781, 2022.

T. H. Meuwissen, B. J. Hayes, and M. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001. doi: 10.1093/genetics/157.4.1819.

O. A. Montesinos López, A. Montesinos López, P. Pérez-Rodríguez, G. de los Campos, K. Eskridge, and J. Crossa. Threshold models for genome-enabled prediction or ordinal categorical traits in plant breeding. *G3: Genes, Genomes, Genetics*, 5:1599–1609, 2015. doi: 10.1534/g3.115.019401.

O. A. Montesinos López, A. Montesinos López, and J. Crossa. *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer Nature, 2022.

O. A. Montesinos-López, A. R. Bentley, C. Saint Pierre, L. Crespo-Herrera, L. Rebollar-Ruellas, P. E. Valladares-Celis, M. Lillemo, A. Montesinos-López, and J. Crossa. Efficacy of plant breeding using genomic information. *The Plant Genome*, 16(2):e20346, 2023.

L. F. Mota, L. M. Arikawa, S. W. Santos, G. A. Fernandes Júnior, A. A. Alves, G. J. Rosa, M. E. Mercadante, J. N. Cyrillo, R. Carvalheiro, and L. G. Albuquerque. Benchmarking machine learning and parametric methods for genomic prediction of feed efficiency-related traits in nellore cattle. *Scientific Reports*, 14(1):6404, 2024.

S. Myles, J. Peiffer, P. J. Brown, E. S. Ersoz, Z. Zhang, D. E. Costich, and E. S. Buckler. Association mapping: critical considerations shift from genotyping to experimental design. *The Plant Cell*, 21(8):2194–2202, 2009.

L. Ornella, S. Singh, P. Perez, J. Burgueño, R. Singh, E. Tapia, S. Bhavani, S. Dreisigacker, H.-J. Braun, K. Mathews, and J. Crossa. Genomic prediction of genetic values for resistance to wheat rusts. *The Plant Genome*, 5, 2012. doi: 10.3835/plantgenome2012.07.0017.

L. Ornella, P. Pérez, E. Tapia, J. M. González-Camacho, J. Burgueño, X. Zhang, S. Singh, F. S. Vicente, D. Bonnett, S. Dreisigacker, R. Singh, N. Long, and J. Crossa. Genomic-enabled prediction with classification algorithms. *Heredity*, 112, 2014. doi: 10.1038/hdy.2013.144.

T. Pook, M. Schlather, G. de Los Campos, M. Mayer, C. C. Schoen, and H. Simianer. Haploblocker: creation of subgroup-specific haplotype blocks and libraries. *Genetics*, 212(4):1045–1061, 2019.

D. Prakapenka, C. Wang, Z. Liang, C. Bian, C. Tan, and Y. Da. GVCHAP: a computing pipeline for genomic prediction and variance component estimation using haplotypes and SNP markers. *Frontiers in Genetics*, 11:282, 2020. doi: 10.3389/fgene.2020.00282.

P. Pérez and G. de los Campos. Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, 198(2), 2014. doi: 10.1534/genetics.114.164442.

A. Rafalski. Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology*, 5(2):94–100, 2002.

J. A. Rafalski and S. V. Tingey. Genetic diagnostics in plant breeding: Rapds, microsatellites and machines. *Trends in Genetics*, 9(8):275–280, 1993.

G. K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6:15–32, 1991.

J. Rutkoski, J. Benson, Y. Jia, G. Brown-Guedira, J. J-L, and M. Sorrells. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. *The Plant Genome*, 5, 2012. doi: 10.3835/plantgenome2012.02.0001.

A. H. Sallam, E. Conley, D. Prakapenka, Y. Da, and J. A. Anderson. Improving prediction accuracy using multi-allelic haplotype prediction and training population optimization in wheat. *G3: Genes, Genomes, Genetics*, 10(7):2265–2273, 2020.

M. C. Shewry and H. P. Wynn. Maximum entropy sampling. *Journal of applied statistics*, 14(2):165–170, 1987.

D. Shipilina, A. Pal, S. Stankowski, Y. F. Chan, and N. H. Barton. On the origin and structure of haplotype blocks. *Molecular Ecology*, 32(6):1441–1457, 2023.

J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

M. M. Tehseen, Z. Kehel, C. P. Sansaloni, M. d. S. Lopes, A. Amri, E. Kurtulus, and K. Nazari. Comparison of genomic prediction methods for yellow, stem, and leaf rust resistance in wheat landraces from Afghanistan. *Plants*, 10, 2021. doi: 10.3390/plants10030558.

V. Tomar, G. S. Dhillon, D. Singh, R. P. Singh, J. Poland, A. A. Chaudhary, P. K. Bhati, A. K. Joshi, and U. Kumar. Evaluations of genomic prediction and identification of new loci for resistance to stripe rust disease in wheat (triticum aestivum l.). *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.710485.

L. Van Der Maaten, E. O. Postma, H. J. Van Den Herik, et al. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(66-71):13, 2009.

P. M. VanRaden. Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423, 2008.

V. Vapnik. The nature of statistical learning theory springer. *New York*, 10:978–1, 1995.

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

R. K. Varshney, A. Graner, and M. E. Sorrells. Genic microsatellite markers in plants: features and applications. *TRENDS in Biotechnology*, 23(1):48–55, 2005.

A. Vignal, D. Milan, M. SanCristobal, and A. Eggen. A review on snp and other types of molecular markers and their use in animal genetics. *Genetics selection evolution*, 34(3):275–305, 2002.

T. M. Villumsen, L. Janss, and M. S. Lund. The importance of haplotype length and heritability using genomic selection in dairy cattle. *Journal of Animal Breeding and Genetics*, 126(1):3–13, 2009.

P. M. Visscher, W. G. Hill, and N. R. Wray. Heritability in the genomics era—concepts and misconceptions. *Nature Reviews Genetics*, 9(4):255–266, 2008.

P. Vos, R. Hogers, M. Bleeker, M. Reijans, T. v. d. Lee, M. Hornes, A. Friters, J. Pot, J. Paleman, M. Kuiper, et al. Aflp: a new technique for dna fingerprinting. *Nucleic acids research*, 23(21):4407–4414, 1995.

J. D. Wall and J. K. Pritchard. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, 4(8):587–597, 2003.

G. Wang, J. Xuan, P. Wang, C. Li, and J. Lu. Lstm autoencoder-based deep neural networks for barley genotype-to-phenotype prediction. *arXiv preprint arXiv:2407.16709*, 2024.

X. Wang, S. Shi, G. Wang, W. Luo, X. Wei, A. Qiu, F. Luo, and X. Ding. Using machine learning to improve the accuracy of genomic prediction of reproduction traits in pigs. *Journal of Animal Science and Biotechnology*, 13(1):60, 2022.

S. E. Weber, M. Frisch, R. J. Snowdon, and K. P. Voss-Fels. Haplotype blocks for genomic prediction: a comparative evaluation in multiple crop datasets. *Frontiers in Plant Science*, 14:1217589, 2023.

Y. Xu, Y. Lu, C. Xie, S. Gao, J. Wan, and B. M. Prasanna. Whole-genome strategies for marker-assisted plant breeding. *Molecular Breeding*, 33(1):1–11, 2014. doi: 10.1007/s11032-013-9912-6.

S. Z. Yasi Wang, Hongxun Yao. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242, 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2015.08.104. RoLoD: Robust Local Descriptors for Computer Vision 2014.

H. Zhang, L. Yin, M. Wang, X. Yuan, and X. Liu. Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. *Frontiers in genetics*, 10:189, 2019.

K. Zhang, P. Calabrese, M. Nordborg, and F. Sun. Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, 71(6):1386–1394, 2002.

H. Zhao, D. J. Schaid, E. D. Wieben, and W. A. McKinney. Haplotype block structure and its applications to association studies. *Nature Reviews Genetics*, 4(8):599–607, 2003.

Chapter 8

Supplementary files

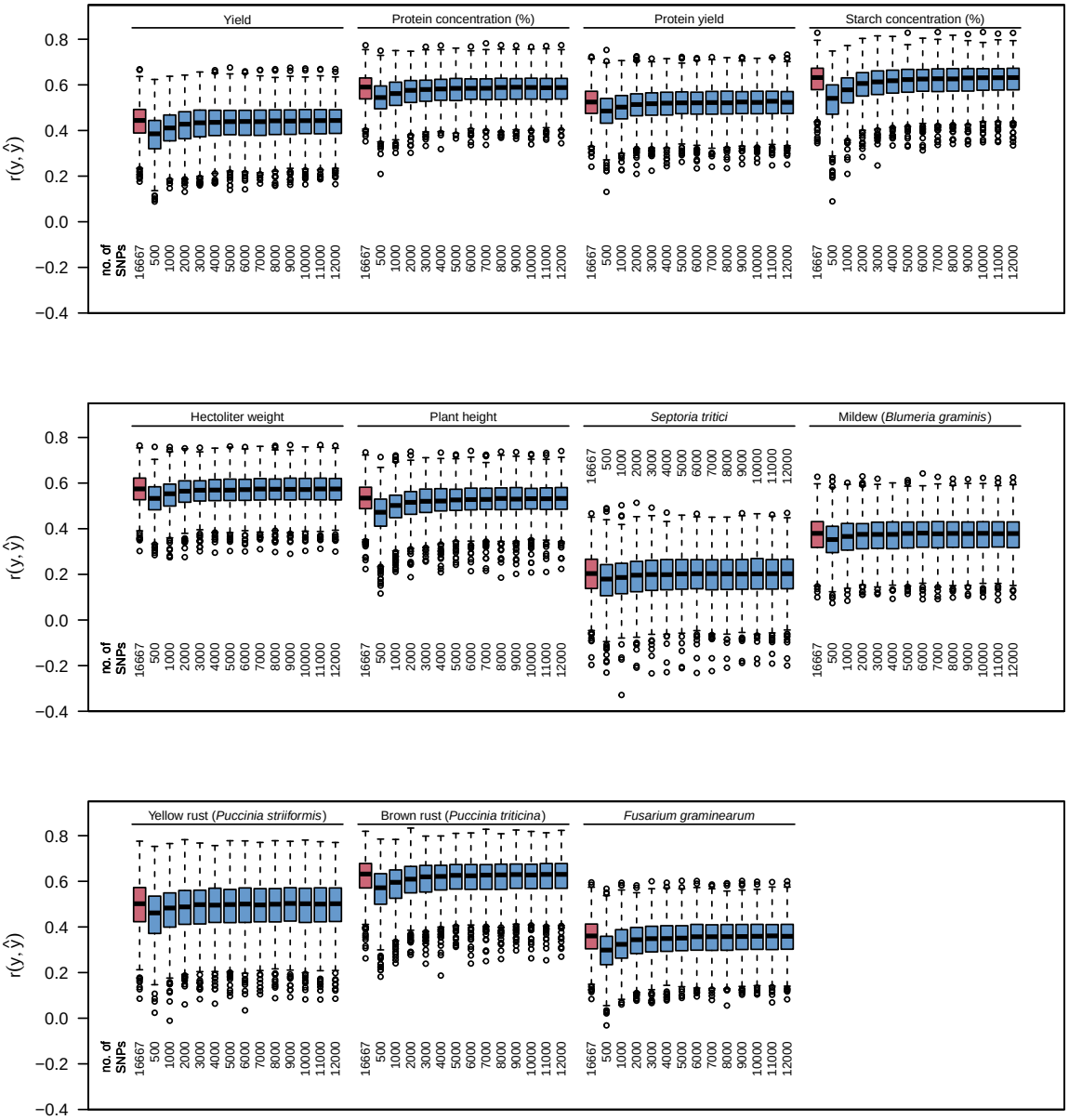

Figure S1. Prediction accuracies for genomic prediction of eleven traits in winter wheat with different numbers of SNP markers. Marker effects were estimated with RR-BLUP. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs.

