Genomic prediction in plants: opportunities for machine learning-based approaches

Muhammad Farooq (qadrikazmi@yahoo.com)
Wageningen University & Research https://orcid.org/0000-0003-0704-3372
Aalt D.J. van Dijk
Wageningen University
Harm Nijveen
Wageningen University
Shahid Mansoor
National Institute for Biotechnology and Genetic Engineering
Dick de Ridder
Wageningen University & Research https://orcid.org/0000-0002-4944-4310

Research Article

Keywords: Genomic selection, genomic prediction, machine learning, GBLUP, Bayesian, random forest, extreme gradient boosting

Posted Date: February 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1315622/v1

Genomic prediction in plants: opportunities for machine learning-based approaches

Muhammad Farooq[1,2], Aalt D.J. van Dijk[1], Harm Nijveen[1], Shahid Mansoor[2] and Dick de Ridder[1,*]

[1] Bioinformatics Group, Wageningen University, The Netherlands

[2] Molecular Virology and Gene Silencing Lab, Agricultural Biotechnology Division, National Institute for Biotechnology and Genetic Engineering (NIBGE), Pakistan

***Correspondence:**

Dick de Ridder

dick.deridder@wur.nl

https://orcid.org/0000-0002-4944-4310

Keywords:

Genomic selection, genomic prediction, machine learning, GBLUP, Bayesian, random forest, extreme gradient boosting

Abbreviations

QTL: Quantitative Trait Loci

QTN: Quantitative Trait Nucleotide

GBLUP: Genomic Best Linear Unbiased Predictor

RKHS: Reproducing Kernel Hilbert Spacing

RF: Random Forest

XGBoost: Extreme Gradient Boosting

SVM: Support Vector Machine

SVR: Support Vector Regression

MLP: Multilayer Perceptron

GP: Genomic Prediction

Key message

Machine learning methods can be the potential choice for genomic prediction of both simple and complex traits; under both low or high SNP-QTL linkage disequilibrium scenarios. Reliable predictions over the course of a breeding programme would require to adjust the potential confounding factors e.g. population structure for all methods during model development.

Abstract

Many studies have demonstrated the utility of machine learning (ML) methods for genomic prediction (GP) of various plant traits, but a clear rationale for choosing ML over conventionally used, often simpler parametric methods is still lacking. Predictive performance of GP models might depend on a plethora of factors including sample size, number of markers, population structure and genetic architecture. Here, we investigate which problem and dataset characteristics are related to good performance of ML methods for genomic prediction. We compare the predictive performance of two frequently used ensemble ML methods (Random Forest and Extreme Gradient Boosting) with parametric methods including genomic best linear unbiased prediction (GBLUP), reproducing kernel Hilbert space regression (RKHS), BayesA and BayesB. To explore problem characteristics, we use simulated and real plant traits under different genetic complexity levels determined by the number of Quantitative Trait Loci (QTLs), heritability (h^2 and h^2_e), population structure and linkage disequilibrium between causal nucleotides and other SNPs. Results show that ML methods are a better choice for nonlinear phenotypes and still comparable to Bayesian methods for linear phenotypes in the case of large effect QTNs. Furthermore, we find that ML methods are susceptible to confounding due to population structure and less sensitive to low linkage disequilibrium than linear parametric methods. Overall, this provides insights into the role of ML in GP as well as guidelines for practitioners.

1. Introduction

Phenotypes of an individual are based on its genetic makeup, the environment and the interplay between them. In plant and animal breeding, the genomic prediction (GP) model, using a genome-wide set of markers, is an integral component of the genomic selection-based approach (Meuwissen, Hayes et al. 2001). A GP model is constructed on a reference population for which both genotypes and corresponding phenotypes are known, mostly employing a cross-validation strategy; and applied to related populations with only genotypes known. The total genomic value, estimated from the GP model, is used as a pseudo-phenotype to select the best parents for the next generation(s). In general, phenotypes differ from each other in terms of their genetic complexity, ranging from simple/monogenic to complex/polygenic. These differences impact the potential performance of GP. Complex traits are predominantly governed by a combination of additive and non-additive (e.g. dominant/recessive, epistatic etc.) allele effects, which makes GP challenging for these traits (Moore, Amos et al. 2015). The genetic architecture of complex traits is characterized by moderate to large numbers of Quantitative Trait Loci (QTLs) with small to medium effect sizes and with or without having large effect QTL(s) (Korte and Farlow 2013). Moreover, the ratio of additive to non-additive genetic variance may differ even for closely related traits. Besides the actual genetic variance level, its distribution over the genome is also a determinant of the trait architecture (Speed and Balding 2019). Next to genetic architecture, population structure play a role as well (*Figure 1*): inconsistent relatedness among samples due to ancestral allele frequency imbalance among sub-populations (population structure) or cryptic structures, e.g. familial relationships; linkage disequilibrium (LD) structure, due to inbreeding or selection pressure; varying relatedness between training and test population, e.g. over the course of breeding cycle; and sizes of reference and effective populations can all significantly impact prediction accuracies (Zhao, Chen et al. 2012).

Technological advancements and statistical frameworks used bring new challenges (*Figure 1*). Genotyping and/or phenotyping technologies can now generate millions of markers and thousands of phenotypic measurements, e.g. in time series, increasing the dimensionality of the prediction problem. For example, using a high-density SNP array (or imputing SNPs based on a low-density array) will increase the likelihood of getting many markers in LD with the true QTL (high SNP-QTL LD). It can increase total explained variance (Ogawa, Matsuda et al. 2016), but may induce multicollinearity among SNPs. In contrast, low-density genotyping can miss important SNPs in LD with, or weakly linked to, the QTLs, leading to inferior prediction performance (de Los Campos, Sorensen et al. 2019). Consequently, SNP selection prior to predictive modelling has been reported to provide superior performance compared to simply using a dense marker set (Veerkamp, Bouwman et al. 2016).

Statistical genetics approaches have traditionally focused on formulating phenotype prediction as a parametric regression of one or more phenotypes on genomic markers, treating non-genetic effects as fixed or random in a linear equation. The resulting GP models are biologically interpretable but might yield poor performance for complex phenotypes, as linear regression fails to capture the more complex relations (Pérez-Rodríguez, Gianola et al. 2012). This approach also requires proper translation of prior knowledge on the genetics underlying phenotypes into parametric distributions. Although statistical distributions can help describe genetic architecture; devising a specific distribution for each phenotype is

impractical. Therefore, many variations of linear regression were proposed by relaxing statistical assumptions; the main differences lie in their estimation framework and prior assumptions on the random effects (for an overview, see 2.4). Alternatively, machine learning (ML) offers a more general set of non-parametric methods that can model phenotypes as (non)linear combinations of genotypes. Moreover, these methods can jointly model the problem, e.g. strong learners can be stacked (Sapkota, Boatwright et al. 2020) or weak learners can be combined in an ensemble. Examples include Support Vector Machines (SVMs), (ensembles of) decision trees and artificial neural networks (ANNs). No statistical assumptions are required in advance; therefore, these methods should be able to pick up more complex genetic signals that are missed by linear models. The downside is the large amount of data required for learning these models from the data.

The performance of ML methods in GP problems has previously been compared using simulated and real phenotypes. Some were found to perform better under non-additive allelic activity (Howard, Carriquiry et al. 2014, Abdollahi-Arpanahi, Gianola et al. 2020); however, a clear link between simulated and real phenotypes is often missing, or only a specific breeding population structure is considered. Moreover, there are conflicting reports on performance of ML (Howard, Carriquiry et al. 2014, Grinberg, Orhobor et al. 2020). For example, ANNs have been reported to perform worse in some applications and comparable to competing methods in others (Bellot, de Los Campos et al. 2018, Abdollahi-Arpanahi, Gianola et al. 2020). Ensemble decision tree methods, combining the output of a large number of simple predictors, have proven better for some traits but not for others (Ogutu, Piepho et al. 2011, Ghafouri-Kesbi, Rahimi-Mianji et al. 2017, Azodi, Bolger et al. 2019). Gradient boosting showed improved performances for many real traits (Li, Zhang et al. 2018, Yan, Xu et al. 2021) but was inferior to random forests on simulated datasets (Ogutu, Piepho et al. 2011). Furthermore, the impact of population structure and low SNP-QTL LD on the performance of ML methods is still unclear.

In this paper, we investigate which GP characteristics (genetic architecture, population properties and genotype/phenotype measurement technology) a priori point to a better performance for either traditional statistical approaches or ML-based methods. We compare GP performance of two ensemble methods, Random Forests (RF) and Extreme Gradient Boosting (XGBoost), to that of linear mixed models, GBLUP, BayesA, BayesB and RKHS regression with averaged multi-Gaussian kernels. We focus on typical applications in plant breeding to explore various GP characteristics, including the ratio of the total number of markers to the number of samples ratio (p/n), genetic complexity, QTN effect size, linear (additive) vs. nonlinear (epistatic) heritabilities, sparse vs. dense genotyping and population structure.

2. Methods

2.1. Data

2.1.1. Simulations

In a first experiment, artificial genotypes were simulated, in combination with associated phenotype values. Genotype data was simulated for a diploid population with a minor allele frequency of 0.4, using a binomial distribution, where each allele was the outcome of a binomial trial. The genotype dataset was coded as {0=AA, 1=Aa, 2=aa}. To explore GP characteristics (*Figure 1*), different levels of genetic complexity and dimensionality, defined as the ratio of total number of SNPs to the sample size ($c = p/n$), were simulated. For the high dimensionality scenarios, sample size was fixed at $n = 500$, because reference populations of this size are feasible for genotyping and phenotyping in genomic selection studies. Using values of $c = \{2, 10, 20, 40, 120\}$, the number of SNPs varied up to $p = 60{,}000$ (120*500). Similarly, for the low dimensionality scenarios, the number of SNPs was fixed at $p = 500$ and sample size was varied up to $n = 3{,}000$ to arrive at $c = \{1, 1/2, 1/4, 1/6\}$. Subsequently, Quantitative Trait Nucleotides (QTNs) were randomly selected from these simulated SNP sets to generate phenotypes. We selected either 5, 50 or 100 QTNs, corresponding to a range of low to high genetic complexity, coupled with narrow-sense heritability ranging from 0.1 to 0.7. A phenotype with a high number of QTNs and low heritability is more complex than one with few QTNs and higher heritability.

Phenotype datasets were generated using the simplePHENOTYPES R package (Fernandes and Lipka 2020). Linear polygenic phenotypes were simulated using additive modes of allele effects, as follows:

$$\mathbf{y} = \beta_1\mathbf{QTN}_1 + \beta_2\mathbf{QTN}_2 + \beta_3\mathbf{QTN}_3 + \cdots + \beta_k\mathbf{QTN}_\mathrm{k} + \boldsymbol{\varepsilon} \quad (\mathbf{1})$$

Here, β_i describes the effect size of the i^{th} QTN, where $\mathbf{QTN}_i$ is a vector containing the allele dosages for the i^{th} QTN for all samples. The residuals (ε) were sampled from a normal distribution $\sim N(0, \sqrt{(1-h^2)})$. Using the narrow-sense heritability (h^2) to determine the effect sizes: each QTN is assigned an effect size of $\beta_i = h / n$. To further explore the genetic complexity, we used equation (1) to generate another set of phenotypes where the first QTN is assigned a larger effect than others, $2h$, and the remainder still $h / (n-1)$.

