JANEO EUSTÁQUIO DE ALMEIDA FILHO

GENOMIC PREDICTION OF ADDITIVE AND NON-ADDITIVE EFFECTS IN A PINE BREEDING AND SIMULATED POPULATION

Thesis presented to the Universidade Federal de Viçosa as part of the requirements of Genetics and Breeding Graduate Program for the achievement of the title of *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS-BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
A447g
2016

Almeida Filho, Janeo Eustáquio de, 1987-
Genomic prediction of additive and non-additive effects in a pine breeding and simulated population / Janeo Eustáquio de Almeida Filho. – Viçosa, MG, 2016.
xiv, 107f. : il. (algumas color.) ; 29 cm.

Orientador: Fabyano Fonseca e Silva.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.98-107.

1. Genética vegetal. 2. Genômica. 3. *Pinus taeda* - Seleção. 4. Locos de caracteres quantitativos. 5. Matemática estatística. I. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 580.35

JANEO EUSTÁQUIO DE ALMEIDA FILHO

GENOMIC PREDICTION OF ADDITIVE AND NON-ADDITIVE EFFECTS IN A PINE BREEDING AND SIMULATED POPULATION

Thesis presented to the Universidade Federal de Viçosa as part of the requirements of Genetics and Breeding Graduate Program for the achievement of the title of *Doctor Scientiae*.

APPROVED: 17th February, 2016.

_____	_____
Camila Ferreira Azevedo	Cosme Damião Cruz
_____	_____
Matias Kirst (Co-adviser)	Messias Gonzaga Pereira

_____

Fabyano Fonseca e Silva
(Adviser)

“We are heirs of our own acts”
André Luiz

To God, source of love.
To my lovely Solange for her adorable company.
To my father Janeo (*In Memoriam*).
To my mother Narah, for her love.
To my brother, my great friend.
To my grandmother Ireny for her unconditional love.
To my grandparents.
To everybody who dream with something better.

ACKNOWLEDGEMENTS

First, I would like to thank God and Jesus for my life and for giving me strength and inspiration during my life, and for being my safe harbor and comfort in times of hardship.

I would like to thank my advisor Dr. Fabyano Fonseca e Silva for his friendship, guidance and concern about my success as a student.

I would like to thank my co-advisor Dr. Marcos Deon Vilela de Resende for all lessons and time spent in teaching me. I appreciate him for his dedication in giving me new knowledge and for being interested in propagating his knowledge.

I am very thankful to my co-advisor Dr. Matias Kirst for receiving me at the Forest Genomics Laboratory in University of Florida, also because he had been attentive with my progress and committed to the success of our research.

I would like to thank Dr. Márcio Resende Jr for his friendship, competent scientific help and all the support during my time at University of Florida.

I would like to thank Dr. Patricio Munoz for his valuable support, interest in our research and his excellent background on genomic selection in breeding.

I would like to thank Dr. Leonardo Lopes Bhering for his friendship and for receiving me at the Laboratory of Biometrics in the Federal University of Viçosa.

I would like to thank Dr. Cosme Damião Cruz for being an example of professor and always be available for the students.

I wish to thank Dr. Cosme D. Cruz, Dr. Camila F. Azevedo, Dr. Matias Kirst and Dr. Messias G. Pereira for their contribution and time in my thesis defense.

I would like to thank the Federal University of Viçosa and my graduate program Genética e Melhoramento for the opportunity to have a great education; I also wish to thank the University of Florida for the great structure and the opportunity to improve my research.

I am very thankful to Solange, for her lovely company, care, and valuable fellowship.

I am very thankful to my grandmother Ireny, because she always took care of me, and for her love in my personal formation.

I am very thankful to my mother Narah, for her love and concern about my personal formation.

I am very thankful to my brother Gabriel and my grandparents Ruth, Juarez (*In Memoriam*) and Pedro for all their support throughout my life; I would also like to thank my father (*In Memoriam*) and my whole family.

I want to thank my friends at the Laboratory of Biometrics: Leonardo Azevedo, Leonardo Correa, Lidiane, Andrei, Lizandra, Rafael, Humberto, Nadson, Bruno Ermelindo, Vinicius, Edson, Michele, Juan, Matoso, Afonso, Ândrea, and Ana Maria for the enjoyable company.

I want to thank my friends at UFV Caillet, Thais, Angélia, also I would like to thank the all friends that I met in Viçosa.

I wish to thank the people from Forest Genomics Laboratory mainly: Rodrigo, Justyna, Cintia, Flora, Annette and Chris Dervinis, for the nice time in the lab.

I would like to give a special thank for João Filipi for his help during my time at UFV and UF. Also, I wish to thank João, Luciano and Hélcio for receiving me in their home for a while.

I am very thankful to the secretaries of the graduate program Genética e Melhoramento at UFV, Edna, Rita, Odilon and Marco Túlio, for the competent support.

I am very thankful to Kelly and Márcio for assisting my personal needs in Gainesville. Also, I wish to thank Márcio, Kelly, Barbara, Leandro, Rodrigo, Ediene, João, Higino for the enjoyable moments in Gainesville.

I thank Dr. Rosana Vianello for her support at Embrapa Arroz e Feijão, I would also like to thank the colleagues from the Laboratory of Biotechnology at Embrapa Arroz e Feijão: Paula, Gabriel, Stella, Wendell, João, Lorraine... for their pleasant company.

I wish to thank Giselle Davi for her friendship and her valuable help.

I wish to thank Jeff's family for having received me at their home in Gainesville, and for the nice moments that I had there.

I would like to thank Ana Amélia e Josi all the support in Gainesville.

I am very thankful for GenMelhor, and the friends that I made there, for trying to improve our education quality at UFV.

I would like to thank Father Bill and his wife Kathy and everybody from covenant church of Gainesville, for the agreeable moments and attention with me.

I wish to thank the people from the English school in the Baptist Church for the lessons.

I want to thank my friends from Camilo Chaves: Geraldo, Aparecida, Angelina, Alexandre, Denizar, Maurício, Eduardo, João Bosco, Nalusa (*In Memoriam*), Suelly for making my Saturdays in Viçosa special.

Everybody who has motivated me and directly or indirectly contributed to this work, my sincere thanks.

Finally, I wish to thank every Brazilian who has worked hard and contributed with my education through their taxes.

CONTENTS

RESUMO

ALMEIDA FILHO, Janeo Eustáquio de, D.Sc., Universidade Federal de Viçosa. fevereiro de 2016. **Predição genômica de efeitos aditivos e não aditivos em uma população de melhoramento de pinus e em populações simuladas.** Orientador: Fabyano Fonseca e Silva. Coorientadores: Marcos Deon Vilela de Resende e Matias Kirst.

A predição do mérito genético dos indivíduos é um dos maiores desafios no melhoremento de plantas e animais. A predição é difícil por que as características importantes possuem natureza complexa, onde alguns caracteres possuem poucos genes de efeito maior, enquanto que outros são controlados por um elevado número de genes de efeito pequeno, além disso, efeitos não-additivos como dominância e epistasia podem ser importantes para o controle da variação genética. Para obter altas acurácias na predição é importante usar o modelo que corresponde com a arquitetura genética da característica e adicionalmente a adequada partição das várias fontes de variação genética (aditiva, dominancia e epistasia) é desejada para várias aplicações como capacidade geral e específica de combinação. No capítulo 1 foi revisado os aspectos gerais da predição genômica (GP), a aplicação dessa abordagem com diferentes propósitos em características com distintas arquiteturas genéticas e no final alguns modelos estatísticos aplicado na GP. No capítulo 2 foi avaliado modelos de regressão genômica (WGR) aditivos e aditivo-dominante com diferentes prioris, essas são premissas sobre a presença ou não de marcas com efeito maior. Adicionalmente no capítulo 3 foi avaliado a inclusão da informação oriunda do pedigree na predição genômica, usando os modelos BayesA aditivo e aditivo-dominante e também com o RKHS, que teoricamente pode predizer os efeitos aditivo e não aditivos confundidos. Esses modelos foram aplicados na altura de árvores (HT) aos 6 anos de idade, diâmetro na altura do peito (DBH) e resistência a ferrugem, mesurados em 923 indivíduos de pinos oriundos de uma população estruturada em 71 irmãos completos e genotipados com 4722 marcadores genéticos. Também foram simulados 6 características com distintas arquiteturas genéticas (poligenica e oligogênica com três leveis de dominância) para esses estudos. As populações simuladas usadas nessas características foram derivadas a partir de um programa de melhoramento padrão de pinos. No capítulo 2 para as

caracteríticas oligogenica simuladas e para resistência a ferrugem o BayesA e BayesB forneceram as melhores acurácias para predição genotípica, porem as diferentes priores usadas em WGR produziram resultados similares para HT e para característica poligênicas simuladas. Contudo a inclusão da dominância nos modelos WGR aumentaram a acurácia apenas para características simuladas com elevado efeito de dominância e para HT. Quando o BayesB foi ajustado em uma geração para predizer na geração seguinte, a inclusão da dominância aumentou as acurácias apenas para características oligogenicas simuladas com elevada dominancia. Independente do modelo adotado, a acurácia da predição genotípica total decresceu com o aumento dos efeitos de dominancia nas características simuladas. Então esses resultados refletem que a predição da dominancia foi complexa quando comparado com a predição dos efeitos aditivos, e para a aplicações posteriores dos efeitos de dominância, algumas propriedades genéticas da população devem ser avaliadas como MAF e número de meios irmãos e irmãos completos. No capítulo 3, a inclusão do informação oriunda do pedigree no modelo genômico, não produziu acurácias mais elevadas quando comparado com os modelos que usaram apenas informações de marcadores, e ambos modelos foram substancialmente mais acurados que o modelo baseado apenas em informação de pedigree. Em HT, DBH e características poligênicas simuladas com efeitos aditivos e dominantes, os modelos baseados em RKHS mostraram acurácias ligeiramente superiores que o BayesA para predição genotípica total, enquanto que o BayesA foi a melhor opção para resistência a ferrugem e características oligogenicas. Para a predição dos valores de melhoramento o BayesA aditivo foi o melhor modelo.

ABSTRACT

ALMEIDA FILHO, Janeo Eustáquio de, D.Sc., Universidade Federal de Viçosa. February, 2016. **Genomic prediction of additive and non-additive effects in a pine breeding and simulated population.** Adviser: Fabyano Fonseca e Silva. Co-advisers: Marcos Deon Vilela de Resende and Matias Kirst.

The prediction of individual genetic merit is one of most important challenges in plant and animal breeding. Prediction is difficult because the important traits have a complex nature, where some traits have few genes with major effects, while others are controlled by a large number of genes with small effects. Non-additive effects such as dominance and epistasis can also be important for controlling the genetic variation. In order to achieve higher accuracies in the prediction, it is important to use the model that matches the genetic architecture of trait. The proper partition of the various sources of genetic variation (additive, dominance and epistasis) is desired for several applications, such as exploring the overall and specific combination ability. In Chapter 1, the general remarks of genomic prediction (GP) are reviewed, with the application of this approach with different proposals in distinct genetic architecture traits, together with some statistic models applied in GP. In Chapter 2, the additive and additive-dominance whole-genomic-regression (WGR) models are evaluated with different priors, together with assumptions regarding the presence or not of markers with major effects. Chapter 3 evaluates the inclusion of pedigree information in genomic prediction with additive- and additive-dominance BayesA and also with RKHS model that can theoretically predict confused additive and non-additive effects. These models were applied in tree height (HT), diameter at breast height (DBH) and rust resistance in 923 loblolly pine individuals at 6 years of age from a structured population of 71 full-sib families genotyped with 4722 genetic markers. Six traits were also simulated with distinct genetic architectures (polygenic and oligogenic traits with three dominance levels) for these studies. The simulated population for these traits was derived from a standard pine breeding program. In the oligogenic simulated traits and rust resistance in chapter 2, BayesA and BayesB provided greater accuracies for genotypic prediction; however, the different priors of WGR yielded similar results for HT and simulated polygenic traits. Therefore, the inclusion of dominance effects in WGR

increases the accuracy only for simulated traits with high dominance effects and HT. When BayesB was fitted in one generation for predicting the next generation, the dominance inclusion increased the accuracies only for the oligogenic simulated trait with high dominance. Regardless of the model adopted, the accuracy of whole genotypic prediction decreased with the increase of dominance effects in simulated traits. Thus, these results reflect that dominance prediction is complex when compared to additive prediction, and for downstream applications of dominance effects, some genetic properties of the population should be evaluated, such as MAF and the number of half and full-sibs. In chapter 3, the inclusion of pedigree information in genomic model did not yield higher accuracies than models based in only marker information, and both models were substantially more accurate than models basedonly on pedigree. In HT, DBH and in polygenic traits simulated with additive-dominance effects, the RKHS-based models showed slightly higher accuracies than BayesA for whole genotypic prediction, while BayesA-based models were the best option for rust resistance and oligogenic simulated traits. For the prediction of breeding values, the BayesA additive was the best model.

GENERAL INTRODUCTION

The selection of superior individuals is one of greatest challenge in breeding programs. Traditionally this selection is based on genetic merit that can be estimated from pedigree information. However, with the dense genetic markers currently available, the genomic prediction (GP) approach proposed by Meuwissen et al. (2001) has received much attention in the breeding of plants (Bernardo 2008; Heffner et al. 2009; Resende Jr et al. 2012a; Resende et al. 2012) and animals (Goddard and Hayes 2009; Hayes et al. 2009; Wiggans et al. 2011), with the possibility of earlier selection, as well as more accurate genetic predictions. This approach has also been applied in human science (Yang et al. 2010; de los Campos et al. 2010a; Wray et al. 2013) for clinical outcomes and/or response to drug treatments.

The prediction models traditionally applied in breeding were based on pedigree, but different from GP that provides genetic markers, the pedigree base-line model considers the expected relationship of individuals that cannot follow Mendelian segregation (de Los Campos et al. 2009). In addition, the models based on pedigree information consider infinitesimal assumptions, whereas the number of genes that control the trait tends to infinite and explains the same portion of genetic variance. However, GP models can directly infer genome variations, which allows for the computing of the locus with major effects (Meuwissen et al. 2001; Gianola 2013; de los Campos et al. 2013).

The prediction of locus effect can be achieved with Whole-Genome Regressions (WGR), there are several Bayesian WGRs, and since letters are used for their differentiation (e.g. BayesA, BayesB, BayesCπ …), such variations are usually known as Bayesian alphabet (Gianola et al. 2009; Gianola 2013). These approaches share the same model but differ regarding prior distribution assumed for marker effects (Gianola 2013; de los Campos et al. 2013). This allows WGRs to be much more flexible in explaining quantitative traits. For instance, the Bayesian Ridge Regression (BRR) assumes that all covariates (markers) have common variance (Pérez and de los Campos 2014) and consequently, markers with the same allele frequency express the same genetic variance portion (Gianola et al. 2009). This assumption matches the infinitesimal model and consequently is desired for polygenic traits. There are other more parametrized priors of WGR, these priors were formulated to better explain the traits with major-effect

genes that explain considerable part of the genetic variation (Meuwissen et al. 2001). Some of these models assume a heterogeneity variance component for marker effects (e.g. Bayes A, BayesB, Bayesian Lasso) while others induce covariate selection (e.g. BayesB and BayesCπ) to markers that are not in linkage disequilibrium with any gene. For oligogenic traits, the correct choice of prior implemented in additive WGR has been considered important to achieve higher accuracies with real and simulated data (Coster et al. 2010; Daetwyler et al. 2010; Clark et al. 2011; Habier et al. 2011; Resende Jr et al. 2012b).

Initially, the GP models accounted only for additive effects (Meuwissen et al. 2001). These models are useful in breeding systems that explore mainly additive effects where the selected individuals are allocated in matting with several other individuals and specific crosses are not explored. For instance, in recurrent intra-population selection, or in cattle breeding where bulls provide semen for a large number of cows, the alleles a given individual carry are more important than the individual itself. However, in large numbers of plant and animal species, breeding presents benefits in the exploitation of hybrid vigor (heterosis) and the inclusion of dominance in prediction models can improve the accuracy of genomic hybrid predictions (Zeng et al. 2013), allowing an extra genetic gain with the best mate-pair allocation (Toro and Varona 2010). Regarding the inclusion of dominance in GP, one should also consider that the genetic architecture. For deal with the different genetic architecture, the different priors formulated primarily for additive models can be extended for additive-dominant models. Therefore, the contribution of these priors to accuracy in models with dominance inclusion and for traits with different genetic architecture has not been extensively explored.

Despite the benefits of dominance prediction, consideration should be given to the statistical challenge. In traditional additive GP models, it is already an issue, since there are much more parameters than observations, and these issues are even greater with the inclusion of non-additive effects (e.g. dominance and epistasis), since the parameters to be estimated increase substantially (Gianola 2013). One option to avoid estimating large numbers of parameters would be to use direct prediction of individuals instead of predictions by marker effects. The direct individual predictions can be achieved with models based on the relationship measured from markers, such as GBLUP (VanRaden

2008) and RKHS. In GBLUP, non-additive effects can also be included (Su et al. 2012; Resende Jr et al. 2012b; Vitezica et al. 2013; Morota et al. 2014), but under the assumptions of no linkage, linkage equilibrium (Morota et al. 2014) and Hardy-Weinberg equilibrium, similar to base-line pedigree models. The RKHS predicts the whole-genotypic value of individuals (Gianola et al. 2006; Gianola and van Kaam 2008), but in a non-explicit manner (Morota et al. 2014), which can be explored for clonal selection. The computational advantage of direct individual prediction is higher if there is an increase in the number of markers. However, unlike GBLUP- and RKHS-based models, considers that loci with the same allele frequency explain the same portion of genetic variance, which is undesired for genes with major effects. Therefore, the performance of these models should be compared with WGR-based models under traits with distinct genetic architecture.

Ideally, in GP, the genome is widely covered by genetic markers and all genes are in LD with at least one marker. However, in practical applications, important genes may not be completely explained by any of the markers, thus the inclusion of polygenic effects is recommended for capturing any genetic variance not associated with the genetic markers, and to impose some selection pressure on low-frequency QTL that may not be captured by the markers (Hayes et al. 2009).

Given the stated above, the first chapter of this work is a review about general remarks of genomic prediction, the application of this prediction approach with distinct proposes, in different genetic architecture traits, and finally the statistic models applied in GP in face of these demands. The second and third chapter are an investigation of genomic prediction of traits with distinct genetic architecture using real traits measured in a standard breeding population of loblolly pine, using simulated data with similar proprieties from the real population. In the second chapter, different genetic-statistics assumptions of WGR using additive- and additive-dominance effects are evaluated. In the third chapter, RKHS based models, traditional BayesA, additive-dominance BayesA and the pedigree inclusion in genomic prediction models are evaluated.

CHAPTER I
LITERATURE REVIEW

GENOMIC SELECTION

Genomic prediction (GP) was proposed by Meuwissen et al. (2001), this approach aims to predict the genetic merit of individuals in a population using information from genetic markers. Most of the GP models currently available may be used to predict the clinical outcome of diseases (Wray et al. 2013), and also to predict other medical science approaches (Yang et al. 2010; de los Campos et al. 2010a). In the breeding program, GP was proposed to select the best individuals using their genomic predicted values.

The use of these predicted values on the selection decision is called Genomic Selection (GS) or Genomic Wide Selection (GWS) (Goddard and Hayes 2009; Hayes et al. 2009). The term “wide” is related to the covering of the entire genome in GWS (Resende et al. 2014), since ideally, GP uses dense SNP panel and all genes that affect the given trait are in linkage disequilibrium (LD) with at least one marker. Consequently, the all-genes effect would be captured by markers in the prediction model.

In the GS seminal paper, Meuwissen et al. (2001) used a reference population where all individuals were marker genotyped and recorded for the trait. From this population, the markers effects were estimated in order to predict the Genomic Expected Breeding Values (GEBV) in subsequent populations, where the individuals were genotyped but not phenotyped. These authors showed high accuracies for breeding value prediction, using only maker information in the following five generations after the reference population.

The results of Meuwissen et al. (2001) changed the paradigm of breeding, because with the GS approach, it is not necessary to have a special mating design, which allows the application in any plant or animal breeding program as a tool to select the best individuals using marker information. After the marker effects estimated in the reference population, the selection can be achieved with only marker effects in some subsequent generations, thus opening the possibility of early selection and substantially increasing the genetic gain per time unit.

Early selection has a huge potential for reducing the cost of breeding program, especially for traits that are expensive to record, which demand a long time to evaluation (Schaeffer 2006; Resende Jr et al. 2012a; Resende et al. 2012) or that requires the destruction of the individual to collect such phenotypes (Fritsche-Neto et al. 2012). Schaeffer (2006) reported a potentially huge saving in dairy breeding program with the possibility of early selection of sires, instead of recording the phenotypes of the female offspring of these sires, and after estimate the sire breeding values. Other important example for reducing time occurs with forest breeding, where most of the important traits are collected after some years in the field, such as wood quality traits. Thus, GS can considerably reduce the time for the breeding program, mainly when associated to techniques for reducing the flowering time (Resende et al. 2012; Resende Jr et al. 2012b).

According to Resende et al. (2014), after the GS proposed in 2001, this approach was discrete for some years, because the number of genetic markers available and the cost to genotype large numbers of individuals at that time were not feasible. However, with the increase in the number of markers available and the reduced cost to genotype, GS became feasible. After the proposal from the seminal paper emphasizing the GS application in animal breeding programs by Schaeffer (2006), Bernardo and Yu (2007) discussed the incorporation in crop breeding and Grattapaglia and Resende (2011) in forest breeding.

With real data, the results showed that GS is in fact feasible in animal breeding (Hayes et al. 2009) and plant breeding (Crossa et al. 2010). The pioneering applications in forest breeding with real data were (Resende Jr et al. 2012a; Resende Jr et al. 2012b) in pine and (Resende et al. 2012) in eucalyptus. These works reported favorable results for the incorporation of GS in breeding programs.

The GS general application requires three populations: reference, validation and selection. The reference or training population is a population where genetic markers and phenotypes are available. This population is used to calibrate the predictive model. The validation population also presents available phenotypes and marker information, and this population is used to validate the model. The individuals from the validation population are not used for calibrating the model. However, the model calibrated with the reference population is used to predict the GEBV of individuals in this population, and since the

phenotypic information is available, it is possible to infer if the predictive model was accurate enough. The last population is the selection population, if the predictive model provided high accuracy, the model will be used in this population in order to practice for selecting individuals with only marker information (Goddard and Hayes 2009; Resende et al. 2014). These three populations should be related, because empirical results showed that models calibrated for one population are not accurate for use in disconnected populations (Resende et al. 2012). According to Resende et al. (2014), these three population can be the same, and/or one population can be used for more than one function (e.g. use the same population for estimating the model and validation), which has been showed in practical applications (Crossa et al. 2010; Resende Jr et al. 2012b).

To date, GS is definitively incorporate in breeding programs (Wiggans et al. 2011; Crossa et al. 2014); however, some questions should be answered for each case such as the number of generations necessary to re-calibrate the model, since each re-calibration is costly and time demanding. Grattapaglia (2014) proposed a forest intra-population breeding scheme, where the reference population had at least 1,000 individuals with effective size (Ne) between 30 and 100 (Grattapaglia and Resende 2011). This suggestion involves early selection using only marker information and a generation for updated the model. The number of generation without updating the model is not specified, however, even in a conservative scenario where the predictive model would be used only for one generation without re-calibration, the breeding cycle would be substantially reduced in forest breeding. It is expected that the number of GS generations without re-calibrating the model depends on the breeding strategy adopted and all factors that affect accuracy (heritability, Ne, Linkage Disequilibrium (LD), length of genome (L), models used, and number of informative markers available). The accuracy issue will be commented later in this review.

The dense markers panel is important to achieve higher accuracies (Grattapaglia and Resende 2011); however, with high number of marker effects to be estimate in statistics viewpoint, it becomes a challenge, since the number of parameters to be estimated would be considerable higher than the number of phenotypes available. Other difficulty is that many markers are in LD, which create higher multicollinearity. One way to deal with this would be to further learn about genetic architecture and use only important

markers instead of all markers. However, in order to achieve such learning, it is necessary to employ an appropriate approach, since GS does not provide trait dissection and gene discovery. In GS, the focus is the genetic value prediction without any inference regarding marker association and, thus, it is anecdotically criticized and called of “Black Box” (Grattapaglia 2014).

Nonetheless, different models to deals with the different genetic architecture are available. These models use different assumptions about the distribution of marker effects, with some models assuming there are markers with major effects, which match with oligogenic traits, and other models assuming that all markers have similar effects, which is more like the infinitesimal assumptions (Gianola 2013; de los Campos et al. 2013). It is also possible to include non-additive effects such as dominance (Toro and Varona 2010; Zeng et al. 2013; Vitezica et al. 2013; Nishio and Satoh 2014) and epistasis (Wittenburg et al. 2011; Su et al. 2012; Muñoz et al. 2014). Finally, some models propose the genomic prediction considering additive and non-additive effects, but in a non-explicit way, such as Reproducing Kernel Hilbert Space (RKHS) (Gianola et al. 2006; Gianola and van Kaam 2008). The different genetic architecture and the models used to deal with each case will be commented later in this paper.

PREDICTION WITH MARKERS AND PEDIGREE

Initially, the traditional genetic evaluations were based only on pedigree information. With special designs (e.g. North Caroline genetic designs), it is possible to estimate the genetic variance components to support breeding strategies (Pereira and Amaral Jr 2001). In addition, with more general population structure, regardless of mating design, but with pedigree information available, the mixed model methodology (Henderson 1984) can be applied to estimate parameters such as breeding values, dominance deviation values, and also genetic variance components.

Unlike pedigree-based models, the genetic markers allow for the follow-up of Mendelian segregation; a term that accounts for 50% genetic variability in additive models and in the absence of inbreeding (de los Campos et al. 2010b). For example, if two full-sib individuals do not have phenotypes, in pedigree-based models their predicted breeding values are calculated as the average of the breeding values from their parents,

but if genetic markers for these full-sibs are available, it would be possible to discriminate then from the genetic merit. The GS overcame the traditional pedigree models in many cases for plants (Crossa et al. 2010; Crossa et al. 2013; Crossa et al. 2014), animal predictions (Hayes et al. 2009; de Los Campos et al. 2009) and also in studies with simulations (Calus and Veerkamp 2007).

Other advantage of GS over traditional pedigree-based models is that GS is more flexible regarding the distribution of gene effects. The prediction based on pedigree considers the infinitesimal model, that is, it considers that a trait is controlled by the number of genes that tend to the infinite, and all genes have equal effect. This assumption is desired in polygenic traits, which are traits controlled by large numbers of genes with small effects. However, in oligogenic traits where some genes have major effects, a prediction model that considers some markers as having higher effects that can be active with some GS models is preferred. This topic will be discussed later.

Therefore, as mentioned before, in the best scenario all genes controlling the quantitative trait are in strong LD with at least one marker. However, it might be that some expressed genes are not in LD with any marker. In order to capture the genetic variation uncovered by markers, it is recommend to combine marker and pedigree information in GS model with what is known as polygenic or infinitesimal effect (Calus and Veerkamp 2007; Heffner et al. 2009; Hayes et al. 2009). The polygenic (or infinitesimal) effect is basically an effect associated for each individual, using the relationship information from pedigree, whereas the pedigree information can be incorporated in the model by traditional A matrix, that is twice the kinship relationship coefficient.

Using low density SNP panels, the inclusion of polygenic effects in GP models provided better predictions in wheat (Crossa et al. 2010), and dairy cattle (Vazquez et al. 2010). However, when SNP density was increased, the polygenic inclusion did not improve predictions in the same study with dairy cattle, and in other work with mice (de los Campos et al. 2009).

In maize studies using large numbers of SNP markers from Genotyping-by-Sequencing (GBS), the models including polygenic effects were the best option (Crossa et al. 2013). These authors also showed that in some cases, the models with only pedigree information outperformed traditional GP models with no pedigree information. A

reasonable explanation is that the GBS technique yielded a large number of missing data, thus some important genes may not have had any informative marker in LD.

In simulation studies, (Calus and Veerkamp 2007) reported that the inclusion of polygenic effects improved the prediction accuracy only in scenarios with low LD, and when the LD was increased, the models with and without polygenic effects showed similar accuracies for GEBV. These authors also reported that the variance components estimated from models with the inclusion of polygenic effects were less biased regardless of LD.

Another way to combine pedigree and marker was proposed by Misztal et al. (2009), but in this case, the combination is used to predict breeding values for all individuals, both genotyped and non-genotyped. This method considers the genomic relationship among individuals with genetic markers, and traditional relationship from pedigree information if one individual does not present any marker information. One modification to this method was proposed by Meuwissen et al. (2011). These methods will not be viewed in details here.

POLYGENIC AND OLIGOGENIC TRAITS

If the number of genes controlling a trait is really small (e.g. four or less), one option would be to develop a pyramid of favorable alleles of these genes in an individual (Bernardo 2008). The discovery of individuals with favorable alleles could be performed using the classical way, evaluating genetic segregation in the progeny, and the incorporation of the favorable alleles in a potential individual could be made with mating and selection. In this case, the discovery of markers associated with favorable alleles could potentially improve the breeding strategy with marker assisted selection (MAS). The association between markers and important genes can be achieve inside the family in a structured population from QTL mapping approach, or in the entire population with the preference for individuals unrelated from Genomic Wide Association Studies (GWAS).

However, if this gene number becomes large, the development of the pyramid of favorable alleles in a single individual is not simple, and the techniques to perform the association have some statistic challenges. In QTL and GWAS mapping, the association of a gene with a marker is determined using a hypothesis test, which presents issues with

type 2 error, mainly for genes with minor effects. This means that it is not rejected a null hypothesis (null association) that should be rejected, and for that reason, many genes remain uncovered, and only a small portion of the genetic variation is explained by markers, which is undesired for MAS. In GS, all makers are considered in the prediction (a hypothesis test is not used), then many makers with small effects that would have been rejected in the hypothesis test are analyzed together, which could explain a large portion of genetic variation, are also considered in GS (Resende et al. 2014).