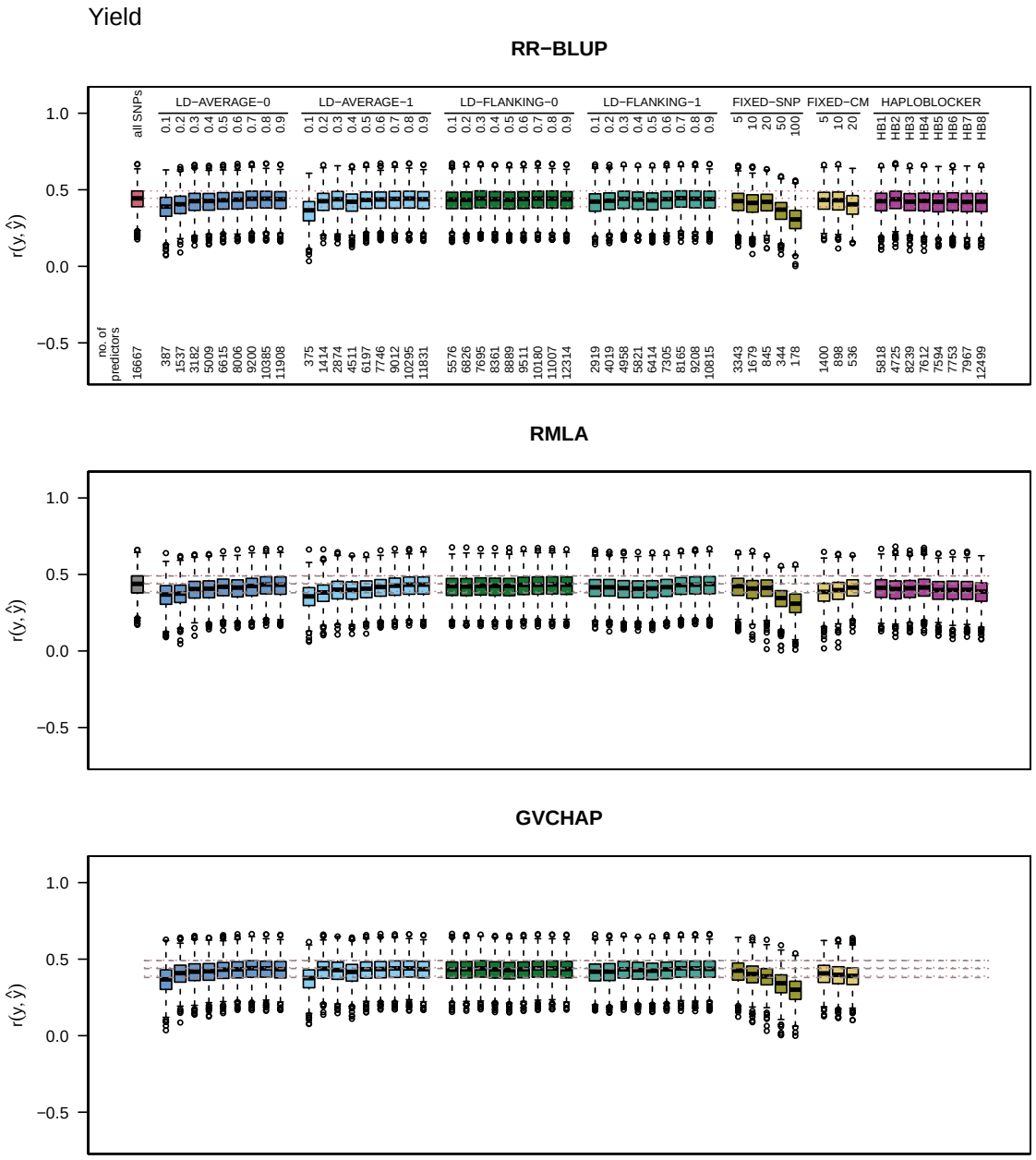

Figure S2. Prediction accuracies for genomic prediction of yield with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

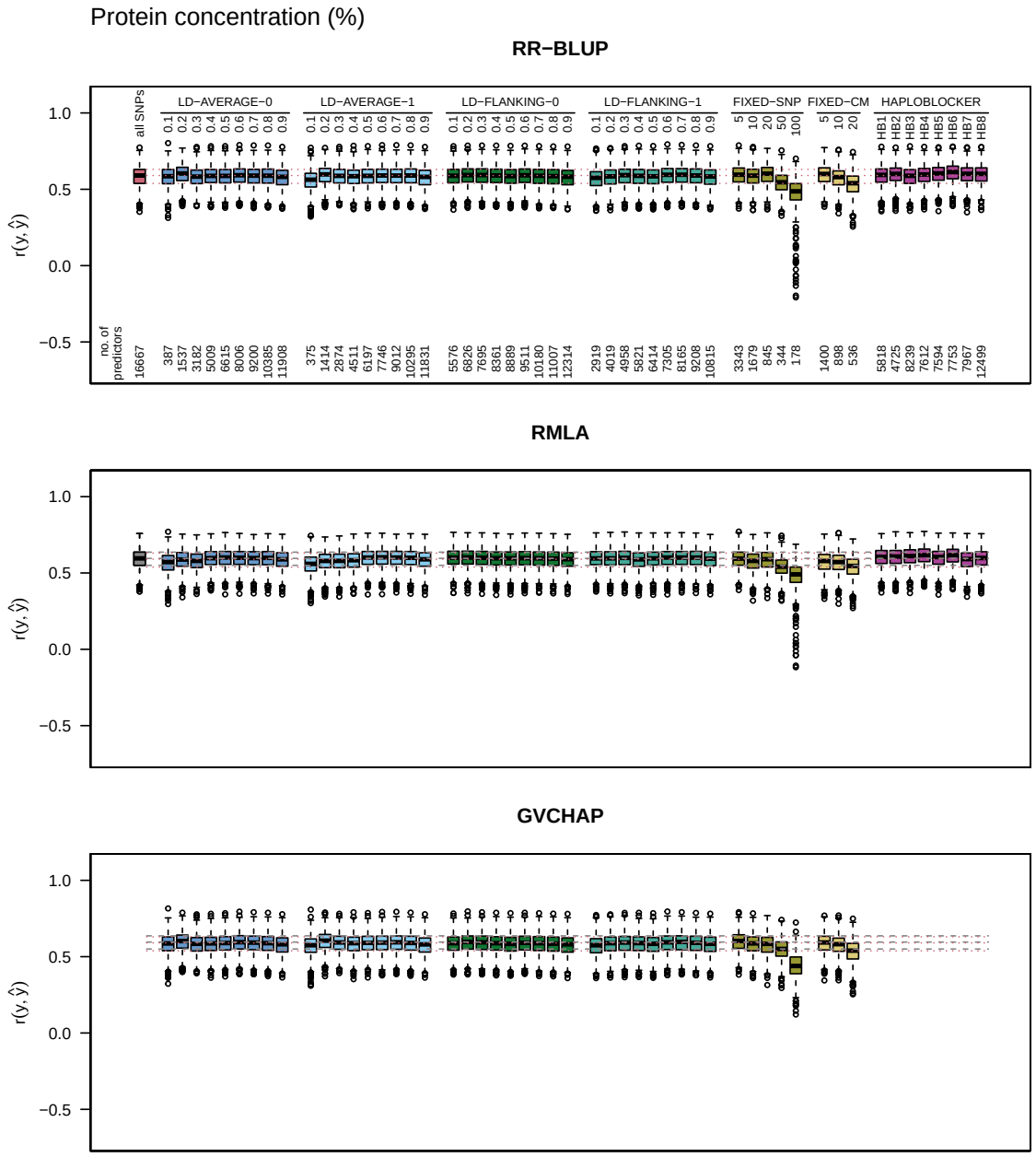

Figure S3. Prediction accuracies for genomic prediction of protein concentration with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD-AVERAGE-0`, `LD-AVERAGE-1`, `LD-FLANKING-0`, `LD-FLANKING-1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED-SNP`), with a fixed block length in cM (`FIXED-CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

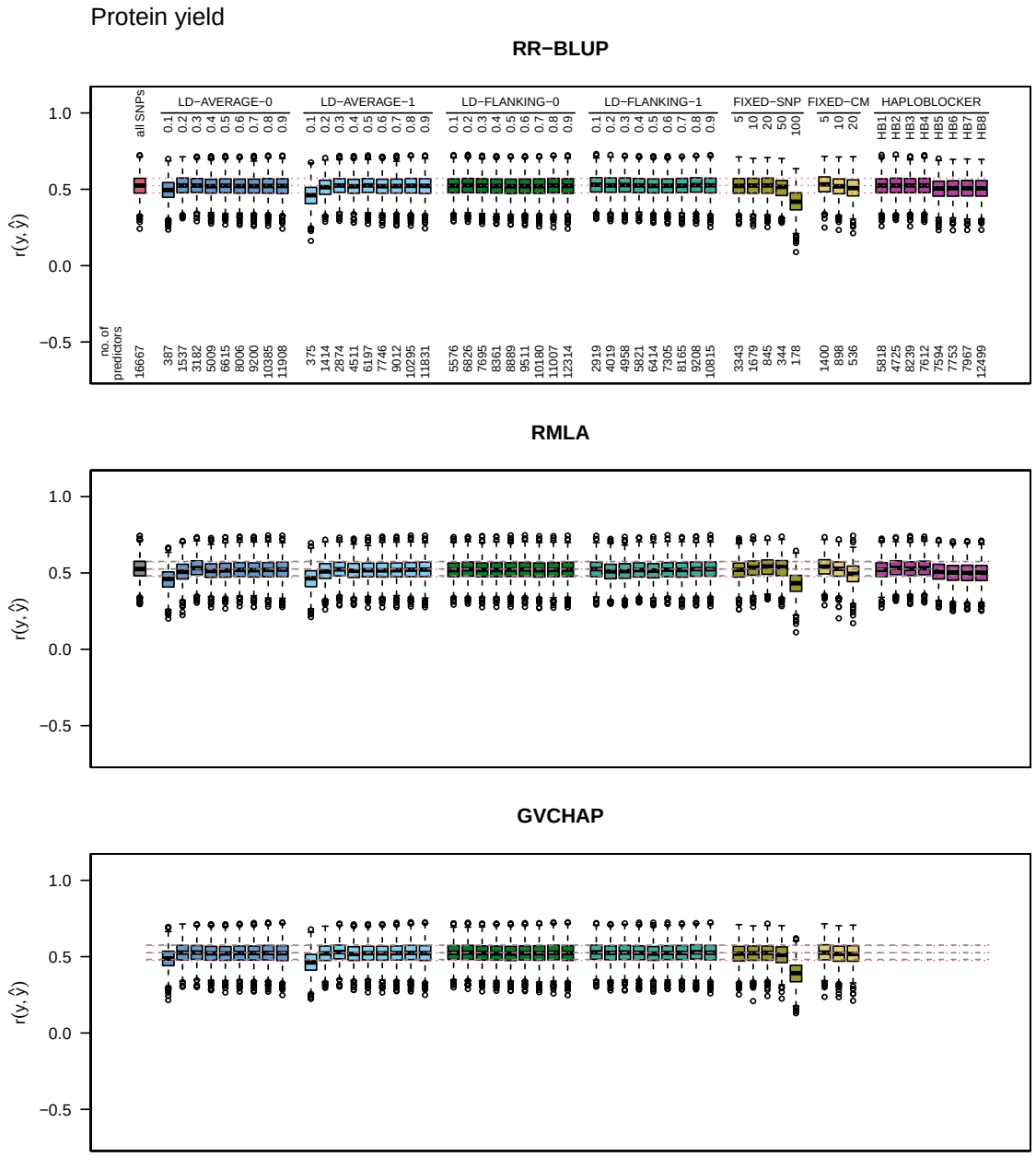

Figure S4. Prediction accuracies for genomic prediction of protein yield with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