For nonlinear phenotypes, broad-sense heritability was set at most to 0.8, so the distribution of residuals is $N(0,\sqrt{0.2})$. We considered only epistasis to induce nonlinearity, ignoring other factors such as dominance. Adding an additional term for epistasis to equation (1) results in:

$$\mathbf{y} = \beta_1\mathbf{QTN}_1 + \beta_2\mathbf{QTN}_2 + \beta_3\mathbf{QTN}_3 + \cdots + \beta_k\mathbf{QTN}_\mathrm{k} + \beta_e(\mathbf{QTN}_{\mathrm{e1}} * \mathbf{QTN}_{\mathrm{e2}}) + \boldsymbol{\varepsilon} \quad (\mathbf{2})$$

The epistatic heritability (h^2_e) was set analogous to the additive heritability (h^2), such that, $H^2 = h^2 + h^2_e$. The *additive x additive* epistasis model was used, with only a single pairwise interaction. The epistatic effect β_e was sampled from $N(0, \sqrt{h^2_e})$, and attributed to a single interacting pair of markers ($e1$, $e2$) such that $\beta_e = \beta_{e1} \times \beta_{e2}$. We sampled this interacting pair from the set of additive QTNs; therefore, each

interacting marker will always have some main effect. As for additive phenotypes, we also created nonlinear phenotypes with one large effect QTN. The total number of settings (scenarios considered in *Table 1*) for the simulated GP characteristics is 81 per phenotype class, i.e. linear and nonlinear. For each class, phenotypes were simulated with and without large effect QTN. Thus, in total 324 (81 x 2 x 2) simulated datasets were generated. These will be referred to as 'simdata' in the text.

Table 1. Simulation scenarios.

			Linear phenotypes ($H^2 = h^2$)	**Nonlinear phenotypes** ($H^2 = h^2 + h^2_e$ = 0.8)
#Markers (*p*) / #Samples (*n*)	**Ratio (*c*)**	**#QTNs**	h^2	$h^2 + h^2_e$
500/3000	0.17	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/2000	0.25	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/1000	0.50	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
500/500	1.00	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
1000/500	2.00	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
5000/500	10.0	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
10000/500	20.0	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
20000/500	40.0	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7
60000/500	120	5, 50, 100	0.1, 0.4, 0.7	0.7+0.1, 0.4+0.4, 0.1+0.7

2.1.2. Real datasets

To compare trends observed in simulations with outcomes obtained with real traits, publicly available wheat genotype and phenotype data were taken from Norman, Taylor et al. (2017). This includes 13 traits: Biomass, Glaucousness, Grain protein, Grain yield, Greenness, Growth habit, Leaf loss, Leaf width, Normalised Difference Vegetative Index (NDVI), Physiological yellows, Plant height, Test weight (TW) and Thousand kernel weight (TKW). This particular dataset was chosen as it contains a fairly large number of genotypes (n = 10,375) each genotyped for p = 17,181 SNPs. The impact of population structure, training set size, marker density and its interaction with population structure was assessed in a study by the same authors (Norman, Taylor et al. 2018) and GBLUP prediction accuracies were reported to saturate when training set size was greater than 8,000. We used the same settings, with 5-fold cross-validation (training

set size 8,300, validation set size 2,075).

The data was generated from a small-plot field experiment for pre-screening of germplasm containing some genotypes that are sown in multiple plots, thus containing spatial heterogeneity with correlation between closely located plots and imbalance in the number of phenotypes per genotype. Soil elevation and salinity, spatial coordinates and virtual blocks (made available on request by the authors) were taken as covariates:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{Z_g g} + \boldsymbol{\varepsilon} \quad (3)$$

Here, **X** is the *n x 4* design matrix for the fixed effects and overall mean, **b** is a *4 x 1* vector of fixed effects, i.e. soil salinity and elevation; **Z** is an *n x 3* design matrix for **u** non-genetic random effects, i.e. range, row and block; $\mathbf{Z_g}$ is the *n x k* design matrix for genotypes **g** for a maximum of k replicates, and is an *n x 1* vector of residuals. The Best Linear Unbiased Predictions (BLUPs) of genotypes were used for GP; in that way, we take care of the unbalanced phenotypic records for some of the genotyped accessions, for example, some genotypes were sown at multiple blocks. Note that equation (3) does not contain any SNP information, instead only genotype accessions are used to obtain their adjusted phenotype.

2.2. Population structure analysis

To analyse the influence of population structure on the performance of different GP methods, we used a population of the *Arabidopsis thaliana* RegMap panel (Horton, Hancock et al. 2012) with known structure, containing 1,307 accessions including regional samples (*Figure S5*). Linear phenotypes were simulated using narrow-sense heritabilities h^2 = 0.1, 0.4 and 0.7, with equal effect QTNs. The genotypes, available from the Arabidopsis 250k SNP array, were further pruned for LD and minor allele frequency (MAF > 5%) using PLINK (Purcell, Neale et al. 2007). LD pruning was carried out using a window size of 500 markers, stride of 50 and pairwise r^2 threshold of 0.1, using the '--indep-pairwise' command in PLINK. This implies that a set of markers in the 500 markers window with squared pairwise correlation greater than 0.1 is greedily pruned from the window until no such pairs remain. This dataset will be referred to as 'STRUCT-simdata' in the text.

The effect of population structure was also assessed on real data: a genotype dataset of 300 out of the 1,307 RegMap accessions, phenotyped for the sodium accumulation trait with a strongly associated gene (Baxter, Brazelton et al. 2010). This should resemble one of our simulation scenarios, i.e. high heritability (e.g. h^2 = 0.7) with few QTNs (e.g. 5) of large effect. This dataset will be referred to as 'STRUCT-realdata' in the text.

To correct for population structure, we used principal components corresponding to the top ten highest eigenvalues as fixed effects in the models for GBLUP, RKHS regression, BayesA and BayesB (Hoffman 2013). Principal component analysis (PCA) was performed on the allele dosage matrix using the prcomp() method in R, with centring and scaling. For random forest and XGBoost, we used these top principal components as additional features in the models.

2.3. Analysis of SNP-QTN Linkage Disequilibrium (LD)

To explore the impact of varying LD between SNP markers and actual QTNs on the performance of GP methods, we used two other datasets: one with real genotypes and simulated phenotypes, the other with real genotypes and real traits.

For the first dataset, we selected a natural population with minimal structure, balanced LD, genotyped at roughly equal genomic spacing and mostly inbred lines: the 360 accessions in the core set of the *Arabidopsis thaliana* HapMap population (Baxter, Brazelton et al. 2010). Genotype data of 344 out of the 360 core accessions was obtained from Farooq, van Dijk et al. (2020), containing 207,981 SNPs. The phenotypes were simulated using one of the scenarios in Section 2.1.1. The total number of SNPs was kept close to the number of samples and genetic complexity was kept low, to study the impact of SNP-QTN LD only. To this end, we simulated linear phenotypes with $h^2 = 0.7$ and 5 QTNs with equal effects. Linkage disequilibrium between SNPs was calculated as squared pairwise Pearson correlation coefficient (r^2) using PLINK (Purcell, Neale et al. 2007). Input sets of 500 SNPs were selected randomly from pairs with either low LD ($r^2 \leq 0.5$) or high LD ($r^2 > 0.9$); these two sets were used to train two prediction models using each GP method: one model that trained on the QTNs that were used to generate the phenotype, another on QTN-linked SNPs (closest on the genome) instead of QTNs themselves, from the low or high LD SNPs pool. To avoid spurious correlations between SNPs in both models, non-QTN-linked SNPs were sampled from a different chromosome. We restricted the sampling of QTNs and the QTN-linked SNPs to chromosome 1, whereas the remaining non-QTN SNPs were sampled from chromosome 2. We refer to this dataset as 'LD-simdata' in the text.

For the second dataset, we used three soybean traits (HT: height, YLD: yield and R8: time to R8 developmental stage) phenotyped for the SoyNam population (Xavier, Muir et al. 2016). This dataset contains recombinant inbred lines (RILs) derived from 40 biparental populations and the set of markers have been extensively selected for the above traits. Moreover, high dimensionality is not an issue as the dataset contains 5,014 samples and 4,235 SNPs. We refer to this dataset as 'LD-soy' in the text. A complete list of datasets used in this study has been provided in *Table 2*.

Table 2. List of datasets

ID	DESCRIPTION
SIMDATA	Simulated dataset used to explore GP characteristics of trait genetic complexity, population properties and dimensionality. See Methods section 2.1.1 for details.
WHEAT	Real wheat dataset from Norman, Taylor et al. (2017) containing 13 traits of varying genetic complexity. These traits are referred to by abbreviations: BM: Biomass, PH: Plant Height, NDVI: Normalised Difference Vegetative Index, LL: Leaf Loss, LW: Leaf Width, GY: Grain Yield, GL: Glaucousness, GP: Grain Protein, Y: Physiological Yellows, TW: Test Weight of grains, TKW: Thousand Kernel Weight, GH: Growth Habit, GR: Greenness
STRUCT-SIMDATA	Real structured RegMap panel genotype data of 1,307 *Arabidopsis thaliana* accessions with simulated phenotypes data used to analyse the effect of population structure

STRUCT-REALDATA	A subset of the real *Arabidopsis thaliana* structured RegMap panel genotype data of 300 accessions with real phenotype data of the sodium accumulation trait used to analyse the effect of population structure
LD-SIMDATA	An unstructured set of 360 accessions from the core set of the *Arabidopsis thaliana* HapMap population with known genotype data and simulated phenotype data to study the impact of LD
LD-SOY	Real soybean dataset of with real phenotypes (R8, HT: height and YLD: yield) for studying the impact of low SNP-QTN LD (Xavier, Muir et al. 2016)

2.4. Models

A wide range of statistical models have been proposed for GP. Most widely applied are Linear Mixed Models (LMMs), which use whole-genome regression to tackle multicollinearity and high-dimensionality with shrinkage during parameter estimation, employing either a frequentist approach, e.g. restricted maximum likelihood (REML), or Bayesian theory (Baek, Natasha Beretvas et al. 2020). Below, we briefly describe the GP methods used in our experiments. For (semi)parametric methods, we used BGLR with default settings of hyperparameters (Pérez and de Los Campos 2014); for Random Forests, the ranger R package (Wright and Ziegler 2017); and for XGBoost, h2o4gpu (Ovsyannikov 2021).