In literature, it is reported that the QTL explained by QTL mapping approach is usually inconsistent (Bernardo 2008), and usually explains only a small portion of genetic variation (Dekkers 2004; Bernardo 2008). In this case, one suggestion is to improve the frequency of favorable alleles by recurrent selection (Bernardo 2008), which can be performed with GS in recurrent cycles.

In order to better perform GS, one important issue is how large is the number of genes and/or how is the gene effect distribution in the quantitative trait. In traits controlled by large numbers of genes with small effects, the infinitesimal assumption is a good approximation, and GS models that tend to assume similar contributions among markers to predictions is theoretically desired.

However, in some traits, the gene number is not necessary large enough to be considered a classical polygenic trait (e.g., 30 genes), or there are few genes with major effects. Thus, the infinitesimal assumption does not really match the nature of these traits. In order to better explain the quantitative trait regarding the number of genes (or gene effects), many models based on different marker effect distributions are available (Gianola 2013; de los Campos et al. 2013), mainly from GS models that input the markers as covariates in Bayesian regression models, as described further below.

Simulated studies have shown that the performance of GS model is dependent of genetic architecture (Coster et al. 2010; Daetwyler et al. 2010; Clark et al. 2011), and it is possible to achieve higher accuracy if choosing the model that better matches the genetic architecture of the target trait. According to de los Campos et al. (2013), the correct choice of prior marker distribution in regression models has been considered important, mainly for oligogenic traits in simulated studies, and the advantages of these marker effect distributions are not always evident in real data.

However, studies with rust resistance in pine (Resende Jr et al. 2012b) and fat in milk (Habier et al. 2011) have showed that the choice of prior distribution of marker effects with the assumptions that some markers have higher effects provide higher prediction accuracies. In polygenic traits, some results showed that the different marker distributions assumed in GS models showed similar results (Resende Jr et al. 2012b; Pérez et al. 2012).

NON-ADDITIVE EFFECTS

In quantitative genetic theory, the additive effect of a gene (a) is defined as the difference in genotypic value of homozygote with favorable alleles and the mid-point between two homozygotes, and dominance (d) is the difference from the heterozygote with homozygote mean. The dominance effect is viewed as an interaction between different alleles from same gene. When considering more than one gene, any kind of interaction among two or more genes is called epistasis (Falconer and Mackay 1996).

In quantitative genetics applied to plant and animal breeding, one of the most important subjects is the prediction of breeding values of individuals in a given population for the desired traits. In order to understand what the breeding value is, imagine a population large enough in Hard-Weinberg equilibrium, where one individual is mated with all individuals (or with all individuals of the opposite gender, in the case of animals) in this population, creating a second population. If the mean of these two populations is taken for a given trait, the difference of these means is half of the breeding value of the common parent for this trait. In others words, the breeding value is a measure of the direct effects that individuals transmit to theirs offspring.

If the genes have only additive effects, the breeding values correspond to the whole genotypic values, but if there are additive and dominance effects, the genotypic value is the sum of the breeding and dominance deviation value (Falconer and Mackay 1996). If two individuals mate, the genotypic value of their offspring is the means of its parents breeding values, plus the dominance deviation that is viewed as a non-explained part by breeding values and is specific for each individual, because it is a function of intra allelic interaction.

Initially, genetic evaluations used models that only included additive effects. These models yield the required information for some breeding system that explores mainly the additive effects, such as in dairy cattle, where the goal is the selection of bulls that provide semen for worldwide distribution. In addition, these models could be also used in general recurrent intra-population selection systems. In general, for any breeding program that strictly depends on the alleles of an individual that will be transmitted to the next generation more than the individual itself, the models based on additive models are a reasonable option. In addition, the models based on additive effects are recommended if the focus is the selection of inbred lines for use as an improved cultivar (e.g. soybean breeding), because in the improved cultivar there is no heterozygote loci.

However, most of the breeding systems explore the non-additive effects, especially if the breeding requests specific mating, or mating of individuals from different breeds or heterotic groups, where the dominance effect is highly explored from heterosis. In these cases, the inclusion of dominance effects in GS is desired. With the estimation of dominance marker effects, it is possible to predict specific combination abilities of a given cross, allowing the breeder to perform only the promising cross (Toro and Varona 2010).

For the breeding program, it is necessary to have at least basic knowledge of trait heritance to better explore it. If the trait is mainly governed by genes with predominance of additive effects, the breeding can be based on the exploration of individuals with higher general combination ability or higher breeding values. Whereas if the genes have important dominance effects, the breeding could explore specific crosses.

In practice, breeding systems dealing with non-additive effects are a challenge because dominance and epistasis are not directly transmitted to progenies (Resende Jr 2014). These effects depend on the genotype of the progeny depending on Mendelian segregation occurring in the parents during the gametogenesis. The direct effect transmitted by additive x additive interaction could be explored only for genes with a low recombination rate. Unraveling the genetic effects and variance components would be good to better understand and consequently find the best model to perform the predictions (Muñoz et al. 2014). In this same context, GS can be used to select the mating with highest probability of providing superior individuals (Resende Jr 2014).

The literature reports that the estimation of non-additive effects is complex; Toro and Varona (2010) argues that in genetic evaluations it is not common to include dominance effects, and the reason behind it can be the computational demands and the estimation usually being inaccurate. These authors believe that it would be necessary to have a large number of data including a high proportion of full sibs (Misztal et al. 1998). It is a fact that in traditional dominance relationship matrixes, information from pedigree is much sparser than in additive relationship matrixes, and for dominance estimation, it is necessary to have appropriate family structures. The results of Nishio and Satoh (2014) showed that dominance deviation prediction using SNP information was not as accurate as breeding values.

Using only pedigree information, the epistasis estimation is more complex than dominance estimation, and the difficulty increases according to the complexity level of the epistasis order, and/or which effects are involved. For example, in the first epistasis order, the epistasis includes two loci, which can be: additive x additive, additive x dominance or dominance x dominance; if more genes are involved, more complex is the combination of effect interactions. Similar to dominance estimations for estimating epistasis, it is necessary to have appropriate family structure; however, in the case of epistasis, more than one kind of family is necessary. An example can be seen if there are only half sibs, it is possible to estimate only additive variance, whereas if there were half and full sibs, it would be possible to estimate additive and dominance variance, and if the main objective is to estimate epistasis, it is necessary to have more complex pedigree, beyond half and full sibs, could be used inbred families. The GS models based on relationship information from individuals cannot go much further from pedigree demands to estimate non-additive variance components.

According to Hill (2010), it is usually impossible to estimate epistasis in random matting populations. With meta-analysis, Hill et al. (2008) concluded that most genetic variation is due to additive effects, and argue that emphasis given on using additive variation for the selection is still the best strategy.

In their study on simulation traits, even with the presence of dominance and epistasis in the traits, Wittenburg et al. (2011) showed that in most results, the model that included additive and dominance effects outperformed the additive models and also

models that matched effect additive, dominance and epistasis. With pig quantitative traits, Nishio and Satoh (2014) tried to match epistasis in additive-dominance model, however, the variance components for these effects were outside the parameter space. These authors also reported that additive-dominant GS model provided better results in real and simulated traits than additive models.

In simulated studies, Toro and Varona (2010) estimated additive and dominance effects for each marker and the expected genotypic values of offspring for all mating options. These authors reported an extra genetic gain with the best allocation of mating instead of only selecting individuals and random mating for traits with dominance presence. Also in simulated studies Zeng et al. (2013) showed that in additive-dominant models, the genetic gain in a crossbreed population was higher than only-additive models in traits with considerable dominance effects.

Some studies showed good results with the inclusion of epistasis. In pine tree heights, Muñoz et al. (2014) concluded that the model with inclusion of dominance and epistasis provided better breeding value predictions, and half of the genetic variation is due to non-additive effects. In addition, Su et al. (2012) also showed that models with additive, dominance and epistasis effects outperformed models without epistasis for average daily gain in pigs.

ACCURACY

Accuracy is the correlation between the estimated value and the parametric value, this is a parameter of the measure of bias and precision together (Resende 2002). One of the most important contributions of quantitative genetics is the prediction of genetic gain (GG) for one given breeding strategy (Ramalho et al. 1993). This GG can be predicted by the general equation $GG = ir_{a\hat{a}}\sigma_A$, where i is the selection intensity, $r_{a\hat{a}}$ is the accuracy of breeding value prediction, and σ_A is the additive standard deviation of the trait. From GG prediction, it is possible to see that accuracy is the most important parameter for comparing selection methods (Resende 2002).

Resende (2008) proposed the equation for the expected accuracy value for GS, and other authors have also proposed similar equations (Daetwyler et al. 2008; Goddard et al. 2011). From these equations, it is argued that accuracy depends on heritability,

genetic architecture of traits, effective size (Ne), number of individuals in reference population and marker density (Resende et al. 2014). The first and second factors cannot be changed by the breeder. However, the other factors can be manipulated to improve the accuracies (Resende et al. 2014).

In all prediction models, higher accuracies are achieved with large number of phenotype records. The Ne and marker density factors are directly related with LD. In order to achieve a high accuracy in GS, it is necessary that the population is in high LD. It is expected to obtain higher LD with small Ne and dense marker panel (Resende et al. 2014). For traits with high heritability, higher accuracies are expected (Resende Jr et al. 2012b). Finally, the genetic architecture involves the number of genes, distribution and presence or absence of non-additive effects; in order to achieve higher accuracy it is necessary to use models that best explain the genetic architecture.

In experimental results, (Legarra et al. 2008) proposed an accuracy estimator from the correlation among phenotypes and GEBV, and parametric heritability.

The accuracy ($r_{g\hat{g}}$) is the correlation between GEBV ($\hat{g}$) and parametric g:

$$r_{g\hat{g}} = \frac{cov(g,\hat{g})}{\sigma_g \sigma_{\hat{g}}}$$

Assuming:

$$h^2 = \sigma_g^2/\sigma_y^2$$

$$cov(y,\hat{g}) = cov(g+e,\hat{g}) = cov(g,\hat{g})$$

then:

$$r_{g\hat{g}} = \frac{cov(y,\hat{g})}{h\sigma_y \sigma_{\hat{g}}} = \frac{r_{y\hat{g}}}{h}$$

From this accuracy equation, it can be concluded that for the selection of the best GS model, the choice from $r_{y\hat{g}}$ leads to the choice of a more accurate method. In literature, it is common to see reports on the method comparison from $r_{y\hat{g}}$ (Pérez et al. 2012; Ertl et al. 2014; Nishio and Satoh 2014; Crossa et al. 2014).

PERSPECTIVE OF PREDICTIONS IN BREEDING: LARGE DATA SET IS COMING

In the GS seminal paper, Meuwissen et al. (2001) simulated a scenario where one marker was available in each cM. At the time, the study could be performed only in

simulated date, because the number of markers was limited for most of the species. However, nowadays, the number of marker available has increased, and it is common to have SNP panels larger than 600K applied to breeding populations (Crossa et al. 2013; Ertl et al. 2014). With the advancement of technologies, this figure tends to increase, creating more statistics and computational challenges.

In order to deal with this large set of data, it will be necessary to filter the data used. In the case of markers, it might not be a good practice to include all markers indiscriminately in the model. The LD pruning approach would be a good option to remove markers that are in high linkage disequilibrium and consequently promote multicollinearity problems, in addition association studies could help to discover important genome regions for a target trait, and in predictive models, important markers could be used. The better understanding of the trait and the inclusion of reasonable knowledge of markers in the predictive model can play an important role in avoiding false positive discoveries (Ioannidis 2005). However, this marker selection should be performed with care, in order to avoid spurious associations (Wray et al. 2013). One option would be to perform a GWAS experiment with a large number of unrelated individuals, and select markers to use in GS in breeding populations. In addition, a gene expression approach, such as transcriptomes, proteomics or epigenetics, could be used as a tool to better understand the target trait in a specific environmental condition.

According to Varshney et al. (2014), the genotyping of a large number of individuals for dense marker panels is not cost limiting. Currently, the bottleneck in plant breeding is phenotyping, which demands a very high cost. For these authors, it is necessary to find cost-effective and precise phenotyping methodologies, which will involve digital image capturing, remote sensing, and many new forms of information and communication technologies to overcome the phenotype cost barrier. Some of these technologies are available but their efficiency should be evaluated on a case-to-case basis.

Even with the issues about phenotyping, to date, a large number of phenotype records is usually used in genetic evaluations. For instance, in the Brazilian Zebu genetic evaluation of 2014-2015, over 1 million phenotyped individuals were used (Silva Personal Communication). In this case, the evaluation was not performed with genetic markers, but it is expected that breeding programs will use large markers and phenotype data sets in

the coming years, which will demand efficient software to be implemented along with efficient algorithms and also high performance computers.

Another point that requests the breeding program to record large numbers of phenotypes is the Genotype by Environment interaction (GE). This interaction makes the best genotype in a given environment not necessarily be the best in other environment conditions, and since the plant breeding programs aim to recommend cultivars for a region, GE is one of most difficult interactions in plant breeding. In order to overcome GE, it is necessary to perform field test on the potential genotypes in a broad range environment in order to provide recommendations of improved cultivars with high general adaptably or define breeding zones where the effects of GE inside these zones are non-significant. In this case, cultivar recommendation should be performed for each breeding zone (Ramalho et al. 1993; Cruz et al. 2012; Ramalho et al. 2012; Cruz et al. 2014). With the dense marker panel available, a model that includes environment variations and genomic information is necessary for cultivar recommendation. The environment effects and GE can be adjusted as two additional effects in regular mixed models, or with multi-environment models with different covariance structures (Burgueño et al. 2012; Lopez-Cruz et al. 2015). Nevertheless, these models could not predict GEBV for new environments where the individuals were not phenotyped. Therefore, with well characterized environments by environmental-covariates, (e.g. weather covariates), the GS models can be extended to predictions in unobserved environments (Heslot et al. 2013; Jarquín et al. 2014).

The reasonable use of molecular data with integrated phenotypic information is already a reality in breeding programs, but such information is coming in a larger scale, and the use of all information that arrives is a challenge for plant and animal breeding applications.

GENOMIC PREDICTION IN PINE BREEDING

The species from the *Pinus* genus are gymnosperms from Coniferophyta (or Pinophyta) division. There are more than one hundred species of this genus, but few of these species are economically exploited (Aguiar et al. 2011). *Pinus taeda L*. is the scientific name of loblolly pine, also known as Arkansas pine. This is the most important

commercial forest specie in the southern United States, where it can be observed in about 11.7 million ha (Baker and Langdon 1990). In Brazil, the loblolly pine is the main forest tree planted in Southern Brazil (Alcantara et al. 2007) due to its potential for growing in low temperature conditions and its specific characteristics of wood (long fiber) (Aguiar et al. 2011).

The wood of loblolly pine is what is exploited, thus for a cultivar, high trees with large diameters at breast height are desirable. It is also important that the plants resistant both pests and diseases. The loblolly pine is a diploid (Zimin et al. 2014) and monoecious (Baker and Langdon 1990), and the breeding consists in recommending improved seeds or clones as cultivar. There are some benefits of clonal cultivars, mainly due that with clonal selection, it is possible to transfer the entire genotypic value, and the resulting commercial forest will be more uniform. Beyond the introduction of plants, the hybridization and selection are fundamental for loblolly pine breeding, and the genomic prediction is a potential approach for selecting the best individuals and recommending any potential crossbreeding.

In order to determine the number of markers and the number of individuals used in a genomic prediction, the factors affecting accuracy should be considered. The loblolly pine has 12 chromosomes and approximately 22Gb were sequenced and assembled (Zimin et al. 2014), the genome size of pine is large when compared to many other species, for instance, the human genome is around 3Gb (Lander et al. 2001; Venter et al. 2001). Moreover, the pine populations usually has a low LD (Brown et al. 2004) and the important traits in pine has small heritability (Resende Jr et al. 2012a; Resende Jr et al. 2012b). All of these factors indicate that a larger number of markers should be necessary in order to cover the entire genome, with makers in strong LD with major part of genes controlling the trait, in addition to greater precision of experimental conditions, being important for record the traits. Therefore, experimental results for pine showed that a number of approximately 5K polymorphic markers and 1K individuals with eight clonal repetitions is enough for achieving reasonable accuracy for traits with low heritability (Resende Jr et al. 2012b; Muñoz et al. 2014). These results agree with (Gratapaglia & Resende 2011) in a study with determinist simulation, under forest breeding conditions,

that demonstrated little accuracy improvement in using more than 1K individuals in many scenarios with distinct heritability, with different number of markers and genes.

Since it is possible to have large-scale cloning in pine, and the whole-genotypic values are transferred in the clonal selection, the genetic prediction models should include non-additive effects to predict these genotypic values. The genotypic values could be predicted by pedigree information, but with marker information, greater genotypic accuracies can be achieved with genomic prediction models, such as additive-dominance whole-genomic regressions, GBLUP models with non-additive effects (Muñoz et al. 2014; Azevedo et al. 2015) and RKHS models, that theoretically predict the additive and non-additive effects confused (Gianola et al. 2006; Gianola and van Kaam 2008).

However, even with the possibility of clonal selection, the crossing among selected individuals is important in pine breeding in order to have new genotype combinations, and consequently for the breeding program to continue having genetic gain. Therefore, it is important to predict the breeding values and the general combination ability (half of the breeding values), in order to select individuals to be used in large number of crosses. It is also important in order to predict the specific combination ability of a particular cross. For these proposes, the genomic prediction approach could be used to predict the breeding values (Meuwissen et al. 2001) and the expected genotypic merit of a progeny (Toro and Varona 2010; Ertl et al. 2014), where the breeder could perform only the selected cross combination, instead of performing all possible combinations.

The reciprocal recurrent selection with full-sib families (Hallauer et al. 2010) can be used in pine breeding using two population from distinct heterotic groups. This process is based in the evaluation of full-sibs families with one parent from each heterotic group, also one genitor could be crossed with more than one individual, thus beyond full-sibs will also would result half-sibs families, what allow estimate additive and dominance variance components. After the evaluation of families, the parents of the best families are intercrossed within their respective heterotic group. However, since it is possible to clone pine, a superior individual can be cloned at anytime, and used as a clonal cultivar. With selection and recombination, the frequency of favorable alleles is expected to improve, and an inter-population hybrid of one generation has higher genotypic value than the inter-population hybrid of previous generations.

In a conservative scenario, genomic prediction can be used to select individuals to be phenotyped. This approach can be applied in this breeding scheme for allocating the potential cross-combination. After performing these combinations, the individuals from these progenies can be genotyped in order to have their genetic merit predicted with the genomic model already fitted for the parent generation, and thus the breeder can select individuals inside families to be phenotyped. In a less conservative scenario, the breeder can use the genomic models for selecting the individuals that will directly contribute to the next generation, without any phenotyping. This less conservative scenario could considerably reduce the time for generations (Schaeffer 2006; Resende et al. 2012), but many issues should be evaluated, mainly related to the accuracy across generation in each situation.

However, even with low LD, selection inside families, as mentioned in the previous example of GS application, requires a lower number of markers than in parental generation, since the LD inside families are strong, and with the genotypes from parents, it is possible to expand the number of markers inside the progeny with imputation (Browning and Browning 2007; Hickey et al. 2012; Sargolzaei et al. 2014).

Other important point is related to the use of low-density SNP panel and the genotyping of more individuals. In advanced generations of pine breeding, the breeder will have a wide pedigree information across the generations, and thus the pedigree information can be combined with markers and even with low-density marker panels, the accuracy would not decrease too much due the large number of individuals and pedigree information. Habier et al. (2009) showed only a small loss in accuracy when selecting one marker at 10 cM, when compared with high-density panel with an average of 1 marker per cM.

In practical applications, genotyping with low-density panel is cheaper, thus it would be possible to genotype a large number of individuals, and even with the reduction in accuracy due the reduced number of markers, depending on the situation, higher genetic gains could be achieved with higher selection intensity due to the higher number of genotyped individuals.

STATISTIC MODELS FOR GENOMIC SELECTION

The GS models can be originated from Frequentist or Bayesian statistics. However, this review will include only Bayesian models. Most statistic models applied in GS are based on regressions where the markers are the covariates or are based on animal models but with the replacement of the traditional pedigree relationship matrix to a relationship matrix based on markers. The general GS model with polygenic effects can be given by:

$$y = X\beta + Zg + Zu + Z\delta + e$$

Where:

y: phenotype vector (or corrected phenotypes); β: systematic effect vector (fixed in a frequentist sense); g: genotypic value vector explained by markers, in general, what differs in GS models is how to model this vector; u: additive polygenic effect; δ: dominant polygenic effect; e: residuals effect vector.

If the g term is ignored, the model is similar to the pedigree base-line model and u and δ are breeding values and dominance deviation, respectively.

In this model, can be assumed that:

$$y|X\beta + Zg + Zu + Z\delta, I\sigma_e^2 \sim MVN(X\beta + Zg + Zu + Z\delta, I\sigma_e^2)$$

$$p(\beta) \propto 1$$

$$u|A\sigma_u^2 \sim MVN(0, A\sigma_u^2)$$

$$\sigma_u^2 \sim \chi^{-2}(\nu_u, S_u)$$

$$\delta|D\sigma_\delta^2 \sim MVN(0, D\sigma_\delta^2)$$

$$\sigma_\delta^2 \sim \chi^{-2}(\nu_\delta, S_\delta)$$

$$e|I\sigma_e^2 \sim MVN(0, I\sigma_e^2)$$

$$\sigma_e^2 \sim \chi^{-2}(\nu_e, S_e)$$

$$g|V_g \sim MVN(0, V_g)$$

The A and D are additive and dominance relationship matrixes, respectively; V_g is the covariance matrix from marker effects on the form and distribution of V_g depending of g adopted.

WHOLE-GENOME REGRESSIONS

The whole-genome regression (WGR) are multiple linear regressions that input markers as covariates. In GS common situations, the number of markers available is much larger than the number of individuals. Thus, the WGR model where the number of covariates is larger than the number of observations is unfeasible by ordinal least square. Then, one alternative is to consider that marker effects are random and use REML/BLUP procedure, or alternatively, estimate marker effects using Bayesian statistics. In these WGR models, effects for each marker are estimated.

ESTIMATION OF A AND D AND EPISTASIS

From the general GS model, simplified to assume that there are no systematic or polygenic effects, the model is given by: $y_j = \mu + g_j + e_j$; if replacing g_j to estimate a, d and the pairwise interaction, the model would be:

$$y_j = \mu + \sum_{i}^{k}\left(x_{ij}a_i + w_{ij}d_i\right) + \sum_{i}^{k}\sum_{\substack{i' \\ i' \neq i}}^{k} x_{ij}x_{i\prime j}aa_i + \sum_{i}^{k}\sum_{\substack{i' \\ i' \neq i}}^{k} x_{ij}w_{i\prime j}ad_i + \sum_{i}^{k}\sum_{\substack{i' \\ i' \neq i}}^{k} w_{ij}w_{i\prime j}dd_i + e_j$$

Where:

a_i and d_i are additive and dominance effects of loci i, respectively; aa_i, ad_i and dd_i are additive-by-additive, additive-by-dominance and dominance-by-dominance effects of loci i, respectively; x_{ij} and w_{ij} are functions of bialleic marker (e.g. SNP) i from individual j; x_{ij} assumes the values 1, 0 and -1 for genotypes AA, Aa and aa, respectively, and $w_{ij} = 1 - \left|x_{ij}\right|$. This model with epistasis effects, has a large number of parameters, mainly with large number of markers, and thus, it is more feasible in WGR to ignore epistasis.

BREEDING AND DOMINANCE DEVIATION VALUES AND CROSS PREDICTION

Considering the WGR with the absence of epistasis, the model would be:

$$y_j = \mu + \sum_{i}^{k}\left(x_{ij}a_i + w_{ij}d_i\right) + e_j$$

After estimating a_i and d_i, the genomic expected breeding values (GEBV) and dominance deviation (GEDD) would be given by:

$$GEBV_j = \sum_i \left[I(x_{ij} = 1) 2q_i + I(x_{ij} = 0)(q_i - p_i) - I(x_{ij} = -1) 2p_i \right] \hat{\alpha}_i$$

$$GEDD_j = \sum_i \left[-I(x_{ij} = 1) 2q_i^2 + I(x_{ij} = 0) 2p_i q_i - I(x_{ij} = -1) 2p_i^2 \right] \hat{d}_i$$

Where p_i is allele frequency of allele A in SNP *i*; q_i*=1-*p_i; $\hat{\alpha}_i$ is the average effect of substitution. $\hat{\alpha}_i = \hat{a}_i + \hat{d}_i(q_i - p_i)$, and I is an indicator function of SNPs. The whole genotypic value of an individual is the sum between GEBV and GEDD, and the whole genotypic value expected for the progeny from the cross between the individuals j and j' (G_{j*}) is:

$$G_{j*} = \frac{GEBV_j + GEBV_{j'}}{2} + \sum_i \left[-P(x_{ij*} = 1) 2q_i^2 + P(x_{ij*} = 0) 2p_i q_i - P(x_{ij*} = -1) 2p_i^2 \right] \hat{d}_i$$

Alternatively:

$$G_{j*} = \sum_i \left[P(x_{ij*} = 1) \hat{a}_i + P(x_{ij*} = 0) \hat{d}_i - P(x_{ij*} = -1) \hat{a}_i \right]$$

DISTRIBUTIONS ASSUMED FOR REGRESSION COEFFICIENTS

Most WGR methods share the same linear model, differing, however, regarding prior distribution adopted for markers effects. These different priors adopted to marker effects allow Bayesian WGR to be very flexible regarding genetic architecture. Some prior distributions for parameterization with a and d are shown below, as well as how to estimate additive and dominance variance under Hardy-Weinberg equilibrium.

Bayesian Ridge Regression (BRR)

The BRR is the Bayesian version of RR-BLUP proposed by Meuwissen et al. (2001). In additive models, BRR assumes that all regression coefficients have the same variance component. In additive-dominant WGR, BRR assumes that the additive effects have common variance σ_a^2 and the dominance effects have the common variance σ_d^2. As the common variance assumed in BRR, the literature usually assumes that BRR performs the same shrinkage, but according to Gianola (2013), this is not the case. Less shrinkage towards zero of markers presents intermediate allelic frequencies. The BRR assumes:

$$p(\mu) \propto 1$$

$$a_i | \sigma_a^2 \sim N(0, \sigma_a^2)$$

$$\sigma_a^2 \sim \chi^{-2}(\upsilon_a, S_a)$$
$$d_i|\sigma_d^2 \sim N(0, \sigma_d^2)$$
$$\sigma_d^2 \sim \chi^{-2}(\upsilon_d, S_d)$$
$$e|\sigma_e^2 \sim N(0, \sigma_e^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e)$$

Bayes A

The original BayesA was proposed by (Meuwissen et al. 2001). This method is similar to BRR, with the difference that the variance component of regression coefficient is heterogeneous for markers that belong to different chromosome segments. This method allows some marker to have higher effects. However, even with heterogeneous variance, the marginal regression coefficient prior is common ($a_i|\upsilon_a, S_a \sim t(0, \upsilon_a, S_a)$) (Gianola et al. 2009). These authors also criticized the original BayesA because this method allows only a few "Bayesian learning" (mainly for small number of individuals) for variance components. This means that the variance in component estimations is strength dependent of prior distribution. This issue can be better analyzed when examining the fully conditional posterior distributions of variance components:

$$\sigma_{a_i}^2|ELSE \sim \chi^{-2}(\tilde{\upsilon}_a, \tilde{S}_a)$$

Where:

$$\tilde{\upsilon}_a = \upsilon_a + 1$$
$$\tilde{S}_a = \frac{\upsilon_a S_a + a_i^2}{\upsilon_a + 1}$$

As in most markers, $a_i^2 \approx 0$, thus, in most cases $\tilde{S}_a \approx S_a[\upsilon_a/(\upsilon_a + 1)]$. The distribution of $\sigma_{a_i}^2|ELSE$ suggests that the shrinkage in marker effects is highly dependent of hyper-parameters that are arbitrarily chosen. However, since in original BayesA (Meuwissen et al. 2001) it is assumed that there is a common variance component for all markers inside a given segment of 1 cM, and in this study the authors reported approximately 50 markers/cM, it is possible to assume that the influence of hyper-parameters on original BayesA was not as high as pointed out in (Gianola et al. 2009). In the original BayesA, the fully conditional posterior distributions of variance components reported by (Meuwissen et al. 2001) were:

$$\sigma_{a_i}^2|ELSE \sim \chi^{-2}(\tilde{\upsilon}_a, \tilde{S}_a)$$

$$\tilde{\upsilon}_a = \upsilon_a + m_i$$

$$\tilde{S}_a = S_a + \sum_{i}^{m_i} a_i^2$$

m_i: is the number of markers in the i[th] segment.

