# GENOMIC SELECTION IN WHITE SPRUCE

by

Omnia Gamal El-Dien Ibrahim

BSc, Alexandria University, 2005

MSc, Alexandria University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Forestry)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

March 2017

## Abstract

Tree improvement programs are long-term and resource-demanding endeavors consisting of repeated cycles of breeding, testing, and selection and suffer from protracted testing phases. Phenotypic selection is commonly practiced and often requires trees reaching certain age and/or size resulting in slow accumulation of genetic gain. Open-pollinated (OP) family testing is the simplest and most economical means for screening, evaluating, and ranking large number of candidate parent trees but suffers from inflated additive genetic variance and heritability estimates. This dissertation investigates genomic selection (GS) and its applicability to forestry in selection and progeny testing evaluation.

To address these two applications, I studied yield and wood traits from two white spruce populations, genotyped using Genotyping-by-Sequencing and SNPs array. I investigated the applicability of GS using the Ridge Regression Best Linear Unbiased Predictor (RR-BLUP) and the Generalized Ridge Regression (GRR)) algorithms and validated the derived predictive models in space across three progeny testing sites in interior British Columbia. Moreover, using principal component analysis (PCA), I fitted a multi-traits GS predictive model to address the inter-correlation among the studied attributes. Additionally, the Genomic Best Linear Unbiased Predictor (GBLUP) was used in genetic variance decomposition framework to unravel additive from non-additive genetic variances and I compared the results to that from the traditional pedigree-based (ABLUP) analysis. Differences between the RR-BLUP and GRR predictive models' accuracies were observed indicating that the studied attributes' genetic architecture is complex. Validating the GS's predictive models in space clearly confirmed multi- to single-site superiority as they account for the genotype x environment interaction, commonly observed in forestry evaluation trials. When PCA scores used as multi-trait representatives, GS prediction models produced surprising results where the concurrent selection of negatively correlated traits such as wood density and growth is possible. The genetic variance decomposition indicated that the genomic-based approach outperformed that of the pedigree-based with the successfully separation of additive from non-additive genetic effects. This approach was demonstrated in a single- and extended to multi-site scenario, propelling OP testing to the forefront of forest trees genetic evaluation. In general, the effectiveness of GS was clearly demonstrated as an alternative selection and evaluation method.

## Preface

I participated with my supervisor in identifying and designing of the research program, I conceived the research questions under the consultation of my supervisory committee. I collaborated with other research team members in the data collection stages. I conducted the data analysis, interpretation, and drafted the manuscripts. Chapters in this dissertation have been published or under review in peer review journals:

- **Chapter 2**

Gamal El-Dien O*, Ratcliffe B, Kĺap̌stˇe J, Chen C, Porth I, El-Kassaby YA. (2015). Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC genomics, 16:370.

- **Chapter 3**

Gamal El-Dien O*, Ratcliffe B, Kĺap̌stˇe J, Porth I, Chen C, El-Kassaby YA. (2016). Implementation of the realized genomic relationship matrix to open-pollinated Interior spruce family testing for disentangling additive from non-additive genetic effects. G3: Genes | Genomes | Genetics, 6(3): 743-753.

- **Chapter 4**

Gamal El-Dien O, Ratcliffe B, Kĺap̌stˇe J, Porth I, Chen C, El-Kassaby YA. (2016). Genetic variance decomposition using genomic relationships in interior spruce multi-site open-pollinated family testing. (under review).

The author, Omnia Ibrahim

University of British Columbia

# Table of Contents

## List of Tables

## List of Figures

## Acknowledgements

**IN THE NAME OF GOD, MOST GRACIOUS, MOST MERCIFUL**

My PhD journey is the first enjoyable educational challenge in my life, challenges are a part of my life but enjoying them is a new feeling for me. I stared to realize the meaning of "UBC a place of mind" after beginning the fourth year of my program. It is not only a place where my mind was growing and developing scientifically but also a place where my understanding to a lot of life perspectives changed. This incredible experience really created a paradigm shift not only in my career but also in my whole life aspects.

On presenting this work, I would like to express my endless thanks to my program supervisor Dr. Yousry El-Kassaby for giving me this valuable chance to join his group and creating the perfect environment to smoothly transition from pharmaceutical to genomics perspective. I couldn't find enough words to express my appreciation for his support during this enjoyable and challenging journey and for trusting my capability of doing this transformation. I would like also to thank my supervisory committee members, Drs. Richard Hamelin (UBC), Michael Stoehr (British Columbia Ministry of Forests, Lands and Resource Operations), Jaroslav Klápště (Scion, New Zealand (formerly known as Forest Research Institute Ltd.)), and Charles Chen (Oklahoma State University) for keeping me on the track to finish my thesis and I would like also to thank Dr. Ilga Porth (Université Laval) for her generous support and advises during my program. My thanks are also extended to Dr. Valerie LeMay (UBC) for her valuable courses and advises, being her student and graduate teaching assistance really helped me to build up my biometrics knowledge and teaching expertise, and to Dr. Robert Kozak (UBC) for giving me the chance to be one of his graduate teaching assistants which greatly enriched my experience. Moreover, I would like to thank my funding agencies; the Johnson's Family Forest Biotechnology Endowment, FPInnovations' ForValueNet, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to Dr. Yousry El-Kassaby.

Also, I want to express my great appreciation to Blaise Ratcliffe, my lab mate who is working on the same project, for sharing data, ideas, and valuable comments. I would like to extend my gratitude to the rest of my lab mates for their friendly support: Yang Liu, Susan Song, Faisal Al-harbi, and Frances Thistlethwaite and to the Departmental stuff specially Natasha and Rosemary. I would also like to thank Irina Fundova and Tomas Funda for phenotyping, Tomas Funda and Jiri Korecky for DNA extraction, and Cornell's Genomic Diversity Facility for genotyping (GBS), and Barry Jaquish for access to progeny test trials and sharing phenotypic data.

No words can express my deep thanks to my special family: especially my tender mother Hala who is always the light that showed me my path with all her care, blessings and support without waiting for something in return, my sisters Yasmine and Menna for their unrelenting encouragements especially in my ebbs, my lovely sons Aly, Ahmed and Adam for giving me the motive to come to Canada looking for better future and giving me the power to never give up, and my aunts, uncles and friends for their prayers and support. Finally, I owe very special endless thanks to my dear husband Moustafa for his continuous help, encouragement and support that gave me the strength to continue this work and for considering my success his own success.

**"This thesis could have never been accomplished without their support"**

*To the spirit of my dearest father*

**Gamal EL-Dien**

*To my dear*
*Supervisor*
*Mother*
*Husband*
*Sisters*
*Sons*

# Chapter 1: Introduction

## 1.1 From traditional tree improvement to Genomic Selection (GS)

The tree improvement programs are long-term and resource demanding endeavors consisting of repeated cycles of breeding, testing, and selection following the recurrent selection scheme (White *et al.* 2007). All conventional tree breeding programs are facing three major challenges; the long term breeding cycle, the large progeny test sites, and the late expression of most economic quantitative trait such as wood density (Grattapaglia 2014). The phenotypic-based selection approach is commonly used in the tree improvement programs and often requires trees reaching certain age and/or size resulting in slow accumulation of genetic gain per unit time and cost (El-Kassaby *et al.* 2014).

In the past two decades due to the development of molecular technologies and the discovery of Next-Generation-Sequencing (NGS), a new approach started to be implemented in breeding programs, which is the molecular breeding or genotype based selection. The main concept of molecular breeding is the linkage between genetic markers with Quantitative Trait Loci (QTLs), which are the genes controlling trait variation of interest (Lande and Thompson 1990; Paterson *et al.* 1991). In 1990's breeders started to use this technology in the form of Marker-Assisted-Selection (MAS) "which is the use of a genetic marker for indirect selection of a trait of interest, but this technique required the prior knowledge of the genes or marker associated with the trait of interest (Lande and Thompson 1990; Paterson *et al.* 1991).

As most traits of interest, in breeding programs, are complex quantitative traits controlled by many genes, each with small effect (Fisher, 1918), thus it is very difficult to identify all these genes and as a result the markers associated with them. For that reason MAS was ineffective in both animal and crop breeding and few successes mostly involving oligo-genic traits with simple inheritance (qualitative traits that are controlled by major genes, each with large effect, e.g., disease resistance (Neale & Williams, 1991; Williams & Neale, 1992)), were reported (Stuber *et al.* 1999; Dekkers 2004).

In 2001, the concept of Genomic Selection (GS) was developed by Theo Meuwissen (Meuwissen *et al.* 2001) with the main idea of using all available marker data from the genome as predictors of the phenotype, in other word the genomic estimated breeding value (GEBV) of an individual. So the major advantage of GS is that it doesn't require the identification of neither the QTLs nor the linked markers, thus making this method suitable for the selection of complex quantitative traits, thus creating a paradigm shift from phenotype-based breeding to phenotype-predicted selection. Using GS, selection can be made at any age and/or size, as long as genotypic information can be available; in addition, it has the potential to replace the testing phase, which will shorten the breeding cycles and increase the genetic gain per unit time. GS has replaced conventional progeny testing in dairy cattle breeding (Goddard *et al.* 2010; Wiggans *et al.* 2016) and it was also successfully applied in some of major crop breeding programs (Heslot *et al.* 2012; Poland *et al.* 2012) but GS is still at its infancy in tree breeding programs (Resende, Muñoz, Acosta, *et al.* 2012; Resende, Muñoz, Resende, *et al.* 2012; Zapata-Valenzuela *et al.* 2012; Beaulieu *et al.* 2014; Grattapaglia 2014; Isik, Bartholomé, *et al.* 2015; Isik, Kumar, *et al.* 2015; Bartholomé *et al.* 2016).

Alternatively a smaller subset of markers can be used to estimate realized genomic relationships matrix, known as (G-matrix), using genotypes shared by individuals and standardizes with allele frequencies (VanRaden 2008). Then, the average numerator relationship matrix (***A***-matrix) derived from pedigree (Wright 1922) is substituted by the genomic relationship matrix (***G***-matrix) to predict genomic estimated breeding values (VanRaden 2008). This analysis is known as Genomic Best Linear Unbiased Predictor (GBLUP) and has the potential to be a powerful tool in forest tree breeding programs. Such genomic-based models can capture the Mendelian segregation effect in half- and full-sib families, which is not the case using the expected additive genetic relationships (***A***-matrix) (Zapata-Valenzuela *et al.* 2013). Moreover, GBLUP can remove hidden relatedness resulting in more accurate genetic parameters and variance components estimates.

Open-pollinated (OP) (also known as wind-pollinated) family testing is, by far, the simplest and most economical means for screening, evaluating, and ranking large number of candidate parent trees. OP testing assumes that the tested material are half-sib families, so only additive genetic variance will be possible to be determined using pedigree information (traditional breeding) (White *et al.* 2007). As the assumption of half-sib families is hardly fulfilled, all OP based estimated genetic parameters (e.g., additive genetic variance, heritability, breeding values, etc.) are unreliable, largely reducing the efficiency of this testing method (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994).

White and Interior spruce are the most economically important forest tree species in BC. Interior spruce is a complex of white spruce (*Picea glauca* (Moench) Voss), Engelmann spruce (*Picea engelmannii* Parry), and their hybrids and, because of their similar growing habitats and silvicultural requirements, they are often collectively treated as one species complex (Sutton *et al.* 1991).

In this study I used two OP testing trials: a multi-site Interior spruce growing in British Columbia (N=1,126) replicated over three sites, and a single-site white spruce population growing in Quebec (N=1,649). The British Columbia's populations were genotyped using genotyping-by-sequencing (GBS), while Quebec population were genotyped using array-based SNP genotyping platform and the phenotypic and genotypic data for this populations are available online at (http://datadryad.org/resource/doi:10.5061/dryad.6rd6f) from another study by Beaulieu *et al.* (2014).

## 1.2 Research objectives

This dissertation has two main objectives:

- Present a proof-of-concept for the applicability of genomic selection (GS) in tree improvement programs.
- Increase reliability of Open-Pollinated testing's estimated genetic parameters by using genomic analysis (GBLUP).

## 1.3 Thesis overview

To achieve these objectives, I divided my thesis into three chapters:

### 1.3.1 Chapter 2: Genomic selection (GS) and its validation in space

Here different GS algorithms and imputation methods (for GBS data) were applied to validate GS's predictability in space (in other words, the applicability of a specific-site predictive models to other sites within the same breeding zone). The possibility of fitting a multi-traits GS model was also investigated. The data used in this chapter included 1,126 trees from 25 OP families of interior spruce families replicated over three sites in British Columbia. The phenotypic data were growth and wood density attributes and genotypic data were generated using Genotyping-by-Sequencing (GBS). This study is the first in using GBS on large scale in forestry's research.

### 1.3.2 Chapter 3: Genomic-based vs. pedigree-based approach to genetic variance decomposition in single-site OP white spruce population

In this chapter, the strength of the GS approach known as Genomic Best Linear Predictor (GBLUP) to decompose dominant and epistatic genetic variance from the additive was tested. Commonly, the additive genetic relationship is estimated using the pedigree-based genetic relationship matrix (***A***-matrix; commonly known as ABLUP). Here, I am substituting the ***A***-matrix with the genomic-based relationship matrix (***G***-matrix) in the mixed effect linear model to estimate genetic parameters, including individuals' breeding values. GBLUP enabled us to separate the additive from non-additive genetic effects (dominance and epistasis) through implementation of additive and non-additive marker based relationship matrices, a situation is not possible with pedigree analysis (ABLUP). Our results have showed reliable genetic variance estimates using GBLUP, particularly for OP families where the additive variance is always inflated when pedigree-based analysis is used. It is noteworthy to mention that this study provided the first attempt of such an analytical approach in OP family testing. For this analysis, I used a single-site population of pure white spruce (N=1,694) representing 214 families growing in Quebec with phenotypic data represented by height and wood density and genotypic data generated from SNPs array. The data used for this chapter were available online from the Dryad Digital Repository (http://datadryad.org/resource/doi:10.5061/dryad.6rd6f) (Beaulieu *et al.* 2014)

### 1.3.3 Chapter 4: Extension of the OP testing genetic analysis to multi-site using interior spruce populations from British Columbia

Here, I extended the model developed in chapter 3 from single-site to multi-site using the same population I studied in chapter 2 (i.e., height and wood density and the same genotypes). The advantage of multi-site analysis exists in its ability to account for the genotype x environment effect which increase the reliability of genetics estimates. Additionally, this model extension demonstrated its potential to use on all similar multi-site OP testing programs around the world.

# Chapter 2: Genomic selection (GS) and its validation in space

## 2.1 Introduction

Tree improvement programs are long-term and resource demanding endeavors requiring repeated cycles of selection, breeding and testing. Most of conventional tree breeding programs face major challenges; including, long breeding cycles, large field experiments planted over vast territory, late expression of economic traits (e.g., wood density), and low to medium heritability of traits (Grattapaglia 2014). The phenotypic selection approach coupled with long testing phase often result in slow accumulation of genetic gain per unit time and cost (El-Kassaby *et al.* 2014). Plant breeders adopted Marker-Assisted-Selection (MAS) to take advantage of the linkage disequilibrium (LD) between genetic markers and Quantitative Trait Loci (QTLs) and realized the method's potential to increase breeding efficiency (Lande and Thompson 1990; Paterson *et al.* 1991). Similarly, tree breeders perceived MAS as a means to reduce the time required for phenotypic selection, increasing selection intensity, and improving selection precision particularly for low heritability and late expressing traits as well as its ability to overcome major conventional breeding obstacles such as the long and costly breeding cycle (Neale and Williams 1991; Williams and Neale 1992). However, MAS faced several challenges; as most associations were limited to only specific genetic background due to the rapidly decaying LD in forest trees, the interaction of QTLs effects with the genetic background, the genotype by environment (GxE) interaction, and the fluctuation of the alleles frequency over generations (Strauss *et al.* 1992). The complex nature of quantitative traits (Fisher 1918) rendered MAS ineffective in both animal and crop breeding and few successes mostly involving traits with simple inheritance (e.g., disease resistance) were reported (Stuber *et al.* 1999; Dekkers 2004).

Meuwissen et al. (Meuwissen *et al.* 2001) introduced Genomic Selection (GS) as a method that collectively uses the genome-wide marker data in predicting the phenotype by estimating the genomic breeding values for each individual. The major advantage of GS is that it does not require the identification of the QTLs or linked markers with target traits as all marker effects are estimated simultaneously and used to develop the prediction model for estimating Genomic Estimated Breeding Values (GEBV) for each individual. Thus, this method is suitable for selection of traits with complex genetic architecture as it does not rely on the identification of a single causal variant, rather it fits the genetic effects of all markers regardless of their known functional relevance (Meuwissen *et al.* 2001; Goddard and Hayes 2009). In forest tree breeding context, GS has the ability to predict the phenotype for selecting elite genotypes at early age and developmental stage, thus substantially shortening the breeding cycle and increasing the selection differential, ultimately raising the genetic gain per unit time (Resende, Muñoz, Acosta, *et al.* 2012; Resende, Muñoz, Resende, *et al.* 2012; Zapata-Valenzuela *et al.* 2012; Beaulieu *et al.* 2014). The time savings involve tree testing (for late expressing traits in particular), which is not needed in the next few generations with GS being implemented in the conifer breeding program, thus providing 15-25 years anticipated savings (Beaulieu *et al.* 2014).

The development of Next-Generation-Sequencing (NGS) technologies and the implementation of genetic markers from sequence data in quantitative genetics related to GS, the Genomic Best Linear Unbiased Predictor (GBLUP) (VanRaden 2008), and the unified single-step evaluation approach (also known as HBLUP, single-step combining pedigree and realized kinship information) (Misztal *et al.* 2009) have created novel opportunities for breeding, including forest trees (El-Kassaby and Lstibůrek 2009; El-Kassaby, Cappa, *et al.* 2011; El-Kassaby *et al.* 2014; Isik 2014). Genotyping-By-Sequencing (GBS) (Elshire *et al.* 2011), of the NGS technologies, offers a promising opportunity in studying non-model species including those with large and complex genomes with no assembled reference sequence such as conifers (Chen *et al.* 2013). GBS uses restriction enzymes to allow the sequencing of a reduced subset of the studied genome and the resulting fragments are DNA barcoded to permit multiplexed sequencing. GBS has made genome-wide population studies possible due to the affordability of the method and its capability of resolving tens of thousands of markers scattered throughout the genome.

In this study, using GBS as a genotyping platform, we developed GS prediction models in a dataset of 1,126 Interior spruce trees representing 25 open-pollinated families replicated over three sites in British Columbia (BC), Canada. White and Interior spruce are one of the most economically important forest tree species in BC. Interior spruce is a complex of white spruce (*Picea glauca* (Moench) Voss), Engelmann spruce (*Picea engelmannii* Parry), and their hybrids and, because of their similar growing habitats and silvicultural requirements, they are often collectively treated as one complex (Sutton *et al.* 1991). While white spruce shows transcontinental distribution, the natural distribution of Engelmann spruce is much more limited and scattered and in BC province is confined to the northern part of central BC. Hybridization occurs mainly at mid elevations, where their distributions overlap. Recently, extensive genetic and genomic resources became available for this species (4.9 million scaffolds from the 20.8 giga base pairs draft genome of Interior spruce individual PG29, Birol et al (Birol *et al.* 2013); 21,840 spruce ESTs microarray employed in genetical genomics of interior spruce progenies (Porth *et al.* 2012)).

The objectives of the present study were to: 1) evaluate the efficiency of GBS as a rapid genetic marker genotyping platform for GS studies, 2) investigate different imputation algorithms for GBS data on GS prediction accuracy, 3) compare two GS approaches (Ridge regression best linear unbiased predictor (RR-BLUP) and generalized ridge regression (GRR)), 4) investigate the heterogeneous GxE effect on GS prediction accuracy in space, and 5) use PCA in the comparisons of multi- vs. single-trait GS prediction models.

## 2.2 Materials and methods

### 2.2.1 Experimental population and DNA sampling

For this study, 1,126 38-year-old Interior spruce trees (*Picea glauca* (Moench) Voss x *Picea engelmanni*i Parry ex Engelm.) were sampled from a progeny test trial established by the Ministry of Forests, Lands and Natural Resource Operations of British Columbia Canada, and planted on three sites [Aleza Lake (Lat. 54° 03' 15.7" N, Long. 122° 06' 35.4" W, Elev. 700 mas), Prince George Tree Improvement Station (PGTIS) (Lat. 53° 46' 17.9" N, Long. 122° 43' 07.6"W, Elev. 610 mas), and Quesnel (Lat. 52° 59' 27.2" N, Long. 122° 12' 30.6" W, Elev. 915 mas)]. The sites were established in 1972/73 and consisted of 181 open-pollinated families using 3-year-old seedlings planted at 2.5x2.5m spacing in a complete randomized block design with five or ten blocks and ten or fifteen tree-row-plots, respectively. Twenty-five families were selected based on their superior growth traits and four trees per family from four blocks per site were randomly sampled (maximum of 32 trees per family). Evidence of similar genetic diversity between selected and unselected populations have been reported for spruces, including white spruce (Chaisurisri and El-Kassaby 1994; Stoehr and El-Kassaby 1997). The differences across all the three sites in the relationship between overall X-ray density and growth traits (see below) indicated that the Quesnel site is most while PGTIS least favorable for growing interior spruce (YA El-Kassaby, pers. obs.).

### 2.2.2 Genotyping and SNP selection

DNA extraction was performed from dormant vegetative buds of the sampled trees using a CTAB procedure modified after Doyle and Doyle (Doyle and Doyle 1990). To generate a high-density

SNP profile for the 1,126 spruce DNA extracts, we conducted a multiplexed, high-throughput Genotyping-by-Sequencing (GBS) following Elshire et al. (Elshire *et al.* 2011) and Chen et al. (Chen *et al.* 2013). A 48-plex GBS library comprising of 47 DNA samples and a negative control (without DNA) was prepared and each of the 47 spruce DNA extracts was barcoded. In brief, each DNA extract (500ng) was digested with restriction enzyme *Ape*KI for 2 hours. The details of oligonucleotide sequences for the *Ape*KI barcode adapters and temperature cycles are provided in Chen et al. (Chen *et al.* 2013). Ligation products from each DNA extract were pooled and purified using QIAquick PCR purification kit (Qiagen). The amplified 48-plex libraries were diluted and sequenced (single-end reads only) twice on the Illumina HiSeq 2000 at the Cornell University Genomics Core Laboratory to achieve the sequencing coverage equivalent to 24-plex. Raw DNA short-read sequences were analyzed with a pipeline, the Universal Network Enabled Analysis Kit (UNEAK), tailored to species lacking reference genome information (Lu *et al.* 2013). This SNP detection pipeline is available in TASSEL v5.0 (Bradbury *et al.* 2007). To reduce sequencing error in genotype determination, we set the error tolerance rate to 0.03 (to pass the expected Illumina sequencing error rate at 0.4%). The resulting SNP table was further filtered using minimum value of inbreeding coefficient (mnF = 0.05) and minimum minor allele frequency (mnMAF = 0.05), and finally, SNPs that are present in less than 40% of the samples were eliminated from further analysis.

### 2.2.3 Missing data imputation

To interpret missing values present in the filtered SNP set, five different imputation algorithms were employed: (1) mean imputation (MI), (2) singular value decomposition imputation (SVD:(Troyanskaya *et al.* 2001)), (3) traditional k nearest neighbor (kNN:(Troyanskaya *et al.*

2001)), (4) expectation maximization imputation (EM:(Dempster *et al.* 1977b)), and (5) k-nearest neighbor imputation but newly derived for half-sib family structure (kNN-Fam).

For SVD, the original SNP matrix was used to obtain a set of the *k* most significant eigenvectors of the SNP markers. The *k* eigenvectors were then used as predictors for linear regression estimation of the missing data. SVD was implemented in R (R Core Team 2014) using the "bcv" pakage (Perry 2009). The resultant numerical SNP values (*x*) were further classified into three separate genotype classes, -1, 0, and 1. The classification algorithm was taken as a modified k-means algorithm (Hartigan and Wong 1979), with the centroids set at -1 ($k_1$), 0 ($k_2$), and 1 ($k_3$). The assignment of genotypes was done by satisfying:

$$argmin(SS) = \sum_{i=1}^{k} \sum_{x \in S_i} \| x - k_i \| \qquad [1]$$

where (1) defines the minimum distance for the SNP value from the centroids.

For traditional kNN, the missing values were replaced with the weighted average of SNP values at the k closest SNP markers. The distances between all possible pairs of markers were computed by Euclidean distance. We selected five families (6, 11, 17, 21, and 47) to test the imputation accuracy, as well as the efficiency of iterations for convergence (2, 3, 5 and 10 iterations for SVD; for EM, we tested the distance between the new estimate and the previous values less than 0.01). K = 10 and 20 were selected for accuracy estimates for kNN imputation. All iterations reached convergence criteria that were used in (Rutkoski *et al.* 2013), however they resulted in different accuracies (Table 2.1).

The kNN-Fam algorithm is derived from the kNN method of Troyanskaya et al.(Troyanskaya *et al.* 2001). Missing values in the SNP table were first replaced with the mean of the locus by MI. A standardized genomic similarity matrix for all samples was calculated based on VanRaden (VanRaden 2008) and the Euclidean distance between SNP markers was defined following  Rutkoski et al.(Rutkoski *et al.* 2013). Instead of the classic k-nearest neighbor method, where

$$\hat{y} = \left(\frac{1}{K}\right) sigma(y) \qquad [2]$$

the missing SNP values were replaced with:

$$\hat{y} = mode\left(\frac{1}{K1+K2} y\right) \qquad [3]$$

where *K1* is the number of neighbors within the half-sib family based on the genomic similarity, *K2* is the number of neighbors from outside the family based on the Euclidean distance, and *y* is the original locus mean. We conducted exhaustive search for the optimal values of *K1* and *K2*, by permutating *K1* through 1 to 30 (the nearest neighbor set as 1, and then 2, 5, 10, 15, 20 to the maximum family size of 30), and *K2* from 1 to 250, as the total sample size of the panel is 1,126. The accuracy of kNN-Fam imputation was conducted for each permutation by randomly masking one million known data points from the filtered SNP table of the 5 selected families, and calculating the percentages of markers being imputed back to the correct SNP values.

### 2.2.4 Phenotypic data

The studied trees were phenotyped for (a) two growth traits (height in m (HT) and diameter at breast height in cm (DBH) which were subsequently used to estimate stem volume in m$^3$ (VOL) following Millman's formula (Millman 1976)) and (b) three wood quality attributes (wood density in kg/m$^3$ using X-ray densitometry (WD$_{\text{X-ray}}$), resistance to drilling (WD$_{\text{Res}}$), and acoustic velocity in km/s (V$_{\text{Dir}}$)) (El-Kassaby, Mansfield, *et al.* 2011). Furthermore, WD$_{\text{X-ray}}$ and V$_{\text{Dir}}$ were used to derive the dynamic modulus of elasticity (MoE$_{\text{d}}$) (Auty and Achim 2008). WD$_{\text{X-ray}}$ is commonly used to estimate wood density using increment cores extracted from the sampled trees, while WD$_{\text{Res}}$ and V$_{\text{Dir}}$ represent indirect (i.e., non-invasive) methods that rely on wood density for either creating resistance during drilling or the speed of transmitting sound though the wood, respectively (El-Kassaby, Mansfield, *et al.* 2011).

### 2.2.5 Estimated breeding values (EBV)

The breeding value for each tree was estimated using ASReml v.3 using two different mixed linear models (Gilmour *et al.* 2009). The first used the pooled populations to estimate multi-site breeding values (MSEBV), while the second was used to estimate single-site breeding values (SSEBV) as follows:

Multi-site model:

$$y = \boldsymbol{X}b + \boldsymbol{Z}_1 a + \boldsymbol{Z}_2 sb + \boldsymbol{Z}_3 sa + e \qquad [4]$$

where *y* is the phenotypic measurement of the analyzed trait, *b* is a vector of fixed effect (i.e., the overall mean and the site effect), *a* is a vector of random additive effect of individual trees ~N(0, $\boldsymbol{A}\sigma^2_a$), *sb* is a vector of the random effect of block within site ~N(0, $\boldsymbol{I}\sigma^2_{sb}$), *sa* is a vector of random site x genotype interaction ~N(0, $\boldsymbol{I}\sigma^2_{sa}$), *e* is a vector of random residual effect ~N(0, $\boldsymbol{I}\sigma^2_e$), and $\boldsymbol{X}$ and $\boldsymbol{Z}_1$-$\boldsymbol{Z}_3$ are incidence matrices assigning fixed and random effects to each observation and $\boldsymbol{I}$ and $\boldsymbol{A}$ are the identity and average numerator relationship matrices, respectively. Narrow-sense heritability was calculated as $h^2 = \sigma_a^2/ (\sigma_a^2+\sigma_{sa}^2+\sigma_e^2)$ for the multi-site model.

Single-site model:

$$y = \boldsymbol{X}b + \boldsymbol{Z}_1 b + \boldsymbol{Z}_2 a + e \qquad [5]$$

This model is identical to the multi-site mixed linear model but without all terms related to site (site, block nested within site, and site x genotype interaction). Narrow-sense heritability was calculated as $h^2 = \sigma_a^2/ (\sigma_a^2 +\sigma_e^2)$. Additionally, Genomic Best Linear Unbiased Predictor (GBLUP) (VanRaden 2008) was used to estimate the narrow-sense heritabilities of the traits for single and multi-site using genotypes from imputed data produced by the EM algorithm with 30% missing data. This analysis was performed by substituting average numerator relationship matrix with marker-based relationship matrix (VanRaden 2008) using observed allele frequencies.

### 2.2.6 Genomic selection analyses

The SNP effects were estimated on the basis of two different methods: 1) Ridge Regression Best Linear Unbiased Predictor (RR-BLUP) implemented in R package rrBLUP (Endelman 2011) and 2) Generalized Ridge Regression (GRR) implemented in R package bigRR (Shen *et al.* 2013). In both cases the following mixed linear models were fitted:

$$y = \boldsymbol{X}\beta \ + \boldsymbol{Z}b + e \qquad [6]$$

where $y$ is the vector of EBV, $\beta$ is the vector of fixed effect which is the overall mean, $b$ is the vector of random SNP effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices for $\beta$ and $b$, respectively, $\boldsymbol{X}$ is a vector of 1 while $\boldsymbol{Z}$ was built from (-1, 0, 1) for aa, Aa and AA, respectively. The codes for $\boldsymbol{Z}$ were standardized according to the allele frequency using VanRaden's method (VanRaden 2008). $\beta$ and $b$ are estimated simultaneously using Henderson's mixed model equation (MME) (Henderson 1953):

$$\begin{pmatrix} \mathrm{X'X} & \mathrm{X'Z} \\ \mathrm{Z'X} & \mathrm{Z'Z} + \lambda \mathrm{I} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} \mathrm{X'y} \\ \mathrm{Z'y} \end{pmatrix} \qquad [7]$$

where $\lambda = \hat{\sigma}_e^2 \ / \ \hat{\sigma}_b^2$ is the shrinkage parameter for the random SNP effects, so all the SNPs will have the same shrinkage magnitude, in other words, all are penalized to the same degree. In GRR, the SNPs with small effects are more penalized. The first step in GRR is an ordinary RR, then it again uses MME to fit the heteroscedastic model:

$$\begin{pmatrix} \mathrm{X'X} & \mathrm{X'Z} \\ \mathrm{Z'X} & \mathrm{Z'Z} + \mathrm{diag}(\lambda) \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} \mathrm{X'y} \\ \mathrm{Z'y} \end{pmatrix} \qquad [8]$$

where diag (**λ**) is the diagonal matrix of SNP specific shrinkage parameters estimated as $\lambda_j = \hat{\sigma}_e^2 \; / \; \hat{\sigma}_{bj}^2$ , where $\hat{\sigma}_{bj}^2$ is variance attributed to j$^{th}$ SNP and is estimated as:

$$\hat{\sigma}_{bj}^2 = \frac{\hat{b}_j^2}{1-h_{jj}} \qquad [9]$$

where, $b_j$ is the SNP effect, and $h_{jj}$ is the $(n + j)^{th}$ diagonal element of the matrix $\boldsymbol{H} = \boldsymbol{T}\,(\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'$, where

$$T = \begin{pmatrix} X & \mathrm{Z} \\ 0 & \mathrm{diag}(\lambda) \end{pmatrix} \qquad [10]$$

$\hat{\sigma}_{bj}^2$ is needed as it represents the form of implemented variable selection.