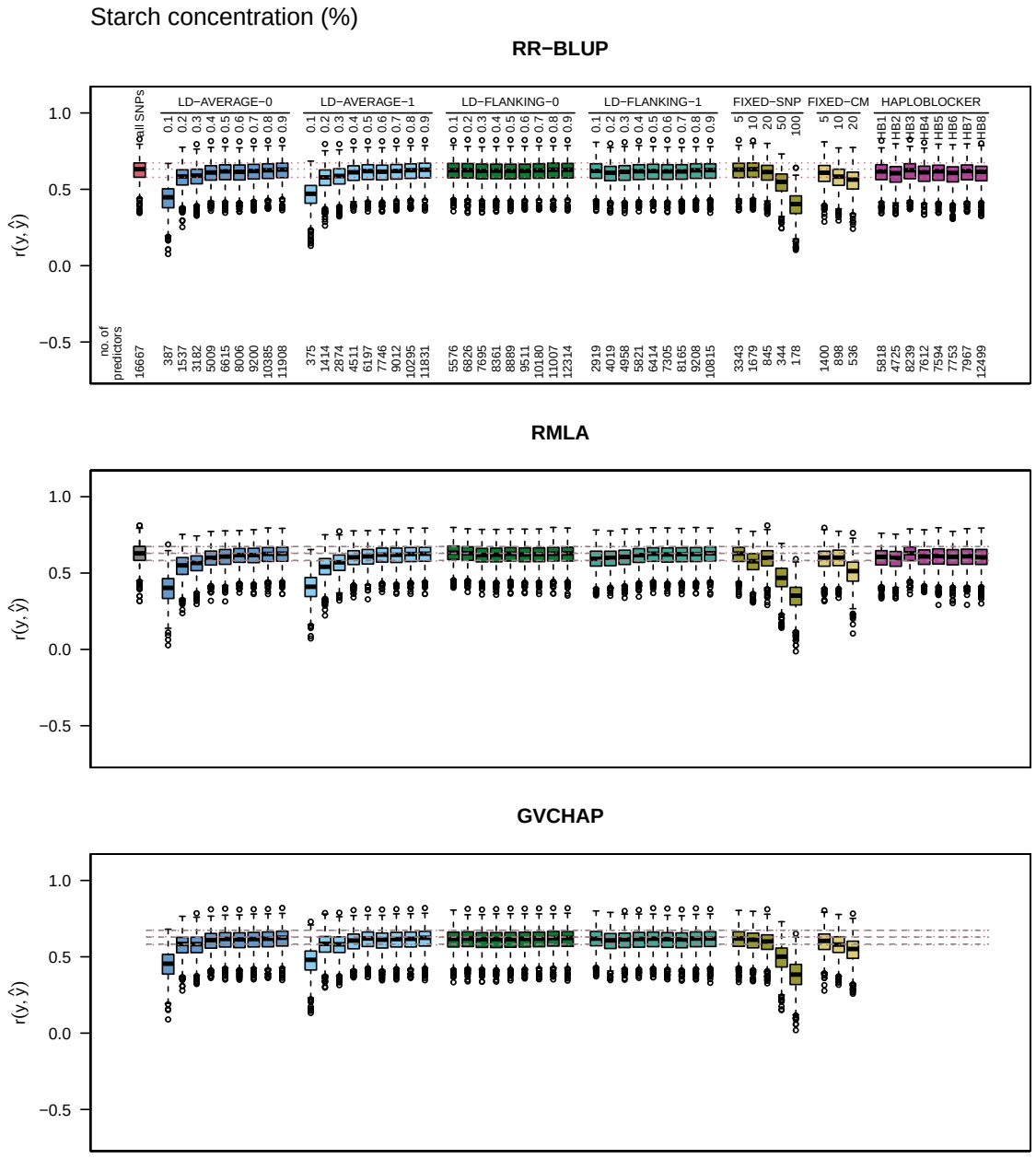

Figure S5. Prediction accuracies for genomic prediction of starch concentration with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD-AVERAGE-0`, `LD-AVERAGE-1`, `LD-FLANKING-0`, `LD-FLANKING-1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED-SNP`), with a fixed block length in cM (`FIXED-CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

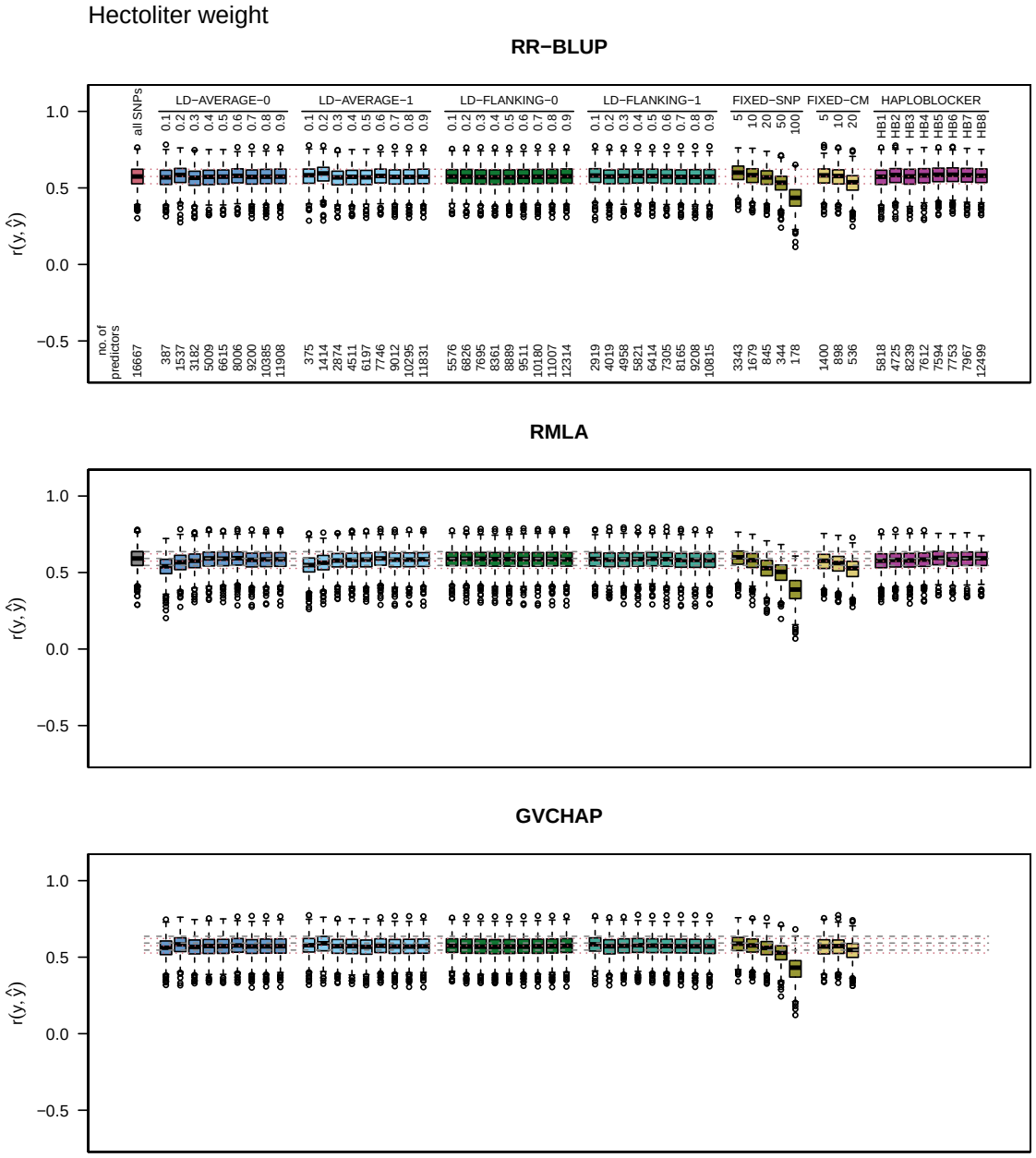

Figure S6. Prediction accuracies for genomic prediction of hectoliter weight with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

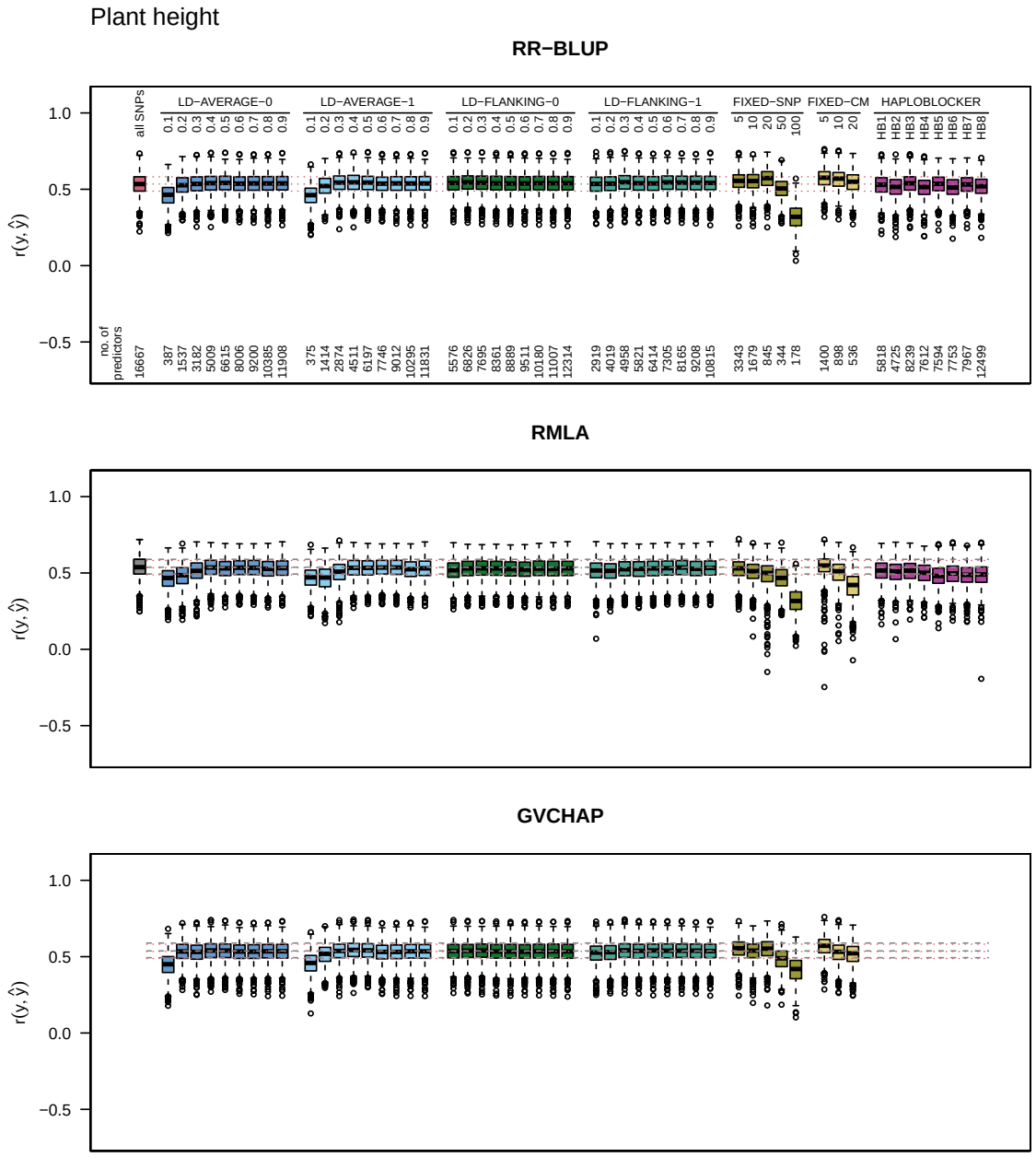

Figure S7. Prediction accuracies for genomic prediction of plant height with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