In order to overcome the hyper-parameters influence, Gianola et al. (2009) suggested that the markers could be grouped and could assume common variance for markers that belong to the same group (similar to the original BayesA), and/or assume that S_a and υ_a are parameters (not hyper-parameter) with non-informative prior distributions. The studies by (de los Campos and Perez 2014; Pérez and de los Campos 2014) have modified BayesA. These authors assumed that shape parameters for χ^{-2} follow the gamma distribution, which meet part of suggestions from (Gianola et al. 2009). The prior distributions in modified BayesA considering additive and dominance effects are:

$$p(\mu) \propto 1$$

$$a_i|\sigma_a^2 \sim N(0, \sigma_{a_i}^2)$$

$$\sigma_{a_i}^2 \sim \chi^{-2}(\upsilon_a, S_a)$$

$$S_a \sim Gamma(r_a, s_a)$$

$$d_i|\sigma_d^2 \sim N(0, \sigma_{d_i}^2)$$

$$\sigma_{d_i}^2 \sim \chi^{-2}(\upsilon_d, S_d)$$

$$S_d \sim Gamma(r_d, s_d)$$

$$e|\sigma_e^2 \sim N(0, \sigma_e^2)$$

$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e)$$

BayesB

In addition to BayesA, Meuwissen et al. (2001) proposed other Bayesian approach called BayesB. Those authors recognize that BayesA has issues, such as the fact that variance distributions of marker effects do not show point mass equal to zero. According to those authors, this is a desired feature, since most of the loci do not contribute to genetic

variance (non-segregating) and only a few loci contribute to this genetic variance. In order to attribute this feature, the original BayesB (additive model) assumes the following genetic parameters:

$$a_i|\sigma^2_{a_i} \sim N(0, \sigma^2_{a_i})$$

$$\sigma^2_{a_i} \sim \chi^{-2}(\upsilon_a, S_a) \quad \text{with probability equal } \textit{1-}\pi$$
$$\sigma^2_{a_i} = 0 \quad \text{with probability equal } \pi$$

According to (Gianola et al. 2009), with the prior assumption of $\sigma^2_{a_i}$=0, it is consequently assumed that $a_i|(\sigma^2_{a_i} = 0) = \theta$, where θ is a given real number, considering the logic in (Meuwissen et al. 2001) $\theta = 0$. Therefore, Gianola et al. (2009) criticized this formulation, because assuming $\sigma^2_{a_i} = 0$ implies determinism about such an effect. In order to overcome this situation, Gianola et al. (2009) suggested the formulation below for genetic parameters:

$$a_i|\sigma^2_{a_i} = \begin{cases} 0 & \text{with probability } \pi \\ \sim N(0, \sigma^2_{a_i}) & \text{with probability } \textit{1-}\pi \end{cases}$$

and

$$\sigma^2_{a_i} \sim \chi^{-2}(\upsilon_a, S_a) \quad \text{If the marker is included in the model}$$

Since BayesA is a special case of BayesB with π*=0*, the same issues regarding Bayesian learning in BayesA are extended to BayesB (Gianola et al. 2009). Therefore, since the prior marginal variance of regression coefficient is reduced by a fraction π, the Bayesian learning is even more difficult in BayesB than in BayesA (Gianola 2013).

Other issue pointed out in BayesB (Gianola et al. 2009; Habier et al. 2011; Gianola 2013) is related to π, which, in original BayesB, is an arbitrary value that drives marker selection. According to those authors, π should be a parameter estimated in the model. Gianola et al. (2009) suggested that π should be a parameter that follows beta distribution. Habier et al. (2011) formulated a method called BayesDπ that assumes uniform distribution to π. The use of beta distribution for π is desired because it is possible to control the prior value to π and the confidence value regarding this prior value from the

expectation and variance of beta. The BayesB addressed here considered beta distribution to π and modification of BayesA:

$$p(\mu) \propto 1$$

$$a_i|\sigma^2_{a_i} = \begin{cases} 0 & \text{with probability } \pi_a \\ \sim N(0, \sigma^2_{a_i}) & \text{with probability } 1-\pi_a \end{cases}$$

$$\sigma^2_{a_i} \sim \chi^{-2}(\upsilon_a, S_a)$$

$$S_a \sim Gamma(r_a, s_a)$$

$$\pi_a \sim Beta(b_{a_1}, b_{a_2})$$

$$d_i|\sigma^2_{d_i} = \begin{cases} 0 & \text{with probability } \pi_d \\ \sim N(0, \sigma^2_{d_i}) & \text{with probability } 1-\pi_d \end{cases}$$

$$\sigma^2_{d_i} \sim \chi^{-2}(\upsilon_d, S_d)$$

$$\pi_d \sim Beta(b_{d_1}, b_{d_2})$$

$$S_d \sim Gamma(r_d, s_d)$$

$$e|\sigma^2_e \sim N(0, \sigma^2_e)$$

$$\sigma^2_e \sim \chi^{-2}(\upsilon_e, S_e)$$

BayesCπ

BayesCπ was proposed by Habier et al. (2011). It is similar to BRR, since BayesCπ also assumes that marker coefficients have homogenous variance. However, similar to BayesB, BayesCπ implemented parameter to select markers that are not associated to any genes. BayesCπ assumes that:

$$p(\mu) \propto 1$$

$$a_i|\sigma^2_a = \begin{cases} 0 & \text{with probability } \pi_a \\ \sim N(0, \sigma^2_a) & \text{with probability } 1-\pi_a \end{cases}$$

$$\sigma^2_a \sim \chi^{-2}(\upsilon_a, S_a)$$

$$\pi_a \sim Beta(b_{a_1}, b_{a_2})$$

$$d_i|\sigma_d^2 = \begin{cases} 0 & \text{with probability } \pi_d \\ \sim N(0, \sigma_d^2) & \text{with probability } 1-\pi_d \end{cases}$$

$$\sigma_d^2 \sim \chi^{-2}(\upsilon_d, S_d)$$
$$\pi_d \sim Beta(b_{d_1}, b_{d_2})$$
$$e|\sigma_e^2 \sim N(0, \sigma_e^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e)$$

Bayesian Lasso (BL)

The Bayesian Lasso (BL) (Lasso - *Least Absolute Shrinkage and Selection Operator*) was proposed initially by Park and Casella (2008) and de los Campos et al. (2009) adapted BL to genomic prediction. Similar to BayesA and BayesB, BL assumes that the covariates (markers) do not have common variance and additionally promote an indirect covariate selection from strong shrinkage in marker effects, since marginal prior of regression coefficients follows double exponential (DE) distribution (or Laplace distribution) (Park and Casella 2008). BL assumes that:

$$a_i|\tau_{a_i}^2, \sigma_e^2 \sim N(0, \tau_{a_i}^2 \sigma_e^2)$$
$$\tau_{a_i}^2|\lambda \sim Exp(0.5\lambda_a^2)$$
$$\lambda_a^2 \sim Gamma(r_a, s_a)$$
$$d_i|\tau_{d_i}^2, \sigma_e^2 \sim N(0, \tau_{d_i}^2 \sigma_e^2)$$
$$\tau_{d_i}^2 \sim Exp(0.5\lambda_d^2)$$
$$\lambda_d^2 \sim Gamma(r_d, s_d)$$
$$e|\sigma_e^2 \sim N(0, \sigma_e^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e)$$

Considerations of lambda prior

The use of Gamma distribution for λ^2 (λ_a^2 and/or λ_d^2) as initially proposed by (Park and Casella 2008) providing the full conditional distribution with closed form ($p(\lambda^2|ELSE)$), which allows the implementation of Gibbs Sampler algorithm. However, according to de

los Campos et al. (2009), with this prior it is not possible to adopt the insufficient reason principle. In others words, this prior does not allow vague prior knowledge for λ^2. Thus, these authors suggested beta distribution for this parameter. Nevertheless, in order to use beta distribution, it is necessary to use Metropolis-Hastings algorithm.

The motivation to consider vague prior knowledge for λ^2 is observed in a report of these same authors, whereas estimated values of λ^2 were highly influenced by prior and hyper-parameters (few Bayesian learning). However, even though de los Campos et al. (2009) demonstrated in the same study that these different priors (and hyper-parameters) did not provide considerable differences in GEBV estimation, which indicated that BL showed a large Bayesian learning for GS.

Other Bayesian Lasso formulation

In BL formulation, the marker variance is a function of residual variance ($a_i|\tau^2_{a_i}, \sigma^2_e \sim N(0, \tau^2_{a_i}\sigma^2_e)$) (Park and Casella 2008; de Los Campos et al. 2009). Legarra et al. (2011) pointed out that marker distribution should not be related to residual variance, and proposed an alternative formulation for the Bayesian Lasso (BL2). In this formulation for the additive model, the authors assumed that:

$$a_i|\tau_i^2, \lambda^2 \sim N(0, \tau_i^2)$$
$$\tau_i^2|\lambda^2 \sim Exp(\lambda)$$
$$\sigma_e^2 \sim \chi^{-2}(\nu, S^2)$$
$$\lambda^2 \sim U(0, 10^6)$$

According to (Gianola 2013), the additive variance in BL is not necessarily dependent of residual variance. This author argues that in a standard additive infinitesimal model of quantitative genetics the additive variance can be given by:

$$V_A = V_F - V_E = h^2 V_F = V_E \frac{h^2}{1 - h^2}$$

Gianola (2013) also argues that if one considers the additive variance of a marker as $Var(a_i|v_e, S_e, r_a, s_a)$, where v_e, S_e, r_a, s_a are hyper-parameters of BL (previous described), thus:

$$Var(a_i|v_e, S^2, \alpha_1, \alpha_2) = \frac{2v_e S_e s_a}{(v_e - 2)(r_a - 1)}$$

Gianola (2013) showed that V_E should be viewed as $\upsilon_e S_e/(\upsilon_e - 2)$ since this term is the expected value of the prior distribution assigned to the residual variance. Then, $2s_a/(r_a - 1)$ plays a role on $h^2/(1-h^2)$. With similar idea, BRR could be formulated as: $a_i|\sigma_a^2 \sim N(0, \sigma_e^2\, h^2/(1-h^2)$, but it would also be necessary to formulate a prior for h^2.

WGR variance components

After fit the WGR model and assuming linkage equilibrium, absence of epistasis and Hardy-Weinberg equilibrium (Gianola et al. 2009), the additive and dominance variance captured by markers could be estimated from:

$$\hat{V}_A = 2(1-\hat{\pi}_a)\sum_i p_i q_i \left[\hat{\sigma}_{a_i}^2 + (q_i - p_i)^2 \hat{\sigma}_{d_i}^2\right]$$

and

$$\hat{V}_D = 4(1-\hat{\pi}_d)\sum_i (p_i q_i)^2\, \hat{\sigma}_{d_i}^2$$

From these general equations of variance estimators, it is possible to estimate V_A and V_D for all WGR. In BayesA, BRR and BL $\hat{\pi}_a = \hat{\pi}_d = 0$; In BRR and BayesCπ $\hat{\sigma}_{a_i}^2 = \hat{\sigma}_a^2$ and $\hat{\sigma}_{d_i}^2 = \hat{\sigma}_d^2$; In BL $\hat{\sigma}_{a_i}^2 = \hat{\tau}_{a_i}^2 \hat{\sigma}_e^2$ and $\hat{\sigma}_{d_i}^2 = \hat{\tau}_{d_i}^2 \hat{\sigma}_e^2$.

The assumptions of linkage equilibrium is a slight paradox, since in GP the linkage disequilibrium (LD) is useful to achieve good predictions. Another factor that play role for achieve high accuracies is the structure formed by relationship among the individuals, since in populations structured in families the predictions tend to be more accurate than a population of unrelated individuals, consequently due the family structure, selection and drift, both assumptions - LD and HW - are not attended in common breeding populations. However, according to (Gianola et al. 2009) these are assumed to make the issue addressable. These authors have also mentioned that accommodating LD explicitly in the prediction models would be a “formidable challenge”.

Individual models

Individual models are based on standard baseline models used in animal breeding, and more recently in plant breeding, but with the replacement of the relationship matrix resulted from pedigree to a relationship matrix from markers. The standard GS based on individual model is the additive GBLUP (VanRaden 2008). This method can be extended

for additive-dominance (Vitezica et al. 2013; Nishio and Satoh 2014) and epistasis can also be included in GBLUP (Su et al. 2012; Morota et al. 2014; Muñoz et al. 2014). The models based on GBLUP and BRR are equivalent models (de los Campos et al. 2013). Moreover, it is also possible to accommodate heterogeneous genetic variance for each marker in the relationship matrix. This method is called het-GBLUP (Legarra et al. 2011; Resende et al. 2014) and depending on the assumption for marker effects, it can be an equivalent model to BayesA, BayesB or BL. Another individual model is the Reproducing Kernel Hilbert Space (RKHS), this model can theoretically predict the entire genotypic value (additive and non-additive effects). This model is described below.

Reproducing Kernel Hilbert Space (RKHS)

The RKHS in GS prediction was proposed by (Gianola et al. 2006; Gianola and van Kaam 2008). From the standard genetic evaluation model $y = \mu + g + e$, where: μ is the intercept, e is the error vector, and g is a vector of unknown function of genetic characterization of individuals, until now any assumption were calculated for g. Therefore, g can be viewed as a non-parametric function. The g can be achieve with penalized regression, and one way to do this is by minimizing the mean square error, under the restriction that g belongs to Hilbert space:

$$\hat{g} = \underset{g \in \mathcal{H}}{\arg\min}\{(y - \mu - g)'(y - \mu - g) + \lambda \|g\|_{\mathcal{H}}^2\}$$

Where: λ is the parameter used to control the trade-off between the model suitability and complexity, and $\|.\|_{\mathcal{H}}^2$ denotes square norm in Hilbert space (de Los Campos et al. 2009). According to (Gianola et al. 2006; Gianola and van Kaam 2008), the solution of this penalized regression leads to model: $y = \mu + K\alpha + e$, ($g = K\alpha$) where: K is a symmetric, positive definite matrix that corresponds to the relationship among individuals. Now, assuming that α is a parameter and $\alpha \sim N(0, K^{-1}\sigma_k^2)$, it can be demonstrated (de Los Campos et al. 2009) that RKHS model is equivalent to animal models where $g \sim N(0, K\sigma_k^2)$. Thus, it can be concluded that pedigree baseline models, and GBLUP are special cases of RKHS, but it is necessary to substitute K for the appropriate kernel matrix (de Los Campos et al. 2009; Morota and Gianola 2014).

According to (Gianola et al. 2006; Gianola and van Kaam 2008; Morota et al. 2014) with kernels, the RKHS model can predict the entire genotypic value; however, in a non-explicit manner. One kernel suggested by these authors is called Gaussian Kernel, which is given by:

$$K = \exp(-\varphi D_e^2)$$

Where: D_e^2 is the squared Euclidean distance using marker incidence matrix. This matrix is the same matrix included in additive WGR models; this matrix assume the values -1 (aa), 0 (Aa) and 1 (AA). The φ is the bandwidth parameters that control how fast the covariance function drops as the points get further apart, as measured by D_e^2 (de los Campos et al. 2010b). For a given distance between two individuals ($D_{e_{jj'}}^2$), if $\varphi \longrightarrow \infty$ the correlation of these individuals is $K_{jj'} \longrightarrow 0$, and if $\varphi \longrightarrow 0$, then $K_{jj'} \longrightarrow 1$.

With the Bayesian approach, φ can be estimated from Metropolis-Hasting algorithm (Gianola et al. 2006; Gianola and van Kaam 2008), but these authors argue that it would demand a lot of time and computational resources. They also suggested to test a grid of values for φ and choose the one maximizing the predict ability. However, testing a grid values for φ also demands a great amount of time and computational resources. In order to overcome the problem of choosing (or estimating) φ, (de los Campos et al. 2010b) proposed the use of multiple kernels with the same distance matrix, but with different bandwidth values. This approach was called “kernel averaging”.

This kernel averaging approach can include many kernels, and this model considering three kernels is:

$$y = \mu + g_1 + g_2 + g_3 + e$$
$$g_k | K_k\sigma_k^2 \sim N(0, K_k\sigma_k^2) : \{k = 1,2,3\}$$
$$\sigma_k^2 \sim \chi^{-2}(\upsilon_k, S_k)$$
$$e | \sigma_e^2 \sim N(0, \sigma_e^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_e, S_e)$$
$$K_k = \exp(-\varphi_k D^2)$$

González-Camacho et al. (2012) suggested that φ_1, φ_2 and φ_3 can be 5/h, 1/h and 1/5h, where h is the 5th percentile of the distribution of squared Euclidean distances between pairs of individuals, with leading local intermediate and global kernels,

respectively (Tusell et al. 2014). These suggestion agree with (Crossa et al. 2010), where these authors argues that the bandwidth parameter should consider the distribution of distance values.

In the genetic point of view, the entire genotypic value is $g = g_1 + g_2 + g_3$, and the entire genotypic variance is $\sigma_g^2 = \sum_k \sigma_k^2$ (de los Campos et al. 2010b). However, it is not possible to split the genotypic variance in additive and non-additive terms.

Other kernels to predict entire genotypic values are available, but this Gaussian kernel with kernel averaging approach has been showed to be a robust choice (Morota et al. 2013; Tusell et al. 2014). Morota et al. (2014) suggested using Gaussian kernel build from dominance incidence matrix, and using this new kernel together with the previous model. However, the inclusion of new kernels has not provided greater accuracies.

Choice of hyper-parameters

According to de los Campos et al. (2013), the hyper-parameters control the extent and strength of shrinkage of marker effect estimates, and they can have important impacts on inferences. These authors suggest the solutions below when dealing with the choice of hyper-parameters:

Heritability-based rules: In this case, the hyper-parameters are chosen based on prior expectation regarding the genetic variance components of trait, similar to the seminal GP paper (Meuwissen et al. 2001) and also (Zeng et al. 2013);

Validation methods: This consists in testing a grid of values for different hyper-parameters in the model and choosing the values that maximize predictive performance. However, this strategy may be unfeasible with a large number of parameters to be defined. As previously mentioned, one way to define the bandwidth parameter in semi-parametric RKHS model is by using the validation approach (Gianola et al. 2006; Gianola and van Kaam 2008);

Full Bayesian treatment: This method basically considers the unknown hyper-parameter as regular parameters, thus increasing the hierarchy in the model. As mentioned before, (Gianola et al. 2009) suggested the definition of one more hierarchical level in the original Bayes A and B (Meuwissen et al. 2001) for dealing with Bayesian

learning problems, and it was considered in BayesA and BayesB, also previously mentioned.

Empirical Bayes methods: This approach suggests the replacement of prior distribution by estimation of parameter using data. It is similar to the common breeding value evaluation, where the unknown genetic variance parameter is replaced by its estimated values from REML procedure (de los Campos et al. 2013).

The hyper-parameters for all methods demonstrated heren are completely explained based on Heritability-based rules and Full Bayesian treatment in Pérez and de los Campos (2014). An alternative for choosing the hyper-parameters in the additive-dominante WGR as provided in Zeng et al. (2013).

Due to the lack of knowledge about genetic architecture of traits in literature, the same hyper-parameter is generally assumed for all marker effects or theirs variance component. However, it drives all markers to the same marginal prior even if the method assumes a heterogeneous variance component (Park and Casella 2008; Gianola et al. 2009). Maybe with the improvement on the knowledge of genetic architecture, from the expression or association studies, markers of some loci could have higher prior genetic variance, or the inclusion of information from previous studies about genetic architecture could make statistic models more realistic.

CHAPTER II

THE CONTRIBUTION OF DOMINANCE TO PHENOTYPE PREDICTION IN A PINE BREEDING AND SIMULATED POPULATION

ABSTRACT

Pedigrees and dense marker panels have been used to predict the genetic merit of individuals in plant and animal breeding, accounting primarily for the contribution of additive effects. However, non-additive effects may also impact trait variation in many breeding systems, particularly when specific combining ability is explored. Here we used models with different priors, and including additive-only and additive plus dominance effects, to predict polygenic (height) and oligogenic (fusiform rust resistance) traits in a structured breeding population of loblolly pine (*Pinus taeda* L.). Models were largely similar in predictive ability, and the inclusion of dominance only improved modestly the predictions for tree height. Next, we simulated a genetically similar population to assess the ability of predicting polygenic and oligogenic traits controlled by different levels of dominance. The simulation showed an overall decrease in the accuracy of total genomic predictions as dominance increases, regardless of the method used for prediction. Thus, dominance effects may not be accounted for as effectively in prediction models, compared to traits controlled by additive alleles only. When the ratio of dominance to total phenotypic variance reached 0.2, the additive-dominance prediction models were significantly better than the additive-only models. However, in the prediction of the subsequent progeny population, this accuracy increase was only observed for the oligogenic trait.

Keywords: genomic prediction, dominance, additive, polygenic, oligogenic

---

Chapter accepted for publication in Heredity

INTRODUCTION

Genomic prediction of complex traits can increase genetic gains per unit of time in plant and animal breeding, by allowing early and more accurate selection than traditional approaches (Heffner *et al.*, 2010; Wiggans *et al.*, 2011; Resende, *et al.*, 2012b). In human genetics, the same methods may be applicable to predict propensity to disease, and response to drug treatments (Yang *et al.* 2010; de los Campos *et al.* 2010; Wray *et al.* 2013). Most of the early development of genomic prediction methods occurred in dairy cattle, with the aim of selecting sires with high breeding value. Thus, prediction models were developed to account for the contribution of additive effects to phenotypic traits, while non-additive effects were typically not considered. Considering non-additive effects in the model could improve predictions, as the genetic architecture of traits is a factor that contributes to the accuracy of models (Hayes et al. 2009). In addition, dominance and epistasis may be confused with the additive effect in genomic predictions. Thus, their specific contribution should be accounted for to avoid the overestimation of genetic parameters in downstream applications (Muñoz et al. 2014).

Prediction of dominance effects is needed in advanced breeding programs that explore specific combining ability (SCA). In those programs, seeds from a small number of crosses known to have superior SCA can be scaled up through controlled mass pollination and deployed in large-scale (White et al. 2007). When dominance contributes to the complex trait, these strategies increase the yield and genetic gain when compared with half-sib, open-pollinated families (McKeand et al. 2006). Recent studies in plants and animals have reported a significant contribution from non-additive effects to phenotypes, adding to a considerable proportion of the genetic variance and improving the accuracy of predictions (Su et al. 2012; Vitezica et al. 2013; Nishio and Satoh 2014; Muñoz et al. 2014). Analysis of simulated data indicated that including dominance is recommended to achieve higher genetic gains in crossbred population (Zeng et al. 2013) and would also allow the application of mate-allocation (Toro and Varona 2010; Sun et al. 2013; Ertl et al. 2014). When only additive effects are considered, predicting the best combination of parents that generate superior families equals the average of their breeding values. Thus, inclusion of dominance is critical to identify complementary individuals and explore heterosis. In species as pine, the additive and non-aditive effects prediction are also

important for clonal propagation, since in this case are explored the whole genotypic value.

Numerous whole-genome regression (WGR) approaches have been proposed for genomic prediction of additive effects. These approaches generally share the same linear model but differ in their assumptions regarding the prior information of markers effects (Gianola 2013; de los Campos et al. 2013). For instance, priors implemented in Bayesian Ridge Regression assume that marker effects follow a normal distribution with a common variance component. This assumption is suitable under the infinitesimal model, where the trait is controlled by a large number of genes with small effect. Others models implement more complex (parameterized) priors that can fit traits with major-effect genes that explain a significant proportion of the genetic variation. These models rely on variable selection (e.g. BayesB) to remove markers that are not in linkage disequilibrium with any quantitative trait loci (QTL), and modeling variance heterogeneity of marker effect (e.g. Bayes A, BayesB, Bayesian Lasso) that assumes that each marker explains a distinct part of genotypic variation. In polygenic traits it was previously observed that the different WGR models and priors usually result in similar accuracies (Heslot *et al.*, 2012; Resende, *et al.*, 2012a; Pérez *et al.*, 2012). However, when WGR was applied to traits that are expected to be oligogenic, such as rust resistance (Resende, *et al.*, 2012a) and milk fat (Habier et al. 2013), the accuracies were superior under priors that assume variable selection, variance heterogeneity or both.

Despite the relevance of different priors in the performance of additive whole-genome prediction models, their contribution to the accuracy of models that incorporate dominance effects, and for traits with distinct genetic architecture, have not been extensively explored. The objective of this study is to address this limitation. We evaluate additive and additive-dominance models in the prediction of traits with a relatively simple (disease resistance) and complex (growth) genetic architecture, measured in a standard breeding population of loblolly pine (Resende, *et al.*, 2012a). Furthermore, to fully explore the advantages and limitations of different models in the prediction of dominance, we extend the analysis to a simulated population with traits controlled by contrasting levels of dominance.

MATERIALS AND METHODS

Loblolly pine population data

The reference loblolly pine (*Pinus taeda* L.) breeding population CCLONES (Comparing Clonal Lines On Experimental Sites) was used in this study. The population was created by crossing 42 parents representing a wide range of accessions from the US Atlantic coastal plain, in a circular mating design with additional off-diagonal crosses (Baltunis et al. 2007). In total, 923 individuals from 71 full sibs families (average of 13 individuals per family, SD=5) were genotyped for 7,216 single-nucleotide polymorphism (SNP) loci using an Illumina Infinium assay (Illumina, San Diego, CA; Eckert *et al.*, 2010). All 4,722 loci that were polymorphic in the population were used in this study, regardless of their minor allele frequency. Missing data was low (<1%) and missing values were replaced by the marker expected value (de los Campos and Perez 2014). Three traits with contrasting genetic architecture were analyzed. Tree height (HT) is a polygenic trait, and was measured in field trials at Nassau (Florida, USA), when the trees were six years old, in eight clonal replicates distributed in an alpha-lattice design (Baltunis et al. 2007). Fusiform rust is an oligogenic trait, controlled by a number of loci of large effect (Resende, *et al.*, 2012a). Fusiform rust incidence was measured as gall volume (RFgall) and as a binary (presence/absence) trait (RFbin) (Quesada *et al*., 2014). Plants were phenotyped for rust in a greenhouse experiment that followed a randomized complete block design, with three repetitions allocated with alpha design, as described previously (Resende, *et al.*, 2012a). The estimated narrow sense heritability of these traits was previously reported as 0.31, 0.21 and 0.12 for HT, RFbin and RFgall, respectively (Resende, *et al.*, 2012a).

Simulated Data

The parametric contribution of dominance to trait variation, and the ratio of dominance to additive effects are unknown in the CCLONES population. In order to fully evaluate the ability of models in predicting dominance effects of different architectures and degrees we proceeded to simulate a population with similar genetic properties as CCLONES, except that trait QTL were manipulated to include dominance, and regulation by different numbers of loci. The simulation of a population with similar properties as CCLONES was carried out in two steps. First, 1,000 diploid individuals were created by

randomly sampling 2,000 haplotypes generated after 1,000 generations of a neutral coalescence model from a population with effective size (N_e) of 10,000 and mutation rate of 2.5×10^{-8} (Willyard et al. 2007). The simulated genome had 12 chromosomes, each with 100 cM, and 10,000 polymorphic loci were randomly selected. This first step was simulated using Macs (Chen et al. 2009). In the second step of the simulation, the 1,000 diploid individuals generated previously were subject to selection and recombination, and used to generate a loblolly pine improvement program in its second breeding cycle (Figure 1). The simulation of the population generated a total of 196,303,656 polymorphic sites. As commonly observed in pine tree breeding populations, the majority of loci had very low minor allele frequencies (Supplementary Figure S1).

Six traits with different genetic architectures (polygenic and oligogenic) and levels of dominance (none, medium or high dominance) were simulated. For the polygenic traits, 1,000 QTL were used in the analysis, and their additive effects were sampled from a standard normal distribution (Hickey and Gorjanc 2012). For the oligogenic traits, 30 QTL were sampled from a gamma distribution with rate 1.66 and shape 0.4, and the QTL effects were sampled to be positive or negative with equal probability (Meuwissen et al. 2001). The dominance effect of the i^{th} QTL, when present, were determined by: $d_i = a_i \times \varphi_i$, where φ_i was sampled from a normal distribution with mean zero and standard deviation of 1 (moderate dominance) and 2 (high dominance) (Table 1). The additive effect (a_i) of the i^{th} QTL was defined as half of the difference between alternative homozygote categories, and the dominance effect (d_i) as the deviation of the heterozygote from the mean of two homozygote classes. The heritability was calculated as $h^2 = V_A/V_P$, and $d^2 = V_D/V_P$, where $V_P = V_A + V_D + V_E$ (additive-dominance scenario) or $V_P = V_A + V_E$ (additive scenario). V_P, V_A, V_D and V_E are the phenotypic, additive, dominance deviation and residual variances, respectively (Falconer and Mackay 1996). The error was simulated from a normal distribution with mean zero, and the variance was defined to result in an h^2 equal to 0.25. The simulation of dominance traits was supervised in order to achieve a d^2 of 0.1 and 0.2 for traits with moderate and large dominance effects, respectively. For traits with moderate dominance, we accepted d^2 between 0.09 and 0.11; for traits with large

dominance, we accepted d^2 between 0.19 and 0.21. When d^2 fell outside the desired range the simulation was discarded.

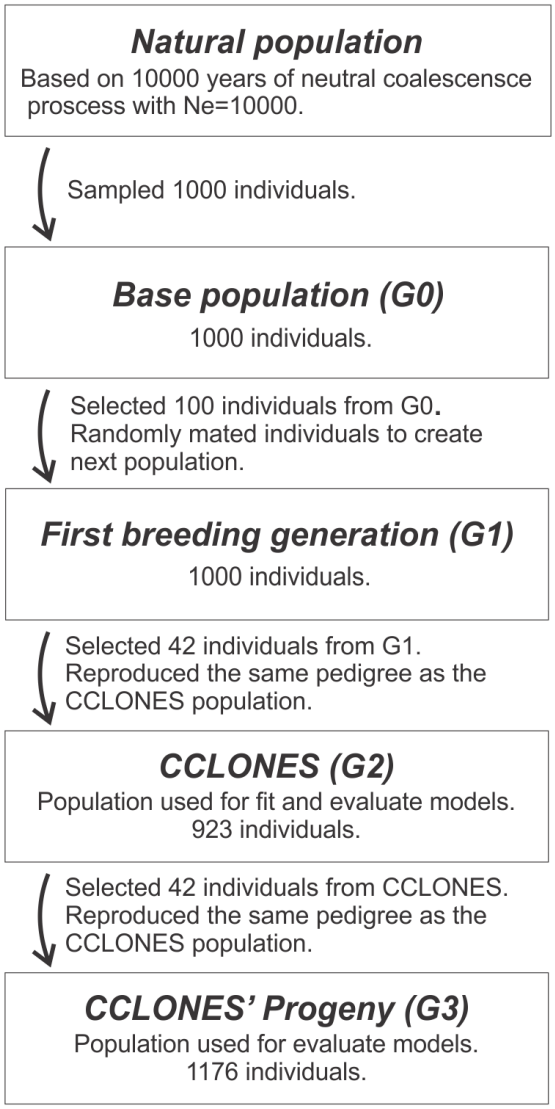

Figure 1. Breeding scheme applied to create the simulated CCLONES population used for analysis of all traits.

