Genomic prediction in polyploid crops

Master Thesis

Student: Jos de Valk
Student nr.: 01050767
MSc plant sciences – Plant Breeding and Genetic Reseources
Supervisors:
Dr. Giorgio Tumino
Dr. Roeland Voorrips

Genomic prediction in polyploids crops

Course

Title:	Thesis-research Practice
Code:	PBR80364
Credits:	36
Submission date:	10-3-2023

Author

Name:	Jos de Valk
Student number:	01050767
Programme:	MSc Plant Sciences – Plant Breeding and Genetic Resources
Email:	jos.devalk@wur.nl

Supervisor

Name:	Dr. Giorgio Tumino
Email:	giorgio.tumino@wur.nl
Name:	DR. Roeland Voorrips
Email:	roeland.voorrips@wur.nl

Institute

Name:	Wageningen UR Plant Breeding
Address:	6708 PB Wageningen

Figure front page: Shutter2U/Getty Images

Abstract

Since its proposition in 2001, genomic prediction (GP) has been the subject of many plant breeding studies. The majority of these studies, however, focussed on diploid organisms, and it is unsure whether their findings can be translated directly into polyploid crops. This study investigated the performance of four GP models when using polyploid data and observed how their predictive ability was affected by trait architecture, heritability, marker density and population structure.

We simulated both genomic and phenotypic data. The genomic data represented an autotetraploid organism with four chromosomes, and the LD structure of the population mimics that of potato varieties from 1945-1974. For the phenotypic data, a total of fifteen traits were simulated; three of which were simulated to correlate with population structure and had their specific trait architectures and heritability, and the remaining twelve traits had different combinations of four trait architectures and three heritability values. The four GP models under investigation were two variables selection models, BayesB and Elastic Net and two shrinkage models, RKHS and RR-BLUP.

We observed a clear distinction between the models that apply variable selection and those that apply shrinkage. The variable selection models were more reactive to trait architecture, heritability and marker density than the shrinkage models. Variable selection models function by identifying markers in LD with the QTL. We observed that this was more beneficial when the trait was governed by few QTLs, when the heritability is high and when there is sufficient LD between marker and QTL. The predictive ability of the shrinkage models was mainly determined by heritability and marker density and it was unaffected by trait architecture.

For the trait simulated to correlate with population structure, there was a clear indication that the population structure inflated the model's predictive ability. This is an important consideration when breeding. By crossing between or selecting within subgroups, the population structure can disappear, and the predictive ability will deflate. If unaccounted for, this deflation of predictive ability may seriously affect selection accuracy in a breeding programme. We recommend thoroughly investigating the effect of population structure for each new trait and adjusting the breeding strategy accordingly.

Especially for the mono- and oligogenic traits, the variable selection models were the best option. The differences between the variable selection models were insignificant, and no model outperformed the other. However, Elastic Net had some problems predicting the polygenic and mix traits with low heritability, making BayesB the best option. For the traits that contained polygenic effects, the predictive ability of the models was on par, but the computation time of RKHS and RR-BLUP was much shorter than that of BayesB and RR-BLUP.

The results found in this study correspond to what was found in literature. However, we also realise that the predictive ability of GP models is affected by many variables, and the results may vary for different populations, marker sets and traits. Therefore, we recommend continuing these comparisons in polyploid data with different population and LD structures and a wider range of traits and marker densities.

Contents

1 Introduction

Marker-assisted breeding (MAB) has evolved significantly since its first applications in the 70's and 80's (Edwards et al., 1987; Paterson et al., 1988; Soller, 1978; Soller & Plotkin-Hazan, 1977). Despite high expectations, early attempts had a weak impact on plant breeding due to high marker costs, low density marker maps, insufficient selection accuracy and limited availability of software packages (Collard & Mackill, 2008). A turning point for MAB was the development and advancements of high-throughput DNA sequencing techniques which made it possible to detect thousands of markers, covering the whole genome, at much lower cost (Baird et al., 2008; Davey et al., 2011; Elshire et al., 2011; Gupta et al., 2008). These new sequencing techniques also gave rise to more comprehensive genome based breeding techniques such as genome wide association studies (Loos, 2020) and genomic prediction (GP; Meuwissen et al., 2001).

1.1 Genomic prediction

GP was first proposed by Meuwissen et al. (2001) as a better way to utilise the increasingly dense marker data. They hypothesised that quantitative traits are affected by many genes and therefore it should be advantageous to use all available marker data, rather than just a few markers linked to the quantitative trait loci (QTL) of the trait. The general idea of GP is to use a training population with both genomic marker data and phenotypic data to estimate marker or genotype effects. With these estimated effects, it is possible to make predictions on the genotypes or breeding values of individuals that are not phenotyped. When these predictions are based on additive effects, they are called genomic estimated breeding values (GEBV) and their accuracy is usually evaluated using cross validation (Crossa et al., 2017; Goddard & Hayes, 2007; Meuwissen et al., 2001). When compared to phenotypic and pedigree-based selection, GP has the potential to increase genetic gain while reducing cost and breeding time (Crain et al., 2020; de Bem Oliveira et al., 2019; Hickey et al., 2017; Matei et al., 2018; Voss-Fels et al., 2018). For example, de Bem Oliveira et al. (2019) managed to reduce the breeding time to release a new blueberry cultivar by three years. In that study GP increased the genetic gain per year by 86% when compared to phenotypic selection and by 32% when compared to pedigree-based selection. A second example is the study of Crain et al. (2020), in which the time of one breeding cycle for intermediate wheatgrass was reduced from 2-7 years to one year through the application of GP.

1.2 Complexities of whole genome regression

Utilising all available marker information has its advantages, but it also means that there are much more markers (p) than phenotypic data points (n) in the prediction model. This is known as the large-p, small-n ($p >> n$) problem and it rules out the use of a basic linear regression model (Crossa et al., 2017; de los Campos et al., 2013; Goddard & Hayes, 2007). In basic linear regression, parameters are estimated using the ordinary least squares (OLS) method which is not suited to deal with large number of variables (p) as it tends to overfit the model. When p exceeds n, using OLS is impossible as there is no longer a unique least squares estimate for every variable and the variability of the model fit will be infinite (Goddard & Hayes, 2007; James et al., 2021).

To deal with the large $p >> n$ problem, many penalized, Bayesian and kernel-based regression models have been developed that apply some sort of penalty to the estimated marker effects. The most common forms of penalization are shrinkage and a combination of variable selection and shrinkage. With shrinkage, it is assumed that all markers are important, therefore the weight of the penalty applied to each estimated marker effect is equal and all markers are included in the model (James et al., 2021; Meher et al., 2022). Applying shrinkage improves upon OLS as it greatly reduces

the variability of the least squares fit at the cost of some minor bias (James et al., 2021). With variable selection, it is assumed that some markers are more important than others. Therefore, the shrinkage applied to each estimated marker effect is variable. This allows important markers to capture more of the genetic variance while unimportant markers are shrunk more severely or left out of the model completely. Variable selection improves upon OLS as it effectively reduces the amount of variables in the model (James et al., 2021). In the next paragraphs it is discussed how different penalized, Bayesian and kernel-based regression models are defined and how they can relate to each other.

1.3 Penalized regression

To estimate marker effects, penalized regression models solve an optimization problem that finds the optimal balance between the OLS fit and the penalty term $\lambda j(\beta)$ (de los Campos et al., 2013). This penalty term consists of a tuning parameters λ which controls the severity of the penalty and a penalty function $j(\beta)$ which defines a particular model. For example, the $\ell 2$ norm $j(\beta) = \sum_{j=1}^{p} \beta_j^2$ defines Ridge Regression (RR) which is a model that applies shrinkage whereas the $\ell 1$ norm $j(\beta) = \sum_{j=1}^{p} \|\beta_j\|$ defines the Least Absolute Shrinkage and Selection Operator (LASSO) which applies a combination of variable selection and shrinkage (de los Campos et al., 2013; Hoerl & Kennard, 1970; Tibshirani, 1996).

1.4 Bayesian regression

Bayesian regression methods estimate marker effects by sampling from a posterior distribution $P(\theta|y)$, which is derived from the conditional distribution of the data $P(y|\theta)$ and the joint prior distribution $P(\theta)$ of model parameters. Where θ represents the unknown model parameters $\{\mu, \beta, \sigma_\varepsilon^2, \omega\}$ (de los Campos et al., 2013; Pérez & de los Campos, 2014). Bayesian regression models are defined by their prior distribution for marker effects $P(\beta)$ as this will determine the type and the severity of the penalty. For example, a gaussian prior for marker effects $\beta \sim N(0, \sigma_\beta)$, will impose a level of shrinkage that is equivalent to that of the penalized RR model when $\lambda = {\sigma^2}_\varepsilon / \sigma_\beta^2$. On the other hand, a Laplace prior for marker effects $Laplace(\lambda)$ applies a combination of variable selection and shrinkage that is equivalent to that of the penalized LASSO model. These Bayesian regression models are called Bayesian Ridge Regression and Bayesian LASSO respectively and there are more Bayesian regression models with a penalized regression counterpart (de los Campos et al., 2013).

1.5 Kernel-based regression

Kernel-based regression is a particular case of regression which allows to capture complex genetic effects like dominance and epistasis by choosing an appropriate kernel function. This is convenient as it removes the need to model them explicitly, which is complex (de los Campos et al., 2010; Endelman, 2011; Gota & Gianola, 2014). In kernel-based regression, models are defined by their choice of kernel K which can take many different forms (Endelman, 2011; Gota & Gianola, 2014; Pérez & de los Campos, 2014). An example of a very popular kernel-based GP model is the genomic best linear unbiased predictor (GBLUP). The kernel that is used in GBLUP is a genomic relationship matrix which captures the covariance between individuals based on the number of marker alleles they share. In GBLUP shrinkage is applied by correcting the genotypic variance σ_g for the covariance between individuals. Under the assumption that σ_g is the same as the marker variance σ_β^2, the shrinkage applied by GBLUP is equivalent to the shrinkage applied by RR (Clark & van der Werf, 2013; Endelman, 2011).

1.6 Variables affecting prediction accuracy

Choosing the correct GP model is not always straightforward as the accuracy of the prediction is affected by a multitude of variables. It has been shown that the predictive accuracy can be increased when the model is chosen according to the genetic architecture of a trait. In general, variable selection models are better for predicting traits with mono- or oligogenic effects, while shrinkage models are better for predicting traits with multigenic effects (Daetwyler et al., 2010; de los Campos et al., 2013; Haile et al., 2020; Meher et al., 2022; Riedelsheimer et al., 2012).

A second variable to take into account is the heritability of a trait. A high heritability will generally increase the predictive ability of a model as there is more genetic variance to capture (Crossa et al., 2017; Meher et al., 2022). However, Meher et al. (2022) found that this increase is not equal for all models, as variable selection models tend to benefit more from an increase in heritability than shrinkage models.