### 2.2.7 Cross-validation, predictive accuracy and type-b genetic correlation

The predictive accuracy was estimated using a 10-fold cross-validation approach with 20 replications. In each replication, the data were randomly divided into 10 subsets (folds) and each one was used as validation population (representing 10% of the data set), while the remaining 9-folds were used as the training population (90% of the data set) to fit the GS model. This process was repeated 20 times with random assignment of the data to the 10 folds (Gianola *et al.* 2011; González-Camacho *et al.* 2012; Crossa *et al.* 2013). One advantage of this scheme is that it provides the degree of uncertainty (i.e., standard error) around these point estimates. In all the replicates, the models were fitted to the training data set and used to predict the GEBV of the validation data set by multiplying the vector of the marker effect estimated from the training population with the incidence matrix ***Z*** of the individuals in the validation population and summing over the estimated general mean:

$$\hat{y}_j = \hat{u} + \sum_i Z_{ij}\, \widehat{m}_i \qquad [11]$$

where *u* is intercept, ***Z*** is genotype at the i$^{th}$ locus of the j$^{th}$ individual and *m* is the marker effect. The accuracy of GS to predict the breeding value (BV) was estimated as the correlation of the vector of GEBV for all individuals (predicted from the validation step) with their estimated BV (MSEBV or SSEBV according to the validation scenario). As we used 20 replicates, we obtained 20 estimates for prediction accuracy and we estimated means and standard errors for these estimates. The developed models were validated under the following four scenarios, namely, (1) within site, (2) in all 6 possible combinations for cross-validation comparisons across sites, (3) as a multi-site population, where training and validation populations were derived from the combined population for cross-validation and (4) again as a multi-site population, but where the entire multi-site population was used as training population and the individual site as validation population. Moreover, we estimated the type-b genetic correlation across sites, which is the additive genetic correlation between the traits measured on different individuals from the same genetic group but present in different environments, using a method described by Burdon (Burdon 1977).

### 2.2.8 Multi-trait GS model

We applied Principle Component Analysis (PCA) to distil the correlated variables (EBV) into a set of linearly independent variables (i.e., the principal components (PCs)). We used HT, DBH, $V_{Dir}$, $WD_{Res}$, and $WD_{X\text{-}ray}$ EBVs as variables to determine the PCs that best express these phenotypes and used their score as a new phenotype in subsequent RR-BLUP GS model for the multi-site scenario using the kNN-Fam imputation.

### 2.2.9 ABLUP vs. GBLUP elite genotype selection comparison

Notwithstanding the relatively small number of 25 open-pollinated families under investigation, to illustrate the benefits of incorporating genomic information in selection, we conducted a selection exercise of 40 elite genotypes for inclusion into a hypothetical production population (seed orchard) following the group merit selection scheme of Lindgren and Mullin (Lindgren and Mullin 1997). Group merit selection is founded on penalizing the average BV of a selected subset by increasing the weight on the entire group co-ancestry (measured by co-ancestry coefficient) to reach a desired ‘‘status number ($N_s$)’’ (Lindgren *et al.* 1996) which is an approximation of the effective number of parents ($N_e$) (i.e., measure of diversity). In this method, the co-ancestry coefficients are estimated from the pedigree values of the selected individuals (ABLUP) while in the GBLUP case, we used the marker-based relationship matrix (VanRaden 2008) to approximate the co-ancestry of the selected individuals and their diversity was estimated by the number of founder genome equivalents ($N_{ge}$: (Caballero and Toro 2000)).

## 2.3 Results

### 2.3.1 Genotyping, missing data imputation, and selection of imputation method

In this study, 1,126 38-year-old Interior spruce trees (*Picea glauca* (Moench) Voss x *Picea engelmannii* Parry ex Engelm.) originating from 25 open-pollinated families selected for their superior growth traits were sampled from the progeny test trial planted on three sites, (1) Aleza Lake, (2) Prince George Tree Improvement Station (PGTIS), and (3) Quesnel. A cost-effective NGS technology, genotyping-by-sequencing (GBS), was employed for genotyping a 20GB unassembled genome such as spruce. After two 48-multiplexed sequencing passes, a total of 4,798,791,310 good barcoded reads was generated, and the median of read depth per site was at 3.92 (averaged 4.58±4.28). TASSEL UNEAK SNP calling pipeline was used to determine SNP polymorphism for these 1,126 spruce trees, resulting in a large genotype table of 1,232,406 SNPs (Lu *et al.* 2013; Chen *et al.* 2013) . Typical to GBS, a low coverage sequence platform, many markers tended to have missing data even after the repeated sequencing of all studied trees (see Discussion, for more details). From the identified 1,232,406 SNPs, the applied imputation methods and filtering (minimum minor allele frequency of 0.05) used produced genotyping files ranged from 8,868 (MI-30% and EM-30%) to 62,618 (kNN-Fam-60%) SNPs (Table 2.2). Imputation accuracy ranges from 0.77 (SVD 10 iterations) to 0.82 (SVD with 2 iterations). On average, SVD with 2 iterations produced the best accuracy in the four currently existing methods: MI, SVD, EM and kNN. Using K's (in K-nearest neighbors) from family versus non-family members, accuracy for kNN-Fam imputation ranged from 0.77 to 0.85. In general, including more family members resulted in higher accuracy (Table 2.3); however, imputation accuracy remained unchanged (and did not improve), when the number of non-family members that was included was larger than the family size. The best imputation accuracy gained was at K1 = 5 and K2 = 20, which represented the K values used in this study for imputing the whole SNP table (Table 2.3). As a result, we chose kNN-Fam over kNN of Troyanskaya et al. (Troyanskaya *et al.* 2001) due to its slight superiority in accuracy. The SNP table imputed with this method is referred to as kNN-Fam.

The selection of specific imputation methods for genomic selection analyses were restricted to the method with greater GS accuracies within the same percentage of missing data class (i.e., 30% vs. 60%). For the 30% missing data, the EM-30% produced greater accuracy than MI-30%, similarly for the 60% missing data, the kNN-Fam-60% and SVD-60% produced better accuracies comparing to MI-60%; however, the kNN-Fam-60% was superior to SVD-60% (see below). This comparison was done based on GS prediction accuracies produced for the two GS models and the seven studied traits for both single- and multi-site scenarios (see below).

### 2.3.2 Traits' heritability

Using genotypes resulting from the EM-30% algorithm imputed data, the narrow-sense heritabilities of the traits estimated from the pedigree (ABLUP, i.e. the conventional BLUP model using the pedigree-based relationship matrix) and genomic best linear unbiased predictors

(GBLUP using the genomic-based realized kinship matrix) produced several broad generalizations that include: 1) single- and multi-site heritabilities were higher for ABLUP than those from their GBLUP counterparts, 2) multi-site heritabilities were lower than that of a single site for both ABLUP and GBLUP, 3) trait heritabilities varied among sites for both ABLUP and GBLUP; however, the differences were lower for the GBLUP than that of the ABLUP, 4) the Quesnel site produced higher heritabilities than PGTIS and Aleza Lake, yet they have some overlapping ranges, and 5) standard error estimates of heritabilities obtained from ABLUP were higher than those from GBLUP for single- and multi-site (Table 2.4). Lower GBLUP heritabilities were expected as ABLUP tended to inflate the estimates as the pedigree based analysis assumptions are often violated due to mating pattern, relatedness built-up due to population history, and inability to separate common environment effect from genetics.

### 2.3.3 Prediction accuracy for different GS models and imputation methods

The accuracy of GS models (RR-BLUP and GRR) in predicting the GEBV were evaluated for the seven studied traits using all imputation methods (30% missing data: MI and EM, and 60% missing data: MI, kNN-Fam, and SVD) and over the four cross-validation scenarios: 1) within each individual site, 2) cross-site (all possible combinations), 3) within multi-site (the three sites combined), and 4) the multi-site population in predicting individual site (see below).

#### 2.3.3.1 Within site GS accuracies

Across all imputation methods (30% and 60% missing data), the RR-BLUP produced higher within site GEBV accuracies than the GRR (Tables 2.5 and 2.6, Figures 2.1 and 2.2). In general, the RR-BLUP produced higher accuracies than the GRR (100 out of the possible 105 comparisons for both GS models) and this was also mirrored by their standard error estimates (Tables 2.5 and 2.6). Within the 30% missing data imputation methods, the EM-30% produced greater accuracy than MI-30% for all traits for RR-BLUP (traits averages were 0.51, 0.50, and 0.46 as opposed to 0.52, 0.51, and 0.46 for PGTIS, Aleza Lake, and Quesnel sites, respectively) and GRR (averages were 0.49, 0.43, and 0.41 vs. 0.49, 0.46, and 0.41 for PGTIS, Aleza Lake, and Quesnel sites, respectively) (Table 2.5). The 60% missing data imputation methods produced similar GS prediction and confirmed the superiority of the RR-BLUP over GRR and additionally highlighting the better accuracies for kNN-Fam-60% compared to MI-60% and SVD-60% (Table 2.6).

#### 2.3.3.2 Cross-site GS accuracies

Unlike within site cross-validation, testing the applicability of a GS model for a specific site to predict the GEBV of other sites generally produced lower accuracies for both models (RR-BLUP and GRR) (Figures 2.1 and 2.2, Table 2.9). This is expected due to the GxE interaction even when the three sites are located within one breeding zone (Prince George Seed Planning Zone (http://www.for.gov.bc.ca/hfd/pubs/docs/mr/annual/ar_1995-96/pspzm.htm)). For simplicity, in this section we will restrict the cross-sites comparisons to the imputation method with the highest number of SNPs (i.e., kNN-Fam-60% (62,198 SNPs)), and the GS model with highest accuracies (i.e., RR-BLUP (Figure 2.1)). Over the seven studied traits, the RR-BLUP model produced cross-site validation accuracies ranging from 0.16 and 0.23 when PGTIS was used to predict the GEBV of Aleza Lake (1→2), 0.13 and 0.24 for 2→1, 0.01 and 0.32 for PGTIS to predict Quesnel (1→3), 0.0 and 0.38 for 3→1, 0.06 and 0.36 for 2→ 3, and 0.03 and 0.39 for 3→ 2 (Figure 2.2, Table 2.9). The estimated type-b genetic correlations between sites mimicked the trend observed for cross sites GS accuracy with their Pearson-product-moment correlations ranging between 0.94 and 0.99 (P<0.05) over the seven studied traits for the kNN-Fam-60% imputation method (Figure 2.3).

#### 2.3.3.3 Within multi-site GS accuracies

Similar to within site assessment, the within multi-site cross-validation produced higher GEBV accuracies for RR-BLUP as compared to GRR and this increase in accuracy persisted across all 30% and 60% missing data imputation methods (Table 2.7). Comparisons between imputation methods revealed that EM-30% and kNN-Fam-60% produced better accuracies (Table 2.7, Figures 2.1 and 2.2). Again, we will restrict the GEBV accuracy comparisons to the kNN-Fam-60% imputation method as it uses the largest number of SNPs (62,198 SNPs). On average and across the seven studied traits, GS accuracies ranged between 0.62 and 0.77 for both RR-BLUB and GRR (Table 2.7). The span of this range is far greater than the one observed within sites and cross-sites validation (Tables 2.4, 2.5 and 2.6). These estimates represent the most realistic accuracies as they accommodated the GxE interaction and, furthermore, were produced with a large training population size (90% of the total N = 1,126).

#### 2.3.3.4 Single- vs. multi-site accuracies

When the meta-population was used to predict the GEBV for each individual site, the observed accuracies were high with Aleza Lake producing the highest accuracies (average over the 7 traits of 0.49 for RR-BLUP and GRR) followed by Quesnel (averages of 0.46 and 0.45 for RR-BLUP and GRR, respectively) and PGTIS which produced the lowest accuracies (average of 0.42 for both RR-BLUP and GRR) (Table 2.8). These accuracies are higher than those observed for the cross-site validation (Table 2.8, Figures 2.1 and 2.2).

### 2.3.4 Multi-trait GS prediction models

The first three principle components, PCA1-3, collectively accounted or 86% of the total phenotypic variation and individually accounted for 44, 25, and 17%, respectively. PCA1 produced significant (*P*<0.002 - 0.0001) loading for all the studied traits and was positive for height (HT) (0.69), diameter at breast height (DBH) (0.80), and acoustic velocity ($V_{Dir}$) (0.09) and negative for wood density using X-ray densitometry ($WD_{X\text{-}ray}$) (-0.71) and wood density using resistance to drilling ($WD_{Res}$) (-0.75). PCA2 produced interesting results with significant ($P<0.0001$) and positive loadings for HT (0.39), $V_{Dir}$ (0.92), and $WD_{X\text{-}ray}$ (0.49). Similarly, PCA3 produced significant ($P<0.0001$) and positive loadings for HT (0.46), DBH (0.38), $WD_{X\text{-}ray}$ (0.19) and $WD_{Res}$ (0.64). The fact that growth and wood quality traits produced significant and positive loadings, even if it is for PCA2 and PCA3, is interesting as it creates concurrent selection opportunities for yield and wood quality traits that are commonly known to be negatively correlated. The two GS models produced high prediction accuracies for PCA1 with 0.72±0.001 and 0.71±0.001 for RR-BLUP and GRR, respectively. Similar results were observed for PCA 2 (RR-BLUP: 0.65±0.001 and GRR: 0.64±0.001) and PCA3 (RR-BLUP: 0.57±0.001 and GRR: 0.55±0.002) using the multi-site GS model.

### 2.3.5 ABLUP vs. GBLUP elite genotype selection comparison

Expectedly, across all the range of genetic gain penalties, the selection of 40 elite individuals yielded ABLUP genetic gain higher than that of the GBLUP with percentage increase between 9.2 and 14.6% for 100 and 1,000 penalty classes, respectively (Figure 2.4). Naturally, any increase in co-ancestry is associated with increase in genetic gain; however, the GBLUP offers greater flexibility for elite genotype selection than the ABLUP as the effective number of genomic equivalent provides a continuum for selection as opposed to the pedigree-based status number which offers only two options of relatedness (unrelated or half-sibs).

## 2.4 Discussion

### 2.4.1 GBS and imputation methods

The utilization of NGS technology, and GBS in particular, provides a low cost opportunity for genomic studies for non-model species (Chen *et al.* 2013). In the present study, GBS produced exceedingly large number of SNPs (1,232,406); however, the low coverage nature of the technique has substantially reduced the available SNPs for analyses due to missing data. Missing data could also result from either the absence of the restriction site in the genomic sequence or due to technical issues associated with DNA digestion or PCR amplification (Wang *et al.* 2011; Pan *et al.* 2015). Out of the five imputation methods used, the expectation maximization (EM-30%: (Dempster *et al.* 1977b)) and the newly developed half-sib family-based k-nearest neighbor (kNN-Fam-60%) method resulted in 8,868 and 62,198 SNPs, respectively, and produced the greatest accuracies (Figure 2.1, for kNN-Fam-60%). We used the EM-30% imputation method in estimating the trait heritabilities employing the GBLUP approach (VanRaden 2008), while all described imputation methods were used to evaluate the GS models across all described scenarios. We believe that the higher GEBV accuracies attained from the kNN-Fam imputation method are attributable to the method's capacity of recovering resemblance among individuals within families. In addition, kNN-Fam method proportionately weights family structure and the underlying LD of SNPs, which is also likely contributing to the slightly higher predictability due to its strength of simultaneously capturing identical-by-state with the variants in LD with the causal genes (Solberg *et al.* 2008).

### 2.4.2 Heritability estimates

Treating the offspring from open-pollinated families as half-sibs is often associated with inflated heritability estimates, resulting in an exaggeration of the expected genetic gain (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994). In the present study, heritability estimates obtained from the ABLUP were higher than those from the GBLUP (Table 2.4), highlighting the advantages of incorporating genomic information in standard quantitative genetic analyses (VanRaden 2008) to obtain realistic estimates of breeding values and genetic gain (see ABLUP vs. GBLUP elite genotype selection comparison below).

Our results are similar to those reported for another open-pollinated white spruce progeny trial in Québec, Canada (Beaulieu *et al.* 2014).While heritability estimates were population-specific, slight differences in GBLUP-based heritability estimates for wood density ($WD_{X\text{-ray}}$) and height (36- vs. 22-year-old height) were observed between the two studies (wood density: 0.18 vs. 0.24 and height: 0.20 vs. 0.16) (Beaulieu *et al.* 2014). Additionally, our results suggest that the trait heritability has only limited effect on the prediction accuracy (PA) as diameter at breast height (DBH) and stem volume (VOL) showed high multi-site RR-BLUP predictability despite their low heritability estimates (DBH: $h^2 = 0.07$ and PA = 0.77; VOL: $h^2 = 0.09$ and PA = 0.73), results consistent with those reported for loblolly pine (*Pinus taeda*) (Grattapaglia and Resende 2011; Zapata-Valenzuela *et al.* 2012).

### 2.4.3 GS models

GS models suffer from the “large *p*, small *n*” problem, where the number of predictor effects *p* exceeds by far the number of observations *n* (*p*>>*n*). A variety of statistical methods were proposed to handle this issue and they can be classified into three major categories: shrinkage models, Bayesian methods (including variable selection), and semi- or non-parametric methods such as support vector regression and random forest regression. Those methods are different in their assumptions regarding the genetic architecture of the tested traits (Lorenz *et al.* 2011; Grattapaglia 2014). RR-BLUP, the most common shrinkage model, assumes that the trait is controlled by many genes each with small effects, thus is suitable for traits following the infinitesimal model (Fisher

1918). RR-BLUP assumes that all marker effects are random, normally, and identically distributed and have a common variance, thus all the effects will be equally shrunken toward zero(Lorenz *et al.* 2011; Shen *et al.* 2013; Grattapaglia 2014). This approach was described previously by Meuwissen et al. (Meuwissen *et al.* 2001) and termed SNP-BLUP. In GS and genome wide association studies (GWAS), it is not realistic to use common shrinkage effects for all fitted SNPs across the genome as not all markers will be linked to functional genes and not all gene effects are normally distributed (Meuwissen *et al.* 2001). To overcome this assumption, the Bayesian methods were developed to provide more flexibility in modeling oligogenic traits (i.e., traits that are controlled by few genes each with large effects) (Lorenz *et al.* 2011); however, these methods are computationally demanding (Hofheinz and Frisch 2014). A new, fast, deterministic, and flexible Ridge regression method was suggested by Shen et al. (Shen *et al.* 2013) known as the generalized Ridge regression (GRR). The main difference between RR-BLUP and GRR is that a SNP-specific shrinkage will be used instead of the common shrinkage effect (Shen *et al.* 2013), which is more realistic and more suitable to model oligogenic traits and represents a viable alternative to Bayesian models (Troyanskaya *et al.* 2001).

Our results showed that GRR produced either similar or even lower prediction accuracies as compared to RR-BLUP, which indicates that marker selection by giving different degree of penalization through the application of different shrinkage effects is inadequate for the tested traits. This provides evidence that the tested traits (growth and wood quality) follow the infinitesimal model. Moreover, experimental results in both plants and animals suggested that RR-BLUP provides the best adjustment/compromise between the computational effort and the prediction efficiency (Lorenz *et al.* 2011). This supports the notion that most of the economically important traits are complex and quantitative in nature (i.e., follow the infinitesimal model). For example, in loblolly pine, Resende et al.(Resende, Muñoz, Resende, *et al.* 2012) evaluated RR-BLUP and three Bayesian models across 17 traits related to growth, development, and fusiform rust resistance and the resulting prediction accuracies were marginally different across the four models, except for rust resistance, an oligogenic trait, where the Bayes A and C models resulted in moderately larger performance than RR-BLUP.

### 2.4.4 Within site vs. within multi-site validation

The multi-site cross-validation produced higher prediction accuracies as compared to single-sites (Tables 2.5, 6 and 2.7, Figure 2.1) as the multi-site training population is three times larger than any of the single-site models, resulting in more accurate estimation of marker effects and this is consequently reflected in higher prediction accuracy and precision (Lorenz *et al.* 2011; Grattapaglia 2014). Previous GS studies conducted on plant and animal populations clearly demonstrated the role of training population size on prediction accuracy and illustrated the importance of the training population size as compared to the number of markers used in the models, thus supporting the present study results (VanRaden *et al.* 2009; Luan *et al.* 2009; Lorenzana and Bernardo 2009). In forestry context, our results are also consistent with prediction accuracies obtained for growth and wood quality attributes in loblolly pine and Eucalyptus (Resende, Muñoz, Acosta, *et al.* 2012; Resende, Resende, *et al.* 2012; Zapata-Valenzuela *et al.* 2012). However, comparing the prediction accuracies between our study and those from the Québec white spruce open-pollinated progeny trial is of interest as the experimental settings were somewhat similar (Beaulieu *et al.* 2014). Height, wood density, and dynamic modulus of elasticity were common traits between the two studies; however, their prediction accuracies were lower than in the present study (height: 0.17 vis. 0.63, wood density: 0.33 vis. 0.64, dynamic modulus of elasticity: 0.21 vs. 0.67). In general, the lower prediction accuracies in the Québec study across all the traits compared to our and other tree species studies, is mainly due to the considerably larger number of tested families (214 vs. 25 families) which resulted in higher $N_e$ (effective population size). It is also worth mentioning that we used the EBV as opposed to the raw phenotype in training our GS models; this could have also contributed to the observed differences.

### 2.4.5 Cross-site validation

The economic and ecological importance of interior spruce to British Columbia promoted thorough understanding of the various ecological regions of the species and subsequently 6 unique Seed Planning Zones (SPZs) were identified (Bukley Valley, East Kootenay, Nelson, Prince George, Peace River, and Thompson Okanagan). To date, most forestry GS studies were conducted within the confines of a single "environment model" similar to those GS studies conducted in animal breeding programs where the assumption of a common environment was invoked. The assumption of "common environment" is not suitable in forestry as estimates of GxE, even within a single breeding zone, are high (Burdon 1977) and this motivated breeders to evaluate the performance of a specific genotype or family across different environments to identify generalists for their inclusion in seed production populations (Annicchiarico 2002). For the successful implementation of GS in tree breeding, it is essential that GS models remain accurate across sites, at least within the dedicated breeding zone. Only two out of the published four GS studies in forest tree tested GxE interaction, these include loblolly pine (Resende, Muñoz, Resende, *et al.* 2012) and white spruce (Beaulieu *et al.* 2014). In the present study, we used data from three sites within the Prince George breeding zone and the observed prediction accuracies of a single site to predict another site were generally low (Figures 2.1 and 2.2). The observed reduced prediction accuracies across sites were lower than those obtained from the white spruce and loblolly pine studies. Thus, it is important to pay considerable attention to the structure of the training population; hence the developed models reflect the underpinning forces affecting trait expression and their response to sites heterogeneities.

### 2.4.6 Multi-trait GS prediction models

GS models are trait-specific and do not lend themselves to multi-trait selection as does index selection method which maximizes the correlation between the index score of an individual and its breeding value (Hazel 1943). Yet, selection indices require prior knowledge about the economic value of the traits for proper scaling before optimum phenotypic weights can be estimated. The use of Principle Component Analysis offered an opportunity to handle a set of correlated variables by reducing the dimensionality to a set of uncorrelated ones (i.e., principal components). Negative genetic correlations between yield and wood quality traits are commonly observed (Bouffier *et al.* 2008) and the results from PC1 which accounts for 44% of the total variation confirmed these observations. However, while yield and wood quality are known to act in antagonizing fashion, the results based on PC2 and PC3, albeit collectively accounting for 42% of the total variation, created interesting opportunities for the concurrent selection for both traits without any adverse effect associated with the known negative correlations. It seems that PC2 and PC3 accessed different combinations of SNPs (i.e., causal genes) that work in the same direction.

While we did not consider any prior economic knowledge for weighing in constructing the PCs, the results from PC2-3 clearly demonstrated that it is (to a certain extent) also possible to artificially co-select such attributes that are commonly known to be negatively correlated in the same positive direction. Considering economic weights for traits during constructing selection indices can result in changing the magnitude of genetic correlation among these traits as a consequence of selection. This change in genetic correlation is expected to change SNP effects and thus frequent training is required for GS model to be effective over generations. Finally, our objective of using PCA is to offer a simple method that accounts for the inter-relation (genetic correlation) between the studied traits and provide an opportunity for further expansions that consider economic weights.

### 2.4.7 ABLUP vs. GBLUP elite genotype selection comparison

The observed genetic gain differences between the ABLUP and GBLUP across all co-ancestry penalties were not surprising as heritability, breeding value of an individual, and genetic gain estimates are expected to be higher in open-pollinated populations due to the ABLUP inability to ascertain the true genetic relationship among offspring (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994). On the other hand, GBLUP relies on estimating the realized kinship which provides a more accurate ascertainment of the genealogical relationships among members of an open-pollinated family and thus, resulting in more realistic gain estimates due to adjustment for Mendelian sampling term (Hayes *et al.* 2009a). Our results are similar to those reported in the Québec white spruce study as they consistently produced higher gains from pedigree- vs. marker-based methods (Beaulieu *et al.* 2014).

It should be pointed out that the Bulmer effect (i.e., reduction in response to selection) would be similar for ABLUP and GBLUP and thus the response to selection for both methods will be similarly affected irrespective of the breeding values estimation method used (Van Grevenhof *et al.* 2012). If genomic selection effectively reduces generation interval, then in the forestry context, a relatively smaller reference (training) population size is needed to attain the same response to selection from larger traditional population (i.e., ABLUP). Conversely, if generation turnover is not possible, then larger training population size is required, therefore defeating GS goals. Bastiaansen et al. (Bastiaansen *et al.* 2012) found similar response to selection for GBLUP and ABLUP but the former accumulated lower level of inbreeding and consequently higher genetic variance than the latter.

## 2.5 Summary

**Background**: Genomic selection (GS) in forestry can substantially reduce the length of breeding cycle and increase gain per unit time through early selection and greater selection intensity, particularly for traits of low heritability and late expression. Affordable next-generation sequencing technologies made it possible to genotype large numbers of trees at a reasonable cost.

**Results**: Genotyping-by-sequencing was used to genotype 1,126 Interior spruce trees representing 25 open-pollinated families planted over three sites in British Columbia, Canada. Four imputation algorithms were compared (mean value (MI), singular value decomposition (SVD), expectation maximization (EM), and a newly derived, family-based k-nearest neighbor (kNN-Fam)). Trees were phenotyped for several yield and wood attributes. Single- and multi-site GS prediction models were developed using the Ridge Regression Best Linear Unbiased Predictor (RR-BLUP) and the Generalized Ridge Regression (GRR) to test different assumption about trait architecture. Finally, using PCA, multi-trait GS prediction models were developed. The EM and kNN-Fam imputation methods were superior for 30 and 60% missing data, respectively. The RR-BLUP GS prediction model produced better accuracies than the GRR indicating that the genetic architecture for these traits is complex. GS prediction accuracies for multi-site were high and better than those of single-sites while multi-site predictability produced the lowest accuracies reflecting type-b genetic correlations and deemed unreliable. The incorporation of genomic information in quantitative genetics analyses produced more realistic heritability estimates as half-sib pedigree tended to inflate the additive genetic variance and subsequently both heritability and gain estimates. Principle component scores as representatives of multi-trait GS prediction models

produced surprising results where negatively correlated traits could be concurrently selected for using PCA2 and PCA3.

**Conclusions**: The application of GS to open-pollinated family testing, the simplest form of tree improvement evaluation methods, was proven to be effective. Prediction accuracies obtained for all traits greatly support the integration of GS in tree breeding. While the within-site GS prediction accuracies were high, the results clearly indicate that single-site GS models ability to predict other sites are unreliable supporting the utilization of multi-site approach. Principle component scores provided an opportunity for the concurrent selection of traits with different phenotypic optima.

**Table 2.1 The comparison of imputation methods' accuracies**

| | MI | SVD_2 | SVD_3 | SVD_5 | EM_0.01 | EM_0.001 | kNN_10 | kNN_30 |
|---|---|---|---|---|---|---|---|---|
| **family 11** | 0.799 | 0.806 | 0.805 | 0.800 | 0.771 | 0.771 | 0.799 | 0.795 |
| **family 17** | 0.808 | 0.812 | 0.810 | 0.803 | 0.778 | 0.777 | 0.788 | 0.809 |
| **family 21** | 0.805 | 0.805 | 0.801 | 0.795 | 0.769 | 0.768 | 0.769 | 0.800 |
| **family 6** | 0.800 | 0.822 | 0.809 | 0.802 | 0.767 | 0.767 | 0.799 | 0.802 |
| **family 47** | 0.803 | 0.802 | 0.800 | 0.798 | 0.770 | 0.770 | 0.773 | 0.809 |
| **Average** | 0.803 | 0.810 | 0.805 | 0.799 | 0.771 | 0.771 | 0.785 | 0.803 |

**Table 2.2 Imputation methods used for genotyping-by-sequencing data**

| Imputation method[1] | Missing Data Threshold | Imputation algorithm | # of SNPs |
|---|---|---|---|
| **MI** | 30% | Mean imputation (MI) | 8,868 |
| **MI** | 60% | Mean imputation (MI) | 47,521 |
| **EM** | 30% | Expectation-maximization (EM) | 8,868 |
| **kNN-Fam** | 60% | Family-based K-nearest neighbor (kNN-Fam) | 62,198 |
| **SVD** | 60% | Singular Value Decomposition (SVD) | 55,618 |

[1] See main text for abbreviations

**Table 2.3 Imputation accuracy of kNN-Fam method with different K1 and K2 values**

| | K2: K value selected from non-family samples | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **K1** | **1** | **2** | **3** | **4** | **5** | **10** | **20** | **50** | **100** | **250** |
| **1** | 0.771 | 0.818 | 0.819 | 0.830 | 0.832 | 0.843 | 0.848 | 0.848 | 0.846 | 0.845 |
| **2** | 0.816 | 0.816 | 0.830 | 0.830 | 0.837 | 0.845 | 0.848 | 0.848 | 0.846 | 0.845 |
| **5** | 0.824 | 0.834 | 0.835 | 0.839 | 0.840 | 0.846 | 0.849 | 0.848 | 0.846 | 0.845 |
| **10** | 0.837 | 0.838 | 0.842 | 0.842 | 0.843 | 0.845 | 0.848 | 0.848 | 0.846 | 0.845 |
| **15** | 0.840 | 0.842 | 0.843 | 0.843 | 0.844 | 0.846 | 0.848 | 0.848 | 0.846 | 0.845 |
| **20** | 0.840 | 0.842 | 0.843 | 0.843 | 0.844 | 0.845 | 0.848 | 0.848 | 0.846 | 0.845 |
| **30** | 0.842 | 0.843 | 0.844 | 0.844 | 0.844 | 0.846 | 0.847 | 0.847 | 0.846 | 0.845 |

**Table 2.4 Multi- and single site heritability estimates and their standard errors using pedigree (ABLUP) and genomic (GBLUP) best linear unbiased predictors.**

| **Trait** | **ABLUP** | | | | **GBLUP (EM-30%)** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Multi-site** | **Single site** | | | **Multi-site** | **Single site** | | |
| | | **PGTIS** | **Aleza L.** | **Quesnel** | | **PGTIS** | **Aleza L.** | **Quesnel** |
| **HT** | 0.35±0.14 | 0.64±0.22 | 0.43±0.19 | 0.98±0.02 | 0.20±0.06 | 0.50±0.15 | 0.32±0.14 | 0.56±0.13 |
| **DBH** | 0.05±0.08 | 0.39±0.17 | 0.28±0.15 | 0.55±0.19 | 0.07±0.06 | 0.37±0.15 | 0.26±0.13 | 0.53±0.15 |
| **VOL** | 0.09±0.10 | 0.45±0.18 | 0.29±0.15 | 0.76±0.23 | 0.09±0.06 | 0.42±0.15 | 0.27±0.13 | 0.60±0.15 |
| **$V_{Dir}$** | 0.28±0.12 | 0.31±0.15 | 0.38±0.17 | 0.78±0.24 | 0.12±0.06 | 0.17±0.11 | 0.37±0.15 | 0.49±0.14 |
| **$WD_{res}$** | 0.27±0.12 | 0.59±0.21 | 0.65±0.22 | 0.42±0.15 | 0.10±0.06 | 0.49±0.15 | 0.28±0.13 | 0.42±0.14 |
| **$WD_{X\text{-}ray}$** | 0.38±0.14 | 0.55±0.20 | 0.48±0.19 | 0.59±0.20 | 0.18±0.06 | 0.28±0.13 | 0.39±0.15 | 0.43±0.13 |
| **$MoE_d$** | 0.28±0.12 | 0.31±0.15 | 0.38±0.17 | 0.78±0.24 | 0.12±0.06 | 0.17±0.11 | 0.37±0.15 | 0.49±0.14 |

Note: Traits are HT: height in m; DBH: diameter at breast height in cm; VOL: stem volume in m3; VDir: acoustic velocity in km/s; WDRes: resistance to drilling; WDX-ray: wood density in kg/m3 using X-ray densitometry; MoEd: dynamic modulus of elasticity.