After sampling individuals from the natural population and creating the base population (G0), two discrete generations of selection and mating were simulated. From 1,000 individuals in the base population (G0), the 10% highest phenotypic values were selected and randomly mated to generate 1,000 individuals that compose the first breeding generation (G1). From G1, 42 individuals were selected and used in a mating

design that reproduced the same pedigree as the CCLONES population (G2). The breeding populations from G2 were simulated with 10 replicates for each trait using the R software (R Core Team 2014). In addition, the 42 individuals with highest phenotypic value from each replicate of G2 were selected to be parents in the subsequent generation (G3). The mating followed again the same design as CCLONES and the top selected individuals were randomly crossed. In G2 were considered the 923 individuals that correspond the genotyped individuals of real CCLONES population, but in G3 were considered all individuals, including those that were not genotyped in the real population.

Table 1. Summary of simulated traits. A heritability of 0.25 was used in all simulated conditions.

Traits description	Number of genes (QTL)	d^2	d^2/h^2
Oligogenic with no dominance	30	0	0
Polygenic with no dominance	1,000	0	0
Oligogenic with medium dominance	30	0.1	0.4
Polygenic with medium dominance	1,000	0.1	0.4
Oligogenic with high dominance	30	0.2	0.8
Polygenic with high dominance	1,000	0.2	0.8

Statistical methods

We used Bayesian WGR models with SNPs as covariates and common priors, including Bayesian ridge regression (BRR, also called SNP-BLUP), BayesA, BayesB, and Bayesian Lasso (BL). All methods used here can be represented by the following base model:

$$y_j = \mu + g_j + e_j$$

Where y_j is the phenotype (clonal mean) of individual j; μ is the intercept; e_j is the error of observation j; g_j is the genotypic value. In all models it was assumed that:

$$y_j|\mu + g_j, \sigma_e^2 \sim IID\ N(\mu + g_j, \sigma_e^2);$$

$$\mu \sim N(0, 10^6);$$

$$e_j|\sigma_e^2 \sim IID\ N(0, \sigma_e^2);$$

$$\sigma_e^2|\nu_e, S_e \sim \chi^{-2}(\nu_e, S_e).$$

For each prior either additive only or additive-dominance effects were considered. Thus, the general additive-dominance whole-genome regression model was replaced by:

$$y_j = \mu + \sum_{i=1}^{k} \left(x_{ij} a_i + w_{ij} d_i \right) + e_j$$

Where k is the number of markers, x_{ij} and w_{ij} are the functions of SNP *i* in individual *j*, for genotypes AA, Aa and aa. We parameterized xij with values 1 (AA), 0 (Aa) and -1 (aa) and w_{ij} with 0 (AA), 1 (Aa) and 0 (aa) (Toro and Varona 2010). The additive and dominance effects of the i^{th} marker were represented by a_i and d_i, respectively. The dominance effect was fitted only in the additive-dominance model. The priors used in linear regression coefficients for additive-dominance and additive models are described below.

Bayesian Ridge Regression (BRR)

The BRR is a Bayesian method in which it is assumed that all regression coefficients have common variance. Thus, for an additive-dominance model, all markers with the same allele frequency explain the same proportion of the additive and dominance variances, and have the same shrinkage effect (Gianola 2013). For BRR it was assumed that:

$a_i|\sigma_a^2 \sim N(0, \sigma_a^2)$; $\sigma_a^2|\nu_a, S_a \sim \chi^{-2}(\upsilon_a, S_a)$;$d_i|\sigma_d^2 \sim N(0, \sigma_d^2)$; $\sigma_d^2|\nu_d, S_d \sim \chi^{-2}(\nu_d, S_d)$.

Bayes A

Bayes A was proposed by Meuwissen et al. (2001) and, contrary to BRR, it considers that markers have heterogeneous variances. Bayes A was further modified (de los Campos and Perez 2014) to estimate the shape parameter of the inverted chi-square distribution. This modification is expected to reduce the influence of the hyperparameter and improve the learning process (Gianola et al. 2009). For Bayes A it was assumed that:

$a_i|\sigma_{a_i}^2 \sim N(0, \sigma_{a_i}^2)$; $\sigma_{a_i}^2|\nu_a, S_a \sim \chi^{-2}(\nu_a, S_a)$; $S_a|r_a, s_a \sim G(r_a, s_a)$;

$d_i|\sigma_{d_i}^2 \sim N(0, \sigma_{d_i}^2)$; $\sigma_{d_i}^2|\nu_d, S_d \sim \chi^{-2}(\nu_d, S_d)$; $S_d|r_d, s_d \sim G(r_d, s_d)$.

Bayes B

Bayes B differs from Bayes A in that it includes the selection of covariates (SNPs) that don't contribute to genetic variance (Meuwissen et al. 2001). Similarly to Bayes A, we adopted a modified version of Bayes B (de los Campos and Perez 2014), where the shape parameter follows a gamma distribution and π is an estimated parameter (Gianola et al. 2009). This implementation of Bayes B is very similar to Bayes Dπ (Habier et al. 2011), and it assumes:

$$a_i|\sigma^2_{a_i} = \begin{cases} 0 & \text{with probability } \pi_a \\ \sim N(0, \sigma^2_{a_i}) & \text{with probability } 1\text{-}\pi_a \end{cases}$$

$$d_i|\sigma^2_{d_i} = \begin{cases} 0 & \text{with probability } \pi_d \\ \sim N(0, \sigma^2_{d_i}) & \text{with probability } 1\text{-}\pi_d \end{cases}$$

$\sigma^2_{a_i}|\nu_a, S_a \sim \chi^{-2}(\nu_a, S_a)$; $\sigma^2_{d_i}|\nu_d, S_d \sim \chi^{-2}(\nu_d, S_d)$; $S_a|r_a, s_a \sim G(r_a, s_a)$; $S_d|r_d, s_d \sim G(r_d, s_d)$; $\pi_a|p_0, \pi_0$ and $\pi_d|p_0, \pi_0 \sim Beta(p_0, \pi_0)$

Bayesian Lasso (BL)

The Bayesian version of Lasso regression was proposed by Park and Casella (2008), and the application in whole genomic prediction was proposed by de los Campos et al. (2009). As in Bayes A and Bayes B, BL presupposes that covariates do not have homogeneous variance. Furthermore, it promotes an indirect marker selection with strong shrinkage in the regression coefficients, since the marginal prior of regression coefficients follows a double exponential distribution (Park and Casella 2008), which drive many marker effects to zero or near zero. The BL assumes:

$a_i|\tau^2_{a_i}, \sigma^2_e \sim N(0, \tau^2_{a_i}\sigma^2_e)$; $d_i|\tau^2_{d_i}, \sigma^2_e \sim N(0, \tau^2_{d_i}\sigma^2_e)$; $\tau^2_{a_i}|\lambda_a \sim Exp(0.5\lambda^2_a)$; $\tau^2_{d_i}|\lambda_d \sim Exp(0.5\lambda^2_d)$; $\lambda_a|r_a, s_a \sim G(r_a, s_a)$ and $\lambda_d|r_d, s_d \sim G(r_d, s_d)$

All analysis with the WGR models were carried out with the R package BGLR (de los Campos and Perez 2014) with default hyperparameter (Supplementary Table S1 and S2) values described previously (de los Campos et al. 2013; de los Campos and Perez 2014; Pérez and de los Campos 2014). In total 30,000 MCMC iterations were used, of which the first 10,000 were discarded as burn-in and every 3rd sample was kept for parameter estimation. We also evaluated the accuracy of additive and additive-dominance

models based exclusively on pedigree information by generating the expected relationship matrix. Although the additive-dominance pedigree model was more accurate for dominance deviation, the genomic models were more accurate for parent and clonal selection. Thus, this study focused on genomic prediction models only (Supplementary Table S3 and S4).

Breeding value and dominance deviation

After fitting each WGR model, the breeding values (u) and dominance deviation of the additive-dominance models (δ) were estimated (Falconer and Mackay 1996) as described below.

$$\hat{u}_j = \sum_i \left[I(x_{ij} = 1) 2q_i + I(x_{ij} = 0)(q_i - p_i) - I(x_{ij} = -1) 2p_i \right] \hat{\alpha}_i$$

$$\hat{\delta}_j = \sum_i \left[-I(x_{ij} = 1) 2q_i^2 + I(x_{ij} = 0) 2p_i q_i - I(x_{ij} = -1) 2p_i^2 \right] \hat{d}_i$$

Where p_i is allele frequency of allele A of SNP *i*, q_i=*1-*p_i, $\hat{\alpha}_i$ is the average effect of substitution, $\hat{\alpha}_i = \hat{a}_i + \hat{d}_i(q_i - p_i)$, and I is an indicator function of SNPs.

Variance components and heritability estimation

For estimation of variance components, linkage equilibrium, absence of epistasis and Hardy-Weinberg equilibrium was assumed (Gianola et al. 2009). Considering these assumptions, the additive variance (σ_A^2) and the variance due to dominance deviation (σ_D^2) were estimated as described previously (Zeng et al. 2013; Ertl et al. 2014):

$$\hat{\sigma}_A^2 = 2 \sum_i p_i q_i \left[\hat{\sigma}_a^2 + (q_i - p_i)^2 \hat{\sigma}_d^2 \right]$$

and

$$\hat{\sigma}_D^2 = 4 \sum_i (p_i q_i)^2 \, \hat{\sigma}_d^2$$

These estimates were used to calculate h^2 and d^2, as previously described.

Validation

A 10-fold cross-validation was used to compare results in the real and simulated populations (Ertl et al. 2014). Briefly, the dataset was separated into ten subsets. In each

cycle, a subset was excluded before models were fitted with the remaining data, and the model was used to predict the excluded subset. The process was repeated ten times, and in each cycle the prediction accuracy was estimated (Pearson's correlation) of parametric values on predicted validation data were calculated. For the simulated population, the accuracies were calculated for breeding values, dominance deviations, total genotypic values and phenotype values of individuals. The results reported are means (and standard errors) of accuracies of parametric values on estimated values across folds. Because in the non-simulated population the true genotypic values are unknown, we used the prediction ability (accuracy of phenotype prediction $r_{y\hat{y}}$), which is the correlation between predicted whole genotypic value and phenotype.

RESULTS

Heritability

Bayesian ridge regression was used to estimate the narrow sense heritability using additive and additive-dominance models. Estimates of h^2 were higher in additive models, for all traits, in the real and the simulated population (Table S5 and S6). For traits measured in the real population, estimates of d^2 ranged from 0.09 to 0.15, while σ_D^2/σ_A^2 (or d^2/h^2) varied from 0.31 to 0.42. Because the parametric values are known in the simulated population, it was possible to evaluate the impact of model selection in the estimation of genetic parameters. For traits without dominance, the estimates of h^2 were similar to the parametric value for additive- and additive-dominance models. The dominance component of the additive-dominance model captured dominance variability and overestimated d^2 as 0.07. For simulated traits with low dominance (d^2=0.1), estimates of d^2 and h^2 were similar to the parametric value. However, in the case of higher dominance (d^2=0.2), these estimates were underestimated for d^2 and modestly overestimated for h^2.

Additive and additive-dominance model prediction in the CCLONES population

We contrasted the predictive ability of linear models with different assumptions regarding prior information of marker effects, and accounting for only additive, or additive-dominance contributions. The models with different prior were similar in absolute value of the predictive ability (Table 2). However, an analysis of variance indicated that the results

were statistically different for HT and RFbin (Supplementary Table S7). The inclusion of dominance effects only increased modestly the predictive ability for HT. For instance, additive Bayes B showed the highest accuracies for RFgall (0.299) and RFbin (0.376). In contrast, the highest accuracies with additive-dominance models were 0.292 and 0.369 for RFgall and RFbin, respectively (Table 2). These results suggest a minor contribution of dominance to tree height. On the other hand, prediction of rust resistance traits show no improvement in accuracy when dominance is considered, possibly because this effect is absent or negligible. Other factors, such as limited marker coverage of rust QTL or insufficient population size to estimate the dominance effect, may have also contributed to the observed results. Overall the results are in agreement with the proportion of variance of dominant deviations relative to total genetic variance, which was estimated to be 50% higher for HT, compared to RFgall and RFbin (Supplementary Table S5).

Genetic properties of the simulated population

To assess the effect of the trait genetic architecture on prediction models that include additive and additive-dominance effects, scenarios considering a polygenic trait (1,000 QTL) and an oligogenic trait (30 QTL) were evaluated. For both types of traits three dominance levels were simulated: no dominance (d^2=0; d^2/h^2=0), moderate dominance (d^2=0.1; d^2/h^2=0.4), and high dominance (d^2=0.2; d^2/h^2=0.8). A set of 10,000 markers randomly distributed across the genome (expected 8.33 markers per cM), and polymorphic in the base population were included in the analysis. In the population that simulated CCLONES (G2), approximately half of QTL (mean=53.92% SD=1.18%) and markers (mean= 55.45% SD=0.56%) were fixed (Supplementary Figure S1). Thus, the two cycles of breeding and selection reduced (or fixed) the frequency of alleles in a large number of loci. The allele frequency distributions of polymorphic SNPs were similar between CCLONES and the simulated population (Supplementary Figure S1). In the simulated base population, the LD among markers and QTL was low. As expected, the LD increased over successive generations, reflecting the lower effective population size relative to the base population (Supplementary Figure S2). On average, two or more markers had an r^2 higher than 0.4 with any QTL for all simulated traits.

Table 2 results of predict ability of whole-genome regressions using different priors and including dominance effects for height (HT) and rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in *Pinus taeda*.

Model	Prior	HT		RFgall		RFbin	
		$r_{y\hat{y}}$	(se)	$r_{y\hat{y}}$	(se)	$r_{y\hat{y}}$	(se)
add-dom	Bayesa	0.415^{ab}	(0.04)	0.291^{a}	(0.03)	0.367^{ab}	(0.02)
	BayesB	0.414^{ab}	(0.04)	0.291^{a}	(0.03)	0.369^{a}	(0.02)
	BL	0.415^{ab}	(0.04)	0.288^{a}	(0.03)	0.338^{c}	(0.02)
	BRR	0.418^{a}	(0.04)	0.292^{a}	(0.03)	0.329^{c}	(0.02)
additive	BayesA	0.401^{bc}	(0.03)	0.296^{a}	(0.03)	0.375^{a}	(0.02)
	BayesB	0.401^{bc}	(0.03)	0.299^{a}	(0.03)	0.376^{a}	(0.02)
	BL	0.392^{bc}	(0.03)	0.292^{a}	(0.03)	0.345^{bc}	(0.02)
	BRR	0.402 ^{abc}	(0.03)	0.291^{a}	(0.03)	0.336^{c}	(0.02)

Average of predict ability with same letter are statistically equal by tukey test. all inferences used type 1 error=0.05.

Genetic properties of the simulated population

To assess the effect of the trait genetic architecture on prediction models that include additive and additive-dominance effects, scenarios considering a polygenic trait (1,000 QTL) and an oligogenic trait (30 QTL) were evaluated. For both types of traits three dominance levels were simulated: no dominance (d^2=0; d^2/h^2=0), moderate dominance (d^2=0.1; d^2/h^2=0.4), and high dominance (d^2=0.2; d^2/h^2=0.8). A set of 10,000 markers randomly distributed across the genome (expected 8.33 markers per cM), and polymorphic in the base population were included in the analysis. In the population that simulated CCLONES (G2), approximately half of QTL (mean=53.92% SD=1.18%) and markers (mean= 55.45% SD=0.56%) were fixed (Supplementary Figure S1). Thus, the two cycles of breeding and selection reduced (or fixed) the frequency of alleles in a large number of loci. The allele frequency distributions of polymorphic SNPs were similar between CCLONES and the simulated population (Supplementary Figure S1). In the simulated base population, the LD among markers and QTL was low. As expected, the LD increased over successive generations, reflecting the lower effective population size relative to the base population (Supplementary Figure S2). On average, two or more markers had an r^2 higher than 0.4 with any QTL for all simulated traits.

Dominance reduces the overall accuracy of prediction models

The suitability of additive and additive-dominance prediction models was assessed by estimating the total genomic accuracy (Figure 2), breeding value (Figure 3), dominance deviation (Figure 4) and phenotypic accuracy (Supplementary Figure S3). In all scenarios, the different WGR provided statically different results (Supplementary Table S8-11). Overall there was a decrease in the accuracy of total genomic predictions as the dominance increased, regardless of the method used for model development. Thus, the data indicates that dominance effects may not be accounted for as effectively in the prediction models, compare to traits controlled by loci that contribute additive effects only.

Models that incorporate dominance are only more accurate when $\mathbf{d^2}$ ***is high***

In the simulated population we detected a very small (mostly non-significant) improvement in accuracy of genomic prediction from additive-dominance models, when d^2 was equal to 0.1 (Figure 2). A much larger and significant improvement was only observed as d^2 increased to 0.2, a relatively high dominance to additive effect ratio. The standard errors were generally higher among oligogenic traits, compared to polygenic traits. This difference was accentuated when dominance was high. This may occur because the oligogenic architecture can exacerbate the inaccuracy in the estimation of dominance. Random sampling of individuals from the population in the cross validation can result in sub-samples with different representations of heterozygous individuals between the training and validation sub-populations.

The accuracy of the total genomic prediction was similar across different methods for polygenic traits, regardless of the presence of dominance (Figure 2). However, BayesA and BayesB had higher accuracy than BL and BRR, for oligogenic traits in all scenarios. This observation is similar to previous reports (Resende, *et al.*, 2012a; Daetwyler *et al.*, 2013) that have shown the limitation of BL and RR-BLUP (frequentist version of BRR) in accounting for few loci of large effect in the predictive model. It suggests that, when the trait architecture is unknown, it may be suitable to evaluate multiple models before adoption of one approach for trait prediction in future generations.

Accuracy of predicting additive and dominance effects, and phenotypes

The inclusion of dominance in the prediction model did not affect the prediction of breeding values, as expected (Figure 3). There was no difference among models in the accuracy of prediction of additive effects in polygenic traits. However, similarly to the prediction of total genetic effects, a significant improvement was detected when BayesA and BayesB were used for prediction of oligogenic traits, over BL and BRR.

The accuracy of dominance prediction improved significantly (over 50%) when its contribution to traits increased from d^2=0.1 to 0.2 (Figure 4). Thus, as the contribution of dominance is higher, the ability to accurately capture it in prediction models improves. However, the overall genetic accuracy decreases as the d^2 increases, as those effects may not be estimated adequately. Accuracies were observed to be more accurate for oligogenic traits predicted with BayesA and BayesB models.

Finally, the accuracy derived by the correlation of phenotypes to the estimated genetic effect (Supplementary Figure S3) showed that, as dominance increases in oligogenic and polygenic traits, accuracy of phenotype prediction also increases. As d^2 increased from 0 to 0.2, the prediction accuracy improved 22%. However, there is only a significant difference in the prediction using the additive-dominance model, when d^2 is 0.2. We expect this difference to increase as dominance increases.

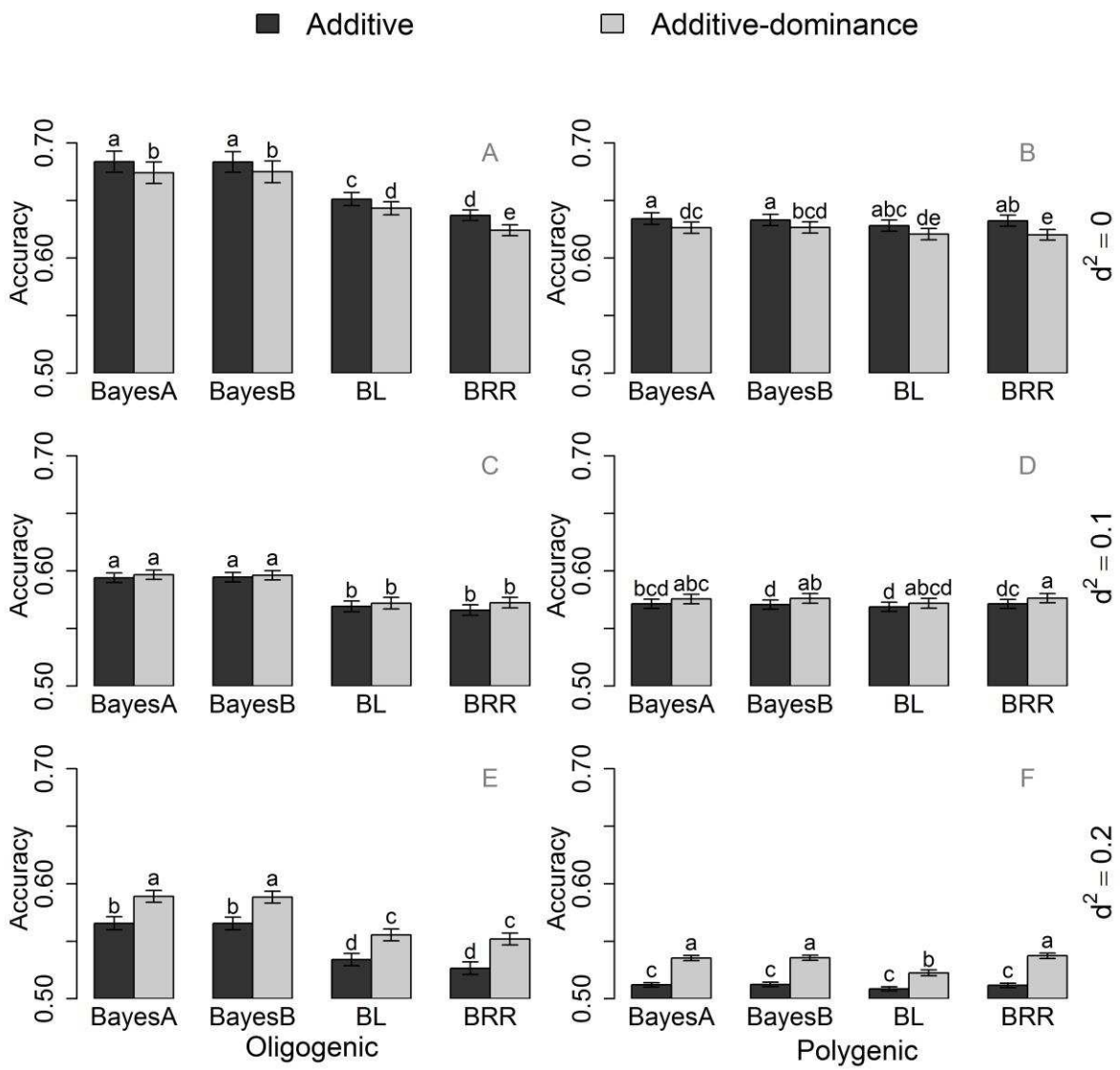

Figure 2. Total genetic accuracies of whole genotypic predictions with additive and additive-dominance WGRs using different priors for six different simulated traits: A and B oligogenic and polygenic respectively traits with h^2=0.25 and non-dominance effects B; C and D oligogenic and polygenic respectively trait with h^2=0.25 and d^2=0.1; E and F oligogenic and polygenic respectively trait with h^2=0.25 and d^2=0.2. Error bars are standard error among 10 replicates. Means with same letter are statistically equal by Tukey test ($p<0.05$)

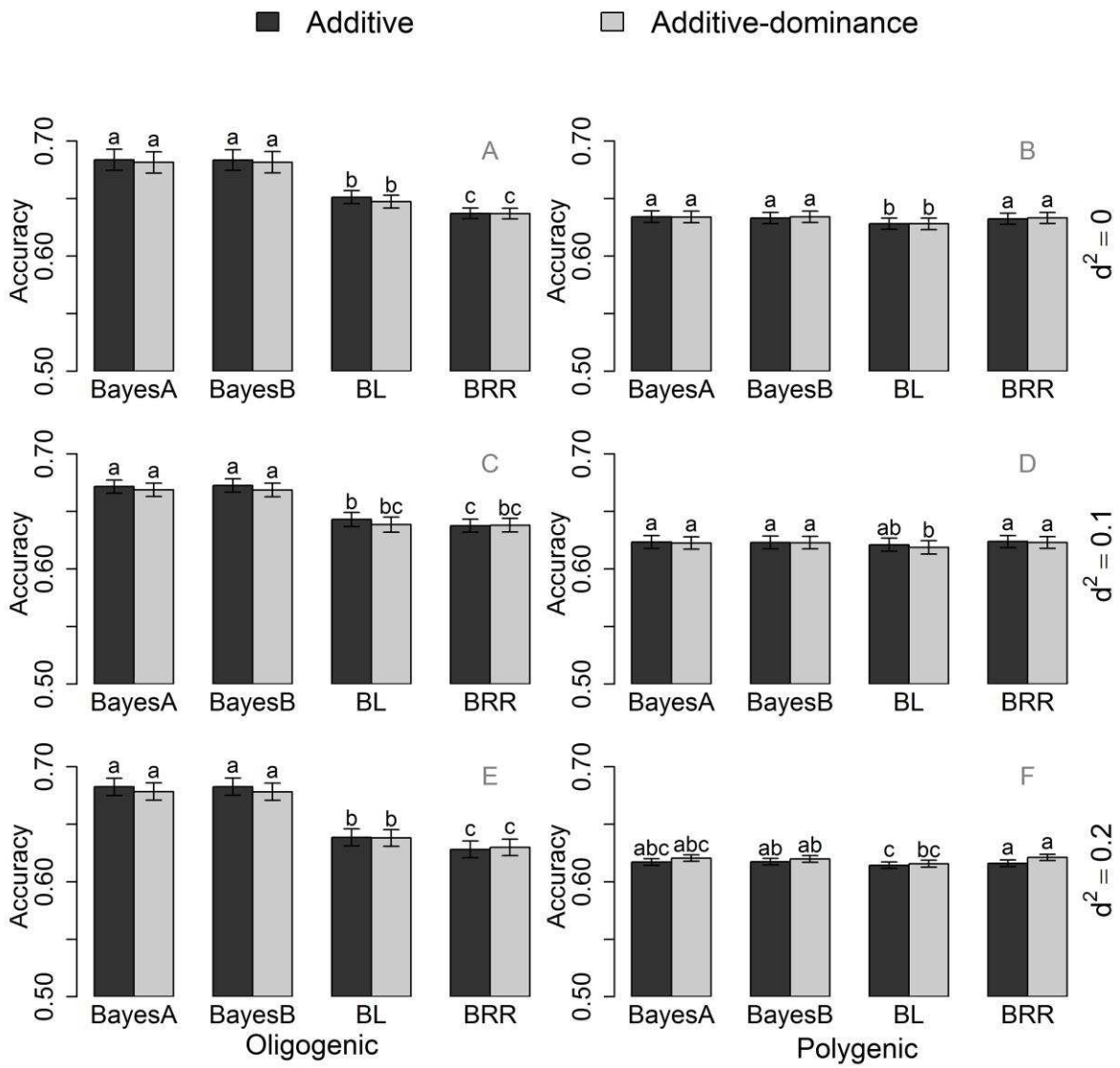

Figure 3. Accuracies of breeding values predictions with additive and additive-dominance WGRs using different priors for six different simulated traits: A and B oligogenic and polygenic respectively traits with h^2=0.25 and non-dominance effects B; C and D oligogenic and polygenic respectively trait with h^2=0.25 and d^2=0.1; E and F oligogenic and polygenic respectively trait with h^2=0.25 and d^2=0.2. Error bars are standard error among 10 replicates. Means with same letter are statistically equal by Tukey test ($p<0.05$).