A similar effect has been observed for linkage disequilibrium (LD). If LD between the markers and the QTL of a trait is sufficient, GP models are able to capture a large part of the genetic variance, which generally results in better predictions (Voss-Fels et al., 2018). When LD between markers and QTL is high, the distribution of genetic effects is centred around those markers. Models that apply variable selection, benefit from this as it simplifies the identification of marker with large genetic effects. Naturally, the effect of low LD is similar to that of low marker density as the chance of finding markers that are in strong LD with the QTL decreases with a reduction in marker density (de los Campos et al., 2013).

The final variable that will be covered in this study is population structure. If population structure is not accounted for, it is known to cause serious inflation of the estimated predictive ability of GP models (Daetwyler et al., 2012; Habier et al., 2007). Daetwyler et al. (2012) showed that, if there is strong population structure, predictions using only one chromosome could be accurate up to 86% when compared to predictions using the whole genome. As it is unlikely that one chromosome is responsible for 86% of the genetic variance of a quantitative trait, it indicates that a significant portion of the predictive ability is attributable to population structure.

In general, the variables described above are unknown when using actual data and it is difficult to estimate their exact effects. Simulation studies are a practical and effective way to overcome this problem as they allow the user to create and control the variables on desire, making it is easier to estimate their effects and to validate results. For this reason simulations are often used in research towards the development of GP methods(Daetwyler et al., 2010; de los Campos et al., 2013; Meher et al., 2022; Meuwissen et al., 2001; Scutari et al., 2016)

Naturally, there are many more variables affecting the predictive ability of GP models, like: the size and composition of the training and test populations (Crossa et al., 2017; Sallam et al., 2020; Sarinelli et al., 2019), the type of markers that are used (Cuyabano et al., 2014, 2015; Sallam et al., 2020) and the implementation of prior knowledge on the genome (Rice & Lipka, 2019; Sarinelli et al., 2019). However, due to time restraints, they are out of the scope of this study.

1.7 GP in polyploid crops

Since the paper by Meuwissen et al. (2001), GP has been the focus of many plant breeding studies and many breeding companies have been implementing it into their breeding programs (Hickey et al., 2017). However, most of these efforts focused on diploid crops and the development and applications of GP in polyploid crops has been lagging behind (Wilson et al., 2021). The reason for this backlog is the complex genetic structure of polyploids. Unlike diploids, polyploids have more than two homologous copies of each chromosome, resulting in a large degree of heterozygosity, a greater number of possible allelic dosages, a wider range of genetic classes and the possibility of multivalent pairing (Dufresne et al., 2014; Lloyd & Bomblies, 2016; Ramsey & Schemske, 2002). The combination of these factors has complicated the implementation of GP and slowed down its development for polyploid species (de Bem Oliveira et al., 2019; Endelman et al., 2018). In recent years, however, the interest for GP in polyploid crops has grown, resulting in promising results in crops like alfalfa (Jia et al., 2018), blueberry (de Bem Oliveira et al., 2019), intermediate wheatgrass (Crain et al., 2020) and potato (Endelman, 2011; Wilson et al., 2021), among others.

Over the past few decades, a great effort has been made to implement and understand GP in breeding programmes. However, most of the developed methods are based on diploid crops and it is unclear whether they can be directly implemented in polyploid crops. The aim of this project was to investigate the performance of different GP models using polyploid data and to see how their predictive ability was affected by: trait architecture, heritability, marker density and population structure.

2 Materials and methods

2.1 Simulation of genomic data

We simulated genomic data using the PedigreeSim software (Voorrips & Maliepaard, 2012). The simulated genome represents an autotetraploid organism with four chromosomes ($2n = 4x = 16$). Each chromosome was 100 centimorgan (cM) long and the centromere was situated 20 cM from the start of each chromosome. In this genome no preferential pairing occurs, the probability of quadrivalent formation is 10% and it is assumed that there is no chiasma interference.

The simulation aims to mimic the LD structure of potato varieties released in the period of 1945 to 1974 as described by Vos et al. (2017). Following their example the LD was calculated as the Pearson correlation (r^2) between markers and the following parameters were used to describe LD structure:

1. Short-range LD: the calculated LD between marker pairs at a certain interval of genetic distance
2. Background LD: the calculated LD between markers at different chromosomes
3. $LD_{1/2}$: the genetic distance in cM at which half of the short-range LD has decayed

The short-range LD was calculated between all marker pairs within 1 cM and the background LD was estimated using 300 randomly chosen marker pairs for each chromosome. These parameters were calculated for each simulated generation and compared to the values reported by Vos et al. (2017) for the potato varieties from 1945 to 1974. To make these comparisons we used the average short-range LD, the 95% percentile of all pairwise correlations for the background LD and the 90% percentile for $LD_{1/2}$. Additionally, Vos et al. (2017) used the minimum LD threshold of 0.1, representing the point at which there is no effective correlation between markers. We used $LD_{0.1,90}$ as a parameter that indicates the genetic distance at which the LD drops below 0.1 for 90% of all marker pairs.

Table 1 Representing values of the 95% percentile of the background LD ($LD_{back,95}$), the average of the short-range LD (LD_{short}), $LD_{1/2,90}$ and $LD_{0.1,90}$ reported by Vos et al. (2017) for the potato varieties that originate from 1945 to 1974 and the values that were measured in the simulation

	Potato varieties 1945-1974 (vos et al. 2017)	**simulation**	**Unit**
$\boldsymbol{LD_{short}}$	0.22	0.233	r^2
$\boldsymbol{LD_{back,95}}$	0.14	0.113	r^2
$\boldsymbol{LD_{1/2,90}}$	3.0	4.071	cM
$\boldsymbol{LD_{0.1,90}}$	NA	16.99	cM

The desired LD structure was created by generating a founder population that was used as a basis for the simulation of all subsequent generations. To create the founder generation, for each chromosome six founder haplotypes were randomly generated with 25% haplotype specific SNPs. These six founder haplotypes were assigned to the chromosomes of ten founder individuals with a frequency of 0.30, 0.25, 0.175, 0.125, 0.10, and 0.05 respectively. The first generation was obtained by random mating between the ten founder individuals, including self-fertilization.

Population structure was created by simulating three subpopulations: X, Y and Z (figure 1). These subpopulations diverged for three similarly called traits (paragraph 2.3). In each generation, the top 50% of individuals within each subpopulation were selected as the parents for the next generation based on their respective traits. For example, the individuals in subpopulation X were selected for trait X. The proportion of subpopulations in each generation was 0.5, 0.3 and 0.2 for X, Y and Z, respectively.

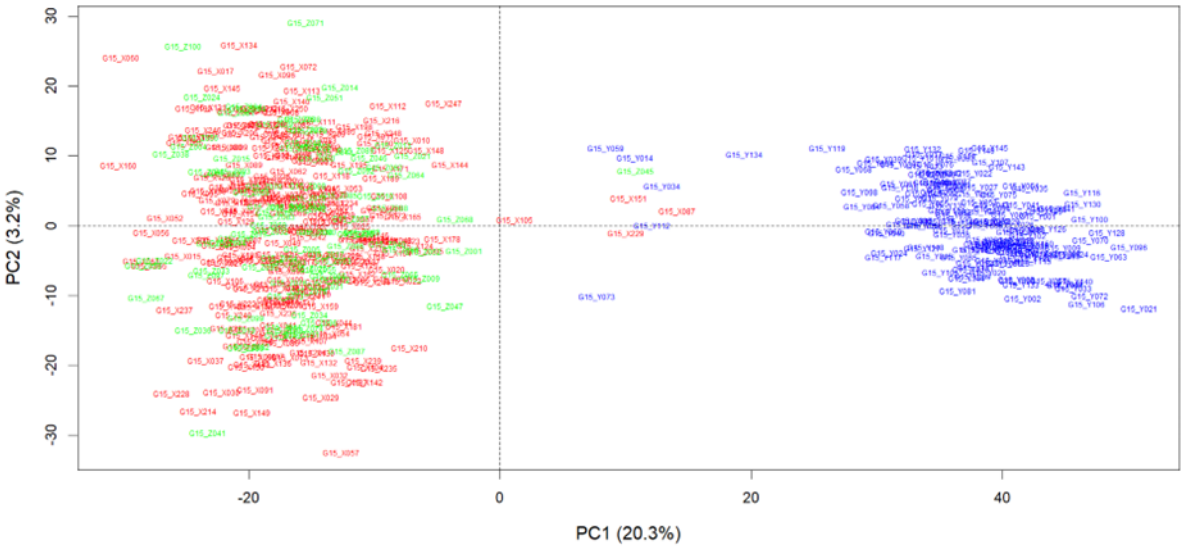

Figure 1 PCA plot highlighting population structure; red = X, bue = Y, green = Z

In total, 15 generations were simulated. In the first 14 generations, 200 offspring were obtained per generation, while in the 15th generation, 500 offspring were obtained. The parents in all 15 generations were crossed through random mating. There was no self-fertilization and crosses between subpopulations were limited to 5%. A SNP marker set of 4004 SNP markers was superimposed on the founder alleles for each generation.

2.2 Simulation of phenotypic data

In total, 15 traits were simulated with different heritability values and trait architectures. All of the simulated traits follow an additive genetic model and the calculation of phenotypic values is described by the following equation:

$$phenotype = genotype + error \quad (1)$$

To calculate the genotype effect of an individual, the allele dosages at the QTL were multiplied by the effects assigned to the alleles, while the error was randomly sampled from the normal distribution $N(0, \sigma_\varepsilon^2)$. The error variance σ_ε^2 was estimated using the following equation:

$$\sigma_\varepsilon^2 = (1 - H^2) * \sigma_g^2 / H^2 \quad (2)$$

Where the genetic variance σ_g^2 was calculated from the genotypes and H^2 is a priorly defined value for the heritability. After the phenotypes were calculated the phenotypic variance (σ_p^2) was calculated and used to determine the realised heritability $H^2 = \sigma_g^2 / \sigma_p^2$. Different heritability values were simulated by specifying a different value for the prior heritability. It is important to note that the realised heritability deviates slightly from the prior heritability due to the variation imposed by the random error effect. To simulate different types of trait architecture, one or more biallelic markers were selected across the genome to act as QTL and specifying their additive effect.

2.3 Traits X, Y and Z

Traits X, Y and Z were the first traits to be simulated, as they were used for selection within subpopulations during the simulation of genomic data. These traits were all affected by eight QTL that were randomly chosen across the genome and the effect for each allele was set to one. The prior heritability of each trait was set to 0.6 for the first generation but the realised heritability increased over generations. This increase in heritability was caused by an overall increase of genetic variation across the population, while the error variation remains the same (table 2).

Table 2 List of quantitative trait loci (QTL), the effects of each allele, the variation due to error that was applied to the trait and the realised heritability in generation 15 for traits X, Y and Z.