**Table 2.5 Within site (PGTIS, Aleza Lake (AL), and Quesnel) genomic selection prediction accuracies and their standard errors for RR-BLUP and GRR models across 30% missing data imputation methods (MI-30% and EM-30%).**

| Trait | GS Model | Imputation method | | | | | |
|---|---|---|---|---|---|---|---|
| | | MI-30% | | | EM-30% | | |
| | | PGTIS | AL | Quesnel | PGTIS | AL | Quesnel |
| **HT** | **RR-BLUP** | 0.48±0.003[1] | 0.46±0.002 | 0.33±0.003 | 0.50±0.003 | 0.48±0.003 | 0.35±0.004 |
| | **GRR** | 0.44±0.003 | 0.45±0.010 | 0.27±0.007 | 0.46±0.005 | 0.45±0.005 | 0.29±0.006 |
| **DBH** | **RR-BLUP** | 0.58±0.002 | 0.55±0.003 | 0.53±0.004 | 0.58±0.003 | 0.55±0.002 | 0.53±0.003 |
| | **GRR** | 0.54±0.003 | 0.47±0.017 | 0.51±0.006 | 0.53±0.004 | 0.49±0.006 | 0.51±0.003 |
| **VOL** | **RR-BLUP** | 0.56±0.002 | 0.54±0.003 | 0.44±0.003 | 0.55±0.004 | 0.54±0.002 | 0.45±0.002 |
| | **GRR** | 0.52±0.003 | 0.50±0.004 | 0.42±0.006 | 0.53±0.004 | 0.49±0.004 | 0.41±0.006 |
| $\mathbf{V_{Dir}}$ | **RR-BLUP** | 0.55±0.002 | 0.54±0.002 | 0.41±0.004 | 0.55±0.003 | 0.55±0.002 | 0.41±0.004 |
| | **GRR** | 0.52±0.003 | 0.48±0.004 | 0.31±0.006 | 0.52±0.013 | 0.50±0.005 | 0.33±0.004 |
| $\mathbf{WD_{Res}}$ | **RR-BLUP** | 0.47±0.003 | 0.37±0.003 | 0.59±0.003 | 0.49±0.003 | 0.39±0.004 | 0.59±0.003 |
| | **GRR** | 0.46±0.005 | 0.34±0.005 | 0.54±0.005 | 0.44±0.009 | 0.33±0.007 | 0.54±0.005 |
| $\mathbf{WD_{X\text{-}ray}}$ | **RR-BLUP** | 0.41±0.003 | 0.49±0.003 | 0.50±0.002 | 0.43±0.003 | 0.48±0.003 | 0.50±0.001 |
| | **GRR** | 0.41±0.004 | 0.25±0.011 | 0.50±0.004 | 0.42±0.003 | 0.46±0.020 | 0.50±0.002 |
| $\mathbf{MoE_d}$ | **RR-BLUP** | 0.55±0.003 | 0.55±0.002 | 0.40±0.004 | 0.55±0.002 | 0.55±0.002 | 0.39±0.003 |
| | **GRR** | 0.53±0.004 | 0.51±0.003 | 0.30±0.006 | 0.55±0.004 | 0.52±0.005 | 0.29±0.006 |
| **Ave.** | **RR-BLUP** | 0.51±0.062 | 0.50±0.067 | 0.46±0.088 | 0.52±0.051 | 0.51±0.060 | 0.46±0.085 |
| | **GRR** | 0.49±0.051 | 0.43±0.097 | 0.41±0.113 | 0.49±0.052 | 0.46±0.063 | 0.41±0.108 |

Ave.: average across all traits.

**Table 2.6 Within site (PGTIS, Aleza Lake (AL), and Quesnel) genomic selection prediction accuracies and their standard errors for RR-BLUP and GRR models across 60% missing data imputation methods (MI-60%, kNN-Fam-60% and SVD-60%).**

| Trait | GS Model | Imputation method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | MI-60% | | | kNN-Fam-60% | | | SVD-60% | | |
| | | PGTIS | AL | Quesnel | PGTIS | AL | Quesnel | PGTIS | AL | Quesnel |
| **HT** | **RR-BLUP** | 0.54±0.002 | 0.51±0.003 | 0.40±0.002 | 0.55±0.002 | 0.56±0.002 | 0.42±0.002 | 0.53±0.003 | 0.50±0.004 | 0.42±0.003 |
| | **GRR** | 0.51±0.005 | 0.45±0.011 | 0.34±0.007 | 0.51±0.005 | 0.51±0.006 | 0.39±0.005 | 0.51±0.004 | 0.47±0.006 | 0.37±0.005 |
| **DBH** | **RR-BLUP** | 0.62±0.002 | 0.60±0.002 | 0.56±0.003 | 0.62±0.001 | 0.63±0.002 | 0.55±0.002 | 0.60±0.002 | 0.59±0.003 | 0.54±0.003 |
| | **GRR** | 0.59±0.009 | 0.58±0.004 | 0.53±0.006 | 0.59±0.005 | 0.62±0.004 | 0.53±0.004 | 0.59±0.002 | 0.57±0.004 | 0.52±0.004 |
| **VOL** | **RR-BLUP** | 0.60±0.002 | 0.58±0.003 | 0.49±0.003 | 0.61±0.002 | 0.63±0.001 | 0.47±0.002 | 0.59±0.002 | 0.57±0.003 | 0.48±0.003 |
| | **GRR** | 0.58±0.005 | 0.55±0.006 | 0.44±0.009 | 0.58±0.003 | 0.59±0.005 | 0.44±0.005 | 0.58±0.003 | 0.56±0.004 | 0.45±0.005 |
| **$V_{Dir}$** | **RR-BLUP** | 0.62±0.002 | 0.57±0.002 | 0.46±0.003 | 0.63±0.002 | 0.61±0.002 | 0.49±0.002 | 0.58±0.002 | 0.55±0.002 | 0.46±0.003 |
| | **GRR** | 0.59±0.005 | 0.51±0.010 | 0.40±0.006 | 0.60±0.003 | 0.57±0.005 | 0.46±0.006 | 0.57±0.004 | 0.53±0.003 | 0.42±0.004 |
| **$WD_{Res}$** | **RR-BLUP** | 0.53±0.002 | 0.44±0.002 | 0.62±0.002 | 0.55±0.002 | 0.49±0.002 | 0.62±0.002 | 0.56±0.003 | 0.46±0.004 | 0.58±0.002 |
| | **GRR** | 0.46±0.007 | 0.36±0.009 | 0.58±0.004 | 0.47±0.005 | 0.44±0.007 | 0.59±0.005 | 0.54±0.003 | 0.43±0.005 | 0.56±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.49±0.002 | 0.51±0.002 | 0.53±0.003 | 0.51±0.002 | 0.53±0.002 | 0.53±0.002 | 0.50±0.002 | 0.50±0.002 | 0.50±0.003 |
| | **GRR** | 0.45±0.006 | 0.47±0.005 | 0.49±0.009 | 0.48±0.005 | 0.50±0.006 | 0.48±0.009 | 0.49±0.005 | 0.49±0.003 | 0.49±0.004 |
| **$MoE_d$** | **RR-BLUP** | 0.62±0.001 | 0.57±0.002 | 0.45±0.002 | 0.64±0.001 | 0.61±0.001 | 0.49±0.002 | 0.59±0.003 | 0.54±0.004 | 0.45±0.004 |
| | **GRR** | 0.60±0.003 | 0.52±0.007 | 0.38±0.007 | 0.61±0.004 | 0.58±0.004 | 0.45±0.004 | 0.58±0.002 | 0.52±0.004 | 0.41±0.005 |
| **Ave.** | **RR-BLUP** | 0.57±0.053 | 0.54±0.056 | 0.50±0.074 | 0.59±0.050 | 0.58±0.054 | 0.51±0.064 | 0.56±0.037 | 0.53±0.045 | 0.49±0.055 |
| | **GRR** | 0.54±0.065 | 0.49±0.073 | 0.45±0.086 | 0.55±0.060 | 0.54±0.063 | 0.48±0.065 | 0.55±0.039 | 0.51±0.050 | 0.46±0.067 |

**Table 2.7 Within multi-site genomic selection prediction accuracies and their standard errors for RR-BLUP and GRR models for the studied five imputation methods.**

| Trait | GS Model | Imputation method | | | | |
|---|---|---|---|---|---|---|
| | | MI-30% | EM-30% | MI-60% | kNN-Fam-60% | SVD-60% |
| **HT** | **RR-BLUP** | 0.56±0.001 | 0.58±0.001 | 0.60±0.001 | 0.63±0.001 | 0.61±0.001 |
| | **GRR** | 0.50±0.002 | 0.48±0.004 | 0.57±0.003 | 0.62±0.002 | 0.58±0.002 |
| **DBH** | **RR-BLUP** | 0.71±0.001 | 0.72±0.001 | 0.75±0.001 | 0.77±0.001 | 0.76±0.001 |
| | **GRR** | 0.71±0.001 | 0.73±0.001 | 0.74±0.001 | 0.77±0.001 | 0.75±0.001 |
| **VOL** | **RR-BLUP** | 0.67±0.001 | 0.68±0.001 | 0.71±0.001 | 0.73±0.001 | 0.72±0.001 |
| | **GRR** | 0.67±0.001 | 0.68±0.001 | 0.70±0.001 | 0.72±0.001 | 0.71±0.001 |
| **$V_{Dir}$** | **RR-BLUP** | 0.59±0.001 | 0.61±0.001 | 0.63±0.001 | 0.67±0.001 | 0.65±0.001 |
| | **GRR** | 0.52±0.004 | 0.50±0.003 | 0.62±0.002 | 0.66±0.001 | 0.62±0.006 |
| **$WD_{Res}$** | **RR-BLUP** | 0.56±0.001 | 0.58±0.001 | 0.62±0.001 | 0.64±0.001 | 0.63±0.001 |
| | **GRR** | 0.48±0.002 | 0.47±0.003 | 0.59±0.003 | 0.64±0.002 | 0.60±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.55±0.001 | 0.56±0.001 | 0.59±0.001 | 0.62±0.001 | 0.61±0.001 |
| | **GRR** | 0.54±0.002 | 0.55±0.001 | 0.59±0.002 | 0.62±0.001 | 0.60±0.002 |
| **$MoE_d$** | **RR-BLUP** | 0.50±0.001 | 0.61±0.001 | 0.63±0.001 | 0.67±0.001 | 0.65±0.001 |
| | **GRR** | 0.50±0.013 | 0.56±0.002 | 0.63±0.002 | 0.66±0.001 | 0.64±0.002 |
| **Ave.** | **RR-BLUP** | 0.59±0.073 | 0.62±0.059 | 0.65±0.060 | 0.68±0.055 | 0.66±0.057 |
| | **GRR** | 0.56±0.091 | 0.57±0.101 | 0.63±0.063 | 0.67±0.056 | 0.64±0.063 |

**Table 2.8 Single site GS prediction accuracies and their standard errors resulting from using the multi-sites as training population for RR-BLUP and GRR models for kNN-Fam-60% imputation method.**

| Traits | GS Model | Cross-validation | | | |
|---|---|---|---|---|---|
| | | Multi-sites | PGTIS | Aleza Lake | Quesnel |
| **HT** | **RR-BLUP** | 0.63±0.001 | 0.37±0.001 | 0.53±0.002 | 0.45±0.001 |
| | **GRR** | 0.62±0.002 | 0.36±0.003 | 0.52±0.003 | 0.45±0.002 |
| **DBH** | **RR-BLUP** | 0.77±0.001 | 0.37±0.001 | 0.50±0.001 | 0.40±0.001 |
| | **GRR** | 0.77±0.001 | 0.37±0.002 | 0.50±0.001 | 0.40±0.001 |
| **VOL** | **RR-BLUP** | 0.73±0.001 | 0.34±0.001 | 0.50±0.001 | 0.41±0.001 |
| | **GRR** | 0.72±0.001 | 0.34±0.002 | 0.50±0.002 | 0.40±0.002 |
| **$V_{Dir}$** | **RR-BLUP** | 0.67±0.001 | 0.50±0.001 | 0.47±0.001 | 0.49±0.001 |
| | **GRR** | 0.66±0.001 | 0.49±0.001 | 0.47±0.001 | 0.48±0.002 |
| **$WD_{Res}$** | **RR-BLUP** | 0.64±0.001 | 0.41±0.001 | 0.48±0.001 | 0.46±0.001 |
| | **GRR** | 0.64±0.002 | 0.41±0.002 | 0.48±0.002 | 0.45±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.62±0.001 | 0.46±0.001 | 0.49±0.002 | 0.50±0.001 |
| | **GRR** | 0.62±0.001 | 0.46±0.002 | 0.49±0.002 | 0.50±0.002 |
| **$MoE_d$** | **RR-BLUP** | 0.67±0.001 | 0.50±0.001 | 0.46±0.001 | 0.48±0.001 |
| | **GRR** | 0.66±0.001 | 0.49±0.002 | 0.45±0.002 | 0.47±0.002 |
| **Ave.** | **RR-BLUP** | 0.68±0.055 | 0.42±0.066 | 0.49±0.023 | 0.46±0.039 |
| | **GRR** | 0.67±0.056 | 0.42±0.063 | 0.49±0.023 | 0.45±0.038 |

**Table 2.9 Cross-site GS prediction accuracies for all studied combinations (GS prediction model, imputation method, and trait).**

| **Traits** | **Imputation** | **MI-60%** | | | | | |
|---|---|---|---|---|---|---|---|
| | **Sites** | **1 & 2** | | **1 & 3** | | **2 & 3** | |
| | **GS model** | **1 --> 2** | **2 --> 1** | **1 --> 3** | **3 --> 1** | **2 --> 3** | **3 --> 2** |
| **HT** | **RR-BLUP** | 0.22±0.002 | 0.21±0.002 | 0.17±0.002 | 0.18±0.002 | 0.32±0.002 | 0.34±0.002 |
| | **GRR** | 0.19±0.003 | 0.20±0.004 | 0.15±0.005 | 0.17±0.006 | 0.28±0.007 | 0.29±0.006 |
| **DBH** | **RR-BLUP** | 0.23±0.002 | 0.19±0.002 | -0.01±0.002 | -0.01±0.003 | 0.06±0.002 | 0.03±0.002 |
| | **GRR** | 0.21±0.005 | 0.18±0.003 | 0.00±0.004 | 0.01±0.004 | 0.05±0.005 | 0.03±0.005 |
| **VOL** | **RR-BLUP** | 0.18±0.001 | 0.15±0.001 | 0.00±0.002 | -0.01±0.001 | 0.15±0.002 | 0.13±0.002 |
| | **GRR** | 0.17±0.005 | 0.14±0.004 | 0.00±0.004 | 0.00±0.004 | 0.14±0.005 | 0.12±0.004 |
| **$V_{Dir}$** | **RR-BLUP** | 0.17±0.001 | 0.14±0.001 | 0.30±0.002 | 0.36±0.002 | 0.22±0.002 | 0.22±0.002 |
| | **GRR** | 0.17±0.005 | 0.12±0.004 | 0.28±0.004 | 0.34±0.004 | 0.21±0.006 | 0.19±0.004 |
| **$WD_{Res}$** | **RR-BLUP** | 0.13±0.001 | 0.20±0.002 | 0.13±0.002 | 0.13±0.001 | 0.27±0.002 | 0.21±0.001 |
| | **GRR** | 0.12±0.005 | 0.17±0.008 | 0.12±0.006 | 0.12±0.003 | 0.23±0.006 | 0.21±0.004 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.20±0.002 | 0.23±0.001 | 0.28±0.002 | 0.31±0.002 | 0.28±0.002 | 0.27±0.002 |
| | **GRR** | 0.19±0.005 | 0.21±0.005 | 0.27±0.004 | 0.29±0.005 | 0.26±0.007 | 0.27±0.004 |
| **$MoE_d$** | **RR-BLUP** | 0.16±0.001 | 0.15±0.002 | 0.28±0.001 | 0.35±0.002 | 0.20±0.002 | 0.20±0.002 |
| | **GRR** | 0.16±0.003 | 0.14±0.004 | 0.27±0.004 | 0.33±0.006 | 0.19±0.005 | 0.18±0.005 |
| **Traits** | **Imputation** | **KNN-60%** | | | | | |
| | **Sites** | **1 & 2** | | **1 & 3** | | **2 & 3** | |
| | **GS model** | **1 --> 2** | **2 --> 1** | **1 --> 3** | **3 --> 1** | **2 --> 3** | **3 --> 2** |
| **HT** | **RR-BLUP** | 0.21±0.002 | 0.19±0.002 | 0.17±0.002 | 0.19±0.002 | 0.36±0.002 | 0.39±0.002 |
| | **GRR** | 0.18±0.005 | 0.17±0.005 | 0.16±0.005 | 0.18±0.005 | 0.33±0.007 | 0.34±0.006 |
| **DBH** | **RR-BLUP** | 0.21±0.002 | 0.17±0.001 | 0.01±0.001 | -0.01±0.002 | 0.06±0.002 | 0.03±0.002 |
| | **GRR** | 0.21±0.005 | 0.16±0.004 | 0.01±0.004 | 0.00±0.004 | 0.05±0.003 | 0.03±0.004 |
| **VOL** | **RR-BLUP** | 0.17±0.001 | 0.13±0.001 | 0.01±0.001 | 0.00±0.002 | 0.16±0.002 | 0.14±0.002 |
| | **GRR** | 0.15±0.004 | 0.14±0.005 | 0.01±0.004 | 0.00±0.003 | 0.15±0.005 | 0.13±0.005 |
| **$V_{Dir}$** | **RR-BLUP** | 0.17±0.001 | 0.14±0.001 | 0.32±0.002 | 0.38±0.002 | 0.22±0.001 | 0.22±0.001 |
| | **GRR** | 0.17±0.005 | 0.14±0.005 | 0.30±0.003 | 0.37±0.003 | 0.22±0.004 | 0.21±0.003 |
| **$WD_{Res}$** | **RR-BLUP** | 0.16±0.002 | 0.23±0.002 | 0.14±0.002 | 0.14±0.001 | 0.29±0.002 | 0.24±0.001 |
| | **GRR** | 0.14±0.004 | 0.22±0.005 | 0.12±0.005 | 0.13±0.002 | 0.26±0.004 | 0.24±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.23±0.002 | 0.24±0.002 | 0.32±0.002 | 0.34±0.002 | 0.30±0.001 | 0.31±0.002 |
| | **GRR** | 0.22±0.004 | 0.22±0.004 | 0.30±0.005 | 0.32±0.004 | 0.29±0.003 | 0.29±0.005 |
| **$MoE_d$** | **RR-BLUP** | 0.16±0.001 | 0.14±0.001 | 0.31±0.001 | 0.38±0.002 | 0.20±0.001 | 0.21±0.002 |
| | **GRR** | 0.16±0.003 | 0.15±0.003 | 0.29±0.004 | 0.36±0.005 | 0.20±0.003 | 0.22±0.004 |

**Table 2.9 Cross-site GS prediction accuracies for all studied combinations (GS prediction model, imputation method, and trait).**

| **Traits** | **Imputation** | **SVD-60%** | | | | | |
|---|---|---|---|---|---|---|---|
| | **Sites** | **1 & 2** | | **1 & 3** | | **2 & 3** | |
| | **GS model** | **1 --> 2** | **2 --> 1** | **1 --> 3** | **3 --> 1** | **2 --> 3** | **3 --> 2** |
| **HT** | **RR-BLUP** | 0.25±0.001 | 0.22±0.001 | 0.20±0.003 | 0.25±0.002 | 0.31±0.003 | 0.32±0.002 |
| | **GRR** | 0.23±0.003 | 0.21±0.003 | 0.19±0.004 | 0.25±0.003 | 0.30±0.004 | 0.30±0.006 |
| **DBH** | **RR-BLUP** | 0.20±0.002 | 0.18±0.002 | 0.03±0.002 | 0.04±0.002 | 0.06±0.002 | 0.01±0.002 |
| | **GRR** | 0.19±0.003 | 0.17±0.003 | 0.03±0.003 | 0.04±0.004 | 0.06±0.003 | 0.02±0.004 |
| **VOL** | **RR-BLUP** | 0.18±0.002 | 0.14±0.002 | 0.05±0.001 | 0.06±0.002 | 0.13±0.002 | 0.11±0.001 |
| | **GRR** | 0.17±0.004 | 0.14±0.003 | 0.04±0.002 | 0.06±0.003 | 0.13±0.003 | 0.11±0.003 |
| **$V_{Dir}$** | **RR-BLUP** | 0.19±0.002 | 0.17±0.002 | 0.30±0.002 | 0.37±0.002 | 0.20±0.002 | 0.24±0.002 |
| | **GRR** | 0.19±0.003 | 0.17±0.004 | 0.30±0.003 | 0.35±0.003 | 0.18±0.003 | 0.24±0.004 |
| **$WD_{Res}$** | **RR-BLUP** | 0.14±0.003 | 0.21±0.003 | 0.11±0.002 | 0.16±0.002 | 0.28±0.002 | 0.19±0.002 |
| | **GRR** | 0.13±0.003 | 0.21±0.005 | 0.11±0.004 | 0.16±0.004 | 0.25±0.004 | 0.19±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.22±0.002 | 0.25±0.002 | 0.26±0.002 | 0.30±0.002 | 0.29±0.002 | 0.27±0.002 |
| | **GRR** | 0.21±0.003 | 0.24±0.004 | 0.26±0.003 | 0.30±0.003 | 0.29±0.004 | 0.26±0.003 |
| **$MoE_d$** | **RR-BLUP** | 0.18±0.002 | 0.16±0.002 | 0.29±0.002 | 0.36±0.001 | 0.18±0.002 | 0.23±0.001 |
| | **GRR** | 0.17±0.003 | 0.16±0.004 | 0.28±0.003 | 0.34±0.004 | 0.17±0.003 | 0.22±0.004 |

| **Traits** | **Imputation** | **MI-30%** | | | | | |
|---|---|---|---|---|---|---|---|
| | **Sites** | **1 & 2** | | **1 & 3** | | **2 & 3** | |
| | **GS model** | **1 --> 2** | **2 --> 1** | **1 --> 3** | **3 --> 1** | **2 --> 3** | **3 --> 2** |
| **HT** | **RR-BLUP** | 0.20±0.001 | 0.18±0.002 | 0.14±0.002 | 0.15±0.002 | 0.30±0.002 | 0.34±0.003 |
| | **GRR** | 0.17±0.004 | 0.13±0.008 | 0.13±0.005 | 0.14±0.007 | 0.27±0.006 | 0.28±0.008 |
| **DBH** | **RR-BLUP** | 0.19±0.002 | 0.20±0.002 | 0.03±0.003 | -0.02±0.003 | 0.04±0.002 | 0.04±0.002 |
| | **GRR** | 0.19±0.003 | 0.17±0.007 | 0.03±0.004 | 0.01±0.006 | 0.03±0.005 | 0.03±0.004 |
| **VOL** | **RR-BLUP** | 0.16±0.002 | 0.15±0.002 | 0.02±0.002 | -0.03±0.003 | 0.12±0.003 | 0.14±0.003 |
| | **GRR** | 0.16±0.003 | 0.13±0.004 | 0.02±0.004 | 0.02±0.004 | 0.11±0.004 | 0.10±0.005 |
| **$V_{Dir}$** | **RR-BLUP** | 0.13±0.002 | 0.14±0.002 | 0.26±0.002 | 0.31±0.002 | 0.18±0.002 | 0.19±0.002 |
| | **GRR** | 0.11±0.004 | 0.14±0.003 | 0.24±0.003 | 0.28±0.004 | 0.17±0.004 | 0.15±0.005 |
| **$WD_{Res}$** | **RR-BLUP** | 0.08±0.002 | 0.12±0.003 | 0.11±0.002 | 0.12±0.002 | 0.25±0.002 | 0.19±0.002 |
| | **GRR** | 0.09±0.004 | 0.10±0.004 | 0.09±0.004 | 0.11±0.004 | 0.20±0.005 | 0.18±0.005 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.17±0.003 | 0.19±0.003 | 0.24±0.002 | 0.26±0.002 | 0.25±0.002 | 0.23±0.003 |
| | **GRR** | 0.17±0.004 | 0.13±0.010 | 0.24±0.003 | 0.26±0.003 | 0.19±0.008 | 0.23±0.004 |
| **$MoE_d$** | **RR-BLUP** | 0.10±0.002 | 0.13±0.002 | 0.25±0.002 | 0.30±0.002 | 0.17±0.002 | 0.17±0.003 |
| | **GRR** | 0.09±0.003 | 0.13±0.003 | 0.23±0.003 | 0.25±0.004 | 0.17±0.004 | 0.11±0.005 |

**Table 2.9 Cross-site GS prediction accuracies for all studied combinations (GS prediction model, imputation method, and trait).**

| | **Imputation** | **EM-30%** | | | | | |
|---|---|---|---|---|---|---|---|
| **Traits** | **Sites** | **1 & 2** | | **1 & 3** | | **2 & 3** | |
| | **GS model** | **1 --> 2** | **2 --> 1** | **1 --> 3** | **3 --> 1** | **2 --> 3** | **3 --> 2** |
| **HT** | **RR-BLUP** | 0.21±0.002 | 0.19±0.002 | 0.14±0.003 | 0.16±0.002 | 0.32±0.002 | 0.37±0.002 |
| | **GRR** | 0.20±0.004 | 0.16±0.005 | 0.14±0.004 | 0.13±0.004 | 0.30±0.005 | 0.31±0.008 |
| **DBH** | **RR-BLUP** | 0.18±0.002 | 0.19±0.002 | 0.03±0.002 | -0.02±0.002 | 0.04±0.002 | 0.06±0.003 |
| | **GRR** | 0.18±0.003 | 0.18±0.003 | 0.03±0.003 | 0.00±0.004 | 0.03±0.005 | 0.05±0.003 |
| **VOL** | **RR-BLUP** | 0.16±0.002 | 0.15±0.002 | 0.02±0.002 | -0.02±0.003 | 0.13±0.001 | 0.16±0.002 |
| | **GRR** | 0.16±0.003 | 0.13±0.003 | 0.02±0.004 | 0.01±0.005 | 0.13±0.003 | 0.15±0.004 |
| **$V_{Dir}$** | **RR-BLUP** | 0.13±0.002 | 0.15±0.002 | 0.28±0.002 | 0.34±0.002 | 0.17±0.002 | 0.19±0.002 |
| | **GRR** | 0.12±0.004 | 0.14±0.005 | 0.26±0.010 | 0.29±0.004 | 0.16±0.003 | 0.14±0.003 |
| **$WD_{Res}$** | **RR-BLUP** | 0.10±0.002 | 0.13±0.002 | 0.12±0.002 | 0.13±0.003 | 0.26±0.002 | 0.20±0.002 |
| | **GRR** | 0.10±0.003 | 0.12±0.005 | 0.11±0.004 | 0.12±0.004 | 0.21±0.005 | 0.19±0.003 |
| **$WD_{X\text{-}ray}$** | **RR-BLUP** | 0.17±0.003 | 0.19±0.003 | 0.25±0.002 | 0.27±0.002 | 0.25±0.002 | 0.22±0.002 |
| | **GRR** | 0.17±0.003 | 0.18±0.004 | 0.25±0.004 | 0.27±0.003 | 0.24±0.008 | 0.22±0.003 |
| **$MoE_d$** | **RR-BLUP** | 0.11±0.002 | 0.14±0.002 | 0.26±0.002 | 0.32±0.002 | 0.16±0.002 | 0.17±0.003 |
| | **GRR** | 0.09±0.003 | 0.13±0.004 | 0.25±0.004 | 0.26±0.004 | 0.15±0.004 | 0.12±0.007 |

For sites: 1→ refers to PGTIS, 2→ refers to Aleza Lake, and 3→ refers to Quesnel

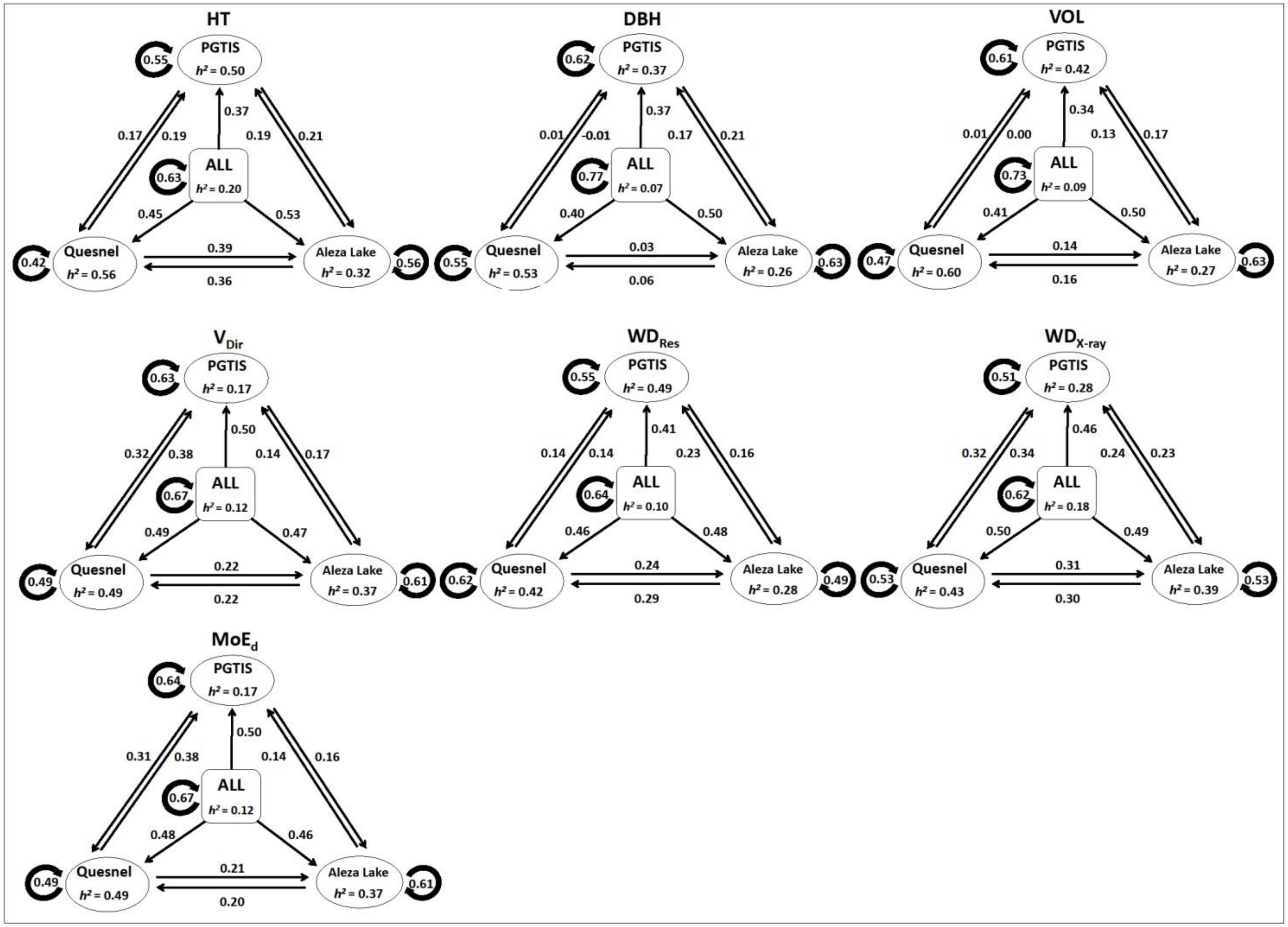


**Figure 2.1 Genomic selection prediction accuracies for the seven traits using the RR-BLUP model and KNN-60% imputation**

Note: (within single site (three values), cross-sites (six values), within multi-site (one value), and for multi-site to single site (three values)), with narrow-sense heritabilities ($h^2$) from single- and multi-site GBLUP analyses. Sites are Prince George Tree Improvement Station (PGTIS), Quesnel, Aleza lake, and multi-site (ALL).