2.4.1. Parametric models

GBLUP

The genomic best linear unbiased prediction (GBLUP) method uses a Gaussian prior with equal variance for all markers and a covariance matrix between individuals, called the genomic relationship matrix (GRM), calculated using identity by state (IBS) distances between markers for each pair of samples (VanRaden 2008). SNP effects are modelled as random effects that follow a normal distribution with zero mean and common variance, and are estimated by solving the mixed model equation:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{g} + \boldsymbol{\varepsilon} \quad (4)$$

Here, $\mathbf{g}$ is an *n x 1* vector of the total genomic value of an individual, captured by all genomic markers; $\boldsymbol{\mu}$ is the overall population mean; and $\boldsymbol{\varepsilon}$ is an *n*-vector of residuals. The genomic values $\mathbf{g}$ and residuals were assumed to be independent and normally distributed as $\mathbf{g} \sim N(0,\mathbf{G}\sigma^2_g)$, $\varepsilon \sim N(0,\mathbf{I}\sigma^2_\varepsilon)$. Here $\mathbf{G}$ is the GRM, calculated using the rrBLUP package (Endelman 2011) in R, providing variance-covariance structure for genotypes and $\mathbf{I}$ is the identity matrix. Due to the small number of estimable parameters, GBLUP is computationally fast but the assumption of normality only holds when most effects are close to zero and only a few are larger. The limitation of this approach is that it captures only linear relationships between individuals and assumption of equal variance for all marker effects may not be truly valid for many traits.

Bayesian methods

Several Bayesian methods with slight variations in their prior distributions have been proposed to model different genetic architectures (Gianola 2013) e.g. BayesA, using a scaled *t*-distribution; Bayesian LASSO or BL (Park and Casella 2008), using a double-exponential; BayesCπ (Habier, Fernando et al. 2011); and BayesBπ (Meuwissen, Hayes et al. 2001), both utilising two-component mixture priors with point mass at

zero and either a Gaussian or scaled *t*-distribution, respectively. To control the proportion of zero effect markers, the hyperparameter 'π' was set equal to 0.5, resulting in a weakly informative prior. For simplicity, we refer to BayesBπ as BayesB in the text. The model in equation (5) was solved for posterior means in both BayesA and BayesB with the only difference in priors of β_j:

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{j}^{J} \mathbf{x}_j \beta_j + \boldsymbol{\varepsilon} \quad (\mathbf{5})$$

Here, $\boldsymbol{\mu}$ is the intercept, $\mathbf{x}_j$ is an *n*-vector of allele dosages for each SNP and β_j is the effect of SNP *j* out of total SNPs *J*.

2.4.2. Semi-parametric models

Reproducing Kernel Hilbert Spaces (RKHS) regression is a general semiparametric method that models pairwise distances between samples by a Gaussian kernel and can therefore better capture nonlinear relationships than GBLUP. In fact, GBLUP is a special case of RKHS regression, with a linear kernel (De los Campos, Gianola et al. 2010, Jiang and Reif 2015). We used RKHS regression as a representative semi-parametric model, because it not only employs prior assumptions for random components in LMM equation (6), but also learns hyperparameters from the data itself:

$$\mathbf{y} = \boldsymbol{\mu} + \sum_{l=1}^{3} \mathbf{g}_l + \boldsymbol{\varepsilon} \quad (\mathbf{6})$$

In contrast to the GBLUP model (4), the RKHS regression model has three random genetic components $\mathbf{g} = \sum_{l=1}^{3} \mathbf{g}_l$, such that $\mathbf{g}_l \sim N(0, \mathbf{K}_l \sigma^2_{gl})$; where $\mathbf{K}_l$ is the kernel evaluated for the l^{th} component. This kernel matrix **K** is used as genomic relationship matrix, where $\mathbf{K} = \{ k(\mathbf{x}_i, \mathbf{x}_j) \}$ is an $n \times n$ matrix of Gaussian kernels applied to the average squared-Euclidean distance between genotypes:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-b\left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2\right)/p\right) \quad (7)$$

The kernel $k(\mathbf{x}_i, \mathbf{x}_j)$ is a covariance function that maps genetic distances between pairs of individuals $\mathbf{x}_i$ and $\mathbf{x}_j$ onto a positive real value. The hyperparameter *b*, called the bandwidth, controls the rate at which this covariance function drops with increasing distance between pairs of genotypes. Tuning this parameter for range of values between 0 and 1 could be computationally inefficient. So, instead of tuning *b*, we used a kernel averaging method (De los Campos, Gianola et al. 2010), such that multiple kernels, corresponding to possible bandwidth values *b* = {0.2, 0.5, 0.8}, were averaged.

2.4.3. Ensemble machine learning models

Random Forest

The Random Forest regressor uses an ensemble of decision trees (DTs) that are each grown using bootstrapping (random sampling with replacement of samples), and a random subset of SNPs. The test sample prediction is made by averaging all unpruned DTs as;

$$\hat{f}_{RF}^{D}(x) = \frac{1}{D}\sum_{k=1}^{D}\tau(x,\psi_k) \quad (8)$$

Here *x* is the test sample genotype using a random forest τ with *D* decision trees, for which ψ_k is the k^{th} tree. An RF has a number of hyperparameters that need to be tuned, for which we have used grid search using the caret R package (Kuhn, Wing et al. 2020). We used 500 trees in the forest for all analyses and tuned 'mtry' and 'nodesize' hyperparameters to control tree shapes. The total number of SNPs randomly selected at each tree node, i.e. mtry, was selected from { *p*/3, *p* /4, *p* /5, *p* /6 } and the minimum size of terminal nodes below which no split can be tried, i.e. nodesize, was selected from {0.01, 0.05, 0.1, 0.2, 0.3} times the number of training samples in each cross-validation fold.

Extreme Gradient Boosting (XGBoost)

We used XGBoost, a specific implementation of the Gradient Boosting (GB) method. Similar to the Random Forest, Gradient Boosting is an ensemble method, using weak learners such as DTs. The main difference is that an RF aggregates independent DTs trained on random subsets of data (bagging), whereas GB grows iteratively (boosting) by selecting samples in the subsequent DTs based on sample weights obtained in previous DTs, related to how well samples are predicted already by these previous DTs.

Hyperparameters were tuned using a grid search through 5-fold cross-validation on each training data fold. We searched over max_depth = {2, 3, 4, 50, 100, 500 }, colsample_bytree = { 0.1, 0.2, 0.3, 0.5, 0.7, 0.9 } and subsample = {0.7, 0.8, 0.9}.

2.4.4. Performance evaluation

Model performance was evaluated based on prediction accuracy, which was measured as the Pearson correlation coefficient (*r*) between observed phenotypic values and predicted genomic values of the test population. For each model, five repeats of 5-fold cross-validation were performed, so in total 25 values of *r* were used to compare performances.

2.5. Assessment of trait nonlinearity

To link GP performance in simulation scenarios with performance on real data, an assessment of the nature of real traits (i.e. linear or nonlinear) was used. To obtain a proxy for linearity of the trait, we assumed that if a trait has a higher proportion of additive variance compared to other traits, estimated with the same model, it will be more linear. To verify this on our simulated dataset scenarios (*Table 1*) for nonlinear phenotypes, we used the linear mixed model:

$$\mathbf{y} = \boldsymbol{\mu} + (\mathbf{g}_a + \mathbf{g}_r) + \boldsymbol{\varepsilon} \quad (9)$$

Here $\mathbf{g}_a$ defines a set of additive genotype effects such that $\mathbf{g}_a \sim N(0,\sigma^2_a\mathbf{G})$, where **G** is the genomic relationship matrix (GRM) calculated as described by VanRaden (2008), $\mathbf{g}_r \sim N(0,\sigma^2_r\mathbf{I})$ is the residual genetic effect and is a vector of residuals. The ratio of additive genetic variance to the residual genetic variance (σ^2_a / σ^2_r) was calculated for both the simulated dataset and real wheat traits. We tested our assumption

on simulated phenotypes (*Figure S1*), showing simulated amounts of non-additive heritability to indeed be negatively related to empirical additive heritability.

3. Results

3.1. ML outperforms traditional methods for GP

Previously, numerous GP methods were tested for different traits of varying genetic architectures using low or high density marker sets, but it is still unclear for which (class of) GP problems applying machine learning can be beneficial (Pérez-Rodríguez, Gianola et al. 2012). To investigate the role of underlying characteristics (*Figure 1*), we generated an extensive set of simulated genotype-phenotype data (simdata: see Section 2.1.1). This data was analysed using the linear parametric methods GBLUP, BayesA and BayesB; nonlinear semi-parametric regression method RKHS, using a Gaussian multi-kernel evaluated as average squared-Euclidean distance between genotypes (De los Campos, Gianola et al. 2010); and popular nonlinear ML methods, i.e. support vector regressor (SVR), random forest regressor (RF), extreme gradient boosting (XGBoost) regression trees and Multilayer Perceptron (MLP). The simulations covered a variety of trait scenarios (from simple to more complex), as shown in *Table 1*. Simple traits correspond to simulation scenarios with larger heritabilities, additive allele effects (linear) and small numbers of QTNs; complex traits can have both additive and non-additive allele effects (nonlinear) with small heritabilities and large numbers of QTNs. For linear phenotypes, narrow-sense heritability was set equal to broad-sense heritability and for the nonlinear phenotypes, the sum of narrow-sense and epistatic heritability was set equal to the broad-sense heritability. The extent of phenotypic linearity in both simulations and real datasets was calculated using the ratio of additive genetic variance to the residual genetic variance (σ^2_a/ σ^2_r) from equation (9). In the results presented below, SVR and MLP were excluded because their performances were significantly lower than the tree-based ensemble ML methods (i.e. RF and XGBoost) on a subset of our simulation scenarios (Appendix-I).

3.1.1. ML methods perform well for simple traits

Many non-mendelian plant traits are fairly simple, where only one or a few QTLs explain a large proportion of phenotypic variance. If these QTLs are identified by the GP model, prediction performance can be pretty high. In our simulations (*Table 1*), this scenario is investigated using additive phenotypes with narrow-sense heritability (h^2) equal to 0.7 and a total number of QTNs equal to 5.We then attribute equal effects to all QTNs or unequal effects by assigning a larger effect to the first QTN in equation (1) than to other QTNs (see Section 2.1.1).

The results in *Figure 2A* and *Figure 2B* illustrate that the performance of Bayesian methods and ML was better than that of genomic relationship-based methods (GBLUP, RKHS). The performance of ML methods was slightly worse than that of Bayesian methods when all QTNs shared equal effects (*Figure 2A*) but comparable when one of them had a larger effect size (*Figure 2B*). Therefore, although not outperforming the other methods, ML methods seem to be reasonable choices for simple traits.

3.1.2. ML methods outperform parametric methods for complex traits

Complex polygenic traits may contain a large effect QTL along with many small to medium effect QTLs (Goddard, Kemper et al. 2016). This makes their detection more challenging through conventional

univariate regression models that are followed by strict multiple testing corrections. Moreover, shrinkage of effect sizes in multivariate regression models restricts them from growing too large. Thus, many true small effects may be ignored in the analysis. SNPs may also have non-additive effects, which could cause a large amount of variance to remain unexplained and narrow-sense heritabilities to be low, when modelled by their linear action only.

