MSc Thesis Report

Wageningen University

MSc Plant Sciences

Laboratory of Plant Breeding

Supervision:

Chris Maliepaard

Roeland Voorrips

Genomic Selection

Incorporating prior QTL information

George Korontzis

881115467050

Wageningen

2015

Acknowledgments

I am grateful to my supervisors Chris Maliepaard and Roeland Voorrips for the guidance through this project.

Also, I thank Albart Coster for providing me helpful script for the use of the "HaploSim" package and Jochen Reif and Yusheng Zhao of IPK Gatersleben for the correspondence that helped me apply their method (W-BLUP).

Finally, I thank the Onassis Foundation for the financial support throughout my MSc studies at Wageningen University.

Contents

Abstract

Genomic selection has been described as a method that can revolutionize plant breeding. A wide series of studies on simulated and real data have demonstrated the potential of this approach for breeding for complex traits. Genomic selection methods use genome-wide marker data to predict the breeding value of unphenotyped individuals by forming a regression model on a training population, but the vast majority of methods ignore prior information on QTLs. Incorporation of prior information could improve the prediction accuracy of genomic selection.

Recently several methods have been developed to meet this objective. These methods propose a special treatment of markers that are known to capture the effects of major genes. In the studied cases, these markers are either known genes or markers identified by Genome-Wide Association Studies.

The main aim of this study is to evaluate the potential of incorporating information derived by QTL mapping experiments that are commonly found in literature. This is achieved by applying two different approaches. The first is by fitting specific markers as fixed effects in an RR-BLUP model (RR-BLUP/FIXED) and the second is to estimate separate shrinkage parameters for these markers (W-BLUP).

For this purpose, a simulation study was conducted. The simulation included generation of marker data under linkage disequilibrium, simulation of a QTL mapping experiment and genomic prediction of hybrid performance.

The results imply that it is possible to increase the prediction accuracy by incorporating prior information but further research is required in order to specify the conditions that promote this increase.

1. Introduction

1.1. Breeding for quantitative traits

1.1.1. Quantitative traits

Quantitative genetics is a major pillar of plant breeding science due to the fact that many important agronomic traits such as yield are quantitative. These traits are complex and regulated by many genes. In most cases, the phenotype for these traits is measured on a continuous scale and is affected strongly by the environment.

1.1.2. Conventional approach

1.1.2.1. Variance components

The genetics of quantitative traits is studied using a biometrical approach. This approach is based on the decomposition of the phenotypic variance in variance components. The phenotypic variance can be divided in genetic and environmental variance, under the assumption of no genotype by environment interaction. This decomposition is easily achieved by replication of the observed genotypes. The genetic variance can be further separated in additive, dominance and epistatic variance. The most important part of the genetic variance is the additive variance as it is the one that can be used to predict the performance of the progeny. The part of the phenotypic value that is due to additive effects is called breeding value. Using the variance components, the broad and narrow sense heritability coefficients can be defined as $h^2_{bs} = \frac{\sigma^2_G}{\sigma^2_P}$ and $h^2_{ns} = \frac{\sigma^2_A}{\sigma^2_P}$ respectively. In most cases of plant breeding, the phenotype is used as a selection criterion rather than the breeding value, but the heritability coefficient is used to predict the response to selection.

1.1.2.2. Best Linear Unbiased Prediction

Breeding values can be estimated by Best Linear Unbiased Prediction (BLUP) (Henderson, 1975). For example, for the mixed model $y = \boldsymbol{X}\beta + \boldsymbol{Z}a + e$, where β is a vector of fixed effects, a is a vector of random genotype effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ the corresponding design matrices, $e \sim N(0, \boldsymbol{I}\sigma^2_e)$, $a \sim N(0, \boldsymbol{A}\sigma^2_a)$ and $\boldsymbol{A}$ is the relationship matrix, BLUP for a can be obtained by solving the Mixed Model Equations (MME):

$$\begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X^T X} & \mathbf{X^T Z} \\ \mathbf{Z^T X} & \mathbf{Z^T Z} + \frac{\sigma^2_e}{\sigma^2_a}\mathbf{A^{-1}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X^T} y \\ \mathbf{Z^T} y \end{bmatrix} \text{[eq. 1]}$$

In this case, $\boldsymbol{A}$ consists of kinship coefficients that are estimated by pedigree data and define the covariance between genotypes. The variance components σ_e^2 and σ_a^2 can be estimated by Restricted Maximum Likelihood (REML).

1.1.2.3. Prediction of hybrid performance

In hybrid breeding, the prediction of hybrid performance can be done in several ways such as: i) line performance per se, ii) prediction based on the general combining ability (GCA) and iii) BLUP (Schrag et al., 2009, Zhao et al., 2015). In the BLUP method, prediction is based on both the GCA and the specific combining ability (SCA) effects by using pedigree information to estimate the kinship between the individuals.

Prediction based on both the general and specific combining ability can be more accurate when dominance effects are significant. The separation of genetic variance in additive and dominance effects is achieved by the analysis of a diallel design when all possible hybrids are of interest or a factorial design when parents are divided in heterotic pools or males and females. A factorial design can be analyzed using REML for the mixed model:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}_F g_{GCAfemale} + \boldsymbol{Z}_M g_{GCAmale} + \boldsymbol{Z}_S g_{SCA} + e \qquad \text{[eq. 2]}$$

where $g_{GCAfemale}$ and $g_{GCAmale}$ are vectors of the female and male GCA effects respectively and g_{SCA} is the vector of SCA effects. The design matrices $\boldsymbol{Z}_F$ and $\boldsymbol{Z}_M$, code for the male and female parents and the design matrix $\boldsymbol{Z}_S$ codes for the identity of the hybrid.

A description of the analysis of a diallel design by REML can be found in literature(Möhring et al., 2011).

1.1.3. Marker-basedapproach

In the last decades, marker-based approaches have been implemented in plant breeding. QTL mapping is able to locate regions of the genome with significant association to the trait. This information can lead to the identification of a gene, or can be used directly in marker-assisted selection. Marker-Assisted Selection (MAS) is used for the introgression of a trait by selecting the individuals that have the desirable allele of a marker that is linked to the trait of interest. MAS has been applied successfully in breeding for simple monogenic or oligogenic traits.

In the case of quantitative traits, QTLs have been identified, major genes have been characterized, but marker-assisted selection has not been effective. The reason is that marker assisted selection is designed for introgressing a small number of genes. Furthermore, QTL mapping is not able to locate minor genes with a small effect that can however be very important in cases such as the breeding of quantitative traits in elite germplasm. This is because QTL mapping includes hypothesis testing to select the markers that are associated to the trait and to avoid false positives.

1.2. Genomic selection

1.2.1. Concept

Driven by the availability of genome-wide marker data, genomic selection was described by Meuwissen et al. (Meuwissen et al., 2001) as a form of marker assisted selection that is based on estimating breeding values.

Genomic selection (GS) or Genome-wide selection or Whole-genome regression is a selection method based on the prediction of the breeding value from genome-wide marker data. The basic steps are: i) phenotype and genotype a training population, ii) construct a GS modelestimating regression coefficients for all markersand iii) use this model to calculate the Genomic Estimated Breeding Values (GEBVs) in a breeding population and select without phenotyping in the following generations.

Unlike MAS, it does not include statistical testing for the allocation of QTL's. GS uses a large number of markers to capture the genetic variance, so that ideally all polygenes affecting the trait will be selected, even the ones with a small effect. For this reason GS is well-suited for quantitative traits.

The last decade, the interest in genomic selection is increasing and many reviews have been published exposing the potential of genomic selection(Bernardo and Yu, 2007, de los Campos et al., 2013a, Desta and Ortiz, 2015, Heffner et al., 2009, Jannink et al., 2010, Nakaya and Isobe, 2012)

1.2.2. Statistical methods for GS

1.2.2.1. The curse of dimensionality

The statistical methods proposed for genomic selection aim at tackling the "large p - small n" problem that is caused by the use of many markers. This problem, mentioned also as "the curse of dimensionality", arises when the number of variables p (markers) is higher than the number of observations n (individuals). The result is over-fitting that decreases the predictive ability of the model for individuals not included in the training population.

In general, this problem can be handled:i) by variable selection or dimension reduction methods (e.g. forward selection, stepwise regression, PLS, PCR etc.)(Bernardo and Yu, 2007, Resende et al., 2014)to decrease the number of variables, or ii) by shrinkage methods. In the second case, that is more suitable for genomic selection, the regression coefficients are shrunk towards zero. This way, a bias is introduced and the variance of the coefficient is reduced in order to balance the goodness of fit and the complexity of the model (de los Campos et al., 2013a). Shrinkage is usually applied either in the context of penalized or Bayesian regression.

1.2.2.2. Penalized regression and BLUP

1.2.2.2.1. Ridge regression and LASSO

The two most important methods of penalized regression are ridge regression (RR) and least absolute shrinkage and selection operator (LASSO). Both methods differ from ordinary least squares estimation, as they impose a penalty term in the error function that is minimized. This way the coefficients are shrunk towards zero. The penalty is regulated by the shrinkage (or regularization) parameter λ. In ridge regression, λis applied on the sum of the squared regression coefficients. In contrast, LASSO applies λ on the sum of the absolute values of the regression coefficients. This way, in RR small coefficients receive less shrinkage than in LASSO, small effects are captured by the model and RR can be more suitable for highly complex traits. LASSO shrinks the coefficients of small effects to zero or very close to zero, and therefore can be used for variable selection. The fact that in LASSO the number of non-zero estimates cannot be higher than the number of observations, makes LASSO less suitable for genomic selection (de los Campos et al., 2013a).

1.2.2.2.2. RR-BLUP

The parameter λ can be chosen by cross-validation, but in the practice of genomic selection λ is estimated as the ratio of the residual variance and the variance of marker

effects, under the assumption of equal variance for all marker effects (Meuwissen et al., 2001).In the second case, the shrinkage imposed by ridge regression is equivalent to the shrinkage imposed by BLUP according to the model:

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}u + e \quad \text{[eq. 3]}$$

Where $u \sim N(0, \sigma_u^2)$is a vector of marker effects fitted as random,$\boldsymbol{Z}$ the design matrix for marker effectsand $\boldsymbol{X}\beta$ the fixed part of the model (environments etc.).

For this reason ridge regression is applied in a mixed model context and is mentioned as RR-BLUP. RR-BLUP is the most common method for genomic selection.