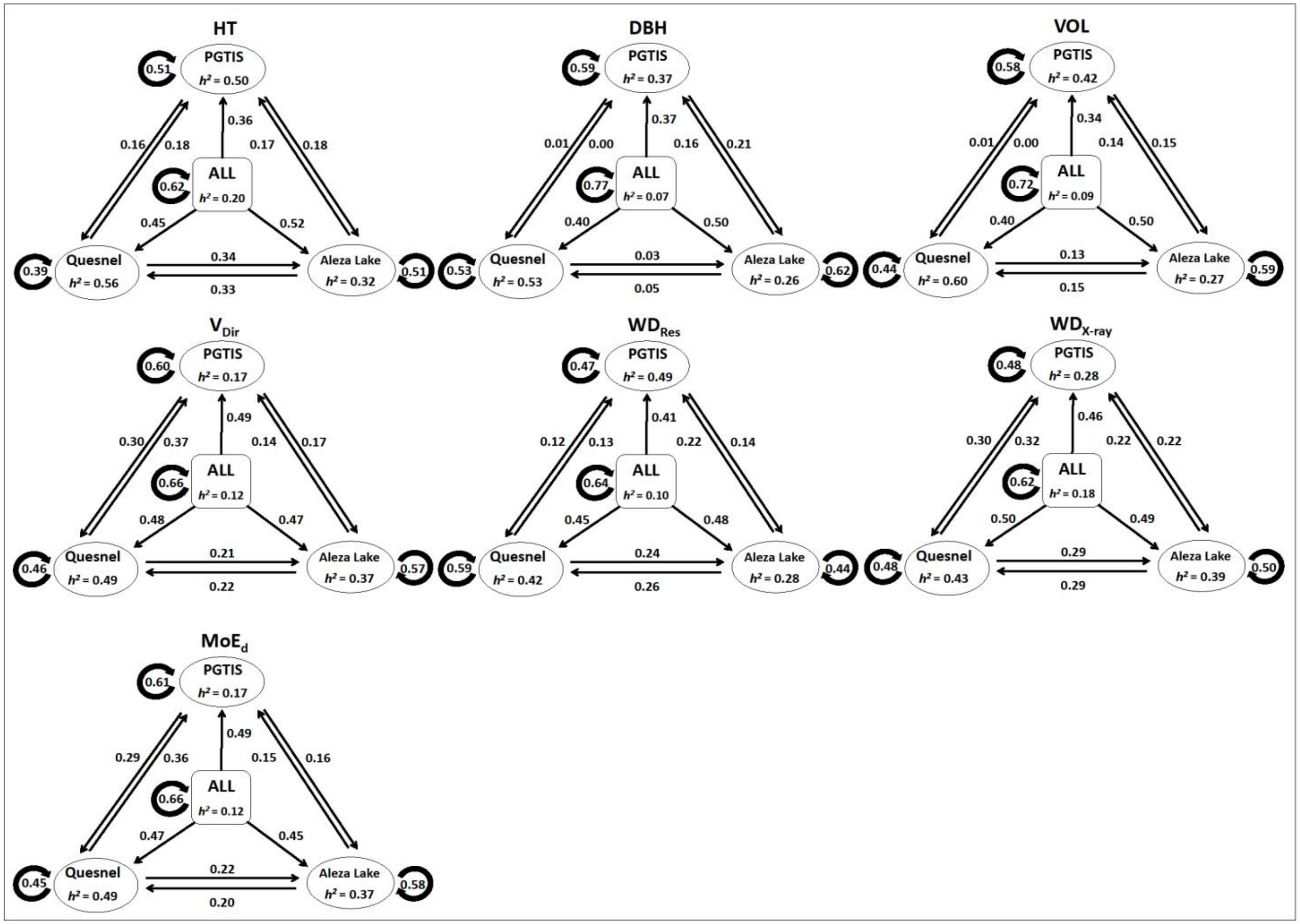
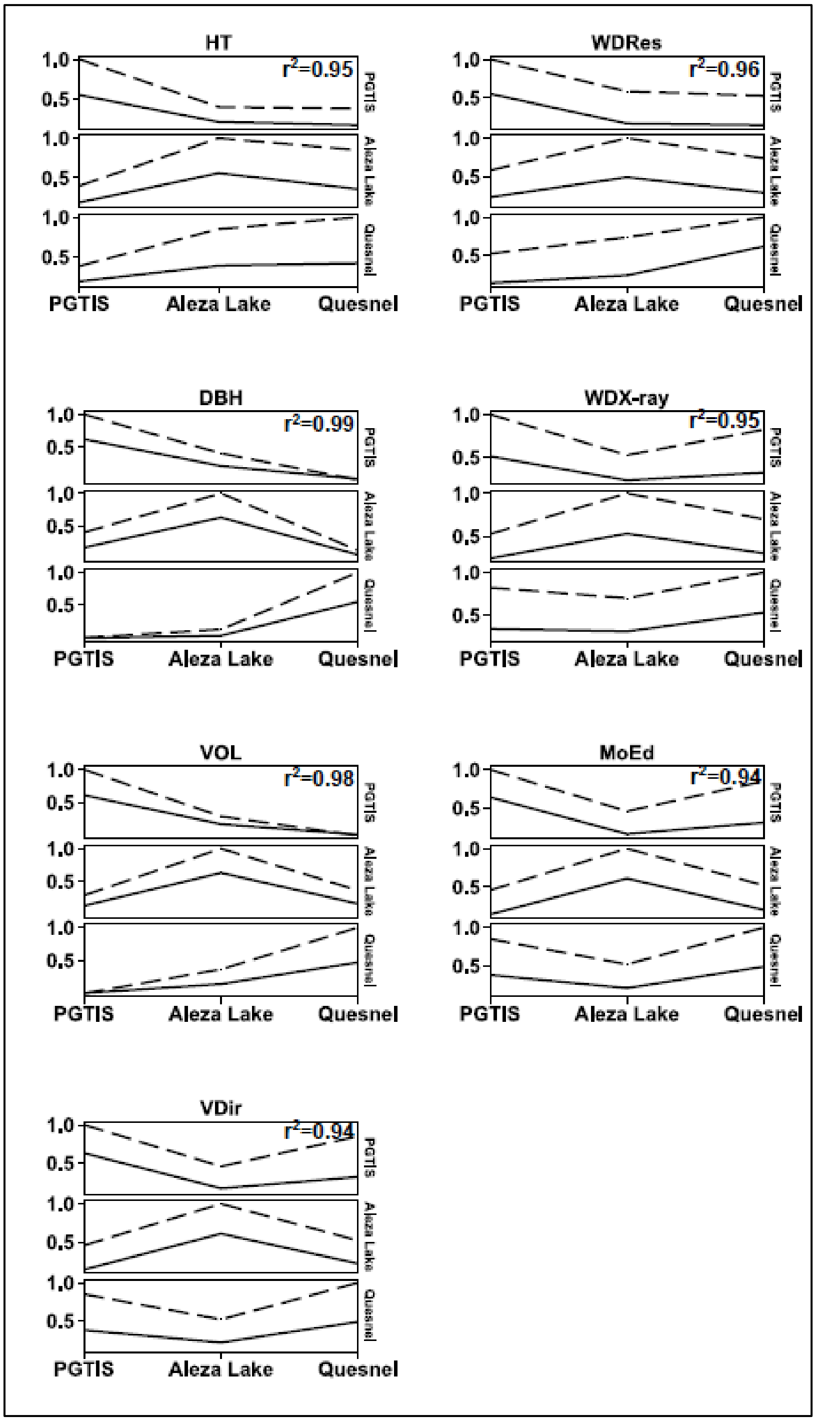


**Figure 2.2 Genomic selection prediction accuracies for the seven traits using GRR model and KNN-60% imputation**

Traits are HT: height in m; DBH: diameter at breast height in cm; VOL: stem volume in m$^3$; $V_{Dir}$: acoustic velocity in km/s; $WD_{Res}$: resistance to drilling; $WD_{X\text{-}ray}$: wood density in kg/m$^3$ using X-ray densitometry; $MoE_d$: dynamic modulus of elasticity. For within single site and within multi-site, single- and multi-site GBLUP heritabilities are presented.



Note: Sites are Prince George Tree Improvement Station (PGTIS), Quesnel, and Aleza lake. Traits are HT: height in m; DBH: diameter at breast height in cm; VOL: stem volume in $m^3$; $V_{Dir}$: acoustic velocity in km/s; $WD_{Res}$: resistance to drilling; $WD_{X\text{-}ray}$: wood density in kg/$m^3$ using X-ray densitometry; $MoE_d$: dynamic modulus of elasticity.

Cross-site GS accuracy ( —— )

Type-b genetic correlations between sites ( _ _ _ )

**Figure 2.3 Cross-site GS accuracy, type-b genetic correlations between sites (Y-axis) and their Pearson-product-moment correlations across sites (X-axis) for the seven traits.**

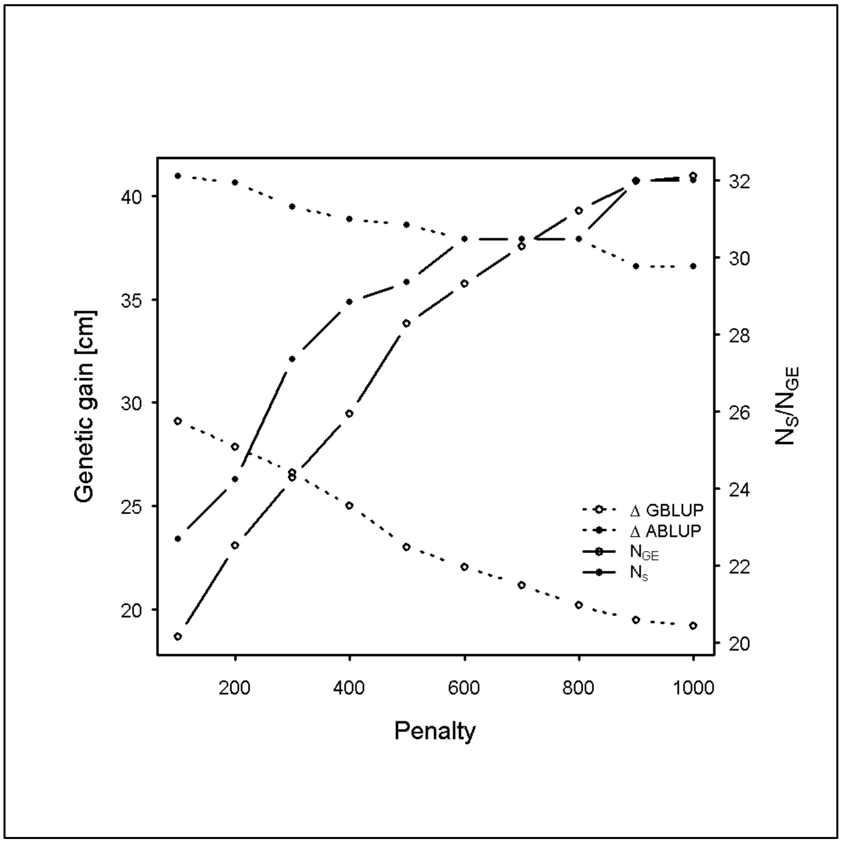


**Figure 2.4 The relationship between height genetic gain and genetic diversity for ABLUP (status number (Ns)) and GBLUP (number of founder genome equivalent (NGE)) across a range of co-ancestry penalties.**

# Chapter 3: Genomic-based vs. pedigree-based approach to genetic variance decomposition in single-site OP white spruce population

## 3.1 Introduction

Open-pollinated (OP) (also known as wind-pollinated) family testing is, by far, the simplest and most economical means for screening, evaluating, and ranking large number of candidate parent trees. Thus, OP testing combines the simplest known field experimental design in pedigree testing as candidate trees enter the test as maternal parents and their offspring are assumed to represent independent half-sib families. OP testing has been widely implemented for several tree species throughout the world (e.g., radiata pine (*Pinus radiata* D. Don) (Burdon and Shelbourne 1971), Interior spruce (*Picea glauca* (Moench) Voss x *P. engelmannii* Parry ex Engelm.) (Kiss and Yanchuk 1991), Douglas-fir (*Pseudotsuga menziesii* (Mirb.) Franco) (El-Kassaby and Sziklai 1982; Johnson 1997), western larch (*Larix occidentalis* Nutt.) (Ratcliffe *et al.* 2013), and Scots pine (*P. sylvestris* L.) (Korecký *et al.* 2013)), and it is often considered as a prelude to full-pedigree testing (Jayawickrama and Carson 2000).

Genealogically speaking, OP testing (i.e., partial pedigree) is positioned between the "no pedigree" provenance testing (Callaham 1964) and the "full-pedigree" mating design-based progeny testing that includes all higher levels of relatedness and connectivity among the created families (Namkoong *et al.* 2012). Thus, the accuracy of all OP testing-based estimated genetic parameters (e.g., additive genetic variance, heritability, breeding values, etc.) is superior to the former yet somewhat limited comparing to the latter (but also see, Hallingbäck and Jansson 2013). In fact, doubts are often raised regarding the accuracy of OP family testing-derived genetic parameters as the assumption of “half-sibling” is hardly fulfilled (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994).

The pedigree-based genetic relationships among individuals (based on the so-called ***A***-matrix: average numerator relationship matrix (Wright 1922)) are often used to estimate the genetic variance components by using the Restricted Maximum Likelihood (Gilmour *et al.* 1995) and predict each individual’s breeding value using the Best Linear Unbiased Prediction algorithms (Henderson 1975, 1976, 1984). However, while effective, this method with its traditional pedigree-based approach does not adjust for the Mendelian sampling term, that is, this method ignores variation among family members of a half- or full-sib family around the family’s average relatedness (as all sibs are not alike (Hill and Weir 2011)). Furthermore, the utilization of the ***A***-matrix, specifically, in the case of the well-known “shallow” pedigree present within most forest tree breeding and testing populations does not permit detecting hidden co-ancestry and inbreeding. Consequently, individuals’ estimated breeding values are inflated by the overestimation of the additive genetic variance.

With the affordability, scalability, and high-throughput nature of Next Generation Sequencing technologies, tens of thousands of single nucleotide polymorphism (SNP) have become available for model and non-model species (Baird *et al.* 2008; Elshire *et al.* 2011; Peterson *et al.* 2012; Poland *et al.* 2012; Truong *et al.* 2012; Chen *et al.* 2013). This technical advancement made it possible to ascertain, with great level of accuracy, the actual fraction of alleles shared between individuals, and the estimates of the individuals’ pairwise realized relationship including potential inbreeding can be easily determined (Santure *et al.* 2010). Therefore, genomic fingerprinting data permit the accurate estimation of the realized relationships among any set of individuals, irrespective of their genealogy, to construct the realized genomic relationship matrix

(***G***-matrix) which can be used to substitute the ***A***-matrix (VanRaden 2008). This advancement represents a clear quantitative genetics watershed as the complete dependency on known pedigree relationships (i.e., ***A***-matrix) for estimating genetic parameters can be circumvented in the so-called “pedigree-free models” using the ***G***-matrix. As already demonstrated in earlier cases, the ***G***-matrix can provide relatively accurate genetic variance components and breeding values estimates without the need for elaborate mating designs (Thomas *et al.* 2002; Frentiu *et al.* 2008; Hayes *et al.* 2009b; El-Kassaby *et al.* 2012; Gay *et al.* 2013; Porth *et al.* 2013; Zapata-Valenzuela *et al.* 2013; Klápště *et al.* 2014; Muñoz *et al.* 2014).

The use of the ***G***-matrix in OP family testing has several implications and is expected to: 1) overcome the drawback of the average numerator relationship matrix (***A***-matrix) as genomic data will unravel any undetectable hidden relatedness such as full-sibs, self-sibs, and self-halfs that inflates the estimated additive genetic variance (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994) (Figure 3.1), 2) provide more accurate genetic co-variances among relatives, thus accounting for the Mendelian sampling term (Visscher *et al.* 2006), and 3) provide higher flexibility in capturing the allele frequency segregation in quantitative trait loci (QTLs) (present vs. absent QTL) (Lippert *et al.* 2013). Additionally, we hypothesize that also for OP families the use of genomic markers will create an opportunity to effectively decompose the genetic variance components, thus separating the additive and non-additive genetic components through the definition of realized genomic relationship matrix related to specific variance components, a so far unattainable feat for OP family testing.

Here, we used 1694 trees representing 214 white spruce OP families grown on one site (Mastigouche Arbortum, Quebec, Canada (Lat. 46° 38’ N, Long. 73° 13’ W, Elev. 230m)) in a randomized complete block design, replicated over six blocks (replications) with each OP family is represented by a 5-tree row plots within each of the six blocks (for complete details, see (Beaulieu *et al.* 2014)).We compared the genetic variance estimates generated from both, the average numerator relationship matrix (the expected relationships) and the realized genomic relationship matrix (the observed relationships), to demonstrate the genomic markers' utility in partitioning the genetic variance components into additive and non-additive effects. To our knowledge, this study provides the first attempt of such an analysis approach in OP families.

## 3.2 Materials and methods

### 3.2.1 White spruce open-pollinated progeny test, phenotype data, and genotyping

The white spruce (*Picea glauca* (Moench) Voss) phenotypic and genotypic data used are available online from the Dryad Digital Repository: doi:10.5061/dryad.6rd6f (Beaulieu *et al.* 2014). Briefly, the study site is a part of a larger, 3-site white spruce provenance-progeny test established in 1979 by the Canadian Forest Service in Quebec, Canada. Each site was planted as randomized complete block design with six blocks and five-tree row plots at 1.2 and 2.4m spacing within and between rows, respectively. The present study is based on a subset of the provenance-progeny test that include 8 individuals per each OP family from 214 families representing a total of 1,694 individuals, the average family representation per block was 1.32 trees since not all families were present in all blocks. It is noteworthy to state that the 214 open-pollinated families were selected from 43 provenances throughout Quebec, thus population effect might be present. Beaulieu et al. (2014), using principal component analysis, reported the presence of weak population structure with no defined geographical pattern. In fact, Beaulieu et al. (2014) estimated that 1.3% of the total variance was explained by the first two principal component analysis eigenvectors and indicated that their lack of population structure is concordant with previous studies using the same populations, thus population structure was not considered in the present study. Wood density was determined using X-ray densitometry from 12-mm increment cores collected at 1.3m from ground (see (Beaulieu *et al.* 2011), for details). Trees were genotyped for 7,338 single nucleotide polymorphic (SNP) loci from 2,814 genes using Illumina Infinium HD iSelect bead chip PgAS1 (Illumina, San Diego, CA, USA) (for details see Rigault et al. 2011). The data used are available from the Dryad Digital Repository: doi:10.5061/ dryad.6rd6f (Beaulieu *et al.* 2014).

### 3.2.2 Relationship matrices

The additive relationship matrix was estimated following:

$$\boldsymbol{G_{add}} = \frac{\boldsymbol{ZZ\prime}}{2\sum p_i(1-p_i)} \qquad [1]$$

where $\boldsymbol{Z}$ is rescaled genotype matrix following $\boldsymbol{M}$ - $\boldsymbol{P}$, $\boldsymbol{M}$ is genotype matrix containing genotypes coded as 0, 1, and 2 according to the number of alternative alleles and $\boldsymbol{P}$ is a vector of twice the allelic frequency, *p* (VanRaden 2008). The dominance genetic variance was fitted by including a marker based dominance relationship matrix following:

$$\boldsymbol{G_{dom}} = \frac{\boldsymbol{WW\prime}}{(2pq)^2} \qquad [2]$$

where $\boldsymbol{W}$ is matrix containing -2q$^2$ for alternative homozygote, 2pq for heterozygote, and -2p$^2$ for reference allele homozygote (Vitezica *et al.* 2013). Similarly, epistatic variance was fitted by including several relationship matrices capturing first order additive x additive, dominance x dominance, and additive x dominance interaction. The relationship matrices were constructed as the Hadamard product of the relationship matrices defined above: $\boldsymbol{G}_{\mathbf{add}}$#$\boldsymbol{G}_{\mathbf{add}}$, $\boldsymbol{G}_{\mathbf{dom}}$#$\boldsymbol{G}_{\mathbf{dom}}$ and $\boldsymbol{G}_{\mathbf{add}}$#$\boldsymbol{G}_{\mathbf{dom}}$ (Su *et al.* 2012; Muñoz *et al.* 2014).

The variance components from pedigree based analysis (ABLUP) were obtained by solving the following mixed model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_i u} + \boldsymbol{Z_j r} + \boldsymbol{Z_k rxf} + \boldsymbol{e} \qquad [3]$$

where $\boldsymbol{y}$ is vector of phenotypic measurements, $\boldsymbol{\beta}$ is vector of fixed effects (overall mean), $\boldsymbol{u}$ is vector of random additive genetic effects following $\boldsymbol{u}$~N(0, $\boldsymbol{A}\sigma_a^2$), where $\boldsymbol{A}$ is average numerator relationship matrix and $\sigma_a^2$ is additive genetic variance, $\boldsymbol{r}$ is vector of random replication effect following $\boldsymbol{r}$~N(0, $\boldsymbol{I}\sigma_r^2$), where $\sigma_r^2$ is replication variance, $\boldsymbol{rxf}$is vector of random replication x family interaction effects following $\boldsymbol{rxf}$~N(0, $\boldsymbol{I}\sigma_{\boldsymbol{rxf}}^2$), where $\sigma_{\boldsymbol{rxf}}^2$ is replication x family interaction variance, and $\boldsymbol{e}$ is a vector of the random residual effects following $\boldsymbol{e}$~N(0, $\boldsymbol{I}\sigma_e^2$) where $\sigma_e^2$ is residual error variance, $\boldsymbol{X}$, $\boldsymbol{Z_i}$, $\boldsymbol{Z_j}$, and $\boldsymbol{Z_k}$ are incidence matrices relating fixed and random effects to measurements in vector $\boldsymbol{y}$.

The variance components from the analysis using marker based additive relationship matrix (GBLUP-A) were obtained from the model described above but the average numerator relationship matrix $\boldsymbol{A}$ is substituted by marker based relationship matrix $\boldsymbol{G_{add}}$. The extended model for the dominance terms are performed as follows:

$$\boldsymbol{y = X\beta + Z_i u + Z_l d + Z_j r \ + Z_k rxf + e} \qquad [4]$$

where $\boldsymbol{d}$ is vector of the random dominance effect following $\boldsymbol{d}$~N(0, $\boldsymbol{G_{dom}}\sigma_d^2$) where $\sigma_d^2$ is the dominance variance. Additional model extension for epistatic terms is performed as follows:

$$\boldsymbol{y = X\beta + Z_i u + Z_l d + Z_m axa + Z_n dxd + Z_p axd + Z_j r + Z_k rxf + e} \qquad [5]$$

where $\boldsymbol{axa}$ is the vector of random additive x additive epistatic interaction effects following $\boldsymbol{axa}$~N(0, $\boldsymbol{G_{add\#add}}\sigma_{axa}^2$) where $\sigma_{axa}^2$ is the additive x additive epistatic interaction variance, $\boldsymbol{dxd}$ is the vector of random dominance x dominance epistatic interaction effects following $\boldsymbol{dxd}$~N(0, $\boldsymbol{G_{dom\#dom}}\sigma_{dxd}^2$) where $\sigma_{dxd}^2$ is dominance x dominance epistatic interaction variance, $\boldsymbol{axd}$ is the vector of random additive x dominance epistatic interaction effects following $\boldsymbol{axd}$~N(0,

$\boldsymbol{G_{add\#dom}}\sigma^2_{axd}$) where $\sigma^2_{axd}$ is the additive x dominance epistatic interaction variance, and $\boldsymbol{Z_l}$, $\boldsymbol{Z_m}$, $\boldsymbol{Z_n}$ and $\boldsymbol{Z_p}$ are incidence matrices relating random effects to measurements in vector $\boldsymbol{y}$.

Narrow-sense heritability was estimated as $\hat{h}^2 = \hat{\sigma}^2_a/\hat{\sigma}^2_p$, where $\hat{\sigma}^2_a$ represents the estimate of the additive variance and $\hat{\sigma}^2_p$ equals the sum of $\hat{\sigma}^2_e$ and all random model effect variance components estimates such as additive, dominance, additive x additive, additive x dominance, dominance x dominance interactions following that of the ABLUP and GBLUPs (GBLUP-A, GBLUP-AD, and GBLUP-ADE) models, respectively (Table 3.1). The analyses and the derived genetic and environmental parameters and their standard errors for the ABLUP and GBLUPs were estimated using ASReml$^{TM}$ v. 3.0 software (Gilmour *et al.* 2009).

### 3.2.3 Models comparison and cross-validation

Models were compared using the AIC estimates obtained from each analysis (Gilmour *et al.* 2009) and the precision of the estimated variance components and their dependence was assessed by investigation of accumulated eigenvalues of the asymptotic sampling correlation matrix of variance component estimates $\boldsymbol{F}$, where $\boldsymbol{F} = \boldsymbol{L}^{-1/2}\boldsymbol{V}\boldsymbol{L}^{-1/2}$ using the asymptotic variance-covariance matrix of estimates of variance components $\boldsymbol{V}$ and its diagonal matrix $\boldsymbol{L}$ (Muñoz *et al.* 2014).

A 10-fold cross-validation scenario with five replications were used to assess prediction accuracy and consistency within and between the various models, respectively. Folding of the training population was either random, block restricted, or family restricted. The latter scenario removes the genetic relatedness between the training and validation populations according to the pedigree information. That is, all individuals belonging to a single OP family were strictly assigned to either the training or validation population. Block restricted folding was performed as a leave one block out scenario. That is, all individuals belonging to single block were assigned as the validation population, while the individuals belonging to remaining five experimental blocks were randomly divided into 10 folds as the training population. Random folding had no prior restriction when assigning the folds.

Prediction accuracy within, and consistency between models was evaluated using the mean Pearson correlation from the five replications. Specifically, the correlation values for each replication were calculated as:

$$r_{EBV_l,PBV_{mn}} = \frac{cov(EBV_l, PBV_{mn})}{\sigma_{EBV_l}\sigma_{PBV_{mn}}} \qquad [6]$$

where, EBV refers to the individual additive breeding value of the validation population obtained using the entire data set (1,694 individuals) for the $l^{\text{th}}$ model (ABLUP, GBLUP-A, GBLUP-AD, GBLUP-ADE), PBV is the individual additive breeding value of the validation population obtained using the $m^{\text{th}}$ model (ABLUP, GBLUP-A, GBLUP-AD, GBLUP-ADE) and $n^{\text{th}}$ cross-validation scenario (random, block, family), $cov$ is the covariance, and $\sigma$ is the standard deviation. Standard error of the mean for the correlations was computed using the following equation:

$$SE = \frac{\sigma}{\sqrt{n}} \qquad [7]$$

where $\sigma$ is the standard deviation of the Pearson correlations and $n$ is the number of replicates.

## 3.3 Results

### 3.3.1 Genetic variance components and heritability estimates

As expected, replication and family x replication interaction produced constant variance components across the four studied models for both height (4.7 and 22%) and wood density (1 and 2-5%), leaving most of the within replication effects residing within the residual terms (Table 3.1). The greatest observed difference between the pedigree- (ABLUP) and the marker-based (GBLUP-A) models was the substantial discrepancy of the additive genetic variance estimates' magnitude

(Table 3.1). The additive genetic variance estimated from GBLUP- A were 64.4 and 46.9% of those from the ABLUP for height and wood density, respectively (Table 3.1). Naturally, this change is reflected in the residual terms as they increased to 110.1 and 168.4% of that of the ABLUP and subsequently resulting in substantial reduction of narrow-sense heritability estimates ($\hat{h}^2$: from 0.25 down to 0.16 for height and from 0.61 down to 0.30 for wood density comparing ABLUP vs. GBLUP-A, respectively). Overall, narrow-sense heritability was reduced by 65% for height, and 50% for wood density, when the genomic relationship matrix GBLUP-A was employed (Figure 3.1), highlighting known caveats of OP progeny testing. Also, the inflation of additive genetic variance observed in ABLUP and the subsequent impact on heritability estimates were expected, thus it is more reasonable to use the results from the GBLUP-A as the basis for comparing the extended analyses that included dominance (GBLUP-AD), and epistasis and dominance (additive x additive, dominance x dominance, and additive x dominance first-order interaction) (GBLUP-ADE).

The GBLUP-AD analysis produced identical results to that of the GBLUP-A confirming the existence of minuscule and non-significant dominance variance estimates, accounting for 1.13 and 2.84% of the total phenotypic variance for height and wood density, respectively (Table 3.1). This is not surprising considering the small sample size of the studied OP families (≈8 individuals/family) or simply due to the fact that these traits do not possess dominance genetic variance (see Discussion). Including dominance variance in the models increased the AIC values for the models, indicating that GBLUP-AD models were over-fitted compared to GBLUP-A models (Table 3.1), and that the simpler GBLUP-A models should be preferred.

The GBLUP-ADE analysis produced the most striking results with further reduction as to the additive genetic and the residual variances compared to the ABLUP and GBLUP-AD models for height and wood density, respectively (Table 3.1). This observed reduction in the additive genetic and residual variances was caused by the presence of significant additive x additive genetic variance within the total phenotypic variance (Table 3.1). This observed additive x additive genetic variance in turn resulted in further reduction of the narrow-sense heritability estimates; from 0.16 to 0.13, and from 0.30 to 0.18 in GBLUP-AD compared to GBLUP-ADE for height and wood density, respectively. Again, the GBLUP-ADE analysis did not cause any change to the dominance variances (Table 3.1). Small and not significant dominance x dominance and additive x dominance first-order interactions were observed for height and wood density in the GBLUP-ADE (Table 3.1). The AIC statistics for this model produced the best fit with value lower than that observed for all tested models for wood density (-9,726.65), supporting the inclusion of the additional epistasis terms in the model, specifically that of the additive x additive (Table 3.1). Unexpectedly, the AIC for GBLUP-A (17,465.80) produced the best fit for height (Table 3.1).

### 3.3.2 Models comparison and cross-validation

Comparing the standard errors for the predictions (SEP) of breeding values (BV) between the ABLUP and GBLUP-A models, all of SEPs for height and wood density BVs were smaller for GBLUP-A compared to ABLUP as all SEPs were below the 45° reference lines, clearly indicating the superiority of the GBLUP-A model (Figure 3.2). GBLUP-A and GBLUP-AD models produced identical results owing to the lack of significant dominance effects and all SEPs for height and wood density BVs resided on the diagonal 45° reference lines. Additionally, SEPs for height and wood density BVs from the GBLUP-ADE model were smaller than the corresponding SEPs produced by the GBLUP-A model indicating the effectiveness of the GBLUP-ADE model (Figure 3.2). When we compared the pedigree- and the marker-based models using the cumulative proportion of variance that was explained by eigenvalues of the sampling variance-covariance matrix of variance component estimates, we found that the GBLUP-A outperformed the pedigree-based (ABLUP) models as indicated by the closeness of their respective lines to the ideal scenario (straight line) where the variance components are completely independent (Figure 3.3). Finally, since the GBLUP-ADE model does not have a corresponding model in the pedigree method, GBLUP-ADE was plotted only against the 45° diagonal for reference (Figure 3.3).

