A THESIS FOR DEGREE OF DOCTOR OF PHILOSOPHY

Genomic selection for growth characteristics in Korean red pine (*Pinus densiflora* Siebold & Zucc.)

소나무의 생장형질에 대한 유전체 선발

August 2022

Program in Forest Environmental Science
Department of Agriculture, Forestry and Bioresources
Graduate School of Seoul National University

Hye-In Kang

Genomic selection for growth characteristics in Korean red pine (*Pinus densiflora* Siebold & Zucc.)

Under the supervision of Professor Kyu-Suk KANG

Submitting a Thesis for a Doctoral Degree in Agriculture

May 2022

Program in Forest Environmental Science
Department of Agriculture, Forestry and Bioresources
Graduate School of Seoul National University

Hye-In Kang

Confirming the Ph.D. Dissertation written by Hye-In Kang

June 2022

Chair	Hyun Seok KIM, Ph.D.	(Seal)
Vice Chair	Kyu-Suk KANG, Ph.D.	(Seal)
Examiner	Tae-Jin YANG, Ph.D.	(Seal)
Examiner	Donghwan SHIM, Ph.D.	(Seal)
Examiner	In-Sik KIM, Ph.D.	(Seal)

Abstract

Korean red pine (*Pinus densiflora*), a native forest tree species of South Korea, is distributed in East Asia and is in high demand for reforestation due to its high value as timber. In order to improve the growth of Korean red pine which is directly related to wood productivity, selective breeding has been traditionally carried out through progeny tests and the generation has been advanced. However, there has been a problem that traditional selection takes a long time, so the annual genetic gain has been limited. Genomic selection (GS) is an alternative to progeny test in the breeding process, estimating genotype-based breeding values of individuals through genomic information using molecular markers. In this study, for the purpose of shortening the breeding cycle of Korean red pine, GS was introduced to the breeding population and the selective efficiency and applicability were evaluated.

At first, in order to identify the characteristics of the target population of the GS, the phenotype of population was analyzed statistically and genetic parameters such as heritability and genetic correlation were estimated. The phenotypes of open- and control-pollinated populations differed in the mean according to the test sites and families. In addition, the superiority of families was different for each test site and the genetic correlation between test sites was found to be low. Thus, it was judged that the interaction between the genotype and the environment was large in the target population, and that phenotype correction was necessary when including multiple regions in GS.

Then, as the stage of training the GS model for Korean red pine, cross-validations under several conditions regarding the marker set, the predictive model, and the training data set were performed in an open-pollinated progeny trial and the prediction accuracy was compared for the model

optimization. Also, the response to the selection of GS was compared to that of traditional selections. As the result, the predictive model or the number of cross-validation folds did not affect, and the heritability, marker selection method, and the environment or family composition of the training and test set did affect the prediction accuracy of GS. Therefore, it was found that training the GS model to include individuals from various environments using GBLUP and markers with a minor allele frequency of larger than 0.05 was effective. In addition, a higher annual genetic gain could be obtained through GS compared to phenotypic selection or family selection.

Finally, in order to verify the practical utility of the Korean red pine GS model, the breeding values of individuals in the control-pollinated progeny trial which is unrelated to training population was predicted using the trained model and the prediction accuracy was evaluated. As the result of comparison according to the relationship of population, the prediction accuracy was higher in the full-sib population, which had a closer genetic relationship with each other, than the half-sib population. Also, predicting the genomic estimated breeding value of a control-pollinated population with the trained GS model was possible. Therefore, it was concluded that the GS model on Korean red pine of this study could be applied to a breeding population with different families and environments.

Through this study, GS in Korean red pine was considered to have a high selective efficiency enough to be able to replace the traditional selection. It was expected that the basis for accelerated breeding of Korean red pine would be laid through the next generation test of selected trees by GS in the future.

Keywords: Korean red pine, progeny test, genomic selection, accelerated breeding, breeding value, genetic gain

Student number: 2018-35767

Contents

List of Tables

List of Figures

List of Abbreviation

ABLUP	additive BLUP
AC	accuracy
BLUP	best linear unbiased prediction
BV	breeding value
CRBT	call rate below threshold
DBH	diameter at breast height
EBV	estimated breeding value
FS	family selection
GBLUP	genomic BLUP
GBS	genotyping-by-sequencing
GCA	general combining ability
GEBV	genomic estimated breeding value
GRM	genomic realized relationship matrix
GS	genomic selection
HBLUP	hybrid genomic BLUP
LD	linkage disequilibrium
MAF	minor allele frequency
MHR	mono high resolution
NMH	no minor homo
NRM	numerator relationship matrix
OTV	off-target variant
PA	predictive ability
PHR	poly high resolution
PS	phenotypic selection
QTL	quantitative trait locus
SNP	single nucleotide polymorphism
TBV	true breeding value

General introduction

Study background

Korean red pine (*Pinus densiflora* Siebold & Zucc.) belongs to the genus *Pinus* of the family Pinaceae and is a native species of South Korea. It is widely distributed throughout East Asia from the Korean Peninsula to Japan and China (Szmidt & Wang 1993). As Korean red pine is highly valued as timber, Korean red pine wood is traded at the second-highest log price after Japanese false cypress (*Chamaecyparis obtusa*) wood in South Korea. Accordingly, reforestation of Korean red pine is in high demand, accounting for about 17% of the annual reforestation area in South Korea as of 2020 (KFS, 2021). Therefore, it is essential to do research for improving wood productivity by breeding the economic traits of Korean red pine.

Tree breeding has been used as one of the important tools to genetically improve the trees by applying genetic principles and techniques (White et al., 2007). Breeding is an activity to increase the frequency of a preferred allele within a population, and tree breeding is usually conducted based on a selective breeding methodology. Selective breeding induces the crosses between selected trees to produce an improved next generation (Wright, 2012). In South Korea, since the tree breeding was started in 1959, the breeding programs including plus tree selection, progeny test, and seed orchard establishment have been implemented until now (Lee et al., 2020). On the other hand, the disadvantage of tree breeding is certainly that it takes a long time. Compared to crops and livestock, forest trees have a longer generation period, and crossing and nurturing for establishment of a breeding population also take a long time. These make the advance of generation take

30 to 45 years in tree breeding. Currently, the breeding program has been advanced up to the fourth generation worldwide in loblolly pine (*Pinus taeda*) (Isik and Mckeand, 2019) and up to the second generation in South Korea in Korean red pine and black pine (*Pinus thunbergii*). As the response to selection which is the expected value of improvement according to the progress of one generation is limited, accelerated breeding is required for rapid generation advancement. With the recent development of molecular breeding technology and genomics, the paradigm of tree breeding is changing towards bioinformatics using big data (El-Kassaby et al., 2014; Isik, 2014).

Since the next generation sequencing (NGS) and statistical analysis methods for large-scale data have been developed, genomic selection (GS) has been proposed as an alternative to traditional selection. GS is a kind of selective breeding that uses molecular marker information instead of phenotype or pedigree information to estimate the genetic value of each individual as a criterion for selection in the breeding population (Meuwissen et al., 2001). Conventional family selection (FS) is only effective at capturing the average effect of the parents and it does not focus on the genetic information that individuals or offspring share with their relatives. While, it is possible to capture the differences among individuals caused by small marker effects in GS where the breeding value is estimated by summing the effects of thousands to hundreds of thousands of markers (Goddard et al., 2011).

Breeding of woody plants is more time- and cost-consuming than the breeding of crops. Trees generally have a long juvenile period, take a long time to flower and produce seeds, and have a large physical size compared to crops. Accordingly, the progeny test of forest trees requires a wide area and long-term observation. In order to improve the efficiency of tree breeding, a

method of shortening the time and reducing the effort for progeny tests is required. In addition, as the utilization of forest trees is diversified and global climate change gets severe, the target traits of tree breeding are changing rapidly. These are why accelerated breeding is important in forestry. The biggest advantage of GS in forest trees is that the selection efficiency could be improved by reducing the generation interval through selection using genetic information before phenotypes are expressed (Grattapaglia and Resende, 2011). Besides, GS would raise the response to selection by increasing the selection intensity (Grattapaglia, 2017; Isik, 2014). Therefore, if GS is introduced into the breeding of Korean red pine, a high improvement could be expected by rapidly advancing the generation at a low cost through early and intensive selection using markers rather than through the progeny test.

Literature review

1. Conventional breeding of genus *Pinus*

Genus *Pinus* belongs to the family Pinaceae, order Pinales. According to the National Standard Plant List, six native pine species including Korean red pine, Korean pine (*Pinus koraiensis*), Ulleungdo white pine (*P. parviflora*), dwarf Siberian pine (*P. pumila*), black pine, and Manchurian red pine (*P. tabuliformis* var. *mukdensis*) are distributed in South Korea. The genus *Pinus* is the most actively studied in worldwide tree breeding (Eo et al., 2020).

The breeding of loblolly pine, one of the most representative tree species in the United States, has been focused on a tree breeding program led by a collaboration of North Carolina State University, the University of Florida, and the Texas Forest Service for 43 years in the southern United States (Li et al., 1999; Mckeand, 2019). The breeding of loblolly pine had begun with the selection of plus trees from natural stands in 1957 and the first breeding cycle continued until the early 1970s. Since 1969, genetically improved seeds from seed orchards have been distributed for reforestation. Then, the artificial crosses were performed and 270 progenies for each full-sib family were planted (Isik and Mckeand, 2019). In the second breeding cycle, selections were made from first-generation progeny tests and non-improved plantations so that the genetic diversity of second-generation was ensured. Afterward, the intensive progeny test including testing various areas was conducted and the best families were selected from the second-generation seed orchard for plantation (Duzan and Willliams, 1988). The loblolly pine breeding program dramatically increased the volume of harvest in the United States. The plantation generated by seeds from first-generation seed orchards showed 7 to 12% larger productivity per hectare than wild stands (Li et al. 1999). Also,

the progenies of open-pollinated families that were selected in third-cycle selection had a 48% better yield than non-breeding trees (Mckeand, 2019).

As loblolly pine has a relatively longer breeding history than other tree species, it became the first target species of GS in forest trees. In the earliest study, 4 predictive models were applied to 17 traits regarding growth, development, disease resistance, and wood quality, showing a predictive ability of 0.17 for lignin content to 0.51 for branch angle average (Resende Jr. et al. 2012a). Since then, loblolly pine has been studied with the highest frequency in the GS researches on forest trees (Levedev et al., 2020).

Monterey pine (*Pinus radiata*), which is native to the western coast of the United States and the islands of Mexico, has been introduced and bred in New Zealand and Australia. The breeding of Monterey pine was also initiated under the leading of the nation represented by the New Zealand Forest Service. The first selection of plus tree was conducted in the 1950s and two generations were advanced by the early 2000s (Jayawickrama and Carson 2000). The target traits of breeding were growth, resistance to *Dothistroma pini*, high wood density, and long internode. In the second-generation breeding in 1992, a total of 83 full-sib families were crossed with 103 parents and the progenies were planted on 3 test sites. Then in the early 2000s, the forward selection was conducted on one of these test sites and the third-generation test sites were established (Dungey et al. 2009). GS has also been studied in Monterey pine, obtaining an accuracy of 0.55 for stem straightness, 0.57 for external resin bleeding, 0.59 for internal checking and 0.70 for branch-cluster frequency using 67,000 SNPs (Li et al., 2019).

Research on the breeding of Korean red pine in South Korea started with a plus tree selection project in 1959 (NIFoS, 2019). At the beginning of the project, plus trees were selected mainly in Gangwon province where Korean

red pine grows well, but later, the selection was expanded to the whole country to secure genetic variation and to test the plantation adaptability. The clone bank was established to preserve the selected 424 plus trees, and 19.57ha area of it is being managed over Chungju, Gangneung, Taean, Suwon, and Jeju as of 2018.

Korean red pine trees were selected through open- and control-pollinated progeny tests, and the first-generation seed orchard of 99ha was established in 1969 (Kang et al., 2003). The open-pollinated progeny trial of 22.66ha for genetic testing was established in 9 regions from 1972 to 1987, but at present, 11.58ha area in 6 regions (Chuncheon, Gongju, Taean, Wanju, Naju, and Kyeongju) are being managed and investigated (NIFoS, 2019). In addition, a control-pollinated progeny trial was established in 8 regions, starting with Naju and Gongju in 1977. In a study on the progeny test, the heritability increased with age from 3 to 8 years old, and the heritability of height was significantly higher than that of the root collar diameter and branch diameter (Yim and Noh, 1979). Also, the response to selection reached 36.6% when the selection intensity was 1/500 in 8-year-old trees.

Since the 2000s, a project for the establishment of a second-generation seed orchard using genetic test results has been promoted. Through the progeny test of 35-year-old trees, the expected response to selection on volume in the second-generation seed orchard is estimated to be about 15% larger compared to the natural stands and 7.5% larger compared to the first-generation seed orchard (NIFoS, 2016). Recently, as interest in breeding using the genome information of Korean red pine is increasing, studies on complete chloroplast genome and SNP chip development have been conducted (Kang et al., 2019; Cheon et al., 2021).

Although the improvement of pine started in the 1950s, the

domestication of the pine tree is still in its infancy because only a few generations have progressed yet. It is a meaningful achievement showing outperformance of improved population over the wild stand, but it is unfortunate that it took a long time and the scale was limited. Therefore, the recent development of tee breeding using genome information is desirable in that it could overcome these limitations.

2. Molecular breeding

As the limitations of tree breeding through phenotypic observation had been revealed, attempts were made to accelerate the selection by genotype evaluation using molecular markers. Marker-assisted selection (MAS), the early method of molecular breeding, has been applied to plants since the 1990s (Desta and Ortiz 2014). The principle of MAS is to use the linkage disequilibrium (LD) between the marker and the quantitative trait locus (QTL) (Muranty et al., 2014). MAS is the most effective for the traits regulated by several QTLs which are responsible for a large portion of phenotypic variation. Therefore, MAS has been usually utilized for the traits that are controlled by a few major genes such as disease resistance in forest trees, and it was achieved to improve selection efficiency and reduce time and effort for progeny test (Grattapaglia, 2017). However, the economic characteristics of trees, such as growth, wood quality, and stem straightness, are defined as quantitative traits and are linked to many QTLs with small effects. For these traits, MAS had not explained genetic variation effectively (Grattapaglia, 2017).

Afterward, selective breeding has been developed toward dealing with the effects of the whole genome, adopting the principle of quantitative

genetics. Nejati-Javaremi et al. (1997) first introduced the concept of applying a genomic relationship estimated by markers instead of an additive relationship derived from pedigree to a mixed model for estimating breeding values. In addition, Meuwissen et al. (2001) suggested that the genetic value could be predicted even for individuals lacking their phenotypes and also suggested the statistical methods for prediction, which had been developed into GS. The difference of GS from MAS is that MAS uses only a few markers associated with QTLs that have large effects on the phenotype, whereas GS uses much more markers associated with QTLs that have small effects on the phenotype. GS assumes that each QTL is likely to exist in the same LD as at least one marker locus (Desta and Ortiz, 2014). Therefore, GS tends to capture most genetic variations for complex quantitative traits explained by a large number of small effects (Grattapaglia, 2014). Besides, unlike MAS, GS does not require prior information such as the association between phenotype and marker, the location of the QTL on the genome, and the relative influence of marker on the phenotype, and does focuse only on the selection efficiency (Isik et al. 2016). Accordingly, the time and effort required for the experiment for finding information on the association of a specific marker and the trait could be saved in GS.

Generally, GS proceeds in the following steps. First, the genotype and phenotype of a training group in a breeding population are evaluated. Second, the two data are combined to create a prediction model that simultaneously estimates the effects of all marker loci. Third, cross-validation is performed to test the applicability of the developed model. Finally, genetic values are predicted for different subgroups of the breeding population, and the individuals for advanced generations are selected based on the genomic estimated values (Grattapaglia, 2014).

Since the GS was first introduced, it has been widely applied to and studied in livestock and crops (VanRaden et al. 2009, Wolc et al. 2011, Zhao et al. 2012). After, researches on GS were started in forestry and have been actively conducted until now (Table 1). GS was studied in loblolly pine and eucalyptus hybrids for the first time in forestry (Resende Jr et al., 2012a; Resende Jr et al., 2012b; Resende et al. 2012). Growth and wood quality were studied in a eucalyptus hybrid full-sib population with a size of 738 to 920, and the prediction accuracy was 0.38~0.60 (Resende et al., 2012). In loblolly pine, a population including 61 full-sib families was studied using 4,850 SNPs, and the prediction accuracy, varied according to the environment, was 0.17~0.51 for 17 traits related to growth, material, and rust resistance (Resende Jr et al., 2012a; Resende Jr et al. 2012b). Afterward, experimental studies of GS in conifers such as *Pinus*, *Picea*, *Pseudotsuga,* and *Cryptomeria* and broadleaf trees such as *Eucalyptus*, *Castanea*, *Fraxinus*, and *Populus* have been continued (Lebedev et al., 2020) (Table 1). Previous researches on forest trees used populations of 25 to 338 half-sib or full-sib families as materials and mainly studied for growth and wood quality. While, the predictive accuracies cannot be compared directly because they were calculated in different ways in each paper.

Molecular breeding using markers has been continuously developed, and the latest technology is GS. The advantages of GS over MAS might be highlighted especially in conifers, of which genomic information is mostly unknown. So far, the most researches on GS in forest trees have been limited to cross-validation within a breeding population, which is the third step for GS. The fourth step, validating GS models using other populations and selecting individuals, should be conducted for practical application of GS.

Table 1. Previous studies on the genomic selection of forest trees.

Species	Population structure	Environ-ment	Population size	Number and type of markers	Traits [a]	Predictive model [b]	Prediction accuracy [c]	Reference
Eucalyptus grandis x *E. urophylla*	43 full-sib families	3	738	3,129 DArT	growth, wood quality	RR-BLUP	0.54-0.60	Resende et al., 2012
E. grandis, E. urophylla, E. globulus and hybrids	75 full-sib families	-	920	3,564 DArT	growth, wood quality	RR-BLUP	0.38-0.55	Resende et al., 2012
Pinus taeda	61 full-sib families	-	951	4,853 SNP	growth, wood quality, disease resistance	RR-BLUP, Bayes A, Bayes Cπ, BL	0.17-0.51	Resende Jr et al., 2012a
Pinus taeda	61 full-sib families	4	800	4,825 SNP	growth	RR-BLUP	0.26-0.37	Resende Jr et al., 2012b
Pinus taeda	61 full-sib families	-	951	4,853 SNP	growth	GBLUP	0.66-0.86	Muñoz et al., 2014
Picea glauca	214 half-sib families	3	1,694	6,385 SNP	growth, wood quality	GBLUP	0.09-0.44	Beaulieu et al., 2014a

Table 1. (Continued)

Species	Population structure	Environ-ment	Population size	Number and type of markers	Traits [a]	Predictive model [b]	Prediction accuracy [c]	Reference
Picea glauca	59 full-sib families	2	1,748	6,932 SNP	growth, wood quality	RR-BLUP, BL	0.0-0.79	Beaulieu et al., 2014b
Picea glauca x *P. engelmannii*	25 half-sib families	3	1,126	8,868-62,198 SNP	growth, wood quality	GBLUP, RR-BLUP, GRR	0.01-0.77	El-Dien et al., 2015
Picea glauca x *P. engelmannii*	25 half-sib families	2	769	34,570-50,803 SNP	growth	RR-BLUP, Bayes Cπ, GRR	0.04-0.47	Ratcliffe et al., 2015
Pinus pinaster	191 half-sib families	-	661	2,500 SNP	growth, tree architecture	GBLUP, BL, BRR	0.43-0.49	Isik et al., 2016
Eucalyptus globulus	40 full-sib and 13 half-sib families	-	310	~12,000 SNP	growth, wood quality	GBLUP, Bayes B, Bayes C, BL	0.58-0.75	Duran et al., 2017
Eucalyptus urophylla x *E. grandis*	338 full-sib families	-	958	41,304 SNP	growth, wood quality	GBLUP, RR-BLUP, BL, RKHS	0.25-0.29	Tan et al., 2017

Table 1. (Continued)

Species	Population structure	Environ-ment	Population size	Number and type of markers	Traits [a]	Predictive model [b]	Prediction accuracy [c]	Reference
Picea mariana	34 full-sib families	2	734	4,993 SNP	growth, wood quality	GBLUP	0.23-0.86	Lenz et al., 2017
Picea abies	128 full-sib families	2	1,370	116,765 SNP	growth, wood quality	GBLUP, BL, BRR, RKHS	0.02-0.81	Chen et al., 2018
Eucalyptus grandis x *E. urophylla*	45 full-sib families	-	999	33,398 SNP	growth, wood quality	GBLUP, HBLUP	0.70-0.73	Cappa et al., 2019
Eucalyptus globulus	Full-sib and half-sib families	-	646	14,422 SNP	growth, wood quality, tree architecture	Bayes A, Bayes B, Bayes C, BL, BRR	0.06-0.58	Ballesta et al., 2019
Pinus contorta	42 full-sib and 57 half-sib families	4	1,569	19,584 SNP	growth, wood quality	HBLUP, GBLUP	0.08-0.85	Ukrainetz and Mansfield, 2020a
Pinus contorta	42 full-sib and 57 half-sib families	4	1,569	19,584 SNP	growth, wood quality	GBLUP, Bayes B, Bayes C	0.27-0.83	Ukrainetz and Mansfield, 2020b

Table 1. (Continued)

Species	Population structure	Environ-ment	Population size	Number and type of markers	Traits [a]	Predictive model [b]	Prediction accuracy [c]	Reference
Pinus radiata	Full-sib families	3	1,105	67,168 SNP	tree architecture	GBLUP	0.55-0.70	Li et al., 2019
Pinus sylvestris	138 full-sib families	-	694	8,719 SNP	growth, wood quality	GBLUP, BL, BRR	0.15-0.84	Calleja-Rodriguez et al., 2019
Pseudotsuga menziesii	37 full-sib families	3	1,321	69,551 SNP	growth	RR-BLUP, Bayes B, GRR	-0.30-0.92	Thistlethwaite et al., 2019
Populus deltoides	473 clones	2	3,784	68,885 SNP	growth	GBLUP	-	Alves et al., 2020
Picea abies	40 full-sib families	2	726	5,660 SNP	growth, wood quality, disease resistance	GBLUP, Bayes Cπ, BRR	0.10-0.46	Lenz et al., 2020a
Picea glauca	136 full-sib families	-	1,516	4,148 SNP	growth, wood quality, disease resistance	GBLUP, Bayes Cπ	0.14-0.48	Beaulieu et al., 2020

Table 1. (Continued)

Species	Population structure	Environ-ment	Population size	Number and type of markers	Traits [a]	Predictive model [b]	Prediction accuracy [c]	Reference
Eucalyptus dunnii	150 half-sib families	5	4,860	11,284 SNP	growth	HBLUP	0.32-0.89	Jurcic et al., 2021
Picea glauca	38 half-sib families	2	560	4,091 SNP	drought response	GBLUP	-	Laverdiere et al., 2022

[a] Growth includes diameter at breast height, tree height, volume growth, and straightness. Wood quality includes pulp yield, wood stiffness, wood density, lignin content, cellulose content, and fiber length. etc.

[b] GBLUP, genomic BLUP; HBLUP, hybrid genomic BLUP; BL, Bayesian LASSO; BRR, Bayesian ridge regression; GRR, generalized ridge regression; RKHS, reproducing kernel Hilbert space.

[c] Predictive accuracies were estimated in different ways in each paper and varied with trait, prediction model, environment, and genomic selection scenario.

3. Markers and genomics

Molecular markers have been widely applied to identifying individuals and determining genotypes in plant breeding and genetic studies. Various types of molecular markers have been developed until recently. The first used molecular marker was restriction fragment length polymorphism (RFLP), which uses a restriction endonuclease that cuts only a region having a specific DNA nucleotide sequence (Tanksley et al., 1989). RFLP determines genotypes using the differences in the size of the DNA fragments caused by a variation in the DNA base sequence when each individual is treated with a restriction enzyme.

As disadvantages of RFLP that it took a lot of time and required a high level of skill were found, genotyping techniques using polymerase chain reaction (PCR) were proposed. Random amplified polymorphic DNA (RAPD) is a method of amplifying only a DNA region of a complementary sequence to a primer which has a random sequence of about 10bp and analyzing the size of the PCR products (Arif et al., 2010). Amplified fragment length polymorphism (AFLP) is a method of cleaving DNA using restriction enzymes, attaching adapters to both ends of the cleaved fragments, amplifying fragments by PCR, and then analyzing the size of the products (Vos et al., 1995). AFLP had the advantages that it can obtain more sensitive PCR products than RAPD and can identify several loci simultaneously, but still had a problem of low reproducibility. Simple sequence repeat (SSR) is a technique for identifying the size of a product after amplifying the position of the repeat sequence using a primer with a fluorescent marker, paying attention to the fact that the number of repeats of the simple nucleotide sequence is different for each individual (Zietkiewicz et al., 1994). SSR is still being

actively used to identify species and individuals.

Afterward, as NGS began to be utilized for genotyping. Single nucleotide polymorphism (SNP), a variation caused by the replacement of a single nucleotide in a DNA sequence, has high density rather than any other molecular marker. Due to these characteristics, SNP has been widely used in genome-wide association study (GWAS), QTL mapping analysis, and inter-individual relationship analysis (Ballesta et al., 2019).

For GS, genetic markers are required to be evenly distributed throughout the genome with high density (Desta and Ortiz, 2014). Compared to other molecular markers, SNP is frequently used in GS because it is capable of obtaining a high-density and stable genotype and is suitable for high-throughput genotyping platforms such as SNP array or SNP chip. SNP chip has relatively low cost and high data reproducibility, and it is accessible as they are provided by a lot of service providers (Grattapaglia et al., 2018).

In previous studies, SNPs of several *Pinus* species were discovered. The SNP chip of loblolly pine consisted of 7,216 SNPs derived from expressed sequence tags (ESTs) (Eckert et al., 2010). Subsequently, the SNP chip was utilized for the GS of loblolly pine (Resende et al., 2012; Muñoz et al., 2014). Also, a chip containing a 12K SNP was developed in maritime pine (*Pinus pinaster*) (Chancerel et al., 2013), and a 50K SNP chip was developed in lodgepole pine (*Pinus contorta*) (Suren et al., 2016). In forest tree species other than pine, SNP chips were developed and utilized in white spruce (*Picea glauca*), black spruce (*Picea mariana*), and Japanese cedar (*Cryptomeria japonica*) (Beaulieu et al., 2014a; Lenz et al., 2017; Mishima et al., 2018). In addition, 60K Illumina Infinium EuCHIP60K was developed using sequencing data covering 12 species in the genus *Eucalyptus* (Silva-Junior et al., 2015).

The SNP chip of Korean red pine has been developed using genotyping-by-sequencing (GBS) (Cheon et al., 2021). GBS, a method that uses restriction enzymes to reduce the complexity of the genome, has a very low cost per sample and does not require genome decoding (Ratcliffe et al., 2015). Therefore, GBS is suitable to be applied to forest trees or coniferous species for which reference genomes have not been analyzed due to large genome sizes (Grattapaglia, 2017). GBS was performed for use in GWAS on lodgepole pine, and the optimized GBS technique for *Pinus* was also reported (Parchman et al., 2012; Pan et al. 2015). Calleja-Rodriguez et al. (2019) pointed out the difficulty of developing a full-genome SNP panel in a tree with large genome size and discovered SNPs from Scots pine (*Pinus sylvestris*) through GBS for GS. GBS was also applied to the discovery of SNPs for GS in interior spruce (*Picea engelmannii* × *P. glauca*) (Ratcliffe et al., 2015).