BLUP estimates can be obtained by solving the MME:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^{\mathbf{T}}\mathbf{X} & \mathbf{X}^{\mathbf{T}}\mathbf{Z} \\ \mathbf{Z}^{\mathbf{T}}\mathbf{X} & \mathbf{Z}^{\mathbf{T}}\mathbf{Z} + \frac{\sigma_e^2}{\sigma_u^2}\mathbf{A}^{-\mathbf{1}} \end{bmatrix}^{-\mathbf{1}} \begin{bmatrix} \mathbf{X}^{\mathbf{T}}y \\ \mathbf{Z}^{\mathbf{T}}y \end{bmatrix} \text{[eq. 4]}$$

There are two ways to implement RR-BLUP. The first is by REML analysis of the mixed model followed by BLUP(Piepho, 2009). This way the variance of marker effects $\sigma_u^2$is directly estimated. This can be done using mixed model software such as ASReml. For mixed models with a single random factor (apart from the residual variance), an R package named rrBLUP has been developed (Endelman, 2011).

An alternative is to use available estimates of the variance of additive genotype effects σ_A^2 in order to estimate σ_u^2(Resende et al., 2014).Several approaches exist for coding the design matrices and estimating σ_u^2 . The simplest approach is to code additive effects as {aa, Aa, AA} = {-1, 0, 1} and estimate σ_u^2 as $\frac{\sigma_A^2}{m}$, where m is the number of markers(Technow et al., 2012, Zhao et al., 2015).Alternatively σ_u^2 can be estimated as $\frac{\sigma_A^2}{\sum_j 2p_j(1-p_j)}$,where p_j is the marker allele frequency. Other approaches include coding as {aa, Aa, AA} = {0, 1, 2} followed by mean centeringand standardization (Resende et al., 2014).

The Genomic Estimated Breeding Values for unphenotyped individuals can be calculated as $\boldsymbol{Z_B}\hat{u}$ where $\boldsymbol{Z_B}$ is the design matrix for marker effects for the individuals of the breeding population.

1.2.2.2.3. G-BLUP

G-BLUP is equivalent to ridge regression, but the effects modeled are the genotype effects. The G-BLUP formulation is similar to kinship-BLUP, but in G-BLUPthe genomic relationship matrix$\boldsymbol{G}$is estimated by marker data(VanRaden, 2008).

$$y = \boldsymbol{X}\beta + g + e \qquad \text{[eq. 5]}$$

Where $g = \boldsymbol{Z}u,\ \ g \sim N(0, \boldsymbol{G}\sigma_g^2),\ \ \boldsymbol{G} = \frac{\boldsymbol{ZZ^T}}{\sum_j 2p_j(1-p_j)}$

1.2.2.3. Bayesian shrinkage

Shrinkage of estimates can be also obtained in Bayesian methods by assigning a prior density to marker effects. Ridge regression can be applied in a Bayesian context (Bayesian Ridge Regression – BRR) by using a Gaussian prior for the marker effects. Also, Bayesian methods can induce variable selection if the prior distribution has high density around zero(de los Campos et al., 2013a).

In genomic selection, a series of Bayesian methods have been applied that, in contrast to RR-BLUP, do not assume normally distributed marker effects. BayesA assigns a scaled-t prior that has thicker tails and higher mass at zero than the Gaussian.The prior in BayesB is a mixture of a scaled-t and a point of mass at zero(Meuwissen et al., 2001). The mixture is regulated by the parameter π that defines the proportion of markers with zero variance. In BayesC, the mixture has a Gaussian distribution instead of a scaled-t(Habier et al., 2011). Parameter πcan be either predefined or considered unknown and estimated from the data.

1.2.2.4. Genomic selection for hybrid prediction

RR-BLUP (and G-BLUP) can be extended to model both additive and dominance effects of the markers (or genotypes). G-BLUP for hybrid prediction is similar to the BLUP method mentioned in 1.1.2.2. with the difference that the relationship matrix is estimated by marker data.

The RR-BLUP model is:

$$y = \mathbf{1_n} + \boldsymbol{Z}_A u_A + \boldsymbol{Z}_D u_D + e \qquad \text{[eq. 6]}$$

where u_A and u_D are vectors of the additive and dominance marker effects respectively and $\boldsymbol{Z}_A$ and $\boldsymbol{Z}_D$ the corresponding design matrices.

The coefficients can be estimated by solving the mixed model equations:

$$\begin{bmatrix} \hat{\mu} \\ \hat{u}_A \\ \hat{u}_D \end{bmatrix} = \begin{bmatrix} \mathbf{1_n^T 1_n} & \mathbf{1_n^T Z_A} & \mathbf{1_n^T Z_D} \\ \mathbf{Z_A^T 1_n} & \mathbf{Z_A^T Z_A} + \lambda_A \mathbf{I_m} & \mathbf{Z_A^T Z_D} \\ \mathbf{Z_D^T 1_n} & \mathbf{Z_D^T Z_A} & \mathbf{Z_D^T Z_D} + \lambda_D \mathbf{I_m} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n^T} y \\ \mathbf{Z_A^T} y \\ \mathbf{Z_D^T} y \end{bmatrix} \text{[eq. 7]}$$

where $\lambda_A = \frac{\sigma_e^2}{\sigma_{u_A}^2}$, $\lambda_D = \frac{\sigma_e^2}{\sigma_{u_D}^2}$

The variance components that define the shrinkage parameters can be estimated by REML analysis of the mixed model:

$$y = \mathbf{1_n} + \boldsymbol{Z}_F u_{GCAfemale} + \boldsymbol{Z}_M u_{GCAmale} + \boldsymbol{Z}_S u_{SCA} + e \quad \text{[eq. 8]}$$

(Zhao et al., 2014, Zhao et al., 2015). The two variance components of GCA effects can then be pooled if the male and female parents do not belong two separate heterotic pools. The variances of additive and dominance marker effects can be estimated by the variance of GCA and SCA effects respectively, as explained in section 1.2.2.2.2.

1.2.3. Prediction accuracy

Prediction accuracy is defined as the correlation between the genomic estimated breeding values and the true breeding values. In the case of simulation studies, this measure can be easily estimated. When real data are used, the phenotypic values are usually the predictands and the prediction accuracy is the correlation between the GEBVs and the phenotypic values divided by the square root of the heritability. In other cases, mainly in animal breeding, the predictand can be the EBVs (estimated breeding values) estimated by BLUP or the deregressed EBVs (Daetwyler et al., 2013).

1.2.4. Factors that affect prediction accuracy

1.2.4.1. Linkage Disequilibrium

Linkage disequilibrium (LD) is the association between alleles at different loci. The most common measure of LD intensity is the r^2 which is the correlation between alleles at the two loci.An r2 value of 0.1 is generally considered to indicate a significant LD.LD is caused by mutation, migration, selection and random drift, and it decays over generations due to recombination.

It is obvious that the prediction accuracy of genomic selection will depend on the LD between the genes and the markers.The level of LD varies between populations largely depending on the effective population size.

1.2.4.2. Number of markers

The minimum number of markers needed depends on the LD span (length of interval with significant LD).

For multiparental populations, the number of markers can be calculated based on the effective population size and the size of the genome(Desta and Ortiz, 2015).The number of markers used for genomic selection in biparental populations is generally smaller.

It is reported that prediction accuracy increases with the number of markers until a plateau is reached. For very large number of markers, the prediction accuracy can decrease as a consequence of the "large p - small n" problem.

1.2.4.3. Size of training population

The size of the training population is reported to be in many cases more important than increasing the number of markers.The optimal ratio of training to breeding population size, depends on the genetic distances, the heritability and the number of QTLs (Nakaya and Isobe, 2012). Furthermore, the smaller the training population size, the stronger the "large p - small n" problem.

1.2.4.4. Relatedness between training and breeding population

The relatedness between the training and the breeding population is shown to be one of the most important factors that affect the prediction accuracy. This issue is related to the decrease of prediction accuracy over selection cycles, which creates the need for updating the model.

In genomic selection of hybrids, where single-cross performance is predicted from the performance of related hybrids, relatedness between training and breeding population is also affected by the structure of the training population. Prediction accuracy is affected by whether none, only one or both parents of a hybrid in the breeding population have been used as parents for the training population.

1.2.4.5. Genetic architecture and statistical model

A high number of QTLs can decrease the prediction accuracy(Daetwyler et al., 2013), butthe efficiency of genomic selection relative to MAS or phenotypic selection should also be considered.

An important aspect is the agreement between the true architecture and the one assumed by the genetic model. If the number of genes is small and major genes are present, variable selection methods might perform better that RR-BLUP, as RR-BLUP assumes equal variance for all markers. Bayesian methods with rather uninformative priors can have a more stable performance across different trait architectures(Daetwyler et al., 2013, de los Campos et al., 2013a).

1.2.4.6. Heritability

Obviously, heritability has a positive correlation with prediction accuracy. The higher the heritability the more accurate the estimation of marker effects in the training population. In a breeding scheme, the potential of genomic selection relative to phenotypic selection is maximized when heritability is high in the training population and low in the breeding population.

1.3. Incorporation of QTL information

1.3.1. Concept

In most cases, the methods used for genomic selection treat markers in a way that ignores prior knowledge of QTLs and candidate genes, although such information is usually available. In recent literature, methods for incorporating prior information have been proposed and studied in simulated and real data.

1.3.2. Methods

1.3.2.1. RR-BLUP/FIXED

Bernardo (Bernardo, 2014) studied in simulation experiments the advantage of modelling effects of known major genes as fixed effects in a RR-BLUP model. The method was compared to RR-BLUP without fixed marker effects. The simulated trait architecture included 100 minor genes and 1, 2, 3 or 10 major genes jointly accounting for 50% of the

genetic variation in all 4 cases. Modelling a subset of the major genes as fixed was also studied.

He concluded that: i) when modelling one gene as fixed, the method is more advantageous for high h^2(heritability on an entry-mean basis) (> 0.5) and high R^2 (percentage of V_G explained by the gene), ii) relative efficiency is higher when the training population is smaller, iii) the rate of decline of prediction accuracy on the next cycles is lower than in ordinary RR-BLUP, iv) when many major genes (10) are modelled as fixed, model was less useful and v) there is difficulty in simultaneously estimating the fixed effects of more than a few major genes.

In his simulations he assumed perfect linkage between the known major genes and the corresponding markers. It would be interesting to expand his simulations in order to evaluate the benefit from using the same approach for mapped QTLs linked to uncharacterized genes.