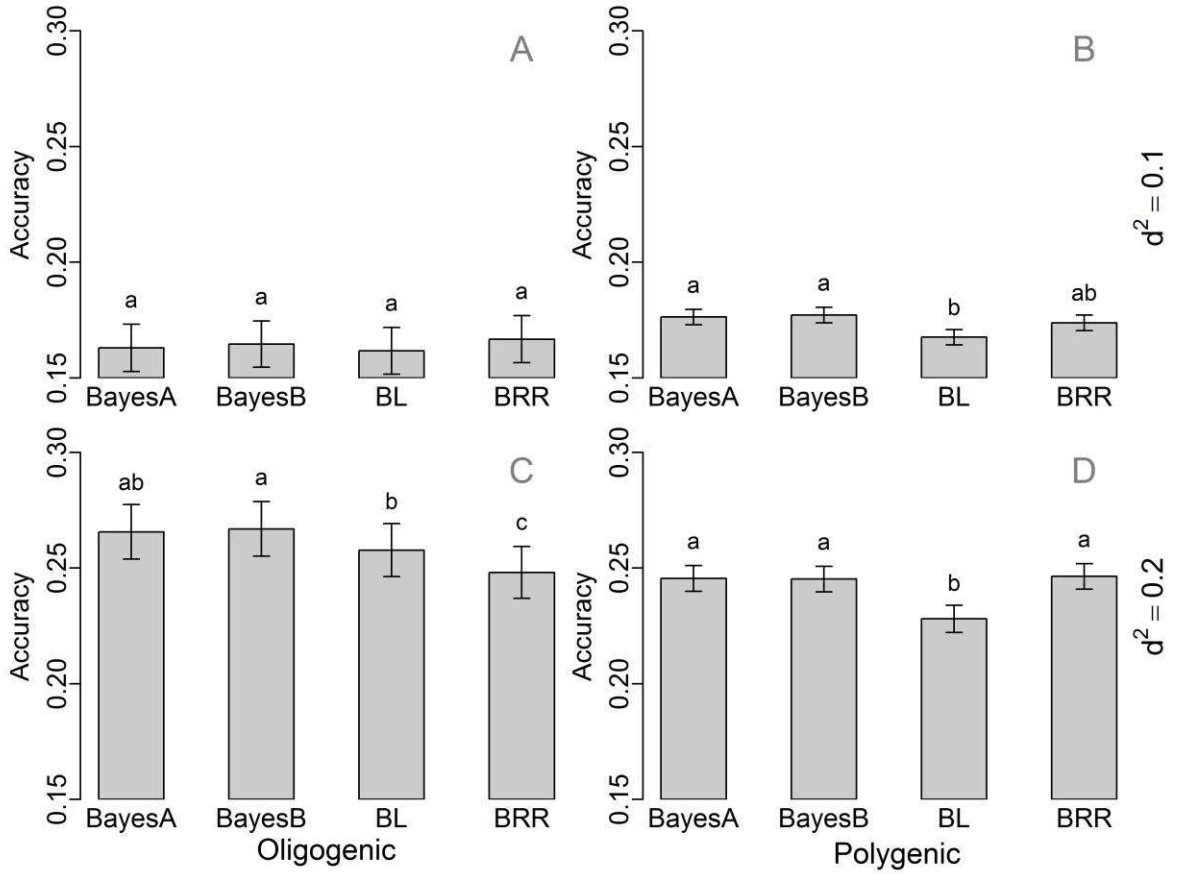

Figure 4. Dominance effect accuracies for dominance deviation predictions with additive-dominance WGRs using different priors for four different simulated traits. A and B are oligogenic and polygenic traits, respectively, with h^2=0.25 and d^2=0.1; C and D are oligogenic and polygenic traits, respectively, with h^2=0.25 and d^2=0.2. Error bars are standard error among 10 replicates. Means with same letter are statistically equal by Tukey test ($p<0.05$).

Additive-dominance models improve accuracy of progeny selection only for oligogenic traits with high dominance

Progeny derived from the real CCLONES population are currently not available, preventing the evaluation of prediction models in generations following the population used for model estimation. However, such progeny can be generated for the simulated population. The first generation (G3) derived from the simulated CCLONES population

was generated by selecting 42 individuals with the highest phenotypic value, which were crossed following the same matting design as CCLONES. The results showed that the accuracy of the prediction in the next generation (Suplementary Figure S4) decreased significantly, when compared to the accuracy in the CCLONES (G2) population (Figure 2-4 and Suplementary Figure S3). The accuracy of the prediction of dominance deviation was almost zero for all characteristics, except for oligogenic trait with high dominance. In all other traits the additive models provided better predictions.

DISCUSSION

Dominance was formulated by Mendel as one of first concepts of genetics (Wilkie 1994). In quantitative genetics, dominance is defined as the interaction between different alleles of a gene, and is measured as the difference of heterozygotes and mean of homozygotes (Falconer and Mackay 1996). Dominance effects contribute to inbreeding depression, and may also play a role in heterosis (or hybrid vigor) (Falconer and Mackay 1996; Hallauer et al. 2010). Expectedly, the presence of dominance is dependent on the trait under consideration, and allele frequencies in the population. Here we analyzed the contribution of dominance effects in the accuracy of genomic prediction, with models that assume different priors, and for traits with different genetic architectures. The assessment was made for traits measured in the reference CCLONES population of loblolly pine, which was previously genotyped and extensively phenotyped for height growth and rust resistance. Next we extended the analysis to a simulated population with similar genetic properties to CCLONES, where traits with different genetic architectures and degrees of dominance were considered. In this study, additive and dominance effects were simultaneously adjusted in genomic prediction models. Epistasis, however, was not considered in the model. Hence, the presence of any epistatic effect could have acted as a confounding effect and affect prediction accuracy.

Previous quantitative genetic analysis of height measured in pine breeding populations indicated that the trait is highly polygenic, and that non-additive effects contribute to its variance (Isik et al. 2003; Muñoz et al. 2014). In the analysis of height measured in the CCLONES population, models that accounted for both additive and dominance effects had higher predictive ability. The analysis of the simulated population

supports these results, as polygenic traits with dominance effects were predicted with significantly higher accuracy in models that included additive and dominance effects. Previous analysis of complex traits reported that inclusion of dominance (and epistasis in some cases) was advantageous for breeding programs, when compared to using models that accounted for only additive effects (Su et al. 2012; Nishio and Satoh 2014; Lopes et al. 2014; Muñoz et al. 2014). The same was observed in simulated populations (Toro and Varona 2010; Denis and Bouvet 2012; Zeng et al. 2013). Contrary to height, the inclusion of dominance effects did not improve the predictive ability of rust resistance related traits in the real population. Other studies previously reported that dominace deviation was not significant for this characteristic in a pinus breeding population (Isik et al. 2003) and in our analysis the additive models were marginally more accurate than additive-dominance models. In summary, the additive-dominance prediction models improved considerably the accuracies in simulated traits with large dominance effects, but showed limited or no improvement when these effects are modest. Thus, inclusion of dominance in genomic prediction will depend on the trait's genetic architecture in each specific population.

Another goal of this study was to evaluate the effect of using WGR methods that adopt distinct priors in the prediction of traits that include dominance effects. These methods differ in their approach to variable selection and the variance of regressions coefficients. As a consequence, WGR differ in the marginal prior of regression coefficients (markers effects) that control the shrinkage of markers effects (Gianola 2013; de los Campos et al. 2013). The identification of the best model or prior is trait-dependent (Resende, *et al.*, 2012a). In the present study, models with different priors did not differ significantly for the trait height measured in the CCLONES population, and for the polygenic traits in the simulated population. In contrast, the accuracy of prediction models for rust resistance traits were higher for BayesA and BayesB, compared to BRR. The same pattern was observed for the simulated oligogenic traits. These results are expected, as the marginal prior of BayesA and BayesB provide more shrinkage than BRR, and BayesB also incorporates variable selection.

The use of dominance in forest breeding programs is desirable for species that are clonally propagated because their entire genotypic value can be translated to commercial plantations. An accurate estimation of dominance effects can also improve the genetic

gain in improvement programs (Falconer and Mackay 1996). Finally, the incorporation of dominance effects is critical for introduction of breeding approaches that aim to create crosses with complementary alleles, in mate-pair allocation (Toro and Varona 2010). Here we showed that including dominance effects in the prediction of traits controlled by loci with additive and dominance effects can result in more accurate models. Improved models will increase genetic gains for clonal selection and in reciprocal recurrent selection of superior mate-pairs. It has to be noted that in the breeding values estimation, the additive-dominance WGR models were not more accurate, even in the presence of a dominance component (see Figure 3). This limitation is likely to occur because dominance variance estimations is less accurate and demands much more information (Toro and Varona 2010). Estimating the contribution of dominance relies on the measurement of phenotypes in heterozygous individuals. In the simulated population, where more than a third of loci have a MAF below 5%, fewer than 10% of the individuals are expected to have the heterozygote genotype. Furthermore, with only 923 individuals, the simulated population used to train the models may not be sufficiently large to support the accurate estimation of these dominance effects. These results suggest that, as dominance increases, the accuracy of predictions will become less suitable for genomic selection. Others have recently reported that the prediction of dominance deviation from SNPs information is not as accurate as that reported for breeding values (Nishio and Satoh 2014). However, the use of larger training populations (Ertl et al. 2014; Wittenburg et al. 2015) or the adoption of training populations where loci with higher MAF occur (and therefore more heterozygotes are available for dominance estimation) may improve predictions. Further investigation is necessary to identify the factors that most improve the accuracy of predicting dominance effects.

Finally, we evaluated the performance of the models estimated in G2 to predict the simulated progeny (G3). The additive-dominance models outperformed the additive models only for simulated oligogenic trait with high dominance effects. Toro and Varona, (2010) also reported that additive-dominance models outperformed additive models only in the first generation, for polygenic simulated traits. These results suggest that the use of additive-dominance models would only be recommended in species that can be vegetative propagated. Further studies combining the use of additive-dominance models

with mate-pair allocation are required to evaluate if the prediction of dominance can improve the accuracy of subsequent generations under sexual propagation schemes.

SUPPLEMENTARY MATERIAL

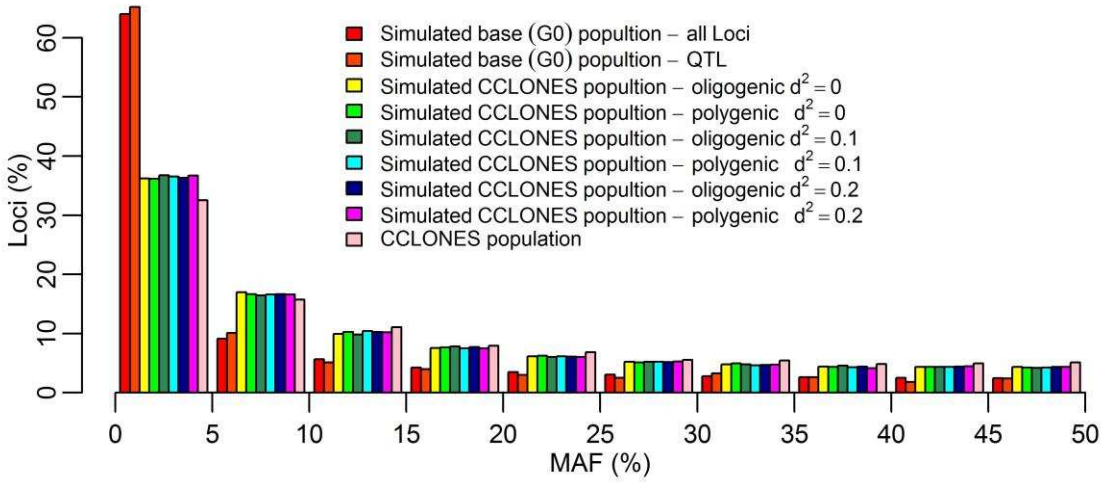

Figure S1. Minor allele frequency distribution of polymorphic loci in the simulated and in the CCLONES population. The base population (G0) corresponds to the unimproved individuals, while the oligogenic and polygenic scenarios reflect the MAF in the population that underwent selection to approximate the genetic composition of CCLONES.

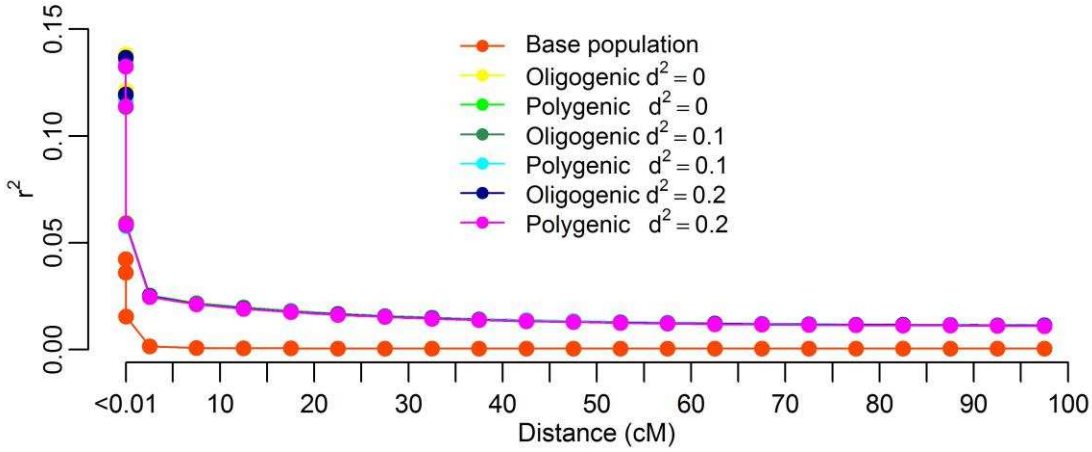

Figure S2. Linkage disequilibrium decay in the simulated populations. The base population (G0) reflects linkage disequilibrium among unimproved, unrelated individuals. The other scenarios reflect linkage disequilibrium the populations that simulates the CCLONES population, after two cycles of breeding and selection.

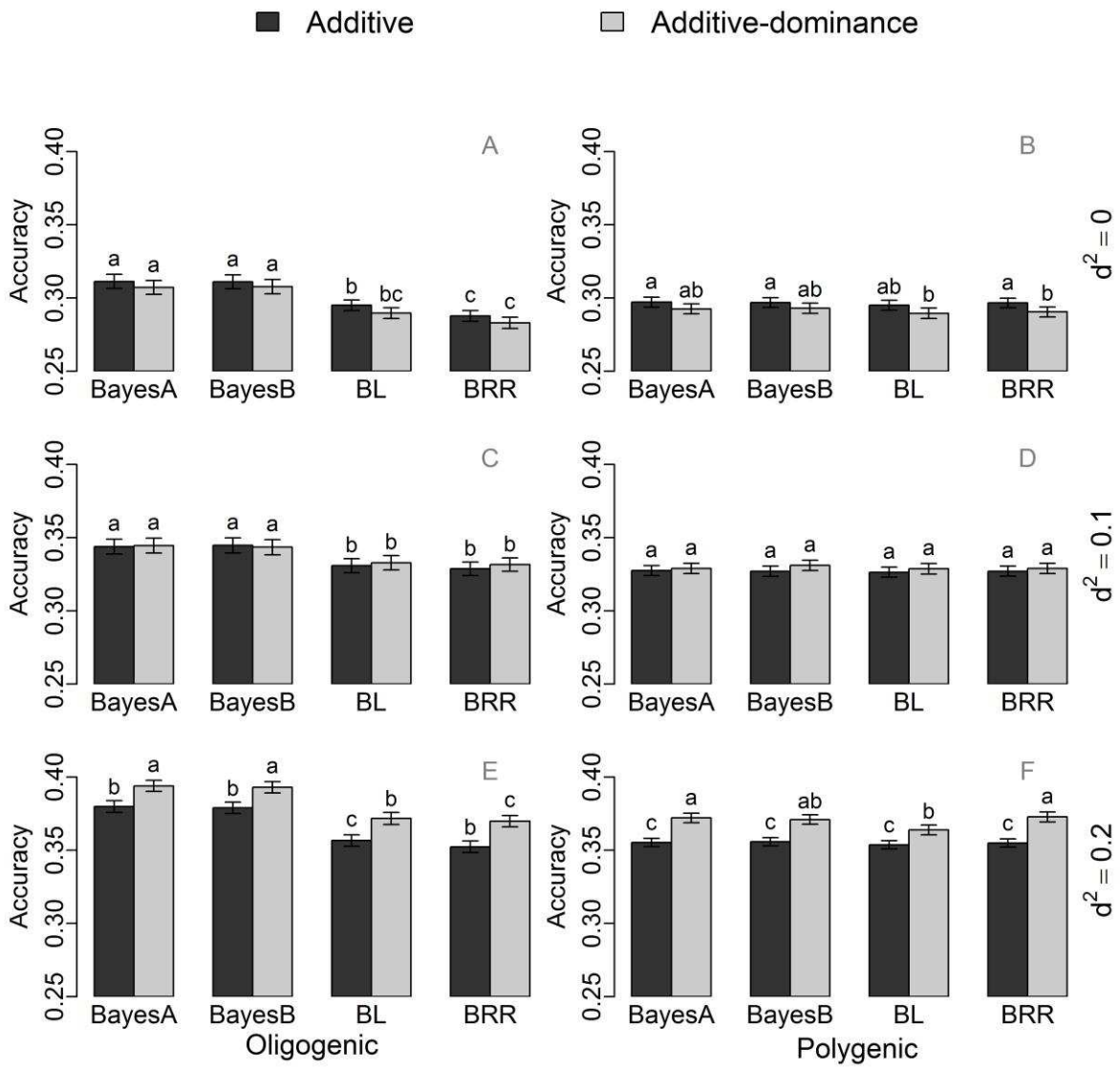

Figure S3. Phenotypic prediction accuracies (or predictive ability $r_{y\hat{g}}$) with additive-dominance WGRs using different priors for four different simulated traits. A and B are oligogenic and polygenic traits, respectively, with h^2=0.25 and d^2=0.1; C and D are oligogenic and polygenic traits, respectively, with h^2=0.25 and d^2=0.2. Error bars are standard error among 10 replicates. Means with same letter are statistically equal by Tukey test ($p<0.05$).

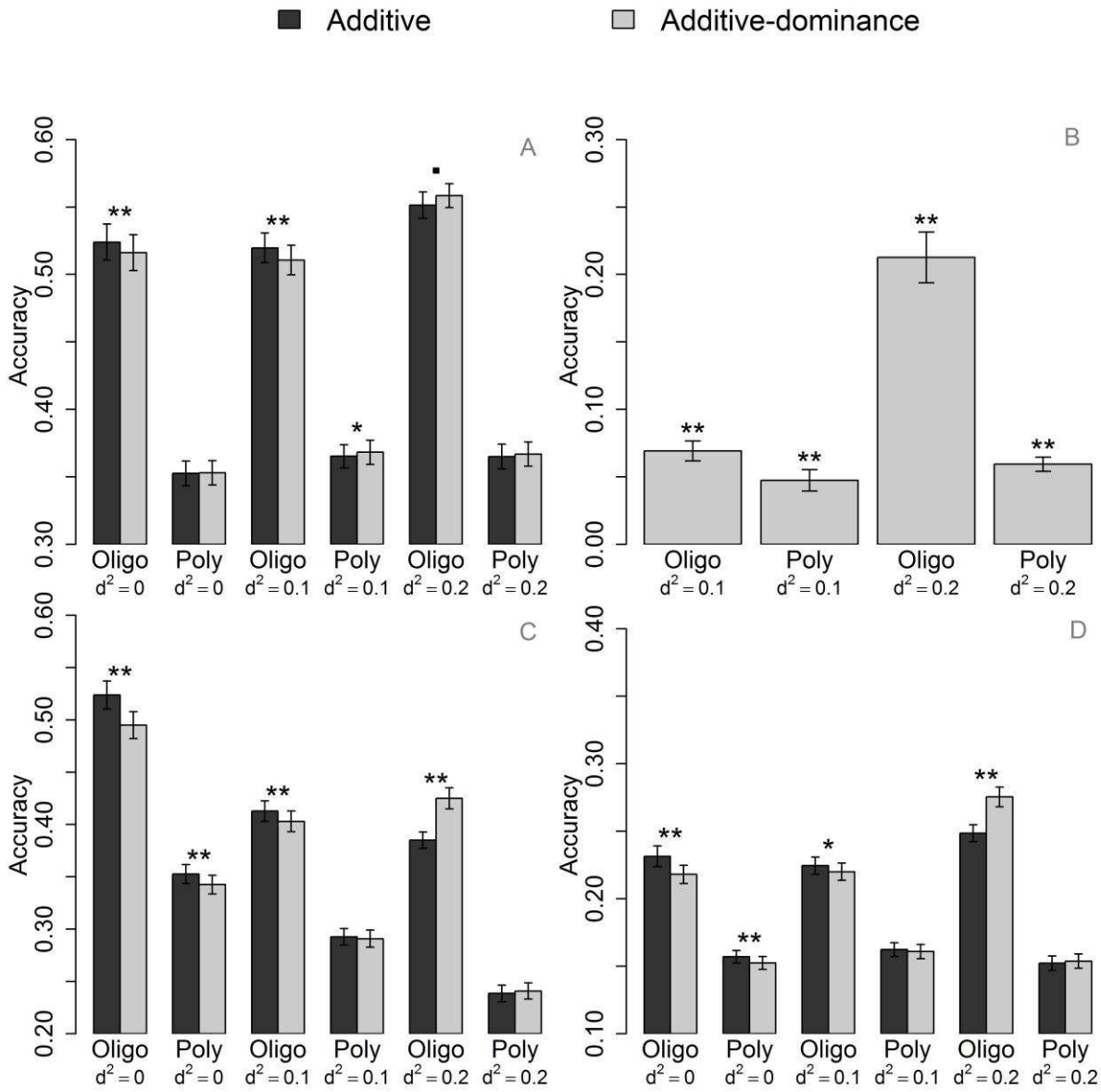

Figure S4. Results of accuracy with model estimated in CCLONES population and validated at the progeny using additive- and additive-dominance BayesB models for A) breeding values prediction, B) dominance deviation prediction C) whole genotypic values prediction and D) phenotypic prediction for six simulated traits with h^2=0.25: oligogenic (oligo) and polygenic (poly) with three different degree of dominance (d^2=0, d^2=0.1 and d^2=0.2). Dot (“.”); ** and *: for A,C and D means additive and additive-dominance models were statistically different P<0.1; P<0.05 and p<0.01 respectively ; ** for B means that the mean were statistically different of zero (p<0.01).

Table S1 Hyperparameters used in Whole-Genomic Regression with different priors considering only additive effects (add) and additive-dominance effects (add-dom). These models were evaluated in three traits, tree height (HT), and two measures of rust resistance: RFbin (presence or absence) and RFgall (gall volume).

Prior	Hyper-parameters	HT		RFbin		RFgall	
		add	add-dom	add	add-dom	add	add-dom
BayesA	ν_e	5	5	5	5	5	5
	S_e	18248.731	18248.731	0.452	0.452	3.744	3.744
	ν_a	5	5	5	5	5	5
	r_a	0.0053	0.0105	212.419	424.837	25.618	51.236
	s_a	1.1	1.1	1.1	1.1	1.1	1.1
	ν_d	-	5	-	5	-	5
	r_d	-	0.007	-	285.751	-	34.46
	s_d	-	1.1	-	1.1	-	1.1
BayesB	ν_e	5	5	5	5	5	5
	S_e	18248.731	18248.731	0.452	0.452	3.744	3.744
	ν_a	5	5	5	5	5	5
	r_a	0.003	0.005	106.209	212.419	12.809	25.618
	s_a	1.1	1.1	1.1	1.1	1.1	1.1
	p_{0_a}	10	10	10	10	10	10
	π_{0_a}	0.5	0.5	0.5	0.5	0.5	0.5
	ν_d	-	5	-	5	-	5
	r_d	-	0.004	-	142.876	-	17.23
	s_d	-	1.1	-	1.1	-	1.1
	p_{0_d}	-	10	-	10	-	10
	π_{0_d}	-	0.5	-	0.5	-	0.5
BL	ν_e	5	5	5	5	5	5
	S_e	18248.731	18248.731	0.452	0.452	3.744	3.744
	r_a	5.2e-5	2.6e-5	5.2e-5	2.6e-5	5.2e-5	2.6e-5
	s_a	1.1	1.1	1.1	1.1	1.1	1.1
	r_d	-	3.9e-5	-	3.9e-5	-	3.9e-5
	s_d	-	1.1	-	1.1	-	1.1
BRR	ν_e	5	5	5	5	5	5
	S_e	18248.731	18248.731	0.452	0.452	3.744	3.744
	ν_a	5	5	5	5	5	5
	S_a	19.019	9.509	0.0005	0.0002	0.0039	0.002
	ν_d	-	5	-	5	-	5
	S_d	-	14.143	-	0.0003	-	0.0029

Table S2 Hyperparameters used in Whole-Genomic Regression with different priors considering only additive effects (add) and addititeve-dominance effects (add-dom). These models were evaluated in six simulated traits: Oligogenic (Oligo) and Poligenic (Poly) with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

Prior	Hyper-Parameters	Oligo d^2=0		Poly d^2=0		Oligo d^2=0.1		Poly d^2=0.1		Oligo d^2=0.2		Poly d^2=0.2	
		add	add-dom	add	add-dom	add	add-dom	add	add-dom	add	add-dom	add	add-dom
BayesA	v_e	5	5	5	5	5	5	5	5	5	5	5	5
	S_e	9.69	9.69	1423.9	1423.9	8.61	8.61	1777	1777	12.62	12.62	3200.5	3200.5
	v_a	5	5	5	5	5	5	5	5	5	5	5	5
	r_a	102.9	205.9	0.08	0.16	28.87	57.73	0.07	0.13	12.12	24.24	0.04	0.07
	s_a	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	v_d	-	5	-	5	-	5	-	5	-	5	-	5
	r_d	-	139.9	-	0.11	-	39.32	-	0.09	-	16.54	-	0.05
	s_d	-	1.1	-	1.1	-	1.1	-	1.1	-	1.1	-	1.1
BayesB	v_e	5	5	5	5	5	5	5	5	5	5	5	5
	S_e	9.69	9.69	1423.9	1423.9	8.61	8.61	1777	1777	12.62	12.62	3200.5	3200.5
	v_a	5	5	5	5	5	5	5	5	5	5	5	5
	r_a	51.48	102.96	0.04	0.08	14.43	28.87	0.03	0.06	6.06	12.12	0.02	0.04
	s_a	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	p_{0_a}	10	10	10	10	10	10	10	10	10	10	10	10
	π_{0_a}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	v_d	-	5	-	5	-	5	-	5	-	5	-	5
	r_d	-	69.95	-	0.05	-	19.67	-	0.04	-	8.27	-	0.02
	s_d	-	1.1	-	1.1	-	1.1	-	1.1	-	1.1	-	1.1
	p_{0_d}	-	10	-	10	-	10	-	10	-	10	-	10
	π_{0_d}	-	0.5	-	0.5	-	0.5	-	0.5	-	0.5	-	0.5
BL	v_e	5	5	5	5	5	5	5	5	5	5	5	5
	S_e	9.69	9.69	1423.9	1423.9	8.61	8.61	1777	1777	12.62	12.6	3200.5	3200.5
	r_a	4.4e-5	2.2e-5	4.4e-5	2.2e-5	4.4e-5	2.2e-5	4.4e-5	2.2e-5	4.4e-5	2.2e-5	4.4e-5	2.2e-5
	s_a	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	r_d	-	3.3e-5	-	3.2e-5	-	3.2e-5	-	3.2e-5	-	3.2e-5	-	3.2e-5
	s_d	-	1.1	-	1.1	-	-	-	1.1	-	1.1	-	1.1
BRR	v_e	5	5	5	5	5	5	5	5	5	5	5	5
	S_e	9.69	9.69	1423.9	1423.9	8.61	8.61	1777	1777	12.62	12.62	3200.5	3200.5
	v_a	5	5	5	5	5	5	5	5	5	5	5	5
	S_a	0.009	0.004	1.26	0.63	0.008	0.004	1.57	0.786	0.011	0.006	2.828	1.414
	v_d	-	5	-	5	-	-	-	5	-	5	-	5
	S_d	-	0.006	-	0.925	-	0.006	-	1.154	-	0.008	-	2.074

Table S3 Average of accuracies of phenotype prediction with pedigree base line modes with only additive effect (Ped-Add), with additive and dominance effects (Ped-Add-dom) and accuracy mean of all genomic models. The comparison between Genomic- and Pedigree-base models were made by contrast estimated as weighted mean of accuracy of genomic models minus pedigree models. The traits evaluated in Pinus were tree height (HT) and two measures of rust resistance

Models	HT	RFbin	RFgall
Ped-Add	0.371	0.335	0.264
Ped-Add-Dom	0.398	0.325	0.259
Genomic	0.407	0.355	0.293
Gen vs Ped	0.023**	0.025**	0.031**

**: means contrast significant at p<0.01.

Table S4. Average of accuracies of pedigree base line modes with only additive effect (Ped-Add), with additive and dominance effects (Ped-Add-dom) and accuracy mean of all genomic models. The comparison between Genomic- and Pedigree-base models were made by contrast estimated as weighted mean of genomic models minus pedigree models.

Accuracy	Model	d^2=0		d^2=0.1		d^2=0.2	
		Oligogenic	Polygenic	Oligogenic	Polygenic	Oligogenic	Polygenic
Breeding values	Ped-Add	0.563	0.575	0.548	0.561	0.533	0.550
	Ped-Add-dom	0.561	0.574	0.551	0.561	0.536	0.556
	Genomic	0.663	0.632	0.655	0.622	0.657	0.618
	Gen vs Ped	0.101**	0.058**	0.105**	0.061**	0.122**	0.065**
Dominance deviation	Ped-Add-dom	-	-	0.170	0.204	0.271	0.257
	Genomic	-	-	0.164	0.174	0.260	0.241
	Gen vs Ped	-	-	-0.006ns	-0.030**	-0.011*	-0.016**
Whole genotypic	Ped-Add	0.563	0.575	0.489	0.521	0.462	0.468
	Ped-Add-dom	0.544	0.553	0.496	0.528	0.493	0.492
	Genomic	0.659	0.628	0.583	0.573	0.560	0.522
	Gen vs Ped	0.105**	0.064**	0.090**	0.048**	0.082**	0.042**
Phenotypic prediction	Ped-Add	0.251	0.264	0.289	0.304	0.306	0.327
	Ped-Add-dom	0.246	0.254	0.291	0.308	0.321	0.343
	Genomic	0.299	0.294	0.337	0.328	0.375	0.362
	Gen vs Ped	0.051**	0.035**	0.047**	0.022**	0.061**	0.027**

**;* and ns: Means contrast significant with p<0.01; p<0.05 and non-significant.

Table S5 Narrow and broad sense heritability, and proportion of proportion of variance of dominant deviations relative to total genetic variance explained by markers using BRR, for height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda.