Cross-validation prediction accuracies (Table 3.2; diagonals) indicated that the ABLUP model was associated with the lowest values among all tested models for both random and block restricted folding (range: 0.451-0.475 and 0.439-0.449 for height and wood density, respectively), while the GBLUP models produced greater prediction accuracies under the same two folding scenarios (range: 0.735-0.772 and 0.748-0.783, for height and wood density, respectively). Prediction accuracies were lowest under the family restricted scenario for the GBLUP models (range: 0.683-0.698 and 0.651-0.658, for height and wood density, respectively), with random folding producing the greatest prediction accuracies. Comparison of prediction accuracies among the GBLUP models using random folding showed that difference between GBLUP-A and GBLUP-AD were not significant (based on standard errors), however the two were significantly greater than GBLUP-ADE for both height and wood density. The family and random folding scenarios both produced no significant differences in prediction accuracy among the GBLUP models.

Pairwise model comparisons (Table 3.2; off-diagonals) showed high consistency between all GBLUP models within the individual folding scenarios. It is also noteworthy to mention that under the family folding scenario, the ability of ABLUP to produce across family prediction challenges the assumption of zero expected relatedness among OP families, thus predictions of individual additive breeding values here would simply be equal to the overall mean.

## 3.4 Discussion

Traditionally, the pedigree-based average numerator relationship matrix (***A***-matrix) is used to estimate the genetic variance components for forest tree progeny test populations. The estimated genetic variance components (e.g., additive and dominance genetic variances, etc.) often are mating design-dependent and the mating scheme determines which component can be obtained. In most cases, this approach is incapable of disentangling the within-family genetic from within-family micro-environment effects. This is even more problematic in OP family screening as separating additive from non-additive genetic variances is limited by shallow pedigrees and lack of connectedness among the tested families; furthermore, as shown in Table 1, the estimated additive genetic variance is inflated as the half-sib assumption is hardly fulfilled (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994). In fact, the estimated genomic pairwise relationships of the studied 214 OP families showed deviation from the expected 0.25 coefficient of relatedness for half-siblings, confirming causes for additive genetic variance overestimation, while the relationships among members of unrelated families clustered around the expected 0.0 (Figure 3.1). The availability of dense genomic marker panels made it possible to genotype individuals for a large numbers of single nucleotide polymorphisms (SNPs) and obtain the realized genomic relationship matrix (***G***-matrix) among these individuals. In turn, the ***G***-matrix can be used as a substitute to the ***A***-matrices to estimate more accurate and precise genetic variance components as the ***G***-matrix represents the realized pedigree as well as having the capacity to exploit the Mendelian sampling/segregation within families (VanRaden 2008; Hayes *et al.* 2009b). It is worthwhile to note that some of the recently reported gain increase in animal breeding programs is mainly due to exploiting the Mendelian sampling term (Avendaño *et al.* 2004).

The utility of the ***G***-matrix in generating improved estimates of the genetic variance parameters from experimental populations of forest trees (e.g., full-sib families) have recently been explored (Zapata-Valenzuela *et al.* 2013; Klápště *et al.* 2014; Muñoz *et al.* 2014). The present study, to our knowledge, represents the first attempt to implement the ***G***-matrix in OP family testing, thus not only, overcoming the common bias associated with the unfulfilled half-sib assumption, but also separating the additive from the non-additive genetic variance components. It is well known that separating the additive from the non-additive (dominance and epistatic variances) genetic components requires elaborate mating designs with large number of inter-connected full-sib families coupled with the inclusion of replicated clonal material (Foster and Shaw 1988; Bradshaw and Foster 1992). Our study accomplished a mixed-model approach for variance decomposition, providing realized estimates of the additive, dominance, and epistatic genetic variances without the need for mating designs to generate inter-connected full-sib families or vegetative propagation for the production of replicated clonal material.

It is interesting to note that the estimated additive genetic variances for the three realized genomic relationship matrix-based analyses (GBLUP-A, GBLUP-AD, and GBLUP-ADE) were lower than those of the average numerator relationship matrix (ABLUP) (Table 3.1), an observation already reported for mice (Lee *et al.* 2010), loblolly pine (Muñoz *et al.* 2014), and Brown Swiss cattle populations (Loberg *et al.* 2015). The improved performance of the GBLUP-A compared to that of the ABLUP indicates that the former model took full advantage of: 1) the within family variation (i.e., Mendelian sampling term), 2) discerning if full-sibs, self-sibs, and self-halfs existed within the studied 214 open-pollinated families, 3) the ability to estimate among-family relationships even if it was as small as seen in Figure 3.1, and 3.4) identifying pedigree errors if present, as shown in Figure 3.1 (i.e., some individuals have a coefficient of relationship of 0.0 within the same OP family). The observed reduction in the additive genetic variance between the two models (ABLUP vs. GBLUP-A) resulted in concomitant increase in the residual error terms and hence considerably reduced narrow-sense heritability estimates for height (0.25 vs. 0.16) and also for wood density (0.61 vs. 0.30) along with improvement in the model fit based on improved AIC values (Table 3.1). Additionally, GBLUP-A produced greater precision for its estimated breeding value (EBV) as indicated by the EBV's smaller standard errors compared to the ABLUP (Figure 3.2; ABLUP vs. GBLUP-A).

It is noteworthy to mention that the present study is based on data collected from one site, thus there is a chance that the estimated genetic parameters could be upwards biased due to the genotype x environment confounding effects specific to this particular site or year. However, results from chapter four for the same species and attributes for a set of 25 open-pollinated families planted in replicated trials over three sites in British Columbia were consistent to that reported here with the added benefits of estimating the additive and dominance x site (environment) interactions.

The observed overall trend in genetic variance decomposition persisted when the dominance genetic variance was estimated using the alternative genotypic approach proposed by Su et al. (2012) and discussed by Vitezica et al. (2013); however, the dominance genetic variance of wood density showed a slight increase (Table 3.3).

Additionally, estimating the dominance genetic variance is only feasible when full-sib families are available (Zapata-Valenzuela *et al.* 2013; Klápště *et al.* 2014; Muñoz *et al.* 2014). This scenario is easily resolved when pedigree- and marker-based models are compared for mating design accommodating full-sib families (Muñoz *et al.* 2014) However, the utility of the GBLUP-AD model in OP family testing is still worth exploring to discern the dominance genetic variance - if existing - as well as separating the genetic variances from the confounding environment effects.

It should be noted that the goodness-of-fit statistics (AIC) for the GBLUP-AD clearly indicated that adding the dominance genetic variance resulted in model over fit and this is expected due to the extremely small and non-significant dominance genetic variance (1.1 and 2.8% for height and wood density, respectively).

The model that included the additive, dominance, and epistatic variances (GBLUP-ADE) offered better partitioning of the variance complements, as the additive x additive epistatic variance became extremely pronounced and accounted for 11 and 52% of the total variance for height and wood density, respectively (Tables 3.1 and 3.3). When we removed the dominance genetic variance from the GBLUP-ADE model, the revised models (GBLUP-AE) produced better model fit for height (17,466.9 vs. 14,472.9) and wood density (-9,732.6 vs. -9,726.6), confirmed that dominance variance was negligible (Table 3.1). Interestingly, both models (GBLUP-ADE and GBLUP-AE) produced similar variance components apportionment and heritability estimates (Table 3.1). Similar magnitude of the additive x additive epistatic variance to that of the additive variance, *per se*, was also observed in loblolly pine (Muñoz *et al.* 2014), a situation meeting theoretical expectations where the additive x additive epistatic variance is commonly absorbed by both the additive and the residual variances (Lynch, M., Walsh 1998; Jannink 2007; Mackay 2014). The power of the GBLUP-ADE and/or GBLUP-AE models in identifying and separating the additive x additive epistasis from the additive genetic variance lies in the genetic background of the tested families for providing a range of options to demonstrate all established interactions between the alleles at the various loci that are affecting the studied traits. The magnitude of the epistatic additive x additive genetic variance observed for height and wood density along with the AIC values produced from the tested models require some reflection. The observed AIC values support GBLUP-A and GBLUP-AE to be the best model for height and wood density, respectively

(Table 3.1). However, in wood density where the additive x additive is ≈3 times as that of the additive variance, the prediction accuracy of the GBLUP-AE and GBLUP-A models were almost similar (Tables 3.1 and 3.2). This indicates that the additive x additive and additive relationship matrices are in a "tug-of-war" state over the same variance. In fact, we estimated the correlation between these two relationship matrices and it was close to perfect correlation (r = 0.988), confirming our notion and makes us believe that while we observed exceedingly large epistatic additive x additive genetic variance, as the impact on predicting the breeding values between GBLUP-A and GBLUP-AE is similar (Figure 3.5).

The subject of genetic epistasis is controversial as all variance components, including epistasis are dependent on the allele frequencies in the studied population. Thus, epistasis could have an allusive and unique effect across different scenarios (Hill *et al.* 2008; Mackay 2014). The role of epistasis on the genetic architecture of quantitative traits is still not clearly determined due to several discrepancies between statistical and functional definition of epistasis. The statistical approach considers the epistatic variance orthogonal to the additive genetic variance and assumes a clear determination (separation) of both components by the implementation of independent terms in the model. Moreover, the epistatic effects are transient and disappear by breaking of linkage disequilibrium (LD) (Hill *et al.* 2008; Crow 2008, 2010). The functional approach assumes that allelic substitution effect depends on the genetic background. Hill et al. (2008) based their empirical evidence on an exhaustive review across a wide range of species. This includes comparisons between narrow- and broad-sense heritability estimates, concluding that complex traits are mainly controlled by additive genetic variance as most studied cases supported the notion that the majority of the genetic variance appeared to be additive (Crow 2010). However, in the present study, if we utilized the heritability estimates derived from the GBLUP-A or GBLUP-AD models alone (without the application of the GBLUP-ADE and/or GBLUP-AE models), then the part of the genetic variance attributable to additive x additive interaction would have been excluded from the calculations, and thus our conclusion would have been mainly based on inflated additive genetic variance. Clearly, the utilization of the marker-based relationship method enabled disentangling the additive from the non-additive genetic component, while effectively accounting for the proper environment variance through the removal of possible confounding effects. Such methodology provides much more realistic breeding value estimations for an individual.

Habier et al. (2007; 2010; 2013) indicated that the realized genomic relationships do not only capture relatedness among individuals but also the LD between SNPs and quantitative trait loci, the deviation from independent segregation of alleles on the same gamete if the loci are linked (co-segregation or classical linkage), as well as the additive genetic relationship. Habier et al. (2013) demonstrated that these types of information collectively have different effects on the accuracy of the EBV. As a result, it is safe to state that there is more to the realized genomic relationship than the straightforward accounting for the Mendelian sampling term, hence resulting in the superior decomposition of the genetic variance components and breeding values estimation.

When ABLUP is used to estimate the genetic variance components from either half- or full-sib families, the above factors are barely considered, except those that were captured through common ancestry. As indicted above, the OP/half-sib structure is incapable of estimating the dominance and the epistatic genetic variances. This situation was clearly demonstrated in the study by Muñoz et al. (2014), as more accurate breeding value estimates and effective partitioning of variance components were obtained from their single site, full-sib, and clonally replicated loblolly pine experiment. The present study, on the other hand, demonstrated the power of the realized genomic relationship in quantitative genetic analyses using a more challenging structure (OP families). As proof-of-concept we compared the rank order among the top 50 performing individuals based on the conventional ABLUP versus the GBLUP-ADE/GBLUP-AE (see details in the interaction plots of Figure 3.4). Only 23 and 33 of the top 50 individuals persisted from ABLUP to GBLUP-ADE/GBLUP-AE for height and wood density, respectively, and overall, the individuals' ranking among the top 50 trees dramatically changed from ABLUP to GBLUP-AE. Interestingly, for the 10 best performing trees, only 2 and 4 individuals persisted for height and wood density, respectively (Figure 3.4). The true estimated breeding value of an individual is commonly determined from experiments with deep pedigree with ample connectedness; however, when the ABLUP approach is used in forestry progeny testing experiments that are characterized by shallow and inadequate connectedness, then the obtained breeding value is expected to greatly deviate from its true value as the assumption of mixed models of error-free covariance matrices is not met (Mrode 2005). The greatest difference between the GBLUP and ABLUP models is the ability of the former to more precisely define the genetic relationship between any two individuals as compared to the latter (Figure 3.1). Our models cross-validation supports this notion as the prediction accuracy for the GBLUP models was greater than those produced by ABLUP, regardless of the folding scenario (Table 3.2; diagonals). This difference is due to the quantity of realized pairwise genetic relationship information used for prediction, wherein ABLUP only the information from the pedigreed OP family is used to predict the breeding value. Conversely in GBLUP all information from related individuals are used regardless of family assignment, this can be seen in the family folding scenario (Table 3.2). Thus, we believe that the estimated breeding values produced by the GBLUP models are closer to the true value as more information is used.

We feel that our results have important and immediate implications for tree improvement programs in forestry as most programs are long-term and resource-dependent (El-Kassaby 1995).

The conventional genetic improvement in forestry follows the classical recurrent selection scheme with repeated cycles of selection, breeding and testing over time and space (Gamal El-Dien *et al.* 2015; Ratcliffe *et al.* 2015). These programs often include: 1) phenotypic selection of untested candidate parents from natural or managed forests, 2) propagation of the selected parents as grafts followed by a period of inactivity until sexual maturity, 3) the sexual production of structure pedigreed offspring from the selected parents using a specific mating designs, 4) field testing over vast geographic areas for a reasonable period to attain meaningful data for target traits, 5) estimation of genetic parameters and ranking of individuals based on their breeding values, and 6) genotypic selection of superior individuals for the second round of breeding and/or seed production from seed orchard populations. Obviously, the completion of a single breeding-testing-selection cycle is a protracted endeavor due to several uncontrollable biological factors; namely, the time needed for reaching sexual maturity for structured pedigree production and reproductive phenology and fertility variation that hinder the mating design completion (El-Kassaby *et al.* 1984; El-Kassaby 1989; El-Kassaby and Barclay 1992). Therefore, the use of OP family testing, as demonstrated in the present study, allows immediate testing and evaluating large number of individuals using their naturally produced offspring through wind-pollination without the need for structured pedigree. The present study also demonstrated the utility of the realized genomic relationship approach in providing a simple and extremely efficient method for generating accurate genetic parameters from a simple OP testing that is characterized by shallow genealogy that is typical of most forest tree testing populations. It is noteworthy to mention that the use of the realized genomic relationship also allowed the generation of genetic parameters comparable to those generated only from elaborate mating designs coupled with cloning approaches. In conclusion, the utility of the realized genomic relationship in a simple, yet extremely efficient testing method, such as OP families, cannot be overlooked and calls for the re-evaluation of present-day conventional elaborate testing methods that are incapable of providing the genetic information produced in the present study.

## 3.5 Summary

**Background:** The open-pollinated (OP) family testing combines the simplest known progeny evaluation and quantitative genetics analyses as candidates' offspring are assumed to represent independent half-sib families. The accuracy of genetic parameter estimates is often questioned as the assumption of "half-sibling" in OP families may often be violated.

**Results:** We compared the pedigree- versus marker-based genetic models by analyzing 22-year height and 30-year wood density for 214 white spruce (*Picea glauca* (Moench) Voss) OP families represented by 1,694 individuals growing on one site in Quebec, Canada. Assuming half-sibling, the pedigree-based model was limited to estimating the additive genetic variances which, in turn, were grossly overestimated as they were confounded by very minor dominance and major additive-by-additive epistatic genetic variances. In contrast, the implemented genomic pairwise realized relationship models allowed the disentanglement of additive from all non-additive factors through genetic variance decomposition.

**Conclusions:** The marker-based models produced more realistic narrow-sense heritability estimates and, for the first time, allowed estimating the dominance and epistatic genetic variances from OP testing. In addition, the genomic models showed better prediction accuracies compared to pedigree models and were able to predict individual breeding values for new individuals from untested families, which was not possible using the pedigree based model. Clearly, the use of marker-based relationship approach is effective in estimating the quantitative genetic parameters of complex traits' even under simple and shallow pedigree structure.

**Table 3.1 Estimates of genetic variance components and their standard errors for height (HT) and wood density (WD) for the Québec white spruce population across the four genetic models.**

| | | ABLUP | | GBLUP-A | | GBLUP-AD | | GBLUP-ADE | | GBLUP-AE | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trait** | **S.O.V.** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** |
| **HT** | $\sigma^2_{Rep}$ | 561.4 (383.72) | 4.70 | 554.8 (379.47) | 4.68 | 555.9 (380.27) | 4.69 | 555.3 (379.85) | 4.69 | 555.2 (3.80E+02) | 4.69 |
| | $\sigma^2_{F*Rep}$ | 2,624.8 (497.90) | 21.97 | 2,653.7 (479.62) | 22.38 | 2,658.6 (479.60) | 22.43 | 2,614.4 (481.24) | 22.08 | 2,613.1 (480.94) | 22.07 |
| | $\sigma^2_{A}$ | 2,178.9 (879.65) | 18.24 | 1,404.0 (413.19) | 11.84 | 1,385.3 (413.98) | 11.69 | 1,160.9 (482.52) | 9.80 | 1,159.0 (480.98) | 9.79 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 133.29 (391.83) | 1.13 | 12.15 (406.64) | 0.10 | N/A | |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 1,334.8 (1664.2) | **11.27** | 1,352.7 (1,595.60) | **11.43** |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 9.86E-03 (2.23E-03) | 0.00 | N/A | |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 9.86E-03 (2.23E-03) | 0.00 | N/A | |
| | $\sigma^2_{E}$ | 6,581.7 (808.23) | 55.09 | 7,243.6 (535.33) | 61.10 | 7,119.8 (640.48) | 60.06 | 6,163.2 (1391.3) | 52.05 | 6,159.1 (1,390.90) | 52.02 |
| | $h^2$ | 0.249 (0.095) | | 0.162 (0.046) | | 0.160 (0.045) | | 0.134 (0.055) | | 0.134 (0.054) | |
| | **AIC** | 17,478.64 | | 17,465.80 | | 17,467.66 | | 17,472.94 | | 17,466.94 | |
| **WD[1]** | $\sigma^2_{Rep}$ | 1.36E-05 (1.11E05) | 1.07 | 1.24E-05 (1.04E-05) | 1.01 | 1.26E-05 (1.05E-05) | 1.02 | 1.34E-05 (1.10E-05) | 1.10 | 1.34E-05 (1.10E-05) | 1.10 |
| | $\sigma^2_{F*Rep}$ | 2.47E-05 (4.77E-05) | 1.95 | 5.89E-05 (4.70E-05) | 4.78 | 5.88E-05 (4.69E-05) | 4.78 | 4.65E-05 (4.65E-05) | 3.83 | 4.65E-05 (4.65E-05) | 3.83 |
| | $\sigma^2_{A}$ | 7.48E-04 (1.28E-04) | 59.01 | 3.51E-04 (5.52-E05) | 28.50 | 3.48E-04 (5.52E-05) | 28.25 | 2.07E-04 (5.85E-05) | 17.05 | 2.07E-04 (5.85E-05) | 17.05 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 3.50E-05 (4.88E-05) | 2.84 | 7.90E-11 (2.78E-11) | 0.00 | N/A | |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 6.32E-04 (1.34E-04) | **52.03** | 6.32E-04 (1.34E-04) | **52.03** |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 5.05E-10 (1.78E-10) | 0.00 | N/A | |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 5.05E-10 (1.78E-10) | 0.00 | N/A | |
| | $\sigma^2_{E}$ | 4.81E-04 (1.12E-03) | 37.96 | 8.10E-04 (6.28E-05) | 69.71 | 7.77E-04 (7.62E-05) | 63.11 | 3.16E-04 (1.11E-04) | 25.98 | 3.16E-04 (1.11E-04) | 25.98 |
| | $h^2$ | 0.609 (0.093) | | 0.303 (0.043) | | 0.300 (0.043) | | 0.179 (0.049) | | 0.179 (0.049) | |
| | **AIC** | -9,687.42 | | -9,716.32 | | -9,714.86 | | -9,726.64 | | -9,732.64 | |

[1] log transformation

**Table 3.2 Correlations for height (HT) and wood density (WD) between estimated individual additive breeding values (EBV) and predicted individual additive breeding values (PBV)**

| | | | **EBV – Full data** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | **HT** | | | | **WD** | | | |
| | | | **ABLUP** | **GBLUP-A** | **GBLUP-AD** | **GBLUP-ADE** | **ABLUP** | **GBLUP-A** | **GBLUP-AD** | **GBLUP-ADE** |
| **PBV – Cross Validation** | **Random Folding** | **ABLUP** | **0.475 (0.003)** | 0.407 (0.003) | 0.407 (0.003) | 0.401 (0.003) | **0.449 (0.004)** | 0.554 (0.004) | 0.554 (0.004) | 0.523 (0.004) |
| | | **GBLUP-A** | 0.331 (0.004) | **0.771 (0.003)** | 0.770 (0.003) | 0.772 (0.003) | 0.402 (0.002) | **0.781 (0.001)** | 0.781 (0.001) | 0.773 (0.001) |
| | | **GBLUP-AD** | 0.334 (0.003) | 0.773 (0.002) | **0.772 (0.002)** | 0.774 (0.002) | 0.405 (0.004) | 0.783 (0.003) | **0.783 (0.003)** | 0.775 (0.003) |
| | | **GBLUP-ADE** | 0.322 (0.004) | 0.762 (0.003) | 0.761 (0.003) | **0.765 (0.003)** | 0.385 (0.002) | 0.765 (0.002) | 0.765 (0.002) | **0.773 (0.002)** |
| | **Block Folding** | **ABLUP** | **0.451 (0.001)** | 0.381 (0.001) | 0.381 (0.001) | 0.374 (0.001) | **0.439 (0.000)** | 0.549 (0.000) | 0.549 (0.000) | 0.518 (0.000) |
| | | **GBLUP-A** | 0.329 (0.000) | **0.735 (0.001)** | 0.735 (0.001) | 0.736 (0.001) | 0.383 (0.000) | **0.748 (0.000)** | 0.748 (0.000) | 0.739 (0.000) |
| | | **GBLUP-AD** | 0.328 (0.001) | 0.734 (0.001) | **0.735 (0.001)** | 0.736 (0.001) | 0.383 (0.000) | 0.748 (0.001) | **0.748 (0.000)** | 0.740 (0.000) |
| | | **GBLUP-ADE** | 0.313 (0.001) | 0.711 (0.001) | 0.712 (0.001) | **0.715** (0.001) | 0.366 (0.000) | 0.728 (0.001) | 0.728 (0.001) | **0.733** (0.001) |
| | **Family Folding** | **ABLUP** | NA[1] | NA | NA | NA | NA | NA | NA | NA |
| | | **GBLUP-A** | 0.178 (0.011) | **0.683 (0.010)** | 0.682 (0.010) | 0.691 (0.009) | 0.249 (0.006) | **0.651 (0.005)** | 0.651 (0.005) | 0.663 (0.005) |
| | | **GBLUP-AD** | 0.188 (0.005) | 0.692 (0.005) | **0.691 (0.005)** | 0.699 (0.005) | 0.254 (0.003) | 0.656 (0.002) | **0.656 (0.002)** | 0.668 (0.002) |
| | | **GBLUP-ADE** | 0.190 (0.006) | 0.691 (0.006) | 0.689 (0.006) | **0.698 (0.006)** | 0.228 (0.007) | 0.627 (0.007) | 0.627 (0.007) | **0.658 (0.006)** |

[1] NA; predicted individual additive breeding value is equal to the overall mean of the model. Note: Validation produced by 10-fold cross-validation for the four models (ABLUP, GBLUP-A, GBLUP-AD, and GBLUP-ADE) using random, block, and family based folding. Prediction accuracies are represented by bold diagonals and pairwise model correlations on the off-diagonals (standard errors in parentheses).

**Table 3.3 Estimates of genetic variance components and their standard errors using the dominance matrix proposed by Su et al. (2012) and discussed by Vitezica *et al.* (2013)**

| | | ABLUP | | GBLUP-A | | GBLUP-AD | | GBLUP-ADE | | GBLUP-AE | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trait** | **S.O.V.** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** |
| **HT** | $\sigma^2_{Rep}$ | 561.40 (383.72) | 4.70 | 554.81 (379.47) | 4.68 | 558.75 (381.80) | 4.71 | 557.67 (381.19) | 4.71 | 555.15 (379.81) | 4.69 |
| | $\sigma^2_{F*Rep}$ | 2,624.8 (497.90) | 21.97 | 2,653.7 (479.62) | 22.38 | 2,663.2 (479.19) | 22.46 | 2,625.2 (481.03) | 22.16 | 2613.1 (480.94) | 22.07 |
| | $\sigma^2_{A}$ | 2,178.9 (879.65) | 18.24 | 1,404.0 (413.19) | 11.84 | 1,259.6 (457.28) | 10.62 | 1,110.2 (500.29) | 9.37 | 1,159.0 (480.98) | 9.79 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 350.24 (535.22) | **2.95** | 215.63 (552.36) | **1.82** | N/A | |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 1134 (1653.9) | 9.57 | 1,352.7 (1595.6) | 11.43 |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 7.79E-03 (1.75E-03) | 0.00 | N/A | |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 5.09E-03 (1.14E-03) | 0.00 | N/A | |
| | $\sigma^2_{E}$ | 6,581.7 (808.23) | 55.09 | 7,243.6 (535.33) | 61.10 | 7,028.6 (622.61) | 59.26 | 6,203.8 (1391.6) | 52.37 | 6,159.1 (1390.9) | 52.02 |
| | $h^2$ | 0.249 (0.095) | | 0.162 (0.046) | | 0.146 (0.051) | | 0.128 (0.057) | | 0.134 (0.054) | |
| | **AIC** | 17,478.64 | | 17,465.80 | | 17,467.30 | | 17,472.76 | | 17,466.94 | |
| **WD[1]** | $\sigma^2_{Rep}$ | 1.36E-05 (1.11E05) | 1.07 | 1.24E-05 (1.04E-05) | 1.01 | 1.24E-05 (1.04E-05) | 1.01 | 1.32E-05 (1.09E-05) | 1.09 | 1.34E-05 (1.10E-05) | 1.10 |
| | $\sigma^2_{F*Rep}$ | 2.47E-05 (4.77E-05) | 1.95 | 5.89E-05 (4.70E-05) | 4.78 | 5.39E-05 (4.66E-05) | 4.37 | 4.38E-05 (4.63E-05) | 3.61 | 4.65E-05 (4.65E-05) | 3.83 |
| | $\sigma^2_{A}$ | 7.48E-04 (1.28E-04) | 59.01 | 3.51E-04 (5.52-E05) | 28.50 | 2.91E-04 (5.93E-05) | 23.63 | 1.78E-04 (6.07E-05) | 14.65 | 2.07E-04 (5.85E-05) | 17.05 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 1.48E-04 (6.78E-05) | **11.99** | 9.64E-05 (6.64E-05) | **7.94** | N/A | |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 5.71E-04 (1.37E-04) | 46.98 | 6.32E-04 (1.34E-04) | 52.03 |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 1.54E-10 (5.36E-11) | 0.00 | N/A | |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 5.00E-10 (1.75E-10) | 0.00 | N/A | |
| | $\sigma^2_{E}$ | 4.81E-04 (1.12E-03) | 37.96 | 8.10E-04 (6.28E-05) | 69.71 | 7.26E-04 (7.20E-05) | 59.00 | 3.13E-04 (1.09E-04) | 25.73 | 3.16E-04 (1.11E-04) | 25.98 |
| | $h^2$ | 0.609 (0.093) | | 0.303 (0.043) | | 0.250 (0.048) | | 0.154 (0.051) | | 0.179 (0.049) | |
| | **AIC** | -9,687.42 | | -9,716.32 | | -9,719.42 | | -9,728.84 | | -9,732,64 | |

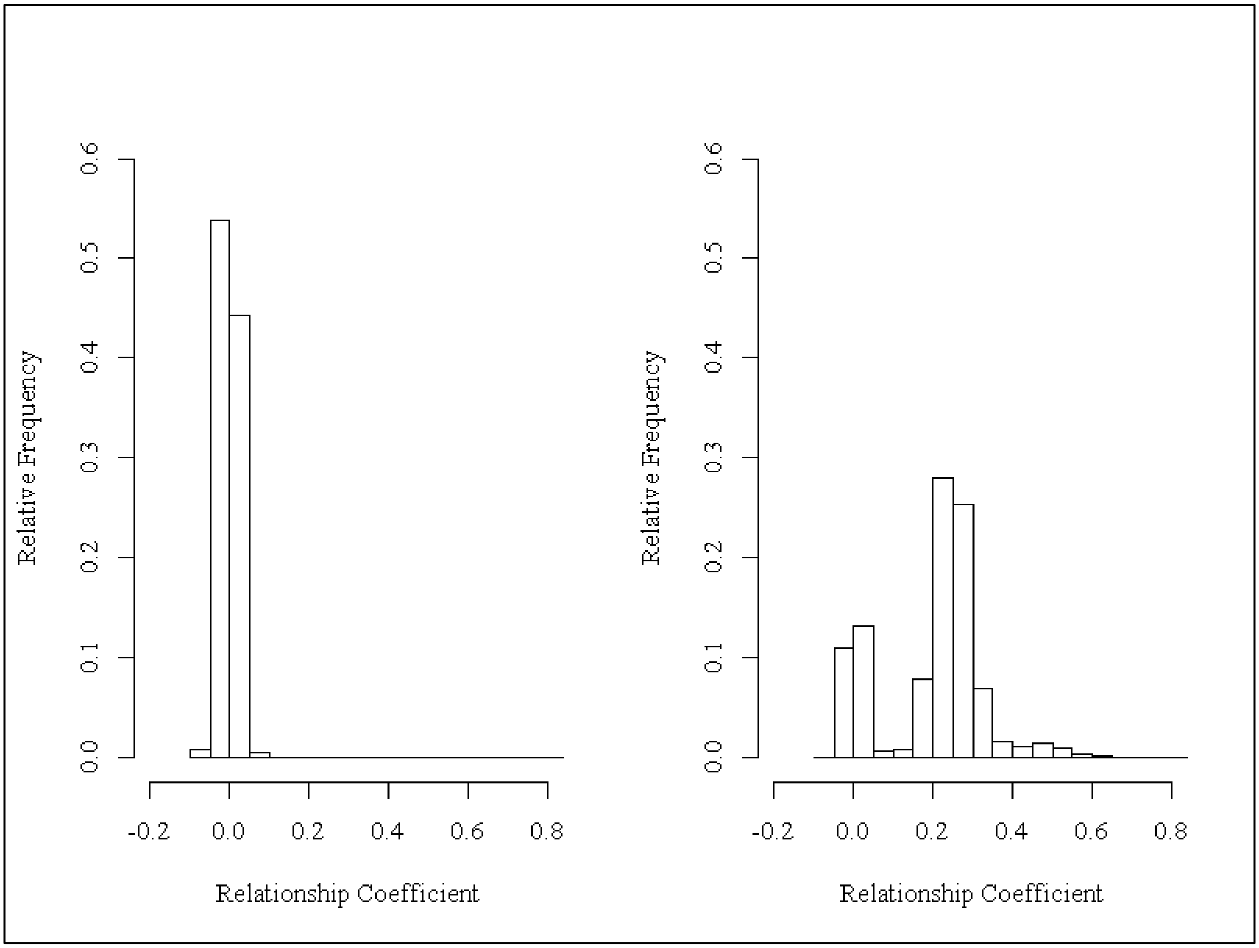


**Figure 3.1 Representative histograms of the genomic pairwise relationship coefficients**

Note: Relationship among (Left panel) and within (Right panel) members of the 214 white spruce open-pollinated families showing relationships clustering around the expected 0.25 with deviations from 0.25 as indicative of imperfect half-sib family (Right panel) and clustering around 0.00 as indicative of no relationship (Left panel).