In the history of genotyping, molecular markers have developed in the direction of improving polymorphism, reproducibility, and manageability. Currently, SNP is the most commonly used marker for GS and has already been widely developed in a lot of species including forest trees. It is a future task to secure high density and accuracy of SNP information for genome research of trees.

4. Genomic prediction models

Genetics traditionally has estimated the genetic value of an individual by combining phenotypic data and relative similarity. Breeding value (BV), the average value of the gene effects that appear in the offspring of an individual, is equal to the sum of the effects of each allele that affects the phenotype of an individual. An individual transmits half of its BV to the next generation.

In selective breeding, individuals with high BVs are selected to advance the generation. Although the true breeding value (TBV) of an individual is impossible to be measured, estimated breeding value (EBV) would be estimated through the phenotype of the offspring in various ways. GS is a method of estimating the BV of an individual using genomic information, and the estimated value by GS is called genomic estimated breeding value (GEBV).

Statistical models of GS for estimating GEBV have been developed over and over, and they vary depending on the assumptions about the distribution and variance of marker effects. Among them, the most popular one is the best linear unbiased prediction (BLUP). BLUP, proposed first by Henderson in 1950, is a statistical model for estimating random effects in a linear mixed model (Henderson, 1975). BLUP maximizes the correlation between the true value and the predicted value (best), and the effects of the model are in linear expression (linear). Also, the expected value of the estimator is the same as the parameter (unbiased). BLUP is similar to the best linear unbiased estimation (BLUE) of the Gauss-Markov theorem, but while BLUE estimates only fixed effects, BLUP additionally estimates random effects. Also, as BLUP was first used in livestock, the BV of individuals without phenotype observations was estimated, and the term 'prediction' started to be used (Henderson, 1975). Diverse BLUPs are depending on what matrix is used to find the solution of BLUP; ABLUP (additive BLUP) using pedigree information, GBLUP (genomic BLUP) using genomic information, and HBLUP (hybrid genomic BLUP) combining the two information. A general mathematical model for BLUP is as follows.

$$y = Xb + Zu + e$$

where y is the vector of the phenotype observation, X is the fixed effect incidence matrix, b is the vector of fixed effect, Z is the random effect incidence matrix, u is the vector of random effect, and e is the vector of random residual.

In order to apply to the GBLUP model, genomic information should be transformed into a genomic realized relationship matrix (GRM), which indicates the similarity between individuals in the population. The numerator relationship matrix (NRM), which has been used in conventional genetics, uses pedigree information. The relationship coefficient of NRM has based on the probability that two individuals inherit the same allele from a common ancestor (Wright, 1922). On the other hand, the relationship coefficient of GRM is based on the ratio of marker alleles shared by two individuals, and is calculated as the covariance of coded genotype data (VanRaden, 2008). The process of writing GRM is as follows.

1) Raw genotyping data is encoded by the gene content or the number of minor alleles (major homozygote, 0; heterozygote, 1; minor homozygote, 2) in the size of (the number of individuals) × (the number of markers).
2) Then, in order to set the center to 0, the Z matrix is created by subtracting twice the minor allele frequency (MAF) which is equal to the average gene content of the corresponding marker in each column.
3) The covariance matrix of the Z matrix is calculated. Then since the covariance increases as the number of markers increases, the covariance matrix is divided by the sum of the variances of each marker to compensate for this.

$$G = \frac{ZZ'}{2\Sigma P_i(1-P_i)}$$

where G is GRM, Z is a matrix with the size of (the number of individuals) × (the number of markers) obtained by subtracting twice the MAF from the number of minor alleles, and P_i is the MAF of the i-th marker.

Although the time and the cost of genotyping have recently been significantly reduced, the cost is still enormous if the number of samples to be analyzed is large. HBLUP or single-step genomic BLUP, one of the solutions, is a method of estimating the BV by combining pedigree information and partial genomic information when it is difficult to obtain genotypes of all individuals (Misztal et al., 2009; Legarra et al., 2009). A blended genetic relationship matrix is written by mixing the GRM of a part of the population with the NRMs of the whole population. The inverse of the blended genetic relationship matrix is calculated as follows (Aguilar et al., 2010; Legarra et al., 2009).

$$H^{-1} = A^{-1} + \begin{vmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{vmatrix}$$

where A^{-1} is the inverse matrix of NRM, G^{-1} is the inverse matrix of GRM, and A_{22}^{-1} is the inverse matrix of NRM of an individual with the genome information.

Different from GBLUP, which assumes that all marker effects follow a normal distribution, the Bayesian models assume that marker effects follow a Student's t-distribution and that markers would have different effects and variances (Wang et al., 2018). The main differences between the various Bayesian methods are the prior distributions assumed and the degree of shrinkage, which normalizes the values to reduce the error (Isik et al., 2017).

Bayes A assumes that the variance of marker effects follows an inverse chi-square distribution, so it is suitable for the analysis of traits controlled by several genes (Meuwissen et al., 2011). Bayes B assumes that the variance of the marker is equal to 0 with a probability of π and is suitable when the trait is strongly influenced by a specific locus. Bayes C assumes that the probability π has a prior uniform distribution (Habier et al., 2011). Bayesian ridge regression adds a small constant to a diagonal matrix to limit the regression coefficients (Hoerl and Kennard, 1970; Gianola, 2013). The Bayesian least absolute shrinkage and selection operator (LASSO) is a model designed not to set prior values (Legarra et al., 2011). The mathematical model for genomic regression using the Bayesian model is as follows (Gianola et al., 2009).

$$y = Xb + Wa_m + e$$

where y is the vector of phenotype observation, X is the incidence matrix of the fixed effect, b is the vector of fixed effect, W is the genotype matrix, a_m is the vector of the additive effect of the molecular marker, and e is the vector of residuals.

GBLUP has been utilized in most GS studies on forest trees, and Bayesian models also have been used in the study of pine, spruce, fir, and eucalyptus. Though statistical models have been developed for GS so far, the attempt to go beyond the frame of statistics such as introducing deep learning to GS would be made.

Purpose of study

The purpose of this study is to accelerate the improvement of Korean red pine by using GS that can replace the traditional technique of selective breeding. To this end, it is necessary to evaluate the genetic and phenotypic properties of the target population, train a GS model and assess the prediction accuracy, and verify its versatility by applying the model to the other population. Therefore, the study was conducted with the following three objectives.

First, the features of the target populations (open- and control-pollinated populations) are identified through statistical analysis of the phenotypes and estimation of genetic parameters such as heritability and genetic correlation. In addition, the individual breeding value which would be used to evaluate the prediction accuracy of GS is estimated.

Second, an efficient GS model for Korean red pine is proposed by comparing the prediction accuracy according to the marker selection method, the predictive model, the size or composition of the training set in the open-pollinated population. In addition, the expected response to selection of GS is evaluated by comparing it to that of traditional selection.

Third, the applicability of the trained GS model is tested using the control-pollinated population. Also, the selection is conducted based on the genomic estimated breeding value using the trained GS model.

There are several differences of this study from previous researches on GS in forest trees. First, this study is a large-scale GS study that includes 7 environments and more than 3,000 trees. Also, this study verifies the applicability of GS by using both half- and full-sib populations. Since this is the first study on GS in Korean red pine, the results of this study are expected to greatly contribute to the accelerated breeding of Korean red pine.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. and Lawlor, T.J. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93(2): 743-752.

Alves, F.C., Balmant, K.M., Resende Jr, M.F., Kirst, M. and de Los Campos, G. 2020. Accelerating forest tree breeding by integrating genomic selection and greenhouse phenotyping. The Plant Genome, 13(3), e20048.

Arif, I. A., Bakir, M. A., Khan, H. A., Al Farhan, A. H., Al Homaidan, A. A., Bahkali, A. H., ... and Shobrak, M. 2010. Application of RAPD for molecular characterization of plant species of medicinal value from an arid environment. Genet. Mol. Res. 9(4): 2191-2198.

Ballesta, P., Maldonado, C., Pérez-Rodríguez, P. and Mora, F. 2019. SNP and haplotype-based genomic selection of quantitative traits in Eucalyptus globulus. Plants 8(9): 331.

Beaulieu, J., Doerksen, T., Clément, S., MacKay, J. and Bousquet, J. 2014a. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113(4): 343-352.

Beaulieu, J., Doerksen, T. K., MacKay, J., Rainville, A. and Bousquet, J. 2014b. Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC genomics 15*(1): 1-16.

Beaulieu, J., Nadeau, S., Ding, C., Celedon, J. M., Azaiez, A., Ritland, C., ... and Bousquet, J. 2020. Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. Evolutionary applications 13(10): 2704-2722.

Calleja-Rodriguez, A., Pan, J., Funda, T., Chen, Z. Q., Baison, J., Isik, F., ... and Wu, H. X. 2019. Genomic prediction accuracies and abilities for growth and wood quality traits of Scots pine, using genotyping-by-sequencing (GBS) data. bioRxiv, 607648.

Cappa, E. P., de Lima, B. M., da Silva-Junior, O. B., Garcia, C. C., Mansfield, S. D. and Grattapaglia, D. 2019. Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Science 284*, 9-15.

Chancerel, E., Lamy, J. B., Lesur, I., Noirot, C., Klopp, C., Ehrenmann, F., ... and Plomion, C. 2013. High-density linkage mapping in a pine tree reveals a genomic region associated with inbreeding depression and provides clues to the extent and distribution of meiotic recombination. BMC biology 11(1): 1-19.

Chen, Z. Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., ... and Wu, H. X. 2018. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC genomics 19(1): 1-16.

Cheon, K. S., Kang, H. I., Park, Y. W., Song, J. H., Kim, I. S. and Shim, D. 2021. Development of SNP chip for Genomic Selection of Korean Red Pine (Pinus densiflora) Trees. Proceedings of The Korean Society of Breeding Science, 406.

Desta, Z. A. and Ortiz, R. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in plant science 19(9): 592-601.

Dungey, H. S., Brawner, J. T., Burger, F., Carson, M., Henson, M., Jefferson, P. and Matheson, A. C. 2009. A new breeding strategy for Pinus radiata in New Zealand and New South Wales. Silvae Genet 58(1-2): 28-38.

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C. and Valenzuela, S. 2017. Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics & Genomes 13*(4): 1-12.

Duzan, H. W. and Williams, C. G. 1988. Matching loblolly pine families to regeneration sites. Southern Journal of Applied Forestry 12(3): 166-169.

Eckert, A. J., Van Heerwaarden, J., Wegrzyn, J. L., Nelson, C. D., Ross-Ibarra, J., González-Martínez, S. C. and Neale, D. B. 2010. Patterns of population structure and environmental associations to aridity across the range of loblolly pine (Pinus taeda L., Pinaceae). Genetics 185(3): 969-982.

El-Dien, O. G., Ratcliffe, B., Klápště, J., Chen, C., Porth, I. and El-Kassaby, Y. A. 2015. Prediction accuracies for growth and wood attributes of interior spruce in space using genotyping-by-sequencing. BMC genomics 16(1): 1-16.

El-Kassaby, Y. A., Isik, F. and Whetten, R. W. 2014. Modern advances in tree breeding. In Challenges and Opportunities for the World's Forests in the 21st Century (pp. 441-459). Springer, Dordrecht.

Eo, S. H., Lee, B. J., Kang, K. S. and Kang, J. W. 2020. Overview of research on forest tree breeding in South Korea based on the keyword analysis in research articles. Korean J. Breed. Sci. Special Issue 189(197): 3.

Gianola, D., de Los Campos, G., Hill, W. G., Manfredi, E. and Fernando, R. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183(1): 347-363.

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

Goddard, M. E., Hayes, B. J. and Meuwissen, T. H. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. Journal of animal breeding and genetics 128(6): 409-421.

Grattapaglia, D. 2014. Breeding forest trees by genomic selection: current progress and the way forward. *Genomics of plant genetic resources* 651-682.

Grattapaglia, D. 2017. Status and perspectives of genomic selection in forest tree breeding. In Genomic selection for crop improvement (pp. 199-249). Springer, Cham.

Grattapaglia, D. and Resende, M. D. 2011. Genomic selection in forest tree breeding. Tree Genetics & Genomes 7(2): 241-255.

Grattapaglia, D., Silva-Junior, O. B., Resende, R. T., Cappa, E. P., Müller, B. S., Tan, B., ... and El-Kassaby, Y. A. 2018. Quantitative genetics and genomics converge to accelerate forest tree breeding. Frontiers in Plant Science 1693.

Habier, D., Fernando, R. L., Kizilkaya, K. and Garrick, D. J. 2011. Extension of the Bayesian alphabet for genomic selection. BMC bioinformatics 12(1): 1-12.

Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 423-447.

Hoerl, A. E. and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1): 55-67.

Isik, F. 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forests 45(3): 379-401.

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., ... and Bouffier, L. 2016. Genomic selection in maritime pine. *Plant*

Science 242, 108-119.

Isik, F., Holland, J. and Maltecca, C. 2017. Genetic data analysis for plant and animal breeding (Vol. 400). Cham, Switzerland: Springer International Publishing.

Isik, F. and McKeand, S. E. 2019. Fourth cycle breeding and testing strategy for Pinus taeda in the NC State University Cooperative Tree Improvement Program. *Tree Genetics & Genomes 15*(5): 1-12.

Jayawickrama, K. J. S. and Carson, M. J. 2000. A breeding strategy for the New Zealand radiata pine breeding cooperative. Silvae Genetica 49(2): 82-89.

Jurcic, E. J., Villalba, P. V., Pathauer, P. S., Palazzini, D. A., Oberschelp, G. P., Harrand, L., ... and Cappa, E. P. 2021. Single-step genomic prediction of Eucalyptus dunnii using different identity-by-descent and identity-by-state relationship matrices. Heredity 127(2): 176-189.

Kang, K. S., El-Kassaby, Y. A., Choi, W. Y., Han, S. U. and Kim, C. S. 2003. Genetic gain and diversity caused by genetic thinning in a clonal seed orchard of Pinus densiflora. Silvae Genetica 52(5-6): 220-223.

Kang, H. I., Lee, H. O., Lee, I. H., Kim, I. S., Lee, S. W., Yang, T. J. and Shim, D. 2019. Complete chloroplast genome of Pinus densiflora Siebold & Zucc. and comparative analysis with five pine trees. Forests 10(7): 600.

Korea Forest Service (KFS). 2021. 2021 Statistical Yearbook of Forestry (no. 51). Daejeon, Korea.

Laverdière, J. P., Lenz, P., Nadeau, S., Depardieu, C., Isabel, N., Perron, M., ... and Bousquet, J. 2022. Breeding for adaptation to climate change: genomic selection for drought response in a white spruce multi-site polycross test. Evolutionary Applications.

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I. and Shestibratov, K. A. 2020. Genomic selection for forest tree improvement: Methods, achievements and perspectives. Forests 11(11): 1190.

Lee, S. W., Kim, I. S., Lee, J. W., Choi, Y. I. and Lee, U. 2020. 60 years of forest tree improvement in Korea: accomplishment and prospects. Korean Journal of Breeding Science 52.

Legarra, A., Aguilar, I. and Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. Journal of dairy science 92(9): 4656-4663.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F. and Fritz, S. 2011. Improved Lasso for genomic selection. Genetics research 93(1): 77-87.

Lenz, P., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M. and Bousquet, J. 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC genomics 18(1): 1-17.

Lenz, P. R., Nadeau, S., Mottet, M. J., Perron, M., Isabel, N., Beaulieu, J. and Bousquet, J. 2020. Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. Evolutionary applications 13(1): 76-94.

Li, B., McKeand, S. and Weir, R. 1999. Impact of forest genetics on sustainable forestry—Results from two cycles of loblolly pine breeding in the US. Journal of sustainable forestry 10(1-2): 79-85.

Li, Y., Klápště, J., Telfer, E., Wilcox, P., Graham, N., Macdonald, L. and Dungey, H. S. 2019. Genomic selection for non-key traits in radiata pine when the documented pedigree is corrected using DNA marker information. BMC genomics 20(1): 1-10.

McKeand, S. E. 2019. The evolution of a seedling market for genetically improved loblolly pine in the southern United States. Journal of Forestry 117(3): 293-301.

Meuwissen, T. H., Hayes, B. J. and Goddard, M. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Meuwissen, T. H. E., Luan, T. and Woolliams, J. A. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. Journal of Animal Breeding and Genetics 128(6): 429-439.

Mishima, K., Hirao, T., Tsubomura, M., Tamura, M., Kurita, M., Nose, M., ... and Watanabe, A. 2018. Identification of novel putative causative genes and genetic marker for male sterility in Japanese cedar (Cryptomeria japonica D. Don). Bmc Genomics 19(1): 1-16.

Misztal, I., Legarra, A. and Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of dairy science 92(9): 4648-4655.

Muñoz, P. R., Resende Jr, M. F., Gezan, S. A., Resende, M. D. V., de Los Campos, G., Kirst, M., ... and Peter, G. F. 2014. Unraveling additive from

nonadditive effects using genomic relationship matrices. Genetics 198(4): 1759-1768.

National Institute of Forest Science (NIFoS). 2016. Selection and genetic test for breeding of main timber species. Research Reports no. 16-38. Seoul, Korea.

National Institute of Forest Science (NIFoS). 2019. Genetic test and improvement of genetic gain of main timber species. Research Reports no. 19-11. Seoul, Korea.

Nejati-Javaremi, A., Smith, C. and Gibson, J. P. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of animal science 75(7): 1738-1745.

Pan, J., Wang, B., Pei, Z. Y., Zhao, W., Gao, J., Mao, J. F. and Wang, X. R. 2015. Optimization of the genotyping-by-sequencing strategy for population genomic analysis in conifers. Molecular Ecology Resources 15(4): 711-722.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W. and Buerkle, C. A. 2012. Genome-wide association genetics of an adaptive trait in lodgepole pine. Molecular ecology 21(12): 2991-3005.

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B. and El-Kassaby, Y. A. 2015. A comparison of genomic selection models across time in interior spruce (Picea engelmannii× glauca) using unordered SNP imputation methods. Heredity 115(6): 547-555.

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., ... and Grattapaglia, D. 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194(1): 116-128.

Resende Jr, M. F. R., Munoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., ... and Kirst, M. 2012a. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190(4): 1503-1510.

Resende Jr, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., ... and Kirst, M. 2012b. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist 193(3): 617-624.

Silva-Junior, O. B., Faria, D. A. and Grattapaglia, D. 2015. A flexible multi-

species genome-wide 60K SNP chip developed from pooled resequencing of 240 Eucalyptus tree genomes across 12 species. New Phytologist 206(4): 1527-1540.

Suren, H., Hodgins, K. A., Yeaman, S., Nurkowski, K. A., Smets, P., Rieseberg, L. H., ... and Holliday, J. A. 2016. Exome capture from the spruce and pine giga-genomes. Molecular ecology resources 16(5): 1136-1146.

Szmidt, A. E. and Wang, X. R. 1993. Molecular systematics and genetic differentiation of Pinus sylvestris (L.) and P. densiflora (Sieb. et Zucc.). Theoretical and Applied Genetics 86(2): 159-165.

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B. and Ingvarsson, P. K. 2017. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology 17*(1): 1-15.

Tanksley, S. D., Young, N. D., Paterson, A. H. and Bonierbale, M. W. 1989. RFLP mapping in plant breeding: new tools for an old science. Bio/technology 7(3): 257-264.

Ukrainetz, N. K. and Mansfield, S. D. 2020a. Prediction accuracy of single-step BLUP for growth and wood quality traits in the lodgepole pine breeding program in British Columbia. Tree Genetics & Genomes 16(5): 1-13.

Ukrainetz, N. K. and Mansfield, S. D. 2020b. Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. *Tree Genetics & Genomes 16*(1): 1-19.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11): 4414-4423.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F. and Schenkel, F. S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of dairy science 92(1): 16-24.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. V. D., Hornes, M., ... and Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. Nucleic acids research 23(21): 4407-4414.

Wang, X., Xu, Y., Hu, Z. and Xu, C. 2018. Genomic selection methods for crop improvement: Current status and prospects. The Crop Journal 6(4): 330-340.

White, T. L., Adams, W. T. and Neale, D. B. (Eds.). 2007. Forest genetics. Cabi.

Wolc, A., Stricker, C., Arango, J., Settar, P., Fulton, J. E., O'Sullivan, N. P., ... and Dekkers, J. 2011. Breeding value prediction for production traits in layer chickens using pedigree or genomic relationships in a reduced animal model. Genetics Selection Evolution 43(1): 1-9.

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56(645): 330-338.

Wright, J. 2012. Introduction to forest genetics. Elsevier.

Yim, K. B. and Noh, E. R. 1979. Study on the heritabilities of Pinus densiflora S. et Z. Journal of Korean Society of Forest Science 42(1): 74-82.

Zhao, Y., Gowda, M., Liu, W., Würschum, T., Maurer, H. P., Longin, F. H., ... and Reif, J. C. 2012. Accuracy of genomic selection in European maize elite breeding populations. Theoretical and Applied Genetics 124(4): 769-776.

Zietkiewicz, E., Rafalski, A. and Labuda, D. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20(2): 176-183.

Chapter 1. Phenotype analysis and genetic parameter estimation in progeny trials of Korean red pine

1.1. Abstract

Understanding the population is fundamental for breeding. Since genomic selection (GS) is also considerably affected by the genetic structure of the population, it is important to figure out the phenotypic and genetic characteristics of the breeding population. In this chapter, in order to identify the features of the population targeted for GS in Korean red pine, phenotypes according to regions and families were analyzed, parameters such as heritability and genetic correlation were estimated, and the breeding values to be used for GS evaluation were estimated. The ’87 open-pollinated progeny trial planted in 6 regions and ’10 control-pollinated progeny trial planted in Hwaseong were targeted as the study population. For phenotype investigation, laser scanning technology was utilized, and for genotype investigation, a recently developed 50K SNP chip of Korean red pine was used. As a result, it was found that there were phenotypic variations by region and family and an interaction between the test sites and the families. Also, it was possible to effectively capture the differences between individuals belonging to the same open-pollinated family through genomic information. In addition, the heritability estimated through genomic information showed generally low values. Therefore, it was concluded that the environment should be considered when the GS being conducted.

1.2. Introduction

Since the prediction accuracy of genomic selection (GS) largely depends on the properties of the population, understanding the phenotype, genotype and genetic parameters of the target breeding population should be the basis. The research subjects for the GS of Korean red pine were the ’87 open-pollinated progeny trial as the half-sib population and ’10 control-pollinated progeny trial as the full-sib population. The reasons for selecting the ’87 open-pollinated progeny trial are as follows. First, comparative studies would be possible because they are planted with identical family composition and design in several regions. Second, a large number of individuals could be secured. Finally, the trial forest is 32 years old which is close to the age of earning of Korean red pine (40 years in the private forest). In addition, the reason for selecting ’10 control-pollinated progeny trial is that the variation in the phenotype is expected to be large. Because, the superior clone (Kyeongbuk4, KB4) and the inferior clone (Gyeonggi1, GG1) in the general combining ability (GCA) analysis result of the open-pollinated progeny test in 2009 (NIFoS, 2009) were used as the female parents for the population.

Since GS uses statistical techniques, accurate phenotype measurement is essential. However, the use of conventional equipment such as calipers and hypsometers, which were used in the forest resource survey, has a large error depending on the skill level of the investigator. Moreover, since straightness is determined by comparison with the reference figure according to the judgment of the investigator, it is inevitably subjective. In addition, manually measuring every standing tree in the field is inefficient because it requires a lot of time and labor. Recently, as an alternative to this, laser scanning (LS) technology has been developed. LS is not only capable of consistently

generating high-precision outputs but also is environmentally friendly due to its non-destructive properties (Chen et al., 2019; Dittmann et al., 2017). In a recent study in South Korea forest, light detection and ranging (LiDAR), one of the LS technologies, showed higher statistical accuracy than the existing survey method and saved the survey time (Ko et al., 2021).

Heritability, the most important parameter in quantitative genetics, refers to the degree to which parental traits are transmitted to offspring. Heritability is expressed as the ratio of genetic variance to the phenotypic variance of a trait. Further, narrow-sense heritability is expressed as a ratio only additive genetic variance separated from genetic variance to phenotypic variances. In forestry, in addition to the individual heritability concerning the phenotype of an individual, the family heritability regarding the mean of a family is used to be taken into consideration. Heritability has been reported to have a significant impact on the prediction accuracy of GS. The prediction accuracy of GS increased as the heritability increased in loblolly pine (Resende et al., 2012). Also, according to Lian et al. (2014), the square root of the heritability has a strong correlation with the accuracy of GS. However, on the other hand, there was a report that when the size of the training population was large (1,000 or more), the effect of heritability on the accuracy of GS was relatively small compared to other factors (Grattapaglia and Resende, 2011). In other words, in the case of traits with low heritability, the decrease in accuracy could be compensated for by increasing the size of the training population or the number of observed phenotypes (Hayes et al., 2009).

In this chapter, the phenotype and genotype of open- and control-pollinated progeny trials of Korean red pine were analyzed for the purpose of understanding the characteristics of the target population prior to GS. Also, the genetic parameters that would be used for the evaluation of GS accuracy

were estimated. For phenotypes, the distribution by trait and the differences according to regions and families were mainly analyzed. The genotypes of the population were investigated using molecular markers and genetic relationship among individuals were analyzed. Then heritability and BV were estimated using an analysis of variance and a mixed model.

1.3. Materials and methods

1.3.1. Study population

(1) Half-sib population

’87 open-pollinated progeny trial was established in 1987 by the National Institute of Forest Sciences (NIFoS) in 6 regions (Taean, Chuncheon, Gongju, Kyeongju, Naju, and Wanju) for progeny test of plus trees (Figure 1-1). The seeds obtained from 49 grafted clones (20 and 29 clones of plus trees from Gangwon and Kyeongbuk province, respectively) of the clone bank located in Taean were nursed to produce 1-1 seedlings. Then seedlings were distributed to 6 regions and planted with 6 repetitions of randomized complete block design (RCBD) at intervals of 1.8m × 1.8m. The number of remained standing trees are different by region and family, and a total of 44 open-pollinated families have been preserved. A total of 3,820 trees, for which the phenotype was investigated and of which the female parent was ensured through parentage analysis using simple sequence repeat (SSR) markers, were the subject of this study (Table S1).

(2) Full-sib population

In 2010, the NIFoS established ’10 control-pollinated progeny trial in Hwaseong (Figure 1-1). The artificial cross was carried out using KB4, which showed most excellent performance in the open-pollinated progeny test, and GG1, which showed most inferior performance in the same test, as the female parents and 15 clones including these two as the male parents (NIFoS, 2009). According to the GCA from open-pollinated progeny test in 2009, half-sib families KB4 (1st), GW84 (3rd), KB33 (7th), KB5 (10th), KB20 (12th),

GW69 (13th), GW44 (17th), and Chungbuk3 (CB3) (74th) were superior out of 232 open-pollinated families. Also, KB26 (168th), GW43 (187th), GW76 (200th), GW39 (219th), GW40 (221st), GW42 (225th), and GG1 (232nd) were inferior. A total of 29 full-sib families were nurtured, and 1-2-3 seedlings were planted in 4 repetitions of the RCBD at 1.8m × 1.8m intervals. 703 trees of which phenotype was investigated were analyzed in this study (Table S2).