This genetic complexity was simulated by increasing the number of QTNs, decreasing the narrow-sense heritability and keeping overall effect sizes equal, thereby letting the effect sizes per QTN become proportionally smaller. The QTNs were randomly chosen from the simulated SNPs pool by setting k equal to 50 or 100 in equation (2), keeping equal effect sizes for all QTNs and h^2 equal to 0.1. Moreover, two QTNs were randomly selected to have a fairly large pairwise interaction effect corresponding to an epistatic heritability h^2_e equal to 0.7. The results in *Figure 2C* illustrate that ML methods outperformed all methods for complex nonlinear phenotypes, although overall performance remained poor.

3.1.3. ML performance is robust to high-dimensional GP

Genomic prediction is usually employed on a genome-wide set of markers but the training population size is limited, i.e. a high dimensional problem. This comes with obscured genetic signal due to an increase in uninformative markers leading to an over/underestimation of allelic effects or genomic relationships, overfitting on training samples and reduced performance on unseen data. To investigate the susceptibility of different GP methods for this issue, we analysed how prediction accuracy varied depending on the ratio of markers vs samples ($c = p/n > 1$). The results in *Figure 2A* and *Figure 2B* illustrate that methods perform differently on high dimensional simple traits GP problems, such that genetic distance-based methods (GBLUP, RKHS) were affected more than ML and Bayesian methods. For complex traits (*Figure 2C*), prediction accuracy was quite low for all methods due to low narrow-sense heritability (h^2: 0.1), so a decreasing trend was not evident, but ML performance was still higher than all other methods. The results with different simulation settings of 'simdata' further show that performance is negatively related to an increase in dimensionality, and less affected by larger heritability (*Figure S2* and *Figure S3)*. In general, this shows that the conclusions drawn in Section 3.1.1 for simple traits and 3.1.2 for complex traits holds under high-dimensionality and when there is sufficient genetic signal (high heritabilities), high-dimensionality is less of an issue for these methods.

3.1.4. ML methods are generally suitable for nonlinear phenotypes

For complex phenotypes, we observed that ML performs well when heritability is small and non-additive allele effects are present (*Figure 2C*). In our simulation scenarios (*Table 1*) we further investigated a range of additive and non-additive fractions of heritabilities, with or without a large effect QTN.

For linear phenotypes with equal effect sizes of all QTNs, performance of ML methods was poorer than that of Bayesian methods under all scenarios; with an increase in genetic complexity (lowering h^2 and increasing the number of QTNs), performance dropped below that of GBLUP and RKHS as well (*Figure S2A*). Therefore, ML methods are not beneficial for this setting. For nonlinear phenotypes however, ML outperformed all methods including the Bayesian methods for all scenarios (*Figure S2B*) with random forests generalizing best. ML methods are thus best suited for nonlinear traits and do not necessarily need

large main effects to be present. Note that although RKHS regression has been reported to better capture epistatic relationships between markers (Jiang and Reif 2015), it did not perform well in our simulations; perhaps it needs more careful tuning of the bandwidth of the Gaussian distributions, rather than using multi-kernel averaging or require matching prior allele effects distributions (see Discussion).

For the linear phenotypes explained by a large effect QTN and many small effect QTNs (*Figure S3A*, *Figure S3B*), Bayesian methods perform comparable to ML methods for both linear and nonlinear phenotypes under all simulation scenarios, although RF gave slightly better performance for nonlinear phenotypes with large epistatic heritability (for $h^2_e = 0.7$) and dimensionality ($p/n > 2$). This could be because the large effect QTN explains most of the additive variance and is easily picked by Bayes and ML methods, but RF has the added advantage of picking up the nonlinear signal, when main effects got smaller with the increase in number of QTNs. XGBoost gave relatively poor performance, especially at smaller heritabilities (0.1 and 0.4) and larger p/n ratios, while GBLUP and RKHS regression performance was consistently poor in all scenarios. Moreover, an increase in dimensionality has a more adverse effect on the performance of genomic relationship matrix-based methods (GBLUP, RKHS) than on that of Bayes and ML.

In conclusion, our simulation results indicate that ML works well when a fair proportion of broad-sense heritability is contributed by nonlinear allele effects or a few large effect QTNs.

3.1.5. Case study in wheat

To see whether our simulation results hold on real traits, we used a dataset of 13 wheat traits (Norman, Taylor et al. 2017) for a fairly large number of samples (10,375 lines) and 17,181 markers ($c \approx 1.6$). These markers have been selected by strict screening criteria, therefore, many of them could be informative. Insights about genetic complexity for some of these traits were previously reported by Norman, Taylor et al. (2017) and Norman, Taylor et al. (2018). For example, Glaucousness was reported be a simple but grain yield to be more complex (Norman, Taylor et al. 2018). The results in *Figure 3* clearly indicate that *5*-fold cross-validated prediction accuracies (r) were higher for both ML methods when the fraction of additive variance was small (i.e. traits were fairly complex) and slightly lower or comparable to both Bayesian and GBLUP/RKHS regression methods otherwise. This is in line to what we observed in our simulations that for simple traits (*Figure 2A&B*), ML was either comparable to Bayesian or slightly poor than them, but for complex traits it was consistently better (*Figure 2C*). For example, leaf width, glaucousness, growth habit, leaf loss, plant height, test weight and thousand kernel weight traits had greater than 80% of their genetic variance explained only by additive variance components and performance of ML relative to Bayesian methods and GBLUP / RKHS regression was either at par or lower than that. On the other hand, biomass, grain protein, grain yield, yellowness and in particular NDVI had smaller fractions of additive variance and, relative to the other methods, ML performed better. Hence, results on this experimental dataset match with the findings in our simulations that ML is best suited for the prediction of more complex traits and a potential candidate for simple traits as well.

3.2. ML methods are sensitive to population structure

Population structure (PS) is a well-known confounding factor that results in decreased diversity in training populations (Norman, Taylor et al. 2018) and unrealistic inflated parameter estimates, e.g. for (co)variances of random effects in LMMs (Visscher, Hemani et al. 2014). Parametric and nonparametric ML methods, based on their modelling assumptions and approaches, may be differently sensitive to PS. To assess the impact of population structure on ML methods, we used real genotype data with a known population structure and combined it with both simulated (STRUCT-simdata) and real phenotypes (STRUCT-realdata). Only linear phenotypes were simulated, with varying complexity and dimensionality scenarios, as described earlier in Section 2.1.1. The STRUCT-simdata contains all 1,307 Arabidopsis RegMap accessions (Horton, Hancock et al. 2012). To exclude the impact of multicollinearity among SNPs, only uncorrelated markers were retained after pruning with pairwise squared correlation coefficient ($r^2 <$ 0.1, see Section 2.2), leaving 15,662 SNPs, but keeping the population structure intact (*Figure S5*). This results in a ratio $c = p/n$ of approximately 12 (15,662/1,307), a setting comparable to the simulation results presented in *Figure 2A.*

Correction for PS was carried out by including the top ten principal components corresponding to the largest eigenvalues as fixed effects into the mixed model equations or as additional features for ML methods. For the simulated phenotypes (*Figure S4*), average pairwise difference of test accuracies before and after correcting for PS was slightly higher for ML methods (RF: 0.03 and XGBoost: 0.04) than for LMMs (GBLUP: 0.01, RKHS: 0.01, BayesA: 0.01 and BayesB: 0.00). Moreover, the correction resulted into relatively elevated accuracies for the scenarios with larger number of QTNs or low heritabilities. This illustrates that with smaller #QTNs and larger heritabilities (h^2=0.7, #QTNs=5), effect sizes per QTN were larger; therefore, confounding due to PS was less of a concern. With the decrease in effect sizes per QTN (increase in #QTNs and decrease in h^2), correction became more important for reliable predictions. From this, we can argue that confounding due to PS should be generally corrected for, but particularly for complex phenotypes having low heritability and large numbers of QTNs with small-medium effect sizes.

To further explore this behaviour, we used real phenotypes of the sodium accumulation trait in *Arabidopsis thaliana* (STRUCT-realdata) using a subset of the same genotypes dataset. Here, we expected to have at least one large effect QTN for this trait, because *AtHKT1;1* locus, encoding a known sodium (Na^+) transporter, has been reported to be a major factor controlling natural variation in leaf Na^+ accumulation capacity (Baxter, Brazelton et al. 2010). Similar to the outcomes on 'STRUCT-simdata', correction for PS increased prediction accuracies of all methods on test data; whereas, GBLUP had the lowest average difference ($\Delta\mu$=0.03) between pair-wise predictions before and after correction (*Figure 4*). In contrast to 'STRUCT-simdata', XGBoost had the largest average difference ($\Delta\mu$=0.1) but RF was comparable to LMMs ($\Delta\mu$=0.05). From the above outcomes, we conclude that ML methods, like other GP methods, are sensitive to confounding due to PS and correcting for this can further improve performance for complex phenotypes. However, it is still unclear to which extent or for which GP problem characteristics different methods are more advantageous or more sensitive to PS.

3.3. ML methods can tackle low SNP-QTN LD

The utility of GP in genomic selection is based on the assumption that there are ample markers within a densely genotyped set of markers which are in LD with the QTLs (Meuwissen, Hayes et al. 2001). The actual QTNs are generally unknown, but SNPs in LD can be used to (partially) capture their effect, depending on the actual correlation and allele frequencies. Therefore, it is worthwhile to investigate the impact of of SNP-QTN correlation levels on GP performance (Uemoto, Sasaki et al. 2015). We used two settings, one with real genotypes and simulated phenotypes (LD-simdata), a second with real genotypes and real traits (LD-soy).

In simulations, GP model performance is evaluated based on the difference in prediction accuracies between a model trained on the actual QTNs and a model trained on SNPs in LD (QTN-linked SNPs). Our results show that when SNPs are highly correlated to QTNs (which is likely the case for densely genotyped markers set and $r^2 > 0.9$), all methods perform equally well and the SNP-based model predictions are very close to those of the actual QTN based models (*Figure S6*). On the other hand, for low LD between SNPs and QTNs, there was in general a difference between median prediction accuracies (Δr) of the QTN and SNP-based models (*Figure 5A*). This difference varied between methods, from 0.18 for RKHS regression to 0.43-0.46 for the Bayesian methods, with GBLUP and ML methods between these (0.32-0.37). The relative robustness of particularly the Random Forest model in these circumstances compared to the Bayesian methods, in combination with its good performance in many simulations, supports its usefulness for GP.

As a real genotype and phenotype dataset, we used three Soybean traits, i.e. height, time to R8 developmental stage and yield (LD-soy). The complete set of markers (4,235 SNPs) had many correlated SNPs, such that only 261 were left with low LD ($r^2 \leq 0.5$). Here, in contrast to LD-simdata where we knew the QTNs in advance, we assumed that many SNPs could be linked to QTNs, because ~94% of all markers had $r^2 > 0.5$. So, we compared two models: one with all markers (the benchmark model), and one with low LD ($r^2 \leq 0.5$). A similar pattern was observed, as shown in *Figure 5B*, i.e. RKHS regression, RF and XGBoost were most robust against low SNP-QTN LD, with negligible differences between median accuracies, where GBLUP and the Bayes methods had higher differences.