Trait	Quantitative trait loci	Effect of each alternative allele	Error variation	Realised heritability of generation 15
X	B0086, B0861, C0216,C0585, C0899, C0992, D0109, D0268	1	5.199	0.688
Y	A0493, A0565, B0105, C0475, C0927, D0049, D0336, D0872	1	4.052	0.877
Z	A0069, A0562, A0575, B0998, D0478, D0529, D0560, D0847	1	2.148	0.869

2.4 Assessing predictive ability

To assess the predictive ability, we used a five fold cross-validation scheme. This cross-validation was set up by randomly splitting the population into five parts. For each cycle of the cross-validation, four parts (training population) were used to train the GP model and estimate marker/genotype effects. These marker/genotype effects were then used to predict the phenotypes of the remaining part of the population (test population). The predictive ability of the training model was determined by calculating a Pearson correlation between the predicted genotypes and the actual phenotypes of the test population. After completion of the cross-validation the predictive ability was averaged over all five folds. This process was repeated ten times, resulting in ten observations per trait for each model.

2.5 GP models

The four GP models compared in this study are BayesB, Elastic Net, Reproducing Kernel Hilbert Space (RKHS) and Ridge Regression Best Linear Unbiased Predictor (RR-BLUP). These models were chosen because they represent a wide range of statistical models and there were statistical packages available that enabled their application in polyploid data.

2.5.1 Bayesian B

BayesB was implemented using the R package BGLR, version 1.1.0 (Pérez & de Los Campos, 2014). Equation 3 represents the model implemented by BGLR for our data:

$$y = \mu + X\beta + \varepsilon \quad (3)$$

$$\sigma_\varepsilon^2 \sim \chi^{-2}(df_\varepsilon, S_\varepsilon)$$

$$P(\theta|y,\omega) \propto P(y|\theta)P(\theta)$$

Where y is a vector of responses, μ is the overall mean, β is a vector of random marker effects, X is an allelic dosage matrix and ε is a vector of residuals. As mentioned in the introduction, different Bayesian models are defined by their prior density for marker effects $P(\beta)$. For Bayes B this is a combination of a point of mass at zero and a scaled-t slab (equation 4; figure 2).

$$P(\beta_j, \sigma^2_\beta, \pi) = \prod_k \beta_{jk} \sim N(0, \sigma^2_\beta) + (1-\pi)1(\beta_{jk} = 0) \quad \{4\}$$

$$\sigma^2_\beta \sim \chi^{-2}(df_\beta, S_\beta)$$

Where $P(\beta_j, \sigma^2_\beta, \pi)$ is a function of the scaled-t slab and the point of mass at zero $(1-\pi)1(\beta_{jk} = 0)$. This function allows for a combination of shrinkage and variable selection. The effects of important markers are sampled from the scaled-t slab which has a large variance that comes from a scaled-inverse χ^2 distribution $\sigma^2_\beta \sim \chi^{-2}(df_\beta, S_\beta)$ while the effects of unimportant markers are equated to zero by sampling from the point of mass at zero.

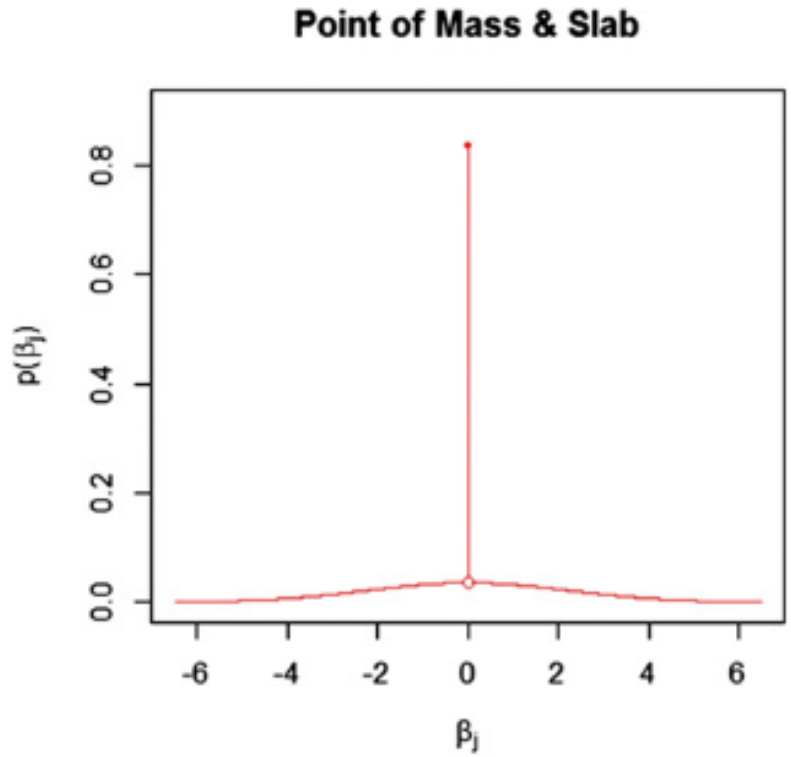

Figure 2 Depiction of a density with a point of mass at zero and a scaled-t slab (de los Campos et al., 2013)

In Bayes B three hyperparameters are used to control the severity of the penalty: df_β is the degrees of freedom for marker effects, S_β is a scaling parameter that is sampled from the gamma distribution $S_\beta \sim G(r, s)$ and π is the proportion of non-zero markers allowed in the model which is sampled from the beta distribution $\pi \sim B(p_0, \pi_0)$ and optimized using an internal validation method. The prior density for the remaining unknown model parameters are a flat prior for the overall mean and a scaled-inverse χ^2 density for the residual variance $\sigma^2_\varepsilon \sim \chi^{-2}(df^2_\varepsilon, S_\varepsilon)$.

The R package BGLR utilizes a Gibbs sampler to optimize the posterior density. In total we used to 12500 iterations of this Gibbs sampler, of which 2500 were burn in iterations.

2.5.2 Elastic Net

Elastic Net was implemented using the R package glmnet, version 4.1-3 (Friedman et al., 2010). It is a penalized regression model that combines LASSO and RR by using a weighted average of the $\ell 1$ and $\ell 2$ norms as a penalty function. The regression model of Elastic Net is the same as equation 3 with a mean and estimated marker effects, but the parameters are estimated by solving the optimization problem that is described in the following equation:

$$\left(\hat{\mu}, \hat{\beta}\right)_{\underset{argmin}{=}} \left\{\sum(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda j(\beta)\right\} \quad (5)$$

$$j(\beta) = (1-\alpha)\frac{1}{2}\left|\left|\beta\right|\right|_{\ell 2}^2 + \alpha||\beta_j||_{\ell 1} = \sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|]$$

Where $\sum(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j)^2$ is the sum of squared residuals and $j(\beta) = (1-\alpha)\frac{1}{2}\left|\left|\beta\right|\right|_{\ell 2}^2 + \alpha||\beta_j||_{\ell 1}$ is the penalty function which is a weighted average of the $\ell 2$ and $\ell 1$ norms. In Elastic Net there are two parameters controlling the penalty function, these are lambda (λ) and alpha (α).

Lambda controls the severity of the penalty and is internally optimized when using the cv.glmnet function. This function fits a grid of 100 values of λ and applies a 10 fold cross-validation to find the optimum value. Cv.glmnet returns two potentially optimal values of λ for the user to choose from: lambda.min and lambda.1se. During our testing phase we found that the Elastic Net penalty was too stringent for certain combinations of training and test sets when predicting for polygenic traits with low heritability. This resulted in model fits that only estimated an intercept, or the estimated marker effects were too small for R to handle. When this happened we were unable to calculate a Pearson correlation and the predictive ability was displayed as NA (from this point we will refer to this problem as the NA problem). To minimize the NA problem it was decided to use lambda.min as it provides a less stringent model.

α controls the mix between RR ($\alpha = 0$) and LASSO ($\alpha = 1$). As there is no internal optimization function for α in glmnet, we set up a double cross-validation scheme (figure 3). This double cross-validation scheme consisted of two cross-validation cycles: an outer cross-validation cycle as described in the paragraph 2.4 and an inner cross-validation cycle. The aim of this inner cycle was to find the optimum value for α for the training population of the outer cycle. To achieve this, the inner cycle carried out five fold cross-validations, on the training population of the outer cycle, for each value in a specified grid of α. We used an evenly spaced grid of eleven values for α ranging from 0 to 1 (0, 0.1, 0.2, …, 1). The predictive ability of each fold was calculated as described in paragraph 2.4 and an average was calculated over all folds of one cross-validation. At the end of the inner cycle the value for α was chosen that provided the highest average predictive ability. However, because this inner cycle only uses the training population of the outer cycle there is a decrease in statistical power, which in turn caused an inflation of the NA problem. This was problematic as it was not possible to calculate an average predictive ability when one of the folds returned an NA. When there was one fold that returns NA in the cross-validation for each value of α, it was not possible to calculate any average predictive ability and no optimal α can be selected. Therefore, it was decided to leave out the folds that returned an NA when calculating the average predictive ability of the inner cross-validations.