Other studies confirmed these conclusions following the same approach in a G-BLUP context. Rutkoski et al. (Rutkoski et al., 2014) analyzed real data of quantitative stem rust resistance in 365 breeding lines of wheat. For the selection of markers with fixed effects, they ranked the markers based on p-values from a genome-wide association study of the same population and added markers stepwise to the model, calculating 5-fold cross validation accuracy within the training set.

In another study, the author also used real data (Fusarium resistance of inbred lines of maize) and conducted a GWAS including the three most significant SNPs as fixed effects in a G-BLUP model achieving higher prediction accuracy (Zila, 2014).

1.3.2.2. W-BLUP

Another study by Zhao et al.(Zhao et al., 2014) introduced a different method also based on RR-BLUP named W-BLUP. In contrast to the previous method, W-BLUP categorizes markers into two groups separating the functional markers from the rest, and shrinks both groups giving larger weight to the functional markers that are linked to known major genes.

The mixed model can be written as:

$$y = \mathbf{1_n} + \boldsymbol{Z}_A u_A + \boldsymbol{Z}_D u_D + \boldsymbol{Z}_{A_F} u_{A_F} + \boldsymbol{Z}_{D_F} u_{D_F} + e \qquad \text{[eq. 9]}$$

Best linear unbiased predictions can be obtained by solving the mixed model equations:

$$\begin{bmatrix} \hat{\mu} \\ \hat{u}_A \\ \hat{u}_{A_F} \\ \hat{u}_D \\ \hat{u}_{D_F} \end{bmatrix} = \begin{bmatrix} \mathbf{1_n^T 1_n} & \mathbf{1_n^T Z_A} & \mathbf{1_n^T Z_{A_F}} & \mathbf{1_n^T Z_D} & \mathbf{1_n^T Z_{D_F}} \\ \mathbf{Z_A^T 1_n} & \mathbf{Z_A^T Z_A} + \lambda_A \mathbf{I_{(m-f)}} & \mathbf{Z_A^T Z_{A_F}} & \mathbf{Z_A^T Z_D} & \mathbf{Z_A^T Z_{D_F}} \\ \mathbf{Z_{A_F}^T 1_n} & \mathbf{Z_{A_F}^T Z_A} & \mathbf{Z_{A_F}^T Z_{A_F}} + \lambda_{A_F} \mathbf{I_f} & \mathbf{Z_{A_F}^T Z_D} & \mathbf{Z_{A_F}^T Z_{D_F}} \\ \mathbf{Z_D^T 1_n} & \mathbf{Z_D^T Z_A} & \mathbf{Z_A^T Z_{A_F}} & \mathbf{Z_D^T Z_D} + \lambda_D \mathbf{I_{(m-f)}} & \mathbf{Z_A^T Z_{D_F}} \\ \mathbf{Z_{D_F}^T 1_n} & \mathbf{Z_{D_F}^T Z_A} & \mathbf{Z_{D_F}^T Z_{A_F}} & \mathbf{Z_{D_F}^T Z_D} & \mathbf{Z_{D_F}^T Z_{D_F}} + \lambda_{D_F} \mathbf{I_f} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{1_n^T} y \\ \mathbf{Z_A^T} y \\ \mathbf{Z_{A_F}^T} y \\ \mathbf{Z_D^T} y \\ \mathbf{Z_{D_F}^T} y \end{bmatrix}$$

[eq. 10]

where n the number of observations, m the number of markers and f the number of functional markers.

The proportion of the additive (or dominance variance) explained by the marker effects is estimated by multiple regression. These proportions are used in order to separate the variance components in the variance of the genome-wide markers effects and the variance of each functional marker effect according to the following procedure:

$$\lambda_A = \frac{\sigma_e^2}{\sigma_{u_A}^2} \ , \quad \lambda_{A_{F_f}} = \frac{\sigma_e^2}{\sigma_{u_{A_{F_f}}}^2} \ , \lambda_D = \frac{\sigma_e^2}{\sigma_{u_D}^2} \ , \ \lambda_{D_{F_f}} = \frac{\sigma_e^2}{\sigma_{u_{D_{F_f}}}^2}$$

$$\sigma_{u_{A_{F_f}}}^2 = P_{A_{F_f}} * \sigma_{GCA}^2 \ , \qquad \sigma_{u_{D_{F_f}}}^2 = P_{D_{F_f}} * \sigma_{SCA}^2 \ ,$$

$$P_{A_{F_f}} = \frac{SSA_{F_f}}{SSGCA} \ , \ P_{D_{F_f}} = \frac{SSD_{F_f}}{SSSCA}$$

$$SSGCA = SST * \frac{\sigma_{GCA}^2}{\sigma_P^2} \ , \quad SSSCA = SST * \frac{\sigma_{SCA}^2}{\sigma_P^2} \ ,$$

where SST relates to the same number of replications as used in the estimation of σ_P^2

$$\sigma_{GCA}^2 = \frac{(N_F - 1) * \sigma_{GCA_F}^2 + (N_M - 1) * \sigma_{GCA_M}^2}{N_F + N_M - 2}$$

$$\sigma_{u_A}^2 = \sum_1^f P_{A_{F_f}} * \sigma_{GCA}^2 \ , \quad \sigma_{u_D}^2 = \sum_1^f P_{D_{F_f}} * \sigma_{SCA}^2$$

The variance components for residual, GCA and SCA are estimated by REML as explained in section 1.2.2.4. (eq. 8).

Zhao et al. used phenotypic and genotypic data of a set of 135 wheat lines (15 male and 120 female) and 1604 (out of 1800 possible) hybrids. The training set included 10 male, 80 female and 610 hybrids derived from these 90 parents and the validation set the rest of the parents and hybrids derived from them. W-BLUP was able to increase the prediction accuracy for heading time and plant height. Furthermore it was shown by simulations that

the gain in prediction accuracy achieved by W-BLUP was much higher when the functional marker was in linkage equilibrium with all other markers.

1.3.2.3. Other methods

De los Campos et al. (de los Campos et al., 2013a) reviewing the methods used in GS, recognized the need to incorporate prior information and suggested the Bayesian context as appropriate for dealing with this challenge. He mentioned two Bayesian approaches: i) grouping the markers using different priors (Calus et al., 2010) and ii) using an antedependence model to allow borrowing of information across markers by specifying spatial correlation between marker effects(Yang and Tempelman, 2012).

Garrick et al. (Garrick et al., 2014) mention a method named BayesRS(Brondum et al., 2012)as more appropriate for including prior information. This approach is not used for introducing information from QTL mapping, but for using the results of a Bayesian model with locus specific variance as prior distribution for another population.

Su et al. (Su et al., 2014) tested G-BLUP models where the G matrix was weighted using data from prior Bayesian models or GWAS. In the case of GWAS the weighting factors were either the square of the estimated SNP effect or the negative ten logarithm of the P-value. The G-BLUP weighted with GWAS results was also used by de los Campos et al. (de los Campos et al., 2013b).

1.4. Aim of the project

For most of the important breeding traits, literature contains QTL information derived by QTL mapping studies. The potential gain from incorporating this information has not been studied so far and it can be affected by a wide series of factors.

Some of these factors are related to the accuracy of the QTL information, either due to limited power of the QTL mapping experiment or due to differences between the mapping and the breeding population in the present QTLs. Obviously, these factors can have a negative impact on the method, but the inclusion of an elimination procedure based on cross-validation can be used to avoid decrease in prediction accuracy.

Other factors influence in general the prediction accuracy of genomic selection methods. These factors could also influence the relative gain from using methods as RR-BLUP/FIXED and W-BLUP compared to standard RR-BLUP.

In this study I will attempt to obtain a general overview of these parameters and their importance. For this reason, marker and phenotype data will be simulated, to be analyzed with these three methods under a series of different scenarios.

2. Materials & Methods

2.1. General scheme

The procedure followed, consisted of: i) simulation of marker data under linkage disequilibrium, ii) simulation of genotype and phenotype for a set of 200 lines (100 male and 100 female) and all 10000 possible hybrids (in a factorial design), iii) simulation of QTL mapping between two contrasting lines, iv) application of genomic selection methods to estimate the regression coefficients in a training population consisting of a subset of the hybrids and v) estimation of prediction accuracy in a different subset of hybrids.

Several different scenarios were simulated with values of a series of parameters varying from one scenario to the other. A base scenario where the parameters were assigned medium values was compared to scenarios where one parameter value at a time was altered.

Table 1: Parameter values of the base scenario

Base scenario	
Training population size	510
Broad sense heritability	0.6
Trait Architecture	20 genes
LD (average r^2 between adjacent markers)	0.096

Table 2: Value of the varying parameter in the 16 other scenarios

16 other scenarios				
Training population size	100	300	700	900
Broad sense heritability	0.3	0.45	0.75	0.9
Trait Architecture	5 genes	10 genes	50 genes	100 genes
LD (average r^2 between adjacent markers)	0.051 (2000markers)	0.071 (1500markers)	0.113 (500markers)	0.123 (300markers)

2.2. Simulation of LD

500 haplotypes of a genome of 6 Morgan divided in 6 chromosomes were simulated. Each Morgan included 10000 monomorphic biallelic loci. Random mating withmutation rate

of $10^{-5}$was simulated for 1000 generations to generate variation in these loci, create a realistic LD profile and reach mutation-drift equilibrium. The number of cross-overs was drawn from a Poisson(5) distribution assuming no interference (Coster et al., 2010).During these 1000 generations the number of haplotypes was 500, implying an effective population size of 250 individuals. Then, 400 of the simulated haplotypes were used to form 200 individuals. Selfing of these individuals for 8 generations led to 200 lines. During selfing mutation rate was set to zero to decrease the number of loci with very low minor allele frequency as done in similar simulations(Coster et al., 2010).

The average r^2 between adjacent markers was used as the measure of linkage disequilibrium. For the estimation of r^2, the calculations were done using the markers that have a minor allele frequency above 0.05. The 200 lines were crossed in a factorial design to simulate the genotypes of 10000 hybrids. The simulation of this population was done using the package "HaploSim" for R. The next steps of the simulation were iterated 20 times.

2.3. Simulation of genotypic and phenotypic values

QTLs were assigned to 20 markers with MAF above 0.1. This level of MAF is used in several other simulations(Daetwyler et al., 2013). The values of the additive effects are presented in Appendix I. All QTLs were assigned positive dominance effects of 50% of the additive values. The QTL genotypes were used to calculate the total genotypic values of the lines and hybrids. A deviate sampled from a normal distribution was added to these values to estimate the phenotypic values. The mean of this normal distribution was zero and the standard deviation calculated in order to simulate specific value of the broad sense heritability coefficient.