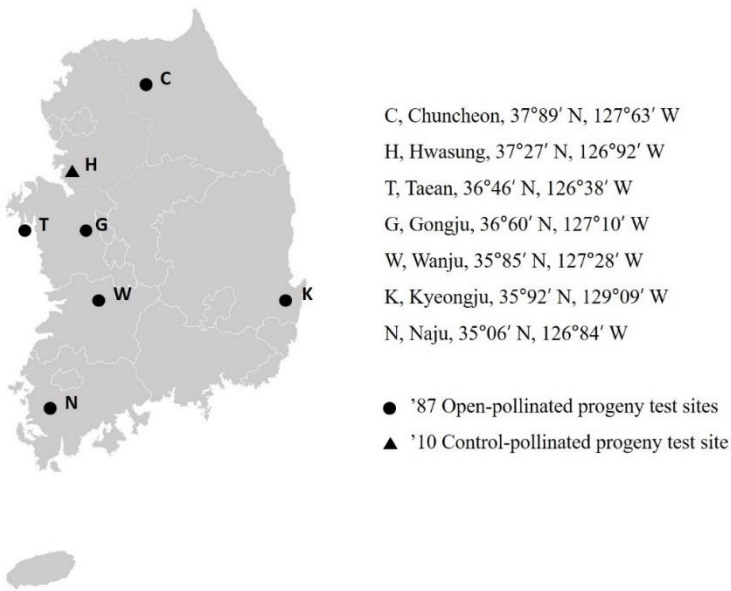

Figure 1-1. Locations of the open- and control-pollinated progeny test sites.

1.3.2. Phenotyping and statistical analysis

(1) Half-sib population

The target trait of GS is growth including diameter at breast height (DBH), height, straightness, and volume. In order to obtain the phenotypes for each individual, point cloud data from six sites of the '87 open-pollinated progeny trial was obtained using a LiDAR device (ScanStation P40, Leica) in 2017

and 2018. Then, for batch data processing, the point cloud data was separated into trees and the ground, and the ground was flattened. DBH (m) was obtained by dividing the circumference of the tree at a height of 1.2 m from the ground by the circumference ratio (π). The tree height (m) was measured as the length from the ground to the top of the canopy. The volume (m^3) of each tree was calculated as $(DBH)^2 \times$ (height). In order to calculate the straightness, an imaginary baseline was written connecting the center point of the tree trunk at the root collar and a height of 6m. Then, the distances from the baseline to the center point of the tree at heights of 0.5m, 1.5m, 2.5m, 3.5m, 4.5m, and 5.5m were calculated. The straightness was digitized by taking the negative natural log ($-\log_e$) of the standard deviation of the six distance values. In other words, the more the stem had a similar shape to the baseline, the greater the straightness was calculated.

(2) Full-sib population

Because the sizes of trees in '10 control-pollinated progeny trial were not large enough to use LiDAR yet, DBH (m) and height (m) of the individual trees were measured manually using a DBH tape and Haglof vertex hypsometer in 2019. Then the volume (m^3) was calculated as $(DBH)^2 \times$ (height).

(3) Statistical analysis

Before statistical analysis, phenotypes were standardized for each trait, and values with $Z \geq 4$ were regarded as outliers and treated as missing data. The equivariance was investigated using Levene's test, and the mean among groups was compared using Welch's F-test or non-parametric Kruskal-Wallis test depending on normality and equivariance. The post-hoc test was

performed using the Games-Howell method after Welch's F-test, and Dunn's test after the Kruskal-Wallis test. All statistical analyzes were performed using R (v3.6.2).

1.3.3. Genotyping and genomic realized relationship matrix

(1) DNA extraction

For the half-sib population, in order to investigate the genotype of each tree in six test sites, cambium was collected from tree stems and DNA was extracted using ExgeneTM Plant SV Kit (Geneall, Seoul). A total of 2,731 DNA samples were obtained and used for the study. For the full-sib population, DNA was extracted from needles of 701 trees using Plant DNA mini kit (Type B) (Onsol, Suwon).

(2) Genotyping

Single nucleotide polymorphism (SNP) probe information for genotyping was obtained using the 50K SNP chip of Korean red pine (Cheon et al., 2021). The SNP chip was developed using the genotyping-by-sequencing (GBS) information of 46 plus trees that have been tested in the '87 open-pollinated progeny trial. Also, Axiom^{TM} analysis suite (v5.0.1) program was used for SNP calling. First, sample quality was investigated and samples with a sample call rate of less than 97% or a DQC value of less than 0.82 based on probe intensity were excluded.

Then, for each SNP marker, the genotypes (AA, BB, AB, NN) were clustered based on the ratio of the intensity of the two allele probes (Figure 1-2). There are six SNP classification categories: poly high resolution (PHR), no minor homo (NMH), mono high resolution (MHR), off-target variant

(OTV), call rate below threshold (CRBT), and others. PHR represents the SNPs that are separated well so that have three genotypes. NMH also represents SNPs with well-separated genotype clusters including a homozygous genotype and a heterozygous genotype. MHR represents SNPs with one well-formed genotype cluster which is homozygous. CRBT is the type of SNPs with low call rates. OTV represents SNPs with a possible off-target cluster. Others are SNPs with more than one problematic issue.

Genotypes were selected based on marker quality. Call rate (CR) refers to the ratio of the number of genotyped samples to all samples and indicates the resolution of the genotype cluster. Fisher's linear discriminant (FLD) indicates whether genotypic clusters are formed in a narrow shape and are well separated from each other. Heterozygous strength offset (HetSO) indicates how high the strength of a heterozygous cluster (AB) is compared to two homozygous clusters (AA, BB). HetSO appears high in normal diploids and it discriminates marker specificity. Homozygote ratio offset (HomRO) is the distance at which the homozygote cluster is away from 0 on the X-axis, indicating the accuracy of clustering.

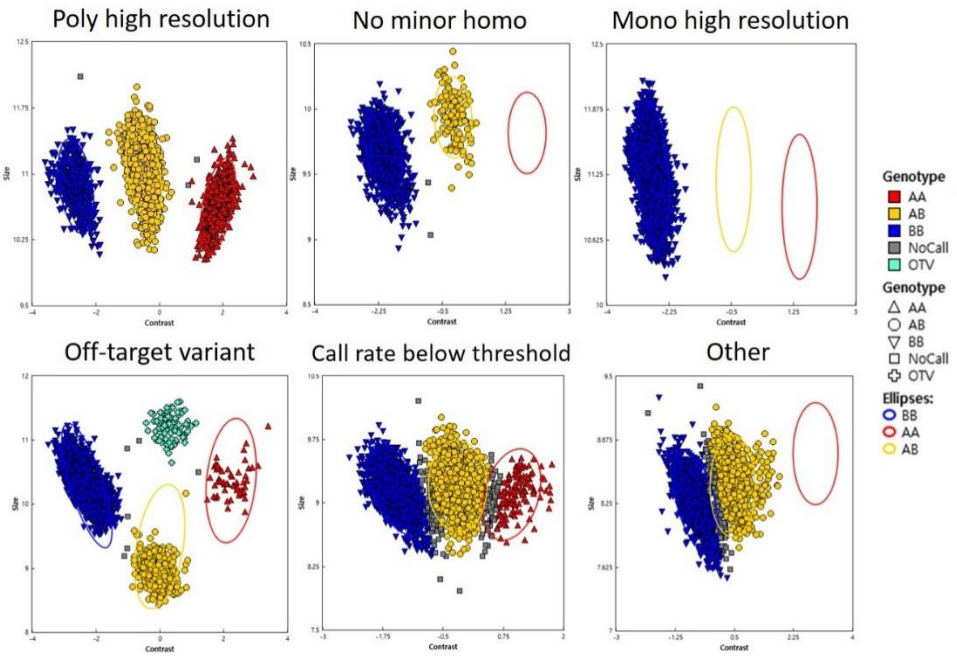

Figure 1-2. The examples of contrast plots representing six SNP genotyping cluster types.

(3) Writing GRM

Since there should be no missing data of genotype in order to write a genomic realized relationship matrix (GRM), imputation was performed using the Beagle (v5.2), which operates even without a reference genome (Browning and Browning, 2009). After, GRM was prepared according to the formula of VanRaden (2008). For comparison with GRM, the numerator relationship matrix (NRM) consisted of the relationship coefficient of two individuals was also written using pedigree information.

1.3.4. Estimation of genetic parameter

(1) Heritability for half-sib population

The mathematical linear model for analyzing the variance component and heritability of the half-sib population was as follows.

$$X_{ijk} = \mu + B_i + F_j + BF_{ij} + \varepsilon_{ijk}$$

where X_{ijk} is the phenotype of the k-th tree of the j-th family in the i-th block. μ is the overall mean, B_i is the effect of the block, F_j is the effect of family, BF_{ij} is the interaction effect of block and family, and ε_{ijk} is the error term. Individual and family heritability by site were calculated using the following formulas (Zobel and Talbert, 1984).

$$h_i^2 = \frac{4\sigma_F^2}{\sigma_w^2 + \sigma_{FB}^2 + \sigma_F^2}, \; h_f^2 = \frac{\sigma_F^2}{\sigma_w^2/NB + \sigma_{FB}^2/B + \sigma_F^2}$$

where h_i^2 is the individual heritability, h_f^2 is the family heritability, σ_F^2, σ_{FB}^2, and σ_W^2 are the variance component for family, family × block, and within plot, N is the number of individuals in iterations, and B is the number of blocks.

The mathematical linear model for the combined analysis of six sites was as follows.

$$X_{ijkl} = \mu + S_i + B_{ij} + F_k + SF_{ik} + BF_{ijk} + \varepsilon_{ijkl}$$

where X_{ijkl} is the phenotype of the l-th tree of k-th family in j-th block of the i-th site, μ is the overall mean, S_i is the effect of site, B_{ij} is the effect of the block, F_k is the effect of family, SF_{ik} is the interaction effect of site and family, BF_{ij} is the interaction effect of block and family, and ε_{ijk} is the error term. Individual and family heritability by combined analysis were calculated by the following formulas (Wright, 2012).

$$h_i^2 = \frac{4\sigma_F^2}{\sigma_w^2 + \sigma_{FB/S}^2 + \sigma_{FS}^2 + \sigma_F^2}, \; h_f^2 = \frac{\sigma_F^2}{\sigma_w^2/NBS + \sigma_{FB/S}^2/BS + \sigma_{FS}^2/S + \sigma_F^2}$$

where σ_F^2, $\sigma_{FB/S}^2$, σ_{FS}^2, and σ_W^2 are the variance components for family, family × block within site, family × site, and within plot, N is the number of individuals in the replicate, B is the number of blocks, and S is the number of sites.

(2) Heritability for full-sib population

The mathematical linear model for analyzing the variance component and heritability of the full-sib population was as follows.

$$X_{ijkl} = \mu + B_i + M_j + F_k + FM_{kj} + MB_{ji} + FB_{ki} + FMB_{kji} + E_{ijkl}$$

where X_{ijkl} is the phenotype of the j-th male parent, the k-th female parent, and the l-th tree in the i-th block. μ is the overall mean, B_i is the effect of block, M_j is the effect of male parent, F_k is the effect of female parent, FM_{ij} is the interaction effect of female and male parents, MB_{ji} is the interaction effect of male parent and block, FB_{ki} is the interaction effect of female parent and block, FMB_{kji} is the interaction effect of female and male parents and block, and ε_{ijkl} is the error term. The individual and family heritability for the full-sib population were calculated using the following formulas (Wright, 2012).

$$h_i^2 = \frac{2(\sigma_F^2 + \sigma_M^2)}{\sigma_W^2 + \sigma_{FMB}^2 + \sigma_{FM}^2 + \sigma_{MB}^2 + \sigma_{FB}^2 + \sigma_F^2 + \sigma_M^2},$$

$$h_f^2 = \frac{\sigma_F^2 + \sigma_M^2 + \sigma_{FM}^2}{\sigma_W^2/NB + \sigma_{FMB}^2/B + \sigma_{FM}^2/FM + \sigma_{FB}^2/B + \sigma_{MB}^2/B + \sigma_F^2 + \sigma_M^2}$$

where σ_F^2, σ_M^2, σ_{FM}^2, σ_{FMB}^2, and σ_W^2 are the variance components for female parent, male parent, female parent × male parent, female parent × male parent × block, and within plot, N is the number of individuals in the replicate, B is the number of blocks, F is the number of female parents, and M is the number of male parents.

(3) Heritability by mixed model

The heritability by the mixed model was obtained as the variance of NRM and GRM was divided by the total variance in best linear unbiased prediction (BLUP).

$$h^2_{NRM} = \frac{\sigma^2_{NRM}}{\sigma^2_{NRM} + \sigma^2_r}\ ,\ h^2_{GRM} = \frac{\sigma^2_{GRM}}{\sigma^2_{GRM} + \sigma^2_r}$$

The site and block were set to fixed effect in the half-sib and full-sib populations, respectively.

(4) Genetic correlation

The Type-A and Type-B genetic correlations, the degree to which family effects between two traits and between two sites coincide with each other, were calculated as follows.

$$r_{12} = \frac{\sigma_{a12}}{\sqrt{\sigma^2_{a1}\sigma^2_{a2}}}$$

where σ_{a12} is the covariance of the additive effects of different traits and sites, and σ_{a1}^2 and σ_{a2}^2 are the additive variances of each trait and site.

(5) Breeding value estimation

The estimated breeding value (EBV) was estimated by phenotypic selection and animal/individual model. In phenotypic selection, the phenotype of an individual is considered to be the most important, and the selection is conducted on an individual basis. The breeding value by phenotypic selection was estimated as follows.

$$EBV_{phenotypic\ selection} = \bar{P} + h^2(P_i - \bar{P})$$

where $\bar{P}$ is the overall phenotype mean, P_i is the individual phenotype, and h^2 is the heritability.

The individual model also assumes that each individual has different genetic abilities and responds to the environment differently. EBV by the individual model was estimated by ABLUP using all phenotypes. The linear

model is as follows.

$$Y = \beta_0 + Zu + e$$

where Y is the phenotype vector, β_0 is the intercept, Z is the random effect incidence matrix, u is the estimated vector for random effects, and e is the vector of random residual. NRM was used as a random effect. EBV by the individual model is as follows.

$$EBV_{individual\ model} = \beta_0 + Z\hat{u}$$

For all BLUPs, the remlf90 function of the R package breedR (v0.12.5) was used. The AI (average information) algorithm was used when calculating variance for heritability and genetic correlation estimation, and the EM (expectation maximization) algorithm was used for breeding value estimation (Muñoz and Sanchez, 2020).

1.4. Result and discussion

1.4.1. Phenotypes according to region and family

(1) Half-sib population

In order to figure out the phenotypic distribution, differences according to the test site and family were analyzed. In phenotype distribution analysis by site, Welch's F test for DBH, height, and straightness and Kruskal-Wallis test for volume were performed. There were very significant differences according to test sites in all traits (p-value <0.001). However, the excellent test sites were different for each trait (Figure 1-3). The DBH in Naju and Wanju, height and straightness in Kyeongju, and volume in Chuncheon and Wanju outperformed other regions. Hence, it is judged that each trait should be analyzed separately when conducting GS. In addition, regional phenotypes were grouped into three or more statistically different post-hoc groups in all traits, and in the case of height, the average of all test sites showed a significant difference. Therefore, if more than two regions are included in GS analysis, phenotype correction should be considered.

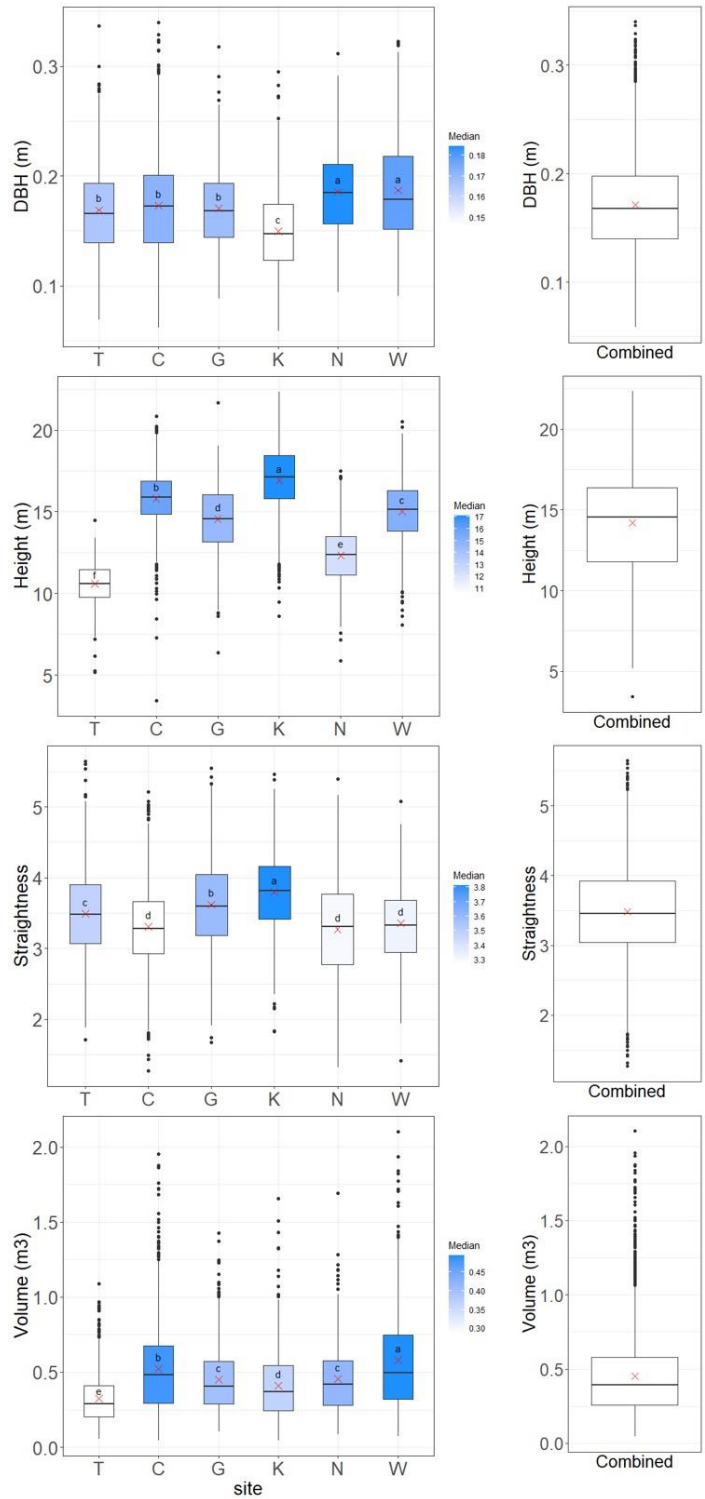

Figure 1-3. Phenotypes by open-pollinated progeny test site. Box colors and red X symbols indicate the median and mean of phenotypes respectively in each site. Alphabets indicate the Games-Howell post-hoc analysis group.

When the Kruskal-Wallis test was performed according to the family, significant differences were found between the families for all traits (p-value <0.01). This indicated that there was a genetic variation between the families. Outstanding families were different for each trait (Figure 1-4). KB89 family showed the highest values in DBH, height, and volume, and the lowest value in straightness. This is thought to be because the number of KB89 family used in the analysis was small and it was mainly distributed in one test site rather than because its genetic characteristics were distinct (Table S1). Even in the case that the KB89 family was excluded from the analysis, there were significant differences between families for all traits (p-value <0.01). Families of KB103, KB89, KB75, KB102, KB82, GW141, KB70, GW99, and GW119 ranked in the top 20% of the '87 open-pollinated progeny test conducted in 2016 (NIFoS, 2016). Of them, six families (KB103, GB89, KB75, KB102, GW141, and GW99) were confirmed to still be in the top 20%.

On the other hand, the performance rankings of open-pollinated families were different by test sites (Figure 1-5). Previous genetic testing studies conducted on the progeny test also showed that the top 20% of the superior families differed depending on the region, supporting the results of this study (NIFoS, 2016). The result suggested that there was an interaction between the test sites and the families.

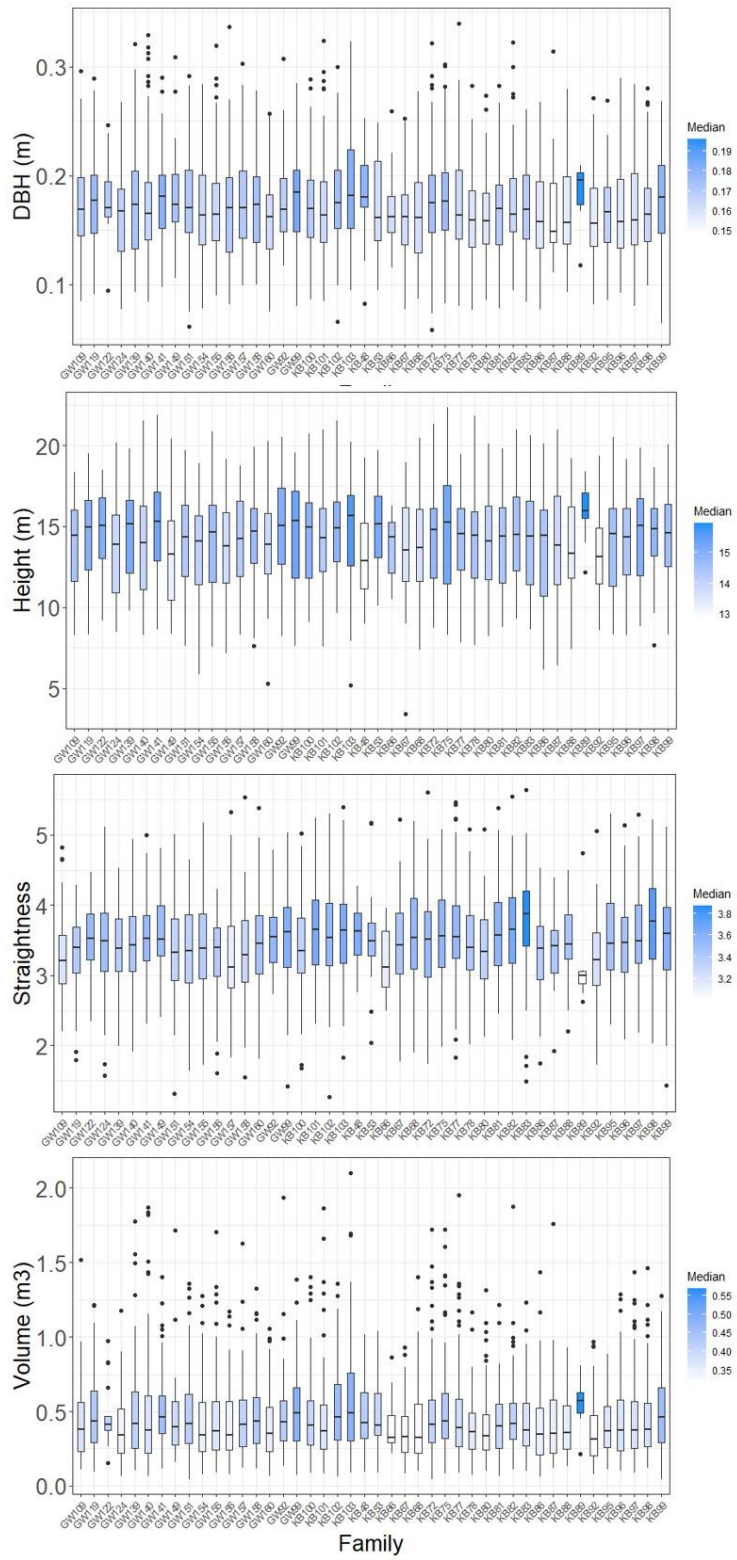

Figure 1-4. Phenotypes by open-pollinated family. Box colors indicate median phenotypes of each family.

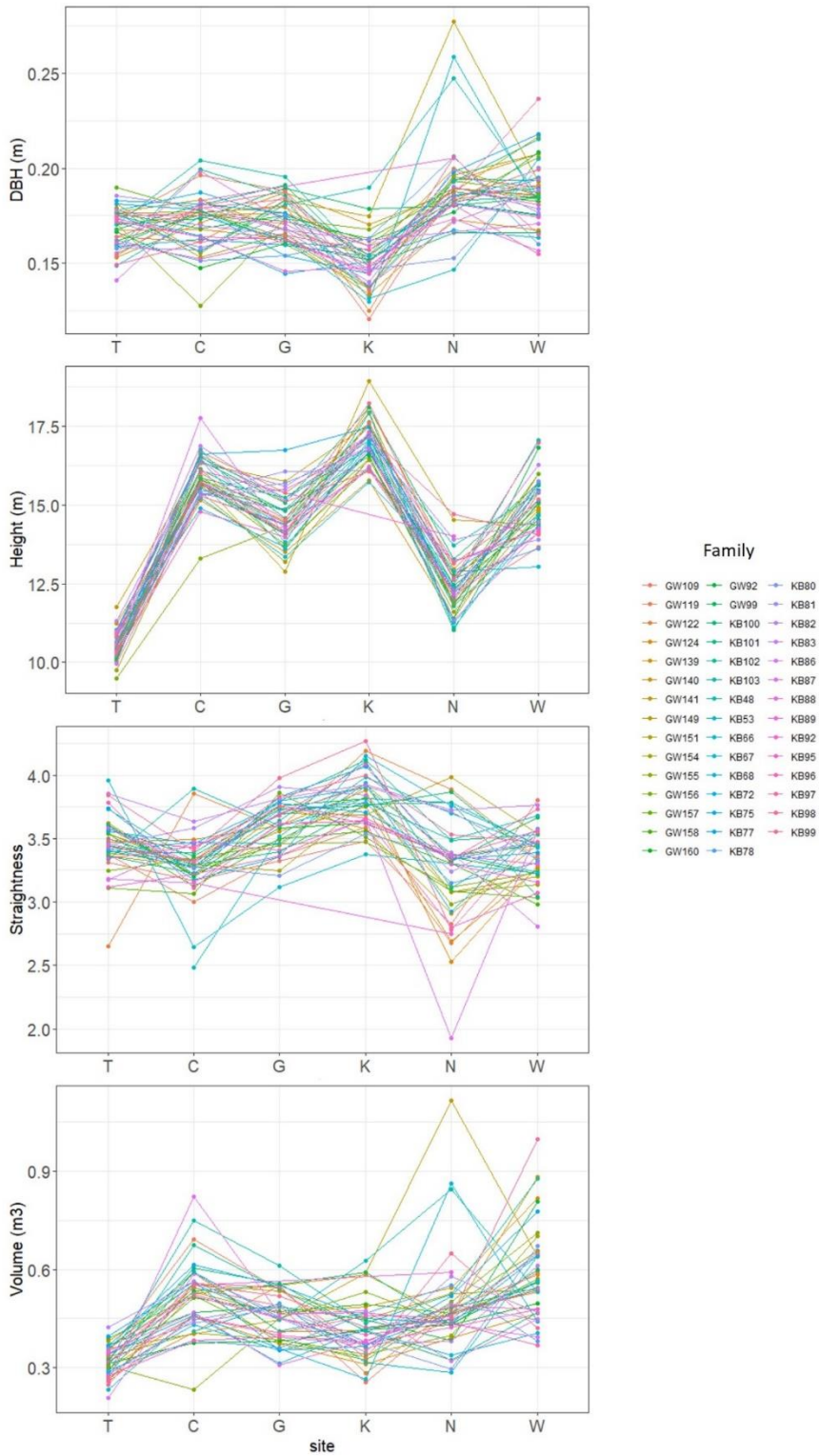

Figure 1-5. Interaction between site and family in open-pollinated progeny test.

(2) Full-sib population

As a result of the Kruskal-Wallis test, all traits showed differences in phenotype according to the family (p-value <0.001), indicating genetic variances among families were expressed on phenotype. KB4 progenies were superior to GG1 progenies (Figure 1-6), indicating the female parent effect is larger than the male parent effect.