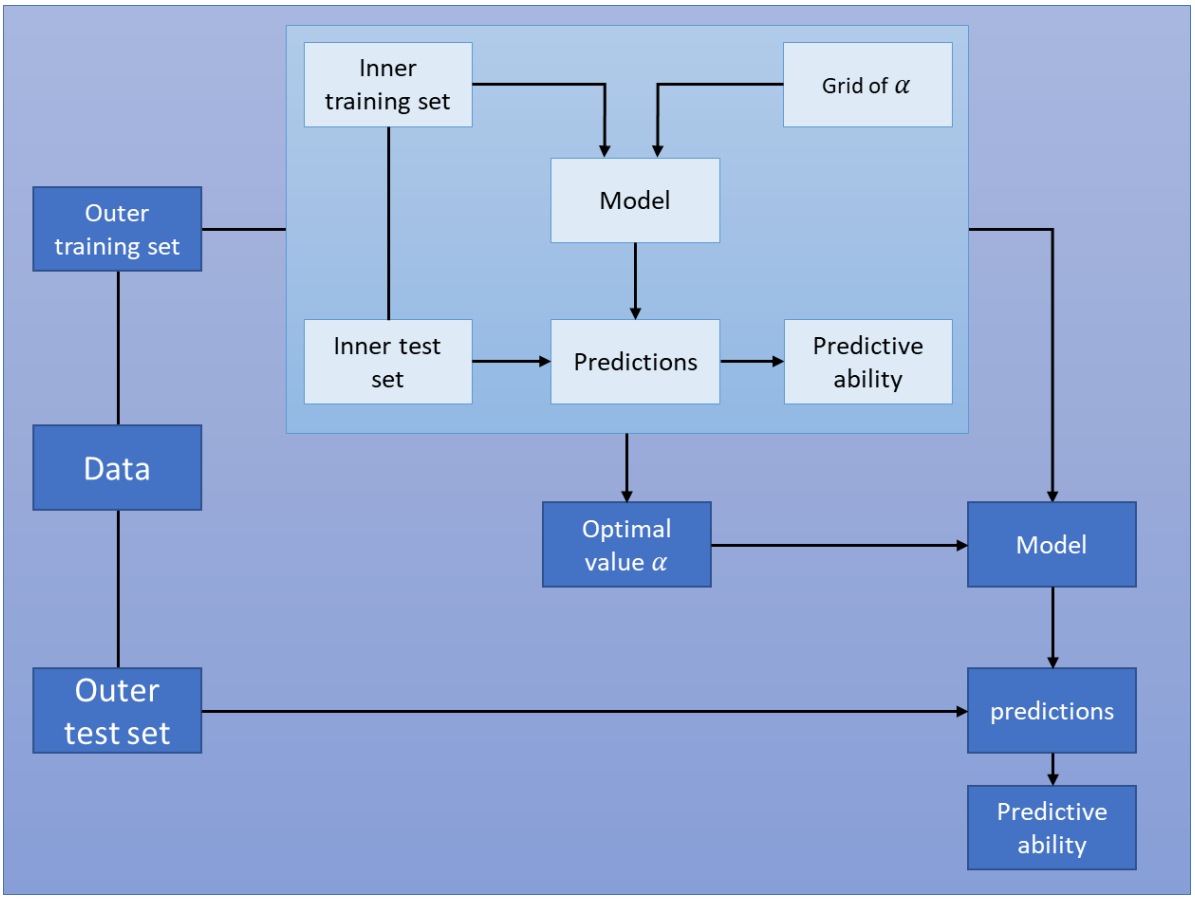

Figure 3 Schematic drawing of the double cross-validation scheme. The data in the outer cycle (dark colours) is split into a training set and a test set. The training set is than used for the inner cycle (light colours) five fold cross-validations for a grid of α's. From these cross-validations in the inner cycle the average predictive ability is calculated and the α which produced the highest average predictive ability is selected to train a model in the outer cycle cross-validation. This figure was inspired by figure 1 in the paper by Hendriks et al. (2007).

2.5.3 Reproducing Kernel Hilbert Space

RKHS was implemented using the kin.blup function in the R package rrBLUP, version 4.6.2 (Endelman, 2011). RKHS is a kernel-based regression model that can be described using the following equation:

$$y = Wg + \varepsilon \quad (6)$$

$$g \sim N\left(0, k\sigma_g^2\right)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Where y is a vector of phenotypes, g is a vector of genotypic values that follows a normal distribution $g \sim N\left(0, K\sigma_g^2\right)$, W is a matrix that links genotypes to individuals and ε is a vector of residuals that follows the normal distribution $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. In RKHS genetic effects such as additivity, dominance and epistasis captured by a Gaussian kernel K:

$$K_{ij} = \exp\left[-(D_{ij}/\theta)^2\right] \quad (7)$$

$$D_{ij} = \left[(1/4M)\sum_{k=1}^{M}(G_{ik} - G_{jk})^2\right]^{1/2}$$

Where D is a Euclidean distance matrix which describes the covariance between genotypes based on the marker data and θ is a scaling parameter that controls the rate at which the covariance between

individuals decays with distance. θ is optimized internally by rrBLUP by fitting a grid of values for theta between 0 and 1. For each value of θ, the log-likelihood is calculated using REML, as a measure for goodness of fit of the model. The value of θ that provides the highest log-likelihood is considered optimal. The marker effects are estimated solving the BLUP equation $\hat{g} = X'(XX' + \lambda K)^{-1}y$. In RKHS the shrinkage parameter λ controls the severity of the shrinkage that is applied. In RKHS this is the ratio of the genotypic variance and the environmental variance $\lambda = \sigma_\varepsilon^2/\sigma_g^2$.

2.5.4 Ridge Regression BLUP

RR-BLUP was implemented using the mixed.solve function of the R package rrBLUP, version 4.6.2. (Endelman, 2011). RR-BLUP is equivalent to the RR model that is discussed in the introduction but the effects are estimated in a different way. The prediction model as implemented for our data is described by the following equation.

$$y = X\beta + \varepsilon \quad (8)$$

$$\beta \sim N\left(0, I\sigma_\beta^2\right)$$

$$\varepsilon \sim N(0, \sigma_\varepsilon^2)$$

Where y is a vector of responses, β is a vector of random marker effects that follows a normal distribution $\beta \sim N\left(0, I\sigma_\beta^2\right)$, X is an allelic dosage matrix with markers in the columns and individuals in the rows and ε is vector of residuals that follows a normal distribution $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. The marker effects are estimated by solving the BLUP equation $\hat{\beta} = X'(XX' + \lambda I)^{-1}y$. Like RKHS the shrinkage parameter λ controls the severity of the shrinkage that is applied. In RR-BLUP λ is the ratio between the residual variance and the marker variance $\lambda = \sigma_\varepsilon^2/\sigma_\beta^2$. The marker effects returned by the mixed.solve function were multiplied by the allelic dosage matrix of the individuals in the test population to get to the actual predictions.

2.6 Experimental setup

Three experiments were conducted to investigate the effects of trait architecture, heritability, marker density and population structure on the predictive ability of the four GP models. For the first experiment, we simulated twelve traits with different trait architectures and heritability values, after which the predictive ability of the models was compared for each trait. In this first experiment we used a high-density SNP marker (HDSM) set with 3622 SNP markers and an approximate marker density of one SNP marker per 0.11 cM. To investigate the effect of marker density on the predictive ability of the models, the first experiment was repeated using a low-density SNP marker (LDSM) set with 362 markers and an approximate density of one SNP marker per 1,1 cM. The final experiment was aimed at investigating the contribution of the population structure to the predictive ability of the models. To achieve this goal, SNP marker sets were created in which most of the LD between the QTL and other markers was removed. As most LD was removed, it was assumed that the remaining predictive ability was mostly attributable to population structure. To verify this assumption, the experiment was carried out on one trait that was strongly correlated to population structure and one trait that was not. These traits were trait X and the oligogenic trait with a heritability of 0.53, respectively. This experiment was also repeated using the HDSM and LDSM sets. For every experiment, the SNP marker data was first normalized to [-1, -0.5, 0, 0.5, 1], then centred and scaled. The phenotypic data was also centred and scaled.

2.6.1 Trait architecture and heritability

The four levels of trait architecture were: monogenic, oligogenic, polygenic and a combination of oligo- and polygenic (mix). All monogenic traits had one QTL that was randomly selected across the genome, the oligogenic traits had one QTL randomly selected on each chromosome and the polygenic traits had 25 QTL randomly selected on each chromosome. The additive effects for each alternative allele at the QTL were set to one. A particular case was the mix trait, which had one QTL on each chromosome that represented an oligogenic effect and 25 QTL on each chromosome that represented a polygenic effect. The effects assigned to alleles were scaled so that the oligo- and polygenic effect explained an equal part of the genotypic value. Within each level of trait architecture all traits had the same QTL (e.g. the QTL for all monogenic traits was A0062) and the mix trait had the same QTL as the oligo- and polygenic traits. The three prior levels of heritability were 0.2, 0.5 and 0.8, and the realised heritability was calculated as explained in paragraph 2.2 (table 3).

Table 3 List of quantitative trait loci (QTL), additive effects assigned to the alleles at the QTL, the variation that is applied due to the error and the realised heritability for all traits.

Trait architecture	Quantitative trait loci	Effect of each alternative allele	Applied error variation	Realised heritability
Monogenic	A0062	1	3.333	0.197
			0.833	0.480
			0.208	0.800
Oligogenic	A0508, B0381, C0173, D0554	1	13.137	0.226
			3.284	0.530
			0.821	0.777
Polygenic	See appendix …	1	316.141	0.203
			79.035	0.501
			19.759	0.798
Combination	Oligogenic: A0508, B0381, C0173,D0554 Polygenic: See appendix	Oligogenic: 4.91 Polygenic: 1	538.156	0.202
			134.539	0.474
			33.634	0.745

2.6.2 Marker sets

The high density SNP marker set was created by excluding the QTL of each trait, and all the SNP markers with a minor allele frequency at or below 0.05, from the 4004 markers generated by PedigreeSim. This resulted in a SNP marker set with 3622 markers. By selecting every tenth marker of the high density marker set, the low density SNP marker set was created for the second experiment.

Removing the LD between QTL and markers was achieved by excluding all markers within 17cM $(LD_{0.1,90})$ of the QTL. This was repeated for the QTL of trait X and the QTL of the oligogenic traits and for the high and low density SNP marker sets. In total there were four SNP marker sets from which the LD was removed.

3 Results

3.1 Effect of trait architecture and heritability

During the first experiment, the predictive ability of the four GP models was compared for 13 traits with different levels of trait architecture and heritability, using a high density SNP marker (HDSM) set (approx. 1 SNP/0.11 cM). The predictive ability ranged from 0.177 to 0.875, and the average predictive ability of the models was 0.578 for BayesB, 0.576 for Elastic Net, 0.526 for RKHS and 0.532 for RR-BLUP.

For the oligogenic traits with low heritability, the polygenic traits and the mix trait and trait X, the predictive ability of the models was on par. BayesB and Elastic Net outperformed RKHS and RR-BLUP for the monogenic traits. Between BayesB and Elastic Net, Elastic Net was the best option for the monogenic trait with low heritability, but the difference was negligible for the other heritability values. When predicting the oligogenic traits with intermediate and high heritability, BayesB and Elastic Net were again performing better than RKHS and RR-BLUP (figure 4).

In general, it can be stated that an increase in heritability improves the accuracy of the predictions. For most traits an increase in heritability had the same approximate effect for all models. Only for the oligogenic trait architecture an increase in heritability was more beneficial for BayesB and Elastic Net than it was for RKHS and RR-BLUP. In some cases the predictive ability of the models was considerably higher than the heritability of the trait. With an average predictive ability 0.799, over all models, and an heritability of 0.688, trait X was the most extreme case. However, this also occurred in the monogenic traits when predicting with BayesB and Elastic Net, and in the mix trait with a heritability of 0.474, for all models.

RKHS and RR-BLUP were mostly unaffected by trait architecture. For the mono-, oligo- and polygenic traits both models show a clear pattern where the predictive ability was stable and close to the heritability of the trait. In general the predictive ability was slightly higher than the heritability for the traits with low heritability, and the reverse was true for traits with high heritability. This pattern was only broken for the traits with a combination of oligo- and polygenic effects (mix traits) and for trait X. In contrast, BayesB and Elastic Net, were clearly affected by trait architecture as they outperformed RKHS and RR-BLUP considerably for the mono- and oligogenic traits, but this difference decreased with the number of QTL.