2.4. QTL mapping

In general, QTL mapping was simulated in an approximate way in order to be automated and applied in every iteration.Two contrasting lines were selected on the basis of their genotypic values.An F3 population was simulated and the genotypic and phenotypic values were estimated as described above. Using the package "qtl" for R, a one qtl scan was performed using on average 200 markers. The average distance between adjacent markers was 3 cM and the maximum 9 cM. The result included the most significant peak position of every chromosome, with a minimum LOD score of three. The marker with MAF above 0.1

that was closer to each peak was selected to be treated like a functional marker. This marker was selected from all the available markers and the distance from the estimated peak of the QTL was on average 0.015cM. These markers are not functional since they do not constitute causative variation (Andersen and Lübberstedt, 2003)and will be referred to in this text as linked markers.

The result roughly represents a case where the set of QTLs that are known does not include all the most important QTLs that segregate in the training population. This is led by identifying only one QTL per chromosome.

2.5. Genomic selection

2.5.1. General

The training population consisted of a random subset of the 10000 hybrids. For genomic selection, 1000 markers were randomly chosen from the available markers that have MAF above 0.05. For every individual of the training population three phenotypic values were simulated assuming no genotype by environment interaction. The data were analyzed by RR-BLUP, RR-BLUP/FIXED and W-BLUP. In all steps described below, analysis was done on the level of the replication, not on genotype means.

2.5.2. RR-BLUP

RR-BLUP was applied as explained in section 1.2.2.4. In the design matrices (eq.6), the additive effects are coded as {aa,Aa,AA}={-1,0,1} and the dominance effects are coded as 0 or 1 for the homozygotes and heterozygotes respectively following the F infinity metric (Technow et al., 2012, Zhao et al., 2014).

Solving the MME (eq. 7), function “make.positive.definite” from Corpcor package for R was used when the system was computationally singular(Technow et al., 2012)

The variance components that define the shrinkage parameters were estimated by REML as described in the introduction using the R package “lme4”.

The two variance components of GCA effects were pooled as no heterotic pools are assumed or simulated. The pooled variance of GCA effects was divided by the number of markers to estimate the variance of the additive marker effects, assuming equal variance for

all markers (Technow et al., 2012, Zhao et al., 2014). In the same way, the variance of dominance marker effects was estimated by the variance of SCA effects.

2.5.3. RR-BLUP/FIXED

The same approach as before was used, but the effects of linkedmarkers were fitted as fixed. Model:

$$y = \boldsymbol{X}_{A_F}\beta_{A_F} + \boldsymbol{X}_{D_F}\beta_{D_F} + \boldsymbol{Z}_A u_A + \boldsymbol{Z}_D u_D + e \qquad \text{[eq. 11]}$$

The mixed model equations are the same as before, with the vector $\mathbf{1}_n$ replaced by the (n×2f) matrix$\mathbf{X}$which is the combination of $\boldsymbol{X}_{A_F}$ and $\boldsymbol{X}_{D_F}$.

$$\begin{bmatrix} \hat{\beta}_{A_F} \\ \hat{\beta}_{D_F} \\ \hat{u}_A \\ \hat{u}_D \end{bmatrix} = \begin{bmatrix} \mathbf{X^T X} & \mathbf{X^T Z_A} & \mathbf{X^T Z_D} \\ \mathbf{Z_A^T X} & \mathbf{Z_A^T Z_A} + \lambda_A \mathbf{I_m} & \mathbf{Z_A^T Z_D} \\ \mathbf{Z_D^T} \boldsymbol{X} & \mathbf{Z_D^T Z_A} & \mathbf{Z_D^T Z_D} + \lambda_D \mathbf{I_m} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X^T} y \\ \mathbf{Z_A^T} y \\ \mathbf{Z_D^T} y \end{bmatrix} \text{[eq. 12]}$$

In the mixed model analysis of the factorial design for the estimation of combining ability effects, the linked markers are also fitted as fixed.

$$y = \boldsymbol{X}_{A_F}\beta_{A_F} + \boldsymbol{X}_{D_F}\beta_{D_F} + \boldsymbol{Z}_F u_{GCAfemale} + \boldsymbol{Z}_M u_{GCAmale} + \boldsymbol{Z}_S u_{SCA} + e \qquad \text{[eq. 13]}$$

This procedure is slightly different from including linked markers as fixed factors in a mixed model of marker effects and estimating the variance components directly. The reason is that the variance of the linked markers is included in the GCA variance. However, it can be argued that when many markers are used, the variance of the linked markers is also captured by the variance of the genome-wide markers.

For the selection of the markers that will be finally fitted as fixed, a backward elimination procedure was followed, without ranking the candidate markers.The training population was divided in training and test set in a 5-fold cross-validation scheme, meaning that the model was build using 80% of the training population and validated in the remaining 20%. The mean prediction accuracy (Pearson correlation coefficient between the GEBVs and the phenotypes of the test set) over the five folds was used as the elimination criterion. When all candidate linked markers were dropped from the model, standard RR-BLUP was performed instead.

2.5.4. W-BLUP

W-BLUP was applied as described in the introduction. In the multiple regression of the linked marker effects, the additive effects were fitted first in the model, followed by the dominance effects. W-BLUP also included the same elimination procedure as RR-BLUP/FIXED.

2.6. Estimation of prediction accuracy

The breeding population consisted of a random set of 3000 hybrids that are not present in the training population. Prediction accuracy is estimated as the Pearson correlation coefficient between the GEBVs and the simulated genotypic values.

2.7. Other simulations

2.7.1. Relatedness between training and breeding population

In additional scenarios, the training population included only hybrids derived by a random sample of 50 male and 50 female parents. This way, the 10000 hybrids were divided in three sets. The set “H2” of 2500 hybrids with both parents used for the training population, the set “H1” of 5000 hybrids with only one parent used and the set “H0” of 2500 hybrids with none of the parents used. According to this division three breeding populations were defined consisting of 1500 hybrids each.

This approach was included in the base scenario, estimating three prediction accuracies for every method.The elimination procedure was based on the cross-validation prediction accuracy in the same way as in the first scenarios, where the test and training sets were complementary random subsets of the training population.

2.7.2. Number of linked markers and distance from the gene

In order to gain more insight in the performance of the methods, more sets of scenarios were formed. These scenarios did not include QTL mapping or elimination, but included the division of the hybrids in the three sets described above. The other parameters were as in the base scenario.

Eight scenarios were run, where the genes were assumed known and were used as functional markers to compare the three methods. Different number of the most important genes (1, 2, 3, 4, 5, 6, 10 or 20) was fitted to test the number of genes and extent of explained variance that maximizes the gain of incorporating this information.

In addition, to test the effect of not using the causative mutation as linked marker, six scenarios with varying distance between the linked marker and the gene were run.

Rare errors in the simulations occurred due tomarkers that were not polymorphic in the training set. In these cases, the corresponding iterations were replaced.

3. Results

3.1. Main Results

The results obtained from 20 iterations of the scenarios described in tables1 and 2 were used to graphically represent the effect of four parameters (training population size, heritability, trait architecture and linkage disequilibrium) on the prediction accuracy.

In general, high values of prediction accuracy were obtained in all scenarios and no further increase was achieved by incorporating QTL information.

The size of the training population had high impact on the prediction accuracy (Figure 1).

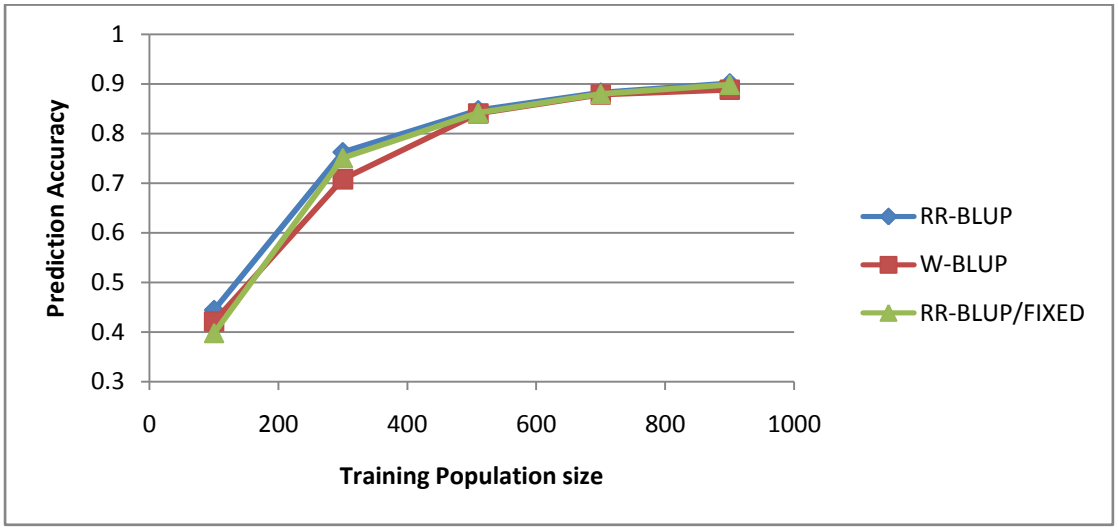

Figure 1: Effect of training population size on prediction accuracy

The effect of heritability is also apparent in Figure 2.