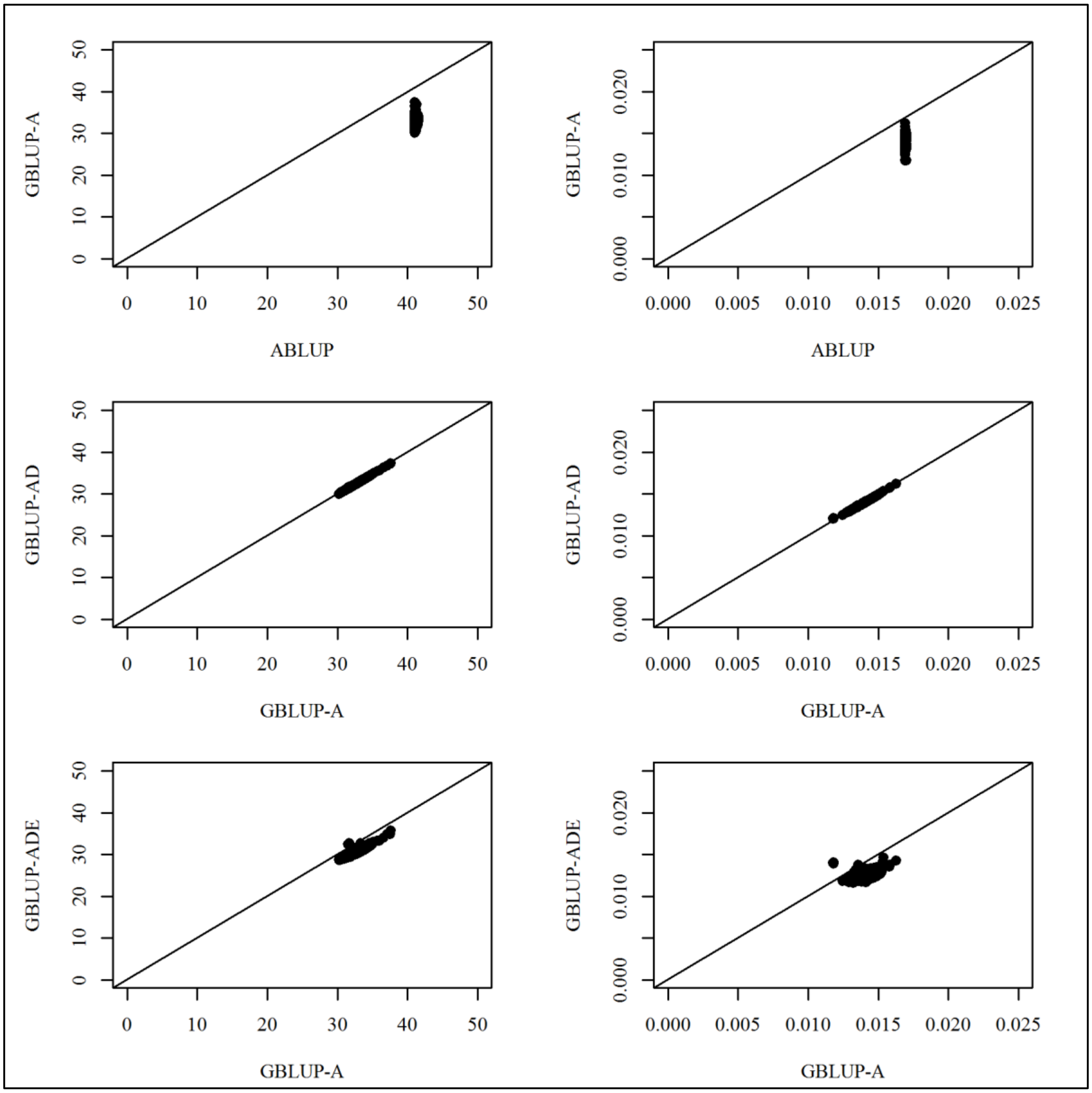


**Figure 3.2 Standard error of the predictions (SEP) of breeding values (BV)**

Note: BV from the ABLUP (X-axis) against that from the GBLUP-A (y-axis) for height (left panel) and wood density (right panel) and that from the GBLUP-A against those from the GBLUP-AD and GBLUP-ADE.

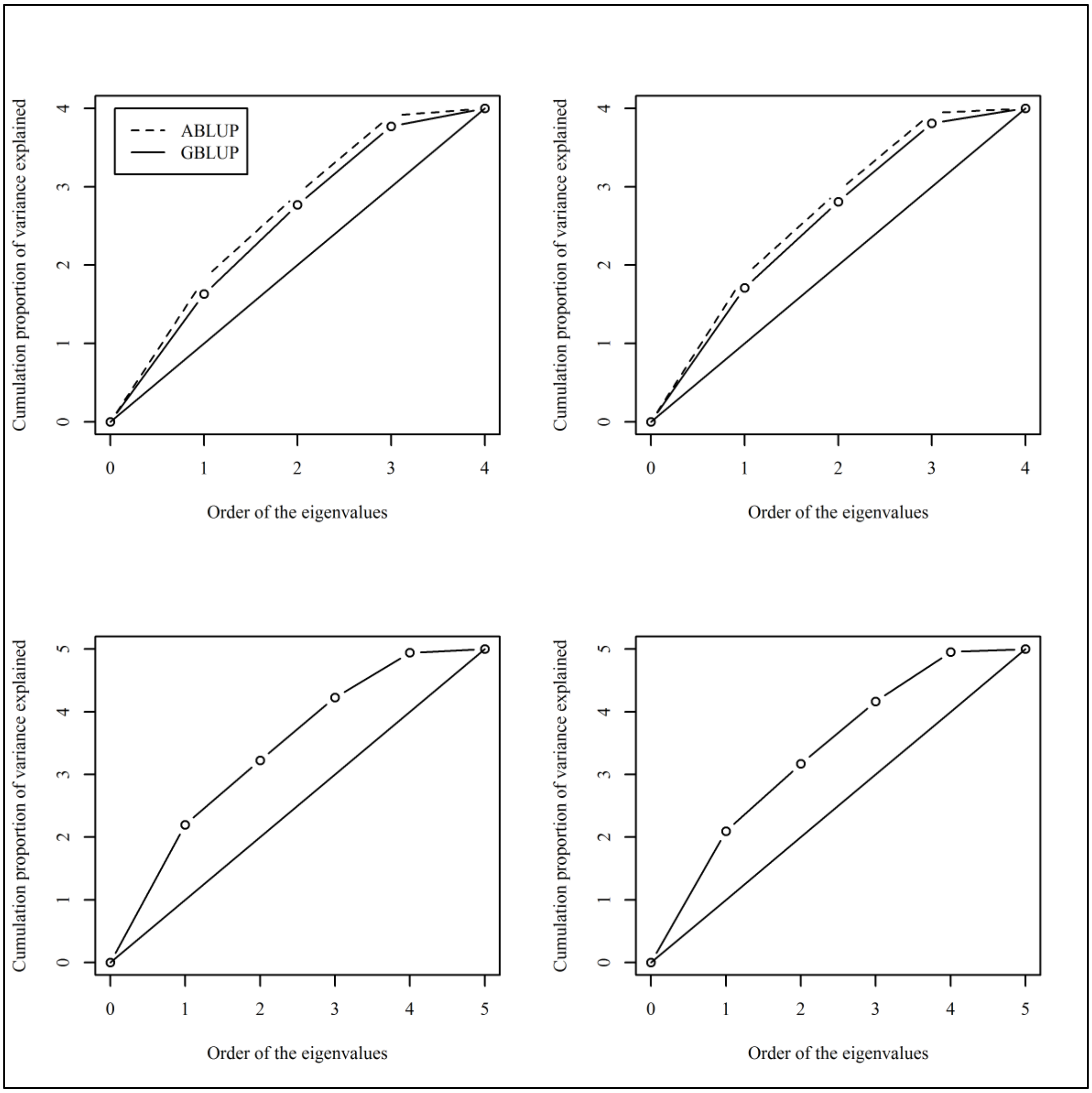


**Figure 3.3 Cumulative proportion of the variance explained by eigenvalues**

Note: Proportion for ABLUP vs. GBLUP-A (top panel) and GBLUP-ADE (bottom panel) for height (left) and wood density (right). Diagonal line represents an orthogonal correlation matrix.

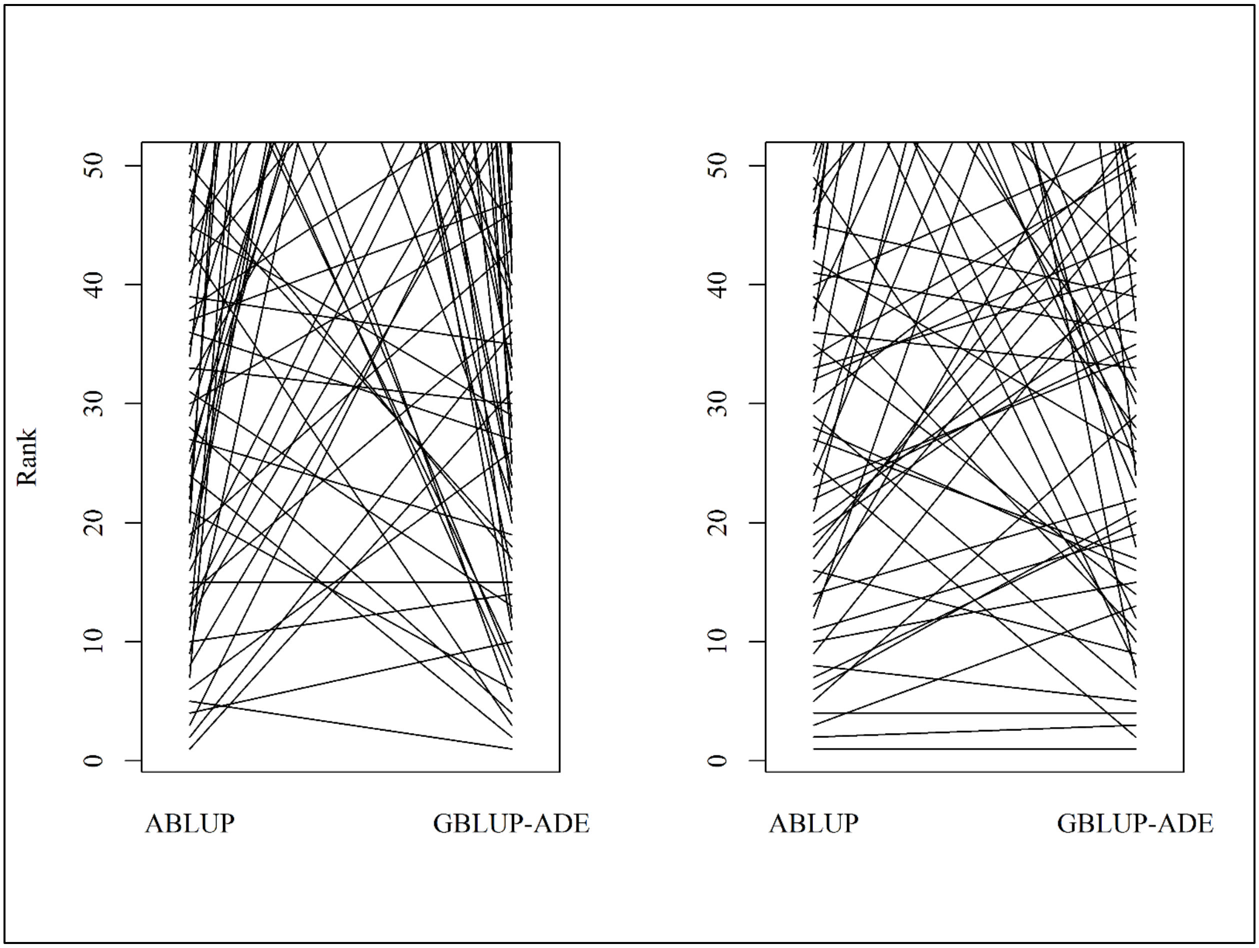


**Figure 3.4 Ranking plots for the top 50 performing white spruce individuals**

Note: Ranking plot for height (left) and wood density (right), respectively, comparing results of ABLUP versus GBLUP-ADE assessments (note; the number of highly ranked individuals in the ABLUP that dropped from the top 50 in the GBLUP-ADE).

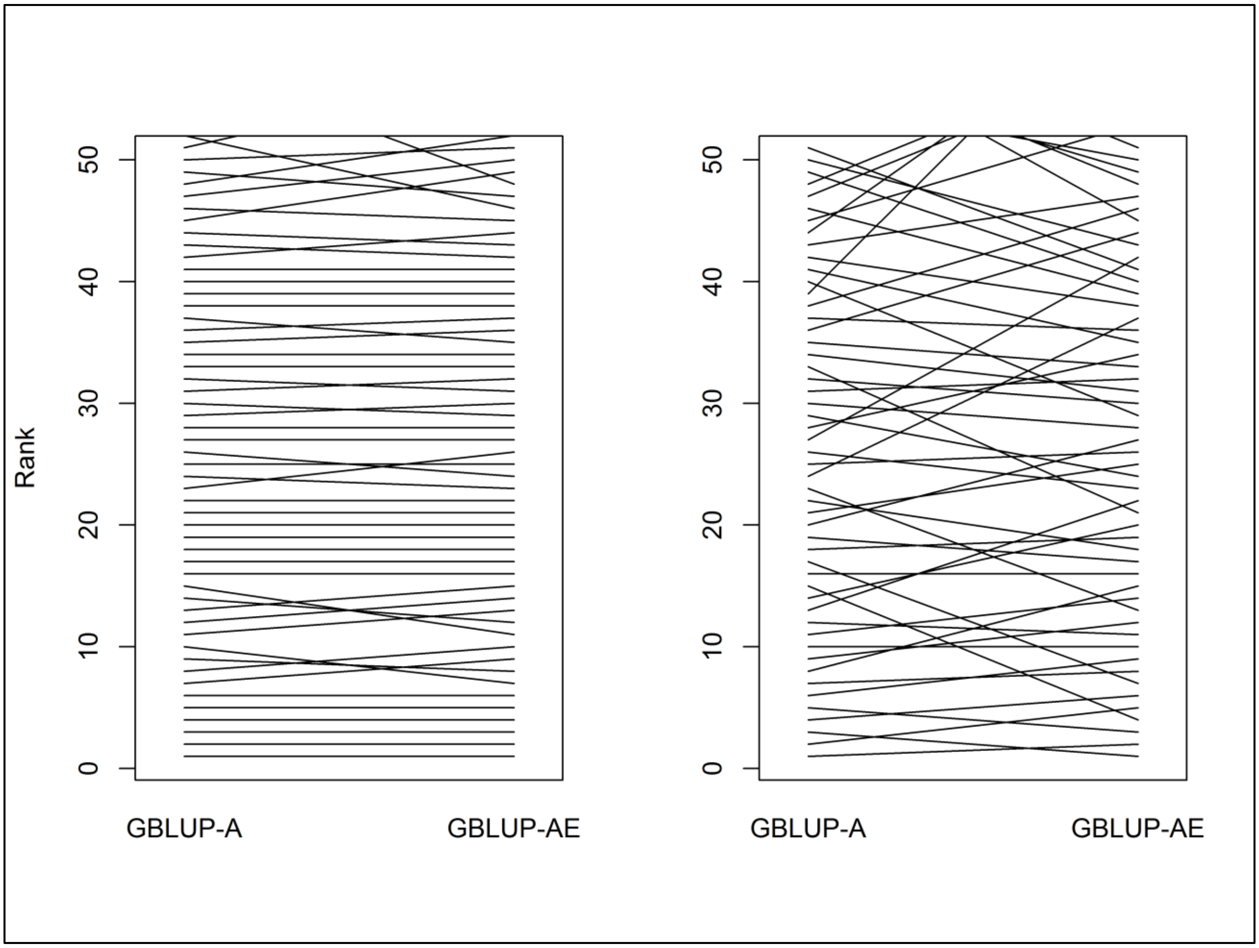


**Figure 3.5 Ranking plots for the top 50 performing white spruce individuals for GBLUP-A versus GBLUP-AE**

Note: Ranking plot for height (left) and wood density (right), revealing the minor change in rank

# Chapter 4: Extension of the OP testing genetic analysis to multi-site using Interior spruce populations from British Columbia

## 4.1 Introduction

Traditional quantitative genetics analyses are mainly pedigree-dependent utilizing the genealogical relationships among individuals for genetic parameters estimation (i.e., the average numerator relationship matrix (***A***-matrix; (Wright 1922)). These methods were effective as evidenced by the gains attained for a substantial number of plant and animal genetic improvement programs (Allard 1999; Lush 2013). This paradigm is changing with the availability of dense Single Nucleotide Polymorphism (SNP) panels through whole genome sequencing (Bentley 2006) and various high-throughput Next Generation Sequencing (NGS) technologies (Schuster 2008). Dense sequencing data permit the accurate determination of the actual fraction of alleles shared between individuals, related or otherwise, and the estimation of their genomic pairwise realized relationship (Santure *et al.* 2010). The resulting genomic relationship between any pair of individuals is more accurate than their assumed pedigree-based as genomic data allow capturing their known contemporary pedigree and their unknown historic pedigree as well (Powell *et al.* 2010). When the genomic pairwise additive relationship is estimated for a group of individuals, the outcome is known as the realized additive genomic relationship matrix (***G***-matrix) which can be used as a substitute to the ***A***-matrix in quantitative genetics analyses (VanRaden 2008). Also, SNP data can be used to construct all types of relationship matrices such as dominance and epistasis genomic relationship matrices regardless of the mating design. The advantage of the genomic-based relationship over that of its counterpart, the traditional pedigree-based, is the ability of the former to adjust for the Mendelian sampling term, while the latter ignores the existing variation among single half- or full-sib family members and treats them equally, thus the ***G***-matrix provides more accurate genetic co-variances among relatives (Visscher *et al.* 2006; Hill and Weir 2011). Additionally, the genomic-based relationship is capable of detecting hidden relatedness among members of a specific family, as commonly observed for open-pollinated (OP) family testing (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994), thus providing unbiased additive genetic variances.

In a previous study (Gamal El-Dien *et al.* 2016), we utilized the additive, dominance and epistasis realized genomic relationship matrices to estimate height and wood density related genetic variances for 214 white spruce OP families growing on one site in Québec and successfully partitioned the genetic variance into its different components, namely, the additive, dominance, and epistatic components. We also demonstrated the presence of a systematic pedigree-based additive genetic variance bias that is commonly observed in OP family testing, as it assumes that family members are all half-sibs and that only additive genetic variance, albeit biased, can be estimated. In this respect, the use of the genomic relationship also permitted estimating both dominance and epistatic genetic variances from testing experiments that do not lend themselves to these genetic components estimation.

Using 25 Interior spruce OP families grown in a replicated block design over three sites in British Columbia (Canada), we compared the genetic variance estimates generated from the average numerator relationship ***A***-matrix (the expected relationships) and the realized genomic relationship ***G***-matrix (the observed relationships). The extended genetic models including dominance and epistasis relationship matrices were added in comparison to assess the genomic markers' utility in partitioning the genetic variance components into additive and non-additive effects. In this study, we also demonstrated the applicability of G-matrix for existing OP programs by extending our previous work to a more generalized multiple sites model that accounts for genotype x environment interaction.

## 4.2 Materials and methods

### 4.2.1 Interior spruce open-pollinated progeny test sites, phenotype data and genotyping

Interior spruce is a complex of white spruce (*Picea glauca* (Moench) Voss), Engelmann spruce (*Picea engelmannii* Parry), and their natural hybrids and, because of their similar growing habitat and silvicultural requirements, they are often collectively treated as one complex species (Sutton *et al.* 1991). A total of 1,126 38-year-old Interior spruce trees, representing 25 open-pollinated (OP) families, growing on three progeny test sites in the Interior of British Columbia, Canada, were phenotyped for total tree height (HT) and wood density (WD). The field trials were established by the British Columbia Ministry of Forests, Lands and Natural Resource and are located in Aleza Lake (Lat. 54° 03’ 15.7” N, Long. 122° 06’ 35.4” W, Elev. 700 mas), Prince George Tree Improvement Station (Lat. 53° 46’ 17.9” N, Long. 122° 43’ 07.6”W, Elev. 610 mas), and Quesnel (Lat. 52° 59’ 27.2” N, Long. 122° 12’ 30.6” W, Elev. 915 mas) and planted in a complete randomized block design with multiple tree-row-plots within each block (see Kiss and Yanchuk (Kiss and Yanchuk 1991) for details). The sampled trees/sites are part of a larger test with 197 OP families with an average family size of 374 trees. From each site, four blocks were sampled and HT (in meters) was measured using an ultrasonic clinometer Vertex$^{TM}$ III (Haglöf$^{®}$, Sweden); WD (g·cm-3) was determined from bark-to-bark wood cores using X-ray scanning (QTRS-01X Tree Ring Scanner, Quintek Measurement Systems Inc., USA); the cores were extracted from each tree at breast height in the north-south direction by 5-mm increment borers.

Genotyping-by-sequencing (GBS) (Elshire *et al.* 2011) was the genotyping platform used. For complete details related to DNA extraction, specific sequencing protocol and SNP detection pipeline see Chen et al. (Chen *et al.* 2013). The SNP data used for estimating the realized genomic relationship matrix were those published previously (Gamal El-Dien *et al.* 2015; Ratcliffe *et al.* 2015); in brief, SNP filtering consisted of constraining individual "missingness" to the best 1,000 of the 1,126 genotyped individuals, resulting in an average of 40 genotyped individuals (range was 32 to 45) across the 25 families. Subsequently, SNPs with no more than 30%missing data were retained. Missing information was imputed using an expectation maximizing (EM) algorithm (Dempster *et al.* 1977a), resulting in a total of 30K SNP markers to infer the genetic relationships.

### 4.2.2 Relationship matrices and genetic models

The additive relationship matrix was estimated as follows:

$$\boldsymbol{G_{add}} = \frac{\boldsymbol{ZZ\prime}}{2\sum p_i(1-p_i)} \qquad [1]$$

where $\boldsymbol{Z}$ is the rescaled genotype matrix following $\boldsymbol{M}$ - $\boldsymbol{P}$, $\boldsymbol{M}$ is the genotype matrix containing genotypes coded as 0, 1, and 2 according to the number of alternative alleles and $\boldsymbol{P}$ is the vector of twice the allelic frequency $p$ (VanRaden 2008). The dominance genetic variance was fitted by including the marker-based dominance relationship matrix following:

$$\boldsymbol{G_{dom}} = \frac{\boldsymbol{WW\prime}}{(2pq)^2} \qquad [2]$$

where $\boldsymbol{W}$ is the matrix containing -2$q^2$ for the alternative homozygote, 2pq for the heterozygote, and -2$p^2$ for the reference allele homozygote (Vitezica *et al.* 2013). Similarly, epistatic variance was fitted by including several relationship matrices capturing first order additive x additive, dominance x dominance, and additive x dominance interaction. The relationship matrices were constructed as the Hadamard product of the relationship matrices defined above: $\boldsymbol{G_{add}\#G_{add}}$, $\boldsymbol{G_{dom}\#G_{dom}}$ and $\boldsymbol{G_{add}\#G_{dom}}$ (Su *et al.* 2012; Muñoz *et al.* 2014).

The variance components from the pedigree based analysis (ABLUP) were obtained by solving the mixed models following:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \ \boldsymbol{Z_1 a} + \boldsymbol{Z_2 axe} + \boldsymbol{Z_3 r(s)} + \boldsymbol{e} \qquad [3]$$

where $\boldsymbol{y}$ is the vector of measurements, $\boldsymbol{\beta}$ is the vector of fixed effects (overall mean and site), $\boldsymbol{a}$ is the vector of random additive genetic effects following $\boldsymbol{a}$ ~ N(0, $\boldsymbol{A}\sigma_a^2$), where $\boldsymbol{A}$ is the average numerator relationship matrix and $\sigma_a^2$ is the additive genetic variance, $\boldsymbol{axe}$ is the vector of random additive x environment (sites) interaction effects following $\boldsymbol{axe}$ ~ N(0, $\boldsymbol{I}\sigma_{axe}^2$), where $\boldsymbol{I}$ is the identity matrix and $\sigma_{axe}^2$ is the additive x environment interaction variance, ***r*(*s*)** is the vector of random replication nested within the site effect following ***r*(*s*)** ~ N(0, $\boldsymbol{I}\sigma_{r(s)}^2$), where $\sigma_{r(s)}^2$ is the replication nested within the site variance, and $\boldsymbol{e}$ represents a vector of the random residual effects following $\boldsymbol{e}$ ~ N(0, $\boldsymbol{I}\sigma_e^2$) where $\sigma_e^2$ is the residual error variance, ***X*** and ***Zs*** are incidence matrices relating fixed and random effects to measurements in the vector ***y***. The variance components from the analysis using the marker-based additive relationship matrix (GBLUP-A) was obtained from the model described above but the average numerator relationship matrix $\boldsymbol{A}$ is substituted by the marker-based relationship matrix $\boldsymbol{G_{add}}$. The extended model for the dominance term (GBLUP-AD) is performed as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \ \boldsymbol{Z_1 a} + \boldsymbol{Z_4 d} + \boldsymbol{Z_2 axe} + \boldsymbol{Z_5 dxe} + \boldsymbol{Z_3 r(s)} + \boldsymbol{e} \qquad [4]$$

where $\boldsymbol{d}$ is the vector of the random dominance effect following $\boldsymbol{d}$ ~ N(0, $\boldsymbol{G_{dom}}\sigma_d^2$) with $\sigma_d^2$ the dominance variance and $\boldsymbol{dxe}$ the random vector of dominance x environment interaction effects following $\boldsymbol{dxe}$ ~ N(0, $\boldsymbol{I}\sigma_{dxe}^2$) where $\sigma_{dxe}^2$ is the dominance x environment interaction variance. Additional model extension for epistatic terms (GBLUP-ADE) is performed as follows:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_1 a} + \boldsymbol{Z_4 d} + \boldsymbol{Z_6 axa} + \boldsymbol{Z_7 dxd} + \boldsymbol{Z_8 axd} + \boldsymbol{Z_2 axe} + \boldsymbol{Z_5 dxe} + \boldsymbol{Z_3 r(s)} + \boldsymbol{e} \quad [5]$$

where ***axa*** is the vector of random additive x additive epistatic interaction effects following ***axa*** ~ N(0, $\boldsymbol{G}_{\boldsymbol{add\#add}}\sigma^2_{axa}$) where $\sigma^2_{axa}$ is the additive x additive epistatic interaction variance, ***dxd*** is the vector of random dominance x dominance epistatic interaction effects following ***dxd*** ~ N(0, $\boldsymbol{G}_{\boldsymbol{dom\#dom}}\sigma^2_{dxd}$) where $\sigma^2_{dxd}$ is dominance x dominance epistatic interaction variance, and ***axd*** is the vector of random additive x dominance epistatic interaction effects following ***axd*** ~ N(0, $\boldsymbol{G}_{\boldsymbol{add\#dom}}\sigma^2_{axd}$) where $\sigma^2_{axd}$ is the additive x dominance epistatic interaction variance.

The narrow-sense heritability estimate was estimated as $\hat{h}^2 = \hat{\sigma}^2_a/\hat{\sigma}^2_p$, where $\hat{\sigma}^2_a$ represents the estimate of the additive variance and $\hat{\sigma}^2_p$ equals $\hat{\sigma}^2_e$ in addition to the other variance components estimates such as additive, dominance, additive x additive, additive x dominance, dominance x dominance, additive x environment, dominance x environment interactions following that of the ABLUP and GBLUPs (termed GBLUP-A, GBLUP-AD, and GBLUP-ADE, respectively) models, respectively (Table 4.1). The estimations of the variance components and their stand errors were performed using a standalone version of ASReml$^{TM}$ v. 3.0 software (Gilmour *et al.* 2009), while the marker-based relationship matrices construction and models' cross-validations were done in R (R Core Team 2014). Additionally, the breeding values (BVs) rank order for the top 50 performing individuals was compared between ABLUP and GBLUP-AD and GBLUP-ADE for HT and WD, respectively.

### 4.2.3 Models comparison and cross-validation

Finally, for comparing the relative quality of the goodness-of-fit for said models, the variance explained by each model (R$^2$) was used (Nakagawa and Schielzeth 2013) that is the summary statistics for the goodness-of-fit of the linear mixed-effects models (LMM) and the fitted line plot (graph of predicted ŷ versus y values), while the standard error (SE) of the predictions (SEPs) of the breeding values (BVs) was used to assess the precision of the BVs.

The predictability (i.e. the Pearson product-moment correlation between phenotypes and the predicted BVs from cross-validation (PBV-CV)) and the prediction accuracy (i.e. the Pearson product-moment correlation between the estimated BVs from full data (EBV-all) and predicted BVs from cross-validation (PBV-CV)) for the four models were estimated using 10-folds CV and five replicates. Two folding scenarios were used, i.e. random and family folding, as in the latter, and to test the effect of removing relatedness between the two sets during CV, the validation set represented families that were absent in the training set. In each replicate, the data was divided into 10-folds according to the used folding scenario, 9-folds was assigned as the training population, while the last fold was used as the validation population to estimate PBV-CV. The five replicates were used to estimate the SE of the correlation. Model pairwise-prediction accuracy was also estimated between the four models in order to evaluate the ability of predicting each other. In this case, accuracy was estimated as the Pearson product-moment correlation between EBV-all of one model and PBV-CV of the other model (see above).

## 4.3 Results

### 4.3.1 Genetic variance components and heritability estimates

Replications of within-site variance components were consistent across the four models and accounted in each case for a relatively small variance component for both height (HT: 1.29-1.66%) and wood density (WD: 6.99-8.86%) (Table 4.1). The main difference between the ABLUP and GBLUP-A was the substantial decrease in the additive and additive x environment interaction (Table 4.1). The additive genetic variances obtained from GBLUP-A were 81 and 66% of the ABLUP additive genetic variance for HT and WD estimates, respectively (Table 4.1). This decrease in the additive genetic variance apportionment subsequently decreased the additive x environment interaction (32.01% vs. 20.99% and 16.72% vs. 13.03%, for height HT and WD, respectively) and increased the residual term (36.61% vs. 53.38% and 40.08% vs. 55.19%, for height HT and WD, respectively), resulting in reduced heritability estimates (0.30 vs. 0.25 and 0.39 vs. 0.26, for height HT and WD, respectively) (Table 4.1).

The GBLUP-AD analysis produced surprising results for HT as the dominance variance component was significant and accounted for 28.74% while it was non-significant for WD (4.14%) (Table 4.1). It is noteworthy to mention that the dominance variance estimates did affect neither the additive genetic variances nor the heritability estimates and that their appearance is mostly reflected in the reduction of the residual term estimates (i.e., the dominance variances were confounded in the residual terms) (Table 4.1).

The GBLUP-ADE produced exactly the same results as GBLUP-AD for HT, indicating absence of first order interactions while WD showed a significant additive x additive interaction accounting for 23.34% of the total variance (Table 4.1). The appearance of additive x additive variance for WD reduced the residual term (50.36 vs. 28.80%) as well as the additive term (23.97 vs. 20.97%), further reducing the WD heritability estimate (from 0.26 to 0.23) (Table 4.1). The additive x additive estimate was confounded mainly within the residuals and to a lesser extent within the additive variances.

### 4.3.2 Models comparison and cross-validation

We used two methods for model comparison, namely, the variance explained by the model ($R^2$) and the fitted line plots (represented by the graph of predicted values ŷ versus observed values y). Moving from ABLUP to the GBLUP-A was characterized by the lack of improvement for the two model comparison methods (Table 4.1 and Figure 4.1). However, this result is not surprising, as the ABLUP models were inaccurate due to the observed inflated additive genetic variance which in turn makes the total variance explained by the model inflated too. The $R^2$ method showed reduced values between ABLUP and GBLUP-A (63.39 vs. 46.62 and 56.92 vs. 44.81, for HT and WD, respectively) (Table 4.1). Comparing GBLUP-A with GBLUP-AD, generally showed improvement, which was more pronounced for HT (80.04 vs. 46.62) than for WD (49.64 vs. 44.81) due to the observed significant dominance variance (Table 4.1). The $R^2$ values for HT did not change between GBLUP-AD and GBLUP-ADE due to the lack of epistatic genetic variances, indicating that GBLUP-AD is the best (and sufficient) model for HT (Table 4.1). WD, on the other hand, showed substantial $R^2$ value improvement (49.64 vs. 71.20), reflecting the presence of significant additive x additive genetic variances and indicating that GBLUP-ADE is the best model for WD, these differences potentially reflecting the two traits' different genetic architecture (Table 4.1). These results collectively indicate that the genomic-based models are superior to the pedigree-based model.