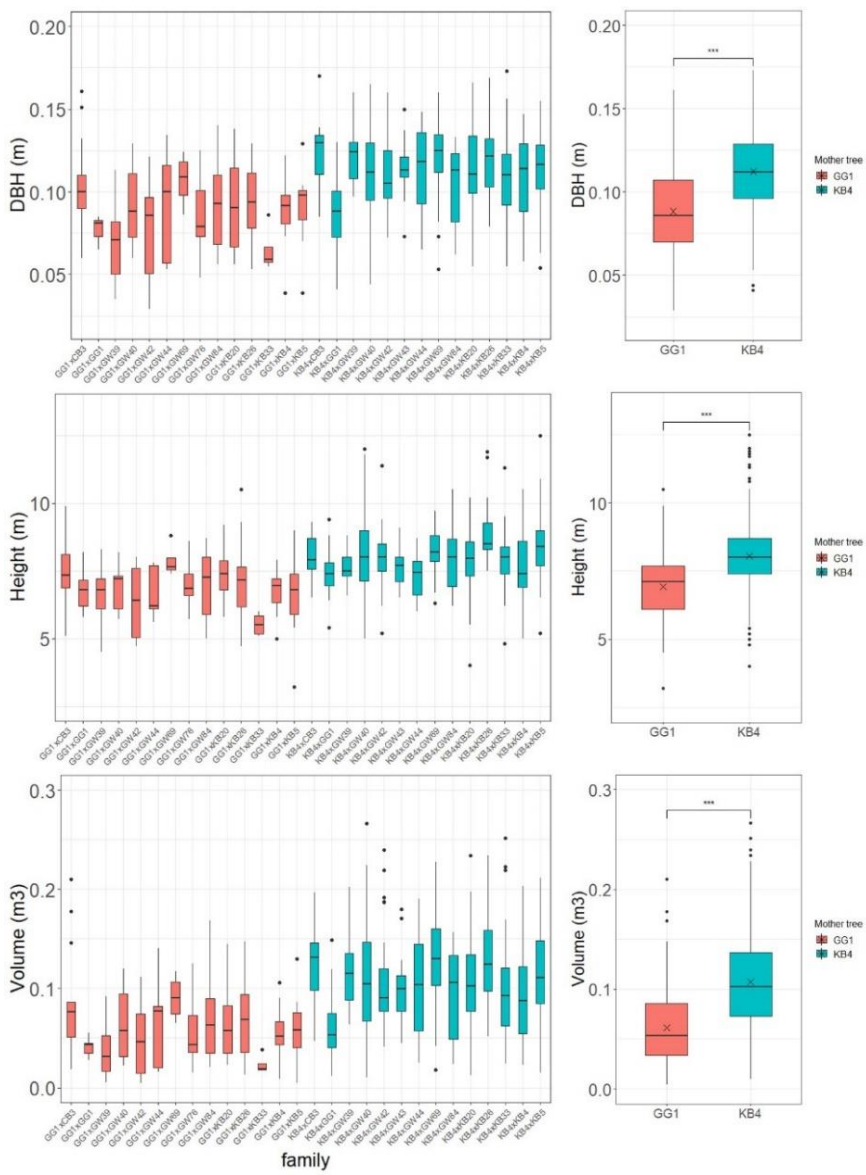

Figure 1-6. Phenotypes by control-pollinated family. Box colors indicate two female parents and the symbol X indicates the mean of phenotypes of each group. GG1, Gyeonggi1; KB4, Kyeongbuk4.

1.4.2. SNP genotype selection and genomic realized relationship matrix

(1) Half-sib population

By examining the quality of the DNA sample of the open-pollinated progeny trial, a total of 2,643 genotype data were obtained excluding 88 samples that did not meet the quality thresholds (Table 1-1, Table S1).

43,655 SNP markers were divided into 6 cluster classifications (PHR 1,120, MHR 21,063, NMH 4,439, CRBT 1,035, OTV 510 and other 15,488), and 26,622 markers of them showed high resolution (PHR, MHR and NMH). SNP genotype selection was made based on SNP quality parameters, SNP cluster classifications, and minor allele frequency (MAF). As the result of marker filtering according to the default threshold for marker quality suggested by the SNP calling program, MAF of 0.05, and classifications of high resolution, a total of 1,164 SNP markers were selected. On average, quality values improved by 1.195% points in CR, 0.199 in MAF, 0.41 in FLD, and 0.285 in HomRO, and decreased by 0.208 in HetSO (Table 1-2). In many markers with high HetSO, one or two samples were heterozygous and had high probe intensity. These markers would not have been selected based on resolution or MAF standards so that HetSO value of the selected markers was lowered.

Table 1-1. Number of samples in each stage

Test sites [a]	T	C	G	K	N	W	Total
Phenotyped	784	947	679	572	445	393	3,820
Sampled	656	735	473	391	249	227	2,731
Genotyped	609	726	456	380	247	225	2,643

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table 1-2. The average quality of 1,164 SNP markers selected in the open-pollinated population

[a] CR (%)	[b] MAF	[c] FLD	[d] HetSO	[e] HomRO
98.971	0.232	5.694	0.193	1.339

[a] CR, call rate [b] MAF, minor allele frequency [c] FLD, Fisher's linear discriminant [d] HetSO, heterozygous strength offset [e] HomRO, homozygous ratio offset

Using the selected markers, the GRM of the half-sib population was prepared and compared with the NRM (Figure 1-7). In contrast to NRM which had the same relationship coefficient when pedigree information is same, GRM showed different relationship coefficients (Figure 1-7a and b). This is because the numerator relationship coefficient is an expected value, whereas the genomic relationship coefficient is a realized value derived from actual marker information (Isik et al., 2017). Another reason is that the kinship coefficient was calculated assuming that the pollen trees are different in an open-pollinated family, but there is a possibility that family members are genetically closer depending on the pollen trees (Askew and El-Kassaby, 1994). Also, if the open-pollinated families are different, it is assumed that there is no relationship, but depending on the relation between the female parent, the progenies could share the genes.

The genomic relationship coefficient between half-sib showed a higher median and mean than the genomic relationship coefficient between individuals of different families, indicating that the numerator relationship coefficient and the genomic relationship coefficient were consistent (Figure 1-7c). However, the average of the genomic relationship coefficient between half-sib did not reach the theoretically expected value of 0.25, which had been also observed in the eucalyptus GS study (Tan et al., 2017). Although the authors explained that it was the result of NRM inaccuracy due to pedigree

labeling error, in this study, the pedigree information was judged to be relatively accurate because the female parent information was confirmed by parentage analysis using the SSR markers. On the other hand, although the use of GBS allows the identification of whole-genome SNPs, the high rate of genotyping errors and data missing poses a problem, potentially lowering the marker density (Wang et al., 2020). According to this, it is possible that the SNP markers did not reflect all genomic information of each individual in this study. Therefore, it is necessary to analyze which marker set is advantageous to obtain high prediction accuracy by comparing the SNP genotype selection method and the number of markers in the stage of GS model training.

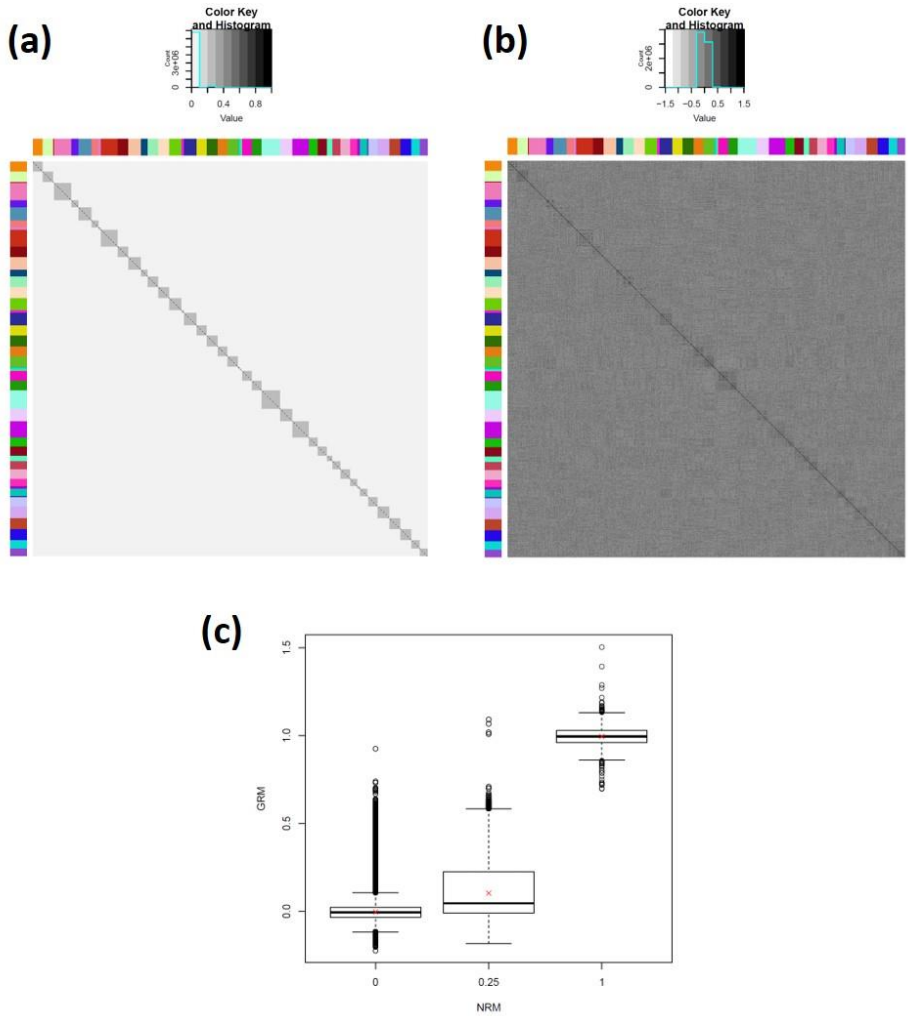

Figure 1-7. Heatmaps of coefficient of (a) numerator relationship matrix and (b) genomic realized relationship matrix ordered by open-pollinated family and (c) distribution of GRM coefficients according to their corresponding NRM coefficients. Symbol X indicates the mean of the genomic realized relationship coefficient.

(2) Full-sib population

In the control-pollinated progeny trial, 691 samples were used for genotyping excluding 10 samples that did not pass the sample quality test (Table S2). SNP genotypes were classified as 986 PHR, 23,385 MHR, 2,277 NMH, 1,133 CRBT, 394 OTV, and 15,480 other, including 26,648 high-resolution genotypes. This is similar to the number in the half-sib population, which clarified that the SNP probes were effectively amplified even in a population with parents different from the plus trees used for the development of the SNP chip. In addition, some of the SNP markers which had a low resolution in the half-sib population showed high resolution in the full-sib population, and the numbers of genotypes by the same SNP marker were different in the two populations. The result indicated that the two populations had different genetic structures such as allele frequency.

When the same criteria for selecting SNP markers as in the half-sib population were applied, a total of 1,277 markers were selected in the full-sib population. Among them, 875 markers were commonly included in both populations. The quality of the selected markers improved by 0.356% points in CR, 0.210 in MAF, 1.196 in FLD, and 0.217 in HomRO (Table 1-3). HetSO decreased by 0.039 as in the half-sib population.

Table 1-3. The average quality of 1,277 SNP markers selected in the control-pollinated population

[a] CR (%)	[b] MAF	[c] FLD	[d] HetSO	[e] HomRO
99.102	0.233	6.162	0.176	1.225

[a] CR, call rate [b] MAF, minor allele frequency [c] FLD, Fisher's linear discriminant [d] HetSO, heterozygous strength offset [e] HomRO, homozygous ratio offset

Using the selected markers, the GRM of the full-sib population was written and compared with the NRM (Figure 1-8). KB4 × GG1 of GRM was found to be closer to the other progenies of GG1 than the other progenies of KB4, which was thought to be regarding the result that the phenotype of KB4 × GG1 showed small DBH and volume (Figure 1-6). Same as in the case of the half-sib population, the mean genomic relationship coefficient was lower than the expected value (Figure 1-8c). In particular, selfing progenies have been expected to have a higher relationship coefficient (0.625) than full siblings (0.5), but the mean of the genomic relationship coefficients in GRM was rather low. In this regard, pollen contamination or pedigree labeling errors might be suspected.

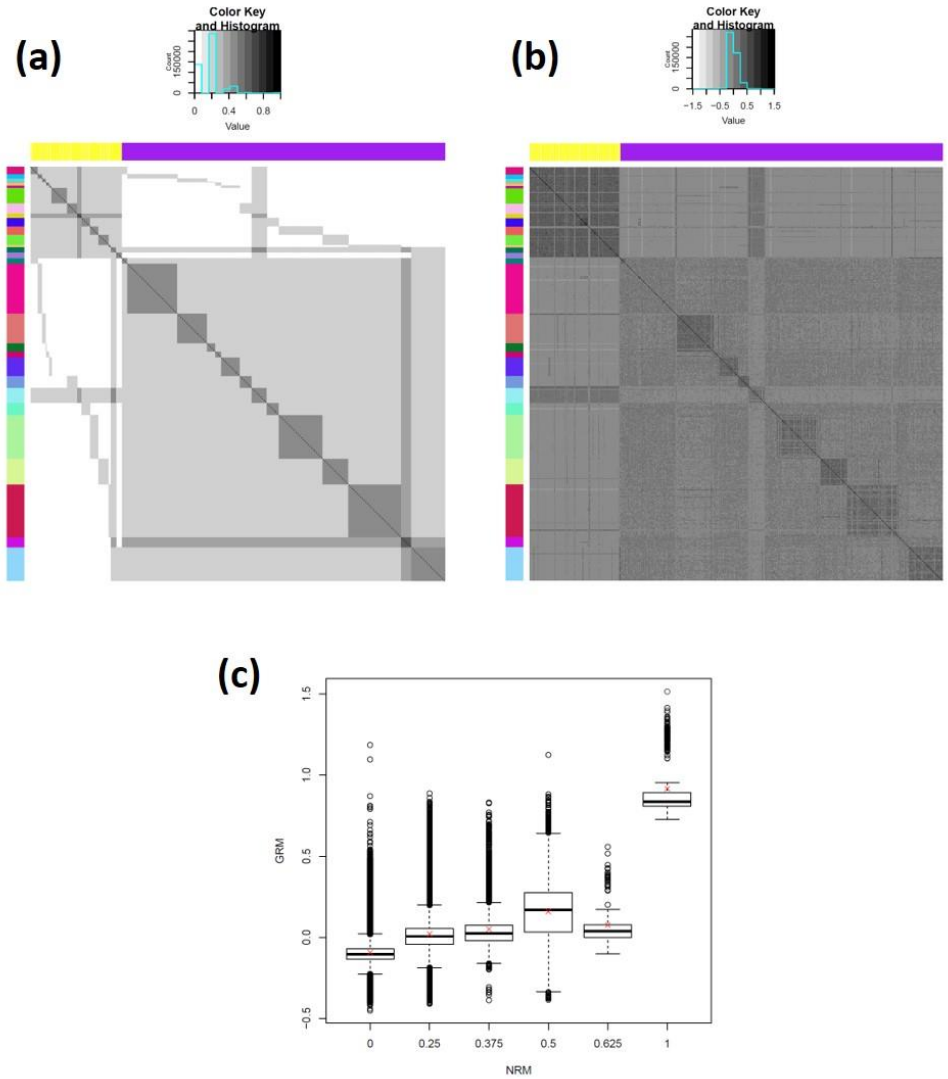

Figure 1-8. Heatmaps of coefficient of (a) numerator relationship matrix and (b) genomic realized relationship matrix ordered by control-pollinated family and mother tree and (c) distribution of GRM coefficients according to their corresponding NRM coefficients. Symbol X indicates the mean of the genomic realized relationship coefficient.

1.4.3. Heritability estimated by ANOVA and mixed model

Traditionally, analysis of variance (ANOVA) has been generally used to divide phenotypic variance into variance components by familial and environmental factors and to estimate heritability using these. However, the formula for estimating heritability by ANOVA assumes balanced data with the same number of replicates in each family and block. At the time of establishment, the progeny trials were designed with RCBD for genetic testing, but as they have been lasted until now, many trees have died due to adaptability, competition, and natural disasters. In such unbalanced data with a different number of observations for each family and block, it is appropriate to use a mixed model rather than ANOVA (Isik et al., 2017). Therefore, ANOVA was used for comparison with previous literature and estimation of family heritability, and a mixed model was used for estimation of individual heritability, genetic correlation, and breeding values.

(1) ANOVA in half-sib population

Individual and family heritability were estimated with the variance components in each test site of the half-sib population (Table 1-4). The heritabilities were generally as low as 0.5 or less, and there was a considerable difference depending on the test site and trait. In Taean, Chuncheon, and Gongju, the heritability of height was the highest, and in Kyeongju, Naju, and Wanju, the heritability of straightness was high. The heritability of the Wanju was calculated as 0 because the variance for family was extremely small. Genetic factors are not properly expressed in the poor locational condition (Falconer and Mackay, 1983). Actually, the test site in Wanju has a steep slope and the orientation of the slope is different in each block. Therefore, the

reason why the variance of family in Wanju was very small was judged that the trees grew in such an unfavorable environment. Moreover, it was thought that the severe inclination might have increased the error during the phenotype measurement using LiDAR. In a study estimating the family heritability in the open-pollinated progeny test of Korean red pine planted in 1981, the heritabilities in Chuncheon and Gongju were 0.17 and 0.27 for DBH and 0.37 and 0.47 for height, respectively (Han et al., 2007). The results of both traits showing high heritability in Gongju were consistent with the result of this study.

Individual and family heritability were obtained by combined analysis of six test sites (Table 1-5). The family heritability was significantly higher than the heritability analyzed for each test site. As the results of the study of heritability in the '81 open-pollinated progeny trial, the combined family heritability in Chuncheon, Gongju, and Naju was 0.36, 0.25, and 0.37 for DBH, height, and volume index, respectively (Han et al., 2007). In comparison with this, the family heritability of this study was relatively high. In particular, the high family heritability of 0.733 for height indicated that an excellent family could be selected by the family unit. However, since individual heritabilities were still low, it is difficult to expect a high improvement when selecting by individual units.

Table 1-4. Variance components, individual heritability, and family heritability in each site

Site	Estimates	DBH	Height	Straightness	Volume
Taean	σ_F^2	0.000014	0.055826	0.007712	0.000538
	σ_{FB}^2	0.000091	0.115826	0.007475	0.001232
	σ_w^2	0.00149	1.02	0.3514	0.02539
	h_i^2	0.036	0.187	0.084	0.079
	h_f^2	0.134	0.435	0.259	0.260
Chuncheon	σ_F^2	0.000049	0.141089	0.005646	0.002014
	σ_{FB}^2	0.00016	0.331017	0	0.011462
	σ_w^2	0.00188	2.28	0.3473	0.0743
	h_i^2	0.094	0.205	0.064	0.092
	h_f^2	0.304	0.471	0.258	0.276
Gongju	σ_F^2	0.000044	0.285873	0.013661	0.002144
	σ_{FB}^2	0.00009	0.800443	0	0.004069
	σ_w^2	0.001158	2.5	0.4057	0.04084
	h_i^2	0.135	0.319	0.130	0.182
	h_f^2	0.342	0.503	0.362	0.408
Kyeongju	σ_F^2	0.000006	0.028073	0.00797	0.000103
	σ_{FB}^2	0.000198	0.349204	0.037397	0.005596
	σ_w^2	0.001097	3.36	0.2905	0.0445
	h_i^2	0.019	0.030	0.095	0.008
	h_f^2	0.053	0.089	0.231	0.025
Naju	σ_F^2	0.000003	0.016762	0.03441	0.000706
	σ_{FB}^2	0.000068	0.477714	0.029347	0.003369
	σ_w^2	0.00151	2.46	0.4295	0.0479
	h_i^2	0.008	0.023	0.279	0.054
	h_f^2	0.023	0.055	0.429	0.134
Wanju	σ_F^2	0.000018	0	0.011869	0.001235
	σ_{FB}^2	0.000187	0.895983	0	0.021689
	σ_w^2	0.002111	3.26	0.285	0.1015
	h_i^2	0.030	0	0.160	0.040
	h_f^2	0.074	0	0.310	0.088

Table 1-5. Variance components, individual heritability, and family heritability estimated by the combined analysis of variance

Estimates	DBH	Height	Straightness	Volume
σ_F^2	0.000022	0.156513	0.011312	0.001369
σ_{FS}^2	0.000015	0	0.005149	0.000599
σ_{FB}^2	0.000174	1.024447	0.001685	0.010264
σ_w^2	0.00154	2	0.352	0.053
h_i^2	0.050	0.197	0.122	0.084
h_f^2	0.427	0.733	0.639	0.545

(2) ANOVA in full-sib population

Individual and family heritability were calculated for the full-sib family in the control-pollinated progeny trial (Table 1-6). The individual heritability was 0.558~0.598, and the family heritability was 0.888~0.962, which was quite high. The variance for the female parent (σ_F^2) is about 5 to 17 times higher than the variance for the male parent (σ_M^2), suggesting that the effect of the female parent on the phenotype of the full-sib progeny was larger than that of the male parent.

Table 1-6. Variance components and individual heritability and family heritability of the full-sib progeny test

Estimates	DBH	Height	Volume
σ_F^2	0.000186	0.441239	0.000738
σ_M^2	0.000034	0.025563	0.000149
σ_{FM}^2	0.000005	0.043944	0.000016
σ_{FB}^2	0.000008	0.041915	0
σ_{MB}^2	0	0.013521	0.000013
σ_{FMB}^2	0.000037	0.032113	0
σ_w^2	0.00052	0.97	0.00205
h_i^2	0.558	0.595	0.598
h_f^2	0.888	0.962	0.924

(3) Mixed model

Narrow-sense heritability was estimated in both the half-sib population and full-sib population by the mixed model (Table 1-7). The heritability of the half-sib population was 0.000-0.723, showing a large difference by test sites and traits. Generally, the heritability of height was the highest. The heritability of the combined analysis was estimated to be low between 0.059 and 0.146 and showed a difference within 0.1 from the results of the ANOVA (Table 1-5). The heritability of the full-sib population was 0.206~0.381, which was higher than that of the half-sib population. Also, the heritability of the full-sib family was significantly reduced compared to the results of the ANOVA (Table 1-6).

On the other hand, the heritability estimated by NRM and GRM showed a tendency to increase and decrease together depending on the test sites and traits, but there was a difference in the values. In the case of regional analysis of the half-sib population, the heritability of GRM was mostly higher than that of NRM (Table 1-7). The reason was thought to be that the relationship coefficient of NRM might not reflect the actual relationship. The kinship coefficient of open-pollinated family progenies was set to 0.25 assuming that the family is half-sib (Wright, 1922). This assumption holds only when both parents are not self-bred or related and the mother tree was pollinated with a sufficiently large effective pollen number (Askew and El-Kassaby, 1994). However, in reality, the kinship relationship between open-pollinated siblings might be higher because the relationships would include not only self-pollination, self-half, half-sib, and full-sib, but also the case that the female and male parents have a common ancestor (Askew and El-Kassaby, 1994).

Table 1-7. Narrow-sense heritability by the mixed model using NRM and GRM in two progeny tests

Population	Estimates	DBH	Height	Straightness	Volume
Half-sib T	h_{NRM}^2	0.103 (0.083)	0.249 (0.124)	0.078 (0.092)	0.143 (0.092)
	h_{GRM}^2	0.186 (0.106)	0.723 (0.103)	0.146 (0.105)	0.276 (0.112)
C	h_{NRM}^2	0.195 (0.101)	0.380 (0.148)	0.045 (0.060)	0.208 (0.103)
	h_{GRM}^2	0.420 (0.099)	0.393 (0.098)	0.227 (0.095)	0.463 (0.105)
G	h_{NRM}^2	0.407 (0.187)	0.457 (0.200)	0.245 (0.144)	0.413 (0.190)
	h_{GRM}^2	0.354 (0.151)	0.720 (0.138)	0.356 (0.146)	0.393 (0.150)
K	h_{NRM}^2	0.055 (0.115)	0.000 (0.000)	0.327 (0.193)	0.001 (0.003)
	h_{GRM}^2	0.051 (0.105)	0.564 (0.197)	0.405 (0.171)	0.004 (0.091)
N	h_{NRM}^2	0.068 (0.155)	0.132 (0.179)	0.366 (0.277)	0.043 (0.151)
	h_{GRM}^2	0.546 (0.228)	0.642 (0.215)	0.395 (0.291)	0.519 (0.225)
W	h_{NRM}^2	0.070 (0.188)	0.108 (0.212)	0.000 (0.003)	0.206 (0.235)
	h_{GRM}^2	0.173 (0.175)	0.105 (0.187)	0.286 (0.245)	0.173 (0.187)
Combined	h_{NRM}^2	0.074 (0.031)	0.146 (0.049)	0.113 (0.042)	0.108 (0.040)
	h_{GRM}^2	0.066 (0.017)	0.104 (0.020)	0.059 (0.017)	0.082 (0.018)
Full-sib	h_{NRM}^2	0.362 (0.157)	0.359 (0.154)	-	0.381 (0.163)
	h_{GRM}^2	0.251 (0.059)	0.206 (0.056)	-	0.251 (0.057)

Standard error of heritability estimation in parenthesis

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Conversely, in the case of the full-sib population, the heritability of NRM was found to be higher than that of GRM. This was consistent with the results in many other pines and spruces (Lenz et al., 2017; Chen et al., 2018; Ukrainetz and Mansfield, 2020a; Beaulieu et al., 2020). If only pedigree information is used, the variation of individuals belonging to the same family is not considered, leading to the overestimation of additive genetic variance (Askew and El-Kassaby, 1994; El-Dien et al., 2016). Because, the variance due to the non-additive effects, including the dominant and epistatic effects, cannot be accurately estimated with the limited family structure (Muñoz et al., 2014). Therefore, it was judged that the heritability estimated by GRM is more accurate, and it can be referred to when evaluating the GS efficiency.

The Type-A genetic correlation indicating the similarity of the female parent effect for the four traits was investigated (Table 1-8). There were significant genetic correlations between the DBH, height, and volume. Straightness had a genetic correlation only with height. Since the half-sib population was planted in several regions, the Type-B genetic correlation indicating whether the female parent effects were similar between environments was also investigated (Table 1-9). The correlations between the test sites were found to be very low, which was consistent with the result that the excellent families for each test site were different in the phenotypic analysis (Figure 1-5).

Table 1-8. Type-A genetic correlation between traits

	DBH	Height	Straightness	Volume
DBH	-			
Height	0.632***	-		
Straightness	0.028	0.417**	-	
Volume	0.966***	0.740***	0.122	-

** 0.001<p-value<0.01, *** p-value<0.001

Table 1-9. Type-B genetic correlation between the open-pollinated progeny test sites

DBH						
	Taean	Chuncheon	Gongju	Kyeongju	Naju	Wanju
Taean	-					
Chuncheon	0.123	-				
Gongju	-0.071	0.472**	-			
Kyeongju	-0.183	0.062	0.241	-		
Naju	0.062	0.088	-0.030	0.198	-	
Wanju	-0.057	0.197	-0.020	0.041	0.031	-
Height						
	Taean	Chuncheon	Gongju	Kyeongju	Naju	Wanju
Taean	-					
Chuncheon	0.307*	-				
Gongju	0.283	0.188	-			
Kyeongju	-0.078	0.211	0.203	-		
Naju	0.296	0.433**	0.335*	0.306	-	
Wanju	0.050	0.322*	0.061	0.114	-0.007	-
Straightness						
	Taean	Chuncheon	Gongju	Kyeongju	Naju	Wanju
Taean	-					
Chuncheon	0.207	-				
Gongju	0.233	0.351*	-			
Kyeongju	0.210	0.136	0.158	-		
Naju	0.190	0.157	0.074	0.076	-	
Wanju	0.250	0.387**	0.278	-0.126	-0.005	-
Volume						
	Taean	Chuncheon	Gongju	Kyeongju	Naju	Wanju
Taean	-					
Chuncheon	0.181	-				
Gongju	-0.042	0.515***	-			
Kyeongju	-0.207	0.185	0.331*	-		
Naju	0.171	0.191	0.090	0.205	-	
Wanju	-0.023	0.278	0.044	0.131	0.017	-

* 0.01<p-value<0.05, ** 0.001<p-value<0.01, *** p-value<0.001

1.4.4. Breeding value estimated by phenotypic selection and individual model

(1) Half-sib population

The breeding value by phenotypic selection (EBV_{ps}) was estimated to be 0.126~0.255 m for DBH, 6.662~19.949 m for height, 2.578~4.238 for straightness, and 0.268~0.874 m^3 for volume in half-sib population. Since EBV_{ps} is calculated as a linear equation for the individual phenotype, its value varies depending on the heritability used as the linear coefficient, but the correlation coefficient with the phenotype always equals 1 regardless of heritability.