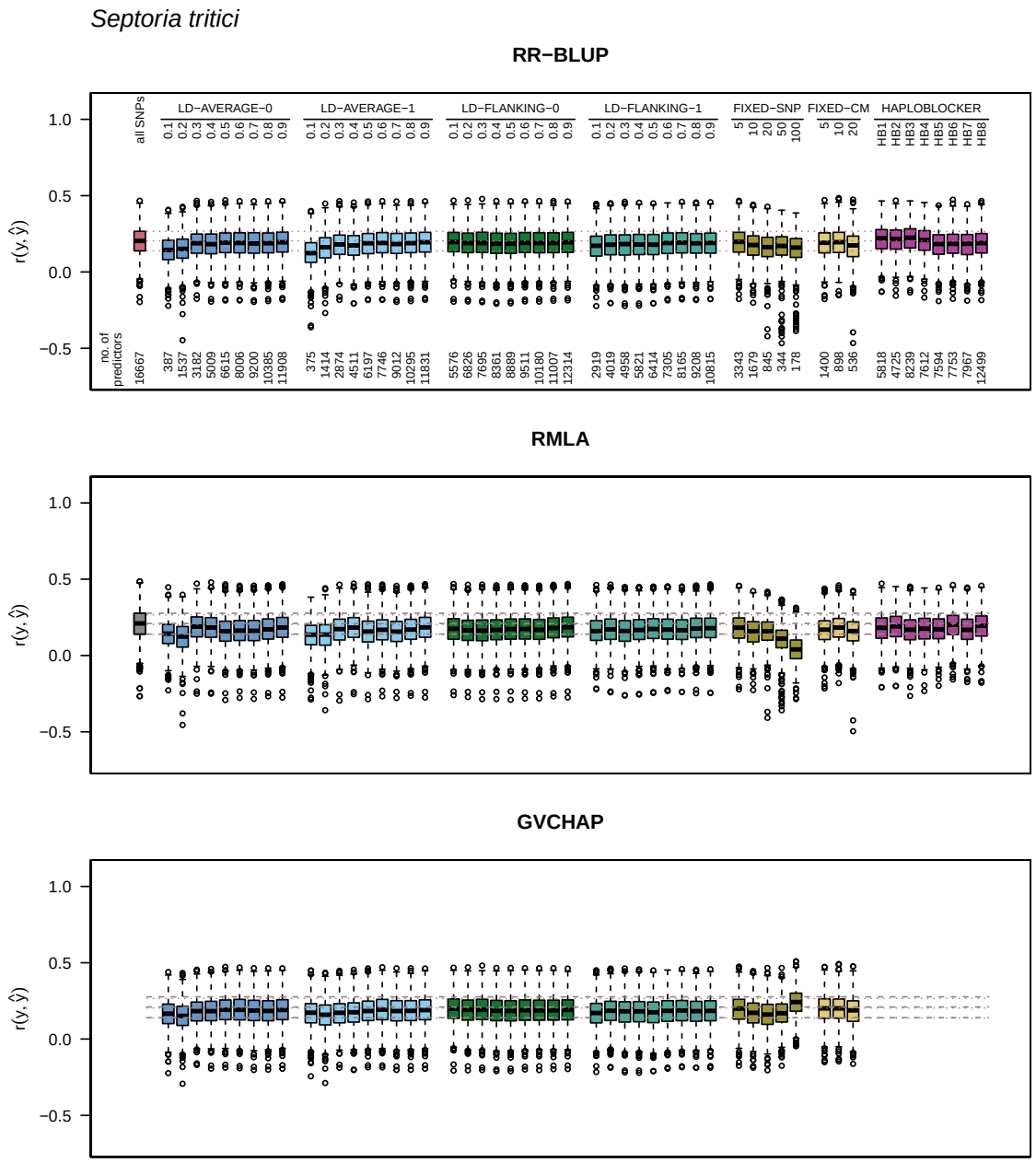

Figure S8. Prediction accuracies for genomic prediction of the resistance score for *Septoria tritici* with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD-AVERAGE-0`, `LD-AVERAGE-1`, `LD-FLANKING-0`, `LD-FLANKING-1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED-SNP`), with a fixed block length in cM (`FIXED-CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

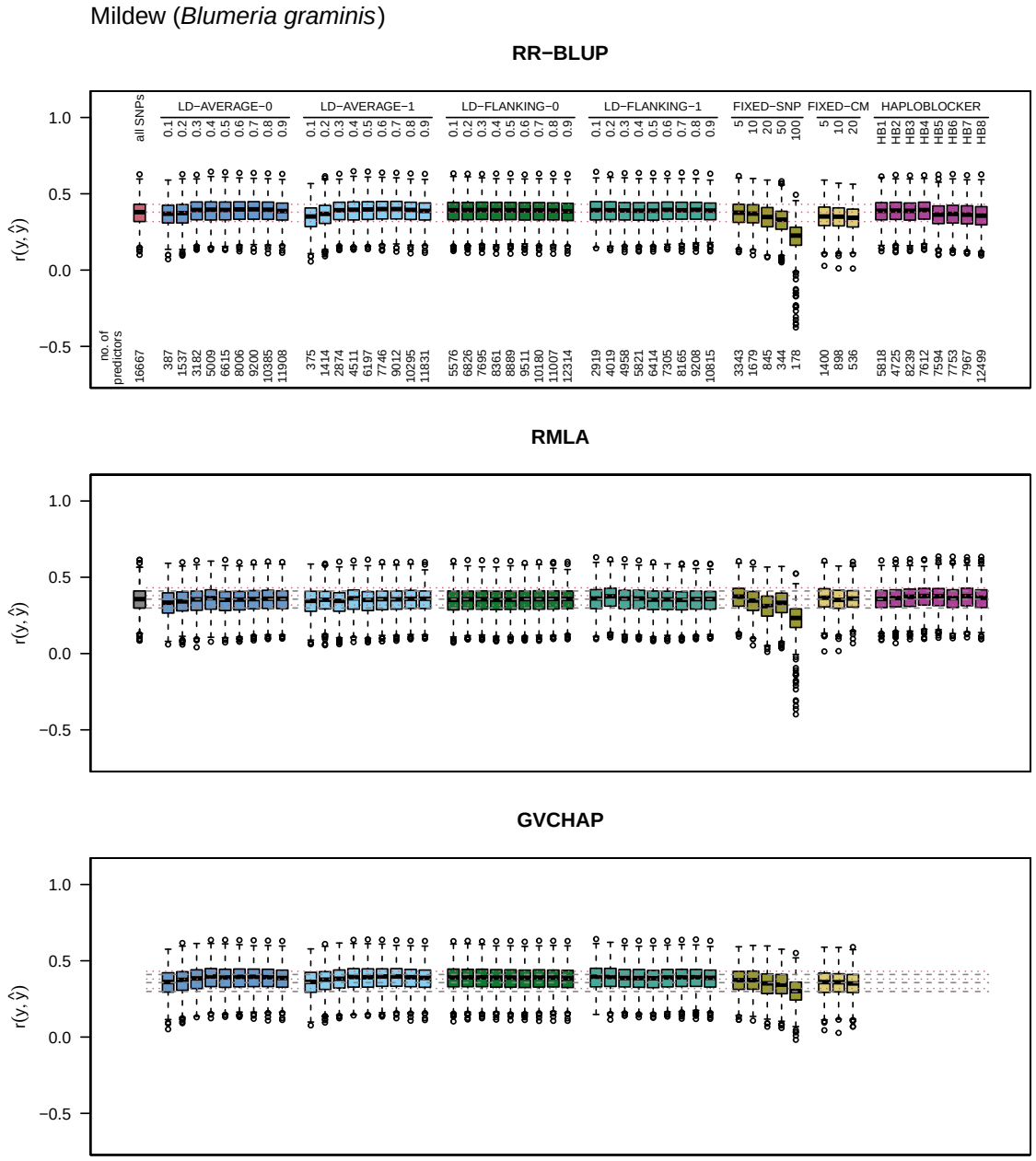

Figure S9. Prediction accuracies for genomic prediction of the resistance score for mildew (*Blumeria graminis*) with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD-AVERAGE-0`, `LD-AVERAGE-1`, `LD-FLANKING-0`, `LD-FLANKING-1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED-SNP`), with a fixed block length in cM (`FIXED-CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

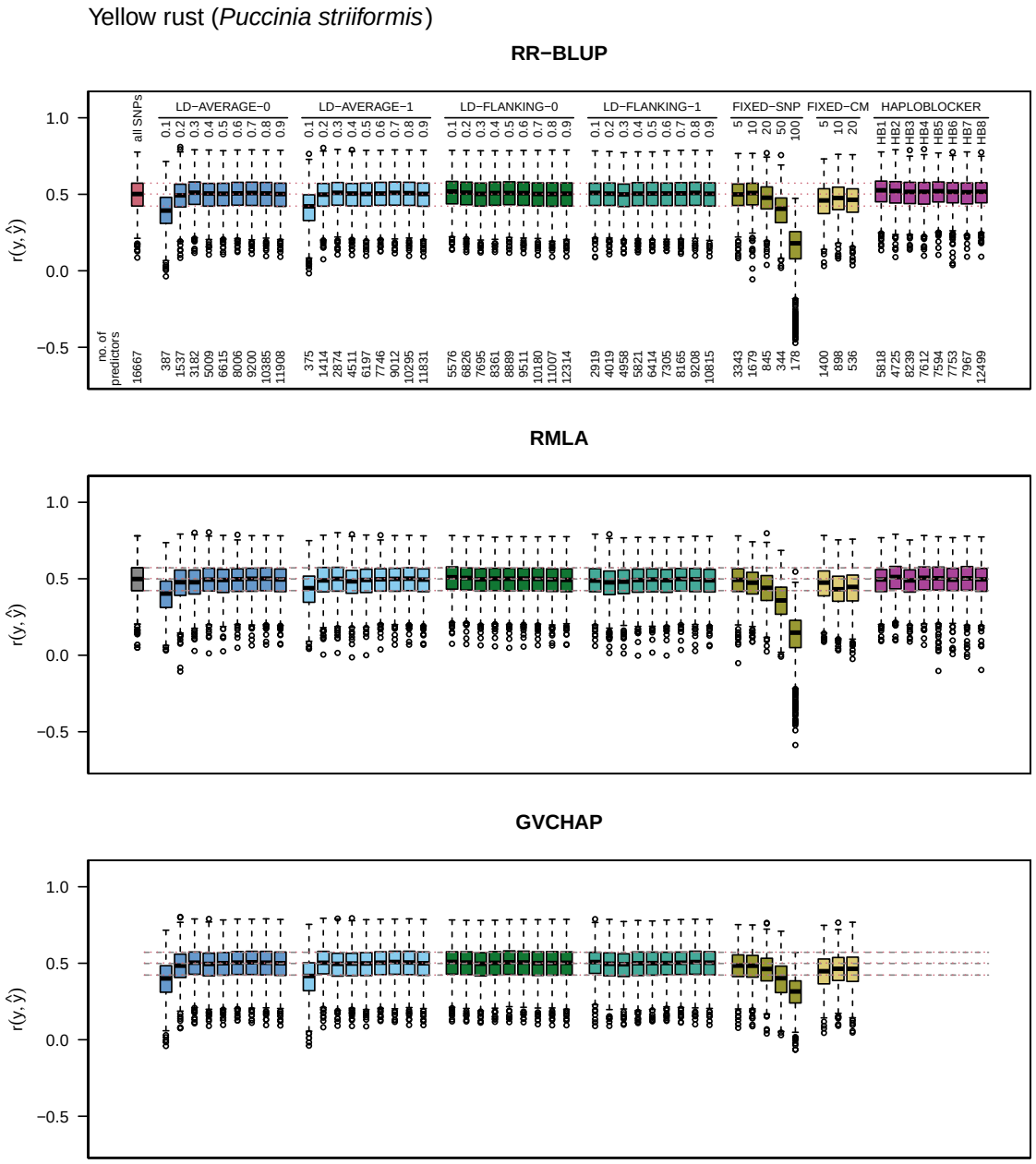

Figure S10. Prediction accuracies for genomic prediction of the resistance score for yellow rust (*Puccinia striiformis*) with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