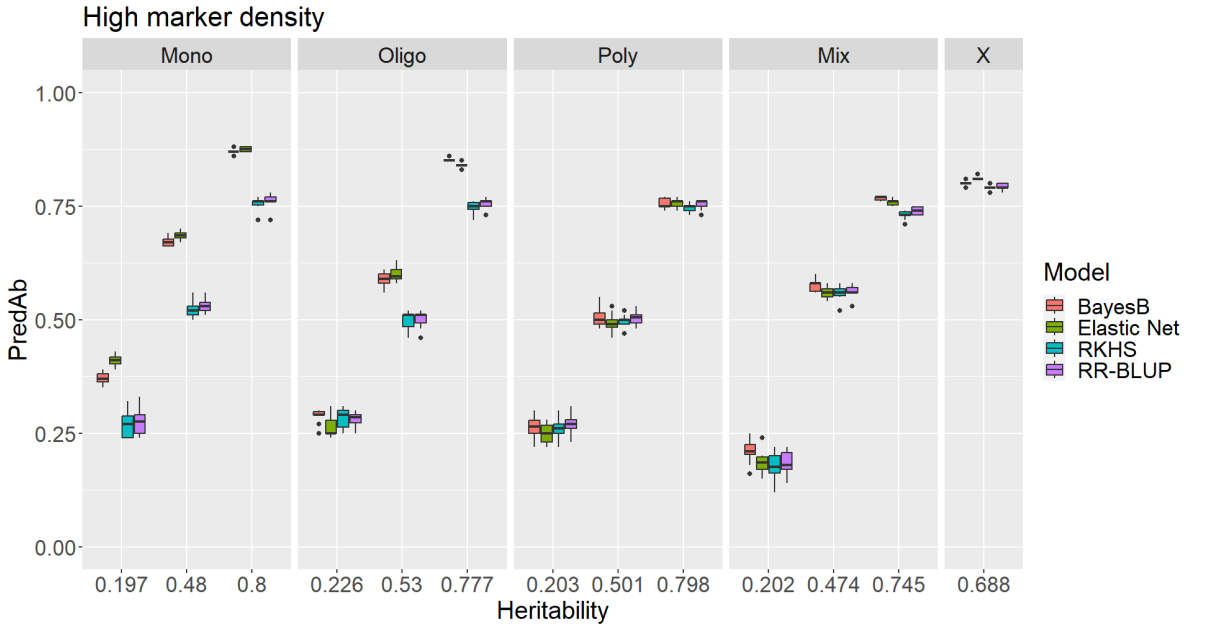

Figure 4 Boxplots representing the results of the genomic prediction analyses when using the high density SNP marker set (ten observations per boxplot), with predicitive ability on the y-axis and the realised heritibality of each trait on the x-axis. Each pannel represents a different trait architecture: mono = monogenic, oligo = oligogenic, poly = polygenic, Mix = an architecture with a combination of oligo- and polygenic effects, and X = an oligogenic architecture, correlated to population structure, that is specific to trait X. The four genomic prediction models represented in this figure are: Bayesian B (BayesB), Elastic Net, Reproducing Kernel Hilbert Space (RKHS) and the Ridge Regression Best Linear Unbiased Predictor (RR-BLUP).

3.2 Marker density

To observe the effect of marker density on the predictive ability of the models, the first experiment was repeated using a low density SNP marker (LDSM) set (approx. 1 SNP/1 cM). When comparig the results of the HDSM set (figure 4) with the results of the LDSM set (figure 5) there was an overall drop in predictive ability. The predictive ability of the LDSM set ranged from 0.155 to 0.796, and the average predictive ability of the models was 0.514 for BayesB, 0.504 for Elastic Net, 0.481 for RKHS and 0.480 for RR-BLUP.

For the oligogenic trait with low heritability, all the polygenic traits, all the mix traits and trait X, the models had approximately the same predictive ability. Only for the mono- and oligogenic traits some clear differences remained. BayesB and Elastic Net still outperformed RKHS and RR-BLUP for the monogenic traits and for the oligogenic traits with intermediate and high heritability. For the monogenic traits with low and high heritability the predictive ability of BayesB and Elastic Net was equal, but for the other three traits BayesB had slightly higher predictive ability than Elastic Net (figure 5).

BayesB and Elastic Net were affected more by the reduction in marker density than RKHS and RR-BLUP. For the mono- and oligogenic traits, BayesB and Elastic Net were still outperforming RKHS and RR-BLUP, but the margins became considerably smaller when compared to the results of the HDSM set. For the other trait architectures the effect of marker density was approximately equal for all models.

The drop in predictive ability of the models was more severe when predicting for traits with high heritability than for traits with low heritability. Averaged over all models and trait architectures, the predictive ability dropped by 0.089 for highly heritable traits, by 0.055 for traits with intermediate heritability and by 0.044 for the low heritability traits.

With an average drop in predictive ability of 0.011, the prediction accuracy for trait X was least affected by the reduction in marker density. Similar to the HDSM set, the predictive ability for trait X was considerably higher than the heritability of the trait, for all models.

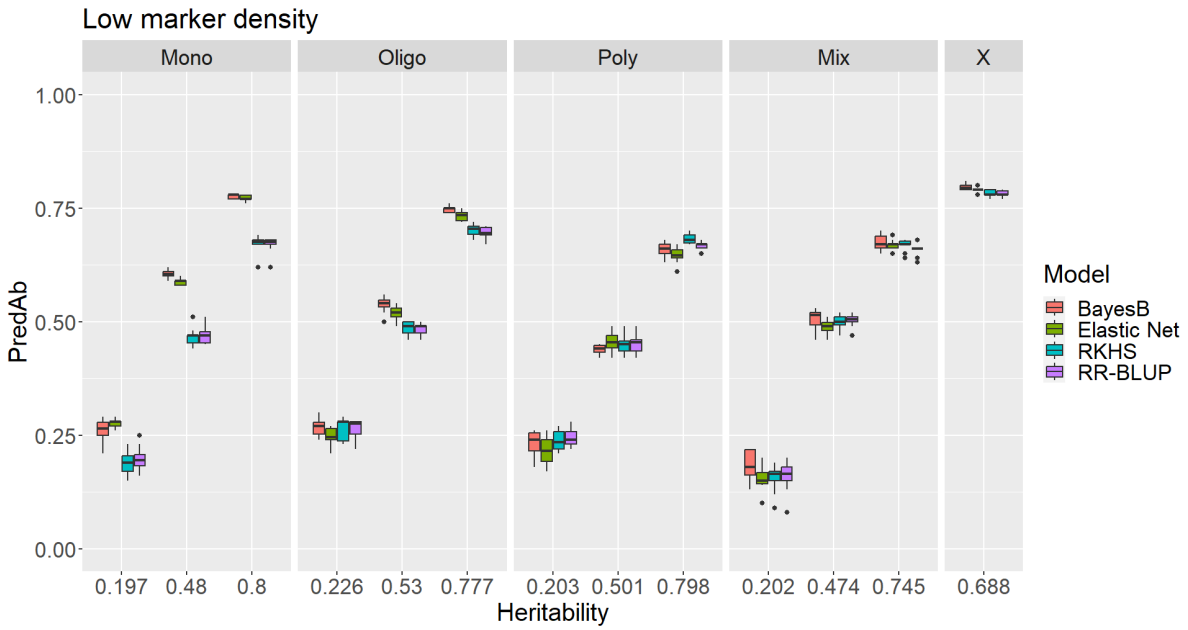

Figure 5 Boxplots representing the results of the genomic prediction analyses when using the low density SNP marker set (10 observations per boxplot), with predicitive ability on the y-axis and the realised heritibality of each trait on the x-axis. Each pannel represents a different trait architecture: mono = monogenic, oligo = oligogenic, poly = polygenic, Mix = an architecture with a combination of oligo- and polygenic effects, and X = an oligogenic architecture, correlated to population structure, that is specific to trait X. The four genomic prediction models represented in this figure are: Bayesian B (BayesB), Elastic Net, Reproducing Kernel Hilbert Space (RKHS) and the Ridge Regression Best Linear Unbiased Predictor (RR-BLUP).

3.3 Population structure

The aim of the final experiment was to explore the effect of population structure on the predictive ability of the models. For this purpose 'low LD' SNP marker sets were created, where all SNP markers in LD with the QTL of a trait were removed. The predictive ability of the models was compared between the low LD SNP marker sets and the original SNP marker sets (containing markers in LD to the QTL). This was repeated four times, once with the HDSM set and once with LDSM set, for trait X and for the oligogenic trait with an heritability of 0.474 (figure 6)

Trait X, which was simulated to correlate to population structure, showed a small drop in predictive ability when LD between marker and QTL was removed. For the HDSM set, the average predictive ability of all models dropped by 0.070, for the LDSM set the average predictive ability dropped by 0.075. For trait X the predictive ability of all four models was very similar, both for the original marker sets and for marker sets with low LD.

The oligogenic trait was simulated irrespective of population structure and the effect of removing LD was more dramatic. On average the predictive ability of the models dropped by 0.222 for the HDSM set and 0.182 for the LDSM set. Using the original HDSM and LDSM sets, BayesB and Elastic Net outperformed RKHS and RR-BLUP for the oligogenic trait. However, for the low LD SNP marker sets, the predictive ability of the models was approximately similar. In fact, the average predictive ability over all models was 0.325 for both the HDSM and LDSM sets from which most LD was removed.