The breeding value by the individual model (EBV_{im}) was estimated using BLUP. The EBV_{im} of the half-sib population ranged 0.134~0.219 m for DBH, 9.306~17.500 m for height, 2.627~4.348 for straightness, and 0.254~0.915 m^3 for volume (Figure 1-9). Since the BV was estimated within the region, each site showed different linear and correlation coefficients in relation to the phenotype. For example, the range of EBV_{im} was very narrow in the height of Kyeongju and the straightness of Wanju, because the narrow-sense heritability by NRM was close to 0 (Table 1-7).

(2) Full-sib population

EBV_{ps} were calculated to be 0.087~0.124 m for DBH, 6.854~8.769 m for height, and 0.074~0.140 m^3 for volume in the full-sib population. The EBV_{im} of the full-sib population was 0.072~0.132 m for DBH, 6.284~9.147 m for height, and 0.038~0.149 m^3 for volume (Figure 1-10).

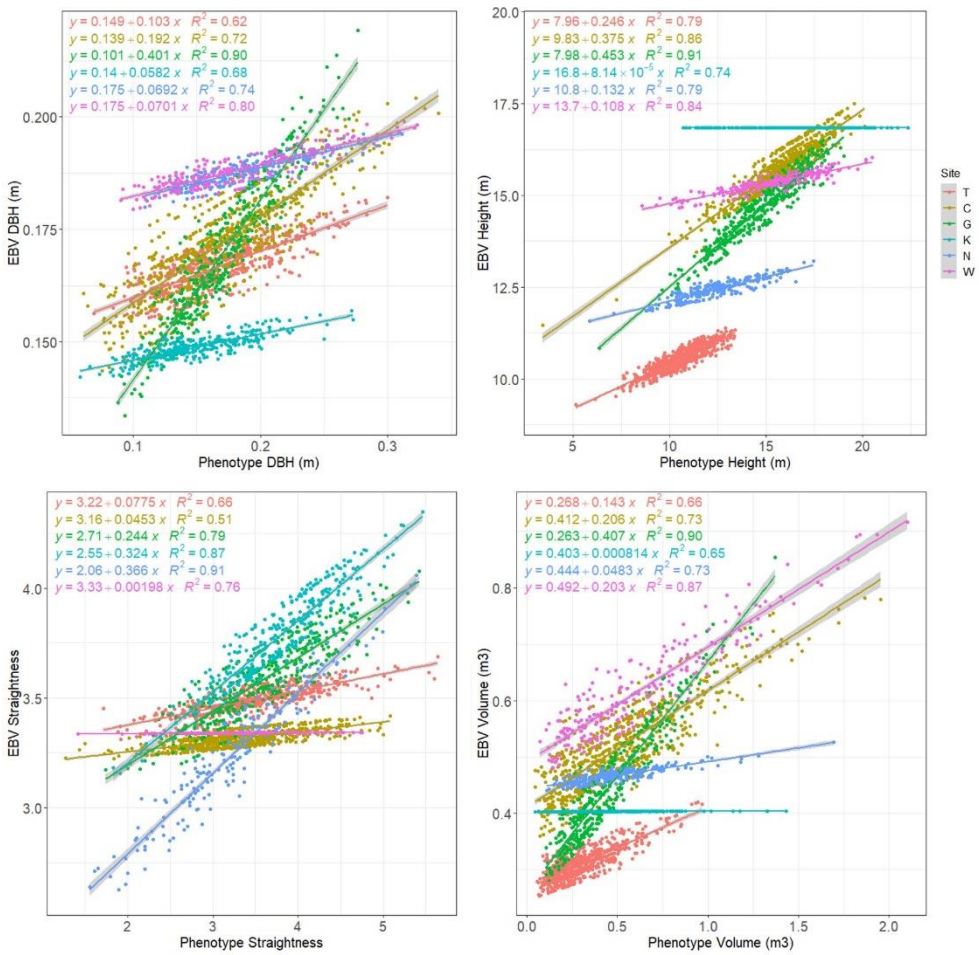

Figure 1-9. Breeding values estimated using the individual model in the open-pollinated progeny test by site

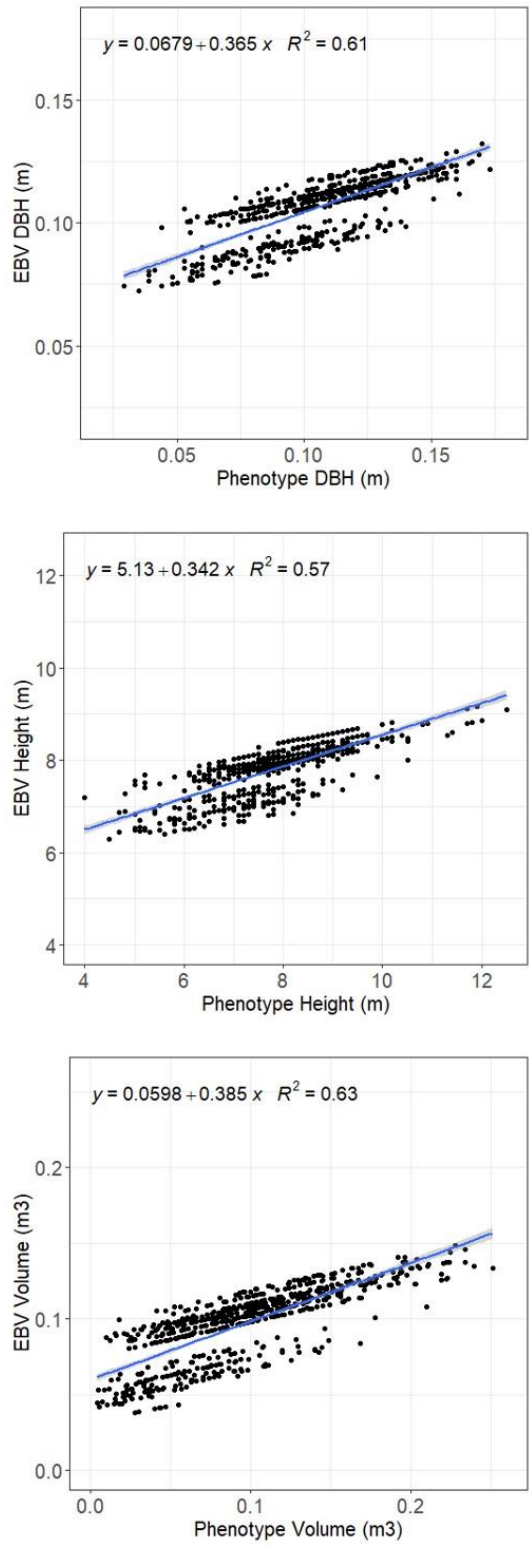

Figure 1-10. Breeding values estimated using the individual model in the control-pollinated progeny test

1.5. Conclusion

The phenotypic and genetic features of the half-sib population and full-sib population, the target populations of GS, were identified. There were genetic variations between families in both populations. Also, the interaction between the genotype and the environment of the half-sib population was observed. In addition, the heritabilities were estimated to be low for four traits in both populations. Thus, it was concluded that the environment and heritability needed to be considered in GS.

References

Askew, G.R. and El-Kassaby, Y.A. 1994. Estimation of relationship coefficients among progeny derived from wind-pollinated orchard seeds. Theoretical and Applied Genetics 88(2): 267-272.

Beaulieu, J., Nadeau, S., Ding, C., Celedon, J. M., Azaiez, A., Ritland, C., ... and Bousquet, J. 2020. Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. Evolutionary applications 13(10): 2704-2722.

Browning, B. L. and Browning, S. R. 2009. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. The American Journal of Human Genetics 84(2): 210-223.

Chen, Z. Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., ... and Wu, H. X. 2018. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC genomics 19(1): 1-16.

Chen, S., Liu, H., Feng, Z., Shen, C. and Chen, P. 2019. Applicability of personal laser scanning in forestry inventory. Plos One 14(2), e0211392.

Cheon, K. S., Kang, H. I., Park, Y. W., Song, J. H., Kim, I. S. and Shim, D. 2021. Development of SNP chip for Genomic Selection of Korean Red Pine (Pinus densiflora) Trees. Proceedings of The Korean Society of Breeding Science, 406.

Dittmann, S., Thiessen, E. and Hartung, E. 2017. Applicability of different non-invasive methods for tree mass estimation: A review. Forest Ecology and Management 398, 208-215.

El-Dien, O. G., Ratcliffe, B., Klápště, J., Porth, I., Chen, C. and El-Kassaby, Y. A. 2016. Implementation of the realized genomic relationship matrix to open-pollinated white spruce family testing for disentangling additive from nonadditive genetic effects. G3: Genes, Genomes, Genetics 6(3): 743-753.

Falconer, D. S. and Mackay, T. F. 1983. Quantitative genetics. Longman.

Grattapaglia, D. and Resende, M. D. 2011. Genomic selection in forest tree breeding. Tree Genetics & Genomes 7(2): 241-255.

Han, S. U., Oh, C. Y., Kim, C. S., Kim, Y. J., Kang, K. N. and Lee, S. M. 2007. Time trends for genetic parameters of growth traits in open-pollinated progenies of Pinus densiflora. Korea Journal of Breeding Science 39(4): 457-463.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J. and Goddard, M. E. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of dairy science 92(2): 433-443.

Isik, F., Holland, J. and Maltecca, C. 2017. Genetic data analysis for plant and animal breeding (Vol. 400). Cham, Switzerland: Springer International Publishing.

Ko, C., Lee, S., Yim, J., Kim, D. and Kang, J. 2021. Comparison of Forest Inventory Methods at Plot-Level between a Backpack Personal Laser Scanning (BPLS) and Conventional Equipment in Jeju Island, South Korea. Forests 12(3): 308.

Lenz, P., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M. and Bousquet, J. 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC genomics 18(1): 1-17.

Lian, L., Jacobson, A., Zhong, S. and Bernardo, R. 2014. Genomewide prediction accuracy within 969 maize biparental populations. Crop Science 54(4): 1514-1522.

Muñoz, P. R., Resende Jr, M. F., Gezan, S. A., Resende, M. D. V., de Los Campos, G., Kirst, M., ... and Peter, G. F. 2014. Unraveling additive from nonadditive effects using genomic relationship matrices. Genetics 198(4): 1759-1768.

Muñoz, F., and Sanchez, L. 2020. breedR: Statistical Methods for Forest Genetic Resources Analysts. R package version 0.12-5. https://github.com/famuvie/breedR

National Institute of Forest Science (NIFoS). 2009. Genetic test of timber species. Research Reports no. 09-12. Seoul, Korea.

National Institute of Forest Science (NIFoS). 2016. Selection and genetic test for breeding of main timber species. Research Reports no. 16-38. Seoul, Korea.

Resende, M. D., Resende Jr, M. F., Sansaloni, C. P., Petroli, C. D., Missiaggia, A. A., Aguiar, A. M., ... and Grattapaglia, D. 2012. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing

heritability and accelerating breeding for complex traits in forest trees. New Phytologist 194(1): 116-128.

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B. and Ingvarsson, P. K. 2017. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology 17*(1): 1-15.

Ukrainetz, N. K. and Mansfield, S. D. 2020a. Prediction accuracy of single-step BLUP for growth and wood quality traits in the lodgepole pine breeding program in British Columbia. Tree Genetics & Genomes 16(5): 1-13.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11): 4414-4423.

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., ... and Zhang, X. 2020. Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. Scientific RepoRtS 10(1): 1-12.

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56(645): 330-338.

Wright, J. 2012. Introduction to forest genetics. Elsevier.

Zobel, B., and Talbert, J. 1984. Applied forest tree improvement. John Wiley & Sons.

Chapter 2. Training of genomic selection model in an open-pollinated progeny trial of Korean red pine

2.1. Abstract

Genomic selection (GS) is the method of estimating breeding values based on all effects of quantitative trait loci (QTLs) using molecular markers. The prediction accuracy of GS, which decides the response to selection, is affected by several factors such as the population size, the number of markers, and trait heritability. In this chapter, the prediction accuracies of GS for growth characteristics (DBH, height, straightness, and volume) in Korean red pine were compared under various conditions in the open-pollinated progeny trial. Also, the selection efficiency of GS was evaluated compared to the traditional selection methods.

As the result, the predictive accuracy was the highest when the model was trained using GBLUP with markers of minor allele frequency 0.05 or more. The accuracy was 0.164~0.498 and the predictive ability was 0.018~0.441 in the within-region scenario. The predictive ability of GBLUP against that of ABLUP was 0.86~5.10 and against the square root of heritability was 0.19~0.76, indicating the GS of Korean red pine was as efficient as GS in previous studies. In the combined region scenario, the accuracy was increased while the predictive ability was decreased compared to the mean of results in each site. The response to the selection of GS was higher than that of traditional selections in the aspect of annual genetic gain. Therefore, the trained GS model was concluded to be effective compared to traditional breeding in Korean red pine.

2.2. Introduction

The essential stage of genomic selection (GS) is to train the model that estimates the effect of all markers. This stage includes the optimization of the model that shows the best prediction efficiency and evaluation of the model as a selective breeding method.

The efficiency of GS is evaluated through prediction accuracy. According to simulation and experimental studies, the prediction accuracy of GS in forest trees is affected by the linkage disequilibrium (LD) range between the marker and quantitative trait locus (QTL), the size of the effective population (N_e), the marker density, the size of the training population, the genetic relevance between the training population and the test population, the genetic structure of the trait (number of loci and effect size), and the trait heritability (Thistlethwaite et al., 2019). These factors affecting GS are interconnected and interdependent. Among these, the heritability and genetic structure of traits are out of the control of the breeder, while others could be controlled in the breeding programs (Lebedev et al., 2020).

The appropriate way to evaluate the prediction accuracy of models is to fit a prediction model with a subset of population and then predict values for the validation set. Cross-validation (CV) is a widely used technique for evaluating the prediction accuracy of prediction models (Hastie et al. 2009). Cross-validation is a method of dividing the target population into several groups and then repeating the predicting the genomic estimated breeding value (GEBV) of one of them by training the remaining groups.

The response to selection can be evaluated by the genetic gain that is achieved by a selection in breeding. Genetic gain is defined as the increase in

performance obtained through selective breeding, and the expected annual improvement is calculated by GS accuracy, selection intensity, genetic standard deviation (the square root of the genetic variance), and the breeding cycle. In order to improve genetic gain, a breeding program could be designed by increasing selection intensity, expanding genetic variance, increasing prediction accuracy or heritability, shortening the breeding cycle, etc. in the consideration of the components of the formula of genetic gain (Xu et al., 2020).

In this chapter, the impact of the single nucleotide polymorphism (SNP) marker set, predictive model, and training data set on prediction accuracy was investigated in the half-sib population (’87 open-pollinated progeny trial) and the efficient GS model for Korean red pine was suggested. Also, the prediction accuracy of the trained GS model was evaluated through standardization. In addition, the efficiency of GS was evaluated by comparing the response to selection of GS and the conventional selections.

2.3. Materials and methods

2.3.1. SNP marker selection

For the comparison of prediction accuracy according to the SNP calling quality of markers, markers selected based on loose, moderate, and strict standards were used for GS. The loose standard was the use of all markers, the moderate standard was to meet the default quality threshold of the SNP calling program (call rate (CR)≥97%, Fisher's linear discriminant (FLD)≥3.6, etc.), and the strict standard was the addition of the threshold of CR≥99%, FLD≥5, heterozygous strength offset (HetSO)≥0, and homozygote ratio offset (HomRO)≥0. In common, markers with minor allele frequency (MAF) of 0.05 or more were used.

For the study of prediction accuracy according to the number of markers, all 17K (17,074) markers showing genotype variation and 2K (2,000), 6K (6,000), and 10K (10,000) markers randomly selected from them were used for GS. In addition, markers with MAF of 0.25, 0.05, and 0.0005 or higher were selected and compared for investigating the impact of MAF on prediction accuracy.

2.3.2. Genomic selection scenario

In order to evaluate the prediction accuracy of GS, cross-validation by region was performed on 2,643 trees, of which the genotypes were investigated in the half-sib population (Table 1-1). In this study, 3, 5, 10, and 20-fold cross-validation were tested and compared in a within-region prediction scenario (Figure 2-1a). In the between-region prediction scenario, to predict the GEBV of one region, all individuals in the remaining five regions were trained

(Figure 2-1b). In addition, in the combined-region prediction scenario, a 10-fold cross-validation was performed by randomly dividing groups regardless of region (Figure 2-1c). In the inter-family prediction scenario, all regions were aggregated, 4 random families were used as the test set, and the remaining 40 families were used as the training set. When multiple regions were included in the scenario, the phenotype was corrected by setting the region as a fixed effect in the linear model.

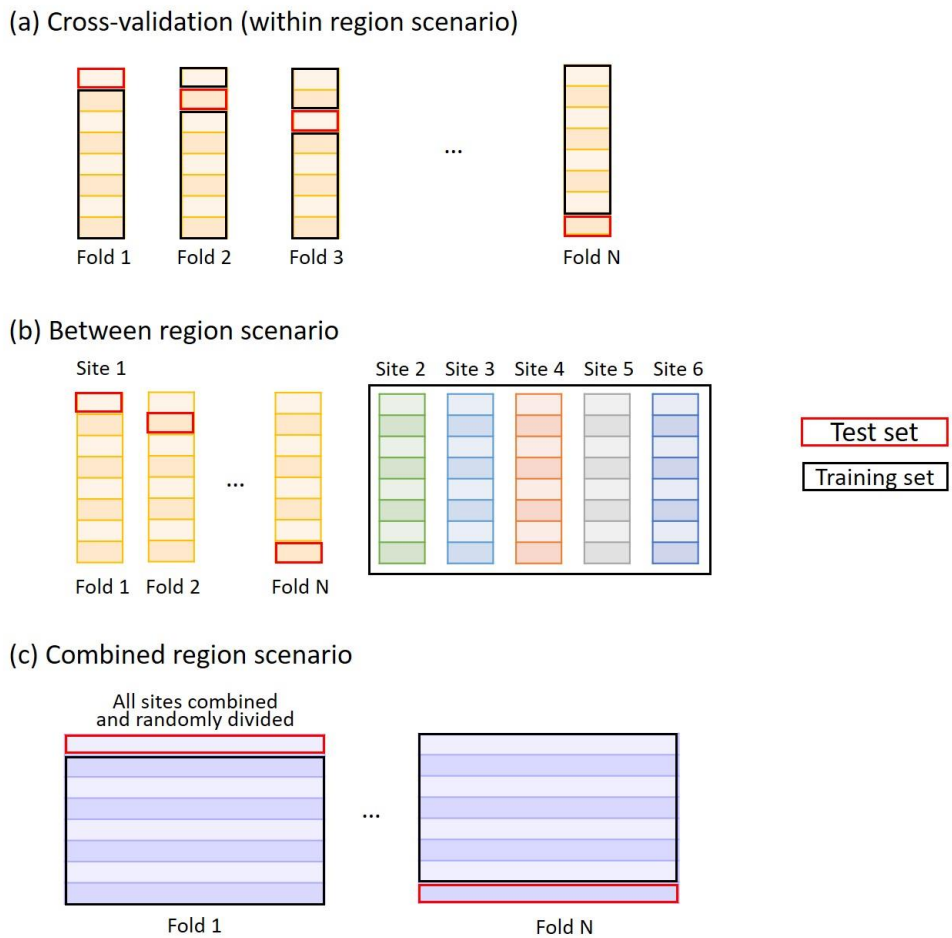

Figure 2-1. Genomic selection scenarios for multiple environment comparison.

2.3.3. Genomic estimated breeding value prediction and prediction accuracy

Six genomic predictive models including genomic best linear unbiased prediction (GBLUP), Bayesian least absolute shrinkage and selection operator (LASSO), Bayesian Ridge Regression, Bayes A, Bayes B, and Bayes C were tried to predict GEBVs. In addition, additive BLUP (ABLUP) which is pedigree-based prediction using numerator relationship matrix (NRM), and hybrid genomic BLUP (HBLUP), a hybrid model, were also performed for comparison. In HBLUP, the genotype data of 10% of individuals were excluded, and genomic information of the remaining individuals was combined with pedigree information to create a blended relationship matrix. The 10% of the population was randomly selected and used as the test set, and the prediction accuracy was averaged after repeating HBLUP 10 times. BLUPs were performed using remlf90 of R package BreedR as in chapter 1. In addition, five Bayesian models were performed with 20,000 iterations using the GBLR function of the R package BGLR (v1.0.8) (Perez and de Los Campos, 2014).

Prediction accuracy was evaluated by accuracy (AC) and predictive ability (PA). Accuracy is the Pearson correlation coefficient of GEBV of the test set and EBV_{im} calculated from pedigree information in the previous chapter.

$$\text{Accuracy} = r(\text{GEBV}, \text{EBV}_{\text{im}})$$

The predictive ability is the Pearson correlation coefficient between GEBV of the test set and EBV_{ps}. Since EBV_{ps} is expressed as a linear equation of individual phenotype, the predictive ability is the same as the correlation coefficient between GEBV and phenotype.

$$\text{Predictive ability} = r(\text{GEBV}, \text{EBV}_{ms}) = r(\text{GEBV}, \text{phenotype})$$

2.3.4. Response to selection

To compare the response to the selection of GS and two traditional selections, the annual genetic gain of each was calculated. Genetic gains from phenotypic selection and family selection were calculated as follows (Voss-Fels et al., 2019).

$$\Delta G_P = \frac{i h^2 \sigma_P}{t}$$

where ΔG_P is the annual genetic gain, i is the selection intensity, h^2 is individual or family heritability, σ_P is the square root of phenotypic variance, and t is breeding cycle.

The genetic gain from GS was measured as follows (Voss-Fels et al., 2019; Isik et al., 2017).

$$\Delta G_A = \frac{ir\sigma_A}{t}$$

where ΔG_A is the annual genetic gain, i is the selection intensity, r is the accuracy of GS, σ_A is the square root of the additive genetic variance, and t is the breeding cycle. For annual genetic gain of GS, genomic selection accuracy (GSAC) was used as accuracy. GSAC is the Pearson correlation coefficient between GEBV and EBV which was estimated by GBLUP with all phenotypic data. It was assumed that the breeding value (BV) estimated by GBLUP and phenotypic data was the closest to the true breeding value (TBV) according to Ukrainetz and Mansfield (2020a). The selection intensity according to the selection ratio was calculated assuming a normal distribution of phenotype.

2.4. Result and discussion

2.4.1. Impact of SNP marker set on predictive accuracy

(1) Marker quality standard

The data quality of the SNP array plays an important role in the accuracy and precision of subsequent analysis, and contaminated data may lead to false-positive or false-negative results (Yang et al., 2011). Therefore, in order to find out the impact of the quality of SNP markers on the prediction accuracy of GS, GBLUP analysis was performed by different standards for marker quality, and the accuracy and predictive ability were compared. The numbers of markers that meet the standards were 6,464 for the loose standard, 1,164 for the moderate standard, and 571 for the strict standard.

As a result of GBLUP analysis according to the marker set, the accuracy was 0.16~0.50 for the loose standard, 0.07~0.44 for the moderate standard, and -0.02~0.38 for the strict standard (Figure 2-2 a-d, Table S3). Also, predictive ability was 0.02~0.44 for the loose standard, -0.03~ 0.26 for the moderate standard, and -0.09~0.26 for the strict standard (Figure 2-2 e-h, Table S3). In all traits and regions, the stricter standard was applied, the lower the accuracy and predictive ability were. While some data showed differences within the error, in the case of Chuncheon, the accuracy of all traits showed significant differences by marker set.

Although the impact of marker quality on the accuracy of GS had not been previously studied, the marker criterion was a call rate of 85~95% in most works of literature on GS of forest trees (Ukrainetz and Mansfield, 2020a; Cappa et al., 2019; Beaulieu et al., 2014). Through this, it could be inferred that the criterion considered important is call rate. However, the

result of the comparison in this study indicated that it is more effective to use a lot of markers even if the markers of a slightly lower quality are included than to increase the quality of the markers for high accuracy of GS.

The reason for this might be that the low call rate could be compensated for through the imputation of missing genotypes. In a previous study on the effect of imputation on the accuracy of GS, the accuracy of prediction including markers with low call rates was higher than that of prediction excluding them, even when there was no marker order information (Rutkoski et al., 2013). The authors insisted that the result was because the markers were not saturated in the whole genome. They also argued that in the case the number of markers was sufficient, the difference in the accuracy of GS by the marker set was not significant. According to this, the results of this study, where prediction accuracy with markers of the loose standard was higher, indirectly indicated that 500 to 1,000 SNP markers of the array were not sufficient to explain the Korean red pine genome.

Another reason for the higher accuracy when imputed was inferred that if there was a similar tendency to be missing at a specific marker among closely related individuals, such associations could be captured well through the imputation (Rutkoski et al., 2013; Weigel et al., 2010). Therefore, in the subsequent analysis, a loose standard was applied in the selection of markers so that relatively high GS accuracy could be obtained.

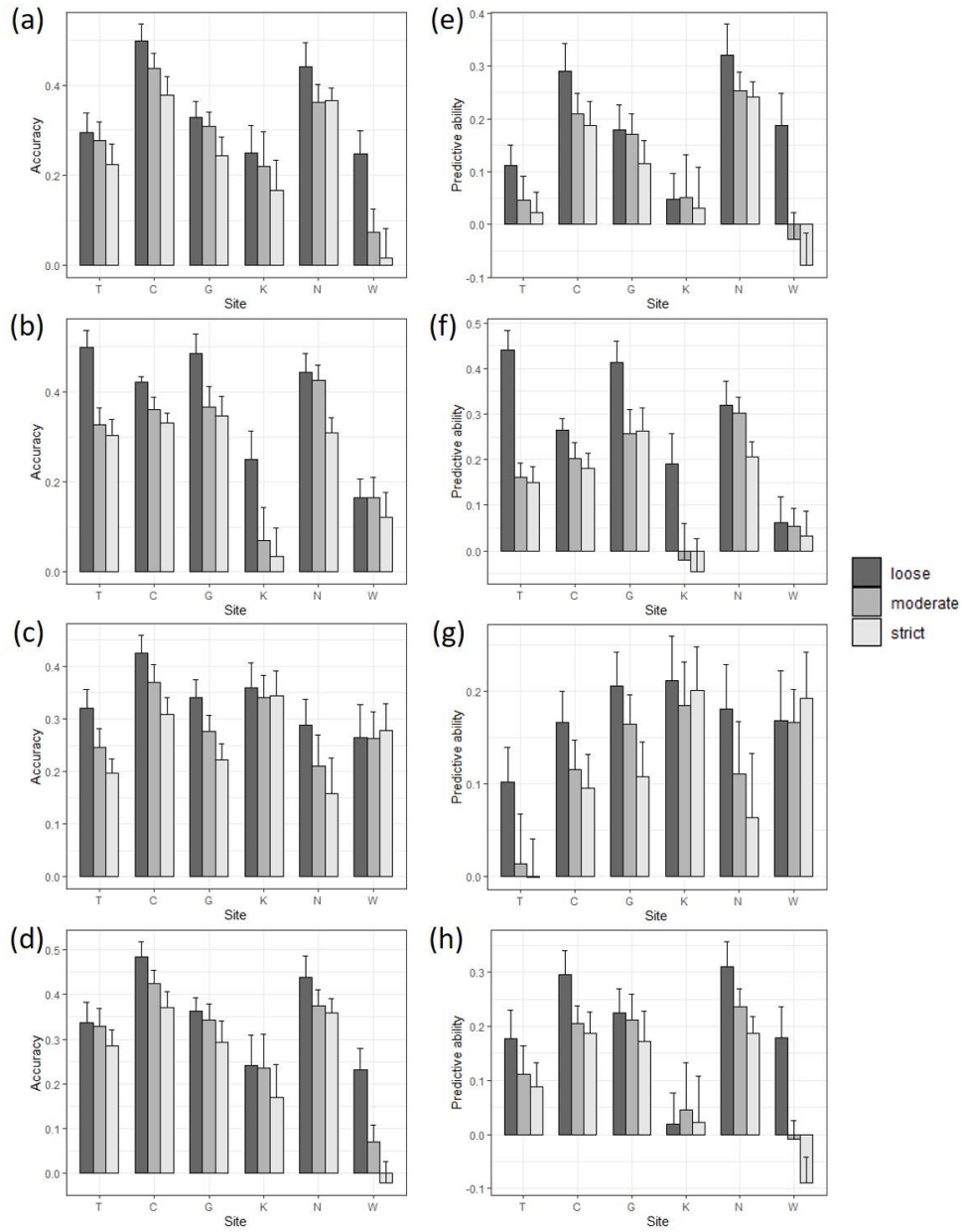

Figure 2-2. GBLUP accuracy and predictive ability using markers selected by the loose, moderate, and strict standards for four traits. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. Bar and error bar are mean and standard error of accuracy and predictive ability from 10-fold cross-validation.