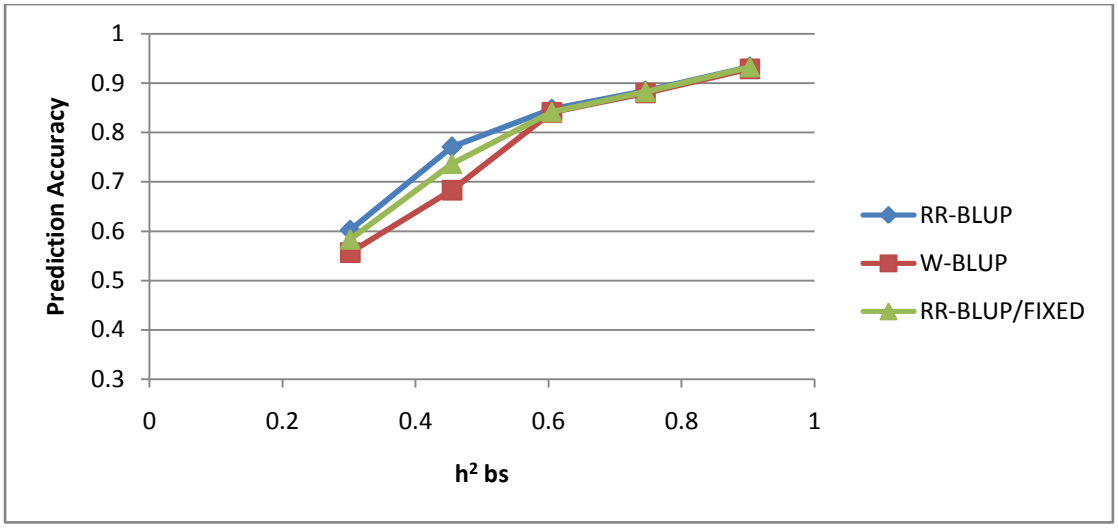

Figure 2: Effect of broad sense heritability coefficient on prediction accuracy

In contrast, prediction accuracy was independent of trait architecture (Figure 3) (Appendix I).

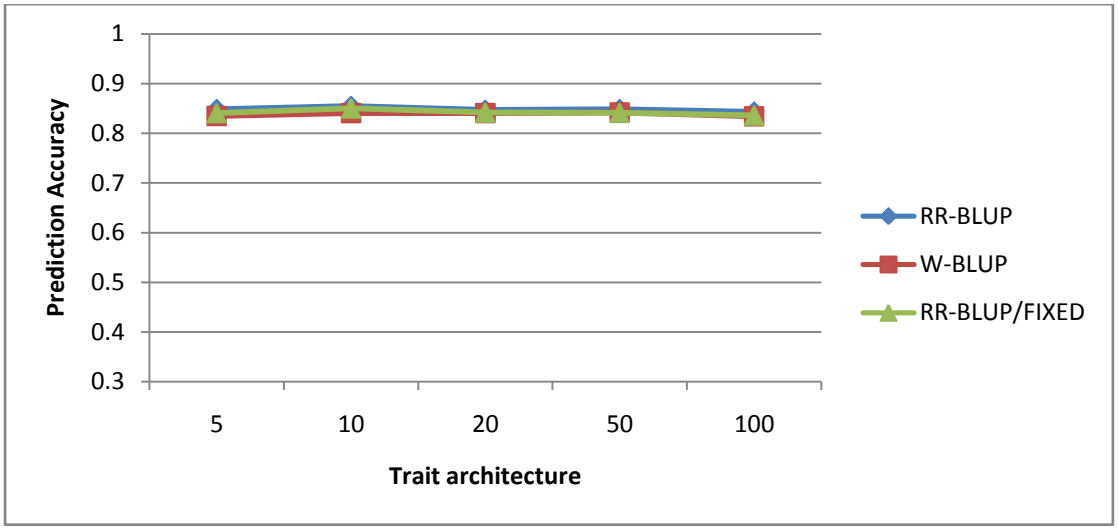

Figure 3: Effect of trait architecture on prediction accuracy

Effect of LD was observed, but prediction accuracy was very high even for low values of r2 between adjacent markers (Figure 4).

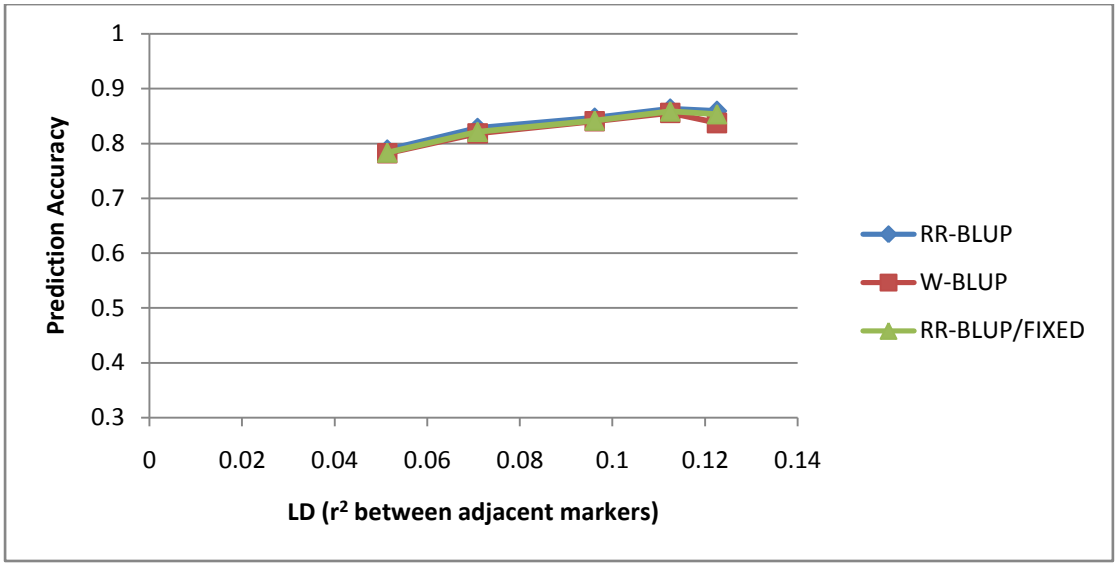

Figure 4: Effect of linkage disequilibrium (r2 between adjacent markers) on prediction accuracy

3.2. Additional Results

3.2.1. H2, H1 and H0 hybrid sets

In order to explain the high accuracy, the results of the alternative base scenario are presented in Figure 5, allowingcomparing the prediction accuracy of the three hybrid sets.

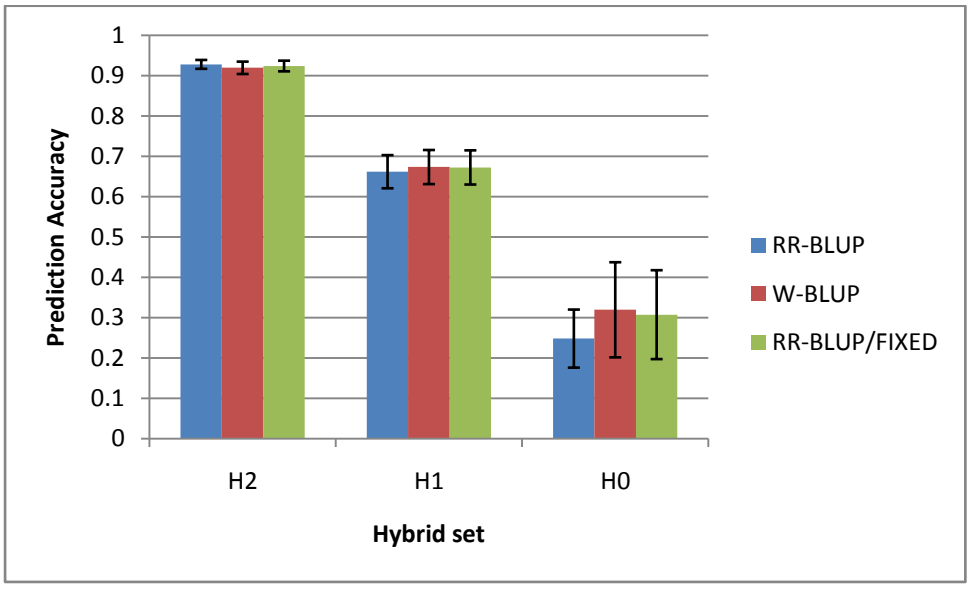

Figure 5: Prediction accuracy of the three methods for the three different sets of hybrids

3.2.2. Known genes

Assuming a varying number of genes known (always the ones with the highest simulated effect), the prediction accuracies observed for the three different hybrid sets are presented separately in the Figures 6 to 8 The prediction accuracy obtained by using a multiple regression model of the functional markers is also depicted in the figures.

In these scenarios W-BLUP run into an error, as the high variance explained by the functional markers led to the estimation of negative variance (and negative λs) for the genome-wide markers.

All methods performed better when applied to the hybrid set "H2".

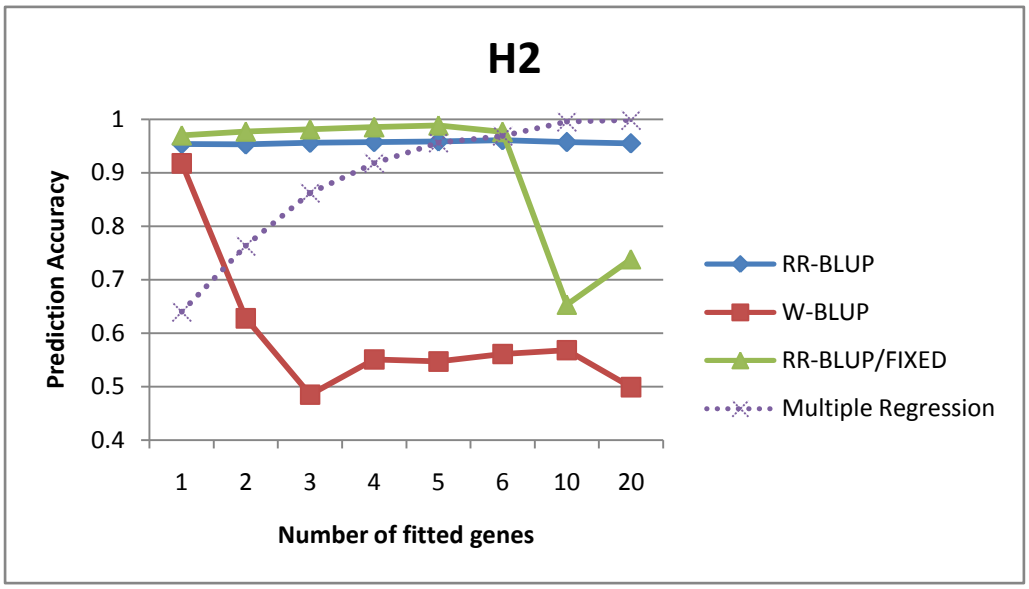

Figure 6: Effect of the number of fitted known genes on prediction accuracy for hybrids with both parents evaluated indirectly in the training population