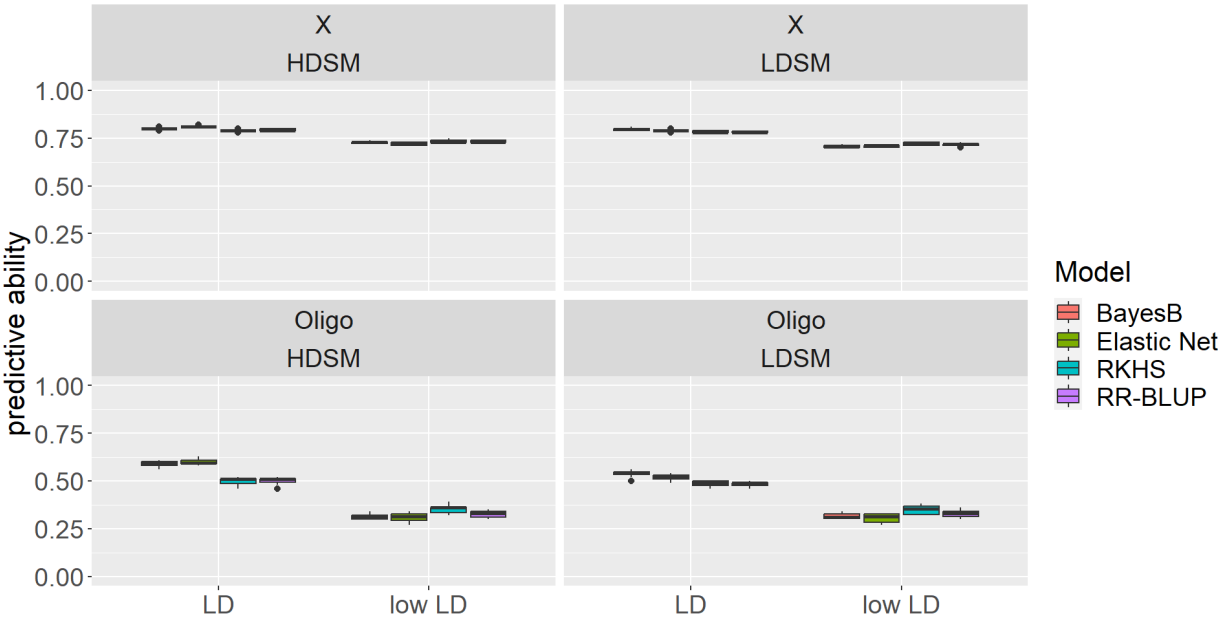

Figure 5 Boxplots representing the effect of population structure on the predictive ability of the models, with predictive ability on the y-axis and the level of LD in the marker set on the x-axis; LD = original SNP marker sets containing marker in LD with the Quantitative trait loci (QTL), noLD = SNP marker sets, derived from the original SNP marker sets, where most of the markers in LD to the QTL were removed. The two traits are: trait X, which was simulated to correlate to population structure (X), and the oligogenic trait, which was simulated irrespective of population structure (Oligo). The heritability of the trait was 0.688 and 0.474, for trait X and the Oligogenic trait, respectively. The two original marker set are the high density SNP marker (HDSM) set and the low density SNP marker (LDSM) set, the low LD SNP marker sets for both traits are derived from these two marker sets. The four genomic prediction models represented in this figure are: Bayesian B (BayesB), Elastic Net, Reproducing Kernel Hilbert Space (RKHS) and the Ridge Regression Best Linear Unbiased Predictor (RR-BLUP).

4 Discussion

4.1 Simulating the LD and population structure

While simulating the genomic data, we observed that the LD structure was affected by population size. There was a clear trend where an increase in population size caused a reduction in the background and short-range LD and an increase in the rate of LD decay. Additionally, genetic diversity in small populations decreased more rapidly due to allele fixation than in larger populations. The amount of fixed alleles was also greatly reduced when the selection intensity was set to zero.

We hypothesise that these effects on the LD structure were caused by two underlying mechanisms that find their origin in population genetics. As the allele fixation accelerates when population size shrinks and because the population was derived from ten founder individuals, we expect that there is some effect of genetic drift. Genetic drift is the change in allele frequency due to the random sampling of alleles and it is known to cause a rapid loss of genetic diversity in small populations and populations with a founder effect (Charlesworth, 2009; Masel, 2011). As the number of fixed alleles was reduced when selection intensity was set to zero, it is likely that selection pressure further accelerated the process that was initiated by the genetic drift.

As it was out of this project's scope, we decided not to continue our investigation towards these effects. However, we think it is beneficial to understand the population genetics that play a part in our simulations and we recommend to further explore them in future research. This will provide a better framework to simulate genomic data that is true to nature and it will improve the ability to translate the results from simulation to GP applications in actual populations. A good start would be to investigate the hypothesis above, whether genetic drift and selection pressure were the only mechanisms playing a role in allele fixation, what causes the genetic drift in the simulation and how it affected the predictive ability of the GP model.

Only subpopulation Y diverged from populations Z and X in the simulated population, while subpopulations Z and X were very similar (figure 1). Why X and Z were so similar is unclear. Traits X and Z, for which the subpopulations should diverge, did not share any QTLs, the genetic distance between QTLs was sufficient, and they did not share any haplotype-specific SNPs.

4.2 Effect of trait architecture on the predictive ability of the GP models

BayesB and Elastic Net were affected by trait architecture. Their predictive ability was highest for the monogenic traits and decreased with an increase in QTL. A similar trend was observed in several studies (Coster et al., 2010; Daetwyler et al., 2010; Meher et al., 2022). Variable selection models can identify and include the markers in LD with the QTL and zero out the uninformative markers (de los Campos et al., 2013). This is most beneficial for traits with few QTL as there is an associated error margin with the identification of each marker. With an increase in the number of QTL these errors accumulate, typically resulting in a loss of predictive ability (Daetwyler et al., 2010). There are also variable selection models that restrict the number of possible markers in the model. If the number of QTL is greater than the maximum number of markers in the model, part of the genetic variance cannot be captured, also resulting in a loss of predictive ability (de los Campos et al., 2013).

However, for the polygenic traits in our simulation, the variable selection models were still on par with the shrinkage models. This is due to two factors. The first factor is the number of QTL in the polygenic traits. In the paper by Daetwyler et al. (2010), the turning point at which GBLUP (equivalent to RR-BLUP) starts to outperform BayesB lies between 600 and 900 QTL, depending on the heritability of the trait. Although, there are more variables that may affect this, like LD between marker and QTL and the

population size, it is safe to say that the 100 QTL in our polygenic traits were not sufficient to reach this turning point.

The second factor is the flexibility of the models that were used in this comparison. The BGLR package (Pérez & de los Campos, 2014), which implements BayesB, optimises the hyperparameter π, which controls the number of markers in the model. When observing the stored values for π (appendix 2) for traits with high heritability, we see that BayesB includes 1.2% and 47.9% of the markers for the oligo- and polygenic trait architectures, respectively. By optimising π for the predicted trait, BayesB effectively increases its flexibility for applications in traits with polygenic trait architectures, when compared to using a fixed value for π. In Elastic Net the penalty function is a weighted average of the $\ell 1$ and $\ell 2$ norms and is controlled by α. By varying α (0 = RR, 1 = LASSO) there is great flexibility between the extent to which variable selection or shrinkage is applied (Friedman et al., 2010). In this study, α is optimised for every trait using a cross-validation scheme, and it is likely that α was very close to zero for the polygenic traits. Sadly, this cannot be checked as we did not store the optimized values of α.

For the mix traits with low and high heritability, BayesB was performing slightly better than the other models, but these differences were negligible. One could argue that the variable selection models should have performed better due to the strong oligogenic effects in the mix traits. However, the polygenic effects that were also in the mix traits seem to have negated the advantages of variable selection.

The predictive ability of RKHS and RR-BLUP was hardly affected by trait architecture. This can be explained by the fact that shrinkage models include all the markers in their estimation of the genotype, irrespective of the number of QTL. As there is no error margin associated with the identification of important markers, the noise in the model depends solely on the heritability of the trait and the extent of LD between markers and QTL (Daetwyler et al., 2010).

4.3 Effect of heritability on the predictive ability of the GP models

A clear correlation was observed between the predictive ability of the models and the heritability of the traits. With an increase in heritability, the predictive ability of the models improved, and similar trends were observed in literature (Daetwyler et al., 2008; Kaler et al., 2022; Meher et al., 2022)

It has also been reported that heritability is an important variable to take into account when choosing a GP model (Crossa et al., 2017; de los Campos et al., 2013; Kaler et al., 2022; Meher et al., 2022). However, only for the oligogenic traits we saw evidence for this claim. All models were on par for the oligogenic trait with low heritability, but BayesB and Elastic Net outperformed RKHS and RR-BLUP for the intermediate and high heritability. It seems that BayesB and Elastic Net could not effectively detect the markers in LD with the oligogenic QTL when heritability was low, as they explained only a tiny portion of the total phenotypic variation. This observation can be explained by examining the optimised values of π for the oligogenic traits (appendix 2). For the low heritability oligogenic trait π was 0.388, indicating that BayesB was unable to effectively identify the markers in LD with the QTL. With an increase in heritability, the value for π decreases, indicating that the markers in LD with the QTL are identified more effectively, resulting in less noise. Presumably, a similar thing happened for Elastic Net, where α was close to zero (similar to RR), increasing the number of markers in the regression model.

4.4 Effect of marker density on the predictive ability of GP models

Decreasing marker density caused an overall reduction in predictive ability. This was expected as low marker density reduces the chance of having markers in LD with the QTL (de los Campos et al., 2013). When LD is low, and the models are unable to effectively capture the genetic variance (Voss-Fels et al., 2018), resulting in a loss of predictive ability.

At first glance, the drop in predictive ability was relatively small for all traits. The low-density SNP marker (LDSM) set had a marker density of approximately one SNP marker per 1.1 cM. In paragraph 2.1, it is reported that $LD_{1/2,90}$ is 4.071 cM. With one SNP per 1.1 cM, the markers in the LDSM set were well within this range. Probably enough LD between the SNP markers and QTLs was left to reach accurate predictions.

Although BayesB and Elastic Net still outclassed RKHS and RR-BLUP for the mono- and oligogenic traits, the difference became considerably smaller when marker density was reduced. Therefore, it can be stated that BayesB and Elastic Net were affected more by a reduction in marker density than RKHS and RR-BLUP, at least for these traits. Similar trends have been reported in the review paper by de los Campos et al. (2013). They state that reducing marker density, or LD, has a greater effect on the prediction ability of variable selection models than that of shrinkage models as they have more trouble to identify the important markers. Still, for none of the traits, BayesB or Elastic Net had lower predictive ability than RKHS or RR-BLUP. As there is such a clear difference in the reaction to marker density we recommend to expand the comparison with a wider range of marker sets.

On average, the drop in predictive ability, when reducing marker density, was more than twice as large for high heritability traits than for low heritability traits, indicating an interaction effect between heritability and marker density. Naturally LD and heritability are connected. Heritability is the proportion of the total phenotypic variation that is explained by the QTLs (Ge et al., 2017) and LD determines how much of the genetic variation is captured by the markers(Goddard & Hayes, 2007; Meuwissen et al., 2001). Both variables introduce noise in the regression model in their own way. When heritability is low, noise is introduced in the form of environmental variation. When LD is low, noise is introduced because the markers are unable to capture the genetic variance, making it more difficult to accurately estimate the QTL effects. One could speculate that, for high heritability traits, a low amount of LD will introduce relatively more noise to the GP model than for low heritability traits. In other words, low heritability already causes so much noise that a decrease in LD will not have as large of an impact.

4.5 Effect of population structure on the predictive ability of GP models

During the first two experiments, we observed some clear indications that the predictive ability for trait X was inflated by the effect of population structure. Despite the oligogenic trait architecture there was very little difference between the models, the predictive ability was consistently higher than heritability and the drop in predictive ability was minimal when marker density was reduced. To further investigate this effect, we removed almost all the markers in LD to the QTL of trait X assuming that the remaining predictive ability was attributable to population structure. As a comparison, we did the same thing for the oligogenic trait with a heritability of 0.53.

Removing the LD from the marker sets resulted in a drop in predictive ability for both trait X and the oligogenic trait. However, the drop in predictive ability was much larger for the oligogenic trait than for trait X. As trait X was simulated to correlate with population structure and the oligogenic trait was not, these results clearly indicate that a large part of the predictive ability for trait X is attributable to population structure.