(2) Number of markers

The number of markers used for analysis is an important factor affecting not only the prediction accuracy of GS but also computational time. This is the reason why it is necessary to search for an efficient number of markers for analysis. In order to examine the impact of the number of markers, all 17K (17,074) markers showing genotype variation and 2K (2,000), 6K (6,000), and 10K (10,000) marker set randomly selected among them were used in GBLUP and the prediction accuracy was compared. As a result, the accuracy was 0.09~0.46 for 2K, 0.18~0.47 for 6K, 0.24~0.48 for 10K, and 0.22~0.51 for 17K marker set (Figure 2-3 a-d, Table S4). Also, the predictive ability was -0.03~0.43 for 2K, 0.01~0.42 for 6K, 0.01~0.44 for 10K, and 0.02~0.46 for 17K marker set (Figure 2-3 e-h, Table S4). In general, as the number of markers decreased, the prediction accuracy decreased.

Then, it was analyzed whether the prediction accuracy decrease even when the number of markers was reduced based on MAF. The prediction accuracy was compared by GBLUP analysis with 2K (2,248), 6K (6,464), and 10K (9,799) markers with MAF of 0.25, 0.05, or 0.0005 or higher, respectively. As a result, the accuracy was 0.14~0.5, and the predictive ability was -0.02~0.46, showing differences within the error range (Figure 2-4, Table S5). In other words, when markers with low MAF were excluded, the prediction accuracy did not decrease, unlike when random markers were excluded. Thus, it was concluded that MAF had a greater effect on the prediction accuracy of GS rather than the simple number of markers.

On the other hand, in the case of height in Chuncheon and Kyeongju, accuracy and predictive ability significantly decreased at the MAF of 0.25 (2K), indicating that markers that were useful for predicting were included between the MAF of 0.05 and 0.25. In the previous study on the GS of forest

trees, markers were selected based on the MAF of 0.005 to 0.05 (Ukrainetz and Mansfield, 2020a; Cappa et al., 2019; Beaulieu et al., 2014). Therefore, it was concluded that selecting markers based on a MAF of 0.05 is efficient according to the results of this study and previous studies.

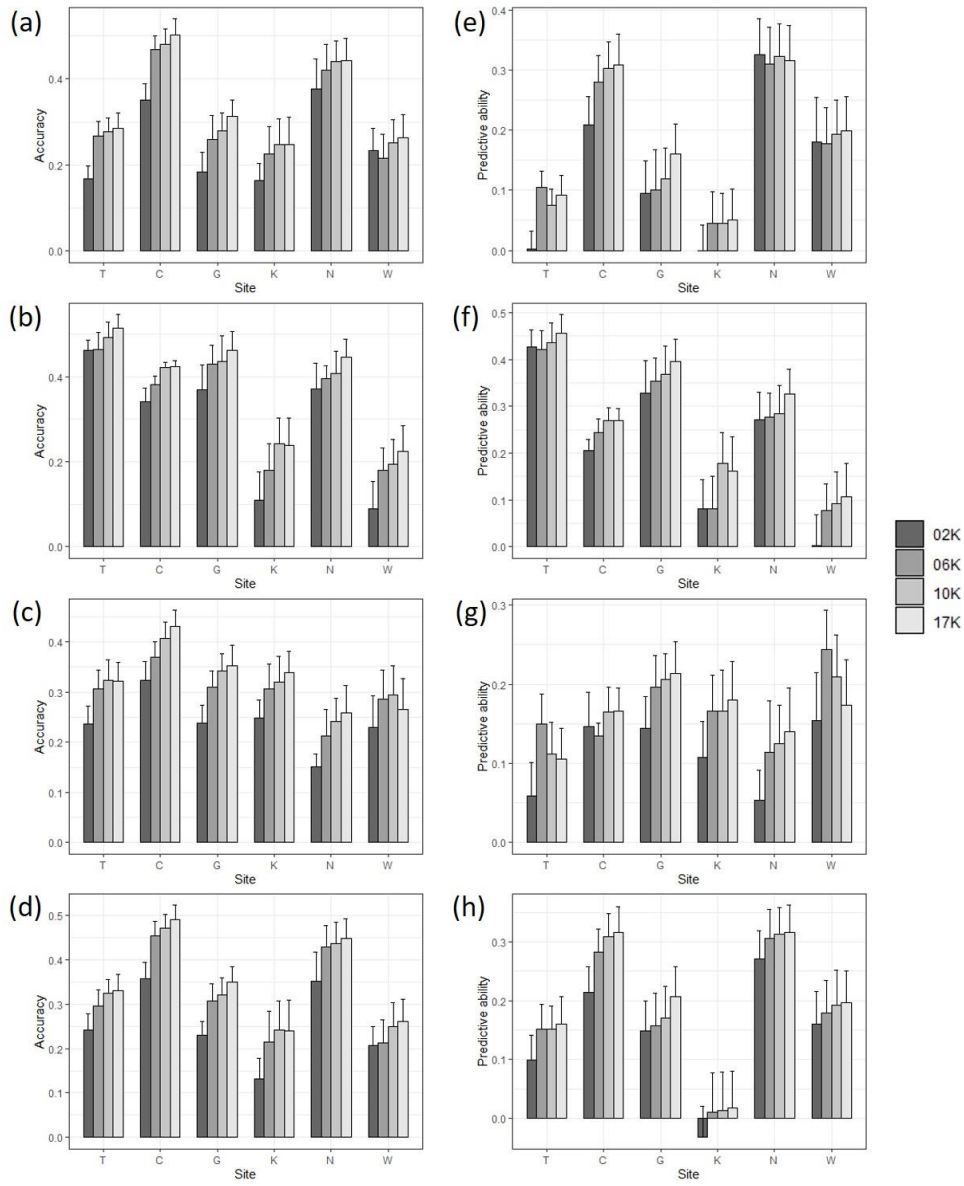

Figure 2-3. GBLUP accuracy and predictive ability using randomly selected 2K, 6K, 10K, and 17K markers. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. Bar and error bar are mean and standard error of accuracy and predictive ability from 10-fold cross-validation.

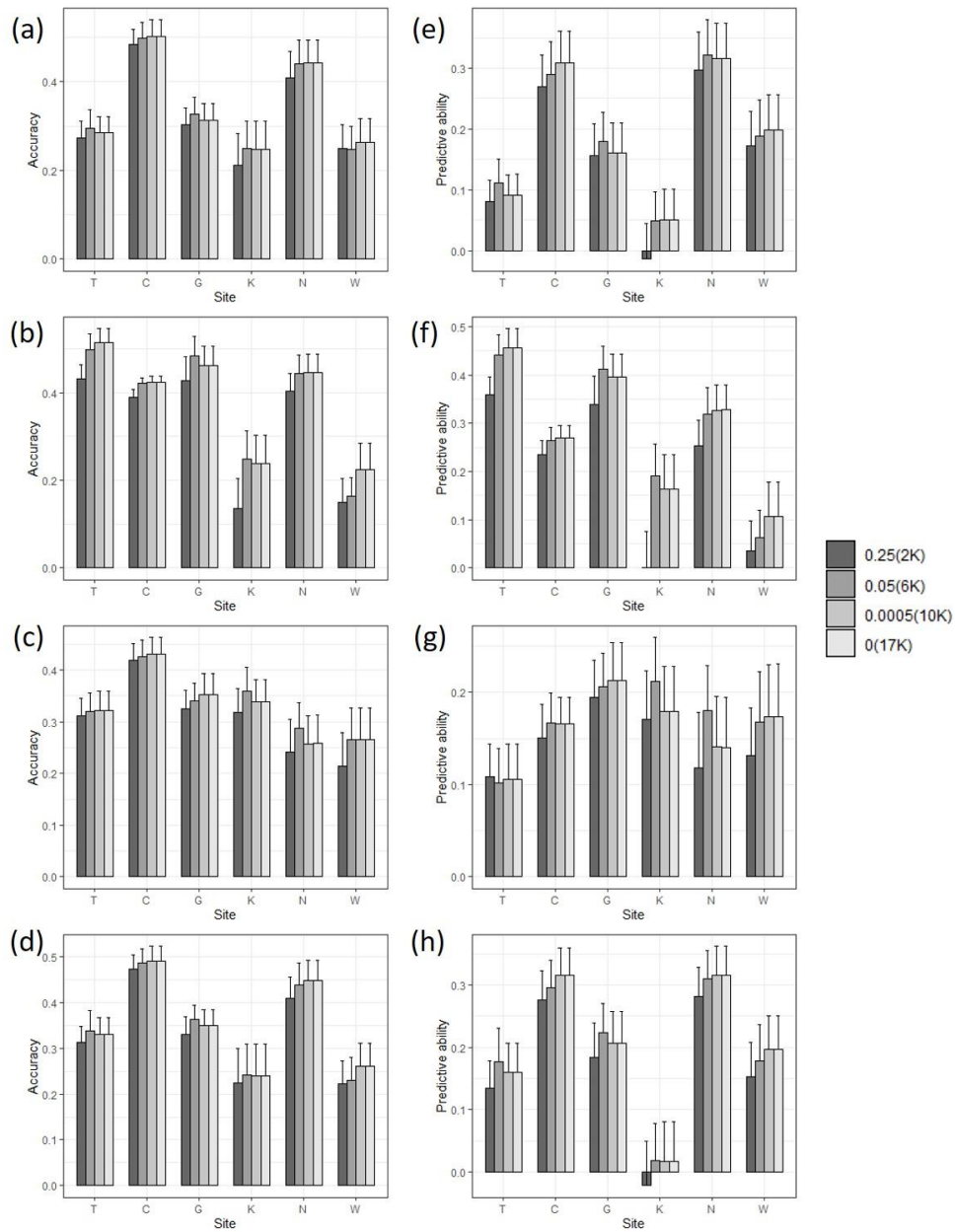

Figure 2-4. GBLUP accuracy and predictive ability using markers selected by minor allele frequency. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. The markers with MAF larger than 0.25, 0.05, 0.0005 and 0 were used. Bar and error bar are mean and standard error of accuracy and predictive ability from 10-fold cross-validation.

2.4.2. Impact of the predictive model on predictive accuracy

(1) Genomic predictive models and ABLUP

In order to search for a suitable predictive model for GS in Korean red pine, the prediction accuracy according to various models was analyzed and compared. In addition, ABLUP was also performed to compare the efficiency of GS with that of pedigree-based selection. As a result, six GS models showed the accuracy of 0.15~0.5 and the predictive ability of 0.01~0.44 and did not bring a significant difference in prediction accuracy among one another (Figure 2-5, Table S6).

In a previous study on the GS in loblolly pine, the Bayesian models (Bayes A, Bayes Cπ, Bayesian LASSO) showed better performance than RR-BLUP (ridge regression BLUP) for disease resistance which is regulated by a small number of loci (Resende et al., 2012a). However, most studies on quantitative traits had not found the advantage of a specific model. For example, RR-BLUP and Bayes Cπ showed similar results in the study of interior spruce (Ratcliff et al., 2015). Likewise, for eucalyptus, GBLUP and Bayesian models (Bayes B, Bayes C, Bayesian LASSO) showed similar prediction accuracy (Duran et al., 2017). It was also reported that there was no significant difference between GBLUP and Bayesian methods for growth and wood qualities in the two *Pinus* species, maritime pine and lodgepole pine (Isik et al., 2016; Ukrainetz and Mansfield, 2020b). In cedar, GBLUP showed higher prediction accuracy than Bayes B in growth, wood qualities, and reproduction ability (Hiraoka et al., 2018).

Insignificant differences in predictive models were also observed in studies of crops as well as forest trees. As the result of a comparison of models including RR-BLUP, Bayesian LASSO, and Bayes Cπ in wheat (*Triticum*

aestivum L.), barley (*Hordeum vulgare* L.), *Arabidopsis thaliana* (L.) Heynh., and maize (*Zea mays* L.), the models showed similar accuracy (Heslot et al., 2012). The authors argued that factors such as overfitting of the training population and computing time should be considered since the accuracy of all models was similar. Finally, they concluded that RR-BLUP was a reasonable choice. RR-BLUP is a model that quantifies the influence of genomic markers in the breeding population. It is the equivalent model to GBLUP when the number of QTLs is large, the QTLs are evenly distributed in the genome, and there is no marker with the main effect (Habier et al., 2007; Goddard, 2009; Lin et al., 2014). Since the accuracy and predictive ability of 6 models were similar as in plenty of previous studies, and GBLUP took about 9 times less computing time than the Bayesian methods in this study, GBLUP was judged to be efficient for GS of Korean red pine.

Meanwhile, in comparison with ABLUP (accuracy 0.32~0.72, predictive ability -0.18~0.25) based on pedigree information, the GS model generally showed low accuracy and high predictive ability (Figure 2-5, Table S6). Accuracy is the correlation coefficient between EBV_{im} and the predicted BV. Since EBV_{im} is obtained from ABLUP calculated with the phenotypes of all individuals, the correlation coefficient with the BV predicted by ABLUP is bound to be higher. Whereas, the predictive ability is the correlation coefficient with EBV_{ps} calculated from the phenotype, so it could be understood as the prediction accuracy for the actual phenotype.

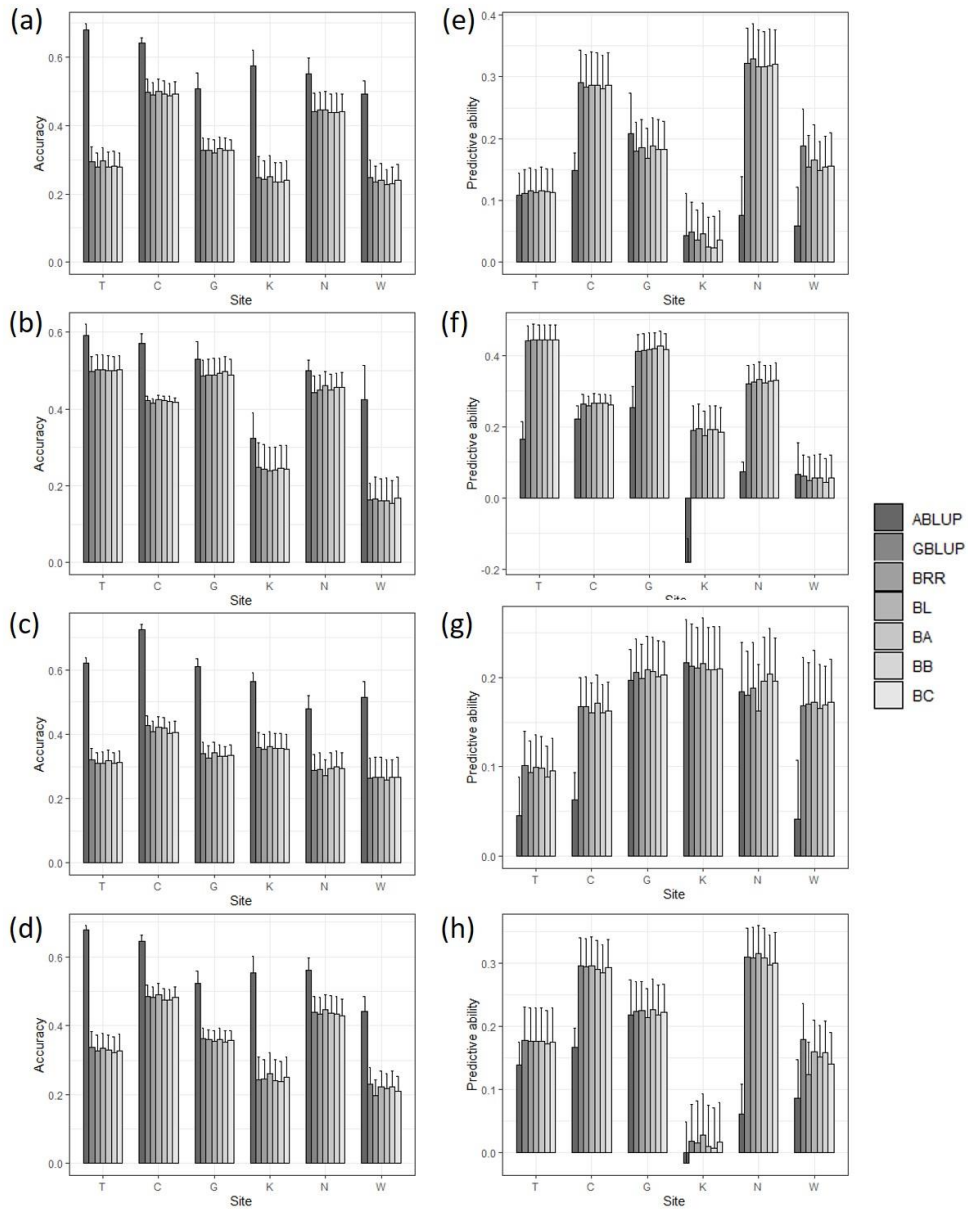

Figure 2-5. Accuracy and predictive ability by ABLUP and genomic selection models including GBLUP and five Bayesian models. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. BRR, Bayesian ridge regression; BL, Bayesian LASSO; BA, Bayes A; BB, Bayes B; BC, Bayes C. Bar and error bar are mean and standard error of accuracy and predictive ability from 10-fold cross-validation.

The results of higher accuracy in ABLUP and higher predictive ability in GS models including GBLUP had also been confirmed in many previous studies. For example, in the GS of black spruce, the accuracy of ABLUP was higher than that of GBLUP, and the predictive ability was vice versa (Lenz et al., 2017). The same result was reported in studies for disease resistance in white spruce (Beaulieu et al., 2020). Another study in white spruce showed that the predictive ability of GBLUP was 22-52% higher than that of ABLUP using only female parent information (Lenz et al., 2020a). Also, in lodgepole pine, the accuracy of ABLUP was higher than that of GBLUP (Ukrainetz and Mansfield, 2020b). Therefore, it was concluded that GS for obtaining GEBV using DNA markers had an advantage in phenotype prediction over traditional selection in Korean red pine.

(2) HBLUP

The prediction accuracy of HBLUP which can predict even without genomic information of some individuals was compared with ABLUP and GBLUP. The genomic information of the test set which corresponds to 10% of the individuals in each region was excluded and the BVs of the test set were predicted through a blended relationship matrix. As a result of the analysis, the accuracy was 0.38~0.73 for ABLUP, 0.23~0.69 for HBLUP, and 0.14~0.49 for GBLUP (Figure 2-6 a-d, Table S7). Also, the predictive ability was -0.23~0.28 for ABLUP, -0.24~0.33 for HBLUP, and -0.03~0.42 for GBLUP (Figure 2-6 e-h, Table S7). The accuracy and predictive ability of HBLUP were more similar to those of ABLUP than those of GBLUP.

In the literature, the accuracy and predictive ability of HBLUP had values between those of ABLUP and GBLUP. For example, in a recent study of lodgepole pine, the prediction accuracy became closer to GBLUP, as the

proportion of individuals with SNP genotypes increased from 20% to 40%, 60%, and 80% of the total population. Also, the HBLUP in the eucalyptus showed an identical predictive ability to GBLUP when the size of the training set was the same (Cappa et al., 2019).

Inconsistent with previous studies, in this study, the prediction accuracy of HBLUP was similar to that of ABLUP, although 90% of the genotypes were used to write the blended relationship matrix. In general, genomic relationship matrix (GRM) uses the allele frequency calculated from the genotypes of the investigated individuals instead of the unknown allele frequency of the population, which might cause incompatibility of GRM and NRM (Oliehoek et al., 2006). Also, the difference between the mean of GRM and NRM affects the accuracy of HBLUP's BV prediction (Powell et al., 2010). In this study, there was a difference in the average relationship coefficient between NRM and GRM, so it is possible that their incompatibility affected the reliability of HBLUP prediction accuracy (Figure 1-7c). In order to solve the incompatibility between NRM and GRM, several correction methods (Meuwissen et al., 2011; Vitezica et al., 2011; Christensen et al., 2012) have been developed. Additional research applying this method is required to utilize HBLUP in GS of Korean red pine.

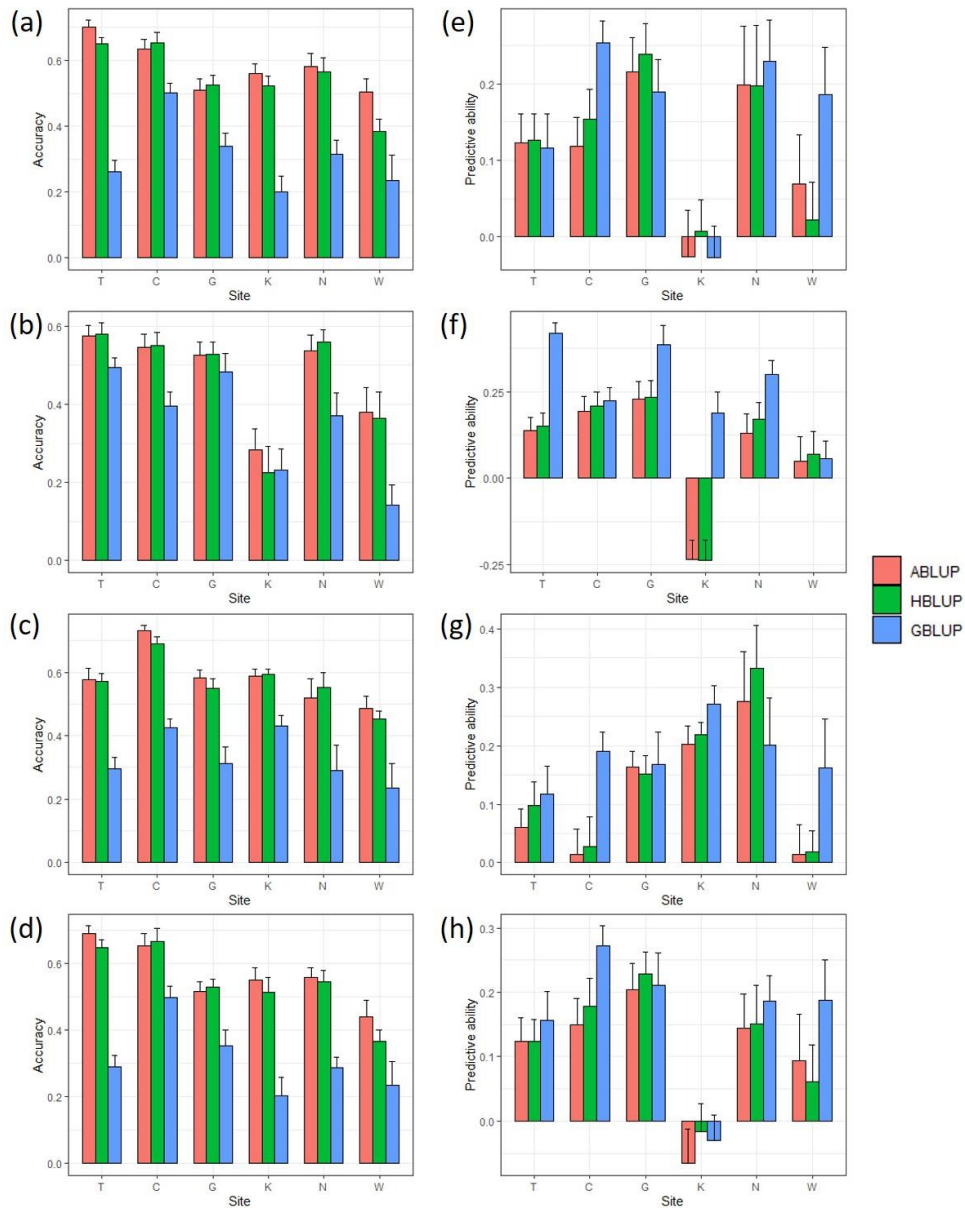

Figure 2-6. ABLUP, HBLUP, and GBLUP accuracy and predictive ability. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. Bar and error bar are mean and standard error of accuracy and predictive ability of 10 replication.

2.4.3. Impact of training data set on predictive accuracy

(1) Training set size

In order to examine whether the size of the training and test set affect the prediction accuracy in the GS of Korean red pine, the accuracy and predictive ability of GBLUP were compared by varying the number of cross-validation folds. As a result of cross-validation with the 3, 5, 10, and 20 folds, the accuracy was 0.14~0.51 and the predictive ability was -0.07~0.44, showing differences within the error range regardless of the size or ratio of the training set (Figure 2- 7, Table S8). For some regions and traits, such as volume in Naju, the prediction accuracy was increased as the training set size got larger.

In Norway spruce (*Picea abies*), the accuracy of the GS according to the ratio of the training set to the test set was compared (Chen et al., 2018). The accuracy increased as the ratio of the training set increased from 1:1 to 3:1, yet there was no significant difference as increasing to 5:1, 7:1, and 9:1. Otherwise, in eucalyptus, as the training set ratio increased from 1:1 to 2:1, 3:1, 4:1, and 9:1, predictive ability was greatly improved in all traits (Tan et al., 2017). However, as in the case of Norway spruce, the prediction accuracy changed the most when increasing from 1:1 to 2:1. The results of the two forest trees were consistent with the results of this study, suggesting that other factors such as population structure or environment were more important than the ratio of the training set when the population size is 200 to 700 in the GS of Korean red pine.