In conclusion, GP methods that model SNP-QTN or SNP-SNP relation as a nonlinear function (RKHS, RF, XGBoost) were more stable under low SNP-QTN LD compared to other methods (GBLUP, BayesA, BayesB). Moreover, RF seems to couple good prediction performance with reliability under low SNP-QTN LD.

4. Discussion

4.1. There is room for ML in genomic prediction

Genomic prediction has long been the realm of parametric methods, but recently nonlinear supervised ML methods have become increasingly popular. Yet literature is unclear on the characteristics of GP problems that warrant application of ML methods. This study fills this gap and concludes that nonlinear tree-based ensemble ML methods, especially Random Forests, can outperform traditional methods for

simple as well as complex polygenic traits where nonlinear allele effects are present. Moreover, ML methods are robust to high dimensionality, although further improvements, e.g. statistical or prior knowledge driven regularization, can further improve performance. ML methods are particularly useful compared to the frequently used GBLUP and RKHS regression given their higher performance. While Bayesian methods often perform on par with ML models, this is mainly when there are large effect QTNs and/or linear phenotypes. Moreover, Bayesian methods are prone to overfitting in case of small sample sizes ($p/n > 1$), which is less of an issue with ML, especially with RF (*Figure S7A*, *Figure S7B*).

4.2. Tree-based ensemble ML methods are a reasonable choice for GP

A wide range of parametric, semi-parametric and nonparametric methods can be used for GP, but it is impractical to test all for a particular application. The choice for a suitable method strongly depends on the GP problem characteristics, described in *Figure 1*. While GP methodology can be compared using various model evaluation metrics (BIC, AIC, log likelihoods), we focused on their utility from a breeder's perspective, so we compared only their prediction accuracies. We found that GP methods based on modelling the distance between genotypes using a covariance structure(s), inferred from genomic markers (GBLUP and RKHS), were generally inferior to Bayesian and ML methods and less robust to high-dimensional problems. The reason could be the increasing ratio of noisy or unrelated markers to informative markers (QTNs) due to increase in marker density; leading to overestimation of genetic distances.

The parametric LMM equations can be solved using Bayesian framework. Bayesian methods define prior SNP effects distribution(s) to model different genetic architectures. Instead of a single distribution for all marker effects (e.g. BRR), it could be defined for each individual marker (e.g. BayesA). Moreover, mixture distributions have also been proposed (e.g. BayesC, BayesB). From the Bayesian alphabet, we used BayesA and BayesB as representatives because the first scenario i.e. single distribution for all markers has been covered by GBLUP. Our results illustrate that these methods outperform GBLUP and RKHS regression when large effect QTNs are present, for both linear and nonlinear phenotypes. On the other hand, tree-based ensemble ML methods had either comparable performance to Bayesian methods (for simple traits) or superior performance (for complex traits). Capitalising on the results from Appendix-I that these ML methods had better performances than other ML methods (SVR and MLP), we can argue that these tree-based ML methods are a reasonable choice to conduct GP.

4.3. Population structure analysis

Population structure can affect GP performance. Our results show that without correcting for population structure, test accuracies were lower than after correction for all methods. However, ML seems to be slightly more sensitive because the average difference between each pairwise test data accuracies was higher than other methods in the simulated data.

Confounding due to population structure can also be due to the frequently employed random cross-validation strategy for predictive modelling (Norman, Taylor et al. 2018). In random cross-validation, the

reference population is randomly divided into subsets, one of which is iteratively selected for testing while the remaining subsets are used to train the model. While samples are all part of a test set once, under population structure some subpopulations may be over or under-represented in the training set. As a result, the model may get overfitted. A solution could be to use stratified sampling instead. On the other hand, parameter estimation may get misguided by within subgroup allele frequency differences rather than the overall true phenotype associated variance.

The impact of population structure can be dealt with in many ways. Conventionally, principal components of the SNP dosages or genomic relationship matrix are introduced as fixed effects in the mixed model equations (Patterson, Price et al. 2006, Guo, Tucker et al. 2014, Bermingham, Pong-Wong et al. 2015). Alternatively, phenotypes and genotypes can be adjusted by the axis of variations before predictive modelling (Zhao, Chen et al. 2012). Nevertheless, some residual structure often remains in the datasets, so it is important to check sensitivity of GP models to this confounding factor. Since ML methods (RF and XGBoost) do not employ any statistical prior and learn the association patterns from the data itself, they may be more sensitive to structure, as evident from our simulation results. But this is not clearly evident from the real phenotypes, so we cannot generalize this conclusion from our simulations.

4.4. Effect of SNP-QTN linkage disequilibrium

Despite technological improvements, low density SNP panels are usually cost-effective for routine genomic selection. Increasing marker density does not necessarily increase prediction accuracy, since accuracy is not a linear function of SNP density only (Technow, Riedelsheimer et al. 2012, Wang, Yu et al. 2017, Zhang, Wang et al. 2017). Instead, many GP problem characteristics (*Figure 1*) jointly affect performance. However, using low density SNP panels can negatively affect prediction performance, since relevant SNPs in LD with the QTLs can either be completely missing or SNPs only in low LD may be present. As a result, allele frequencies between SNPs and QTNs can be quite different, resulting in incorrect predictions (Uemoto, Sasaki et al. 2015). Despite this, low SNP density can still be sufficient for populations with larger LD blocks, e.g. F2 populations, where QTL detection power is highest and in this case, we shouldn't expect much improvement by increasing marker density. But it becomes an important consideration when LD starts to decay and population relatedness decreases in the subsequent crosses of the breeding cycle. In this context, our study addresses the question of whether certain GP methods, especially ML, are more sensitive to low SNP-QTL LD. The results using both simulated and real traits indicate that SNP-QTL LD could also be an important determinant of suitable GP methodological choice and that ML is robust against low LD.

A weak SNP-QTL correlation implies that the SNP is a weak predictor of phenotype and there is an imperfect match between the genotypic distribution and the actual underlying genetic distribution of the phenotype. When using penalised regressions, this can result in different shrinkage for the SNP than that required by the actual QTN, thereby leading to a low genetic variance attribution to that SNP. Therefore, we may expect better prediction by nonparametric ML methods, as they may better learn weak genetic signals and are more robust to low SNP-QTL LD problems. On the other hand, the semiparametric RKHS regression method measures genetic similarity between individuals by a nonlinear Gaussian kernel of SNP markers and also performed better than GBLUP and Bayesian methods under low SNP-QTN LD. The reason

could be that under low SNP-QTN LD, true pair-wise genetic covariance estimation would be less accurate due to losing many important markers and considering all of them equal contributors towards total genetic covariance. In case of RKHS regression, a Gaussian distribution defines a SNP's probable contribution towards total genetic covariance, which becomes more realistic in this scenario because fewer important SNPs are left than in the high SNP-QTN LD case. The Bayesian methods (BayesA and BayesB) had the largest decrease in test performance under low SNP-QTN LD compared to high SNP-QTN LD. This could be due to the application of penalties on individual marker effects, which shrinks the weak SNP-QTN associations towards zero for each SNP.

4.5. ML outperformed parametric methods for predicting complex wheat traits

Bread wheat breeding has huge impact on worldwide food security and socio-economic development (Tessema, Liu et al. 2020). Therefore, minor improvements in GP methodology leading to overall genetic gain can have high impact. In this study, we used a large (10,375 lines) Australian germplasm panel, genotyped with a high quality custom Axiom^{TM} Affymetrix SNP array and phenotyped for multiple traits with varying complexity levels (Norman, Taylor et al. 2017). The authors showed that genomic selection was superior to marker-assisted selection (MAS) by employing GBLUP with two random genetic components (referred to as full-model in their text). Our results clearly indicate that ML can perform well for complex bread wheat traits, e.g. grain yield, yellows, greenness, biomass and NDVI. All of these traits except grain yield can be measured using high-throughput automated phenotyping (Rabab, Breen et al. 2021). This is an interesting finding since, with the rapid advances in low cost high-throughput phenotyping systems, attention is shifting towards measuring component traits, e.g. vegetative indices, rather than final yields. ML methods can predict these traits more accurately, as evident from our analysis.

5. Conclusion and outlook

Based on simulated and real data, we conclude that ML can be useful for GP for both simple and complex traits. Moreover, ML can work for both low- and high-density genotyped populations and can offer methods of choice for practical plant breeding. However, proper correction for population structure should be applied to obtain stable accuracies. Tree-based ML methods, especially RF, are a good first choice for GP given their generalization performance and their ability to work with high dimensional genotype data. It would be interesting to investigate to what extent ML methods can benefit from statistical or prior knowledge-based regularization techniques.

6. Acknowledgments

The authors are grateful for the support of both WUR and NIBGE to conduct this study.

List of Tables

List of Figures

Factors affecting genomic prediction performance, often measured as correlation between true phenotype values and those predicted by a model.

Performance of parametric (GBLUP), semi-parametric regression (RKHS), parametric Bayesian (BayesA, BayesB) and nonparametric ML (RF and XGBoost) methods as average accuracy over *5*-fold cross-validation of test data. Here accuracy is defined as Pearson correlation coefficient between true and predicted values. Each panel is a subset of the simulated scenarios in 'simdata' for a particular heritability and #QTNs. The ratio of the number of markers to the number of samples ($c = p / n$) increases from left to right in each subplot. A) Simple traits, simulated as linear polygenic phenotypes with only additive effects such that #QTNs is equal to 5 and h^2 is 0.7, using equation (1), with all QTNs having equal effects. The largest standard error of mean for all values of *c* for each of the model was 0.023, 0.018, 0.007, 0.008, 0.018 and 0.009 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; B) similar to A, except one of the QTN has a large effect than others. The largest standard error of mean for all values of *c* for each of the model was 0.022, 0.022, 0.006, 0.007, 0.006 and 0.008 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; C) Complex traits have been simulated as nonlinear polygenic phenotypes with both additive and epistatic effects such that #QTNs equal to 50 and h^2 is equal to 0.1, using equation (2), such that all QTNs had equal additive effects. Two of the QTNs were attributed to the epistatic effect such that Broad-sense heritability was set to 0.8 ($H^2 = h^2 + h^2_e = 0.8$). The largest standard error of mean for all values of *c* for each of the models was 0.03 (not shown).

Top: prediction accuracies for GP models on wheat traits, reported as the mean Pearson correlation coefficient (*r*) of 5-fold cross-validation. Trait abbreviations are given in *Table 2*. Bottom: fraction of additive to residual genetic variance calculated using equation (9) for each trait. Traits were sorted in ascending order of additive variance fraction (left to right); therefore, the leftmost trait (NDVI) can be considered more complex than those to the right.

Boxplots present Pearson correlation coefficients (*r*) found in *5*-fold cross-validation, on test data from 'STRUCT-realdata'. Here $\Delta\mu$ is the average difference between pairwise predictions before and after correction and for each model, the nonparametric Wilcoxon rank sum test was used to assess statistical significance.