Trait	Additive-model	Additive-dominance-model			
	h^2	h^2	d^2	H^2	d^2/h^2
HT	0.40 [0.30; 0.51]	0.35 [0.26; 0.45]	0.15 [0.08; 0.22]	0.49 [0.38; 0.60]	0.42 [0.22;0.68]
RFbin	0.37 [0.26; 0.49]	0.32 [0.23; 0.44]	0.10 [0.05; 0.17]	0.42 [0.32; 0.55]	0.31 [0.12;0.57]
RFgall	0.29 [0.19; 0.41]	0.27 [0.18; 0.38]	0.09 [0.05; 0.14]	0.36 [0.25; 0.48]	0.33 [0.16;0.56]

Values between brackets are Bayesian credibility interval (95%).

Table S6 Narrow and broad sense heritability, and proportion of proportion of variance of dominant deviations relative to total genetic variance explained by markers using BRR, for six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

Traits		Additive-model	Additive-dominance-model			
		h^2	h^2	d^2	H^2	d^2/h^2
d^2=0	Oligenic	0.27 [0.18;0.36]	0.25 [0.18;0.33]	0.07 [0.04;0.12]	0.32 [0.23;0.42]	0.28 [0.14;0.48]
	Polygenic	0.26 [0.18;0.35]	0.24 [0.18;0.33]	0.07 [0.04;0.12]	0.32 [0.24;0.41]	0.29 [0.14;0.49]
d^2=0.1	Oligenic	0.30 [0.21;0.40]	0.28 [0.20;0.37]	0.09 [0.05;0.15]	0.37 [0.28;0.47]	0.32 [0.15;0.54]
	Polygenic	0.28 [0.20;0.38]	0.26 [0.19;0.35]	0.09 [0.05;0.13]	0.35 [0.26;0.45]	0.35 [0.20;0.55]
d^2=0.2	Oligenic	0.31 [0.23;0.41]	0.30 [0.22;0.39]	0.11 [0.06;0.17]	0.41 [0.31;0.51]	0.37 [0.19;0.59]
	Polygenic	0.31 [0.22;0.40]	0.29 [0.22;0.37]	0.11 [0.06;0.17]	0.40 [0.30;0.51]	0.38 [0.21;0.58]

Values between brackets are Bayesian credibility interval (95%).

Table S7 Summary of ANOVA for result of correlation between predicted whole genotype values with phenotypes ($r_{y\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure for height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda.

SV	df	HT		RFgall		RFbin	
		MS	p	MS	p	MS	p
Fold	9	0.09808	<0.01	0.06212	<0.01	0.02940	<0.01
Method	7	0.00086	<0.01	0.00011	F<1	0.00370	<0.01
error	63	0.00014		0.00021		0.00025	
R^2 (%)		99.02		97.73		94.93	
CV (%)		2.90		4.92		4.43	

MS and p: means Mean Square and p-value respectively.

Table S8 Summary of ANOVA for result of correlation between predicted whole genotype values with phenotypes ($r_{y\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p	MS	p	MS	p
Fold(Rep)	90	0.06200	<0.01	0.08681	<0.01	0.07760	<0.01	0.06830	<0.01	0.04460	<0.01	0.06919	<0.01
Rep	9	0.13030	<0.01	0.09294	<0.01	0.18737	<0.01	0.09600	<0.01	0.10720	<0.01	0.07218	<0.01
Method	7	0.01326	<0.01	0.00087	<0.01	0.00506	<0.01	0.00023	0.192	0.02295	<0.01	0.00718	<0.01
Rep x Met	63	0.00262	<0.01	0.00013	F<1	0.00109	<0.01	0.00020	0.118	0.00233	<0.01	0.00046	<0.01
error	630	0.00027		0.00019		0.00033		0.00016		0.00055		0.00028	
R^2 (%)		97.60		98.62		97.66		98.58		93.82		97.55	
CV (%)		5.53		4.73		5.41		3.87		6.28		4.60	

MS and p: means Mean Square and p-value respectively.

Table S9 Summary of ANOVA for result of correlation between predicted breeding values values with parametric breeding values ($r_{u\hat{u}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p	MS	p	MS	p
Fold(Rep)	90	0.0234	<0.01	0.03573	<0.01	0.02763	<0.01	0.03764	<0.01	0.0416	<0.01	0.03184	<0.01
Rep	9	0.3792	<0.01	0.19291	<0.01	0.24932	<0.01	0.23936	<0.01	0.3911	<0.01	0.06380	<0.01
Method	7	0.0467	<0.01	0.00065	<0.01	0.02809	<0.01	0.00027	<0.01	0.0639	<0.01	0.00062	<0.01
Rep x Met	63	0.0087	<0.01	0.00013	<0.01	0.00432	<0.01	0.00010	<0.01	0.0061	<0.01	0.00009	F<1
error	630	0.0002		0.00008		0.00014		0.00005		0.0002		0.00010	
R^2 (%)		98.47		99.03		98.27		99.40		98.39		98.19	
CV (%)		1.90		1.39		1.84		1.17		2.20		1.63	

MS and p: means Mean Square and p-value respectively.

Table S10 Summary of ANOVA for result of correlation between predicted whole genotype values with parametric whole genotypes values ($r_{g\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	P	MS	p	MS	p	MS	p
Fold(Rep)	90	0.02355	<0.01	0.0361	<0.01	0.03407	<0.01	0.03182	<0.01	0.0389	<0.01	0.05059	<0.01
Rep	9	0.38690	<0.01	0.1905	<0.01	0.13364	<0.01	0.13273	<0.01	0.1971	<0.01	0.03154	<0.01
Method	7	0.05274	<0.01	0.0029	<0.01	0.01907	<0.01	0.00081	<0.01	0.0513	<0.01	0.01536	<0.01
Rep x Met	63	0.00917	<0.01	0.0001	F<1	0.00335	<0.01	0.00007	F<1	0.0043	<0.01	0.00068	<0.01
error	630	0.00024		0.0002		0.00027		0.00012		0.0005		0.00028	
R^2 (%)		97.70		97.58		96.47		98.18		95.39		96.53	
CV (%)		2.37		2.23		2.81		1.91		3.80		3.23	

MS and p: means Mean Square and p-value respectively.

Table S11 Summary of ANOVA for result of correlation between predicted dominance deviation values with parametric dominance deviation values ($r_{\delta\hat{\delta}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of four simulated additive-dominance traits: : Oligogenic and Poligenic with two degree of dominance (d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0.1		Polygogenic d^2=0.1		Oligogenic d^2=0.2		Polygogenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p
Fold(Rep)	90	0.05415	<0.01	0.03266	<0.01	0.05405	<0.01	0.05012	<0.01
Rep	9	0.40807	<0.01	0.04380	<0.01	0.52494	<0.01	0.12490	<0.01
Method	3	0.00047	0.0909	0.00187	<0.01	0.00752	<0.01	0.00781	<0.01
Rep x Meth	27	0.00051	<0.01	0.00032	F<1	0.00317	<0.01	0.00032	F<1
error	270	0.00022		0.00034		0.00061		0.00046	
R^2 (%)		99.32		97.29		98.32		97.86	
CV (%)		9.00		10.69		9.53		8.88	

MS and p: means Mean Square and p-value respectively.

CHAPTER III
GENOMIC PREDICTION OF ADDITIVE AND NON-ADDITIVE EFFECTS USING GENETIC MARKERS AND PEDIGREES

ABSTRACT

The genetic merit of individuals have been predict using models with dense markers panels and pedigree information for breeding proposes. Initially, models accounted only for additive effects. However, the prediction of non-additive effects is important for many plant breeding systems. In this study we evaluated prediction models that include or ignore non-additive effects, for traits with different genetic architectures. The models tested were based either on genetic markers or pedigree information, or both. The models used to compute the genetic marker information were: Reproducing Kernel Hilbert Spaces (RKHS), additive- and additive-dominance-BayesA. Theoretically RKHS can predict additive and non-additive effects confused (whole genotypic values). Model performance was assessed for the traits tree height (HT) at 6 years of age, diameter at breast height (DBH) and rust resistance, measured in 923 pine individuals from a structured population of 71 full-sib families genotyped with 4,722 genetic markers. We also simulated a population with similar genetic properties, and evaluated the performance of models for six simulated traits with distinct genetic architectures (polygenic and oligogenic traits with three dominance levels). The simulated population were derived from a pine breeding program, originated from selections made in a natural population with Ne=10,000. The inclusion of pedigree information in genomic prediction models did not yield higher accuracies in most part of cases. Both models also provided substantially better predictions than pedigree-only models. The additive-BayesA provided higher accuracies for rust resistance and in simulated additive-oligogenic traits. On the other hand, the inclusion of dominance in BayesA leads to higher accuracies in simulated additive-dominant oligogenic traits. For DBH, HT and additive-dominance polygenic traits the RKHS based models showed slight higher accuracies than BayesA. Our results indicate that the capacity of prediction using genomic information is dependent on the number of genes controlling the trait. Considering that, BayesA performs the best for traits with few

genes with major effects. We also show that the presence of non-additive effects influences the prediction accuracy since additive-dominant-BayesA and RKHS overcame additive-BayesA in traits with non-additive effects.

Keywords: Non-additive, Polygenic, Oligogenic, RKHS, BayesA.

INTRODUCTION

Pedigree and whole-genetic markers have been used to predict genetic merit of individuals in animal (Wiggans et al. 2011) and plant breeding (Resende Jr et al. 2012a; Resende et al. 2012; Crossa et al. 2014). Initially, genetic prediction models only included additive effects (Meuwissen et al. 2001), which yield the required information for animal breeding systems that explore additive effects for the selection of sires that provide semen for worldwide distribution. However, the prediction of dominance effects represents an important feature of models designed for breeding program that focus on cross-bred populations and/or hybrid productions (Zeng et al. 2013; Nishio and Satoh 2014). In forest breeding, non-additive effects are especially relevant because breeders can transfer whole-genotypic values of individuals to the next generation through clonal selection strategies.

Many models predict non-additive effects, and differ among them with regards to genetic architecture assumptions (Gianola 2013; de los Campos et al. 2013). In whole-genomic regression (WGR) models such as the BayesA, the markers are regression coefficients with different variances (Meuwissen et al. 2001). This model provides a good fit for oligenic traits where few genes explain a large proportion of the observed genetic variation (Meuwissen et al 2001). However, in BayesA models, prediction of non-additive effects includes new SNP-covariates associated with these effects (Toro and Varona 2010), which may prohibitively increase the number of parameters with the increasing number of SNP available. The semi-parametric Reproducing Kernel Hilbert Space (RKHS) models can also predict non-additive effects, and demand less computation than WGR, especially when the number of individuals is substantially lower than the number of markers. The RKHS models differ among them depending on number and type of

kernels used to improve the predictions in face of the specific genetic trait architecture (Morota et al. 2013; Morota and Gianola 2014; Tusell et al. 2014).

In addition to marker information in genomic prediction, the inclusion of pedigree effect information has improved prediction accuracy in the case of annual crops (Crossa et al. 2010; Crossa et al. 2013). Vazquez et al. (2010) also showed that, with lower SNP density, the inclusion of pedigree in the model became an important factor for genomic prediction in dairy cattle.

To date, no studies on forest breeding have compared the predictions from models using only genetic marker inputs with others using a combination of genetic markers and pedigree information. Moreover, also in the context of forest breeding, no reports have assessed the RKHS models for the prediction of whole genotypic values. Therefore, in this study, we evaluated RKHS with different kernels, traditional BayesA, additive-dominant BayesA and pedigree inclusion in genomic prediction models applied to forest breeding. To this end, we applied these methods to loblolly pine traits with distinct genetic architecture, and to simulated traits with different genetic architectures.

MATERIAL AND METHODS

Data used

The real trait used were tree height (HT), Diameter at Breast Height (DBH) and two measure of rust: presence or absence of rust (RFbin) and gall volume (RFgall). From previous studies is expected that HT and DBH are polygenic traits (Resende Jr et al. 2012b), and also HT have important non-additive effects (Muñoz et al. 2014) and rust resistance is governed for few genes with higher effects (Resende Jr et al. 2012b; Quesada et al. 2014). The population where these traits were measured was created from 42 founders, and after selections and matting with overlap generation 40 selected individuals were crossed and created 71 full sib families, with average of 13 individuals per family (SD=5). In total 923 individuals from these families were genotyped for 7,216 SNP, where 4,722 loci that were polymorphic in the population were used in this study, regardless of their minimum allele frequency. The HT and DBH were measured in field trials, when the pants were six years old, in eight clonal replicates, this field experiment

were implanted at Nassau (Florida, USA). Whereas the rust resistance (RFbin and RFgall) were measured in green house with three repetitions.

The phenotypes for these traits were adjusted with the linear model:

$$y_{ijk} = \mu + b_k(r_j) + r_j + g_i + e_{ijk}$$

Where: y_{ijk} is phenotype of ith clone evaluated in jth repetition and kth incomplete block, μ is the intercept, $b_k(r_j)$ random effect of kth incomplete block nested jth repetition $b_k(r_j) \sim N(0, \sigma_b^2)$, r_j is the fixed effect of jth repetition and g_i is the effect of jth clone considered as fixed to estimate the least-square means (adjusted means) and e_{ijk} is the error of observation ijk $e_{ijk} \sim N(0, \sigma^2)$. This model were used for DBH and HT, for rust resistance traits, the incomplete block term was dropped. The analysis of variance of these linear models are in Table S1.

We also simulated six traits with different genetic architecture: two different number of genes (oligogenic and polygenic) and three dominance levels (none, median and high). The simulated population was created with similar features of standard forest breeding program that usually start with sample individuals in a natural population and after the breeding provide matting among selected individuals. Here the simulation carried out in two steps, the first were created the base population with 1,000 individuals, these created by randomly sample of 2,000 haplotypes from a population with effective size of 10,000 during 1,000 generations of neutral coalescence model, with mutation rate 2.5 x 10^{-8} per generation (Willyard et al. 2007). Since this first steep had main of simulate the sample in a natural population, this steep were common for all traits. The second step consist in create the breeding population, 100 individuals from base population were phenotypic based selected and after random matting created 1,000 individuals to the first breeding cycle, from these 1,000 individuals of first breeding cycle, 42 individuals were phenotypic based selected and were reproduced exactly the same pedigree of real population used in this study. In the second steep was done with ten independently replicate for each simulated trait.

The genome simulated had 12 chromosomes with 100 cM, the 10,000 non-gene loci were bi-allelic markers (e.g. SNP) used to predictions, and the number of genes were 30 and 1,000 for oligogenic and polygenic traits respectively. All traits had narrow sense

heritability 0.25 and three levels of d^2: 0, 0.1 and 0.2 were considered for create traits with none- , median- and high-dominance-levels respectively, where $d^2 = V_d/V_p$; V_d and V_p are dominance deviation and phenotypic variance respectively (Falconer and Mackay 1996), with the combination of two number genes and three dominance levels, the study had six simulated traits.

The additive effect of a gene (a) were defined as half difference of alternative homozygotes, and dominance effect (d) difference between heterozygote and mean of homozygotes. The distribution used to *a* in oligogenic traits were gamma(rate=1.66,shape=0.4) with signal (positive or negative) sampled with equal probability (Meuwissen et al. 2001), whereas for polygenic traits a where simulated with standard normal distribution (mean=0,sd=1). The dominance when present were simulated by: $d_i = a_i \times \tau_i$, where τ_i were sample from normal distribution with mean zero and standard deviation 1 and 2 for traits with medium- and high-dominance-levels respectively. To achieve the desired values of d^2 were consider just simulations that provided d^2 between 0.9 and 0.11 for medium-dominance traits and between 0.19 and 0.21 for high-dominance traits.

Statistical methods

We used models that consider just SNP or pedigree information, and models that combined SNP and pedigree. In the genomic component (from SNP information), were used semi parametric Reproducing Kernel Hilbert Space models (RKHS) using different kernels (Ka and Ka-Kd) and BayesA that is a whole-genome regression (WGR) with SNP as covariates, considering additive- and additive-dominant effects. The BayesA were used here because overcame other modes in previous studies with these real and simulated traits, this model provided similar results than BayesB and both were better than Bayesian Lasso, BayesCπ, Bayesian Ridge Regression and frequentist RR-BLUP for oligogenic traits, with polygenic traits all models provided similar results. The full base model can be represented by:

$$y_j = \mu + g_j + u_j + \delta_j + e_j$$

Where y_j is the phenotype (adjusted clonal mean in real traits) of individual *j*; μ is the intercept; e_j is the error of observation j; g_j is genotypic value from SNP information that change with the models adopted; u_j additive polygenic effects (when included); δ_j dominant polygenic effect (when included). Except the g_j that depends of model adopted, for the other terms when present were assumed:

$$y_j|\mu + g_j + u_j + \delta_j, \sigma_e^2 \sim IID\ N(\mu + g_j + u_j + \delta_j, \sigma_e^2);$$
$$\mu \sim N(0, 10^6);$$
$$u|A\sigma_u^2 \sim N(0, A\sigma_u^2);$$
$$\sigma_u^2|\nu_u, S_u \sim \chi^{-2}(\nu_u, S_u);$$
$$\delta|D\sigma_\delta^2 \sim N(0, D\sigma_\delta^2);$$
$$\sigma_\delta^2|\nu_\delta, S_\delta \sim \chi^{-2}(\nu_\delta, S_\delta);$$
$$e|I\sigma_e^2 \sim N(0, I\sigma_e^2);$$
$$\sigma_e^2|\nu_e, S_e \sim \chi^{-2}(\nu_e, S_e).$$

Where A is additive relationship matrix that is twice Mallecot's relationship coefficient, D is dominance relationship matrix that is probably of two individuals to be identical by descent, details of A and D matrix can be finding at Henderson (1984). In only-pedigree model u and δ are the breeding values and dominance deviation vectors respectively.

Full BayesA

The full BayesA consider additive and dominant effects from SNP and pedigree. This model can be represented by:

$$y_j = \mu + \sum_{i=1}^{k} (x_{ij}a_i + w_{ij}d_i) + u_j + \delta_j + e_j$$

Where x_{ij} and w_{ij} are the functions of SNP *i* in individual *j*, for genotypes AA, Aa and aa. x_{ij} take values 1 (AA), 0(Aa) and -1 (aa) and w_{ij} is 0 (AA), 1 (Aa) and 0 (aa). a_i and d_i are the additive dominance effect of marker *i*, respectively, p_i is the allele frequency of A in SNP i and $q_i = 1 - p_i$. The dominance effect was fitted only in the additive-dominance model. The priors used in linear regressions coefficients for additive-dominance and additive models are described below.

$a_i|\sigma^2_{a_i} \sim N(0, \sigma^2_{a_i})$; $\sigma^2_{a_i}|\nu_a, S_\alpha \sim \chi^{-2}(\nu_a, S_\alpha)$;$S_a|s_a, r_a \sim G(s_a, r_a)$; $d_i|\sigma^2_{d_i} \sim N(0, \sigma^2_{d_i})$;
$\sigma^2_{d_i}|\nu_d, S_d \sim \chi^{-2}(\nu_d, S_d)$; $S_d|s_d, r_d \sim G(s_d, r_d)$.

RKHS Kernel averaging model

The RKHS model are able to predict together the whole genotypic values (Gianola et al. 2006; Gianola and van Kaam 2008), what include additive and non-additive effects such dominance and gene interactions. The full RKHS here can be represented by:

$$y = \mu + g + u + \delta + e$$

The *g* is the function of markers that correspond the whole genotypic values confused. The g was modeled in two forms called here as RKHS-Ka and RKHS-Ka-Kd. The others terms were already explained.

RKHS-Ka:

$$g|K_a\sigma^2_g \sim N(0, K_a\sigma^2_g)$$
$$\sigma^2_g|\nu_g, S_g \sim \chi^{-2}(\nu_g, S_g)$$
$$K_a = \exp(-\varphi_a D^2_a)$$

D^2_a is squared Euclidean distance matrix among the individuals using the traditional SNP incidence matrix for additive (X), this matrix is the SNP-covariates used in BayesA. The φ_a is bandwidth parameter that control the relationship measure between individuals j and j', for a given distance (squared Euclidean in this case) big positive values of bandwidth drop the relationship of j and j' close (or equal) 0, whereas positive small values drop the relationship of j and j' close (or equal) 1. These bandwidth parameters can be estimated from metropolis-hasting algorithm (Gianola et al. 2006; Gianola and van Kaam 2008), or determined a grid of values, or by kernel averaging approach (de los Campos et al. 2010b). The kernel averaging were used in this study.

In kernel averaging approach each SNP function g is replaced for two or more SNP functions with the same distance (squared Euclidean in this case), however with different bandwidth parameters. Here the *g* were replaced by sum of three functions, thus, $g = \sum_r^3 g_r$, and $\sigma^2_g = \sum_r^3 \sigma^2_{g_r}$, whereas $var(g_{1r}) = \exp(-\varphi_r D^2_a)$.

The bandwidth parameters (φ_{a_r}) used in g_1, g_2, and g_3 are 5/h, 1/h and 1/5/h respectively, where h is 5th percentile of D_a^2 leading to local, intermediate and global kernels, respectively (González-Camacho et al. 2012; Tusell et al. 2014).

RKHS-Ka-Kd:

In RKHS-Ka-Kd beyond the information of X matrix for predict the whole genotypic values, is also included W, that is the SNP incidence matrix for dominance effects. The g in this case is:

$$g = g_a + g_d$$

$$[g_a \quad g_d]' | K_a\sigma_{g2}^2, K_d\sigma_{g2}^2 \sim N([\mathbf{0} \quad \mathbf{0}]', K_a\sigma_g^2 \oplus K_a\sigma_g^2)$$

$$g_d | K_d\sigma_{g_2}^2 \sim N(0, K_d\sigma_{g_d}^2)$$

$$\sigma_{g_d}^2 | \nu_{g_d}, S_{g_d} \sim \chi^{-2}(\nu_{g_d}, S_{g_d})$$

$$K_d = \exp(-\varphi_d D_d^2)$$

The g_a is formulated equally the whole g in RKHS-Ka, also in RKHS-Ka-Kd were considered kernel averaging approach, thus the whole genotypic value is sum of six terms $g = \sum_r^3 (g_{a_r} + g_{d_r})$ and $\sigma_g^2 = \sum_r^3 (\sigma_{g_{a_r}}^2 + \sigma_{d_r}^2)$, whereas $var(g_{1r}) = \exp(-\varphi_{a_r} D_a^2)$ *and* $var(g_{2r}) = \exp(-\varphi_{d_r} D_d^2)$.

The same bandwidth parameters used in RKHS-KA were used for g_{a_r}, and same idea were used in g_{d_1}, g_{d_2} and g_{d_3}, where the bandwidth parameters (φ_{d_r}) were 5/h_d, 1/h_d and 1/5/h_d respectively, where h_d is 5th percentile of D_d^2, similar in (Morota et al. 2014). In both RKHS models (KA and Ka-Kd) is predicted the whole genotypic value, however is not possible split the whole genotypic value in breeding values, dominance deviation and epistasis.

Table 1 Summary of models tested, whereas 'x' means presence of a given effect.

Model code		SNP covariates		Pedigree Kernels		Gaussian Kernels	
Method	Pedigree inclusion	Add	Dom	Add	Dom	Ka	Kd
BayesA Add	None	x					
	Add	x		x			
BayesA Add-Dom	None	x	x				
	Add	x	x	x			
	Add-Dom	x	x	x	x		
RKHS Ka	None					x	
	Add			x		x	
	Add-Dom			x	x	x	
RKHS Ka-Kd	None					x	x
	Add			x		x	x
	Add-Dom			x	x	x	x
Pedigree	Add			x			
	Add-Dom			x	x		

Models validation

In order to compare the prediction results were used 10-fold cross-validation. Each individual were allocated in one of ten groups, each group were dropped once and had their genotypic values predicted with the model fitted using just with the remaining data (other nine groups), this process had 10 loops, each loop was calculated predictions accuracies and regression coefficients of parametric values on predicted of hidden data. This validation were performed in each one of ten replicates of simulated data, in real data the 10-fold process were applied 10 times with independently rearrange of individuals in each fold. The accuracies and regression coefficients showed are means of 100 values, ten-fold x ten-replicates or rearrange for simulated and real data respectively.

Breeding values and dominance deviation

The expected breeding value (EBV) and the expected dominance deviation (EDD) were estimated as described below:

$$E\hat{B}V_j = \sum_i \left[I(x_{ij} = 1) 2q_i + I(x_{ij} = 0)(q_i - p_i) - I(x_{ij} = -1) 2p_i \right] \hat{\alpha}_i + \hat{u}_j$$

and

$$E\hat{D}D_j = \sum_i \left[-I(x_{ij} = 1) 2q_i^2 + I(x_{ij} = 0) 2p_i q_i - I(x_{ij} = -1) 2p_i^2 \right] \hat{d}_i + \hat{\delta}_j$$

Where p_i is allele frequency of allele A of SNP *i*, q_i=1-p_i, $\hat{\alpha}_i$ is the average effect of substitution, $\hat{\alpha}_i = \hat{a}_i + \hat{d}_i(q_i - p_i)$, and I is an indicator function of SNPs; $\hat{u}_j$ and $\hat{\delta}_j$ are terms from additive and polygenic effects respectively when present, or breeding values and dominance deviation in pedigree based models. The whole genotypic value is the sum of $E\hat{B}V_j$ and $E\hat{D}D_j$. In RKHS based models is predicted the whole genotypic value confounded.

Variance components

The variance components from WGR used here are extension of estimators reported in (Zeng et al. 2013; Ertl et al. 2014), these estimators assume absence of epistasis and Hardy-Weinberg equilibrium (Gianola et al. 2009). The general estimator of additive variance (V_A) and the variance due dominance deviation (V_D) are:

$$\hat{V}_A = 2\sum_i p_i q_i \left[\hat{\sigma}^2_{a_i} + (q_i - p_i)^2 \hat{\sigma}^2_{d_i}\right] + \hat{\sigma}^2_u$$

and

$$\hat{V}_D = 4\sum_i (p_i q_i)^2 \, \hat{\sigma}^2_{d_i} + \hat{\sigma}^2_\delta$$

The first part of $\hat{V}_A$ and $\hat{V}_D$ are due marker effects and the second polygenic effects. All of these components were described earlier, the whole genotypic variance is the sum of additive and dominance variance. The h^2, d^2 and H^2 are the proportion of additive, dominance and genotypic variance in phenotypic variance similar the previous explanation. In RKHS models the genetic variance estimated by markers is the whole genotypic variance confounded, and in addition when present there are genetic variance explained by polygenic effects.

All models were fitted with R package BLGR (de los Campos and Perez 2014), using 100,000 iterations, burning of 20,000, thin of 3 and default hyperparameters previously described (Pérez and de los Campos 2014).

RESULTS

There are not dependence between accuracy and rearranges

In real data were considered 10 different rearrange of ten-fold cross validation, the aim of these rearranges were to avoid influence of groups allocations. The results of correlations of predicted genotypic values on phenotypic values showed that there were not dependence of groups allocation and prediction results (Supplementary Tables S2), for that reason in simulated study were considered the ten-fold without rearrange in different groups. For real and simulated traits, the analysis of variance indicated that the models provided results statistically different (Supplementary Table S2-6) in our study conditions.

Prediction bias

The regression coefficient (slope) of observed values versus predicted values was used as a measure of the bias built into the model, where a slope of one indicates the absence of any bias. The linear regression of simulated data included parametric genotypic values and predictions, whereas for real data, with unknown parametric values, we calculated the slope using phenotype values. The predictions in real data (Supplementary Table S7) yielded regression coefficients near one. In most part of predictions in simulated data the slope were close than one, however in dominance deviation predictions the slope were not close than one what reflect that dominance deviation prediction is complex (Supplementary Table S8).

Pedigree information in model predictions

The use of pedigree information often improves the accuracy of genotypic predictions, especially when the latter involve dense SNP information. Here, we evaluated predictions based on correlation using parametric genetic values, for simulated data, and phenotype values, for real data. Traditional pedigree models presented lower accuracy than any model containing SNP information, with real traits (Table 2) for phenotypic prediction and simulated traits (Table 3) for breeding values, genotypic values and phenotypic prediction. When we combined pedigree and SNP information, resulting models did not yield considered better predictions than models that only included SNP

information in most part of cases for breeding values prediction, and genotypic predictions (Table 3). In dominance deviation prediction, the inclusion of polygenic effects improved the accuracy. However in selection based on breeding or genotypic values, altogether, these results indicate that markers-only models are a reasonable option under our study conditions.

Table 2. Average of accuracies of phenotypic values prediction of all models based in only pedigree information, in only markers information and for models that combined pedigree and markers. C1 are contrasts between models with pedigree against others models and C2 are contrasts with models with only markers information against models with markers and pedigree. These contrasts were estimated as difference of weighted means, and were evaluated for diameter at breast height (DBH), height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda.

Models	DBH	HT	RFbin	RFgall
Pedigree	0.536	0.450	0.331	0.255
Markers	0.545	0.459	0.361	0.288
Mar+Ped	0.548	0.465	0.356	0.279
C1: M and MP vs Ped	0.011**	0.013**	0.027**	0.028**
C2: M vs MP	-0.003**	-0.005**	0.005**	0.009**

**: Means contrast significant with p<0.01

Genotypic predictive model strength depends on non-additive effects

The prediction of whole genotypic values provides important information for forest breeding, because the breeder can clone non-additive effects. Alternatively, prediction of dominance effects for each loci is crucial for optimum cross design. The RKHS based models supplied with the appropriate kernels, can theoretically explain additive and non-additive (whole genotypic values). The inclusion of Kd in RKHS did not improve predictions of real traits (Supplementary Table S9) and simulated traits (Supplementary Table S10). In the work with real data, the inclusion of dominance effects in the BayesA model provided better prediction for HT only, for RFbin the additive-BayesA showed better results, and for RFgall and DBH additive- and additive-dominance-BayesA were similar (Figure 1). In the work with simulated traits, the additive-dominance-BayesA showed considerably stronger genotypic (Figure 2) and phenotypic (Supplementary Table S10) prediction accuracy than the additive-BayesA only for traits with high dominance level. These results of whole genotypic prediction indicate that the inclusion of dominance, specifically in the BayesA model should take into account trait dominance levels.