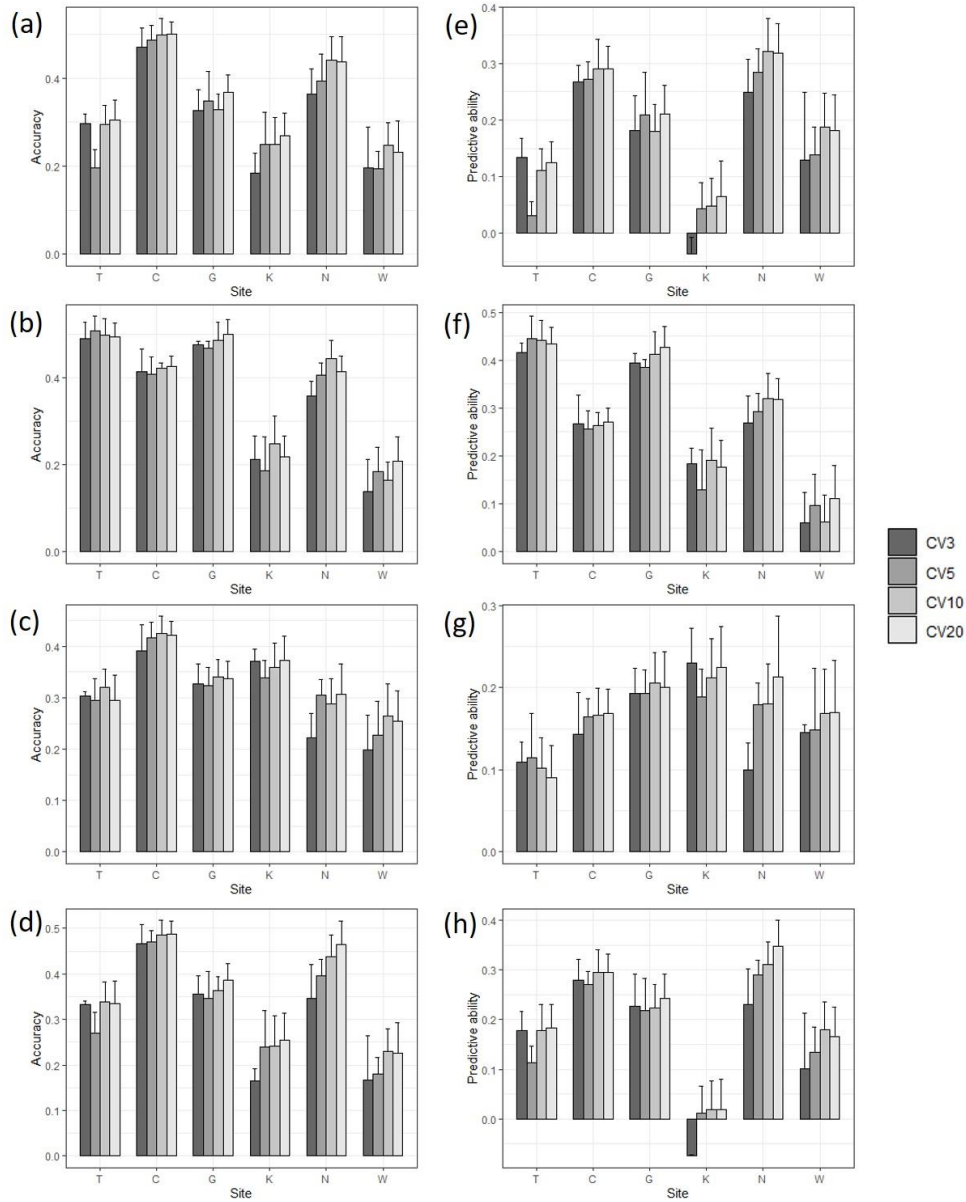

Figure 2-7. GBLUP accuracy and predictive ability by cross-validation fold number for four traits. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. CV3, CV5, CV10, and CV20 indicate cross-validation with 3, 5, 10, and 20 folds, respectively. Bar and error bar are mean and standard error of accuracy and predictive ability from cross-validation.

(2) Environment of training set

To find out whether GS can be applied to populations in different environments, the within-region, between-region, and combined-region GS scenarios were compared (Figure 2-8, Table S9). The accuracy did not show consistent results depending on whether the analysis was within or between regions. However, the predictive ability was generally higher in the within-region prediction (0.02~0.41) than in the between-region prediction (0.05~0.24). In the combined-region scenario, the accuracy was 0.38~0.48, which was higher than the average of six regions (0.33~0.38), and the predictive ability was 0.07~0.18, which was lower than the average of the six regions (0.17~0.28).

A previous study showed consistent results with this study. GS in black spruce was studied in two test sites, and the accuracy of the scenarios in which they were trained and tested each other increased or decreased compared to the accuracy within each region (Lenz et al., 2017). However, the predictive ability was lowered, and the accuracy increased when the two regions were combined as in this study. In addition, in Norway spruce, the prediction accuracy of GS was lowered in the analysis between populations with different environments, and the accuracy was increased when combined analysis (Chen et al., 2018).

The interaction of genotype and environment (G×E) refers to inconsistency in the expression of traits when trees grow up in different environments. In particular, it is considered that there is a greater interaction when the ranks of clones or families are changed in different environments. Since GS ranks the individuals according to GEBV, if G×E is discovered, it must be considered in GS strategies (Grattapaglia, 2017).

Although the G×E was a factor that lowered the prediction accuracy of

GS in the study of loblolly pine, the prediction accuracy was maintained high within the breeding zone (Resende Jr et al., 2012b). The authors argued that it is desirable to ensure that GS is performed within the breeding zone, even if it includes multiple environments or test sites. According to the recently investigated four genetic zones of Korean red pine in South Korea, Taean and Chuncheon belong to a genetic zone, and Gongju, Kyeongju, Naju, and Wanju belong to the other genetic zone (Ahn et al., 2021). To test whether predictions within the same genetic zone were more accurate, predictions between single test sites were performed. As a result, predictions between sites belonging to the same genetic zone had an accuracy of 0.03~0.435 and a predictive ability of -0.08~0.215 (Table S10). Also, predictions between sites belonging to different genetic zones had an accuracy of 0.037~0.474 and a predictive ability of -0.05~0.283 (Table S10). There were no significant differences between the two analyses. Therefore, for Korean red pine in South Korea, it was concluded that whether the training population and the test population belong to the same genetic zone did not significantly affect the GS prediction accuracy.

In summary, the similarity of the environments of training and test population was important in the GS of Korean red pine. Also, referring to the results of this study and previous studies in two spruce species and wheat (Bungueno et al., 2012), higher accuracy could be obtained when the GS model was trained by combining multiple environments. Therefore, it was concluded that training as many environments as possible is advantageous for obtaining high prediction accuracy when predicting GEBV of a new population.

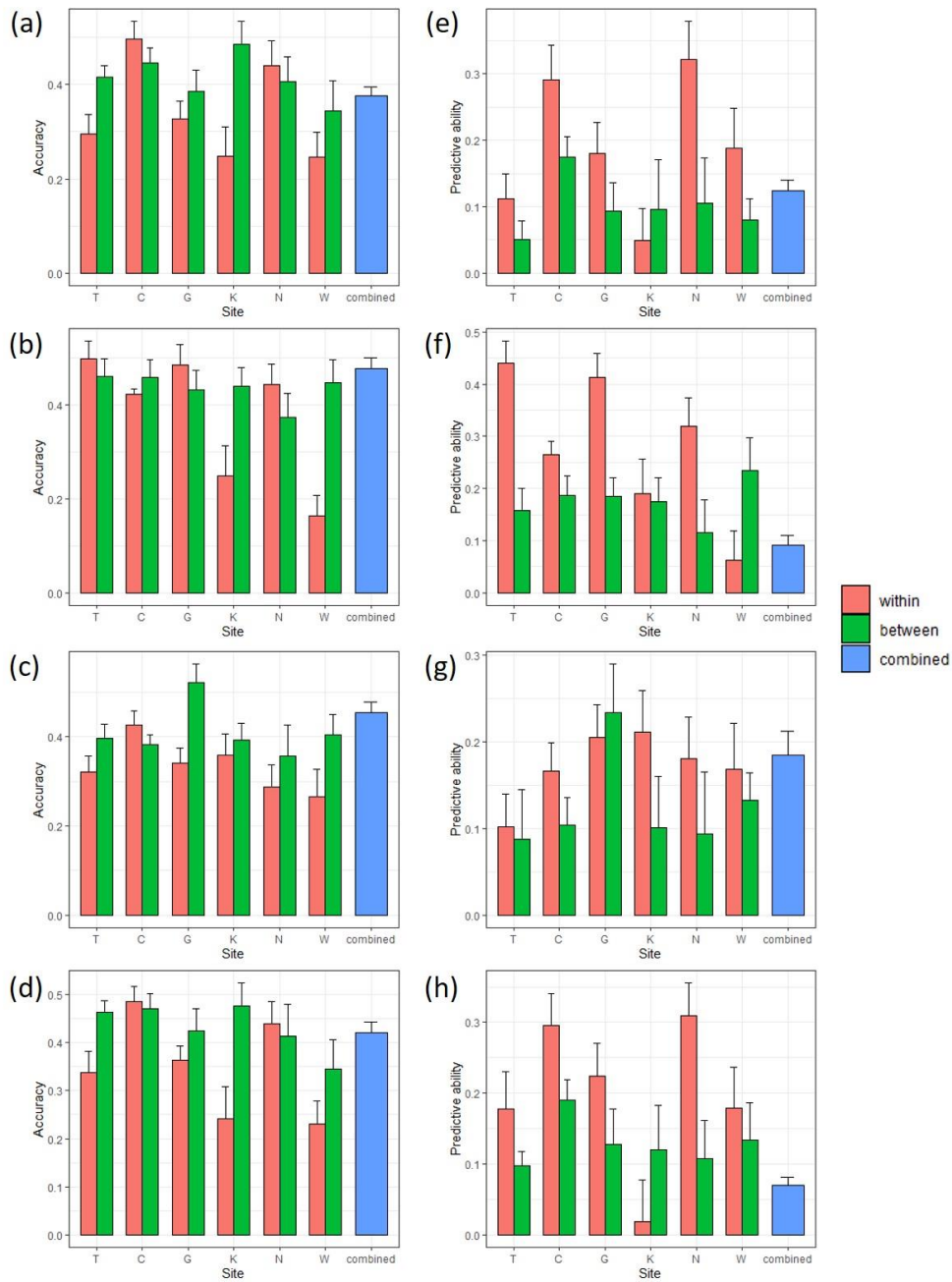

Figure 2-8. GBLUP accuracy and predictive ability of within- and between-region analysis and combined region analysis for four traits. (a, e) DBH (b, f) height (c, g) straightness and (d, h) volume. (a, b, c, d) accuracy (e, f, g, h) predictive ability. Bar and error bar are mean and standard error of accuracy and predictive ability from 10-fold cross-validation.

(3) Family composition of training set

In order to evaluate the genomic prediction accuracy between different open-pollinated families, the GBLUP analysis was performed by configuring the families so that they did not overlap in the training and test set. Forty-four families were randomly divided into 11 groups of 4 each and cross-validation was conducted. As a result, the accuracy of inter-family prediction was 0.112~0.166 and the predictive ability of that was -0.007~0.124 (Table 2-1). These were lower values compared to the results of combining all regions and families (accuracy of 0.38~0.48, predictive ability of 0.07~0.18) (Figure 2-8, Table S9).

In a previous study on Norway spruce, the half-sib scenario, which included individuals with the same female parent in the training and test set, and the non-related scenario where unrelated families were included in the training and test set were compared (Chen et al., 2018). The accuracy and predictive ability of GS were significantly reduced in the non-related scenario compared to the half-sib scenario, which was consistent with this study, revealing that the genetic structure of the training population and test population had a significant impact on GS. Therefore, when evaluating the prediction accuracy of GS, it is essential to consider the genetic association between the training population and the test population.

Table 2-1. GBLUP prediction accuracy of the open-pollinated population using inter-family analysis for four traits.

Trait	Accuracy	Predictive ability
DBH	0.146 (0.049)	0.080 (0.020)
Height	0.166 (0.039)	0.031 (0.017)
Straightness	0.112 (0.047)	0.124 (0.017)
Volume	0.160 (0.047)	-0.007 (0.018)

Mean (standard error) of 11 replications

2.4.4. Prediction accuracy evaluation

Summarizing the study results, it was analyzed that predicting GEBV with the GBLUP model using the genotype of 6,464 markers with a MAF of 0.05 or higher was effective when performing GS using the 50K SNP chip in Korean red pine. Finally, the prediction accuracy of the GS performed by this model was found to be 0.164~0.498 for accuracy and 0.018~0.441 for predictive ability (Table 2-2). The prediction accuracy was the highest for height in Taean. The mean prediction accuracy of all regions was the highest for height. The population size and the prediction accuracy did not show a correlation.

Table 2-2. Accuracy and predictive ability for four traits using GBLUP of 10-fold cross-validation

	DBH		Height		Straightness		Volume	
	AC [a]	PA [b]	AC	PA	AC	PA	AC	PA
Taean	0.294	0.111	0.498	0.441	0.320	0.102	0.338	0.177
Chuncheon	0.497	0.290	0.422	0.264	0.426	0.167	0.485	0.296
Gongju	0.328	0.180	0.485	0.413	0.341	0.206	0.363	0.224
Kyeongju	0.249	0.049	0.249	0.190	0.360	0.212	0.242	0.018
Naju	0.441	0.322	0.443	0.320	0.288	0.180	0.439	0.310
Wanju	0.267	0.188	0.164	0.062	0.265	0.168	0.231	0.179
Mean	0.346	0.190	0.377	0.282	0.333	0.173	0.350	0.201
Combined	0.377	0.124	0.476	0.091	0.453	0.185	0.421	0.070

[a] AC, accuracy, r(GEBV, EBV_{im}), EBV_{im} is estimated by ABLUP with all phenotype data

[b] PA, predictive ability, r(GEBV, phenotype)

On the other hand, the accuracy was always higher than the predictive ability (Table 2-2). Accuracy is the correlation between values indicating the additive effects, whereas predictive ability is the correlation between the GEBV and the phenotype which includes the influence of the environment. Therefore, the result that the predictive ability was lower than the accuracy could be interpreted that the interaction between the genotype and the environment or the non-additive genetic variance played a large role. The accuracy is useful when evaluating the efficiency of GS compared to the conventional pedigree-based selection. While the predictive ability could be used when evaluating prediction of a population with a simple or unknown pedigree. This advantage of predictive ability is favored in the breeding of forest trees in which the history of commercial production is short (Li et al., 2019).

In order to reveal the correlation between the heritability and the prediction accuracy of GS in Korean red pine, a Pearson correlation analysis between them in each region and trait was performed. The correlation coefficient between accuracy and heritability was 0.735 (p-value<0.001), and the correlation coefficient between predictive ability and heritability was 0.924 (p-value <0.001), showing a strong correlation (Figure 2-9). Therefore, it was found that the prediction accuracy of GS in Korean red pine was strongly influenced by heritability.

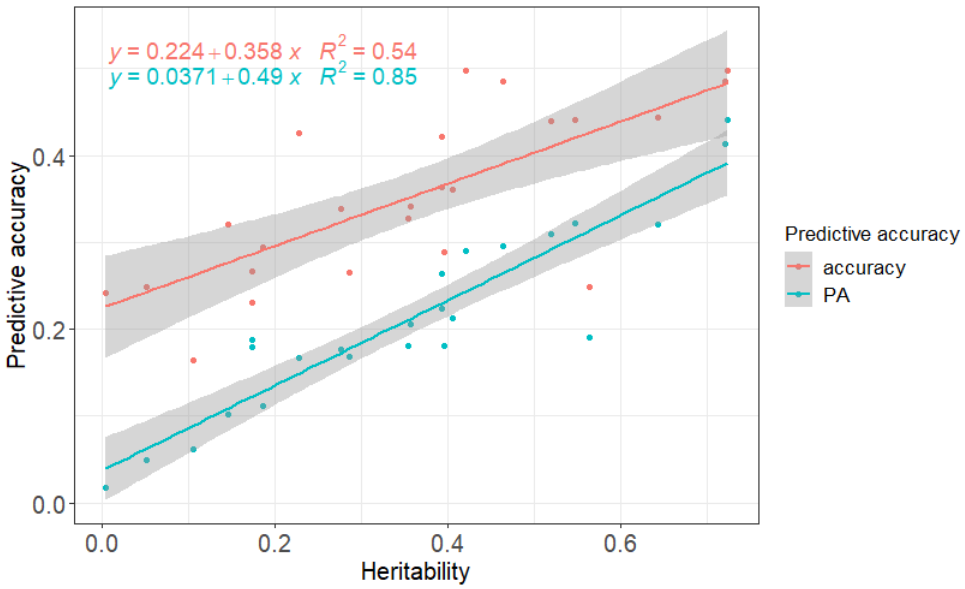

Figure 2-9. Prediction accuracies according to heritability.

Since the features of the population such as heritability affect the prediction accuracy of GS, the direct comparison of prediction accuracies in different populations is not recommended. Instead, standardization using prediction accuracy of ABLUP or heritability could be tried for comparison. The predictive ability of GBLUP against that of ABLUP ranged from 0.864 to 5.098 in within-region analysis (Table 2-3). This result is higher compared to the result of 0.80~0.95 in Norway spruce (Chen et al. 2018). Also, the predictive ability of GBLUP against the square root of heritability was 0.19~0.76 (Table 2-4), which is lower on average than the result in Norway spruce (higher than 0.69) but is similar to the result in white spruce (Lenz et al., 2020b; Beaulieu et al., 2020). Therefore, GS in Korean red pine was as much as efficient in previous studies on forest trees.

Table 2-3. Predictive ability of GBLUP against that of ABLUP

	DBH	Height	Straightness	Volume
Taean	1.022	2.664	2.258	1.274
Chuncheon	1.948	1.189	2.626	1.775
Gongju	0.864	1.633	1.048	1.029
Kyeongju *	1.117	-	0.981	-
Naju	4.197	4.410	0.981	5.098
Wanju	3.230	0.940	4.038	2.067
Combined	1.223	1.159	1.558	0.539

* The values for height and volume were omitted because the predictive ability of ABLUP was negative.

Table 2-4. Predictive ability of GBLUP against the square root of heritability

	DBH	Height	Straightness	Volume
Taean	0.257	0.519	0.267	0.337
Chuncheon	0.447	0.421	0.351	0.435
Gongju	0.303	0.487	0.345	0.357
Kyeongju	0.217	0.253	0.333	0.285
Naju	0.436	0.399	0.286	0.430
Wanju	0.452	0.191	0.314	0.430
Combined	0.483	0.282	0.762	0.244

2.4.5. Response to selection

To test the efficiency of GS versus traditional breeding, genetic gains that are expected to be obtained by GS, phenotypic selection (PS), and family selection (FS) were compared. The breeding cycle of GS was assumed to be 15 years in consideration of the age of reproduction of Korean red pine. For FS, the time for the progeny test was added. The progeny test usually takes about 15 to 20 years for conifers, and selection is sometimes performed at 1/3 of the harvesting age (Isik, 2014; Muranty et al., 2014). In this study, since phenotypic data of about the 30 years old trees from the open-pollinated progeny trial were used, the time required for the progeny test was considered as 30 years. Accordingly, a total breeding cycle of 45 years was assumed for FS. Also, for PS, the breeding cycle was assumed as 30 years due to the time for growing up to 30-year-olds.

As for the annual genetic gain of combined region analysis, GS was the highest, followed by FS, and PS was the lowest in DBH, height, and volume (Figure 2-10). The considerable superiority of the efficiency of GS to the that of two traditional selection methods was observed for DBH. GS showed an annual genetic gain equivalent up to 6.8 times of PS and 1.6 times of FS for DBH when 20% selection was conducted. Also, the annual genetic gain of GS was the highest at 0.19~2.53% for volume. The annual genetic gain of the straightness was highest in FS, but there was no significant difference from that in GS.

In addition, the genetic gains of 20% selection (selection intensity 1.4) by GS, PS, and FS were compared within a region (Table 2-5). As the result, the GS was the most efficient among the three selection methods in the aspect of annual genetic gain. In GS within the region, a response to selection of up

to 2.4% per year could be obtained. Moreover, the selection intensity in GS would be enhanced. Since the selection can be conducted in a seedling stage in GS whereas PS and FS are conducted in the age after phenotype expression, the selection intensity could be increased in GS for the same number of selected trees (Grattapaglia, 2017). Therefore, in terms of the annual genetic gain, the GS of Korean red pine was judged to be more efficient compared to the two traditional selection methods in both within- and combined-region scenario.

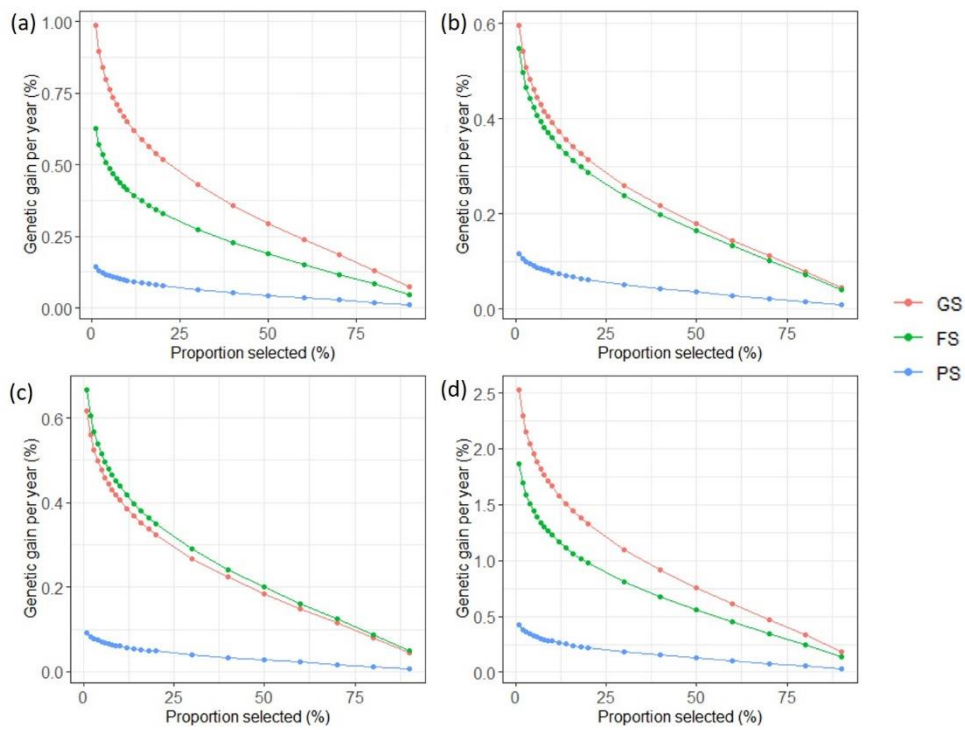

Figure 2-10. Annual genetic gain of genomic selection and two traditional selections by proportion selected for four traits. (a) DBH (b) height (c) straightness and (d) volume. GS, genomic selection; FS, family selection; PS, phenotypic selection.

Table 2-5. Annual genetic gain from phenotypic selection, family selection, and genomic selection in each site for four traits.

Trait	Site [a]	GSAC [b]	Genetic gain [c]		
			ΔG_{PS} (%)	ΔG_{FS} (%)	ΔG_{GS} (%)
DBH	T	0.578	0.209	0.100	0.559
	C	0.653	0.556	0.268	1.121
	G	0.512	0.354	0.228	0.610
	K	0.657	0.058	0.040	0.338
	N	0.521	0.530	0.015	0.748
	W	0.521	0.228	0.065	0.570
Height	T	0.606	0.407	0.163	0.580
	C	0.655	0.202	0.161	0.422
	G	0.568	0.450	0.210	0.602
	K	0.399	0.351	0.037	0.373
	N	0.490	0.450	0.026	0.551
	W	0.490	0.072	0	0.217
Straight-ness	T	0.637	0.122	0.144	0.406
	C	0.659	0.192	0.146	0.531
	G	0.545	0.297	0.201	0.543
	K	0.556	0.298	0.113	0.521
	N	0.449	0.040	0.029	0.058
	W	0.449	0.217	0.157	0.365
Volume	T	0.593	0.677	0.425	1.528
	C	0.628	1.319	0.524	2.434
	G	0.552	0.909	0.629	1.602
	K	0.649	0.012	0.043	0.224
	N	0.534	1.200	0.207	1.780
	W	0.534	0.528	0.179	1.357

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

[b] GSAC, GS accuracy, r(GEBV, EBV) EBV was estimated by GBLUP with all phenotype data

[c] Ratio of genetic gain per year to mean. ΔG_{PS}, the genetic gain of phenotypic selection; ΔG_{FS}, the genetic gain of family selection; ΔG_{GS}, the genetic gain of genomic selection.

2.5. Conclusion

An efficient GS model of Korean red pine was presented and the selection efficiency was evaluated in this chapter. As a result of comparing various marker subsets, predictive models, and scenarios to optimize the GS model in Korean red pine, training the model with markers of MAF of 0.05 or more, using GBLUP as a predictive model, and including as many environments as possible was effective. The GS of Korean red pine was as much efficient as in other tree species. Also, it was evaluated to have a high response to selection compared to the traditional PS and FS. Thus, GS was concluded to be an appropriate alternative to the traditional selection of Korean red pine. The result of this chapter added evidence that GS is efficient in forest trees.

References

Ahn, J. Y., Lee, J.W. and Hong, K.N. 2021. Genetic Diversity and Structure of *Pinus densiflora* Siebold & Zucc. Populations in Republic of Korea Based on Microsatellite Markers. Forests 12(6): 750.

Beaulieu, J., Doerksen, T. K., MacKay, J., Rainville, A. and Bousquet, J. 2014. Genomic selection accuracies within and between environments and small breeding groups in white spruce. *BMC genomics 15*(1): 1-16.

Beaulieu, J., Nadeau, S., Ding, C., Celedon, J. M., Azaiez, A., Ritland, C., ... and Bousquet, J. 2020. Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. Evolutionary applications 13(10): 2704-2722.

Burgueño, J., de los Campos, G., Weigel, K. and Crossa, J. 2012. Genomic prediction of breeding values when modeling genotype× environment interaction using pedigree and dense molecular markers. Crop Science 52(2): 707-719.

Cappa, E. P., de Lima, B. M., da Silva-Junior, O. B., Garcia, C. C., Mansfield, S. D. and Grattapaglia, D. 2019. Improving genomic prediction of growth and wood traits in Eucalyptus using phenotypes from non-genotyped trees by single-step GBLUP. *Plant Science 284*, 9-15.

Chen, Z. Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., ... and Wu, H. X. 2018. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC genomics 19(1): 1-16.

Christensen, O. F., Madsen, P., Nielsen, B., Ostersen, T. and Su, G. 2012. Single-step methods for genomic evaluation in pigs. Animal 6(10): 1565-1571.

Durán, R., Isik, F., Zapata-Valenzuela, J., Balocchi, C. and Valenzuela, S. 2017. Genomic predictions of breeding values in a cloned Eucalyptus globulus population in Chile. *Tree Genetics & Genomes 13*(4): 1-12.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica 136*(2): 245-257.

Grattapaglia, D. 2017. Status and perspectives of genomic selection in forest tree breeding. In Genomic selection for crop improvement (pp. 199-249). Springer, Cham.

Habier, D., Fernando, R. L. and Dekkers, J. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177*(4): 2389-2397.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. 2009. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

Heslot, N., Yang, H. P., Sorrells, M. E. and Jannink, J. L. 2012. Genomic selection in plant breeding: a comparison of models. *Crop science 52*(1): 146-160.

Hiraoka, Y., Fukatsu, E., Mishima, K., Hirao, T., Teshima, K. M., Tamura, M., ... and Watanabe, A. 2018. Potential of genome-wide studies in unrelated plus trees of a coniferous species, Cryptomeria japonica (Japanese cedar). *Frontiers in plant science* 1322.

Isik, F. 2014. Genomic selection in forest tree breeding: the concept and an outlook to the future. New Forests 45(3): 379-401.

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., ... and Bouffier, L. 2016. Genomic selection in maritime pine. *Plant Science 242*, 108-119.

Isik, F., Holland, J. and Maltecca, C. 2017. Genetic data analysis for plant and animal breeding (Vol. 400). Cham, Switzerland: Springer International Publishing.