Prediction accuracy of different GP methods on simulated (A) and real soybean (B) datasets for high and low LD between SNPs and actual QTNs. The difference in median accuracies between these scenarios is indicated as Δr.

Supplementary Information

Supplementary Figures

Appendix-I

Figures

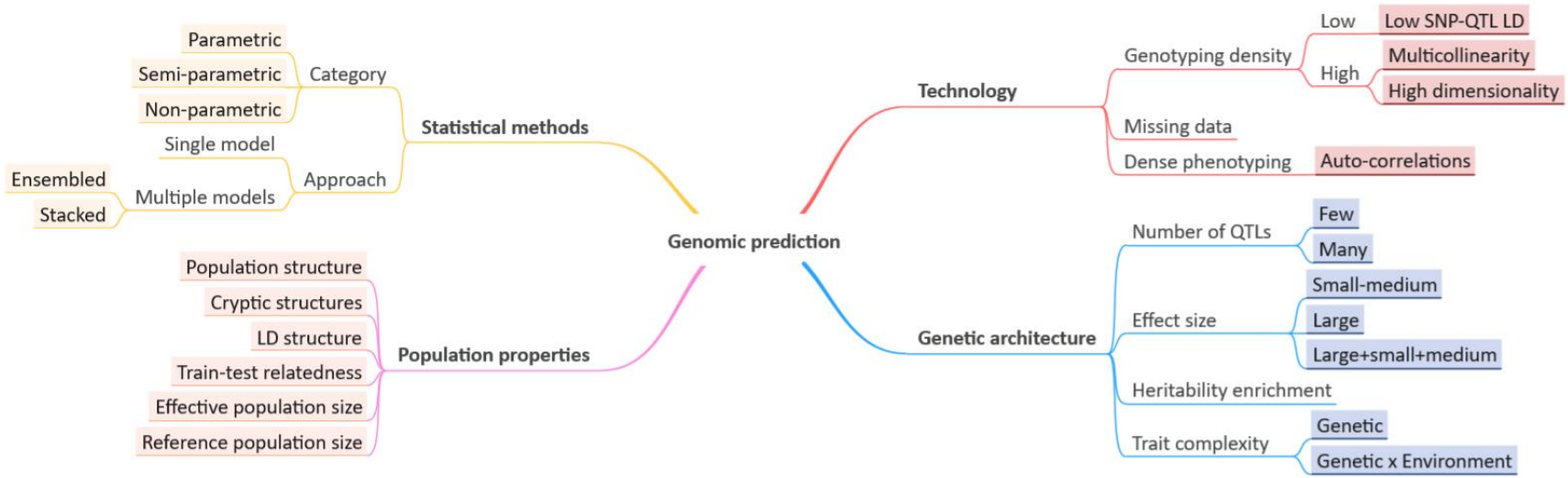


Figure 1. Genomic prediction characteristics. Factors affecting genomic prediction performance, often measured as correlation between true phenotype values and those predicted by a model.

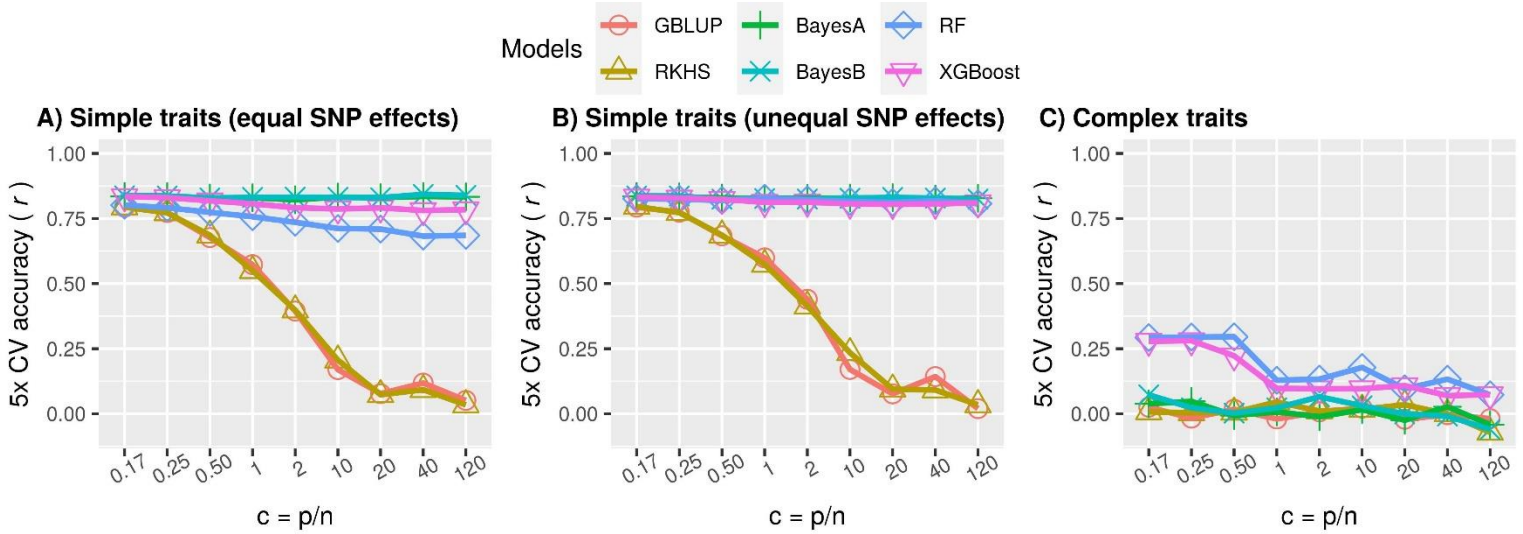


Figure 2. Comparison of prediction performances using simulated simple and complex phenotypes. Performance of parametric (GBLUP), semi-parametric regression (RKHS), parametric Bayesian (BayesA, BayesB) and nonparametric ML (RF and XGBoost) methods as average accuracy over *5*-fold cross-validation of test data. Here accuracy is defined as Pearson correlation coefficient between true and predicted values. Each panel is a subset of the simulated scenarios in 'simdata' for a particular heritability and #QTNs. The ratio of the number of markers to the number of samples ($c = p / n$) increases from left to right in each subplot. A) Simple traits, simulated as linear polygenic phenotypes with only additive effects such that #QTNs is equal to 5 and h^2 is 0.7, using equation (1), with all QTNs having equal effects. The largest standard error of mean for all values of *c* for each of the model was 0.023, 0.018, 0.007, 0.008, 0.018 and 0.009 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; B) similar to A, except one of the QTN has a large effect than others. The largest standard error of mean for all values of *c* for each of the model was 0.022, 0.022, 0.006, 0.007, 0.006 and 0.008 for GBLUP, RKHS, BayesA, BayesB, RF and XGBoost respectively; C) Complex traits have been simulated as nonlinear polygenic phenotypes with both additive and epistatic effects such that #QTNs equal to 50 and h^2 is equal to 0.1, using equation (2), such that all QTNs had equal additive effects. Two of the QTNs were attributed to the epistatic effect such that Broad-sense heritability was set to 0.8 ($H^2 = h^2 + h^2_e = 0.8$). The largest standard error of mean for all values of *c* for each of the models was 0.03 (not shown).

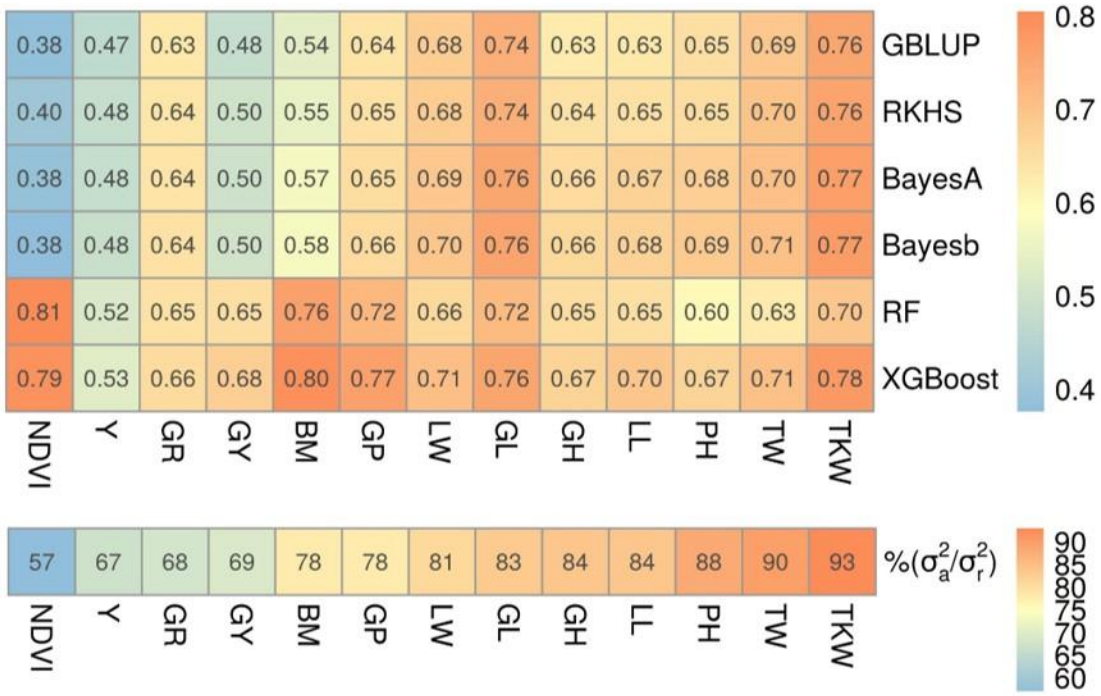


Figure 3. Prediction accuracies of wheat traits. Top: prediction accuracies for GP models on wheat traits, reported as the mean Pearson correlation coefficient (*r*) of 5-fold cross-validation. Trait abbreviations are given in *Table 2*. Bottom: fraction of additive to residual genetic variance calculated using equation (9) for each trait. Traits were sorted in ascending order of additive variance fraction (left to right); therefore, the leftmost trait (NDVI) can be considered more complex than those to the right.