The fitted line plot comparisons (shown in Figure 4.1) reflected the conclusions based on $R^2$ while the differences between the ABLUP and GBLUP-A models for HT and WD showed worse fitting, supporting the notion that the ABLUP models harbor inflated additive genetic variance. Similarly, the plots show that the GBLUP-AD and GBLUP-ADE are the best fit for HT and WD, respectively, and this is illustrated by the points' distribution and their closeness to the 45° reference lines (Figure 4.1).

Comparing breeding values' (BVs) precision, using the standard errors for predictions (SEP), between the ABLUP and GBLUP-A models, indicated that the SEPs of HT and WD were universally smaller for GBLUP-A as compared to ABLUP (as all SEP values were below the 45° reference lines (Figure 4.2; GBLUP-A#ABLUP)), clearly confirming the superiority of the

GBLUP-A model. For this reason, we used the GBLUP-A model as a reference for the extended models' comparisons. GBLUP-AD and GBLUP-ADE were proven to be the best models for HT and WD, respectively (Figure 4.2; GBLUP-AD#GBLUP-A (left panel) and GBLUP-ADE#GBLUP-A (right panel) for HT and WD, respectively).

Random folding cross-validation prediction accuracy was the lowest for ABLUP for both traits (0.620 and 0.624 for HT and WD, respectively) compared to the GBLUPs models which gave a range of 0.676 (GBLUP-AD) to 0.685 (GBLUP-A) and 0.689 (GBLUP-ADE) to 0.692 (GBLUP-AD) for HT and WD, respectively (Table 4.2; diagonal values). On the other hand, the pairwise prediction accuracy between ABLUP and GBLUPs (HT: 0.562 to 0.615; WD: 0.540 to 0.643) was lower than between the GBLUPs models themselves (HT: 0.676 to 0.683; WD: 0.685 to 0.693) (Table 4.2; off-diagonal values). When GBLUPs models were used to predict ABLUP, the prediction accuracies ranged from 0.610 to 0.615 (HT) and from 0.640 to 0.643 (WD), while when the ABLUP was used to predict GBLUPs, the range was significantly lowered (from 0.562 to 0.573 and from 0.540 to 0.547, for HT and WD, respectively) (Table 4.2). Regarding predictability, expressed as the correlation between the predicted BVs from cross-validation (PBV-CV) and the phenotype, GBLUP-A and ABLUP showed the highest values (0.228 and 0.233 for HT and WD, respectively) (Table 4.2; random folding, first column).

For family folding, the predictability and prediction accuracy were generally much lower as compared to random folding (Table 4.2). The use of the ABLUP model for individual breeding value prediction for members of new families is not applicable as the relatedness is equal to zero and the predicted value will be simply the overall mean of the model (Table 4.2).

## 4.4 Discussion

Current tree improvement programs depend mainly on phenotypic selection and the pedigree-based average numerator relationship (***A***-matrix) for estimating genetic parameters and variance decomposition. The utilized mating design determines mainly which genetic component can be generated and, in some cases, additional efforts such as combining full-sib families with replicated clonal trials is attempted to estimate dominance and epistatic genetic variances (Foster and Shaw 1988; Bradshaw and Foster 1992). OP family testing represents the most efficient method for screening large numbers of individuals in terms of low cost and less time; however, it suffers from inflated additive variance estimates due to the impossibility of meeting the commonly assumed half-sib structure (Namkoong 1966; Squillace 1974; Askew and El-Kassaby 1994). The availability and affordability of DNA high-throughput fingerprinting methods, such as Genotyping-by-sequencing (GBS), made it possible to use single nucleotide polymorphisms (SNPs) to estimate the realized relationship matrix (***G***-matrix) among individuals and substitute the ***A***-matrix in estimating genetic variance components particularly in forest trees population (Zapata-Valenzuela *et al.* 2013; Klápště *et al.* 2014; Muñoz *et al.* 2014; de Almeida Filho *et al.* 2016; Gamal El-Dien *et al.* 2016). These studies illustrated the superiority of the GBLUP and resulted in generating more precise genetic parameters, mainly due to the method's efficiency in separating the additive from non-additive (dominance and epistasis) genetic variances as well as accounting for the Mendelian sampling within families (VanRaden 2008; Hayes *et al.* 2009b). In our previous study conducted to parse out additive and non-additive genetic variances (Gamal El-Dien *et al.* 2016), we used data from a single and pure white spruce site and demonstrated the presence of non-significant dominance as well as significant epistatic genetic variances; however, the study might have produced biased , because G x E (genotype x environment interaction) component was not able to be assessed in a single site study. Here, we extended the model to include multiple sites to be able to account also for the G x E, using an Interior spruce OP testing population growing in British Columbia, Canada.

Predictably, the results from the present study produced different additive variance estimates across the tested models (ABLUP vs. GBLUPs). The three GBLUP models produced lower additive genetic variance than the ABLUP model, results concur with those reported for the single-site (Gamal El-Dien *et al.* 2016) and other forest tree studies (Muñoz *et al.* 2014; de Almeida Filho *et al.* 2016). The reduced additive genetic variance subsequently lowered the heritability estimates; however, this observed reduction in the present study was smaller than the one observed in the single-site study (Gamal El-Dien *et al.* 2016), highlighting the benefits of using the multi-site approach in producing realistic estimates (i.e., G x E inclusion). Notwithstanding the better $R^2$ and fitted line plot of the ABLUP model (Table 4.1; Figure 4.1) compared to GBLUP-A, the obtained precise genetic variance and breeding value (Figure 4.2) estimates from the GBLUP-A demonstrate the added value of the realized relationship-based models as their estimates are devoid of hidden relatedness inflating additive genetic variance and un-accounting the Mendelian term (VanRaden 2008; Hayes *et al.* 2009b; Gamal El-Dien *et al.* 2016).

The GBLUP-AD model produced surprising results with a significant dominance variance for HT relative to the additive variance with a higher $R^2$ value supporting better model fit (Table 4.1). This was also illustrated by the fitted line plot and the breeding values' SEPs graph (GBLUP-AD: Figure 4.1 and 4.2 left panels). This trend was not observed for WD as the dominance variance was not significant (based on SE) and only accounted for a small amount of the total genetic variance (Table 4.1 and Figure 4.1 and 4.2, right panels), supporting similar observations on Douglas-fir and white spruce (El-Kassaby and Park 1993; Gamal El-Dien *et al.* 2016). The significant dominance genetic variance for HT in Interior spruce mirrored that reported for loblolly pine (Muñoz *et al.* 2014; de Almeida Filho *et al.* 2016), but see our previous study (Gamal El-Dien *et al.* 2016). The observed significant dominance variance for HT is unexpected as the GBS fingerprinting is expected to under-represent heterozygosity estimates (Nielsen *et al.* 2011; Glaubitz *et al.* 2014). Indeed, the heterozygous individuals' under-representation is also detected in our study and we postulate that it is a by-product of the GBS' low coverage (Figure 4.3). Additionally, the ability to detect dominance variance is also dependent on the nature of the population and the type of markers used to construct the dominance (fraternity) relationship matrix. In a simulation study, García-Cortés et al. (García-Cortés *et al.* 2014) reported that the presence of multi-allelic markers is a prerequisite for the precise estimation of the dominance coefficients, a condition, which can potentially affect the ability to estimate the dominance variance component when using exclusively biallelic markers such as SNPs. It is interesting that the HT additive genetic variance and heritability estimates did not change between GBLUP-A and GBLUP-AD, which means that the additive variance was accurately estimated in the GBLUP-A model and was not confounded with the dominance effect for this trait. Probably this is the reason why the prediction accuracy of GBLUP-AD did not improve when compared with GBLUP-A (Table 4.2; diagonal).

The full model (GBLUP-ADE), which was extended to include first order interaction, gave exactly the same results as GBLUP-AD for HT indicating the absence of all kind of epistatic interactions and furthermore did not show any improvement in all goodness-of-fit measures and precision estimates (Table 4.1, Figures 4.1 and 4.2). Results were distinct from our previous study on *P. glauca* (Gamal El-Dien *et al.* 2016), where HT showed significant additive x additive interaction and non-significant dominance, while here, HT showed a significant dominance component which was extracted from the residual variance without any effect on the additive variance. The hybrid nature of Interior spruce (*P. glauca x P. engelmannii)* in British Columbia (De La Torre *et al.* 2014) can explain such distinct results as hybridization is reflected in higher diversity and higher heterozygosity which may make the dominance effect pronounced, and, additionally, dominance variance is also known to be population specific (Falconer *et al.* 1996). For the WD trait, GBLUP-ADE resulted in improved genetic variance partitioning and showed a relatively larger additive x additive component that was extracted mainly from the residual variance and to a some extent also from the additive variance component (Table 4.1), supporting the theoretical expectation that additive x additive variance is absorbed by additive and residual variances (Lynch, M., Walsh 1998; Jannink 2007; Mackay 2014). The superiority of GBLUP-ADE model for WD was supported by the $R^2$ estimates (Table 4.1), the fitted line plot and the SEP graph (Figures 4.1 and 4.2). A significant additive x additive component was also observed in our previous study (Gamal El-Dien *et al.* 2016) and previously in a full-sibs based population of loblolly pine (Muñoz *et al.* 2014). Thus, the WD results were consistent with our first study in pure white spruce (Gamal El-Dien *et al.* 2016); both studies showed non-significant dominance in addition to a significant additive x additive interaction that was extracted from the additive and residual variances. Also in both studies, substantial epistasis was detected in the genetic architecture of WD in spruce, and therefore, this result cannot be an artifact based on the population sampling and/or genotyping methodology as the two studies used completely different genotyping platforms.

The advantage of GBLUP models is their use of the realized genomic relationship among individuals regardless of their genealogy, while the ABLUP is mainly dependent on the pedigree-structure created by the used mating design. In addition to capturing the additive relatedness among individuals, the realized genomic relationship matrix is also capturing the linkage disequilibrium (LD) between the SNPs and quantitative trait loci (QTLs) and their co-segregation (Habier *et al.* 2007, 2010, 2013). These factors, collectively, affect the accuracy of the genomic estimated BVs (Habier *et al.* 2013). Most tree improvement breeding programs are in their early stage of tree domestication, thus they suffer from their shallow and simple pedigrees which making ABLUP's estimates questionable. Our cross-validation results support this notion as the GBLUP models produced higher prediction accuracy than the ABLUP (Table 4.2). Additionally, using the GBLUPs to predict ABLUP produced better results than the reverse scenario. This is expected as the GBLUP models are capable of capturing contemporary as well as historical relatedness (Table 4.2; see the off-diagonal estimates). The GBLUP models' superiority was already illustrated by Muñoz et al. (Muñoz *et al.* 2014). In their study on loblolly pine, Muñoz et al. successfully estimated the epistatic genetic variance from a full-sib mating design with clonally replicated trials using the GBLUP approach, while the ABLUP failed to estimate the epistatic genetic variance despite having full-sib families and clonal replications.

It is noteworthy to mention that extending the GBLUP models to include the dominance (GBLUP-AD in the case of HT) and dominance as well epistasis variances (GBLUP-ADE in the case of WD) resulted in improving the breeding values' estimates precision (Figure 4.2); however, these adjustments did not improve the prediction accuracy in the cross-validation compared to the GBLUP-A (Table 4.2; diagonal). Such scenario was also observed in a similar genetic variance decomposition study in the context of genomic selection for milk production in cattle (Ertl *et al.* 2014). This discrepancy can be explained by the fact that both dominance and epistatic genetic variances were mainly extracted from the residual term, thus resulting in no or minimal impact on the additive variance component.

The reported multi-site genetic variance decomposition along with the selection of the best model for each studied trait is expected to improve the genetic variance partition (see above) as well as the individuals' breeding values. We compared the ranking of the top 50 individuals for HT and WD between the pedigree-based ABLUP and the genomic-based GBLUP models (GBLUP-AD and GBLUP-ADE, for HT and WD, respectively) (Figure 4.4). Only 76% and 72%, respectively, of the top 50 individuals persisted between the ABLUP (HT) and the GBLUP-AD (HT), and between the ABLUP (WD) and GBLUP-ADE (WD), and both rankings indicated that some of the top ranked individuals from ABLUP have completely dropped out, warrantying potential genetic gain loss when applying only traditional ABLUP approach (Figure 4.4).

## 4.5 Summary

**Background:** The simplicity of open pollinated (OP) family testing made it an ideal method for screening and ranking a large number of parents and their offspring without the reliance on any mating design with structured-pedigree testing. OP testing assumes that the tested material are half-sib families, an assumption that is hardly fulfilled, thus additive variance estimates are often inflated and ranking and gain calculations are unreliable.

**Results:** Here, we extend the OP testing genetic variance decomposition from single- to multi-site using height and wood density measurements from 1,126 38-year-old Interior spruce (*Picea glauca* (Moench) Voss x *P. engelmannii* Parry ex Engelm.) trees, representing 25 OP families, growing on three sites in interior British Columbia, Canada. The advantage of multi-site testing is its ability to account for the genotype x environment effect. Individuals were fingerprinted for 30k SNPs using genotyping-by-sequencing technology, which in turn were used to estimate the genomic realized relationship among the studied individuals. The genomic-based model was extended to account for additive, dominance, epistatic genetic variances and their interactions with the environment.

**Conclusions:** Compared to the pedigree-based OP model, the genomic-based models produced more realistic narrow-sense heritability, breeding value estimates, and better prediction accuracy. Such higher precision resulting into different ranking for the tested individuals compared to the pedigree-based model. Moreover, the marker-based models were able to predict the breeding values for individuals from families that were not included in the developed models, which was not possible with the pedigree-based model. By extending the genomic-based models from single- to multi-site, the developed models are applicable to OP testing programs and offer a more reliable genetic variance decomposition and reliable individuals' ranking and gain estimates.

**Table 4.1 Estimates of genetic variance components (source of variation (S.O.V.) and their standard errors (SE) for height (HT) and wood density (WD) across the four genetic models**

| | | ABLUP[1] | | GBLUP-A | | GBLUP-AD | | GBLUP-ADE | |
|---|---|---|---|---|---|---|---|---|---|
| **Trait** | **S.O.V.[1]** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** | **Value (SE)** | **%** |
| **HT** | $\sigma^2_{R/S}$ | 0.04 (0.04) | 1.29 | 0.04 (0.04) | 1.35 | 0.04 (0.04) | 1.66 | 0.04 (0.04) | 1.66 |
| | $\sigma^2_{A}$ | 1.00 (0.49) | 30.09 | 0.72 (0.25) | 24.29 | 0.60 (0.26) | 24.30 | 0.60 (0.26) | 24.30 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 0.70 (0.39) | 28.74 | 0.70 (0.39) | 28.74 |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 0.00 (0.00) | 0.00 |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 0.00 (0.00) | 0.00 |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 0.00 (0.00) | 0.00 |
| | $\sigma^2_{AxE}$ | 1.06 (0.39) | 32.01 | 0.62 (0.27) | 20.99 | 0.57 (0.29) | 23.34 | 0.57 (0.29) | 23.34 |
| | $\sigma^2_{DxE}$ | N/A | | N/A | | 0.05 (0.07) | 2.00 | 0.05 (0.07) | 2.00 |
| | $\sigma^2_{E}$ | 1.21 (0.44) | 36.61 | 1.59 (0.29) | 53.38 | 0.49 (0.66) | 19.96 | 0.49 (0.66) | 19.96 |
| | $h^2$ | 0.30 (0.14) | | 0.25 (0.08) | | 0.25 (0.10) | | 0.25 (0.10) | |
| | $R^2$ | 63.39 | | 46.62 | | 80.04 | | 80.04 | |
| **WD** | $\sigma^2_{R/S}$ | 4.93E-05 (2.64E-05) | 6.99 | 5.04E-05 (2.71E-05) | 7.77 | 5.09E-05 (2.73E-05) | 8.08 | 5.14E-05 (2.76E-05) | 8.86 |
| | $\sigma^2_{A}$ | 2.55E-04 (1.04E-04) | 36.22 | 1.56E-04 (4.82E-05) | 24.01 | 1.51E-04 (5.03E-05) | 23.97 | 1.22E-04 (5.52E-05) | 20.97 |
| | $\sigma^2_{D}$ | N/A | | N/A | | 2.61E-05 (7.66E-05) | 4.14 | 2.47E-05 (7.56E-05) | 4.25 |
| | $\sigma^2_{AA}$ | N/A | | N/A | | N/A | | 1.36E-04 (1.22E-04) | 23.34 |
| | $\sigma^2_{DD}$ | N/A | | N/A | | N/A | | 2.68E-10 (2.92E-10) | 0.00 |
| | $\sigma^2_{AD}$ | N/A | | N/A | | N/A | | 1.69E-11 (1.85E-11) | 0.00 |
| | $\sigma^2_{AxE}$ | 1.18E-04 (5.97E-05) | 16.72 | 8.45E-05 (4.78E-05) | 13.03 | 5.30E-05 (5.25E-05) | 8.41 | 6.34E-05 (5.20E-05) | 10.91 |
| | $\sigma^2_{DxE}$ | N/A | | N/A | | 3.17E-05 (1.31E-05) | 5.04 | 1.67E-05 (1.83E-05) | 2.88 |
| | $\sigma^2_{E}$ | 2.83E-04 (8.78E-05) | 40.08 | 3.58E-04 (5.72E-05) | 55.19 | 3.17E-04 (1.31E-04) | 50.36 | 1.67E-04 (1.83E-04) | 28.80 |
| | $h^2$ | 0.39 (0.15) | | 0.26 (0.08) | | 0.26 (0.08) | | 0.23 (0.09) | |
| | $R^2$ | 59.92 | | 44.81 | | 49.64 | | 71.20 | |

[1] see Material and Methods for S.O.V. explanation

**Table 4.2 Height (HT) and wood density (WD) predictability (Pearson product-moment correlations between PBV-CV and phenotype) and prediction accuracy (Pearson product-moment correlation between PBV-CV and EBV-all) within and among models (ABLUP, GBLUP-A, GBLUP-AD, and GBLUP-ADE) using random and family folding (standard errors)**

| | | | **HT** | | | | | **WD** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **Phenotypes** | **EBV – all[2]** | | | | **Phenotypes** | **EBV - all** | | | |
| | | | | **ABLUP** | **GBLUP** | | | | **ABLUP** | **GBLUP** | | |
| | | | | | **-A** | **-AD** | **-ADE** | | | **-A** | **-AD** | **-ADE** |
| **PBV – CV[1]** | **Random Folding** | **ABLUP** | 0.222<br>(0.002) | **0.620**<br>**(0.001)** | 0.615<br>(0.001) | 0.610<br>(0.001) | 0.610<br>(0.001) | 0.233<br>(0.002) | **0.624**<br>**(0.002)** | 0.643<br>(0.001) | 0.643<br>(0.001) | 0.640<br>(0.001) |
| | | **GBLUP-A** | 0.228<br>(0.004) | 0.573<br>(0.004) | **0.685**<br>**(0.003)** | 0.683<br>(0.003) | 0.683<br>(0.003) | 0.212<br>(0.002) | 0.545<br>(0.002) | **0.690**<br>**(0.002)** | 0.690<br>(0.002) | 0.691<br>(0.002) |
| | | **GBLUP-AD** | 0.224<br>(0.003) | 0.562<br>(0.004) | 0.677<br>(0.003) | **0.676**<br>**(0.003)** | 0.676<br>(0.003) | 0.213<br>(0.003) | 0.547<br>(0.002) | 0.692<br>(0.002) | **0.692**<br>**(0.002)** | 0.693<br>(0.002) |
| | | **GBLUP-ADE** | 0.225<br>(0.002) | 0.570<br>(0.002) | 0.683<br>(0.002) | 0.680<br>(0.002) | **0.680**<br>**(0.002)** | 0.215<br>(0.004) | 0.540<br>(0.006) | 0.685<br>(0.005) | 0.686<br>(0.005) | **0.689**<br>**(0.005)** |
| | **Family Folding** | **ABLUP** | NA[3] | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| | | **GBLUP-A** | 0.060<br>(0.008) | 0.061<br>(0.012) | **0.222**<br>**(0.014 )** | 0.226<br>(0.014) | 0.226<br>(0.014) | -0.034<br>(0.007) | 0.007<br>(0.013) | **0.191**<br>**(0.016)** | 0.193<br>(0.016) | 0.210<br>(0.016) |
| | | **GBLUP-AD** | 0.063<br>(0.004) | 0.083<br>(0.012) | 0.245<br>(0.013) | **0.251**<br>**(0.013)** | 0.251<br>(0.012) | -0.021<br>(0.010) | 0.032<br>(0.017) | 0.210<br>(0.016) | **0.213**<br>**(0.016)** | 0.232<br>(0.016) |
| | | **GBLUP-ADE** | 0.064<br>(0.009) | 0.082<br>(0.016) | 0.241<br>(0.012) | 0.246<br>(0.013) | **0.246**<br>**(0.013)** | -0.031<br>(0.005) | 0.019<br>(0.008) | 0.201<br>(0.008) | 0.203<br>(0.008) | **0.221**<br>**(0.008)** |

[1]PBV-CV: Predicted breeding values using cross-validation; [2]EBV-all: Estimated breeding values using all data; [3]NA: predicted individual additive breeding value is equal to the overall mean of the model.

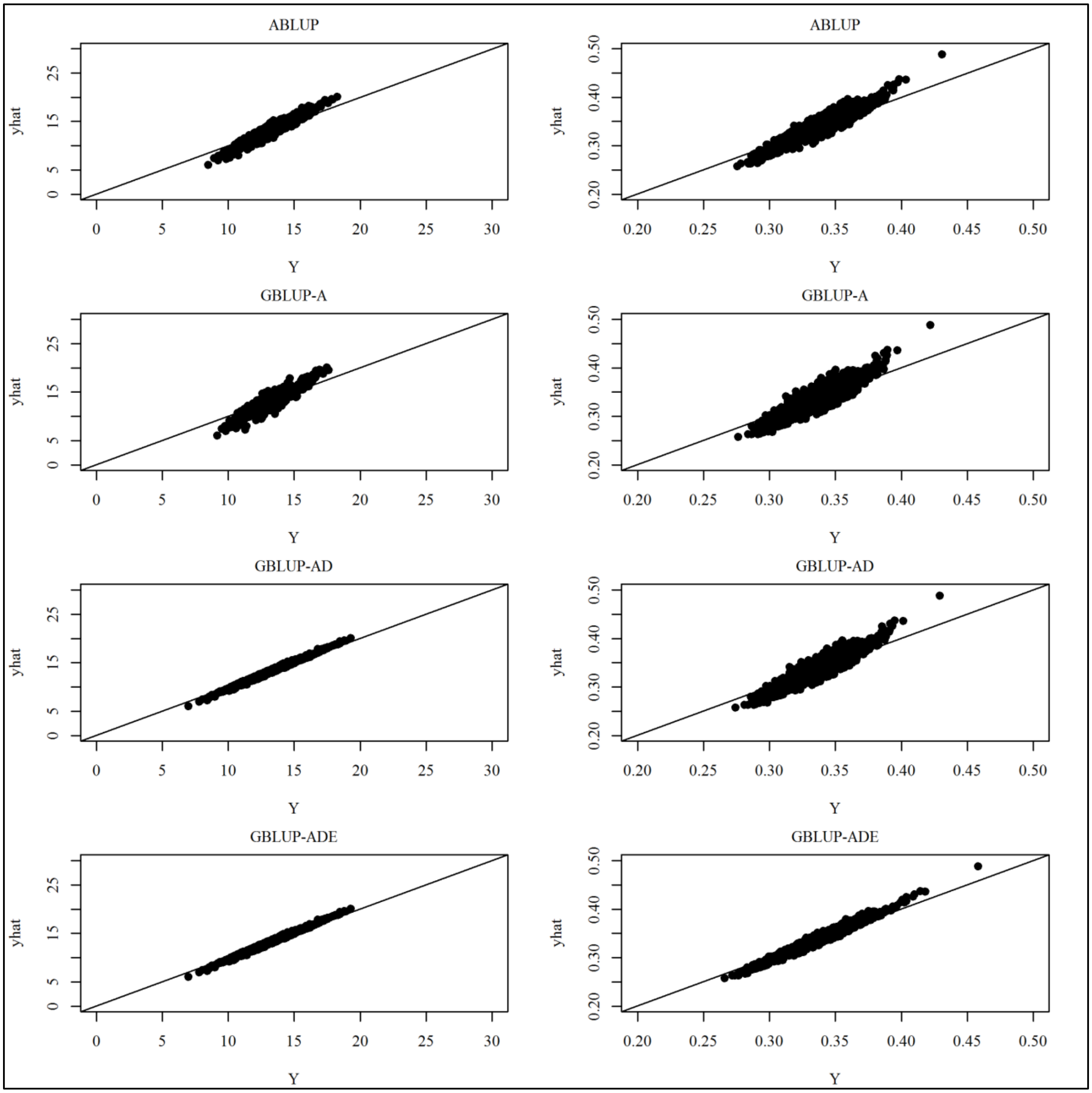


**Figure 4.1 Height (left) and wood density (right) fitted line plot (predicted ŷ vs observed y values) for the four models.**

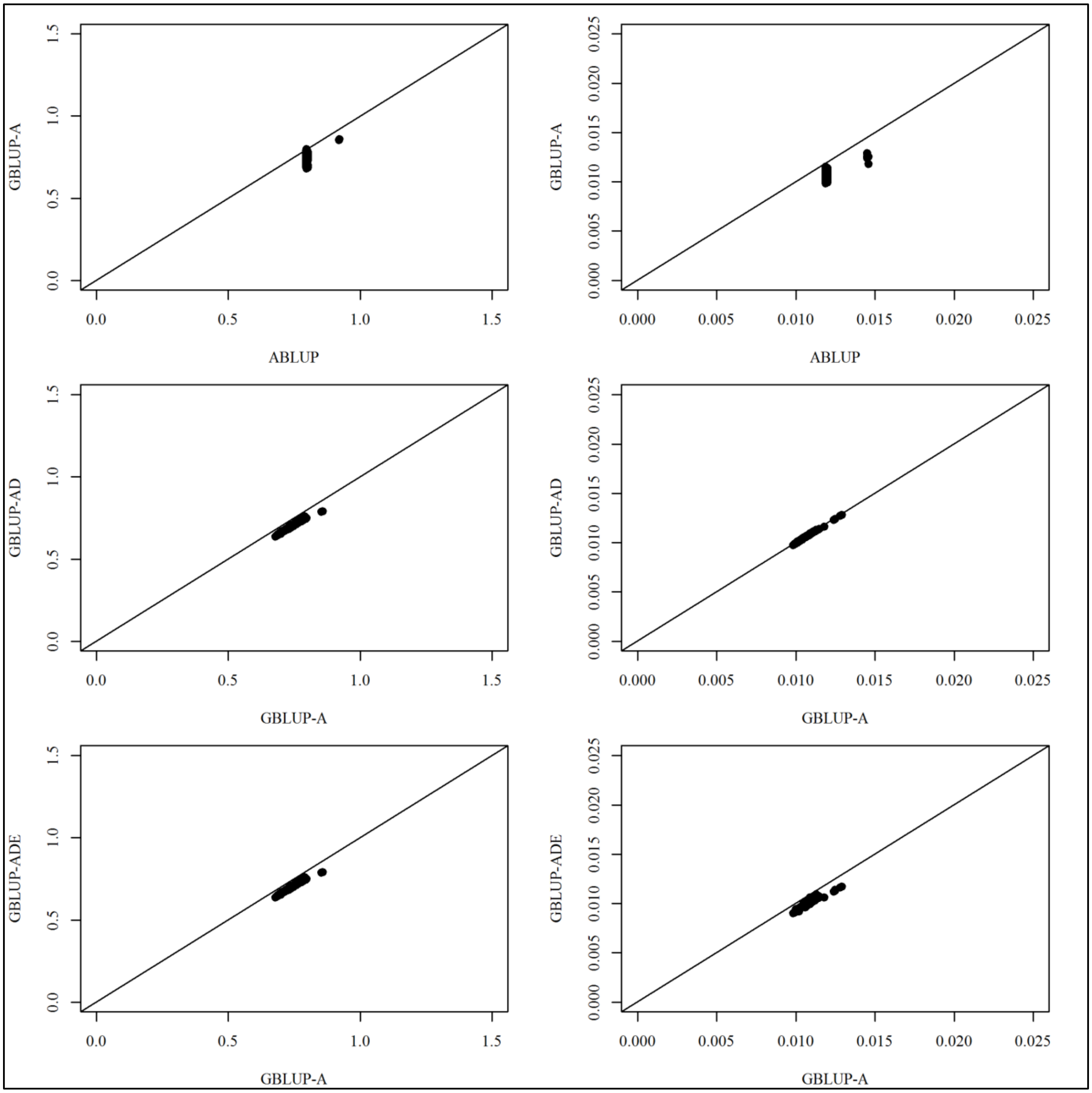


**Figure 4.2 Height (left) and wood density (right) standard error for the predictions of breeding values.**

Note: Comparisons for GBLUP-A vs ABLUP, GBLUP-AD vs GBLUP-A and GBLUP-ADE vs GBLUP-A.

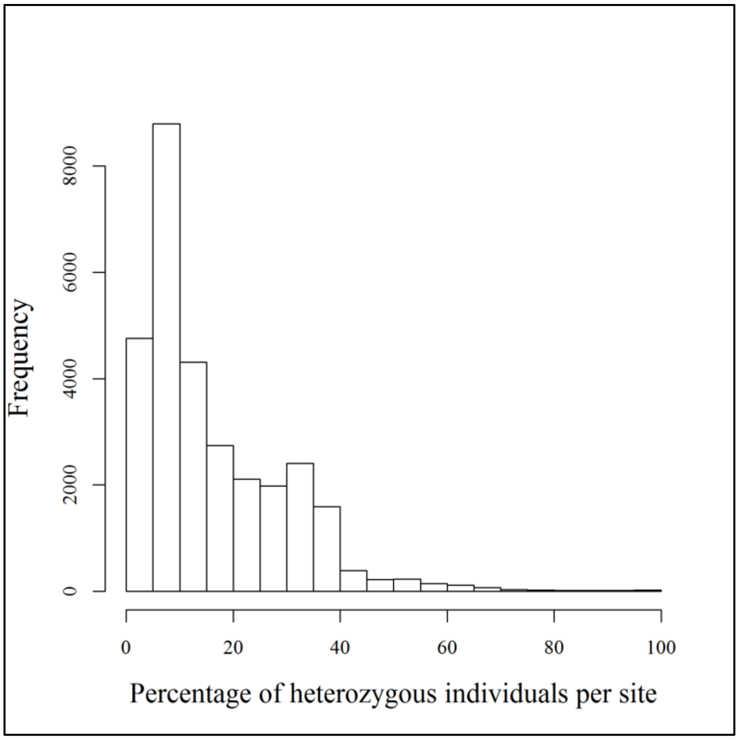


**Figure 4.3 Histogram showing the frequency of observed heterozygosity for GBS derived SNP sites.**

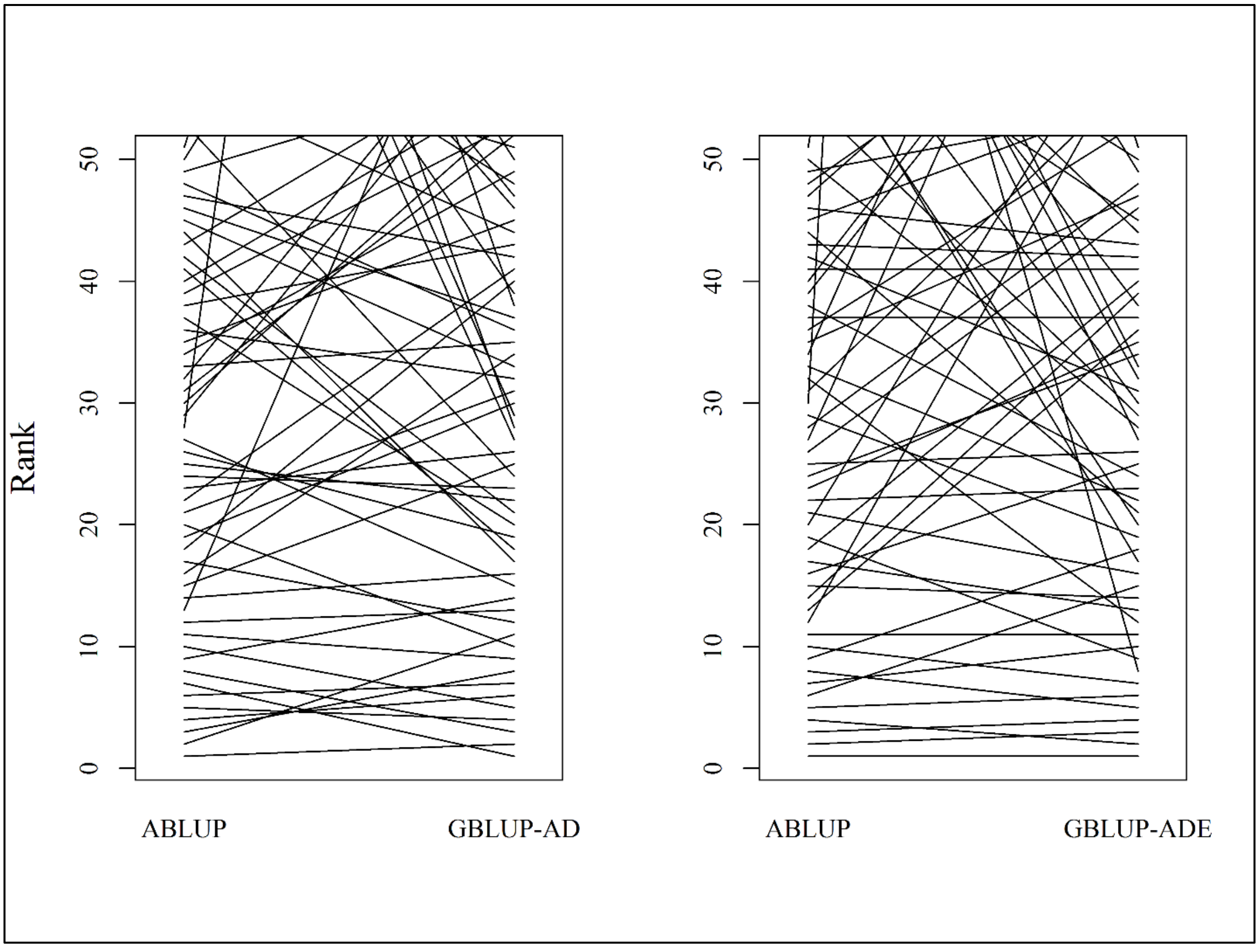


**Figure 4.4 Height (left) and wood density (right) breeding value ranking plots comparing ABLUP versus GBLUP-ADE.**

Note: this assessment is for forward selection of the top performing top 50 individuals.