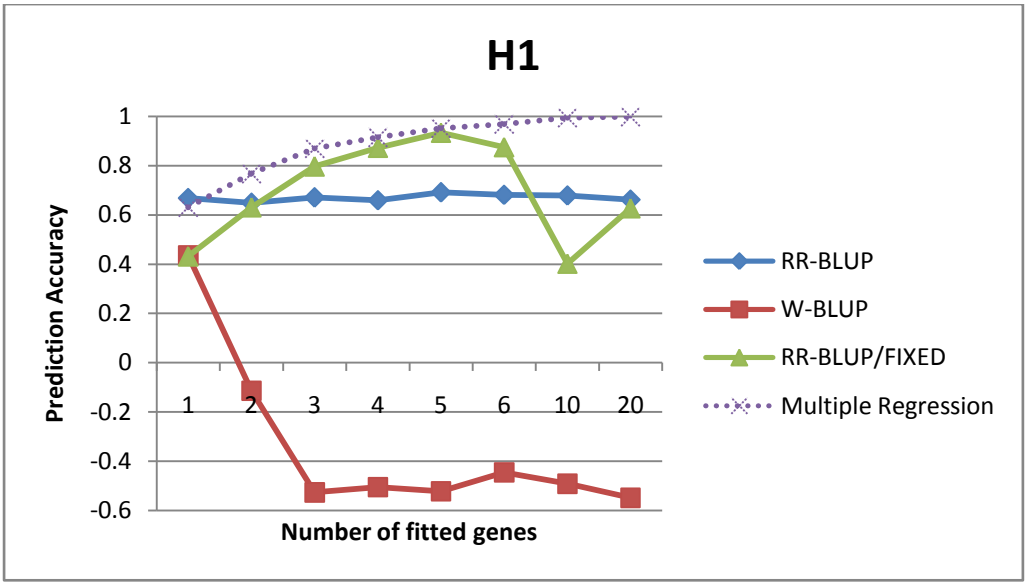

Figure 7: Effect of the number of fitted known genes on prediction accuracy for hybrids with only one parent evaluated indirectly in the training population

While the prediction accuracy of RR-BLUP was strongly affected by the relation between breeding and training population, the other methods performed almost the same when applied to hybrid set "H1" or "H0".

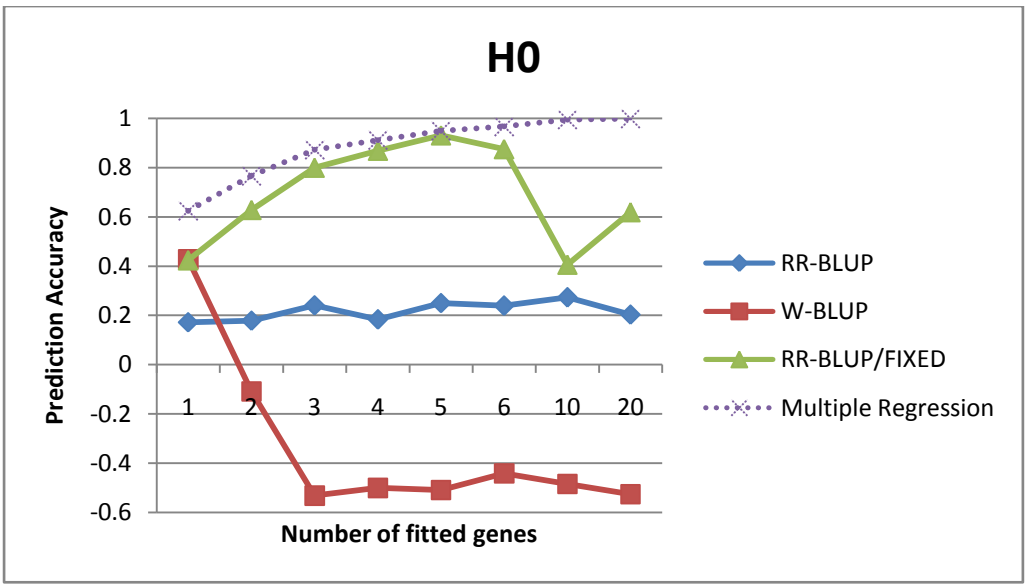

Figure 8: Effect of the number of fitted known genes on prediction accuracy for hybrids with none of the parents evaluated indirectly in the training population

3.2.3. Specified distance

In the next three graphs (Figures 9 to 11), the effect of the distance between the linked marker and the gene is presented for each hybrid set. When the linked marker is not

the gene itself, W-BLUP is not facing a problem and its performance is similar to RR-BLUP/FIXED.

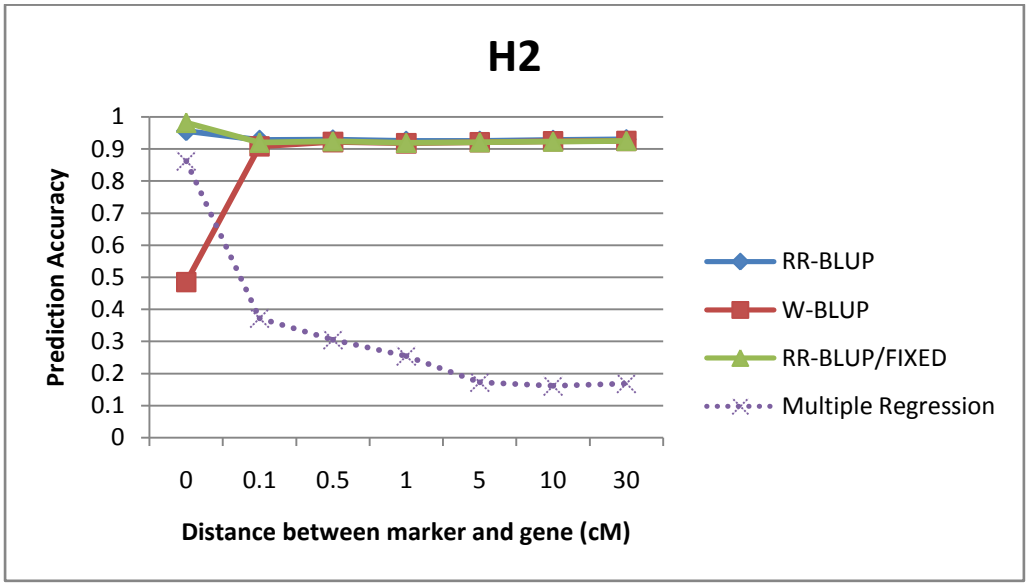

Figure 9: Effect of the distance between the linked marker and the gene on prediction accuracy for hybrids with both parents evaluated indirectly in the training population

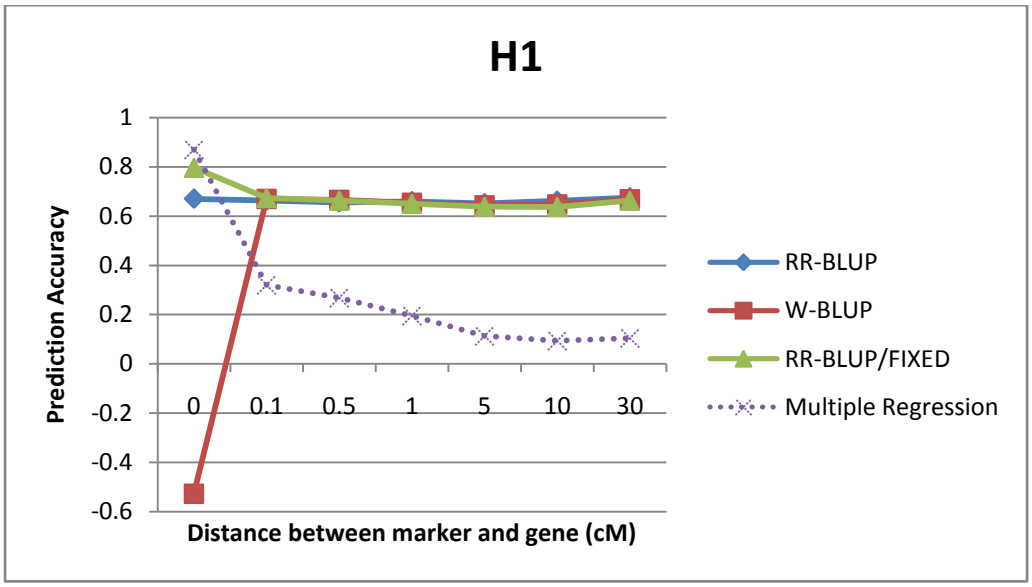

Figure 10: Effect of the distance between the linked marker and the gene on prediction accuracy for hybrids with only one parent evaluated indirectly in the training population

W-BLUP and RR-BLUP/FIXED can be beneficial for prediction in the "H0" hybrid set where the prediction accuracy of RR-BLUP is very low.

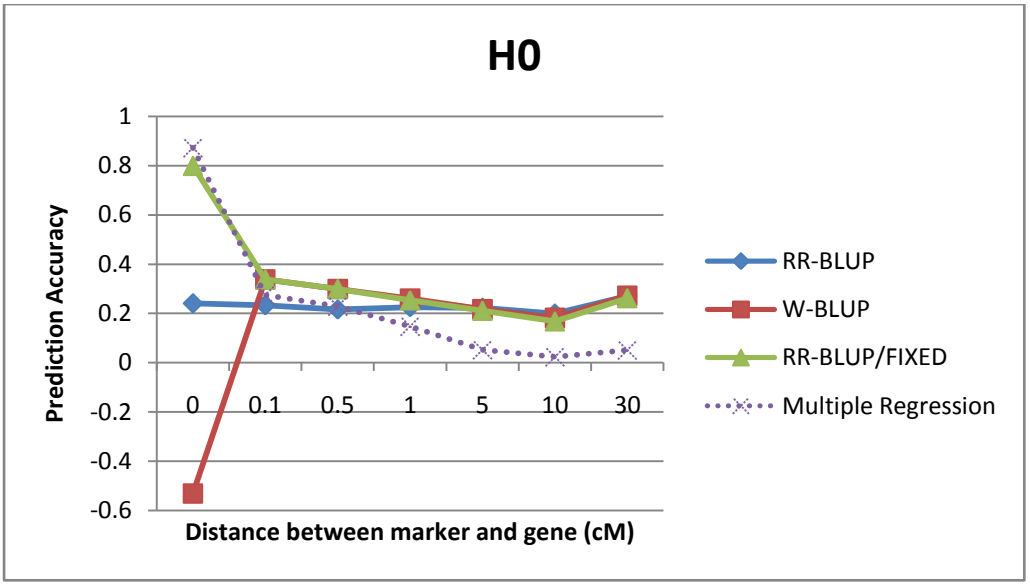

Figure 11: Effect of the distance between the linked marker and the gene on prediction accuracy for hybrids with none of the parents evaluated indirectly in the training population

4. Discussion

4.1. Parameters that affect prediction accuracy

Among the four parameters that were studied, the architecture of the trait did not have any effect on prediction accuracy (Figure 3).On the other hand, the size of the training population and heritability obviously affected prediction accuracy (Figures 1 and 2).