Table 3 Average of accuracies of Breeding values, dominance deviation, genotypic values and phenotypic values prediction of all models based in only pedigree information, in only markers information and for models that combined pedigree and markers. C1 are contrasts between models with pedigree against others models and C2 are contrasts with models with only markers information against models with markers and pedigree. These contrasts were estimated as difference of weighted means, and were evaluated in six simulated traits (Polygenic and Oligogenic traits with three dominance levels).

Accuracy	Models	d^2=0		d^2=0.1		d^2=0.2	
		Olig	Poly	Olig	Poly	Olig	Poly
Breeding Value	Pedigree	0.567	0.576	0.545	0.560	0.538	0.554
	Markers	0.653	0.627	0.645	0.618	0.645	0.613
	Mar+Ped	0.646	0.626	0.639	0.615	0.638	0.610
	C1:M and MP vs Ped	0.085**	0.051**	0.096**	0.056**	0.102**	0.057**
	C2: M vs MP	0.008**	0.001ns	0.006**	0.004**	0.007**	0.003**
Dominance Deviation	Pedigree	-	-	0.179	0.203	0.271	0.259
	Markers	-	-	0.175	0.170	0.273	0.244
	Mar+Ped	-	-	0.186	0.185	0.284	0.258
	C1:M and MP vs Ped	-	-	0.003**	-0.022**	0.010ns	-0.006ns
	C2: M vs MP	-	-	-0.011**	-0.016*	-0.011*	-0.014**
Genotypic Value	Pedigree	0.556	0.567	0.488	0.521	0.481	0.479
	Markers	0.652	0.626	0.586	0.575	0.569	0.537
	Mar+Ped	0.638	0.619	0.578	0.571	0.566	0.536
	C1:M and MP vs Ped	0.087**	0.055**	0.093**	0.051**	0.087**	0.057**
	C2: M vs MP	0.014**	0.007**	0.008**	0.004**	0.003*	0.000ns
Phenotypic Value	Pedigree	0.251	0.259	0.284	0.306	0.313	0.335
	Markers	0.300	0.286	0.338	0.331	0.378	0.373
	Mar+Ped	0.290	0.282	0.335	0.331	0.373	0.373
	C1: M and MP vs Ped	0.0414**	0.025**	0.052**	0.025**	0.0622**	0.038**
	C2: M vs MP	0.007**	0.004**	0.004*	0.000ns	0.005**	0.000ns

**,* and ns means: contrast significance at 1%, 5% and non-significant.

Genotypic predictive model strength is gene-number dependent

Genomic prediction models differ essentially on the assumptions regarding the genetic architecture of traits. The BayesA represents a linear regression model that assumes that each marker has different variance, thus some markers could explain major gene variations in oligogenic traits, such as rust resistance. On the other hand, the RKHS directly yields individual values, in these models all marker with the same MAF contribute equally for relationship measure among individuals, what math more with polygenic assumptions. In whole genotypic and phenotypes predictions the RKHS yielded slight higher accuracy for DBH, HT (Figure 1) and in additive-dominance polygenic simulated

traits (Figure 2, Supplementary Table S11). Regardless of the inclusion of pedigree information in genomic prediction models, the BayesA, when compared to the RKHS, provides higher correlation for RFbin and all oligogenic simulated traits. The difference of accuracies among RKHS and BayesA models were small. Altogether, the results suggested that for whole genotypic prediction, the BayesA based models were the best models for oligogenic traits, where inclusion of dominance effects in BayesA is trait dependent, and for polygenic traits with presence of non-additive effects, the RKHS are a potential option because these models can predict the non-additive effects with much less parameters.

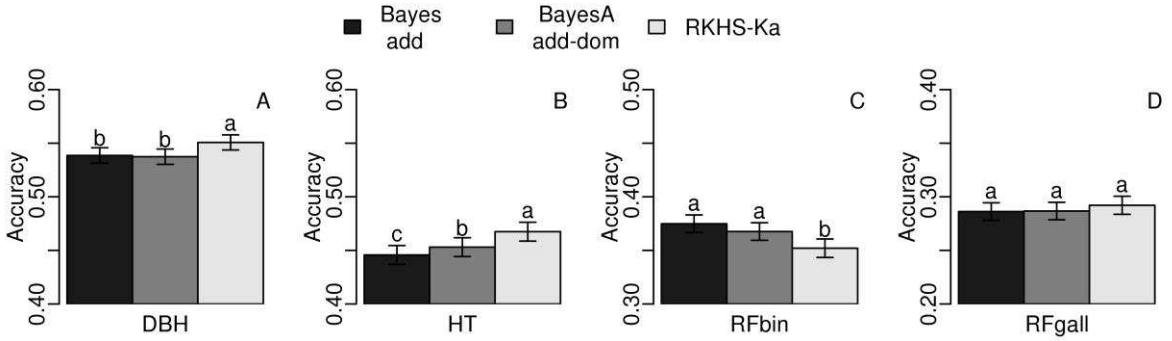

Figure 1. Results of Average of phenotypic prediction accuracies (or predictive ability $r_{\hat{g}y}$) and standard error (error bars) of three models: addive- and additive-dominance-BayesA and RKHS-KA for diameter at breast height (DBH), height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda. Means with same letter are statistically equal by Tukey test ($p<0.05$).

BayesA models provided higher accuracies in breeding value prediction

The breeding value of one individual is the part of it genotypic value that is directly transmitted to progeny, the genotypic value of a progeny from the matting of two individuals is the average of breeding values of parents plus the effects due interactions of alleles from the same locus (dominance) and alleles from alleles of distinct locus (epistasis). Thus the breeding values should be used to select individuals to use extensively in mattings with different individuals. With the BayesA and pedigree based models were possible to predict directly the breeding values. However in RKHS models,

were used the correlation between predicted genotypic values and parametric breeding values to check what is the accuracy for select individuals for explore theirs breeding values. The traditional additive-BayesA and additive-dominance-BayesA based only in markers information provided the higher accuracy for breeding value prediction for all simulated traits (Figure 2), and the pedigree based models showed the worst accuracies. The results suggested that additive-BayesA were the best model for breeding values selection, since this model were statically equal to additive-dominance-BayesA with much less parameters.

Variance components and heritability

One of most important task for breeder is take decision regards breeding strategy, this decision can be supported from: variance components and the proportion of the genetic variance over the phenotypic variance as narrow sense heritability (h^2), broad sense heritability (H^2) and the proportion of dominance variance over phenotypic variance (d^2). Here these parameters were estimated using genetic marker or pedigree information and both in several real and simulated traits. In simulated trait, there is advantage of known the parametric values.

Considering the parametric values of h^2 and d^2 in simulated studies, the only-markers BayesA based models provided the less biased estimated of these genetic parameters, which in traits with non-additive effects the additive-BayesA were the best model (Supplementary Table S12), because the results of h^2 close than parametric value and did not considered dominance effects. While in additive-dominance traits the inclusion of dominance is desired, and the only-markers additive-dominance-BayesA provided the more reasonable results of heritabilities in most part of cases, even though the d^2 were underestimated what reflect the complexity of dominance estimation.

The inclusion of pedigree information on BayesA based models increased the estimates of heritabilities, and in most part of cases these parameters were overestimated. Also the only pedigree based models provided overestimates in majority cases. The RKHS based models predict the whole genotypic values confounded, thus unlike with these models is not possible estimate h^2 and d^2, only H^2. The results of H^2 in simulated traits (Supplementary Table S13) showed that all RKHS based models,

regardless pedigree inclusion, overestimated substantially the H^2 mainly with for models with inclusion of Kd (RKHS Ka-Kd).

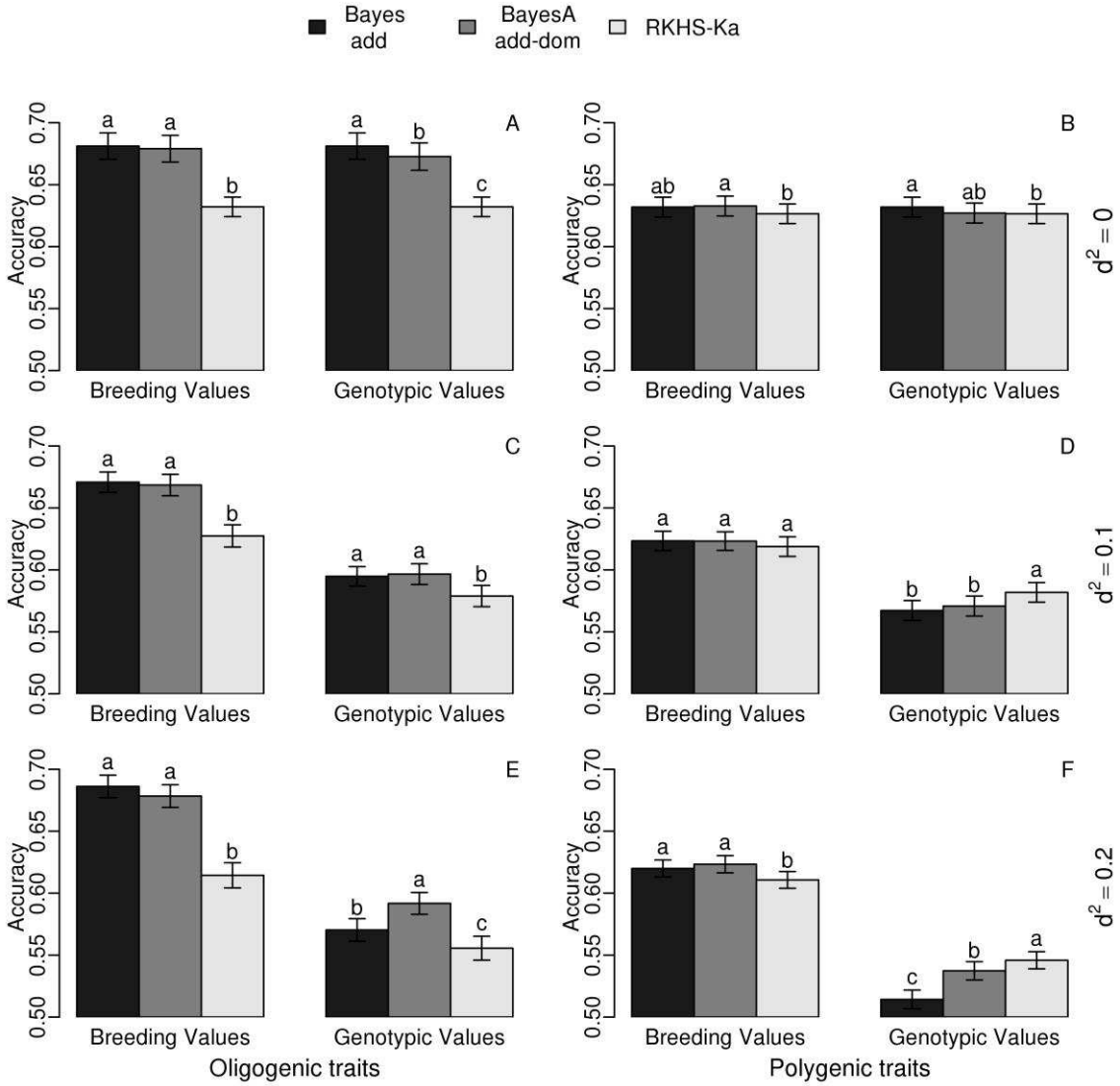

Figure 2. Results of Average of breeding values and genotypic values prediction of three models: addive- and additive-dominance-BayesA and RKHS-KA for of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2). Error bars are standard error. Means with same letter are statistically equal by Tukey test ($p<0.05$).

The results of simulated data suggested the use of BayesA with only markers information for estimate variance components and consequently heritabilities, and the inclusion of dominance effects can be justified if the additive-dominance model provide higher accuracies than additive model. Therefore the for HT is recommended use additive-dominant BayesA and what provided estimates of h^2 and d^2 equal 0.37 and 0.17 respectively, while for DBH, RFbin and RFgall the additive model is suggested and theirs respective h^2 are estimated in 0.52, 0.39 and 0.29 (Table S13).

DISCUSSION

In this study, we tested the strength of genomic- and pedigree-models with and without non-additive effects for the prediction of genetic values in pine. We used real data from a standard forest breeding program that started with the collection of sample trees from natural populations, followed by breeding cycles encompassing basic selection and mating. Pine traits used for model testing included plant height (HT), diameter at breast height (DBH), and the rust resistance measures RFbin and RFgall, whose narrow sense heritability were previous reported in 0.31, 0.31, 0.21 and 0.12, respectively (Resende Jr et al. 2012b). These traits have different genetic architecture, DBH and HT probably represent polygenic traits whereas rust resistance is an oligogenic trait (Resende Jr et al. 2012b; Quesada et al. 2014). Moreover, HT has important non-additive effects (Muñoz et al. 2014). To expand and validate the conclusions we drew from real data, we simulated six distinct genetic architecture traits, polygenic and oligogenic with three dominance levels, considering the same breeding program design. Detailed features of the simulated-data populations included allele frequency and linkage disequilibrium are in previous chapter.

Pedigree information in genomic predictions

Pedigree and marker information was used separately and in combination to predict genetic values. In prediction of breeding and whole genotypic values, the model that combines pedigree and SNP information did not yield higher accuracy than the model that only had markers as an input, and both models provided substantially better predictions than pedigree-only models. In models with low SNP density, the combination of pedigree and markers provided better predictions in simulated studies (Calus and

Veerkamp 2007) wheat (Crossa et al. 2010), and dairy cattle (Vazquez et al. 2010). However, if SNP density was increased, pedigree information did not improve model predictions in the same study with dairy cattle (Vazquez et al. 2010), simulated data (Calus and Veerkamp 2007), and in other work with mice (de los Campos et al. 2009). In maize, models using a high-density SNP panel from Genotyping-by-Sequencing (GBS), and combining pedigree and marker information provided the best option (Crossa et al. 2013). In some cases, these authors also recorded better predictions using pedigree-only models rather than marker-only models. A potential explanation for these different results lie in the fact that the GBS yielded a large number of missing data (Crossa et al. 2013), and important genes may not have had linkage disequilibrium with any informative SNP.

Altogether, these results including real and simulated traits indicate that markers provide sufficient input for total genotypic and breeding values predictions. We suggest that when a large number of SNP information is available, most important gene effects can be captured with marker-only prediction models. On the other hand, when the SNP panel available does not provide enough information on important genes, pedigree inclusion represents a good option for better predictions. Since the cost of genotype depends of number of markers, from these results we can speculate the possibility of genotype more individuals with a low-density panel of markers, and include pedigree information in genomic model. These strategy could provide higher genetic gain, than genotype fewer individuals with high-dense panel of markers, since with large number of individuals the accuracy tend to be higher, and the select intense could be higher.

In dominance deviation prediction the models based in combination of markers and pedigree provided the highest accuracies, and the only-pedigree models also overcame the models with only markers. These results suggested that only markers were not enough to predict dominance deviation effects, and since the accuracies of dominance deviation is much lower than breeding and genotypic values, these results also suggested that the prediction of dominance effects is more complex than breeding values as reported in (Toro and Varona 2010; Nishio and Satoh 2014) and in addition these results indicate that the breeding value correspond the most part of the whole genotypic values (Hill et al. 2008). The cross prediction depends of higher accuracy prediction of breeding values and dominance deviation; in all cases the dominance deviation accuracy is low what could

compromise this prediction, however for cross prediction strategy, the inclusion the dominance polygenic effect should be examined.

Semi-parametric kernel choice

Different kernels are available to improve the predictions of complex traits in semi-parametric RKHS models. In the current study, the genomic predictions of RKHS Ka-Kd models and the simpler RKHS Ka models yielded similar results. These findings are in agreement with those previously reported by Morota et al. (2014) for dairy cattle. These authors did not find additional model strength with the inclusion of extra kernels in the RKHS Ka. Other kernel comparisons in RKHS showed that RKHS Ka is a robust choice for the prediction of additive and non-additive effects (Morota et al. 2013; Tusell et al. 2014).

BayesA provided the highest breeding values accuracy

In simulated study were evaluated the accuracy of breeding values prediction, in all cases the BayesA models with only markers information provided the highest accuracy of breeding values prediction regardless the dominance effects, what consequently would provide higher genetic gain in an intra-population breeding scheme. Beyond additive effects and allele frequency, the breeding value is also function of dominance effects (Falconer and Mackay 1996), thus unexpected that additive-BayesA and additive-dominance-BayesA and models would provide similar accuracies for breeding values in additive-dominance traits, one possible reason for this result is that the accuracy of dominance were small. These results were similar with others simulated studies that showed higher genetic gain with additive model instead additive-dominance in intra-population breeding scheme even in presence of higher dominance effects (Denis and Bouvet 2012). Nishio and Satoh (2014) showed that dominance inclusion did not provided higher accuracies in breeding value prediction, even in traits with dominance effects. Perhaps with increase of accuracy in dominance prediction, the additive-dominance models would be the best model for breeding values prediction in traits with dominance effects. However, for this study case, in breeding values prediction is recommended the additive model because is the simplest model and yielded high accuracies when

compared with others more complex models, and RKHS were not recommended for that prediction.

Prediction whole-genotypic values of distinct genetic architecture traits

Models with built-in assumptions that some markers have major effects, usually provide better genomic predictions for simulated oligogenic traits (de los Campos et al. 2013), and also for real traits controlled by few genes, e.g. fat percentage in milk (Habier et al. 2011). For whole genotypic prediction in this study, the BayesA provided better predictions than the RKHS regarding RFbin and simulated oligogenic traits. This finding is in agreement with other studies that suggest that rust resistance is an oligogenic trait (Resende Jr et al. 2012b; Quesada et al. 2014). Regarding RFgall, the BayesA and RKHS provided equally good predictions, regardless of dominance effects or extra kernels, respectively. In this case, the inclusion of pedigree information did not improve the accuracy of predictions. A possible reason for this similarity between models resides in the RKHS model may have captured epistasis related to RFgall, which the BayesA model would not capture. The inclusion of dominance effects in BayesA did not improve the predictions regarding rust-resistance traits, and additive-BayesA was better than RKHS for RFbin. Together, these results indicate that non-additive effects are less important for rust resistance than additive effects. In the case of simulated additive-dominance oligogenic traits, the additive-dominant-BayesA model provided the best predictions. Thus, the use of models such as the additive-BayesA, which takes into account major genes with additive effects, provides a good option for predictions involving additive-oligogenic traits. In parallel, we can speculated that the use of the additive-dominant BayesA would fit nicely in situations involving additive-dominant traits.

Our analyses of polygenic simulated traits showed that, regarding additive-dominant polygenic traits, the RKHS models were better than the additive-BayesA for whole genotypic predictions. These findings agree with those of other authors who argue that RKHS addresses non-additive variation in a non-explicit manner (Gianola et al. 2006; Gianola and van Kaam 2008; Morota and Gianola 2014). In addition, the RKHS provided slight higher accuracy when compared to the additive-dominant-BayesA, what confirms that RKHS can be explored for predictions in polygenic traits with dominance presence.

In genomic prediction the RKHS models provided slight higher accuracies than additive- and additive-dominant-BayesA for HT and DBH, indicating the presence of important gene interactions for these traits. The results regarding HT agree with those by Muñoz et al. (2014) who suggested the existence of important epistasis effects for this trait. When we compared additive- with additive-dominant-BayesA, results showed that dominance inclusion in BayesA provided better predictions for HT only. These findings suggest that dominance effects are less important for DBH than for HT, and that the better performance of RKHS in comparison with BayesA for DBH could result from additive x additive gene interactions. Altogether, our results indicate that RKHS provides an important tool for the prediction of whole genotypic values of traits with non-additive effects.

Variance components and heritabilities

Using RKHS based models and pedigree information the parameters h^2, d^2 and H^2 were overestimated in most part of cases. The BayesA models that imputes only markers, provided more reasonable estimated variance components values. However even with BayesA models is necessary be aware with dominance inclusion, since the best estimations were achieved with the correct inclusion of genetic effects. Calus and Veerkamp (2007) reported that the inclusion of polygenic effects provided better estimation of variance components when compared with models that included only markers, however in this study the authors considered only additive models and additive traits, and here the estimation of h^2 were close than parametric models for BayesA with only markers information.

Using the additive-dominance-BayesA model, the h^2 estimated for HT in this report were similar the value found in Resende Jr et al. (2012b), but higher than reported in Resende Jr et al. (2012a). While the estimative of d^2 for HT were similar of previous report (Muñoz et al. 2014). For DBH, and rust resistance the h^2 here, were higher than other authors (Resende Jr et al. 2012a; Resende Jr et al. 2012b). In DBH the accuracy for predictions suggested that additive gene interactions can be important, thus a model that include additive epistasis could be tested for estimate variance component for this trait. The simulated results showed that the estimation of variance due dominance deviation is

a challenge, this result agree (Wittenburg et al. 2011) what reported that the estimate of additive variance component were closer than parametric values when compared with variance component due dominance effect. The difficulties in d^2 estimation for traits with high dominance, may be explained by allele frequency, since in the populations used here there are large number of loci with low MAF (see previous chapter), and consequently, there are few heterozygotes Loci what can affect dominance estimation. Perhaps, an investigation in a population with large number of Loci with higher MAF, and structured in reasonable number of half and full sibs could provide the in less unbiased d^2 estimates.

SUPLEMENTARY MATERIAL

Table S1. Summary of Analysis of variance of linear model used for adjust the clonal means.

Source	DBH	HT	RFbin	RFgall
Incomplete Block Variance	0.24**	2448**	-	-
F for Repetition	184.87**	397.35**	3.2795*	20.7005**
F for Clones	2.41**	2.18**	3.851**	2.5576**
Error Variance	3.05	7575.11	0.1342	1.866
Mean	11.40	841.59	0.3531	0.8261
CV(%)	15.33	10.34	103.78	165.37
R^2(%)	63.84	81.41	55.72	46.43

**,*: means significant with p<0.01 and p<0.05 respectively, with F test for fixed effect and LRT for variance components.

Table S2. Summary of ANOVA for result of correlation between predicted whole genotype values with phenotypes (phenotype prediction - $r_{y\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure, with ten different rearrangements of genotype allocation in folds, for diameter at breast height (DBH), height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda.

SV	GL	DBH		HT		RFbin		RFgall	
		MS	p	MS	p	MS	p	MS	p
Modelo	12	0.00394	<0.01	0.010271	<0.01	0.023045	<0.01	0.014257	<0.01
Rearr	9	0.00625	<0.01	0.011367	<0.01	0.015917	<0.01	0.014446	<0.01
Fold(Rearr)	90	0.06322	<0.01	0.108625	<0.01	0.09816	<0.01	0.101926	<0.01
Rearr x Mod	108	5.24E-05	F<1	0.000119	F<1	0.000227	F<1	0.000112	F<1
error	1080	0.00026		0.000363		0.000592		0.000457	
R^2(%)		95.44		96.22		93.56		95.06	
CV(%)		2.94		4.14		6.88		7.69	

MS and p: means Mean Square and p-value respectively.

Table S3. Summary of ANOVA for result of correlation between predicted breeding values with parametric breeding values, with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p	MS	p	MS	p
Rep	9	0.4696	<0.01	0.5090	<0.01	0.6297	<0.01	0.5527	<0.01	0.7890	<0.01	0.1009	<0.01
Fold(Rep)	90	0.0582	<0.01	0.0463	<0.01	0.0627	<0.01	0.0404	<0.01	0.0567	<0.01	0.0560	<0.01
Model	12	0.1664	<0.01	0.0401	<0.01	0.1849	<0.01	0.0481	<0.01	0.2652	<0.01	0.0531	<0.01
Rep x Mod	108	0.0154	<0.01	0.0010	<0.01	0.0113	<0.01	0.0013	<0.01	0.0152	<0.01	0.0005	<0.01
error	1080	0.0003		0.0002		0.0004		0.0002		0.0006		0.0003	
R^2(%)		97.25		97.34		97.27		97.50		96.12		95.67	
CV(%)		2.92		2.49		3.12		2.45		4.04		2.77	

MS and p: means Mean Square and p-value respectively.

Table S4. Summary of ANOVA for result of correlation between predicted whole genotype values with parametric whole genotypes values ($r_{g\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p	MS	p	MS	p
Repl	9	0.4801	<0.01	0.5108	<0.01	0.4137	<0.01	0.2758	<0.01	0.4514	<0.01	0.0475	<0.01
Fold(Rep)	90	0.0585	<0.01	0.0459	<0.01	0.0714	<0.01	0.0633	<0.01	0.0747	<0.01	0.0639	<0.01
Model	12	0.1760	<0.01	0.0495	<0.01	0.1380	<0.01	0.0397	<0.01	0.1345	<0.01	0.0606	<0.01
RepxMod	108	0.0157	<0.01	0.0012	<0.01	0.0083	<0.01	0.0008	<0.01	0.0099	<0.01	0.0007	<0.01
error	1080	0.0004		0.0003		0.0005		0.0003		0.0006		0.0004	
R^2(%)		97.16		97.05		95.92		96.69		95.42		94.32	
CV(%)		3.02		2.66		3.95		2.95		4.41		3.74	

MS and p: means Mean Square and p-value respectively.

Table S5. Summary of ANOVA for result of correlation between predicted whole genotype values with phenotypes ($r_{y\hat{g}}$), with different genomic methods, using ten-fold cross validation procedure in ten repetitions of six simulated traits: Oligogenic and Poligenic with three degree of dominance (d^2=0; d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0		Polygenic d^2=0		Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p	MS	p	MS	p
Fold(Rep)	90	0.1097	<0.01	0.1194	<0.01	0.1199	<0.01	0.1130	<0.01	0.1196	<0.01	0.1050	<0.01
Rep	9	0.1275	<0.01	0.2039	<0.01	0.3781	<0.01	0.1606	<0.01	0.2346	<0.01	0.0753	<0.01
Model	12	0.0423	<0.01	0.0104	<0.01	0.0409	<0.01	0.0099	<0.01	0.0694	<0.01	0.0263	<0.01
Repl x Mod	108	0.0051	<0.01	0.0006	<0.01	0.0035	<0.01	0.0012	<0.01	0.0056	<0.01	0.0007	<0.01
error	1080	0.0004		0.0003		0.0007		0.0004		0.0007		0.0004	
R^2(%)		96.35		97.44		95.46		96.60		95.08		96.11	
CV(%)		7.20		6.31		7.85		6.01		7.17		5.41	

MS and p: means Mean Square and p-value respectively.

Table S6. Summary of ANOVA for result of correlation between predicted dominance deviation values with parametric dominance deviation values, with different genomic methods, using ten-fold cross validation procedure in ten repetitions of four simulated additive-dominance traits: Oligogenic and Poligenic with two degree of dominance (d^2=0.1 and d^2=0.2).

SV	df	Oligogenic d^2=0.1		Polygenic d^2=0.1		Oligogenic d^2=0.2		Polygenic d^2=0.2	
		MS	p	MS	p	MS	p	MS	p
Fold(Repl)	90	0.0548	<0.01	0.0373	<0.01	0.0387	<0.01	0.0343	<0.01
Rep	9	0.3944	<0.01	0.0415	<0.01	0.6073	<0.01	0.1138	<0.01
Model	3	0.0082	<0.01	0.0382	<0.01	0.0217	<0.01	0.0244	<0.01
Mod x Rep	27	0.0042	<0.01	0.0036	0.02	0.0178	<0.01	0.0008	F<1
error	270	0.0018		0.0022		0.0021		0.0018	
R^2(%)		94.73		87.16		94.35		89.82	
CV(%)		23.25		24.97		16.48		16.51	

MS and p: means Mean Square and p-value respectively.

Table S7. Results of Slope of predicted Genotypic values and phenotypes for diameter breast height (DBH), plant height (HT), Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in *Pinus taeda*, using different methods according with table 1.

Method	Pedigree	DBH	HT	RFbin	RFgall
BayesA Add	None	1.000	0.991	0.986	1.029
	Add	1.013	0.996	0.971	0.962
BayesA Add-Dom	None	0.989	0.971	0.965	1.007
	Add	1.007	0.980	0.951	0.953
	Add-Dom	1.010	0.992	0.935	0.908
RKHS Ka	None	1.063	1.066	1.066	1.103
	Add	1.066	1.064	1.038	1.037
	Add-Dom	1.070	1.076	1.021	1.001
RKHS Ka-Kd	None	1.152	1.173	1.172	1.227
	Add	1.116	1.144	1.108	1.124
	Add-Dom	1.117	1.148	1.092	1.084
Pedigree	Add	1.026	1.013	0.985	0.958
	Add-Dom	1.035	1.039	0.961	0.914

Table S8. Results of Slope of prediction of Breeding, Dominance deviation, Genotypic and phenotypic values of for six simulated traits (Polygenic and Oligogenic traits with three dominance levels), using different methods according with table 1.