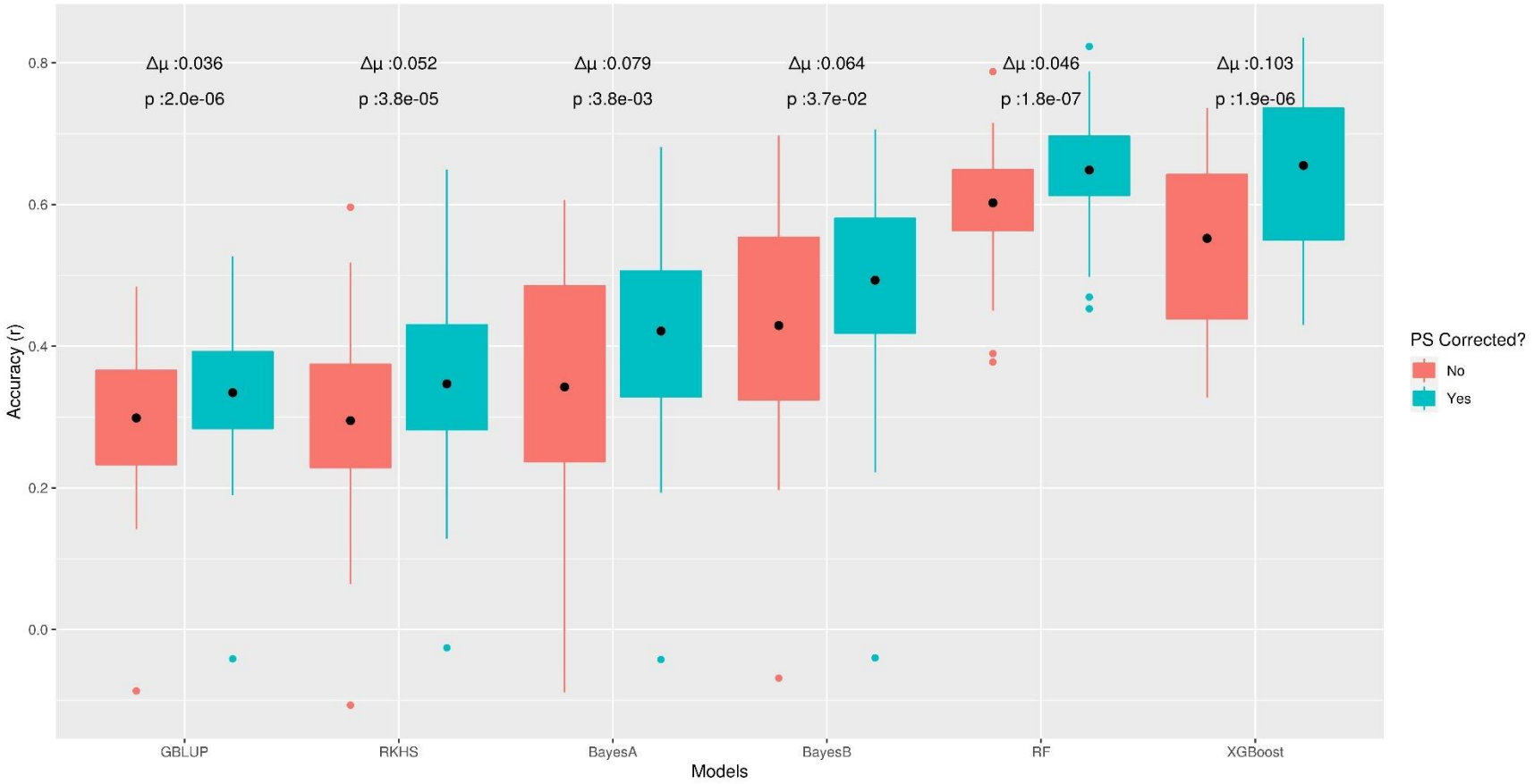
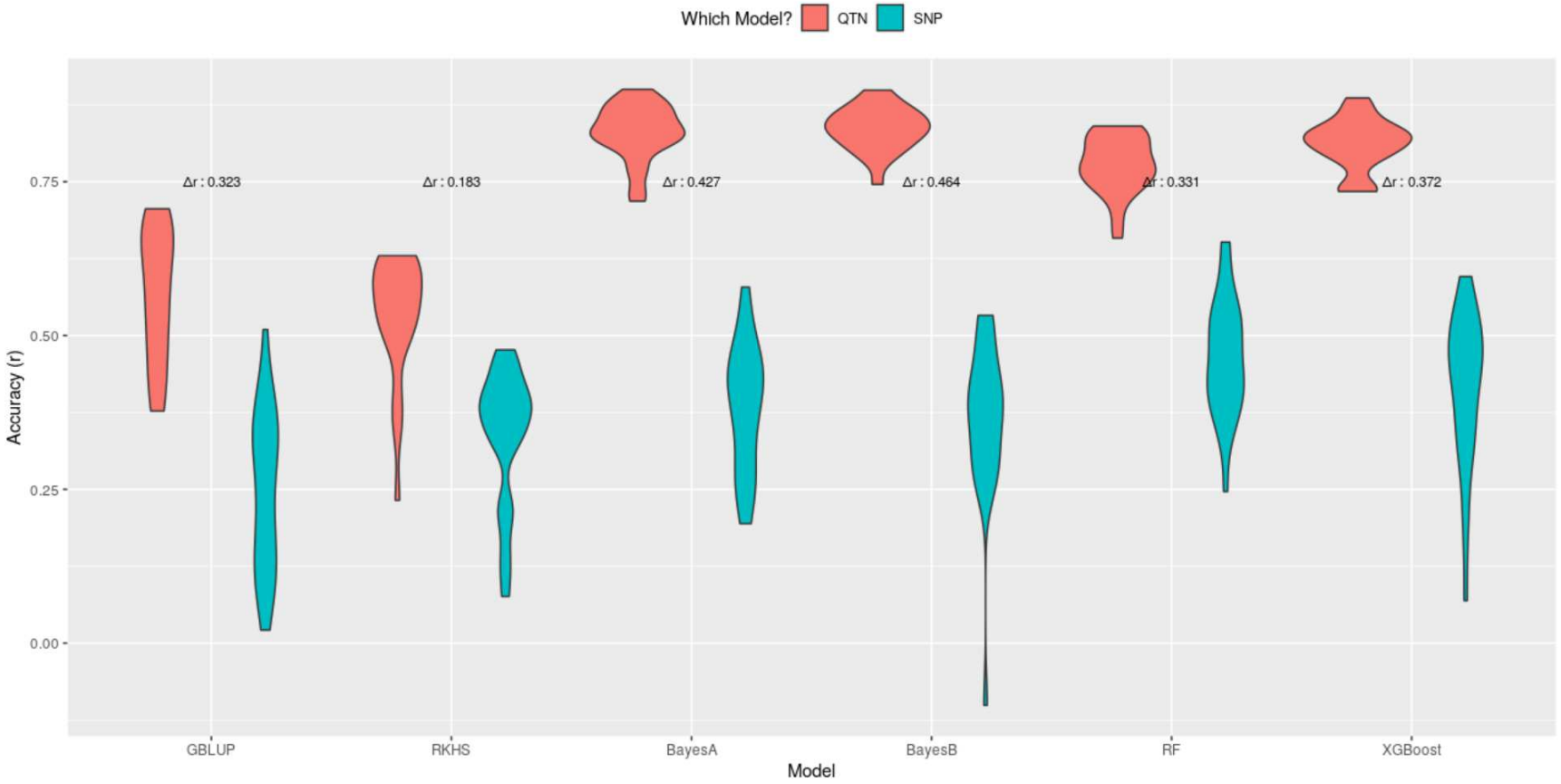


Figure 4. Effect of correction for population structure for the sodium accumulation trait in *Arabidopsis thaliana*. Boxplots present Pearson correlation coefficients (*r*) found in *5*-fold cross-validation, on test data from 'STRUCT-realdata'. Here Δμ is the average difference between pairwise predictions before and after correction and for each model, the nonparametric Wilcoxon rank sum test was used to assess statistical significance.

A. LD-sim data, low SNP-QTN LD ($r^2 \leq 0.5$).



B. LD-soy data, low ($r^2 \leq 0.5$) SNP-QTN LD vs. all SNPs (high LD).

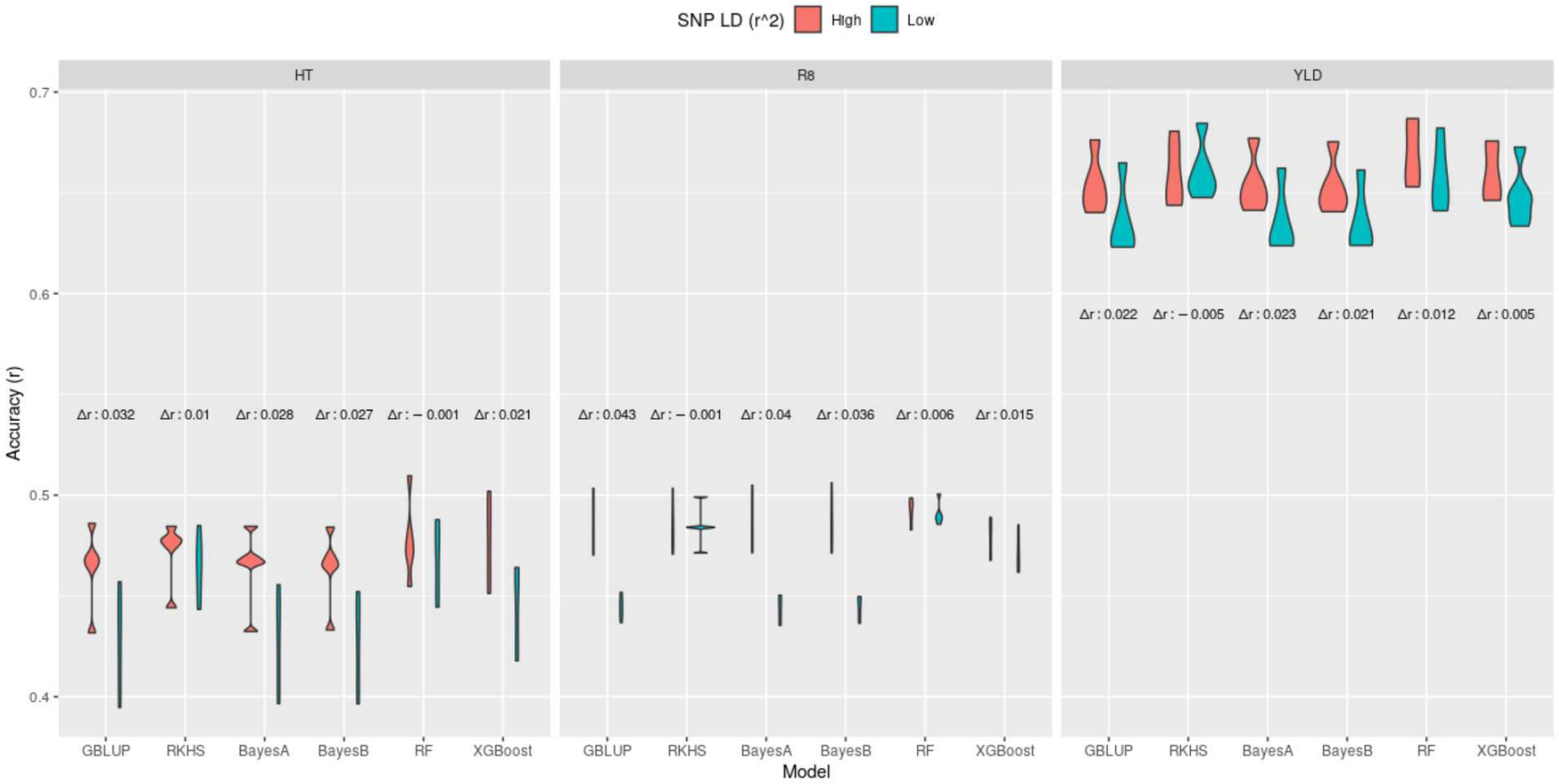


Figure 5. Effect of SNP-QTN LD on prediction accuracy. Prediction accuracy of different GP methods on simulated (A) and real soybean (B) datasets for high and low LD between SNPs and actual QTNs. The difference in median accuracies between these scenarios is indicated as Δr.