The effect of linkage disequilibrium was surprisingly small and high prediction accuracy could be achieved even for low levels of LD. This can be explained by the fact that both parents of all hybrids of the breeding population were used as parents of hybrids evaluated in the training population. In this case, the individuals of the breeding population are not unrelated to the ones in the training population and both their paternal and maternal haplotypes are evaluated in the training population in various background. Also, in a training population of 510 hybrids derived from a set of 100 male and 100 female lines, every parent is represented by approximately 5 hybrids. This results in prediction accuracies that are closer to the ones achieved when predicting the breeding values of the training population. Under these circumstances, the LD levels needed are much lower.

The same explanation can answer the generally high values achieved in all of the first scenarios.

When the training population was not random but only a subset of the parents was used, very high differences were observed between the three sets of hybrids. The prediction accuracy of the “H2” set was even higher thanin the first scenarios because fewer parents were used for the same size of training population resulting in the evaluation of 10 hybrids of each parent. The other sets of hybrids had significantly lower prediction accuracy. The difference between the different sets of hybrids is much higher thanin similar studies(Technow et al., 2012). This can be due to the lower LD and the smaller size of the training population.

4.2. Comparison of methods

In the first scenarios, no gain in selection accuracy could be achieved by incorporating QTL information. The explanation is obviously related with the fact that the breeding population consisted of hybrids with both parents evaluated indirectly in the

training population. The contribution of the linked markers in the training population was already captured by the standard RR-BLUP model.

It should be mentioned that the error of W-BLUP mentioned above that led to negative λs for the genome-wide markers, was also observed in 5 of the first scenarios. Specifically, when h2=0.3 this error occurred with frequency 0.3, for h2=0.45 with frequency 0.2, for training population size of 100 with frequency 0.25, for TP size=300 with frequency 0.1 and for r2=0.123 with frequency 0.05. This led to the lower average prediction accuracy presented in Figures 1, 2 and 4.

The elimination procedure was considered part of W-BLUP and RR-BLUP/FIXED. It can be concluded that in the studied scenarios elimination successfully prevented decrease in prediction accuracy, that could have been a possible result of including information of QTLs with low explained variance in the breeding population. This was apparent in the results when heritability was not below 0.6 and training population size not below 300. For low heritability or small training population size, the 5-fold cross-validation prediction accuracy was not an accurate estimator of the prediction accuracy in the breeding population.However, generalization should be avoided for cases with different relatedness between training and breeding population.

Comparing the three methods specific for each set of hybrids (Figure 5), it is apparent that incorporating QTL information can be advantageous when the prediction accuracy of RR-BLUP is low. Thisdifference between RR-BLUP and the other methods could be underestimated because the eliminationcan drop from the modellinked markers that cannot increasethe cross-validation prediction accuracy, even if these markers would be useful for prediction in the “H1” or “H0” hybrid sets. For example in this scenario that included QTL mapping, elimination and prediction specific for every hybrid set, in RR-BLUP/FIXED (W-BLUP) in 7 (2) out of 20 iterations no linked markers were included and in 9 (12) out of 20 only one was included. Furthermore, the standard deviations depicted in the Figure 5 are derived from only 20 iterations.

Fitting genes that are assumed known (Figures 6 to 8), the results for the “H2” set are in accordance to literature showing that fitting genes as fixed factors is beneficial, when the explained variance is high. Furthermore, RR-BLUP/fixed does not perform well when many genes are fitted(Bernardo, 2014).

Interestingly, in “H1” set, RR-BLUP/FIXED is also outperformed by RR-BLUP when only one gene is fitted. This is in contrast to prediction of set “H2”, probably because in that case the random genome-wide part of the model compensates for the low explained variance of the fitted gene. Also, in contrast to RR-BLUP,the methods that incorporate QTL information perform the same forthe hybrid sets “H1” and “H0”. These results stress the strong effect of predicting hybrids of “H2” type, implying that prediction of “H1” hybrids is closer to prediction of “H0” than of “H2” type, at least when LD is not high.

Varying the distance of the linked marker from the gene reveals that it can be advantageous to incorporate QTL information when the distance is less than 1cM and the predictive ability of the random genome-wide part of the model is very low (Figure 11).The importance of this finding should be further investigated, because if the genome-wide part is not capturing much information, phenotypic selection could be a more efficient method.

Including the prediction accuracy of a multiple regression model, shows how RR-BLUP/FIXED combine the random genome-wide part of the model and the fixed part. In the case of fitting known genes it seems that predicting only based on the known genes would be optimal. Of course such prediction would ignore the rest of the QTLs, with possible consequences on the decay of prediction accuracyor, in the case of hybrid selection, the accuracy of predicting the best hybrids.

Concerning W-BLUP, the problem of estimating negative λs occurs when the variance of the additive (or dominance) functional markers effects is higher than the GCA (or SCA) variance. A naïve solution would be to set the negative variances to ~0, leading to very high λfor the genome-wide markers. This was attempted (results are not presented), but the problem persisted when only the additive (or only the dominance) effect was artificially set to zero. Another solution could involve a different method for estimating the GCA and SCA variance. If these components are estimated by a fixed-effects model instead of a mixed model, the obtained estimates are slightly higher but the problem still remains for the additive variance. A possible explanation for this is that there is correlation between the additive and dominance effects, and fitting the additive effects first when the variance of the linked markers is estimated leads to overestimation of additive and underestimation of dominance effects. It should be also studied whether there is a different metric for coding the effects instead of the F infinity metric (that was used in this study and in the study that introduced W-BLUP) that could decrease this correlation. Alternatively, a more straightforward approach that could be equivalent of W-BLUP is to estimate the variance

components by REML analysis of a mixed model that includes the linked markers as random but not with the common variance of the genome-wide marker effects.

RR-BLUP/FIXED is clearly not able to include more than 6 effects. W-BLUP is worth improving because it may be able to accommodate more information. This is supported by the fact that when elimination was included in the scenario, W-BLUP was finally fitting more linked markers than RR-BLUP/FIXED, resulting in an average number of 1.53 markers for W-BLUP and 1.13 markers for RR-BLUP/FIXED.

To summarize the results concerning the incorporation of information coming from QTL mapping experiments, the scenario that includes QTL mapping and prediction in the three hybrid sets shows that it is possible to increase the prediction accuracy. The results from the scenarios with varying distance between the linked marker and the gene, imply that the increase in prediction accuracy depends on the accuracy of the QTL mapping experiment in estimating the location of the QTL. The distance of 1cM reported above as an approximate threshold, will depend on the LD span. It is not easy to predict the results for a population with higher LD. If LD span is longer, it will lead to higher explained variance of a linked marker and at the same time it will increase the prediction accuracy of standard genomic selection methods. Further research will be mandatory in order to answer this question.

4.3. Further research

First of all, the effect of LD, heritability, population size and trait architecture on the performance of these methods should be studied for all types of hybrids. Also, other types of breeding schemes should be used including the long-term response to selection that is reported to be influenced by fitting known genes as fixed effects(Bernardo, 2014).

RR-BLUP/FIXED can be applied in more straightforward procedure using more powerful REML software that can estimate the variance of effects directly.

W-BLUP should be improved to avoid the error that was observed, because possible superiority of W-BLUP over RR-BLUP/FIXED could be observed in case of including many linked markers that combined will have high explained variance.

These methods can be compared with other possible alternatives like the weighted G-BLUP (de los Campos et al., 2013b, Su et al., 2014).

It would be interesting to compare any method that can incorporate QTL information with methods that do not assume common variance for marker effects. Bayesian methods could have a better performance than RR-BLUP when major genes are present.

Finally, GWAS has been applied in identifying markers that could be specially treated in genomic selection. It would be interesting to combine results from GWAS on the training population and QTL mapping results reported in literature. This way, the conditions under which inclusion of QTL mapping results could be beneficial, can be identified.

5. Conclusions

The results of this study in combination with the brief available literature imply thatconditionally, prediction accuracy can increase by including available QTL information. These conditions can be generally defined as high variance of the linked marker effects and low predictive ability of the genome-wide markers.

It can be also concluded that RR-BLUP/FIXED is apotent method for introducing prior information. The verified limitations of this method when many markers are specially treated, creates the need for the development of alternatives. In order for W-BLUP to serve this role, the way of its implementation needs to be strengthened and clarified.

Further research may specify cases where the gain in prediction accuracy relative to standard methods is maximized. Furthermore, the methods of collecting and introducing prior information can be improved to combine GWAS, QTL mapping, characterized genes and previous GS models.

6. Literature

1. Andersen, J.R. and Lübberstedt, T. (2003) Functional markers in plants. *Trends in Plant Science* 8, 554-560.
2. Bernardo, R. (2014) Genomewide Selection when Major Genes Are Known. *Crop Sci.* 54, 68-75.
3. Bernardo, R. and Yu, J. (2007) Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Sci.* 47, 1082-1090.
4. Brondum, R.F., Su, G., Lund, M.S., Bowman, P.J., Goddard, M.E., and Hayes, B.J. (2012) Genome position specific priors for genomic prediction. *BMC genomics* 13, 543.
5. Calus, M.P.L., Mulder, H.A., and Veerkamp, R.F. (2010) Estimation of Breeding Values for Haploid Chromosomes. In *Proceedings of the 9th World Congress on Genetic Applied to Livestock Production (WCGALP), Leipzig, Germany, 1-6 August 2010*.
6. Coster, A., Bastiaansen, J.W., Calus, M.P., Maliepaard, C., and Bink, M.C. (2010) QTLMAS 2009: simulated dataset. *BMC proceedings* 4 Suppl 1, S3.
7. Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., de los Campos, G., and Hickey, J.M. (2013) Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics* 193, 347-365.
8. de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L. (2013a) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193, 327-345.
9. de los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., and Sorensen, D. (2013b) Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLoS Genet* 9, e1003608.
10. Desta, Z.A. and Ortiz, R. (2015) Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science* 19, 592-601.
11. Endelman, J.B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Gen.* 4, 250-255.
12. Garrick, D., Dekkers, J., and Fernando, R. (2014) The evolution of methodologies for genomic prediction. *Livestock Science* 166, 10-18.
13. Habier, D., Fernando, R., Kizilkaya, K., and Garrick, D. (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* 12, 1-12.
14. Heffner, E.L., Sorrells, M.E., and Jannink, J.-L. (2009) Genomic Selection for Crop Improvement. *Crop Sci.* 49, 1-12.
15. Henderson, C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423-447.
16. Jannink, J.L., Lorenz, A.J., and Iwata, H. (2010) Genomic selection in plant breeding: from theory to practice. *Briefings in functional genomics* 9, 166-177.
17. Meuwissen, T.H., Hayes, B.J., and Goddard, M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157, 1819-1829.
18. Möhring, J., Melchinger, A.E., and Piepho, H.P. (2011) REML-Based Diallel Analysis All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Sci.* 51, 470-478.
19. Nakaya, A. and Isobe, S.N. (2012) Will genomic selection be a practical method for plant breeding? *Annals of Botany* 110, 1303-1316.
20. Piepho, H.P. (2009) Ridge Regression and Extensions for Genomewide Selection in Maize All rights reserved. No part of this periodical may be reproduced or