# Chapter 5: Conclusion

## 5.1 Research novelties and potential applications

Genomic selection (GS) has created a paradigm shift in animal and crops breeding but it still in its infancy in forestry. GS can substantially reduce the length of breeding cycle and increase gain per unit time through early selection and greater selection intensity in tree improvement programs, particularly for traits of low heritability and late expression. Affordable next-generation sequencing technologies also have made it possible to genotype large numbers of trees at a reasonable cost.

Open-pollinated family testing is a formidable and economically viable option for screening a larger number of candidate parents without the development of "structured pedigree" that represents the backbone of most conventional tree breeding methods. The simplicity of the method made it an attractive first step before starting a full-blown tree improvement program. However, the commonly used assumption of treating open-pollinated offspring as half-sib family is by far the greatest drawback of this method as most genetic parameters (e.g., breeding values, trait heritabilities, and gain estimates) are upwardly biased and this was clearly demonstrated in many studies including the present one.

The introduction of genomic data (e.g., SNP markers) has provided the means to overcome this drawback and the genealogical relationship among open-pollinated family members is clearly and accurately ascertained. At present, many open-pollinated family testing trials have reached an advanced age but often abandoned, though they could provide badly needed information for late expressed traits that could not be obtained from younger conventional trials. The present study (Chapter 2) and that of Beaulieu *et al.*, 2014 provided the first examples for the application of GS for producing yield and wood quality attributes data with unprecedented accuracy in OP testing in tree improvement programs. This study is also one of the first studies to focus on the validation of GS model in space. The GBLUP analyses (Chapters 3 and 4) were the first attempt of such an analytical approach in OP families aiming at increasing the reliability of this kind of testing. To our knowledge, this study represents the first large-scale use of GBS (affordable genotyping technique) in a forest tree species known to a have complex genome and for which no reference sequence has been assembled yet.

## 5.2 Conclusions regarding goals and future research directions

In the present study (Chapter 2), the accuracy of GS model in predicting breeding values varied across the different studied validation scenarios with within multi-site being the highest and cross sites being the lowest (Figures 2.1 and 2.2). The high within multi-site GS prediction accuracies offer an opportunity to obtain reliable results for difficult traits such as wood density and yield and point towards considering "old" open-pollinated tests as a valuable source of information. The developed predictive models could be used for selecting elite genotypes with unprecedented selection intensity for their inclusion in future seed production populations, and this can be accomplished without the creation of a single cross. The results reported here suggest that GBS can be used as a genotyping platform for the application of GS in forestry. The use of proper missing marker data imputation algorithms is needed to overcome the commonly observed problem of missing data with GBS. Greater GS prediction accuracies were obtained for RR-BLUP as compared to GRR indicating that the studied traits follow the infinitesimal model of complex traits. Greater accuracies were obtained for multi-site GS model and point to the inherent lack of reliability for cross-site prediction. The use of principle component analysis as a multi-trait GS approach was proven to be effective in dealing with negatively correlated traits.

The GBLUP analyses (Chapters 3 and 4) also demonstrated the utility of the realized genomic relationship approach in providing a simple and extremely efficient method for generating more accurate genetic parameters from the simple OP testing compared to pedigree analysis (ABLUP) and overcame the drawback of this simple kind of testing. It is noteworthy to mention that the use of the realized genomic relationship also allowed the decomposition of dominance and epistasis genetic effects which requires full sibs mating designs coupled with cloning approaches. Persistence of the superiority of GBLUP over ABLUP after extending the GBLUP analysis from single-site (Chapter 3) to multi-site (Chapter 4), regardless the spruce populations and genotyping platforms (SNPs array and GBS, respectively) used, highlight the robustness of the GBLUP analyses and demonstrates the efficiency of the cost-effective GBS genotyping techniques.

Furthermore, the application of the ***A***-matrix, specifically, in the case of the well-known "shallow" pedigree present within most forest tree breeding and testing populations does not permit detecting hidden co-ancestry and inbreeding. Consequently, individuals' estimated breeding values are inflated by the overestimation of the additive genetic variance. GBLUP made it possible to ascertain, with great level of accuracy, the actual fraction of alleles shared between individuals, and the estimates of the individuals' pairwise realized relationship including potential inbreeding can be easily determined (Santure *et al.* 2010). Furthermore, the ***G***-matrix offers a unique opportunity for better genetic management as it provides information such as inbreeding and degree of relatedness among individuals.

In conclusion, the utility of genomic data in a simple, yet extremely efficient testing method, such as OP families calls for the re-evaluation of present-day conventional elaborate testing methods that are incapable of providing the genetic information produced in the present study. In general, the effectiveness of GS was clearly demonstrated as an alternative selection and evaluation method.

## 5.3 Strengths and limitations

In the present study, GBS successfully provided the information needed for genomic-based quantitative genetics analyses at reasonable cost. It is noteworthy to mention that this study was initiated before the release of Norway and white spruce genome sequences (Nystedt *et al.* 2013; Birol *et al.* 2013). However, as the assemblies of the two spruce genomes are not anchored and ordered along the chromosomes, there is little advantage over de novo SNP markers discovery used in this study.

In the single-site GBLUP analysis (Chapter 3), the collected data was limited to a single site of pure white spruce OP testing growing in Quebec, which resulted in biased results due to inability to account for genotype x environment interaction. At the same time, the more expensive better quality genotyping technique (SNPs array vs GBS), the bigger sample size (N =1,694 vs N ≈ 330/site) and the different population nature (pure white spruce vs hybrid white spruce) encouraged us to study the applicability of GBLUP in OP in this population and to compare it with our first studied population (Chapters 2 and 4). This drawback was circumvented by extending the model to multi-site analysis using a different population (Chapter 4). The sample size in chapter 4 (N= 1,000 for the three sites collectively) didn’t enable us to fit a single-site model (N ≈ 330/site).

## Bibliography

Allard, R. W., 1999 *Principles of plant breeding*. John Wiley & Sons.

de Almeida Filho, J. E., J. F. R. Guimaraes, F. F. e Silva, M. D. V de Resende, P. Munoz *et al.*, 2016 The contribution of dominance to phenotype prediction in a pine breeding and simulated population. Heredity (Edinb). 117: 33–41.

Annicchiarico, P., 2002 Genotype x environment interactions - Challenges and opportunities for plant breeding and cultivar recommendations. Fao Plant Prod. Prot. Pap. 174 174: 132.

Askew, G. R., and Y. A. El-Kassaby, 1994 Estimation of relationship coefficients among progeny derived from wind-pollinated orchard seeds. Theor. Appl. Genet. 88: 267–272.

Auty, D., and A. Achim, 2008 The relationship between standing tree acoustic assessment and timber quality in Scots pine and the practical implications for assessing timber quality from naturally regenerated stands. Forestry 81: 475–487.

Avendaño, S., J. a Woolliams, and B. Villanueva, 2004 Mendelian sampling terms as a selective advantage in optimum breeding schemes with restrictions on the rate of inbreeding. Genet. Res. 83: 55–64.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.*, 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3: e3376.

Bartholomé, J., J. Van Heerwaarden, F. Isik, C. Boury, M. Vidal *et al.*, 2016 Performance of genomic prediction within and across generations in maritime pine. BMC Genomics 17: 604.

Bastiaansen, J. W. M., A. Coster, M. P. L. Calus, J. a M. van Arendonk, and H. Bovenhuis, 2012 Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. Genet. Sel. Evol. 44: 3.

Beaulieu, J., T. Doerksen, B. Boyle, S. Clément, M. Deslauriers *et al.*, 2011 Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. Genetics 188: 197–214.

Beaulieu, J., T. Doerksen, S. Clement, J. MacKay, and J. Bousquet, 2014 Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity (Edinb). 113: 343–352.

Bentley, D. R., 2006 Whole-genome re-sequencing. Curr. Opin. Genet. Dev. 16: 545–552.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope *et al.*, 2013 Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29: 1492–1497.

Bouffier, L., A. Raffin, P. Rozenberg, C. Meredieu, and A. Kremer, 2008 What are the consequences of growth selection on wood density in the French maritime pine breeding programme? Tree Genet. Genomes 5: 11–25.

Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*, 2007 TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633–2635.

Bradshaw, J. H. D., and G. Foster, 1992 Marker-aided selection and propagation system in trees: advantages of cloning for studying quantitative inheritance. Can. J. For. Res. 22: 1044–1049.

Burdon, R. D., 1977 Genetic correlation as a concept for studying genotype-environment interaction in forest tree breeding. Silvae Genet. 26: 168–175.

Burdon, R. D., and C. J. A. Shelbourne, 1971 Breeding populations for recurrent selection conflicts and possible solutions. New Zeal. J. For. Sci. 1: 174–193.

Caballero, A., and M. Toro, 2000 Interrelations between effective population size and other pedigree tools for the management of conserved populations. Genet Res 75: 331–343.

Callaham, R. Z., 1964 Provenance research: investigation of genetic diversity associated with geography. Unasylva 18: 40–50.

Chaisurisri, K., and Y. A. El-Kassaby, 1994 Genetic diversity in a seed production population vs . natural populations of Sitka Spruce. Biodivers. Conserv. 3: 512–523.

Chen, C., S. E. Mitchell, R. J. Elshire, E. S. Buckler, and Y. A. El-Kassaby, 2013 Mining conifers' mega-genome using rapid and efficient multiplexed high-throughput genotyping-by-sequencing (GBS) SNP discovery platform. Tree Genet. Genomes 9: 1537–1544.

Crossa, J., Y. Beyene, S. Kassa, P. Pérez, J. M. Hickey *et al.*, 2013 Genomic prediction in maize breeding populations with genotyping-by-sequencing. G3 Genes| Genomes| Genet. Genomes| Genet. 3: 1903–1926.

Crow, J. F., 2008 Maintaining evolvability. J. Genet. 87: 349–353.

Crow, J. F., 2010 On epistasis: why it is unimportant in polygenic directional selection. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365: 1241–4.

Dekkers, J. C. M., 2004 Commercial application of marker- and gene-assisted selection in livestock : Strategies and lessons. J. Anim. Sci. 82: 313–328.

Dempster, A., N. Laird, and D. Rubin, 1977a Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc Ser B 39: 1–38.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977b Maximum Likelihood from Incomplete Data via the EM Algorithm. J. R. Stat. Soc. Ser. B(Methodological) 39: 1–38.

Doyle, J. J., and J. L. Doyle, 1990 Isolation of plant DNA from fresh tissue. Focus (Madison). 12: 13–15.

El-Kassaby, Y. A., 1995 Evaluation of the tree-improvement delivery system: factors affecting genetic potential. Tree Physiol. 15: 545–550.

El-Kassaby, Y. A., 1989 Variation in fruitfulness in a Douglas-fir seed orchard and its effect on crop-management decisions. Silvae Genet. 38: 3–4.

El-Kassaby, Y. A., and H. J. Barclay, 1992 Cost of reproduction in Douglas-fir. Can. J. Bot. 70: 1429–1432.

El-Kassaby, Y. A., E. P. Cappa, C. Liewlaksaneeyanawin, J. Klápště, and M. Lstibůrek, 2011 Breeding without breeding: is a complete pedigree necessary for efficient breeding? PLoS One 6: e25737.

El-Kassaby, Y. A., A. M. K. Fashler, and O. Sziklai, 1984 Reproductive phenology and its impact on genetically improved seed production in a Douglas-fir seed orchard. Silvae Genet. 33: 120–125.

El-Kassaby, Y. A., F. Isik, and R. W. Whetten, 2014 Modern advances in tree breeding, pp. 441–459 in *Challenges and Opportunities for the World's Forests in the 21st Century*, edited by T. Fenning. Springer Science+Business Media, Dordrecht.

El-Kassaby, Y. A., J. Klápště, and R. D. Guy, 2012 Breeding without breeding: Selection using the genomic best linear unbiased predictor method (GBLUP). New For. 43: 631–637.

El-Kassaby, Y. A., and M. Lstibůrek, 2009 Breeding without breeding. Genet. Res. (Camb). 91: 111–120.

El-Kassaby, Y. A., S. Mansfield, F. Isik, and M. Stoehr, 2011 In situ wood quality assessment in Douglas-fir. Tree Genet. Genomes 7: 553–561.

El-Kassaby, Y. A., and Y. S. Park, 1993 Genetic variation and correlation in growth, biomass traits, and vegetative phenology of a 3-year-old Douglas-fir common garden at different spacings. Silvae Genet 42: 289–297.

El-Kassaby, Y. A., and O. Sziklai, 1982 Genetic variation of allozyme and quantitative traits in a selected Douglas-fir population. For. Ecol. Manage. 4: 115–126.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.*, 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6: e19379.

Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250–255.

Ertl, J., A. Legarra, Z. G. Vitezica, L. Varona, C. Edel *et al.*, 2014 Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. Genet. Sel. Evol. 46: 40.

Falconer, D. S., T. F. C. Mackay, and R. Frankham, 1996 *Introduction to Quantitative Genetics*. Pearson Education Limited, Essex.

Fisher, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. R. Soc. Edinb. 52: 399–433.

Foster, G. S., and D. V Shaw, 1988 Using Clonal Replicates to Explore Genetic-Variation in a Perennial Plant-Species. Theor. Appl. Genet. 76: 788–794.

Frentiu, F. D., S. M. Clegg, J. Chittock, T. Burke, M. W. Blows *et al.*, 2008 Pedigree-free animal models: the relatedness matrix reloaded. Proc. Biol. Sci. / R. Soc. 275: 639–647.

Gamal El-Dien, O., B. Ratcliffe, J. Klapste, C. Chen, I. Porth *et al.*, 2015 Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC Genomics 16: 370.

Gamal El-Dien, O., B. Ratcliffe, J. Klapste, I. Porth, C. Chen *et al.*, 2016 Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from non-additive genetic effects. G3 Genes, Genomes, Genet. 6: 743–753.

García-Cortés, L. A., A. Legarra, and M. A. Toro, 2014 The coefficient of dominance is not (always) estimable with biallelic markers. J. Anim. Breed. Genet. 131: 97–104.

Gay, L., M. Siol, and J. Ronfort, 2013 Pedigree-Free Estimates of Heritability in the Wild: Promising Prospects for Selfing Populations. PLoS One 8: e66983.

Gianola, D., H. Okut, K. a Weigel, and G. J. Rosa, 2011 Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 12: 87.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, and R. Thompson, 2009 ASReml User Guide Release 3.0. VSN Int. Ltd.

Gilmour, A. R., R. Thompson, and B. R. Cullis, 1995 Average Information REML : An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models Author ( s ): Arthur R . Gilmour , Robin Thompson and Brian R . Cullis Published by : International Biometric Society Stable URL : http://www.jstor. 51: 1440–1450.

Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS: A high capacity genotyping by sequencing analysis pipeline. PLoS One 9: e90346.

Goddard, M. E., and B. J. Hayes, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10: 381–391.

Goddard, M. E., B. J. Hayes, and T. H. E. Meuwissen, 2010 Genomic selection in livestock populations. Genet. Res. (Camb). 92: 413–421.

González-Camacho, J. M., G. de Los Campos, P. Pérez, D. Gianola, J. E. Cairns *et al.*, 2012 Genome-enabled prediction of genetic values using radial basis function neural networks. Theor. Appl. Genet. 125: 759–771.

Grattapaglia, D., 2014 Breeding Forest Trees by Genomic Selection: Current Progress and theWay

Forward, pp. 651–82 in *Genomics of Plant Genetic Resources*, edited by R. Tuberosa, A. Graner, and E. Frison. Springer Netherlands.

Grattapaglia, D., and M. D. V Resende, 2011 Genomic selection in forest tree breeding. Tree Genet. Genomes 7: 241–255.

Van Grevenhof, E. M., J. a M. Van Arendonk, and P. Bijma, 2012 Response to genomic selection: the Bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. Genet. Sel. Evol. 44: 26.

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

Habier, D., R. L. Fernando, and D. J. Garrick, 2013 Genomic BLUP decoded: A look into the black box of genomic prediction. Genetics 194: 597–607.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller, 2010 The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42: 5.

Hallingbäck, H. R., and G. Jansson, 2013 Genetic information from progeny trials: A comparison between progenies generated by open pollination and by controlled crosses. Tree Genet. Genomes 9: 731–740.

Hartigan, J. A., and M. A. Wong, 1979 Algorithm AS 136: A K-Means Clustering Algorithm. J. R. Stat. Soc. Ser. C (Applied Stat. 28: 100–108.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009a Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 91: 47–60.

Hayes, B. J., P. M. Visscher, and M. E. Goddard, 2009b Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. (Camb). 91: 47–60.

Hazel, L. N., 1943 The genetic basis for constructing selection indices. Genetics 28: 476–490.

Henderson, C. R., 1976 A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. Biometrics 32: 69–83.

Henderson, C. R., 1984 Applications of Linear Models in Animal Breeding, pp. 258–289 in *University of Guelph, Ontario,*.

Henderson, C. R., 1975 Best Linear Unbiased Estimation and Prediction under a Selection Model. Biometrics 31: 423–447.

Henderson, C. R., 1953 Estimation of Variance and Covariance Components. Biometrics 9: 226–252.

Heslot, N., H.-P. Yang, M. E. Sorrells, and J. L. Jannink, 2012 Genomic Selection in Plant Breeding: A Comparison of Models. Crop Sci. 52: 146–160.

Hill, W. G., M. E. Goddard, and P. M. Visscher, 2008 Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet. 4: e1000008.

Hill, W. G., and B. S. Weir, 2011 Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. (Camb). 93: 47–64.

Hofheinz, N., and M. Frisch, 2014 Heteroscedastic ridge regression approaches for genome-wide prediction with a focus on computational efficiency and accurate effect estimation. G3 Genes| Genomes| Genet. 4: 539–546.

Isik, F., 2014 Genomic selection in forest tree breeding: the concept and an outlook to the future. New For. 45: 379–401.

Isik, F., J. Bartholomé, A. Farjat, E. Chancerel, A. Raffin *et al.*, 2015 Genomic selection in maritime pine. Plant Sci. 242: 108–119.

Isik, F., S. Kumar, P. J. Martínez-García, H. Iwata, and T. Yamamoto, 2015 Acceleration of Forest and Fruit Tree Domestication by Genomic Selection, pp. 93–124 in *Advances in Botanical Research*, edited by A.-B. A.-F. Plomion C. Elsevier Ltd.

Jannink, J. L., 2007 Identifying quantitative trait locus by genetic background interactions in association studies. Genetics 176: 553–561.

Jayawickrama, K. J. S., and M. J. Carson, 2000 A Breeding Strategy for the New Zealand Radiata Pine Breeding Cooperative. Silvae Genet. 49: 82–90.

Johnson, G. R., 1997 Site-to-site genetic correlations and their implications on breeding zone size and optimum number of progeny test sites for coastal Douglas-fir. Silvae Genet. 46: 280–285.

Kiss, G. K., and A. D. Yanchuk, 1991 Preliminary evaluation of genetic variation of weevil resistance in interior spruce in British Columbia. Can. J. For. Res. 21: 230–234.

Klápště, J., M. Lstibůrek, and Y. A. El-Kassaby, 2014 Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. Tree Genet. Genomes 10: 241–249.

Korecký, J., J. Klápště, M. Lstibůrek, J. Kobliha, C. D. Nelson *et al.*, 2013 Comparison of genetic parameters from marker-based relationship, sibship, and combined models in Scots pine multi-site open-pollinated tests. Tree Genet. Genomes 9: 1227–1235.

De La Torre, A. R., T. Wang, B. Jaquish, and S. N. Aitken, 2014 Adaptation and exogenous selection in a Picea glauca x Picea engelmannii hybrid zone: Implications for forest management under climate change. New Phytol. 201: 687–699.

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

Lee, S. H., M. E. Goddard, P. M. Visscher, and J. H. van der Werf, 2010 Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits. Genet. Sel. Evol. 42: 22.

Lindgren, D., L. Gea, and P. Jefferson, 1996 Loss of Genetic Diversity Monitored by Status Number. Silvae Genet. 45: 52–59.

Lindgren, D., and T. J. Mullin, 1997 Balancing Gain and Relatedness in Selection. Silvae Genet. 46: 124–129.

Lippert, C., G. Quon, E. Y. Kang, C. M. Kadie, J. Listgarten *et al.*, 2013 The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. Sci. Rep. 3: 1815.

Loberg, A., J. W. Dürr, W. F. Fikse, H. Jorjani, and L. Crooks, 2015 Estimates of genetic variance and variance of predicted genetic merits using pedigree or genomic relationship matrices in six Brown Swiss cattle populations for different traits. J. Anim. Breed. Genet. 132: 376–385.

Lorenz, A. J., S. Chao, F. G. Asoro, E. L. Heffner, T. Hayashi *et al.*, 2011 Genomic Selection in Plant Breeding. Knowledge and Prospects. Adv. Agron. 110: 77–123.

Lorenzana, R. E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theor. Appl. Genet. 120: 151–161.

Lu, F., A. E. Lipka, J. Glaubitz, R. Elshire, J. H. Cherney *et al.*, 2013 Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genet. 9: e1003215.

Luan, T., J. a Woolliams, S. Lien, M. Kent, M. Svendsen *et al.*, 2009 The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. Genetics 183: 1119–1126.

Lush, J. L., 2013 *Animal breeding plans*. Read Books Ltd.

Lynch, M., Walsh, B., 1998 *Genetics and Analysis of Quantitative Traits. Vol.1*. Sinauer, Sunderland, MA.

Mackay, T. F. C., 2014 Epistasis and quantitative traits: using model organisms to study gene-gene interactions. Nat. Rev. Genet. 15: 22–33.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Millman, M., 1976 Metric volume and V-bar tables derived from the British Columbia Forest Service whole stem cubic metre volume equations.(unpublished Report, Vancouver, BC).:

Misztal, I., A. Legarra, and I. Aguilar, 2009 Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92: 4648–4655.

Mrode, R. A., 2005 *Linear Models for the Prediction of Animal Breeding Values*.

Muñoz, P. R., M. F. R. Resende, S. A. Gezan, M. D. V. Resende, G. de los Campos *et al.*, 2014 Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198: 1759–1768.

Nakagawa, S., and H. Schielzeth, 2013 A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods Ecol. Evol. 4: 133–142.

Namkoong, G., 1966 Inbreeding effects on estimation of genetic additive variance. For. Sci. 12: 8–13.

Namkoong, G., H. C. Kang, and J. S. Brouard, 2012 *Tree Breeding: Principles and Strategies*. Principles and Strategies. Vol. 11. Springer Science & Business Media.

Neale, D. B., and C. G. Williams, 1991 Restriction-Fragment-Length-Polymorphism mapping in conifers and applications to forest genetics and tree improvement. Can J Res 21: 545–554.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12: 443–51.

Nystedt, B., N. R. Street, A. Wetterbom, A. Zuccolo, Y.-C. Lin *et al.*, 2013 The Norway spruce genome sequence and conifer genome evolution. Nature 497: 579–584.

Pan, J., B. Wang, Z. Y. Pei, W. Zhao, J. Gao *et al.*, 2015 Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Mol. Ecol. Resour. 15: 711–722.

Paterson, A. H., S. D. Tanksley, and M. E. Sorrells, 1991 DNA Markers in Plant Improvement. Adv. Agron. 46: 39–90.

Perry, P. O., 2009 “bcv: Cross-Validation for the SVD (bi-cross-validation).” R package version 1.

Peterson, B. K., J. N. Weber, E. H. Kay, H. S. Fisher, and H. E. Hoekstra, 2012 Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7: e37135.

Poland, J., J. Endelman, J. Dawson, J. Rutkoski, S. Y. Wu *et al.*, 2012 Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. Plant Genome 5: 103–113.

Porth, I., J. Klápště, O. Skyba, B. S. K. Lai, A. Geraldes *et al.*, 2013 Populus trichocarpa cell wall chemistry and ultrastructure trait variation, genetic control and genetic correlations. New Phytol. 197: 777–90.

Porth, I., R. White, B. Jaquish, R. Alfaro, C. Ritland *et al.*, 2012 Genetical Genomics Identifies the Genetic Architecture for Growth and Weevil Resistance in Spruce. PLoS One 7: e44397.

Powell, J. E., P. M. Visscher, and M. E. Goddard, 2010 Reconciling the analysis of IBD and IBS in complex trait studies. Nat. Rev. Genet. 11: 800–5.

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

Ratcliffe, B., O. Gamal El-Dien, J. Klápště, I. Porth, C. Chen *et al.*, 2015 A comparison of genomic selection models across time in interior spruce (Picea engelmannii × glauca) using unordered

SNP imputation methods. Heredity (Edinb). 5: 547–555.

Ratcliffe, B., F. J. Hart, J. Klápště, B. Jaquish, S. D. Mansfield *et al.*, 2013 Genetics of wood quality attributes in Western Larch. Ann. For. Sci. 71: 415–424.

Resende, M. F. R., P. Muñoz, J. J. Acosta, G. F. Peter, J. M. Davis *et al.*, 2012 Accelerating the domestication of trees using genomic selection: Accuracy of prediction models across ages and environments. New Phytol. 193: 617–624.

Resende, M. F. R., P. Muñoz, M. D. Resende, D. J. Garrick, R. L. Fernando *et al.*, 2012 Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190: 1503–1510.

Resende, M. D. V, M. F. R. Resende, C. P. Sansaloni, C. D. Petroli, A. a. Missiaggia *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: 116–128.

Rigault, P., B. Boyle, P. Lepage, J. E. K. Cooke, J. Bousquet *et al.*, 2011 A White Spruce Gene Catalog for Conifer Genome Analyses. Plant Physiol. 157: 14–28.

Rutkoski, J. E., J. Poland, J.-L. Jannink, and M. E. Sorrells, 2013 Imputation of unordered markers and the impact on genomic selection accuracy. G3 Genes| Genomes| Genet. 3: 427–439.

Santure, A. W., J. Stapley, A. D. Ball, T. R. Birkhead, T. Burke *et al.*, 2010 On the use of large marker panels to estimate inbreeding and relatedness: Empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. Mol. Ecol. 19: 1439–1451.

Schuster, S. C., 2008 Next-generation sequencing transforms today ' s biology. Nat. Methods 5: 16–18.

Shen, X., M. Alam, F. Fikse, and L. Rönnegård, 2013 A novel generalized ridge regression method for quantitative genetics. Genetics 193: 1255–1268.

Solberg, T. R., a K. Sonesson, J. a Woolliams, and T. H. E. Meuwissen, 2008 Genomic selection using different marker types and densities. J. Anim. Sci. 86: 2447–2454.

Squillace, A. E., 1974 Average genetic correlations among offspring from open-pollinated forest trees. Silvae Genet. 23: 149–156.

Stoehr, M. U., and Y. A. El-Kassaby, 1997 Levels of genetic diversity at different stages of the domestication cycle of interior spruce in British Columbia. Theor. Appl. Genet. 94: 83–90.

Strauss, S. H., R. Lande, and G. Namkoong, 1992 Limitations of molecular-marker-aided selection in forest tree breeding. Can J Res 22: 1050–1061.

Stuber, C. W., M. Polacco, and M. L. Senior, 1999 Synergy of empirical breeding, marker-assisted selection, and genomics to increase crop yield potential. Crop Sci. 39: 1571–1583.

Su, G., O. F. Christensen, T. Ostersen, M. Henryon, and M. S. Lund, 2012 Estimating Additive and Non-Additive Genetic Variances and Predicting Genetic Merits Using Genome-Wide Dense Single Nucleotide Polymorphism Markers. PLoS One 7: e45293.

Sutton, B. C. S., D. J. Flanagan, J. R. Gawley, C. H. Newton, D. T. Lester *et al.*, 1991 Inheritance of chloroplast and mitochondrial DNA in Picea and composition of hybrids from introgression zones. Theor. Appl. Genet. 82: 242–248.

Thomas, S. C., D. W. Coltman, and J. M. Pemberton, 2002 The use of marker-based relationship information to estimate the heritability of body weight in a natural population: A cautionary tale. J. Evol. Biol. 15: 92–99.

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie *et al.*, 2001 Missing value estimation methods for DNA microarrays. Bioinformatics 17: 520–525.

Truong, H. T., A. M. Ramos, F. Yalcin, M. de Ruiter, H. J. A. van der Poel *et al.*, 2012 Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7: e37565.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414–23.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.*, 2009 Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92: 16–24.

Visscher, P. M., S. E. Medland, M. A. R. Ferreira, K. I. Morley, G. Zhu *et al.*, 2006 Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2: e41.

Vitezica, Z. G., L. Varona, and A. Legarra, 2013 On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics 195: 1223–1230.

Wang, W., Z. Wei, T.-W. Lam, and J. Wang, 2011 Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. Sci. Rep. 1: 1–7.

White, T. L., W. T. Adams, and D. B. Neale, 2007 *Forest Genetics*.

Wiggans, G. R., J. B. Cole, S. M. Hubbard, and T. S. Sonstegard, 2016 Genomic Selection in Dairy Cattle: The USDA Experience. Annu. Rev. Anim. Biosci. 5: 13.1-13.19.

Williams, C. G., and D. B. Neale, 1992 Conifer wood quality and marker-aided selection—a case-study. Can J Res 22: 1009–1017.

Wright, S., 1922 Coefficients of Inbreeding and Relationship. Am. Nat. 56: 330–338.

Zapata-Valenzuela, J., F. Isik, C. Maltecca, J. Wegrzyn, D. Neale *et al.*, 2012 SNP markers trace familial linkages in a cloned population of Pinus taeda-prospects for genomic selection. Tree Genet. Genomes 8: 1307–1318.

Zapata-Valenzuela, J., R. W. Whetten, D. Neale, S. McKeand, and F. Isik, 2013 Genomic estimated breeding values using genomic relationship matrices in a cloned population of loblolly pine. G3 (Bethesda). 3: 909–16.