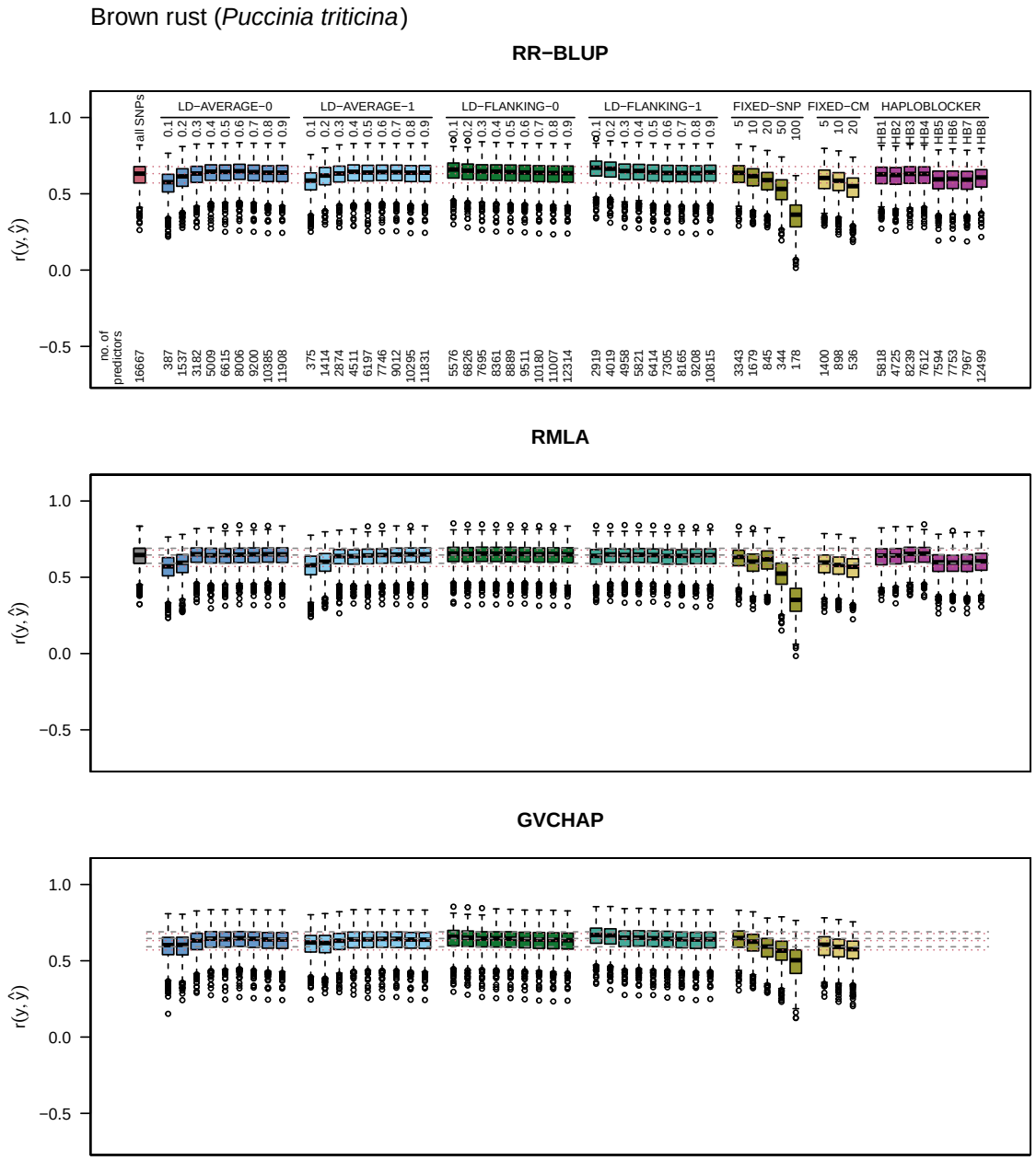

Figure S11. Prediction accuracies for genomic prediction of the resistance score for brown rust (*Puccinia triticina*) with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD−AVERAGE−0`, `LD−AVERAGE−1`, `LD−FLANKING−0`, `LD−FLANKING−1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED−SNP`), with a fixed block length in cM (`FIXED−CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

Fusarium graminearum

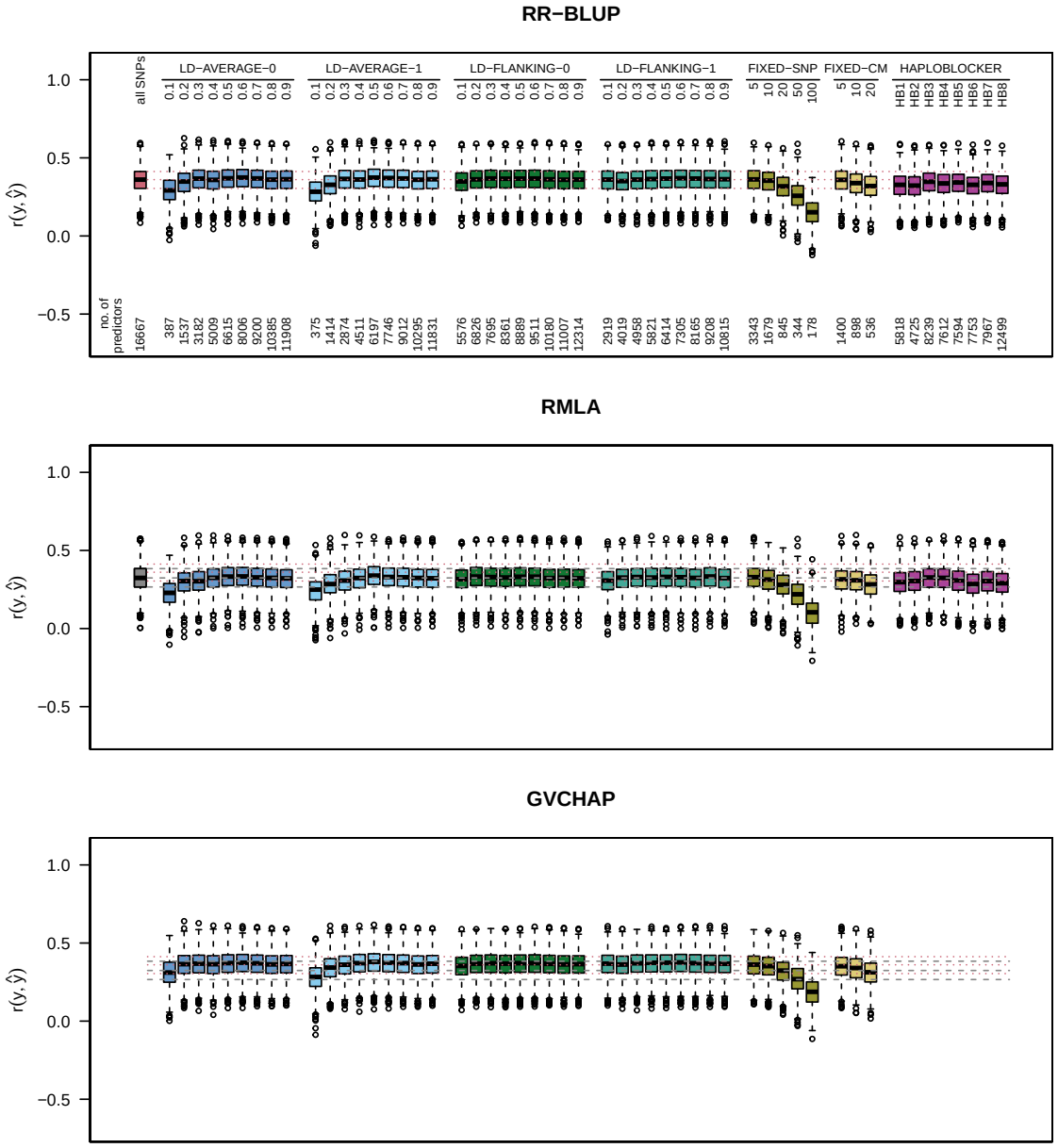

Figure S12. Prediction accuracies for genomic prediction of the resistance score for *Fusarium graminearum* with different types of haplotype blocks and estimation methods. The boxplots show the correlations $r(y, \hat{y})$ between the observed phenotypic values y and the predicted phenotypic values $\hat{y}$ in the validation set for 1000 cross-validation runs. Haplotype blocks were built based on linkage disequilibrium (`LD-AVERAGE-0`, `LD-AVERAGE-1`, `LD-FLANKING-0`, `LD-FLANKING-1`) with different threshold values $t = 0.1, 0.2, ..., 0.9$ for r^2, with fixed numbers of SNPs per block (`FIXED-SNP`), with a fixed block length in cM (`FIXED-CM`), or with the R package HaploBlocker (`HAPLOBLOCKER`). Red dotted lines: Quartiles from RR-BLUP with 16,667 SNPs (baseline). Gray dashed lines: Quartiles from RMLA with 16,667 SNPs. The number of predictors is the combined number of haplotype blocks and unassigned SNPs.

Supplementary Material

1 MARKER DATA

The following example illustrates (1) how haplotype blocks are built from marker data and (2) how re-parameterized design matrices are constructed from the haplotype blocks.

The example data is for 10 genotypes ($G01, \ldots, G10$) and 10 consecutive SNP markers ($m01, \ldots, m10$) on a single chromosome. The original matrix with phased SNP data (A: 1, C: 2; G: 3; T: 4; missing: -1) is

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$G01$	4/4	1/1	3/3	1/1	4/4	2/2	4/4	2/2	3/3	3/3
$G02$	2/4	−1/−1	3/3	1/1	3/3	2/4	4/4	2/2	3/3	3/3
$G03$	2/2	1/3	3/3	1/1	3/3	4/4	4/4	2/2	3/3	3/3
$G04$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G05$	2/2	1/3	1/1	2/2	4/4	4/4	4/4	2/2	3/3	3/3
$G06$	2/2	3/3	1/1	2/2	3/3	4/4	2/4	2/2	2/2	3/3
$G07$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G08$	4/4	3/3	1/1	2/2	3/3	4/4	4/4	2/2	3/3	3/3
$G09$	4/4	1/3	1/1	1/1	4/4	4/4	2/2	4/4	2/2	4/4
$G10$	2/4	3/3	1/1	2/2	3/3	4/4	2/2	2/4	2/2	4/4

which is recoded to the design matrix $\mathbf{Z}$ of a marker model

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$G01$	2	0	2	0	2	0	2	0	2	0
$G02$	1	NA	2	0	0	1	2	0	2	0
$G03$	0	1	2	0	0	2	2	0	2	0
$G04$	0	0	2	0	2	0	0	2	0	2
$G05$	0	1	0	2	2	2	2	0	2	0
$G06$	0	2	0	2	0	2	1	0	0	0
$G07$	0	0	2	0	2	0	0	2	0	2
$G08$	2	2	0	2	0	2	2	0	2	0
$G09$	2	1	0	0	2	2	0	2	0	2
$G10$	1	2	0	2	0	2	0	1	0	2

$$\text{where } X = \begin{cases} 0 & \text{if the genotype is homozygous for the major allele at the SNP} \\ 1 & \text{if the genotype is heterozygous at the SNP} \\ 2 & \text{if the genotype is homozygous for the minor allele at the SNP} \end{cases}$$

2 LD-BASED HAPLOTYPE BLOCKS

r^2 as a measure for pairwise LD is calculated between all the marker pairs, resulting in the following matrix (r^2 values in the upper diagonal):

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.04	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.00	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.06	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.17	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.07	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.81	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67	1.00
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Haplotype blocks are built with method `LD-FLANKING-0` with a threshold value $t = 0.3$ for r^2. The following steps are applied:

1. *Search for the pair of adjacent loci on the chromosome that has the greatest LD value r^2 among all pairs of loci that are not yet assigned to a haplotype block.*

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

   The r^2 value for markers $m09$ and $m10$ is 0.67. Note that r^2 between markers $m07$ and $m09$ is higher (0.82) but those markers are not adjacent. Neither are the markers $m07$ and $m09$, or $m07$ and $m10$.