References

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

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**(11): 3691-3702.

Baek, E., S. Natasha Beretvas, W. Van den Noortgate and J. M. Ferron (2020). "Brief Research Report: Bayesian Versus REML Estimations With Noninformative Priors in Multilevel Single-Case Data." The Journal of Experimental Education **88**(4): 698-710.

Baxter, I., J. N. Brazelton, D. Yu, Y. S. Huang, B. Lahner, E. Yakubova, Y. Li, J. Bergelson, J. O. Borevitz and M. Nordborg (2010). "A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1; 1." PLoS Genetics **6**(11): e1001193.

Bellot, P., G. de Los Campos and M. Pérez-Enciso (2018). "Can deep learning improve genomic prediction of complex human traits?" Genetics **210**(3): 809-819.

Bermingham, M. L., R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro and C. S. Haley (2015). "Application of high-dimensional feature selection: evaluation for genomic prediction in man." Sci Rep **5**: 10312.

De los Campos, G., D. Gianola, G. J. Rosa, K. A. Weigel and J. Crossa (2010). "Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods." Genetics Research **92**(4): 295-308.

de Los Campos, G., D. A. Sorensen and M. A. Toro (2019). "Imperfect linkage disequilibrium generates phantom epistasis (& perils of big data)." G3: Genes, Genomes, Genetics **9**(5): 1429-1436.

Endelman, J. B. (2011). "Ridge regression and other kernels for genomic selection with R package rrBLUP." The plant genome **4**(3).

Farooq, M., A. D. van Dijk, H. Nijveen, M. G. Aarts, W. Kruijer, T.-P. Nguyen, S. Mansoor and D. d. Ridder (2020). "Prior biological knowledge improves genomic prediction of growth-related traits in Arabidopsis thaliana." Frontiers in Genetics **11**: 1810.

Fernandes, S. B. and A. E. Lipka (2020). "simplePHENOTYPES: SIMulation of Pleiotropic, Linked and Epistatic PHENOTYPES." bioRxiv: 2020.2001.2011.902874.

Ghafouri-Kesbi, F., G. Rahimi-Mianji, M. Honarvar and A. Nejati-Javaremi (2017). "Predictive ability of Random Forests, Boosting, Support Vector Machines and Genomic Best Linear Unbiased Prediction in different scenarios of genomic evaluation." Animal Production Science **57**(2): 229-236.

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

Goddard, M., K. Kemper, I. MacLeod, A. Chamberlain and B. Hayes (2016). "Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture." Proceedings of the Royal Society B: Biological Sciences **283**(1835): 20160569.

Grinberg, N. F., O. I. Orhobor and R. D. King (2020). "An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat." Mach. Learn. **109**(2): 251-277.

Guo, Z., D. M. Tucker, C. J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang and G. Gay (2014). "The impact of population structure on genomic prediction in stratified populations." Theor Appl Genet **127**(3): 749-762.

Habier, D., R. L. Fernando, K. Kizilkaya and D. J. Garrick (2011). "Extension of the bayesian alphabet for genomic selection." BMC Bioinformatics **12**: 186.

Hoffman, G. E. (2013). "Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions." PLOS ONE **8**(10): e75707.
Horton, M. W., A. M. Hancock, Y. S. Huang, C. Toomajian, S. Atwell, A. Auton, N. W. Muliyati, A. Platt, F. G. Sperone, B. J. Vilhjalmsson, M. Nordborg, J. O. Borevitz and J. Bergelson (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel." Nature Genetics **44**(2): 212-216.
Howard, R., A. L. Carriquiry and W. D. Beavis (2014). "Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures." G3 (Bethesda) **4**(6): 1027-1046.
Jiang, Y. and J. C. Reif (2015). "Modeling Epistasis in Genomic Selection." Genetics **201**(2): 759-768.
Korte, A. and A. Farlow (2013). "The advantages and limitations of trait analysis with GWAS: a review." Plant Methods **9**: 29.
Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer and B. Kenkel (2020). "caret: Classification and Regression Training. R package version 6.0-86." Avaliable at: https://cran. r-project. org/web/packages/caret/caret. pdf (accessed March 20, 2020).
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.
Meuwissen, T. H. E., B. Hayes and M. Goddard (2001). "Prediction of total genetic value using genome-wide dense marker maps." Genetics **157**(4): 1819-1829.
Moore, J. H., R. Amos, J. Kiralis and P. C. Andrews (2015). "Heuristic identification of biological architectures for simulating complex hierarchical genetic interactions." Genet Epidemiol **39**(1): 25-34.
Norman, A., J. Taylor, J. Edwards and H. Kuchel (2018). "Optimising Genomic Selection in Wheat: Effect of Marker Density, Population Size and Population Structure on Prediction Accuracy." G3 (Bethesda).
Norman, A., J. Taylor, E. Tanaka, P. Telfer, J. Edwards, J. P. Martinant and H. Kuchel (2017). "Increased genomic prediction accuracy in wheat breeding using a large Australian panel." Theor Appl Genet **130**(12): 2543-2555.
Ogawa, S., H. Matsuda, Y. Taniguchi, T. Watanabe, Y. Sugimoto and H. Iwaisaki (2016). "Estimation of variance and genomic prediction using genotypes imputed from low-density marker subsets for carcass traits in Japanese black cattle." Anim Sci J **87**(9): 1106-1113.
Ogutu, J. O., H. P. Piepho and T. Schulz-Streeck (2011). "A comparison of random forests, boosting and support vector machines for genomic selection." BMC Proceedings **5 Suppl 3**: S11.
Ovsyannikov, Y. T. a. N. G. a. E. L. a. V. (2021). "h2o4gpu: Interface to 'H2O4GPU'."
Park, T. and G. Casella (2008). "The Bayesian Lasso." Journal of the American Statistical Association **103**(482): 681-686.
Patterson, N., A. L. Price and D. Reich (2006). "Population Structure and Eigenanalysis." PLOS Genetics **2**(12): e190.
Pérez-Rodríguez, P., D. Gianola, J. M. González-Camacho, J. Crossa, Y. Manès and S. Dreisigacker (2012). "Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat." G3: Genes, Genomes, Genetics **2**(12): 1595-1605.
Pérez, P. and G. de Los Campos (2014). "Genome-wide regression and prediction with the BGLR statistical package." Genetics **198**(2): 483-495.
Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker and M. J. Daly (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American journal of human genetics **81**(3): 559-575.
Rabab, S., E. Breen, A. Gebremedhin, F. Shi, P. Badenhorst, Y.-P. P. Chen and H. D. Daetwyler (2021). "A New Method for Extracting Individual Plant Bio-Characteristics from High-Resolution Digital Images." Remote Sensing **13**(6): 1212.

Sapkota, S., J. L. Boatwright, K. Jordan, R. Boyles and S. Kresovich (2020). "Multi-Trait Regressor Stacking Increased Genomic Prediction Accuracy of Sorghum Grain Composition." Agronomy **10**(9): 1221.
Speed, D. and D. J. Balding (2019). "SumHer better estimates the SNP heritability of complex traits from summary statistics." Nat Genet **51**(2): 277-284.
Technow, F., C. Riedelsheimer, T. A. Schrag and A. E. Melchinger (2012). "Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects." Theor Appl Genet **125**(6): 1181-1194.
Tessema, B. B., H. Liu, A. C. Sørensen, J. R. Andersen and J. Jensen (2020). "Strategies Using Genomic Selection to Increase Genetic Gain in Breeding Programs for Wheat." Frontiers in Genetics **11**(1538).
Uemoto, Y., S. Sasaki, T. Kojima, Y. Sugimoto and T. Watanabe (2015). "Impact of QTL minor allele frequency on genomic evaluation using real genotype data and simulated phenotypes in Japanese Black cattle." BMC Genetics **16**(1): 134.
VanRaden, P. M. (2008). "Efficient methods to compute genomic predictions." Journal of Dairy Science **91**(11): 4414-4423.
Veerkamp, R. F., A. C. Bouwman, C. Schrooten and M. P. Calus (2016). "Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle." Genet Sel Evol **48**(1): 95.
Visscher, P. M., G. Hemani, A. A. E. Vinkhuyzen, G.-B. Chen, S. H. Lee, N. R. Wray, M. E. Goddard and J. Yang (2014). "Statistical Power to Detect Genetic (Co)Variance of Complex Traits Using SNP Data in Unrelated Samples." PLOS Genetics **10**(4): e1004269.
Wang, Q., Y. Yu, J. Yuan, X. Zhang, H. Huang, F. Li and J. Xiang (2017). "Effects of marker density and population structure on the genomic prediction accuracy for growth trait in Pacific white shrimp Litopenaeus vannamei." BMC Genet **18**(1): 45.
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**(1): 1 - 17.
Xavier, A., W. M. Muir and K. M. Rainey (2016). "Assessing Predictive Properties of Genome-Wide Selection in Soybeans." G3 **6**(8): 2611-2616.
Yan, J., Y. Xu, Q. Cheng, S. Jiang, Q. Wang, Y. Xiao, C. Ma, J. Yan and X. Wang (2021). "LightGBM: accelerated genomically designed crop breeding through ensemble learning." Genome Biology **22**(1): 271.
Zhang, A., H. Wang, Y. Beyene, K. Semagn, Y. Liu, S. Cao, Z. Cui, Y. Ruan, J. Burgueno, F. San Vicente, M. Olsen, B. M. Prasanna, J. Crossa, H. Yu and X. Zhang (2017). "Effect of Trait Heritability, Training Population Size and Marker Density on Genomic Prediction Accuracy Estimation in 22 bi-parental Tropical Maize Populations." Front Plant Sci **8**: 1916.
Zhao, Y., F. Chen, R. Zhai, X. Lin, Z. Wang, L. Su and D. C. Christiani (2012). "Correction for population stratification in random forest analysis." International Journal of Epidemiology **41**(6): 1798-1806.

Statements & Declarations

Funding

MF was supported by the sandwich PhD programme of Wageningen University & Research (WUR).

Competing Interests

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

Author Contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Muhammad Farooq. The first draft of the manuscript was written by Muhammad Farooq and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability

All datasets analysed during the current study are already published and publicly available and references to their authors or repositories have been mentioned in the text. All data and scripts have been uploaded to the Wageningen University & Research git server (https://git.wur.nl/faroo002/pub2).

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent to publish

Not applicable.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AppendixIHNDdRADJvDDdRHN.docx
- SupplementaryFiguresDdRHNADMFADDdRHN.docx