21. Resende, M.D.V.d., Silva, F.F.e., Resende Júnior, M.F.R., and Azevedo, C.F. (2014) Genome-Wide Selection (GWS). In *Biotechnology and Plant Breeding* (Borem, A. and Fritsche-Neto, R., eds), pp. 105-133, Academic Press.
22. Rutkoski, J.E., Poland, J.A., Singh, R.P., Huerta-Espino, J., Bhavani, S., Barbier, H., . . . Sorrells, M.E. (2014) Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *Plant Gen.* 7, -.
23. Schrag, T.A., Frisch, M., Dhillon, B.S., and Melchinger, A.E. (2009) Marker-based prediction of hybrid performance in maize single-crosses involving doubled haploids. *Maydica* 54, 353-362.
24. Su, G., Christensen, O.F., Janss, L., and Lund, M.S. (2014) Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. *Journal of dairy science* 97, 6547-6559.
25. Technow, F., Riedelsheimer, C., Schrag, T.A., and Melchinger, A.E. (2012) Genomic prediction of hybrid performance in maize with models incorporating dominance and population specific marker effects. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik* 125, 1181-1194.
26. VanRaden, P.M. (2008) Efficient methods to compute genomic predictions. *Journal of dairy science* 91, 4414-4423.
27. Yang, W. and Tempelman, R.J. (2012) A Bayesian antedependence model for whole genome prediction. *Genetics* 190, 1491-1501.
28. Zhao, Y., Mette, M.F., Gowda, M., Longin, C.F.H., and Reif, J.C. (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity* 112, 638-645.
29. Zhao, Y., Mette, M.F., and Reif, J.C. (2015) Genomic selection in hybrid breeding. *Plant Breeding* 134, 1-10.

30.Zila, C. T., I.I.I. (2014). *Traditional and genomic methods for improving fusarium ear rot resistance in maize* (Order No. 3586257). . (1513601062). Retrieved from http://search.proquest.com/docview/1513601062?accountid=27871

7. Appendices

7.1. Appendix I: Size of simulated QTL effects

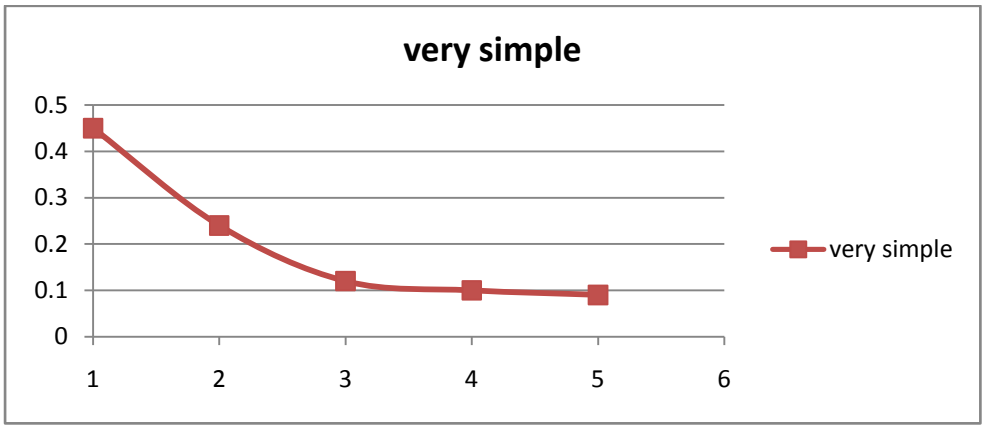

Figure12: Size of simulated QTL effects of a very simple trait regulated by 5 QTLs

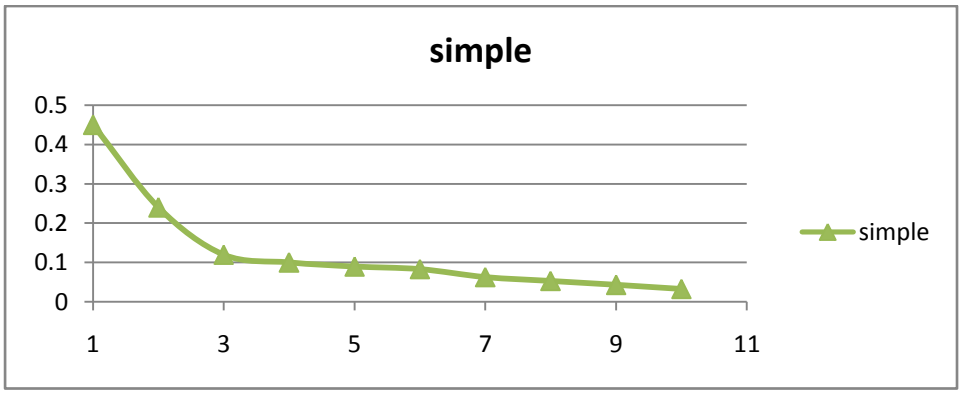

Figure13: Size of simulated QTL effects of a simple trait regulated by 10 QTLs

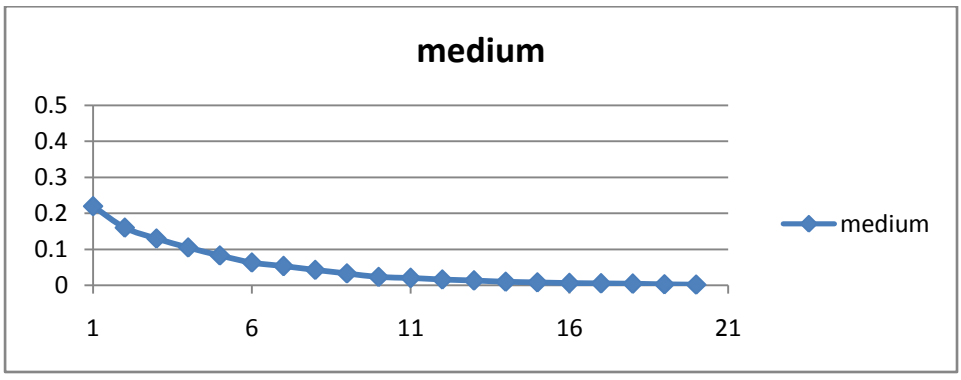

Figure14: Size of simulated QTL effects of a medium trait regulated by 20 QTLs

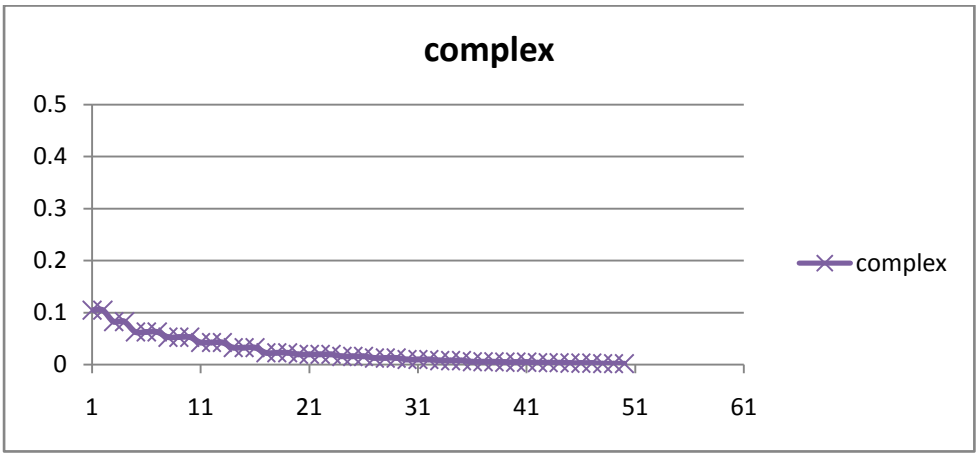

Figure15: Size of simulated QTL effects of a complex trait regulated by 50 QTLs

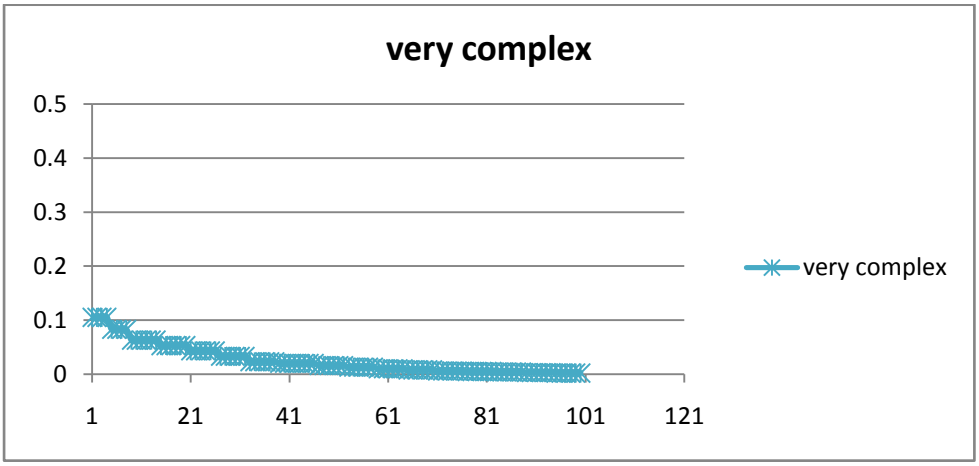

Figure26: Size of simulated QTL effects of a complex trait regulated by 100 QTLs