2. *Check whether r^2 is greater than the defined threshold t.*
   $0.67 > 0.3$. It follows that markers $m09$ and $m10$ are grouped into a block.

3. *Check whether the block can be extendend to the left or to the right by comparing t with the LD between the new marker and the marker flanking the block. If $r^2 > t$, extend the block. If $r^2 < t$, then each locus on the chromosome is an individual block.*

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

The new block is indicated by a vertical line in the column names. To the left, t has to be compared with an LD value of $r^2 = 0.81$ between markers $m10$ (the marker flanking the new block on the right) and $m08$ (the marker flanking the new block on the left). Since $0.81 > 0.3$, the block is extended to the left. Therefore, markers $m08$, $m09$, and $m10$ are grouped into one haplotype block. The block cannot be extended to the right because the chromosome ends with marker $m10$.

4. *Repeat the previous steps until all loci on a chromosome are assigned to a block.*

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.81 > 0.3$. The block can be extended to the left and marker $m07$ is included.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.07 < 0.3$. The block cannot be extended to the left. Now search for the pair of adjacent loci on the chromosome that are not yet assigned to a haplotype block.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.67 > 0.3$. Markers $m03$ and $m04$ are grouped into a haplotype block.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.35 > 0.3$. The block can be extended to the left and marker $m02$ is included.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0 < 0.3$. The block cannot be extended to the left.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.34 > 0.3$. The block can be extended to the right and marker $m05$ is included.

	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$m01$	1.00	0.01	0.04	0.00	0.00	0.03	0.02	0.00	0.04	0.00
$m02$	0.00	1.00	0.34	0.35	0.34	0.36	0.02	0.09	0.01	0.04
$m03$	0.00	0.00	1.00	0.67	0.04	0.54	0.01	0.01	0.04	0.00
$m04$	0.00	0.00	0.00	1.00	0.17	0.36	0.02	0.15	0.00	0.06
$m05$	0.00	0.00	0.00	0.00	1.00	0.27	0.09	0.27	0.04	0.17
$m06$	0.00	0.00	0.00	0.00	0.00	1.00	0.03	0.12	0.01	0.07
$m07$	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.66	0.82	0.81
$m08$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.54	0.81
$m09$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.67
$m10$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

$0.36 > 0.3$. The block can be extended to the right and marker $m06$ is included. The haplotype block ends here because a new one begins with marker $m07$. On the left of the haplotype block, only marker $m01$ is left. This marker is converted in a haplotype block with just one marker.

The final assignment of the markers to the haplotype blocks:

hb1: $m01$
hb2: $m02$, $m03$, $m04$, $m05$, $m06$
hb3: $m07$, $m08$, $m09$, $m10$

The corresponding SNPs are

	***hb*1**	***hb*2**					***hb*3**			
	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$G01$	4/4	1/1	3/3	1/1	4/4	2/2	4/4	2/2	3/3	3/3
$G02$	2/4	−1/−1	3/3	1/1	3/3	2/4	4/4	2/2	3/3	3/3
$G03$	2/2	1/3	3/3	1/1	3/3	4/4	4/4	2/2	3/3	3/3
$G04$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G05$	2/2	1/3	1/1	2/2	4/4	4/4	4/4	2/2	3/3	3/3
$G06$	2/2	3/3	1/1	2/2	3/3	4/4	2/4	2/2	2/2	3/3
$G07$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G08$	4/4	3/3	1/1	2/2	3/3	4/4	4/4	2/2	3/3	3/3
$G09$	4/4	1/3	1/1	1/1	4/4	4/4	2/2	4/4	2/2	4/4
$G10$	2/4	3/3	1/1	2/2	3/3	4/4	2/4	2/4	2/2	4/4

For $m01$ which forms its own haplotype block $hb1$, there are only two variants, 2 and 4. However, for $hb2$ and $hb3$, more combinations of the original SNP alleles have to be considered. The following matrix shows the possible combinations of marker alleles for $hb2$. There is only one line for each homozygous genotype and two lines for the genotypes that were heterozygous at one of the markers.

	***hb*2**				
	$m02$	$m03$	$m04$	$m05$	$m06$
$G01$	1	3	1	4	2
$G02$	−1	3	1	3	2
	−1	3	1	3	4
$G03$	1	3	1	3	4
	3	3	1	3	4
$G04$	1	3	1	4	2
$G05$	1	1	2	4	4
	3	1	2	4	4
$G06$	3	1	2	3	4
$G07$	1	3	1	4	2
$G08$	3	1	2	3	4
$G09$	1	1	1	4	4
	3	1	1	4	4
$G10$	3	1	2	3	4

When combinations that occur more than once are removed, the following sequences of marker alleles are left for *hb*2:

$$\begin{array}{c} \\ \end{array} \begin{array}{c} \boldsymbol{hb2} \\ \begin{array}{ccccc} m02 & m03 & m04 & m05 & m06 \end{array} \end{array} \begin{pmatrix} 1 & 3 & 1 & 4 & 2 \\ -1 & 3 & 1 & 3 & 2 \\ -1 & 3 & 1 & 3 & 4 \\ 1 & 3 & 1 & 3 & 4 \\ 3 & 3 & 1 & 3 & 4 \\ 1 & 1 & 2 & 4 & 4 \\ 3 & 1 & 2 & 4 & 4 \\ 3 & 1 & 2 & 3 & 4 \\ 1 & 1 & 1 & 4 & 4 \\ 3 & 1 & 1 & 4 & 4 \end{pmatrix}$$

Each of these sequences can be considered a unique variant (or "allele") of haplotype block *hb*2. Note that (1) SNP data must be phased to define meaningful variants and (2) missing marker data are considered as separate alleles.

For haplotype block *hb*03, the derivation of the variants looks as follows:

$$\begin{array}{c} \\ \end{array} \begin{array}{c} \boldsymbol{hb3} \\ \begin{array}{cccc} m07 & m08 & m09 & m10 \end{array} \end{array} \begin{array}{c} G01 \\ G02 \\ G03 \\ G04 \\ G05 \\ G06 \\ \\ G07 \\ G08 \\ G09 \\ G10 \\ \end{array} \begin{pmatrix} 4 & 2 & 3 & 3 \\ 4 & 2 & 3 & 3 \\ 4 & 2 & 3 & 3 \\ 2 & 4 & 2 & 4 \\ 4 & 2 & 3 & 3 \\ 2 & 2 & 2 & 3 \\ 4 & 2 & 2 & 3 \\ 2 & 4 & 2 & 4 \\ 4 & 2 & 3 & 3 \\ 2 & 4 & 2 & 4 \\ 2 & 2 & 2 & 4 \\ 2 & 4 & 2 & 4 \end{pmatrix}$$

The following unique combinations remain:

$$\begin{array}{c} \\ \end{array} \begin{array}{c} \boldsymbol{hb3} \\ \begin{array}{cccc} m07 & m08 & m09 & m10 \end{array} \end{array} \begin{pmatrix} 4 & 2 & 3 & 3 \\ 2 & 4 & 2 & 4 \\ 2 & 2 & 2 & 3 \\ 4 & 2 & 2 & 3 \\ 2 & 4 & 2 & 4 \end{pmatrix}$$

Consequently, there are two variants for haplotype block $hb1$, ten variants for haplotype block $hb2$, and five variants for haplotype block $hb3$.

Overview over distinct variants for haplotype blocks built with method `LD-FLANKING-0` for an r^2 threshold of 0.3:

Haplotype block	Variant	Sequence of SNP alleles
$hb1$	1	2
$hb1$	2	4
$hb2$	1	1;3;1;4;2
$hb2$	2	3;1;2;3;4
$hb2$	3	1;3;1;3;4
$hb2$	4	3;3;1;3;4
$hb2$	5	1;1;2;4;4
$hb2$	6	3;1;2;4;4
$hb2$	7	-1;3;1;3;4
$hb2$	8	-1;3;1;3;2
$hb2$	9	1;1;1;4;4
$hb2$	10	3;1;1;4;4
$hb3$	1	4;2;3;3
$hb3$	2	2;4;2;4
$hb3$	3	4;2;2;3
$hb3$	4	2;2;2;4
$hb3$	5	2;2;2;3

The following matrix assigns the haplotype blocks variants to the genotypes according to the SNP sequence that they show for the respective block.

$$\begin{array}{c} \\ G01 \\ G02 \\ G03 \\ G04 \\ G05 \\ G06 \\ G07 \\ G08 \\ G09 \\ G10 \end{array} \begin{array}{c} hb1 \\ \begin{pmatrix} 2/2 \\ 1/2 \\ 1/1 \\ 1/1 \\ 1/1 \\ 1/1 \\ 1/1 \\ 2/2 \\ 2/2 \\ 1/2 \end{pmatrix} \end{array} \begin{array}{c} hb2 \\ \begin{pmatrix} 1/1 \\ -1/-1 \\ 3/4 \\ 1/1 \\ 5/6 \\ 2/2 \\ 1/1 \\ 2/2 \\ 9/10 \\ 2/2 \end{pmatrix} \end{array} \begin{array}{c} hb3 \\ \begin{pmatrix} 1/1 \\ 1/1 \\ 1/1 \\ 2/2 \\ 1/1 \\ 5/3 \\ 2/2 \\ 1/1 \\ 2/2 \\ 4/2 \end{pmatrix} \end{array}$$

Compare this to the sequences of SNP alleles of the genotypes in haplotype block $hb3$:

	hb3			
	$m07$	$m08$	$m09$	$m10$
$G01$	4	2	3	3
$G02$	4	2	3	3
$G03$	4	2	3	3
$G04$	2	4	2	4
$G05$	4	2	3	3
$G06$	2	2	2	3
	4	2	2	3
$G07$	2	4	2	4
$G08$	4	2	3	3
$G09$	2	4	2	4
$G10$	2	2	2	4
	2	4	2	4

The matrix with the haplotype variants then has to be re-parametrized in order to obtain a design matrix **Z** with encoding 0,1,2 for the mixed linear model. In this matrix, each variant for each haplotype block gets one column. The variants are then encoded with 0 (haploblock variant is absent), 1 (one copy of the haploblock variant present), or 2 (two copies of the haploblock variant present).

	hb1		***hb2***								***hb3***				
	$v01$	$v02$	$v01$	$v02$	$v03$	$v04$	$v05$	$v06$	$v09$	$v10$	$v01$	$v02$	$v03$	$v04$	$v05$
$G01$	0	2	2	0	0	0	0	0	0	0	2	0	0	0	0
$G02$	1	1	0	0	0	0	0	0	0	0	2	0	0	0	0
$G03$	2	0	0	0	1	1	0	0	0	0	2	0	0	0	0
$G04$	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0
$G05$	2	0	0	0	0	0	1	1	0	0	2	0	0	0	0
$G06$	2	0	0	2	0	0	0	0	0	0	0	0	1	0	1
$G07$	2	0	2	0	0	0	0	0	0	0	0	2	0	0	0
$G08$	0	2	0	2	0	0	0	0	0	0	2	0	0	0	0
$G09$	0	2	0	0	0	0	0	0	1	1	0	2	0	0	0
$G10$	1	1	0	2	0	0	0	0	0	0	0	1	0	1	0