In GP models, the accuracy of the predictions depends on the amount of genetic variation that is captured by the markers. In distantly related individuals, the only source of genetic variation that is captured by markers is due to their LD with the QTL. However, in populations with a strong population structure, the genetic relationship between individuals is also captured by the markers, providing an additional source of genetic variation, which inflates the predictive ability of GP models (Daetwyler et al., 2012; Habier et al., 2007). This is an important insight for the practical application of GP in breeding. If the population structure disappears due to crossing between or selection within subpopulations, the markers that capture genetic variation due to the relationship between individuals will lose their meaning and the predictive ability of the GP models will "deflate". When this goes unnoticed because the breeder is unaware of the effect of population structure in a particular trait, it can significantly affect the selection accuracy in subsequent generations and hamper genetic gain. Therefore, we recommend investigating population structure thoroughly for every new trait and reassessing the predictive ability regularly. A recommendation for future research would be to develop a protocol that effectively estimates population structure and how it affects the trait of interest.

Although the predictive ability of the oligogenic trait dropped further than the predictive ability of trait X, a substantial portion remained. As almost all LD between SNP markers and the QTLs of the oligogenic trait was removed, there seems to be an additional source of genetic variation that contributes to the predictive ability of the oligogenic trait. One hypothesis was that the oligogenic trait correlated to population structure by chance. The means of the three subpopulations were compared for the oligogenic trait to see if this was true (appendix 3). When examining the means, a similar trend was observed in trait X, where the means of subpopulations X and Z were higher than that of subpopulation Y. However, these differences were unconvincing. To explain the remaining part of the predictive ability for the oligogenic trait, additional research is needed.

4.6 Model performance

BayesB and Elastic Net performed either better than, or on par with, RKHS and RR-BLUP for all traits. Based on predictive ability alone one could conclude that BayesB and Elastic Net were the better options for our population, irrespective of the trait. However, there are more factors to take into consideration. Computational time was considerably longer for both BayesB and Elastic Net than for RKHS and RR-BLUP. For the HDSM set with 3622 markers, BayesB took up to three hours (12500 iterations) to complete one trait, Elastic Net with the double cross-validation scheme could take up to two hours to complete one trait and RKHS and RR-BLUP only took about 15 minutes to complete all 13 traits. This is an important difference and with larger marker sets this difference will become even larger. Especially for the polygenic and mix traits, for which the models gave similar prediction abilities, this computational time is an important consideration.

When trait architecture and heritability are unknown we recommend to use BayesB over all models. Although BayesB and Elastic Net performed similar, Elastic Net (implemented by glmnet version 4.1-3) had some problems when predicting the polygenic and mix trait with low heritability. The Elastic Net penalty could become too stringent with an increase in α, resulting in model fits with only an intercept or estimated marker effects that were too small for R to handle. Under these circumstances, all predicted genotypes were equal, and no correlation could be calculated with the actual phenotypic values, resulting in a NA for predictive ability. This makes it a less appropriate model for these traits.

One of the benefits of RKHS is that it is able to capture non-additive genetic effect such as dominance and epistasis without the need to model for them explicitly (Gota & Gianola, 2014). As all the traits in this study were simulated to follow an additive genetic model, we were not able to exploit this benefit. For further comparisons we recommend to include traits with non-additive effect as well to see how the models hold up.

5 Conclusion

This research aimed to compare several GP models in polyploid data and we observed how their predictive ability was affected by trait architecture, heritability, marker density and population structure. Our results correspond to what is found in literature. For the models that were able to apply variable selection, it was more beneficial to predict traits governed by few QTL than to predict traits with many QTL. In contrast, the shrinkage models were mostly unaffected by trait architecture. All models were positively affected by an increase in heritability. However, this effect was more dramatic for the variable selection models when predicting the oligogenic traits than for the shrinkage models. We observed a similar trend for marker density, as a reduction in marker density caused an overall drop in predictive ability. When predicting the mono- and oligogenic traits, the drop in predictive ability was more severe for the variable selection models than for the shrinkage models. Of course, it is important to note that although trait architecture, heritability and marker density affected the predictive ability of the variable selection models more, there was not one trait for which the shrinkage models outcompeted the variable selection models. The predictive ability for trait X was inflated by the genetic variation explained by the population structure, negating most of the effects of traits architecture, heritability and marker density.

We would like to rank the models per trait architecture. BayesB and Elastic Net are preferred over RKHS and RR-BLUP for mono- and oligogenic traits as their predictive ability was consistently higher. For the traits with polygenic effects, RKHS and RR-BLUP are preferred over BayesB and Elastic Net, as their predictive ability was similar, but the computation time of RKHS and RR-BLUP was much shorter. As Elastic Net had trouble with predicting for polygenic and mix traits with low heritability, we recommend BayesB when trait architecture and heritability are unknown.

However, we also showed that the predictive ability of these GP models is affected by many variables and this ranking may vary for different populations, traits and marker densities. Therefore, we recommend to continue comparing GP models and gather more information for polyploid data. An excellent start would be to increase the variety in traits, e.g. more QTLs and non-additive effects, and to expand on marker density. Further, we recommend investigating the matter of genetic drift and selection intensity to see how they affect LD structure in the simulation and the predictive ability of the models. Finally, we recommend developing a protocol that can quantify the effect of population structure on the predictive ability of GP models.

Literature

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, *3*(10), e3376. https://doi.org/10.1371/JOURNAL.PONE.0003376

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics 2009 10:3*, *10*(3), 195–205. https://doi.org/10.1038/nrg2526

Clark, S. A., & van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods in Molecular Biology*, *1019*, 321–330. https://doi.org/10.1007/978-1-62703-447-0_13

Collard, B. C. Y., & Mackill, D. J. (2008). 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. https://doi.org/10.1098/RSTB.2007.2170

Coster, A., Bastiaansen, J. W. M., Calus, M. P. L., Van Arendonk, J. A. M., & Bovenhuis, H. (2010). Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genetics Selection Evolution*, *42*(1), 1–11. https://doi.org/10.1186/1297-9686-42-9/TABLES/8

Crain, J., DeHaan, L., Poland, J., Jesse Poland, C., & Genet-, W. (2020). *Genomic prediction enables rapid selection of high-performing genets in an intermediate wheatgrass breeding program*. https://doi.org/10.1002/tpg2.20080

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., De Los Campos, G., Burgueño, J., González-Camacho, J. M., Pérez-Elizalde, S., Beyene, Y., Dreisigacker, S., Singh, R., Zhang, X., Gowda, M., Roorkiwal, M., Rutkoski, J., & Varshney, R. K. (2017). Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends in Plant Science*, *22*(11), 961–975. https://doi.org/10.1016/j.tplants.2017.08.011

Cuyabano, B. C. D., Su, G., & Lund, M. S. (2014). Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics*, *15*(1), 1–11. https://doi.org/10.1186/1471-2164-15-1171/TABLES/9

Cuyabano, B. C. D., Su, G., & Lund, M. S. (2015). Selection of haplotype variables from a high-density marker map for genomic prediction. *Genetics Selection Evolution*, *47*(1), 1–11. https://doi.org/10.1186/S12711-015-0143-3/TABLES/6

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J., & Hayes, B. J. (2012). Components of the accuracy of genomic prediction in a multi-breed sheep population. *Journal of Animal Science*, *90*(10), 3375–3384. https://doi.org/10.2527/JAS.2011-4557

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics*, *185*, 1021–1031. https://doi.org/10.1534/genetics.110.116855

Daetwyler, H. D., Villanueva, B., & Woolliams, J. A. (2008). Accuracy of Predicting the Genetic Risk of Disease Using a Genome-Wide Approach. *PLOS ONE*, *3*(10), e3395. https://doi.org/10.1371/JOURNAL.PONE.0003395

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics 2011*, *12*(7), 499–510. https://doi.org/10.1038/nrg3012

de Bem Oliveira, I., Resende, M. F. R., Ferrão, L. F. v., Amadeu, R. R., Endelman, J. B., Kirst, M., Coelho, A. S. G., & Munoz, P. R. (2019). Genomic Prediction of Autotetraploids; Influence of Relationship Matrices, Allele Dosage, and Continuous Genotyping Calls in Phenotype Prediction. *G3 Genes|Genomes|Genetics*, *9*(4), 1189–1198. https://doi.org/10.1534/G3.119.400059

de Los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., & Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research*, *92*(4), 295–306. https://doi.org/10.1017/S0016672310000285

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. L. (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, *193*(2), 327–345. https://doi.org/10.1534/GENETICS.112.143313

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: an overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69. https://doi.org/10.1111/MEC.12581

Edwards, M. D., Stuber, C. W., & Wendel, J. F. (1987). Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. *Genetics*, *116*(1), 113–125. https://doi.org/10.1093/GENETICS/116.1.113

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/JOURNAL.PONE.0019379

Endelman, J. B. (2011). Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, *4*(3), 250–255. https://doi.org/10.3835/PLANTGENOME2011.08.0024

Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., de Jong, W. S., Douches, D. S., Frederick, C. M., Haynes, K. G., Holm, D. G., Miller, J. C., Muñoz, P. R., Navarro, F. M., Novy, R. G., Palta, J. P., Porter, G. A., Rak, K. T., Sathuvalli, V. R., … Yencho, G. C. (2018). Genetic variance partitioning and Genome-Wide prediction with allele dosage information in autotetraploid Potato. *Genetics*, *209*(1), 77–87. https://doi.org/10.1534/GENETICS.118.300685/-/DC1

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *JSS Journal of Statistical Software*, *33*(1). http://www.jstatsoft.org/

Ge, T., Holmes, A. J., Buckner, R. L., Smoller, J. W., & Sabuncu, M. R. (2017). Heritability analysis with repeat measurements and its application to resting-state functional connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(21), 5521–5526. https://doi.org/10.1073/PNAS.1700765114/SUPPL_FILE/PNAS.201700765SI.PDF

Goddard, M. E., & Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323–330. https://doi.org/10.1111/J.1439-0388.2007.00702.X

Gota, M., & Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: A review. *Frontiers in Genetics*, *5*(OCT), Article 363. https://doi.org/10.3389/FGENE.2014.00363/BIBTEX

Gupta, P. K., Rustgi, S., & Mir, R. R. (2008). Array-based high-throughput DNA markers for crop improvement. *Heredity*, *101*(1), 5–18. https://doi.org/10.1038/HDY.2008.35

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics*, *177*(4), 2389–2397. https://doi.org/10.1534/GENETICS.107.081190

Haile, T. A., Heidecker, T., Wright, D., Neupane, S., Ramsay, L., Vandenberg, A., & Bett, K. E. (2020). Genomic selection for lentil breeding: Empirical evidence. *The Plant Genome*, *13*(1). https://doi.org/10.1002/TPG2.20002

Hickey, J. M., Chiurugwi, T., Mackay, I., & Powell, W. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature Genetics 2017 49:9*, *49*(9), 1297–1303. https://doi.org/10.1038/ng.3920

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55–67.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R Second Edition*.