Domiance Level	Method	Pedigree inclusion	Bredding Value		Dominance Deviation		Genotypic Value		Phenotypic Value	
			Olig	Poly	Olig	Poly	Olig	Poly	Olig	Poly
d^2=0	BayesA Add	None	1.073	1.052	_	_	1.073	1.052	0.993	0.969
		Add	1.015	1.002	_	_	1.015	1.002	0.94	0.918
	BayesA Add-Dom	None	1.089	1.096	_	_	1.032	1.023	0.965	0.937
		Add	1.044	1.038	_	_	1.001	0.988	0.921	0.903
		Add-Dom	1.115	1.101	_	_	0.973	0.942	0.888	0.882
	RKHS Ka	None	1.11	1.069	_	_	1.11	1.069	0.997	0.978
		Add	1.059	1.049	_	_	1.059	1.049	0.963	0.947
		Add-Dom	1.178	1.147	_	_	1.028	1.006	0.936	0.919
	RKHS Ka-Kd	None	1.183	1.161	_	_	1.183	1.161	1.087	1.061
		Add	1.13	1.103	_	_	1.13	1.103	1.024	1.008
		Add-Dom	1.239	1.21	_	_	1.094	1.076	0.996	0.981
	Pedigree	Add	1.001	1.009	_	_	1.001	1.009	0.919	0.926
		Add-Dom	1.238	1.231	_	_	0.952	0.958	0.872	0.873
d^2=0.1	BayesA Add	None	0.984	0.932	-	-	1.029	1.013	0.995	0.987
		Add	0.947	0.887	-	-	0.987	0.971	0.959	0.955
	BayesA Add-Dom	None	1.024	0.974	1.195	1.010	1.004	0.979	0.963	0.962
		Add	0.97	0.924	2.986	1.188	0.974	0.959	0.935	0.939
		Add-Dom	1.047	1.019	0.648	0.605	0.962	0.944	0.923	0.925
	RKHS Ka	None	0.982	0.933	-	-	1.069	1.047	1.043	1.027
		Add	0.958	0.908	-	-	1.038	1.017	1.013	1.001
		Add-Dom	1.081	1.028	-	-	1.022	0.994	0.998	0.983
	RKHS Ka-Kd	None	1.037	0.977	-	-	1.155	1.114	1.12	1.098
		Add	1.011	0.962	-	-	1.109	1.087	1.079	1.067
		Add-Dom	1.137	1.049	-	-	1.101	1.056	1.057	1.046
	Pedigree	Add	0.921	0.861	-	-	0.953	0.955	0.919	0.951
		Add-Dom	1.145	1.083	0.506	0.539	0.933	0.941	0.914	0.937
d^2=0.2	BayesA Add	None	0.892	0.867	-	-	1.01	0.963	1.008	1.008
		Add	0.865	0.832	-	-	0.986	0.935	0.978	0.98
	BayesA Add-Dom	None	0.96	0.926	1.262	1.193	0.996	0.95	0.987	0.985
		Add	0.91	0.877	1.358	1.264	0.978	0.933	0.965	0.971
		Add-Dom	0.98	0.969	0.954	0.936	0.977	0.934	0.958	0.969
	RKHS Ka	None	0.867	0.852	-	-	1.067	1.018	1.051	1.063
		Add	0.856	0.841	-	-	1.049	0.998	1.028	1.041
		Add-Dom	0.988	0.968	-	-	1.048	0.997	1.018	1.039
	RKHS Ka-Kd	None	0.911	0.895	-	-	1.159	1.095	1.139	1.143
		Add	0.893	0.879	-	-	1.129	1.065	1.105	1.112
		Add-Dom	1.007	0.988	-	-	1.12	1.059	1.091	1.104
	Pedigree	Add	0.826	0.814	-	-	0.977	0.923	0.954	0.974
		Add-Dom	1.08	1.055	0.856	0.853	0.991	0.937	0.947	0.979

Table S9. Average of accuracies of Breeding values, genotypic values and phenotypic values prediction of RKHS models based in only markers information (without pedigree inclusion) with different kernels: RKHS-Ka and RKHS-Ka-Kd. The comparison between these models was made by contrasts for diameter at breast height (DBH), height (HT) and Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda.

Models	DBH	HT	RFbin	RFgall
RHKS-KA	0.551	0.467	0.352	0.292
RKHS-KA-KD	0.552	0.472	0.349	0.287
Contrast: KA vs KA-KD	-0.001ns	-0.004ns	0.003ns	0.005ns

ns: Means contrast non-significant.

Table S10. Average of accuracies of Breeding values, genotypic values and phenotypic values prediction of RKHS models based in only markers information (without pedigree inclusion) with different kernels: RKHS-Ka and RKHS-Ka-Kd. The comparison between these models was made by contrasts for six simulated traits: Polygenic (Poly) and Oligogenic (Olig) traits with three dominance levels.

Accuracy	Models	d^2=0		d^2=0.1		d^2=0.2	
		Olig	Poly	Olig	Poly	Olig	Poly
Breeding Value	RHKS-KA	0.632	0.627	0.627	0.619	0.614	0.611
	RKHS-KA-KD	0.620	0.618	0.614	0.608	0.601	0.600
	C: KA vs KA-KD	0.012**	0.009**	0.014ns	0.010**	0.014**	0.011*
Genotypic Value	RHKS-KA	0.632	0.627	0.579	0.582	0.556	0.546
	RKHS-KA-KD	0.620	0.618	0.574	0.579	0.560	0.549
	C: KA vs KA-KD	0.012**	0.0087**	0.0053ns	0.002ns	-0.005ns	-0.003ns
Phenotypic Value	RHKS-KA	0.286	0.286	0.336	0.336	0.367	0.380
	RKHS-KA-KD	0.281	0.282	0.332	0.335	0.370	0.382
	C: KA vs KA-KD	0.005ns	0.004ns	0.0032ns	0.0003ns	-0.003ns	-0.002ns

**,* and ns: Means contrast significant with p<0.01, p<0.01 and non-significant.

Table S11. Average of phenotypic prediction accuracies (or predictive ability $r_{\hat{g}y}$) of three models: addive- and additive-dominance-BayesA and RKHS-KA for six simulated traits (Polygenic and Oligogenic traits with three dominance levels).

Model	d^2=0		d^2=0.1		d^2=0.2	
	Oligogenic	Polygenic	Oligogenic	Polygenic	Oligogenic	Polygenic
BayesA-add	0.313a	0.290a	0.343a	0.325b	0.382b	0.358c
BayesA-add-dom	0.307a	0.286a	0.342a	0.327b	0.394a	0.372b
RKHS-KA	0.286b	0.286a	0.336b	0.336a	0.367c	0.380a

Means with same letter are statistically equal by Tukey test (p<0.05). SE<0.01.

Table S12 Results of results of narrow sense heritability (h^2), proportion of phenotypic variance due dominance (d^2), and broad sense heritability (H^2) for six simulated traits (Polygenic and Oligogenic traits with three dominance levels).

Trait	Method	Pedigree inclusion	Without dominance			Medium dominance			High dominance		
			h^2	d^2	H^2	h^2	d^2	H^2	h^2	d^2	H^2
Oligo-genic	BayesA Add	None	0.21	-	0.21	0.27	-	0.27	0.27	-	0.27
		Add	0.34	-	0.34	0.41	-	0.41	0.42	-	0.42
	BayesA Add-Dom	None	0.20	0.04	0.25	0.26	0.06	0.32	0.25	0.11	0.36
		Add	0.30	0.04	0.34	0.37	0.06	0.43	0.38	0.09	0.47
		Add-Dom	0.28	0.14	0.42	0.33	0.16	0.49	0.33	0.22	0.55
	RKHS Ka	None	-	-	0.60	-	-	0.67	-	-	0.70
		Add	-	-	0.59	-	-	0.66	-	-	0.69
		Add-Dom	-	-	0.60	-	-	0.67	-	-	0.70
	RKHS Ka-Kd	None	-	-	0.70	-	-	0.74	-	-	0.76
		Add	-	-	0.68	-	-	0.73	-	-	0.75
		Add-Dom	-	-	0.69	-	-	0.73	-	-	0.75
	Pedigree	Add	0.31	-	0.31	0.34	-	0.34	0.37	-	0.37
		Add-Dom	0.21	0.18	0.39	0.23	0.20	0.43	0.24	0.24	0.48
Poly-genic	BayesA Add	None	0.23	-	0.23	0.25	-	0.25	0.29	-	0.29
		Add	0.35	-	0.35	0.39	-	0.39	0.43	-	0.43
	BayesA Add-Dom	None	0.22	0.03	0.26	0.24	0.07	0.31	0.27	0.11	0.38
		Add	0.31	0.04	0.35	0.35	0.05	0.40	0.39	0.09	0.48
		Add-Dom	0.28	0.13	0.41	0.30	0.17	0.48	0.34	0.21	0.55
	RKHS Ka	None	-	-	0.60	-	-	0.66	-	-	0.70
		Add	-	-	0.58	-	-	0.64	-	-	0.69
		Add-Dom	-	-	0.59	-	-	0.65	-	-	0.70
	RKHS Ka-Kd	None	-	-	0.70	-	-	0.73	-	-	0.77
		Add	-	-	0.68	-	-	0.72	-	-	0.75
		Add-Dom	-	-	0.68	-	-	0.72	-	-	0.75
	Pedigree	Add	0.31	-	0.31	0.36	-	0.36	0.38	-	0.38
		Add-Dom	0.22	0.16	0.38	0.25	0.21	0.45	0.26	0.25	0.51

Table S13 Results of s of narrow sense heritability (h^2), proportion of phenotypic variance due dominance (d^2), and broad sense heritability (H^2). These results are from evaluation of the diameter breast height (DBH), plant height (HT), Rust resistance evaluated as gall volume (RFgall) and presence or absence (RFbin) in Pinus taeda, using different methods according with table 1.

Method	Pedigree inclusion	DBH			HT			RFbin			RFgall		
		h^2	d^2	H^2	h^2	d^2	H^2	h^2	d^2	H^2	h^2	d^2	H^2
BayesA Add	None	0.517	-	0.517	0.447	-	0.447	0.391	-	0.391	0.287	-	0.287
	Add	0.680	-	0.680	0.602	-	0.602	0.509	-	0.509	0.412	-	0.412
BayesA Add-Dom	None	0.462	0.097	0.559	0.367	0.166	0.533	0.378	0.026	0.403	0.285	0.044	0.329
	Add	0.621	0.055	0.676	0.527	0.109	0.636	0.481	0.037	0.518	0.363	0.072	0.435
	Add-Dom	0.550	0.170	0.720	0.430	0.254	0.684	0.432	0.154	0.587	0.290	0.137	0.426
RKHS Ka	None	-	-	0.844	-	-	0.807	-	-	0.723	-	-	0.650
	Add	-	-	0.814	-	-	0.795	-	-	0.679	-	-	0.617
	Add-Dom	-	-	0.822	-	-	0.794	-	-	0.682	-	-	0.625
RKHS Ka-Kd	None	-	-	0.841	-	-	0.827	-	-	0.764	-	-	0.712
	Add	-	-	0.819	-	-	0.809	-	-	0.734	-	-	0.691
	Add-Dom	-	-	0.816	-	-	0.803	-	-	0.726	-	-	0.685
Pedigree	Add	0.677	-	0.677	0.616	-	0.616	0.422	-	0.422	0.318	-	0.318
	Add-Dom	0.513	0.202	0.715	0.392	0.290	0.682	0.296	0.186	0.481	0.221	0.180	0.401

CHAPTER IV

GENERAL CONCLUSIONS

One of the most important task in plant and animal breeding is to predict the individuals with the highest genetic merit. The challenge in this prediction is due to complex nature of the traits which are important in farm production. Some traits are controlled by a large number of genes with small effects, while in other traits, only a small number of genes is liable for the major part of genotypic variation, and regardless the number of genes involved in the control of traits. The interaction between alleles from the same gene and/or from different genes may also play a role in genotypic variation. Genomic prediction (GP) can be more accurate, providing a good knowledge of the genetic architecture of the trait, with the choice of a model that matches this genetic architecture.

The first GP approaches were based on additive models and breeding value prediction; these models are useful for breeding systems that explore mainly the overall combination ability. However, in many breeding systems that require specify crosses, the dominance effects should be taken into consideration. Nonetheless, the dominance prediction is a challenge, since a reasonable number of heterozygotes is necessary for each marker, and at least two kinds of families (e.g. half and full-sibs) are recommended. In our study conditions, dominance prediction provided little accuracy when compared to additive prediction; one potential explanation is that most markers had low MAF and consequently, few heterozygotes per marker.

The inclusion of dominance effects in GP models should be trait-dependent; the results showed that the additive-dominance effects in GP provided higher accuracies for phenotype prediction in plant height, which is a trait with previous knowledge that non-additive effects are important. While, for rust resistance, it is a trait with knowledge that the additive effects were much more important than the non-additive effects, and consequently, the additive models were considered as being the best ones. The results of simulated traits for phenotype and genotypic prediction support the conclusion of real traits. For additive-simulated traits, the additive-GP models provided the highest accuracies and the additive-dominance GP models were most accurate for traits with high

dominance effects. For the prediction of breeding values, the additive models were the best option in most part of cases studied, even for traits with high dominance, these results could be explained by the fact that dominance prediction was not as accurate as additive prediction, and consequently, for the selection of one individual for a large number of crosses, the additive model should be preferred. Also, for heterosis exploration in additive-dominance traits, the additive-dominance model could be explored if the additive and dominance effects were predicted with reasonable accuracy.

In a simulated study, the prediction accuracies with additive- and additive-dominance-BayesB fitted in one previous generation was assessed. The dominance inclusion provided higher accuracies than its additive counterpart model only for oligogenic trait with high dominance, which suggests that it is necessary to have a higher accuracy in dominance prediction for use in the additive-dominance models in next generations.

In addition, there are different assumptions for marker contributions in GP, some models assumes that all markers with the same MAF contribute with the same portion of genetic variation, and other models assume that the markers have heterogeneous variance components and consequently assume that markers with the same MAF can contribute differently for genetic variation. These assumptions play a role in active higher accuracies, with the assumption that some marker can have major BayesA and BayesB effects with more accurate models for simulated oligogenic and rust-resistance traits. For simulated polygenic traits, plant height and diameter at breast height in the RKHS models provided slightly higher accuracy for phenotype and genotype prediction. These RKHS-based models can predict the entire genotypic value confounded. In others words, with RKHS, it is not possible to predict breeding values and dominance deviation separately. In addition, it is not possible to estimate the additive and dominance effects of markers; what restrict RKHS models for cases where the selection is based in genotypic values such as clone selection. Therefore, in models based in whole-genome regressions (WGR), such as BayesA, where it is possible to estimate additive and non-additive effects , it can be used for exploring cross allocation.

Finally, the pedigree information for prediction were investigated. The models with only marker information and models that combined marker with pedigree information

provided similar accuracies, and both were more accurate than the model based only on pedigree. These results suggest that for the real and simulated population used in this study, the markers available were enough for prediction, and it can be speculated that the combination of marker and pedigree can be useful in genomic prediction with low-density marker panel.

REFERENCES

Aguiar A V, Souza VA, Fritzsons E, Pinto Junior JE (2011) Programa de melhoramento de pinus da Embrapa Florestas. Embrapa Florestas, Colombo, PR

Alcantara GB de, Ribas LLF, Higa AR, et al (2007) Efeito da idade da muda e da estação do ano no enraizamento de miniestacas de Pinus taeda L. Rev Árvore 31:399–404. doi: 10.1590/S0100-67622007000300005

Azevedo CF, de Resende MDV, e Silva FF, et al (2015) Ridge, Lasso and Bayesian additive-dominance genomic models. BMC Genet 16:105. doi: 10.1186/s12863-015-0264-2

Baker JB, Langdon OG (1990) Loblolly Pine. In: Burns RM, Honkala BH (eds) Silvics of North America. U.S. Department of Agriculture, Forest Service, Agriculture Handbook 654, Washington, D.C., p 1383

Baltunis BS, Huber D a, White TL, et al (2007) Genetic analysis of early field growth of loblolly pine clones and seedlings from the same full-sib families. Can J For Res 37:195–205. doi: 10.1139/x06-203

Bernardo R (2008) Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. Crop Sci 48:1649. doi: 10.2135/cropsci2008.03.0131

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

Brown GR, Gill GP, Kuntz RJ, et al (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci U S A 101:15255–60. doi: 10.1073/pnas.0404231101

Browning SR, Browning BL (2007) Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet 81:1084–1097. doi: 10.1086/521987

Burgueño J, de los Campos G, Weigel K, Crossa J (2012) Genomic Prediction of Breeding Values when Modeling Genotype × Environment Interaction using Pedigree and Dense Molecular Markers. Crop Sci 52:707. doi: 10.2135/cropsci2011.06.0299

Calus MPL, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring

the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. J Anim Breed Genet 124:362–8. doi: 10.1111/j.1439-0388.2007.00691.x

Chen GK, Marjoram P, Wall JD (2009) Fast and flexible simulation of DNA sequence data. Genome Res 19:136–42. doi: 10.1101/gr.083634.108

Clark SA, Hickey JM, van der Werf JHJ (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43:18. doi: 10.1186/1297-9686-43-18

Coster A, Bastiaansen JWM, Calus MPL, et al (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. Genet Sel Evol 42:9. doi: 10.1186/1297-9686-42-9

Crossa J, Beyene Y, Kassa S, et al (2013) Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 (Bethesda) 3:1903–26. doi: 10.1534/g3.113.008227

Crossa J, Campos G de L, Pérez P, et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–24. doi: 10.1534/genetics.110.118521

Crossa J, Pérez P, Hickey J, et al (2014) Genomic prediction in CIMMYT maize and wheat breeding programs. Heredity (Edinb) 112:48–60. doi: 10.1038/hdy.2013.16

Cruz CD, Carneiro PCS, Regazzi AJ (2014) Modelos Biométricos Aplicados ao Melhoramento Genético - vol II. UFV, Viçosa

Cruz CD, Regazzi AJ, Carneiro PCS (2012) Modelos Biométricos Aplicados ao Melhoramento Genético. UFV, Viçosa

Daetwyler HD, Calus MPL, Pong-Wong R, et al (2013) Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics 193:347–65. doi: 10.1534/genetics.112.147983

Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–31. doi: 10.1534/genetics.110.116855

Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS One 3:e3395. doi:

10.1371/journal.pone.0003395

de los Campos G, Gianola D, Allison DB (2010a) Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat Rev Genet 11:880–6. doi: 10.1038/nrg2898

de Los Campos G, Gianola D, Rosa GJM (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J Anim Sci 87:1883–7. doi: 10.2527/jas.2008-1259

de los Campos G, Gianola D, Rosa GJM, et al (2010b) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genet Res (Camb) 92:295–308. doi: 10.1017/S0016672310000285

de los Campos G, Hickey JM, Pong-Wong R, et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–45. doi: 10.1534/genetics.112.143313

de los Campos G, Naya H, Gianola D, et al (2009) Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–85. doi: 10.1534/genetics.109.101501

de los Campos G, Perez PR (2014) BGLR: Bayesian Generalized Linear Regression.

Dekkers JCM (2004) Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci 82 E-Suppl:E313–328.

Denis M, Bouvet J-M (2012) Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genet Genomes 9:37–51. doi: 10.1007/s11295-012-0528-1

Eckert AJ, van Heerwaarden J, Wegrzyn JL, et al (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics 185:969–82. doi: 10.1534/genetics.110.115543

Ertl J, Legarra A, Vitezica ZG, et al (2014) Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet Sel Evol 46:40. doi: 10.1186/1297-9686-46-40

Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Longman

Fritsche-Neto R, Resende MDV, Miranda GV, DoVale JC (2012) Seleção genômica ampla e novos métodos de melhoramento do milho. Rev Ceres 59:794–802. doi:

10.1590/S0034-737X2012000600009

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

Gianola D, de los Campos G, Hill WG, et al (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–63. doi: 10.1534/genetics.109.103952

Gianola D, Fernando RL, Stella A (2006) Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–76. doi: 10.1534/genetics.105.049510

Gianola D, van Kaam JBCHM (2008) Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178:2289–2303. doi: 10.1534/genetics.107.084285

Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet 10:381–391. doi: 10.1038/nrg2575

Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128:409–21. doi: 10.1111/j.1439-0388.2011.00964.x

González-Camacho JM, de Los Campos G, Pérez P, et al (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet 125:759–71. doi: 10.1007/s00122-012-1868-9

Grattapaglia D (2014) Genomics of Plant Genetic Resources. pp 467–487

Grattapaglia D, Resende MD V (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255. doi: 10.1007/s11295-010-0328-4

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. Genetics 182:343–53. doi: 10.1534/genetics.108.100289

Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194:597–607.

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12:186. doi: 10.1186/1471-2105-12-186

Hallauer AR, Carena MJ, Miranda Filho J. (2010) Quantitative Genetics in Maize Breeding. Springer

Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: Genomic

selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–43. doi: 10.3168/jds.2008-1646

Heffner EL, Lorenz AJ, Jannink J-L, Sorrells ME (2010) Plant Breeding with Genomic Selection: Gain per Unit Time and Cost. Crop Sci 50:1681. doi: 10.2135/cropsci2009.11.0662

Heffner EL, Sorrells ME, Jannink J-L (2009) Genomic Selection for Crop Improvement. Crop Sci 49:1. doi: 10.2135/cropsci2008.08.0512

Henderson CR (1984) Applications of Linear Models in Animal Breeding. University of Guelph

Heslot N, Akdemir D, Sorrells ME, Jannink J-L (2013) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor Appl Genet 127:463–480. doi: 10.1007/s00122-013-2231-5

Heslot N, Yang H-P, Sorrells ME, Jannink J-L (2012) Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci 52:146. doi: 10.2135/cropsci2011.06.0297

Hickey JM, Gorjanc G (2012) Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 (Bethesda) 2:425–7. doi: 10.1534/g3.111.001297

Hickey JM, Kinghorn BP, Tier B, et al (2012) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol 44:9. doi: 10.1186/1297-9686-44-9

Hill WG (2010) Understanding and using quantitative genetic variation. Philos Trans R Soc Lond B Biol Sci 365:73–85. doi: 10.1098/rstb.2009.0203

Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008. doi: 10.1371/journal.pgen.1000008

Ioannidis JPA (2005) Why Most Published Research Findings Are False. doi: 10.1371/journal.pmed.0020124

Isik F, Li B, Frampton J (2003) Estimates of Additive, Dominance and Epistatic Genetic Variances from a Clonally Replicated Test of Loblolly Pine. For Sci 49:77–88.

Jarquín D, Crossa J, Lacaze X, et al (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor Appl Genet 127:595–607. doi: 10.1007/s00122-013-2243-1

Lander ES, Linton LM, Birren B, et al (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921. doi: 10.1038/35057062

Legarra A, Robert-Granié C, Croiseau P, et al (2011) Improved Lasso for genomic selection. Genet Res (Camb) 93:77–87. doi: 10.1017/S0016672310000534

Legarra A, Robert-Granie C, Manfredi E, Elsen J-M (2008) Performance of Genomic Selection in Mice. Genetics 180:611–618. doi: 10.1534/genetics.108.088575

Lopes MS, Bastiaansen JWM, Harlizius B, et al (2014) A genome-wide association study reveals dominance effects on number of teats in pigs. PLoS One 9:e105867. doi: 10.1371/journal.pone.0105867

Lopez-Cruz M, Crossa J, Bonnett D, et al (2015) Increased Prediction Accuracy in Wheat Breeding Trials Using a Marker x Environment Interaction Genomic Selection Model. G3: Genes|Genomes|Genetics 5:569–82. doi: 10.1534/g3.114.016097

McKeand SE, Jokela EJ, Huber DA, et al (2006) Performance of improved genotypes of loblolly pine across different soils, climates, and silvicultural inputs. For Ecol Manage 227:178–184. doi: 10.1016/j.foreco.2006.02.016

Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157:1819–1829.

Meuwissen THE, Luan T, Woolliams JA (2011) The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J Anim Breed Genet 128:429–39. doi: 10.1111/j.1439-0388.2011.00966.x

Misztal I, Legarra A, Aguilar I (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J Dairy Sci 92:4648–55. doi: 10.3168/jds.2009-2064

Misztal I, Varona L, Culbertson M, et al (1998) Studies on the value of incorporating the effect of dominance in genetic evaluations of dairy cattle, beef cattle and swine. Biotechnol Agron Soc Env 2:227–233.

Morota G, Boddhireddy P, Vukasinovic N, et al (2014) Kernel-based variance component estimation and whole-genome prediction of pre-corrected phenotypes and progeny

tests for dairy cow health traits. Front Genet 5:56. doi: 10.3389/fgene.2014.00056

Morota G, Gianola D (2014) Kernel-based whole-genome prediction of complex traits: a review. Front Genet 5:363. doi: 10.3389/fgene.2014.00363

Morota G, Koyama M, Rosa GJM, et al (2013) Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data. Genet Sel Evol 45:17. doi: 10.1186/1297-9686-45-17

Muñoz PR, Resende MFR, Gezan SA, et al (2014) Unraveling Additive from Non-Additive Effects Using Genomic Relationship Matrices. Genetics genetics.114.171322–. doi: 10.1534/genetics.114.171322

Nishio M, Satoh M (2014) Including dominance effects in the genomic BLUP method for genomic evaluation.

Park T, Casella G (2008) The Bayesian Lasso. J Am Stat Assoc 103:681–686. doi: 10.1198/016214508000000337

Pereira MG, Amaral Jr AT (2001) Estimation of Genetic Components in Popcorn Based on the Nested Design. Crop Breed Appl Biotechnol 1:3–10.

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

Pérez P, Gianola D, González-Camacho JM, et al (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. G3 (Bethesda) 2:1595–605. doi: 10.1534/g3.112.003665

Quesada T, Resende M, Muñoz P, et al (2014) Mapping Fusiform Rust Resistance Genes within a Complex Mating Design of Loblolly Pine. Forests 5:347–362. doi: 10.3390/f5020347

R Core Team (2014) R: A Language and Environment for Statistical Computing.

Ramalho MAP, Abreu A, Santos JB, Nunes JAR (2012) Aplicações da Genética Quantitativa no Melhoramento de Plantas Autógamas. UFLA, Lavras

Ramalho MAP, Santos JB, Zimmermann MO (1993) Genética Quantitativa Em Plantas Autógamas Aplicações Ao Melhoramento do Feijoeiro. UFG, Goiânia

Resende Jr MFR (2014) Genomic Selection in Plant Breeding: Predicting Breeding Values, Non-Additive Effects And Application To Mate-Pair Allocation. University of Florida

Resende Jr MFR, Muñoz P, Acosta JJ, et al (2012a) Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. New Phytol 193:617–624.

Resende Jr MFR, Muñoz P, Resende MD V, et al (2012b) Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190:1503–10. doi: 10.1534/genetics.111.137026

Resende MDV (2002) Genética Biométrica e Estatística no Melhoramento de Plantas Perenes. Embrapa, Colombo

Resende MD V (2008) Genomica Quantitative e Seleção no Melhoramento de Plantas e Animais. Embrapa Florestas

Resende MD V, Resende MFR, Sansaloni CP, et al (2012) Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytol 194:116–28. doi: 10.1111/j.1469-8137.2011.04038.x

Resende MD V, Silva FF, Azevedo CF (2014) Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categorias e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTI-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. 881.

Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478. doi: 10.1186/1471-2164-15-478

Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet 123:218–23. doi: 10.1111/j.1439-0388.2006.00595.x

Su G, Christensen OF, Ostersen T, et al (2012) Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7:e45293. doi: 10.1371/journal.pone.0045293

Sun C, VanRaden PM, O'Connell JR, et al (2013) Mating programs including genomic relationships and dominance effects. J Dairy Sci 96:8014–23. doi: 10.3168/jds.2013-6969

Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic

selection. Genet Sel Evol 42:33. doi: 10.1186/1297-9686-42-33

Tusell L, Pérez-Rodríguez P, Forni S, Gianola D (2014) Model averaging for genome-enabled prediction with reproducing kernel Hilbert spaces: a case study with pig litter size and wheat yield. J Anim Breed Genet 131:105–115. doi: 10.1111/jbg.12070

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–23. doi: 10.3168/jds.2007-0980

Varshney RK, Terauchi R, McCouch SR (2014) Harvesting the promising fruits of genomics: applying genome sequencing technologies to crop breeding. PLoS Biol 12:e1001883. doi: 10.1371/journal.pbio.1001883

Vazquez AI, Rosa GJM, Weigel KA, et al (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. J Dairy Sci 93:5942–9. doi: 10.3168/jds.2010-3335

Venter JC, Adams MD, Myers EW, et al (2001) The sequence of the human genome. Science 291:1304–51. doi: 10.1126/science.1058040

Vitezica ZG, Varona L, Legarra A (2013) On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195:1223–30. doi: 10.1534/genetics.113.155176

White TL, Adams WT, Neale DB (2007) Forest Genetics. CABI Pub.

Wiggans GR, Vanraden PM, Cooper TA (2011) The genomic evaluation system in the United States: past, present, future. J Dairy Sci 94:3202–11. doi: 10.3168/jds.2010-3866

Wilkie AOM (1994) The molecular basis of genetic dominance. J Med Genet 31:89–98.

Willyard A, Ann W, Syring J, et al (2007) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for pinus. Mol Biol Evol 24:90–101. doi: 10.1093/molbev/msl131

Wittenburg D, Melzer N, Reinsch N (2015) Genomic additive and dominance variance of milk performance traits. J Anim Breed Genet = Zeitschrift für Tierzüchtung und Züchtungsbiologie 132:3–8. doi: 10.1111/jbg.12103

Wittenburg D, Melzer N, Reinsch N (2011) Including non-additive genetic effects in Bayesian methods for the prediction of genetic values based on genome-wide markers. BMC Genet 12:74. doi: 10.1186/1471-2156-12-74

Wray NR, Yang J, Hayes BJ, et al (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14:507–15. doi: 10.1038/nrg3457

Yang J, Benyamin B, McEvoy BP, et al (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–9. doi: 10.1038/ng.608

Zeng J, Toosi A, Fernando RL, et al (2013) Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. Genet Sel Evol 45:11. doi: 10.1186/1297-9686-45-11

Zimin A, Stevens KA, Crepeau MW, et al (2014) Sequencing and Assembly of the 22-Gb Loblolly Pine Genome. Genetics 196:875–890. doi: 10.1534/genetics.113.159715