This is the R code that can be used to obtain the matrices:

```
> library("SelectionTools")
> # Package can be downloaded from http://population-genetics.uni-giessen.de/~software/
>
> # Make the marker data file (assumption: phased marker data)
> marker <- data.frame(matrix(c("4/4", "2/4", "2/2", "2/2", "2/2", "2/2", "2/2", "4/4", "4/4", "2/4",
+                               "1/1", "-1/-1", "1/3", "1/1", "1/3", "3/3", "1/1", "3/3", "1/3", "3/3",
+                               "3/3", "3/3", "3/3", "3/3", "1/1", "1/1", "3/3", "1/1", "1/1", "1/1",
+                               "1/1", "1/1", "1/1", "1/1", "2/2", "2/2", "1/1", "2/2", "1/1", "2/2",
+                               "4/4", "3/3", "3/3", "4/4", "4/4", "3/3", "4/4", "3/3", "4/4", "3/3",
```

```
+                                "2/2", "2/4", "4/4", "2/2", "4/4", "4/4", "2/2", "4/4", "4/4", "4/4",
+                                "4/4", "4/4", "4/4", "2/2", "4/4", "2/4", "2/2", "4/4", "2/2", "2/2",
+                                "2/2", "2/2", "2/2", "4/4", "2/2", "2/2", "4/4", "2/2", "4/4", "2/4",
+                                "3/3", "3/3", "3/3", "2/2", "3/3", "2/2", "2/2", "3/3", "2/2", "2/2",
+                                "3/3", "3/3", "3/3", "4/4", "3/3", "3/3", "4/4", "3/3", "4/4", "4/4"),
+                              byrow=T, ncol=10))
> colnames(marker) <- sprintf("G%02i", 1:10)
> rownames(marker) <- sprintf("m%02i", 1:10)
> write.table(marker, "example-marker.txt", quote=F)
>
> # Make the map file
> map <- data.frame(name = sprintf("m%02i", 1:10),
+                   chrom = 1,
+                   pos = seq(from=1, to=100, by=10))
> write.table(map, "example-map.txt", quote=F, row.names=F, sep=" ")
>
> # Read marker data and map file
> st.read.marker.data("example-marker.txt", format="m", data.set="default")
M (data set 'default'): No. of individuals: 10, no. of markers: 10
> st.read.map("example-map.txt", format="mcp", skip=1, data.set="default")
M (data set 'default'): No. of individuals: 10, no. of markers: 10
>
> # Marker data in SelectionTools
> xx <- st.marker.data.statistics()
M (data set 'default'): No. of individuals: 10, no. of markers: 10
> xx$genotypes # note the swap of rows and columns!
   Mar/Ind G01   G02 G03 G04 G05 G06 G07 G08 G09 G10
1      m01 4/4   2/4 2/2 2/2 2/2 2/2 2/2 4/4 4/4 2/4
2      m02 1/1 -1/-1 1/3 1/1 1/3 3/3 1/1 3/3 1/3 3/3
3      m03 3/3   3/3 3/3 3/3 1/1 1/1 3/3 1/1 1/1 1/1
4      m04 1/1   1/1 1/1 1/1 2/2 2/2 1/1 2/2 1/1 2/2
5      m05 4/4   3/3 3/3 4/4 4/4 3/3 4/4 3/3 4/4 3/3
6      m06 2/2   2/4 4/4 2/2 4/4 4/4 2/2 4/4 4/4 4/4
7      m07 4/4   4/4 4/4 2/2 4/4 2/4 2/2 4/4 2/2 2/2
8      m08 2/2   2/2 2/2 4/4 2/2 2/2 4/4 2/2 4/4 2/4
9      m09 3/3   3/3 3/3 2/2 3/3 2/2 2/2 3/3 2/2 2/2
10     m10 3/3   3/3 3/3 4/4 3/3 3/3 4/4 3/3 4/4 4/4
>
> # Design matrix with single markers
> ZZ <- gs.build.Z(data.set="default", out.filename="Z.matrix", auxfiles=T)
> ZZ
    m01.4 m02.3 m03.3 m04.2 m05.4 m06.4 m07.4 m08.4 m09.3 m10.4
G01     2     0     2     0     2     0     2     0     2     0
G02     1     1     2     0     0     1     2     0     2     0
G03     0     1     2     0     0     2     2     0     2     0
G04     0     0     2     0     2     0     0     2     0     2
G05     0     1     0     2     2     2     2     0     2     0
G06     0     2     0     2     0     2     1     0     0     0
G07     0     0     2     0     2     0     0     2     0     2
```

```
G08     2     2     0     2     0     2     2     0     2     0
G09     2     1     0     0     2     2     0     2     0     2
G10     1     2     0     2     0     2     0     1     0     2
>
> # Calculate LD
> ld <- st.calc.ld ( ld.measure="r2",
+                    data.set="default" )
> head(ld, 10)
   Chrom Locus1 Locus2 Name1 Name2       LD
1      1      1      2   m01   m02 0.009848
2      1      1      3   m01   m03 0.041667
3      1      1      4   m01   m04 0.001736
4      1      1      5   m01   m05 0.000000
5      1      1      6   m01   m06 0.029304
6      1      1      7   m01   m07 0.015152
7      1      1      8   m01   m08 0.001832
8      1      1      9   m01   m09 0.041667
9      1      1     10   m01   m10 0.001736
10     1      2      3   m02   m03 0.336364
>
> # Pairwise LD
> pairwise.ld <- xtabs(LD ~ Name1 + Name2, data=ld)
> round(pairwise.ld, 2)
     Name2
Name1  m02  m03  m04  m05  m06  m07  m08  m09  m10
  m01 0.01 0.04 0.00 0.00 0.03 0.02 0.00 0.04 0.00
  m02 0.00 0.34 0.35 0.34 0.36 0.02 0.09 0.01 0.04
  m03 0.00 0.00 0.67 0.04 0.54 0.01 0.01 0.04 0.00
  m04 0.00 0.00 0.00 0.17 0.36 0.02 0.15 0.00 0.06
  m05 0.00 0.00 0.00 0.00 0.27 0.09 0.27 0.04 0.17
  m06 0.00 0.00 0.00 0.00 0.00 0.03 0.12 0.01 0.07
  m07 0.00 0.00 0.00 0.00 0.00 0.00 0.66 0.82 0.81
  m08 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.54 0.81
  m09 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.67
>
> # Define haplotype blocks
> hb <- st.def.hblocks ( ld.threshold = 0.3, # Minimum LD
+                        ld.criterion = "flanking", # between markers
+                        data.set="default" ) # flanking the block
M (data set 'default'): LD of markers flanking haplotype blocks > 0.30
> hb
  Chrom Pos    Name Class              Markers
1     1   1 b000000     b                 m01;
2     1  31 b000001     b m02;m03;m04;m05;m06;
3     1  76 b000002     b     m07;m08;m09;m10;
>
> # Recode data set
> rb <- st.recode.hil(data.set="default")
M (data set 'default'): No. of individuals: 10, no. of markers: 3
```

```
> rb
     Block AlleleNr  AlleleDef
1  b000000        1          2
2  b000000        2          4
3  b000001        1  1;3;1;4;2
4  b000001        2  3;1;2;3;4
5  b000001        3  1;3;1;3;4
6  b000001        4  3;3;1;3;4
7  b000001        5  1;1;2;4;4
8  b000001        6  3;1;2;4;4
9  b000001        7 -1;3;1;3;4
10 b000001        8 -1;3;1;3;2
11 b000001        9  1;1;1;4;4
12 b000001       10  3;1;1;4;4
13 b000002        1    4;2;3;3
14 b000002        2    2;4;2;4
15 b000002        3    4;2;2;3
16 b000002        4    2;2;2;4
17 b000002        5    2;2;2;3
>
> # Variants for haplotype blocks
> xx <- st.marker.data.statistics("default")
M (data set 'default'): No. of individuals: 10, no. of markers: 3
> xx
$individual.list
   Name    InMis
1   G01 0.000000
2   G02 0.333333
3   G03 0.000000
4   G04 0.000000
5   G05 0.000000
6   G06 0.000000
7   G07 0.000000
8   G08 0.000000
9   G09 0.000000
10  G10 0.000000

$marker.list
     Name NoAll MaMis ExHet AM A1 A2 A3 A4 A5 A6 A9 A10
1 b000000     2   0.0 0.480  0 12  8  0  0  0  0  0   0
2 b000001     8   0.1 0.759  2  6  6  1  1  1  1  1   1
3 b000002     5   0.0 0.620  0 10  7  1  1  1  0  0   0

$genotypes
  Mar/Ind G01   G02 G03 G04 G05 G06 G07 G08  G09 G10
1 b000000 2/2   1/2 1/1 1/1 1/1 1/1 1/1 2/2  2/2 1/2
2 b000001 1/1 -1/-1 3/4 1/1 5/6 2/2 1/1 2/2 9/10 2/2
3 b000002 1/1   1/1 1/1 2/2 1/1 5/3 2/2 1/1  2/2 4/2
```

```
>
> # Write data file and read it in again - this way, one column is created
> # for each variant of each haplotype block
> st.write.marker.data(nfilename="example-hb-ld", format="n", data.set="default")
> XX <- read.table("example-hb-ld.npo", header=T)
> XX <- data.frame(t(XX))
> rownames(XX) <- sprintf("G%02i", 1:10)
> XX
    b000000.1 b000000.2 b000001.1 b000001.2 b000001.3 b000001.4 b000001.5 b000001.6 b000001.9
G01         0         1         1         0         0         0         0         0         0
G02         1         1         0         0         0         0         0         0         0
G03         1         0         0         0         1         1         0         0         0
G04         1         0         1         0         0         0         0         0         0
G05         1         0         0         0         0         0         1         1         0
G06         1         0         0         1         0         0         0         0         0
G07         1         0         1         0         0         0         0         0         0
G08         0         1         0         1         0         0         0         0         0
G09         0         1         0         0         0         0         0         0         1
G10         1         1         0         1         0         0         0         0         0
    b000001.10 b000002.1 b000002.2 b000002.3 b000002.4 b000002.5
G01          0         1         0         0         0         0
G02          0         1         0         0         0         0
G03          0         1         0         0         0         0
G04          0         0         1         0         0         0
G05          0         1         0         0         0         0
G06          0         0         0         1         0         1
G07          0         0         1         0         0         0
G08          0         1         0         0         0         0
G09          1         0         1         0         0         0
G10          0         0         1         0         1         0
>
> # The matrix must be re-coded so that homo- and heterozygous individuals
> # can be distinguished.
>
> no.alleles <- xx$marker.list$NoAll
> no.hb <- length(no.alleles)
> geno <- xx$individual.list$Name
> counter <- 1
>
> for (ii in 1:no.hb) {
+   alleles <- no.alleles[ii]
+   XX.subset <- XX[,counter:(counter+alleles-1)]
+
+   for (GEN in geno) {
+
+     rowsum <- sum(XX.subset[GEN,])
+
+     if (rowsum == 1) {
+       XX[GEN,counter:(counter+alleles-1)][XX[GEN,counter:(counter+alleles-1)]==1] <- 2
```

```
+     }
+
+   }
+
+   counter <- counter + alleles
+ }
>
> XX
    b000000.1 b000000.2 b000001.1 b000001.2 b000001.3 b000001.4 b000001.5 b000001.6 b000001.9
G01         0         2         2         0         0         0         0         0         0
G02         1         1         0         0         0         0         0         0         0
G03         2         0         0         0         1         1         0         0         0
G04         2         0         2         0         0         0         0         0         0
G05         2         0         0         0         0         0         1         1         0
G06         2         0         0         2         0         0         0         0         0
G07         2         0         2         0         0         0         0         0         0
G08         0         2         0         2         0         0         0         0         0
G09         0         2         0         0         0         0         0         0         1
G10         1         1         0         2         0         0         0         0         0
    b000001.10 b000002.1 b000002.2 b000002.3 b000002.4 b000002.5
G01          0         2         0         0         0         0
G02          0         2         0         0         0         0
G03          0         2         0         0         0         0
G04          0         0         2         0         0         0
G05          0         2         0         0         0         0
G06          0         0         0         1         0         1
G07          0         0         2         0         0         0
G08          0         2         0         0         0         0
G09          1         0         2         0         0         0
G10          0         0         1         0         1         0
```

3 HAPLOTYPE BLOCKS BASED ON A FIXED NUMBER OF SNPS

For method `FIXED-SNP` with n =5, five consecutive SNPs are grouped into one haplotype block.