Jia, C., Zhao, F., Wang, X., Han, J., Zhao, H., Liu, G., & Wang, Z. (2018). Genomic prediction for 25 agronomic and quality traits in alfalfa (medicago sativa). *Frontiers in Plant Science*, *9*, 1220. https://doi.org/10.3389/FPLS.2018.01220/BIBTEX

Kaler, A. S., Purcell, L. C., Beissinger, T., & Gillman, J. D. (2022). Genomic prediction models for traits differing in heritability for soybean, rice, and maize. *BMC Plant Biology*, *22*(1), 1–11. https://doi.org/10.1186/S12870-022-03479-Y/FIGURES/3

Lloyd, A., & Bomblies, K. (2016). Meiosis in autopolyploid and allopolyploid Arabidopsis. *Current Opinion in Plant Biology*, *30*, 116–122. https://doi.org/10.1016/J.PBI.2016.02.004

Loos, R. J. F. (2020). 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications 2020 11:1*, *11*(1), 1–3. https://doi.org/10.1038/s41467-020-19653-5

Masel, J. (2011). Genetic drift. *Current Biology*, *21*(20). https://doi.org/10.1016/J.CUB.2011.08.007

Matei, G., Woyann, L. G., Milioli, A. S., de Bem Oliveira, I., Zdziarski, A. D., Zanella, R., Coelho, A. S. G., Finatto, T., & Benin, G. (2018). Genomic selection in soybean: accuracy and time gain in relation to phenotypic selection. *Molecular Breeding*, *38*(9), 1–13. https://doi.org/10.1007/S11032-018-0872-4/TABLES/3

Meher, P. K., Rustgi, S., & Kumar, A. (2022). Performance of Bayesian and BLUP alphabets for genomic prediction: analysis, comparison and results. *Heredity*, *128*(6), 519–530. https://doi.org/10.1038/S41437-022-00539-9

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

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., & Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, *335*(6192), 721–726. https://doi.org/10.1038/335721A0

Pérez, P., & de los Campos, G. (2014). Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, *198*(2), 483–495. https://doi.org/10.1534/GENETICS.114.164442

Ramsey, J., & Schemske, D. W. (2002). NEOPOLYPLOIDY IN FLOWERING PLANTS. *Annu. Rev. Ecol. Syst*, *33*, 589–639. https://doi.org/10.1146/annurev.ecolsys.33.010802.150437

Rice, B., & Lipka, A. E. (2019). Evaluation of RR-BLUP Genomic Selection Models that Incorporate Peak Genome-Wide Association Study Signals in Maize and Sorghum. *The Plant Genome*, *12*(1), 180052. https://doi.org/10.3835/PLANTGENOME2018.07.0052

Riedelsheimer, C., Technow, F., & Melchinger, A. E. (2012). Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines. *BMC Genomics*, *13*(1), 1–9. https://doi.org/10.1186/1471-2164-13-452/FIGURES/3

Sallam, A. H., Conley, E., Prakapenka, D., Da, Y., & Anderson, J. A. (2020). Improving Prediction Accuracy Using Multi-allelic Haplotype Prediction and Training Population Optimization in Wheat. *G3 Genes|Genomes|Genetics*, *10*(7), 2265–2273. https://doi.org/10.1534/G3.120.401165

Sarinelli, J. M., Murphy, J. P., Tyagi, P., Holland, J. B., Johnson, J. W., Mergoum, M., Mason, R. E., Babar, A., Harrison, S., Sutton, R., Griffey, C. A., & Brown-Guedira, G. (2019). Training population selection and use of fixed effects to optimize genomic predictions in a historical USA winter wheat panel. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *132*(4), 1247–1261. https://doi.org/10.1007/S00122-019-03276-6

Scutari, M., Mackay, I., Balding, D. ;, Valdar, W., Solberg, L. C., Gauguier, D., Burnett, S., Klenerman, P., & Cookson, W. O. (2016). Genome-Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice. *Theor Appl Genet*, *12*(9), 879–887. https://doi.org/10.1371/journal.pgen.1006288

Soller, M. (1978). The use of loci associated with quantitative effects in dairy cattle improvement. *Animal Science*, *27*(2), 133–139. https://doi.org/10.1017/S0003356100035960

Soller, M., & Plotkin-Hazan, J. (1977). The use marker alleles for the introgression of linked quantitative alleles. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *51*(3), 133–137. https://doi.org/10.1007/BF00273825

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Source: Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Voorrips, R. E., & Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, *13*(1), 1–12. https://doi.org/10.1186/1471-2105-13-248/TABLES/3

Voss-Fels, K. P., Cooper, M., & Hayes, B. J. (2018). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics 2018 132:3*, *132*(3), 669–686. https://doi.org/10.1007/S00122-018-3270-8

Wilson, S., Zheng, C., Maliepaard, C., Mulder, H. A., Visser, R. G. F., van der Burgt, A., & van Eeuwijk, F. (2021). Understanding the Effectiveness of Genomic Prediction in Tetraploid Potato. *Frontiers in Plant Science*, *12*, Article 672417. https://doi.org/10.3389/fpls.2021.672417

Appendix 1. list with QTL of the polygenic traits

Table 4 List with QTL of the polygenic trait

QTL polygenic traits									
"A0095"	"A0357"	"A0809"	"B0143"	"B0619"	"C0074"	"C0404"	"C0874"	"D0256"	"D0593"
"A0127"	"A0396"	"A0839"	"B0189"	"B0669"	"C0082"	"C0439"	"C0901"	"D0312"	"D0622"
"A0157"	"A0579"	"A0983"	"B0312"	"B0691"	"C0086"	"C0477"	"C0903"	"D0313"	"D0631"
"A0267"	"A0629"	"A0989"	"B0317"	"B0731"	"C0113"	"C0526"	"C0972"	"D0333"	"D0670"
"A0281"	"A0637"	"A1000"	"B0347"	"B0780"	"C0129"	"C0566"	"C0974"	"D0407"	"D0678"
"A0291"	"A0643"	"B0015"	"B0394"	"B0806"	"C0224"	"C0639"	"D0004"	"D0438"	"D0837"
"A0305"	"A0705"	"B0046"	"B0508"	"B0829"	"C0262"	"C0666"	"D0040"	"D0452"	"D0862"
"A0308"	"A0711"	"B0070"	"B0518"	"B0889"	"C0346"	"C0818"	"D0152"	"D0487"	"D0911"
"A0309"	"A0757"	"B0079"	"B0533"	"B0960"	"C0366"	"C0821"	"D0183"	"D0524"	"D0912"
"A0319"	"A0782"	"B0093"	"B0612"	"B0970"	"C0401"	"C0830"	"D0244"	"D0546"	"D0991"

Appendix 2 List with values for π (BayesB) and α (Elastic Net) for the oligogenic traits

Table 5 Stored values of π for all traits in the high density marker set and low density marker set

Trait architecture	Realised heritability	π (high density marker set)	π (low density marker set)
monogenic	0.197	0.205	0.299
	0.480	0.045	0.041
	0.800	0.055	0.065
oligogenic	0.226	0.388	0.466
	0.530	0.169	0.269
	0.777	0.012	0.204
Polygenic	0.203	0.479	0.439
	0.501	0.354	0.538
	0.798	0.461	0.642
Mix (architecture with both oligo- and polygenic effects)	0.202	0.332	0.477
	0.474	0.277	0.516
	0.745	0.246	0.586
X	0.688	0.145	0.316

Appendix 3 subpopulation means for trait X and the oligogenic trait with an heritability of 0.53

Table 6 Mean and standard deviation of the phenotypes of each subpopulation for the oligogenic trait with an heritabiliy of 0.53 and trait X

Trait	Subpopulation	mean	Standard deviation
X	X	0.458	0.715
	Y	-1.068	0.705
	Z	0.456	0.734
Oligo	X	0.128	0.992
	Y	-0.312	0.945
	Z	0.148	1.007

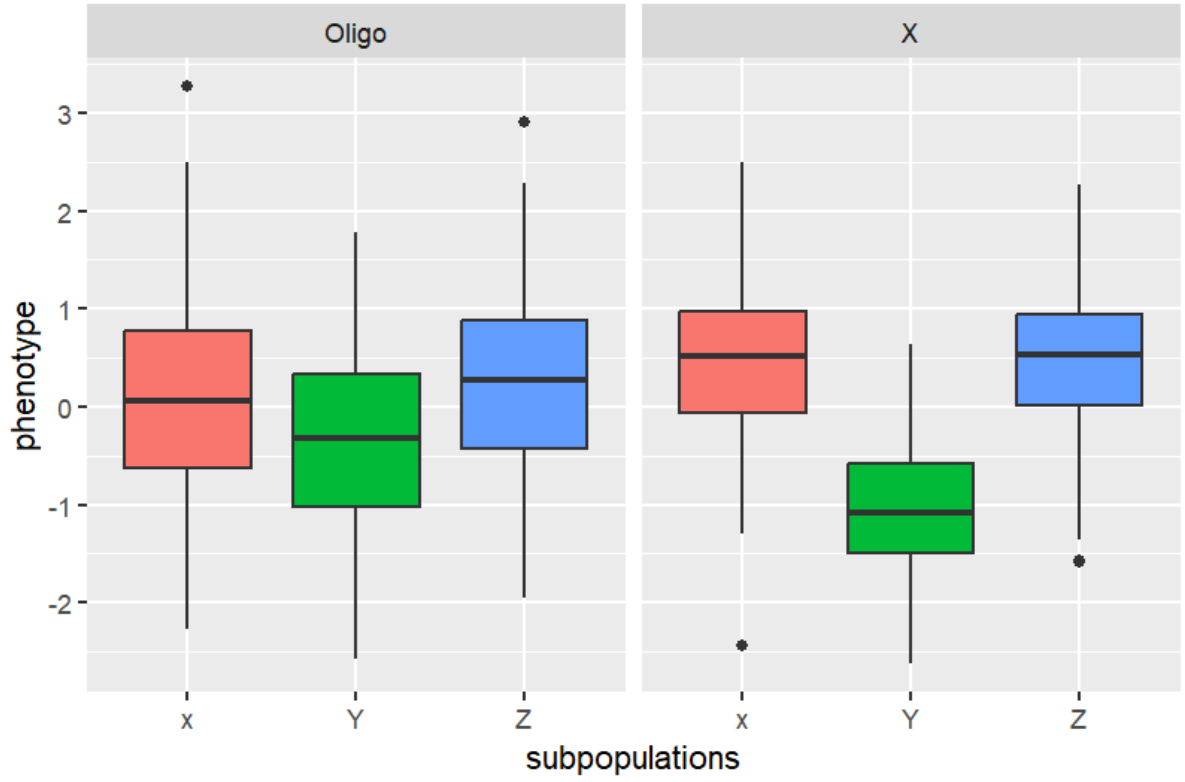

Figure 6 Boxplots representing the phenotypes of each subpopulation for the oligogenic trait with an heritability of 0.53 and trait X