			***hb*1**					***hb*2**		
	$m01$	$m02$	$m03$	$m04$	$m05$	$m06$	$m07$	$m08$	$m09$	$m10$
$G01$	4/4	1/1	3/3	1/1	4/4	2/2	4/4	2/2	3/3	3/3
$G02$	2/4	−1/−1	3/3	1/1	3/3	2/4	4/4	2/2	3/3	3/3
$G03$	2/2	1/3	3/3	1/1	3/3	4/4	4/4	2/2	3/3	3/3
$G04$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G05$	2/2	1/3	1/1	2/2	4/4	4/4	4/4	2/2	3/3	3/3
$G06$	2/2	3/3	1/1	2/2	3/3	4/4	2/4	2/2	2/2	3/3
$G07$	2/2	1/1	3/3	1/1	4/4	2/2	2/2	4/4	2/2	4/4
$G08$	4/4	3/3	1/1	2/2	3/3	4/4	4/4	2/2	3/3	3/3
$G09$	4/4	1/3	1/1	1/1	4/4	4/4	2/2	4/4	2/2	4/4
$G10$	2/4	3/3	1/1	2/2	3/3	4/4	2/4	2/4	2/2	4/4

Considering the sequence of markers within the two haplotype blocks, there are 12 variants for haplotype block $hb1$ and seven variants for haplotype block $hb2$.

Overview over distinct variants for haplotype blocks built with method `FIXED-SNP`, n = 5 (five SNPs per haplotype block):

Haplotype block	Variant	Sequence of SNP alleles
hb1	1	2;1;3;1;4
hb1	2	2;3;1;2;3
hb1	3	4;1;3;1;4
hb1	4	4;3;1;2;3
hb1	5	4;-1;3;1;3
hb1	6	2;1;1;2;4
hb1	7	2;3;1;2;4
hb1	8	2;-1;3;1;3
hb1	9	2;3;3;1;3
hb1	10	2;1;3;1;3
hb1	11	4;1;1;1;4
hb1	12	4;3;1;1;4
hb2	1	4;4;2;3;3
hb2	2	2;2;4;2;4
hb2	3	2;4;2;3;3
hb2	4	4;2;4;2;4
hb2	5	4;4;2;2;3
hb2	6	4;2;2;2;4
hb2	7	4;2;2;2;3

The following matrix assigns the haplotype blocks variants to the genotypes according to the SNP sequence that they show for the respective block. Note that haplotype blocks that contain missing SNP marker data are encoded as missing.

$$\begin{array}{c} \\ G01 \\ G02 \\ G03 \\ G04 \\ G05 \\ G06 \\ G07 \\ G08 \\ G09 \\ G10 \end{array} \begin{array}{c} hb1 \\ \left(\begin{array}{c} 3/3 \\ -1/-1 \\ 10/9 \\ 1/1 \\ 6/7 \\ 2/2 \\ 1/1 \\ 4/4 \\ 11/12 \\ 2/4 \end{array}\right. \end{array} \begin{array}{c} hb2 \\ \left.\begin{array}{c} 3/3 \\ 3/1 \\ 1/1 \\ 2/2 \\ 1/1 \\ 7/5 \\ 2/2 \\ 1/1 \\ 4/4 \\ 6/4 \end{array}\right) \end{array}$$

Lastly, the re-parametrized design matrix for the haplotype blocks:

$$\begin{array}{c} \\ G01 \\ G02 \\ G03 \\ G04 \\ G05 \\ G06 \\ G07 \\ G08 \\ G09 \\ G10 \end{array} \begin{array}{c} \begin{array}{cccccccccc} \multicolumn{10}{c}{\boldsymbol{hb1}} \\ v01 & v02 & v03 & v04 & v06 & v07 & v09 & v10 & v11 & v12 \end{array} \\ \left(\begin{array}{cccccccccc} 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \right. \end{array} \begin{array}{c} \begin{array}{ccccccc} \multicolumn{7}{c}{\boldsymbol{hb2}} \\ v01 & v02 & v03 & v04 & v05 & v06 & v07 \end{array} \\ \left.\begin{array}{ccccccc} 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{array} \right) \end{array}$$

Machine learning for prediction of resistance scores in wheat (*Triticum aestivum* L.) - Supplementary figures

Philipp Georg Heilmann | Yohannes Fekadu Difabachew | Matthias Frisch | Anna Luise Moritz | Andreas Stahl | Benjamin Wittkop | Rod J Snowdon | Michael Koch | Martin Kirchhoff | László Cselényi | Markus Wolf | Jutta Förster | Carola Zenke-Philippi

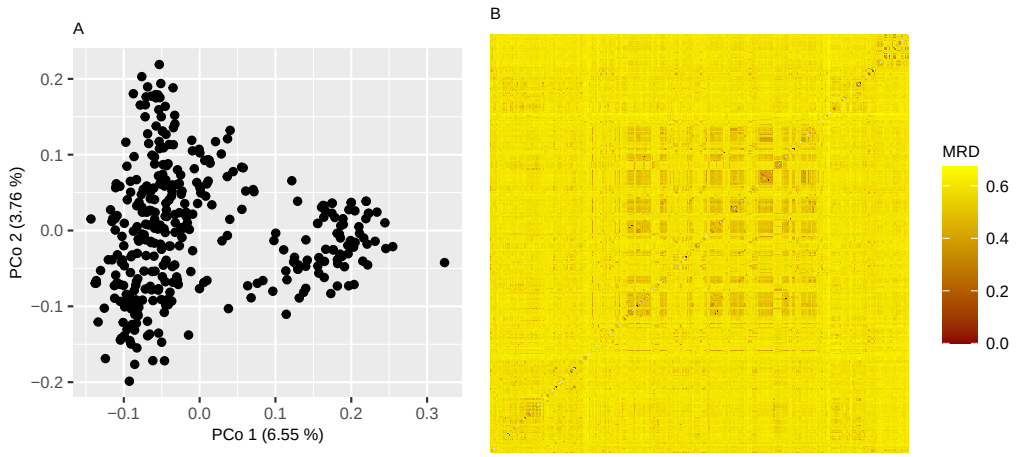

FIGURE S1 (A) Principal coordinate analysis based on the pairwise modified Roger's distances (MRD) and (B) heatmap showing the distances between the 361 elite winter wheat lines.

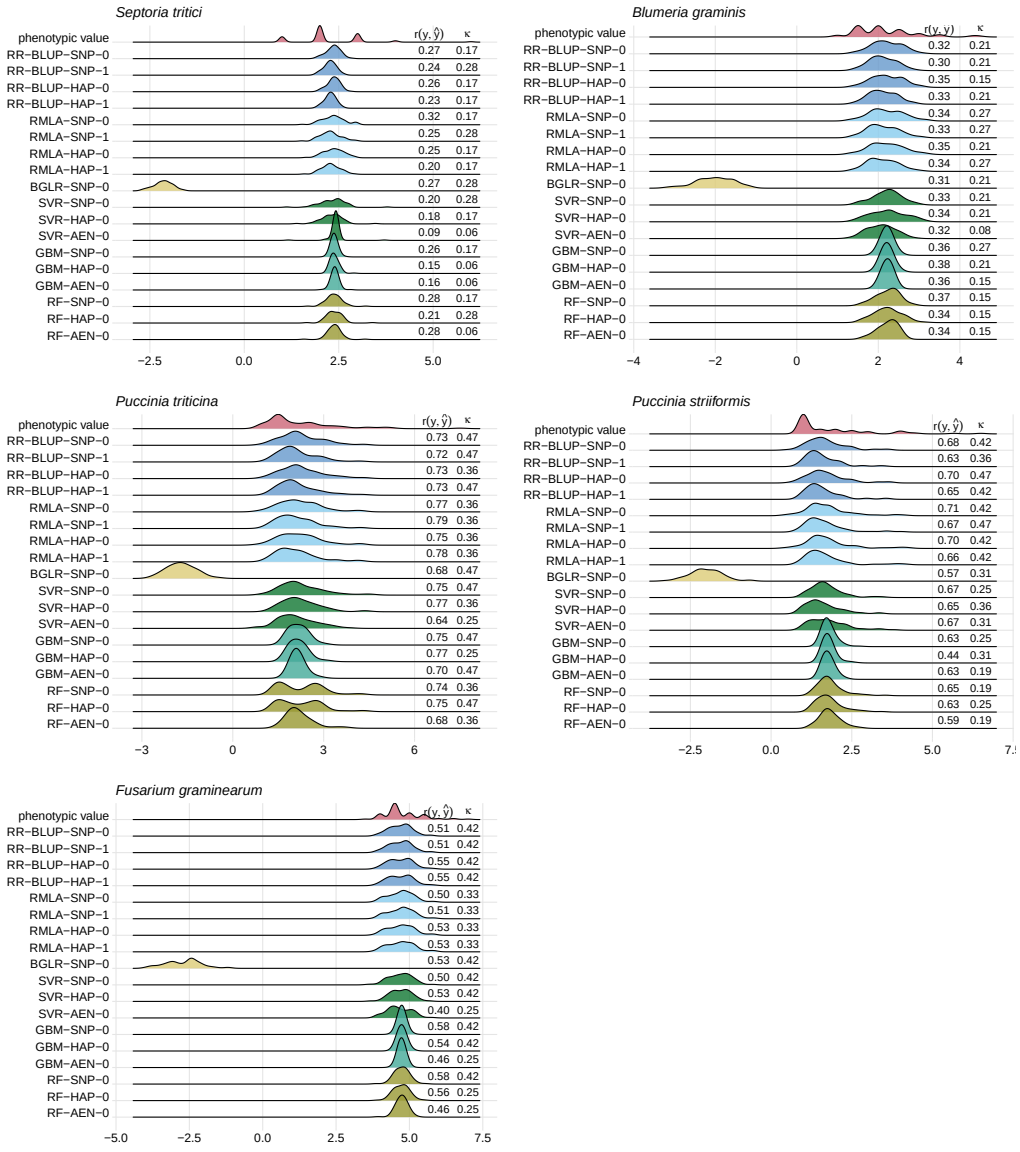

FIGURE S2 Distributions of observed phenotypic values and predicted phenotypic values of resistance scores for *S. tritici*, *B. graminis*, *B. triticina*, *P. striiformis*, and *F. graminearum* with different prediction approaches in the validation set (72 genotypes) in cross-validation run 126. Predictions were made with methods ridge regression BLUP (`RR-BLUP-...`), estimation of the error and genetic variance components with restricted maximum likelihood and partitioning according to ANOVA variance components (`RMLA-...`), Bayesian generalized linear regression (`BGLR-...`), support vector regression (`SVR-...`), gradient boosting machine (`GBM-...`), and random forest (`RF-...`). Predictors were either the full set of 16,667 SNP markers (`...-SNP-...`), haplotype blocks based on linkage disquilibrium (`...-HAP-...`), or 250 autoencoder features (`...-AEN-...`). The response values were either the untransformed resistance scores (`...-...-0`) or the logit-transformed resistance scores (`...-...-1`).

Acknowledgments

First and foremost, I would like to thank God for giving me the strength and guidance throughout my studies and the completion of this thesis.

I would also like to express my sincere gratitude to my academic supervisor, Dr. Carola Anna Luise Zenke-Philippi, for her support, valuable suggestions, and constructive feedback throughout the course of my study and thesis work.

Many thanks to Prof. Dr. Matthias Frisch for providing this opportunity, and to Prof. Dr. Rod Snowdon for being my second supervisor.

I am also thankful to my colleagues and the staff in the Biometry and Population Genetics Department for their support and for fostering a positive working atmosphere. Special thanks to Mr. Philipp Georg Heilmann and Dr. Tesfahun Alemu Stotaw for proof-reading this thesis.

Finally, I would like to extend my appreciation to my families and friends for their understanding and support.

Declaration of academic integrity

"I declare: I have completed this dissertation independently and without unauthorised outside help and only with the help that I have indicated in the dissertation. All text passages taken literally or analogously from published works and all information based on verbal information are labelled as such. In the research I conducted and mentioned in the dissertation, I adhered to the principles of good scientific practice as laid down in the "Statutes of the Justus Liebig University Giessen for Safeguarding Good Scientific Practice."

Gießen, 25 October 2024

Yohannes Fekadu Difabachew