Lebedev, V. G., Lebedeva, T. N., Chernodubov, A. I. and Shestibratov, K. A. 2020. Genomic selection for forest tree improvement: Methods, achievements and perspectives. Forests 11(11): 1190.

Lenz, P., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M. and Bousquet, J. 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC genomics 18(1): 1-17.

Lenz, P., Nadeau, S., Azaiez, A., Gérardi, S., Deslauriers, M., Perron, M., ... and Bousquet, J. 2020a. Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs: an example from white spruce. Heredity 124(4): 562-578.

Lenz, P. R., Nadeau, S., Mottet, M. J., Perron, M., Isabel, N., Beaulieu, J. and Bousquet, J. 2020b. Multi-trait genomic selection for weevil resistance, growth, and wood quality in Norway spruce. Evolutionary applications 13(1): 76-94.

Li, Y., Klápště, J., Telfer, E., Wilcox, P., Graham, N., Macdonald, L. and Dungey, H. S. 2019. Genomic selection for non-key traits in radiata pine when the documented pedigree is corrected using DNA marker information. BMC genomics 20(1): 1-10.

Lin, Z., Hayes, B. J. and Daetwyler, H. D. 2014. Genomic selection in crops, trees and forages: a review. *Crop and Pasture Science 65*(11): 1177-1191.

Meuwissen, T. H. E., Luan, T. and Woolliams, J. A. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. Journal of Animal Breeding and Genetics 128(6): 429-439.

Muranty, H., Jorge, V., Bastien, C., Lepoittevin, C., Bouffier, L. and Sanchez, L. 2014. Potential for marker-assisted selection for forest tree breeding: lessons from 20 years of MAS in crops. Tree Genetics & Genomes 10(6): 1491-1510.

Oliehoek, P. A., Windig, J. J., Van Arendonk, J. A. and Bijma, P. 2006. Estimating relatedness between individuals in general populations with a focus on their use in conservation programs. Genetics 173(1): 483-496.

Pérez, P. and de Los Campos, G. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198(2): 483-495.

Powell, J. E., Visscher, P. M. and Goddard, M. E. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nature Reviews Genetics 11(11): 800-805.

Ratcliffe, B., El-Dien, O. G., Klápště, J., Porth, I., Chen, C., Jaquish, B. and El-Kassaby, Y. A. 2015. A comparison of genomic selection models across time in interior spruce (Picea engelmannii× glauca) using unordered SNP imputation methods. Heredity 115(6): 547-555.

Resende Jr, M. F. R., Munoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., ... and Kirst, M. 2012a. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190(4): 1503-1510.

Resende Jr, M. F. R., Muñoz, P., Acosta, J. J., Peter, G. F., Davis, J. M., Grattapaglia, D., ... and Kirst, M. 2012b. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist 193(3): 617-624.

Rutkoski, J. E., Poland, J., Jannink, J. L. and Sorrells, M. E. 2013. Imputation of unordered markers and the impact on genomic selection accuracy. *G3:*

Genes, Genomes, Genetics 3(3): 427-439.

Tan, B., Grattapaglia, D., Martins, G. S., Ferreira, K. Z., Sundberg, B. and Ingvarsson, P. K. 2017. Evaluating the accuracy of genomic prediction of growth and wood traits in two Eucalyptus species and their F1 hybrids. *BMC plant biology 17*(1): 1-15.

Thistlethwaite, F. R., Ratcliffe, B., Klápště, J., Porth, I., Chen, C., Stoehr, M. U. and El-Kassaby, Y. A. 2019. Genomic selection of juvenile height across a single-generational gap in Douglas-fir. *Heredity 122*(6): 848-863.

Ukrainetz, N. K. and Mansfield, S. D. 2020a. Prediction accuracy of single-step BLUP for growth and wood quality traits in the lodgepole pine breeding program in British Columbia. Tree Genetics & Genomes 16(5): 1-13.

Ukrainetz, N. K. and Mansfield, S. D. 2020b. Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. *Tree Genetics & Genomes 16*(1): 1-19.

Vitezica, Z. G., Aguilar, I., Misztal, I. and Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genetics Research 93(5): 357-366.

Voss-Fels, K. P., Cooper, M. and Hayes, B. J. 2019. Accelerating crop genetic gains with genomic selection. Theoretical and Applied Genetics 132(3): 669-686.

Weigel, K. A., de Los Campos, G., Vazquez, A. I., Rosa, G. J. M., Gianola, D. and Van Tassell, C. P. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of dairy science 93*(11): 5423-5435.

Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., ... and Zhang, A. 2020. Enhancing genetic gain through genomic selection: from livestock to plants. Plant Communications 1(1): 100005.

Yang, H. C., Lin, H. C., Kang, M., Chen, C. H., Lin, C. W., Li, L. H., ... and Pan, W. H. 2011. SAQC: SNP array quality control. *BMC bioinformatics 12*(1): 1-14.

Chapter 3. Validation of genomic selection model using a control-pollinated progeny trial of Korean red pine

3.1. Abstract

The last step in genomic selection (GS) is to select excellent individuals based on genomic estimated breeding value (GEBV) by applying the trained model to a new population. In this chapter, the prediction accuracy in the full-sib population was evaluated and the GS model of Korean red pine trained in the previous chapter was verified. Also, excellent individuals based on GEBV were selected for the full-sib population. A control-pollinated progeny trial planted in Hwaseong was used as the study population. As the result of a 5-fold cross-validation for DBH, height, and volume in the full-sib population, accuracy was 0.766~0.781 and predictive ability was 0.408~0.463, which were higher than that of the half-sib population analyzed in the previous chapter due to closer relationship within the population. In addition, when the full-sib population was tested with the model trained in the previous chapter, it was effective when all regions were trained, showing the genomic selection accuracy of 0.657 for DBH, 0.501 for height, and 0.896 for volume. The prediction accuracy of GS was improved when markers showing polymorphism in the test population as well were used. It was investigated that 10% of individuals selected based on GEBV belonged to 18 full-sib families, ensuring genetic diversity rather than family selection. Therefore, the GS of Korean red pine was concluded to be predictable in a new population without a relationship.

3.2. Introduction

Validating the genomic selection (GS) model is examining whether predictions can be made even in other populations that are not related to the training population. The relationship of individuals within a population has been dealt with as a determinant of predictive accuracy in GS in early simulation studies and has been highlighted in plant and livestock breeding and recent reviews (Heslot et al., 2015; Lin et al., 2014; Van Eenennaam et al., 2014).

Individuals that are closely related to the training population are expected to have an advantage in the prediction accuracy of GS over those who are less relevant. In order to increase prediction accuracy in the maize study, it was found that it is more effective to increase the relevance between the training and test population than to increase the size of the training population with low relevance to the prediction target population (Riedelsheimer et al., 2013). Also, in the study of GS in forest trees, the prediction accuracy significantly decreased when the average estimated kinship coefficient was zero (Beaulieu et al., 2014a). Therefore, since the genetic association between individuals is an important component of prediction accuracy, the relationship with the training population should be reflected when selecting the test population (Daetwyler et al., 2013).

However, the accuracy of genomic estimated breeding value (GEBV) is not equal to zero even in the absence of linkage disequilibrium (LD) because LD between the marker and the quantitative trait locus (QTL) of the trait of interest is not required in GS (Grattapaglia, 2017). This showed the possibility of expanding the use of GS even in a population that is not related to the training population.

Most of the studies on GS of forest trees carried out cross-validation within a population or prediction between two populations of which family composition is identical. However, only a few studies on the GS of forest trees validated the GS model in unrelated populations. For example, an unrelated scenario, where no close relatedness was observed between the training and validation population through realized relationship analysis, was performed in lodgepole pine (Ukrainetz and Mansfield, 2020b).

In this chapter, in order to investigate the association between prediction accuracy and the genetic relationship of the population, the prediction accuracy of GS in the full-sib population (’10 control-pollinated progeny trial) was assessed and compared to that of the half-sib population (’87 open-pollinated progeny trial) studied in chapter 3. In addition, as a verification step for the broad application of the GS model, the model trained in the half-sib population was validated in a full-sib population that has no relation by evaluating the predictive accuracy.

3.3. Materials and methods

3.3.1. Prediction in full-sib population

The predictive ability of additive best linear unbiased prediction (ABLUP) and genomic BLUP (GBLUP) in ’10 control-pollinated progeny trial was analyzed using the genomic realized relationship matrix (GRM) and numerator relationship matrix (NRM) written in chapter 1. The markers had been filtered with quality of default value and minor allele frequency (MAF) of 0.05 (1,277, Table 1-3). Also, the predictive accuracy of ABLUP.FP, which was analyzed only with female parent information and ABLUP.MP, which was analyzed only with male parent information was compared to ABLUP and GLBUP. In addition, in order to examine the impact of the composition of families included in the training and test set on the prediction accuracy of GS, the 5-fold cross-validation where the families were randomly divided (CV_r), and the 5-fold cross-validation where the families were evenly distributed (CV_f), were performed and compared to each other. Then, the progenies of the mother trees of Gyeonggi1 (GG1) and Kyeongbuk4 (KB4) were predicted by training each other.

Also, the full-sib population was predicted by training the half-sib population that was established in the same sites for comparison. The half-sib population consists of 71 trees of 5 open-pollinated families, including 13 progenies of GG1, 14 progenies of KB4, 14 progenies of Gangwon30 (GW30), 16 progenies of KB15, and 13 progenies of KB50. In the total of 1,547 markers that were selected with quality of default value and MAF of 0.05 in the half-sib population were used. For BLUP analysis, as in chapters 1 and 2, the BreedR package of R was used.

3.3.2. Validation of model

The genotypes of the full-sib population were investigated for 6,464 markers (MAF $\geq$ 0.05) selected during the training of the half-sib population in chapter 2. Afterward, the GS model trained with the half-sib population was verified with the full-sib population. Also, referring to chapter 1, the subset of markers commonly selected from both populations and subsets of all markers selected from both populations were used for GBLUP to enhance the efficiency. The verification of the model was performed in as same as the GBLUP training and testing process in chapter 2. For validation, genomic selection accuracy (GSAC) and predictive ability (PA) were evaluated. As in chapters 1 and 2, GBLUP was performed using the BreedR package of the R program.

3.4. Result and discussion

3.4.1. Prediction accuracy in full-sib population

The 5-fold cross-validation of GBLUP was performed in '10 control-pollinated progeny trial to measure the prediction accuracy of GS in the full-sib population. As a result of CV_r, where cross-validation groups were randomly divided regardless of full-sib families, the accuracy of GBLUP was 0.766~0.781, and the predictive ability was 0.408~0.463 (Table 3-1). To investigate the impact of the family composition of the training and test set, CV_f that included all families evenly in each cross-validation group was performed (Table 3-2). As a result, CV_f showed higher accuracy and predictive ability than CV_r on average, but mostly showed only differences within the error range. However, when the progeny of GG1 and the progeny of KB4 were trained and tested each other, the prediction accuracy of GBLUP was significantly lowered to 0.028~0.257 for accuracy and 0.013~0.121 for predictive ability (Table 3-3).

Table 3-1. Prediction accuracy of ABLUP and GBLUP using random 5-fold CV (CV_r) for the control-pollinated population.

Trait	Accuracy		Predictive ability	
	ABLUP	GBLUP	ABLUP	GBLUP
DBH	0.888 (0.006)	0.768 (0.021)	0.412 (0.015)	0.444 (0.017)
Height	0.909 (0.005)	0.781 (0.012)	0.421 (0.021)	0.408 (0.023)
Volume	0.883 (0.004)	0.766 (0.022)	0.426 (0.015)	0.463 (0.012)

Mean (standard error) of 5-fold cross-validation

Table 3-2. Prediction accuracy of ABLUP and GBLUP using family distributed 5-fold CV (CV_f) for the control-pollinated population.

Trait	Accuracy		Predictive ability	
	ABLUP	GBLUP	ABLUP	GBLUP
DBH	0.890 (0.010)	0.780 (0.011)	0.418 (0.048)	0.456 (0.031)
Height	0.908 (0.004)	0.784 (0.006)	0.417 (0.027)	0.406 (0.019)
Volume	0.884 (0.011)	0.779 (0.012)	0.430 (0.046)	0.475 (0.025)

Mean (standard error) of 5-fold cross-validation

On the other hand, the prediction accuracy of GS in the full-sib population was significantly high compared to the half-sib population. The accuracy of the full-sib population was 1.6 times higher than that of the half-sib population with the value of 0.164~0.498 (Table 2-2, 3-1).

Previous studies conducted in forest trees also reported that the accuracy of GS was higher in the full-sib population than the half-sib population. In the study of Norway spruce, the full-sib scenario in which the training and test population were composed of identical families and the half-sib scenario in which the training and test population included individuals with the same female parents were compared (Chen et al. 2018). The accuracy and predictive ability of Norway spruce were higher in the full-sib scenario than in the half-sib scenario. In addition, the same level of accuracy as the half-sib population was obtained using fewer markers in the full-sib population. Similar results were also found in black spruce, where not only the accuracy and predictive ability were decreased in the half-sib population, but also the error was larger in the half-sib population compared to the full-sib population (Lenz et al., 2017). Also, in the study of the white spruce, when the GS was performed in half-sib and full-sib population whose parents were shared, the predictive ability of the full-sib was higher (Lenz et al., 2020). Although the

previous studies showed that the full-sib population had higher prediction accuracy, the efficiency against the pedigree-based ABLUP was similar so that GS was able to be applied to both populations.

The prediction accuracies of GBLUP and three types of ABLUP with different amounts of pedigree information were compared (Figure 3-1). ABLUP.FP and ABLUP.MP, which used only female and male parent information, respectively, showed lower accuracy and predictive ability than ABLUP with full pedigree. Prediction accuracy of ABLUP.FP was higher than that of ABLUP.MP, which was consistent with the result that the difference according to the female parent was larger than the difference according to the male parent in the phenotypic analysis in chapter 1 (Figure 1-6). On the other hand, in the comparison between ABLUP and GBLUP, the accuracy of ABLUP and predictive ability of GBLUP were high, as in the half-sib population. Therefore, the full-sib population could be used for the verification of the GS model of Korean red pine trained with the half-sib population.

Table 3-3. Prediction accuracy of ABLUP and GBLUP of two genomic selection scenarios.

GS scenario*	Trait	Accuracy		Predictive ability	
		ABLUP	GBLUP	ABLUP	GBLUP
KB4 → GG1	DBH	0.439	0.035	0.130	0.073
	Height	0.466	0.072	0.041	0.030
	Volume	0.572	0.069	0.187	0.068
GG1 → KB4	DBH	0.493	0.257	0.136	0.121
	Height	0.181	0.028	0.025	0.013
	Volume	0.494	0.198	0.163	0.095

*Training set → test set

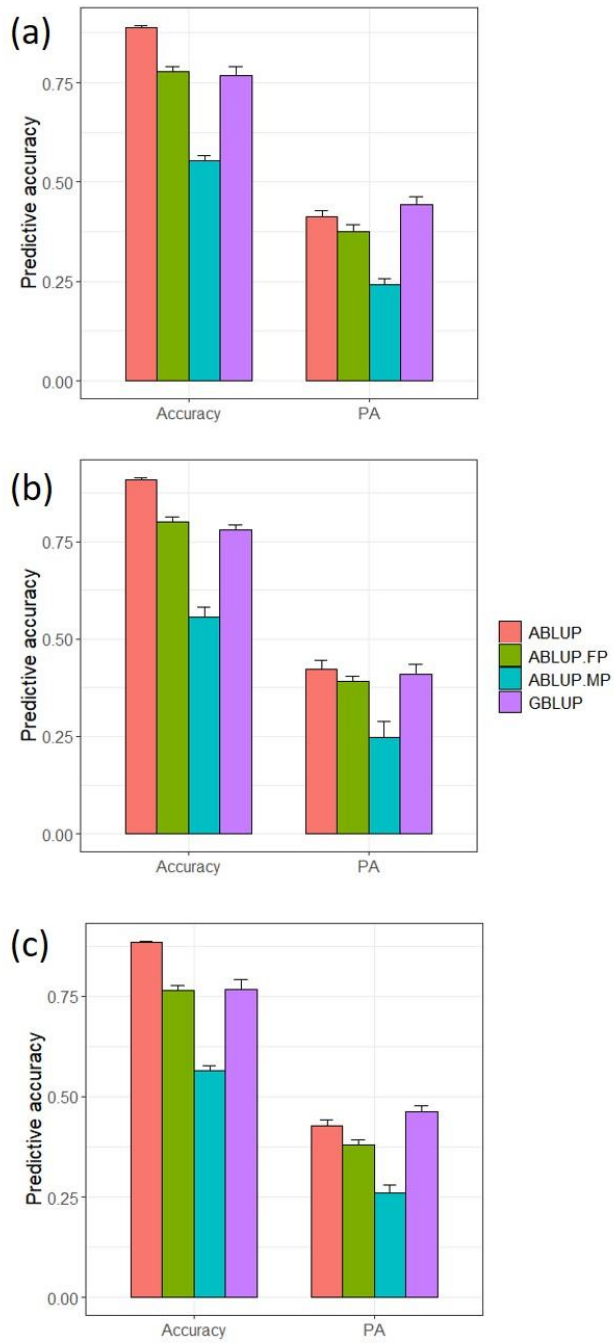

Figure 3-1. Prediction accuracy by pedigree and marker information of control-pollinated population for three traits. (a) DBH (b) height and (c) volume. ABLUP.FP is ABLUP where only female parent information is used. ABLUP.MP is ABLUP where only male parent information is used. Bar and error bar are mean and standard error of accuracy and predictive ability from 5-fold cross-validation.

Table 3-4. GBLUP prediction accuracy of the control-pollinated population when open-pollinated progeny in the same trial was trained.

DBH		Height		Volume	
GSAC [a]	PA [b]	GSAC	PA	GSAC	PA
0.675	0.308	0.597	0.262	0.707	0.344

[a] GSAC, GS accuracy, r(GEBV, EBV), EBV was estimated by GBLUP with all phenotype data of open- and control-pollinated population.
[b] PA, predictive ability, r(GEBV, phenotype)

The results of training with the half-sib population and testing the full-sib population in the same test site (Hwaseong) were 0.597~0.707 for accuracy and 0.262~0.344 for predictive ability (Table 3-4). It was a lower predictive ability compared to the results of 0.408~0.463 for the cross-validation in the full-sib population (Table 3-1). Compared with the average of within-region predictive abilities in the open-pollinated progeny trial (0.19 for DBH, 0.282 for height, 0.201 for volume), the predictive ability was high in DBH and volume (Table 2-2). Predictive ability remained high even though the half-sib population used for training included the progeny of GW30, KB15, and KB50, which were not used as parents of the full-sib population. Therefore, it was possible that the open-pollinated population could be trained to predict the GEBV of the control-pollinated population.

3.4.2. Validation of genomic selection model of Korean red pine

In chapter 2, the genotypes of ’10 control-pollinated progeny trial were analyzed with 6,464 markers selected based on the ’87 open-pollinated progeny trial. As a result, 1,957 monomorphic genotypes and 4,507 polymorphic genotypes were observed (Table 3-5). In addition, the higher the MAF was, the lower the marker frequency was, but the frequency decreased rapidly after MAF 0.15 (Figure 3-2). Also, the mean of the MAF of all

markers was 0.191.

Table 3-5. SNP genotypes of 6,464 markers in open- and control-pollinated population.

Genotype		'87 open-pollinated population	'10 control-pollinated population
Monomorphic	AA	-	555
	CC	-	374
	GG	-	389
	TT	-	639
Polymorphic	A/C	741	526
	A/G	1,979	1,392
	A/T	369	248
	C/G	602	487
	C/T	2,000	1,360
	G/T	773	494
Total		6,464	6,464

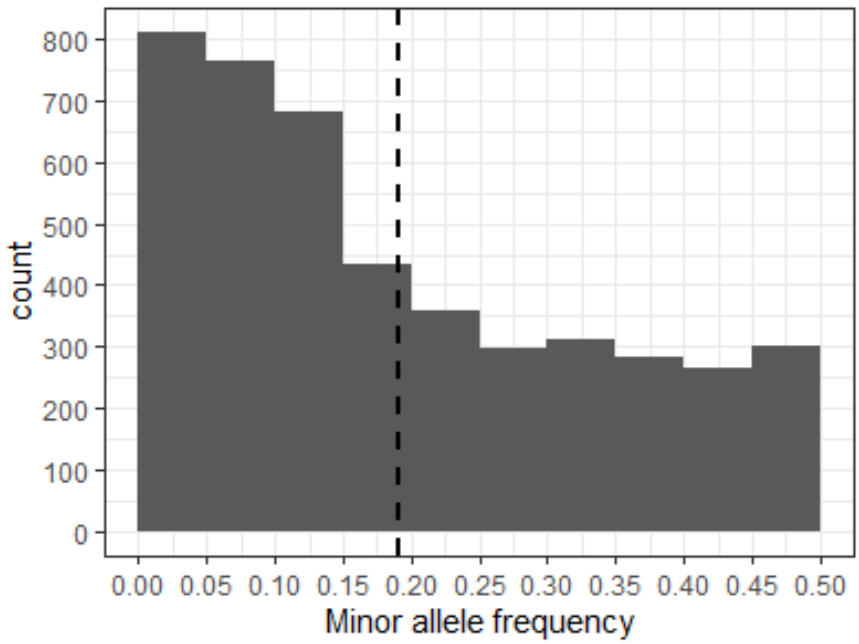

Figure 3-2. Histogram of minor allele frequencies of 4,507 polymorphic markers in the control-pollinated test population. The dashed line means the mean of minor allele frequencies.

In chapter 2, estimated breeding value (EBV) calculated through ABLUP was used as a standard for accuracy as in most other studies in order to compare the result with previous literature. However, the pedigree-based approach uses a kinship relationship matrix that has a lot of sources of error, whereas the marker-based approach is concerned with identity-by-descent (IBD) and identity-by-state (IBS) between individuals. This is the reason why marker-based estimates are accurate and close to the true breeding value (TBV) (Ukrainetz and Mansfield, 2020a; 2020b). Therefore, in the step to verify the actual application of GS, GSAC where EBV was estimated with GBLUP and all of phenotypic data was used. As the training population and test population for the validation of the GS model were not related to each other, the prediction by ABLUP was impossible.

As the result of the verification of the GS model of Korean red pine, the

prediction accuracy when the model was trained by a single half-sib population varied (Table 3-6). Hwaseong is bound to the same genetic region as Gongju, Kyeongju, Naju, and Wanju (Ahn et al., 2021). Even when GS model was trained with these regions, the prediction accuracy was not always improved.

However, when all test sites were trained, a high GSAC of 0.501~0.896 was shown (Table 3-6). As discussed in chapter 2, the highest prediction accuracy was obtained when the training population included various environments. It was a high accuracy result even compared with the results of the early study using simulation data (GSAC 0.17~0.51) in the loblolly pines (Resende Jr et al., 2012a). In particular, it is meaningful that GSAC of volume of Korean red pine showed a higher value (0.896) than that of lodgepole pine reaching 0.82~0.83 (Ukrainetz and Mansfield, 2020a; 2020b) and that of white spruce reaching 0.85 (Beaulieu et al. 2020).

The predictive ability was low at 0.054 to 0.208 when all the test sites were trained, but considering the low heritability of 0.206~0.251, it was difficult to say that the predictive ability of the model was lowered (Table 1-7, Table 3-6). Therefore, it was concluded that it was possible to efficiently predict the GEBV between two genetically distant populations using GS in Korean red pine.

Table 3-6. GBLUP prediction accuracy of the control-pollinated population when the model was trained by the single or combined site of open-pollinated population.

Trained population	DBH		Height		Volume	
	GSAC [a]	PA [b]	GSAC	PA	GSAC	PA
Taean	0.296	0.058	-0.505	-0.270	0.313	0.021
Chuncheon	0.307	0.047	0.390	0.100	0.585	0.102
Gongju	0.505	0.253	0.267	0.071	0.390	0.173
Kyeongju	0.560	0.276	0.678	0.262	0540	0.286
Naju	-0.240	-0.112	-0.414	-0.253	-0.112	-0.156
Wanju	-0.028	-0.030	-0.347	0.092	0.201	-0.047
Combined	0.657	0.208	0.501	0.054	0.896	0.181

[a] GSAC, GS accuracy, r(GEBV, EBV), EBV was estimated by GBLUP with all phenotype data of open- and control-pollinated population.
[b] PA, predictive ability, r(GEBV, phenotype)

In the previous polymorphism analysis, about 30% of markers were monomorphic (Table 3-5). In order to further improve the prediction accuracy of two genetically distant populations, it is necessary to use polymorphic markers in the test population as well. The numbers of markers filtered by the default quality and MAF of 0.05 in chapter 1 were 1,164 for the half-sib population and 1,277 for the full-sib population (Tables 1-2, 1-3). Among them, 875 overlapping markers and 1,566 combined markers were used for GBLUP (Table 3-7). As a result, the accuracy was 0.675~0.952 and the predictive ability was 0.122~0.227, which were enhanced values compared to the result of using markers that were polymorphic only in the half-sib population (Tables 3-6, 3-7). Therefore, it was found that it would be advantageous to use polymorphic markers by analyzing the genotypes of the test population as well as the training population for improvement of the prediction accuracy.

Table 3-7. GBLUP prediction accuracy of the control-pollinated population when the model was trained by combined site using common and total marker set.

Markers	DBH		Height		Volume	
	GSAC [a]	PA [b]	GSAC	PA	GSAC	PA
Common (875)	0.737	0.226	0.678	0.122	0.952	0.207
Total (1,566)	0.675	0.227	0.676	0.169	0.907	0.192

[a] GSAC, GS accuracy, r(GEBV, EBV), EBV was estimated by GBLUP with all phenotype data of open- and control-pollinated population.
[b] PA, predictive ability, r(GEBV, phenotype)

3.4.3. Selection based on genomic estimated breeding value

Individual selection of 10% intensity in '10 control-pollinated progeny trial was conducted for volume based on GEBV using the GS model trained in this study (Table 3-8). Selected 69 individuals belonged to 18 full-sib families, which was much greater number of families compared to 3 families that would have been selected if 10% selection had been carried out in family selection (FS) (Table 3-9). Therefore, it was considered that genetic diversity could be raised by individual selection using GS compared to FS. In addition, half of the selected 69 individuals belonged to 8 families of combinations of two superior clones according to general combining ability (GCA) analysis, indicating that the GS was coincident with open-pollinated progeny tests (NIFoS, 2009).

Table 3-8. Sixty-nine trees selected from '10 control-pollinated progeny trial based on GEBV for volume

Rank	GEBV (m^3)	Family	Rank	GEBV(m^3)	Family
1	0.0930	KB4×GW69	36	0.0617	KB4×KB26
2	0.0904	KB4×GW40	37	0.0604	KB4×GW44
3	0.0863	KB4×GW40	38	0.0604	KB4×KB26
4	0.0832	KB4×GW42	39	0.0602	KB4×CB3
5	0.0779	KB4×KB20	40	0.0590	KB4×KB4
6	0.0779	KB4×GW69	41	0.0587	KB4×GW40
7	0.0762	GG1×GW84	42	0.0585	KB4×GW69
8	0.0761	KB4×GW69	43	0.0584	KB4×GW69
9	0.0714	KB4×GW69	44	0.0584	KB4×KB5
10	0.0705	KB4×GW39	45	0.0583	KB4×GW69
11	0.0705	KB4×GW42	46	0.0577	KB4×GW69
12	0.0695	KB4×GW42	47	0.0570	KB4×GW42
13	0.0691	KB4×KB5	48	0.0563	KB4×GW40
14	0.0690	KB4×GW69	49	0.0561	KB4×KB33
15	0.0685	KB4×GW69	50	0.0559	KB4×GW39
16	0.0677	KB4×KB5	51	0.0558	GG1×CB3
17	0.0675	KB4×KB26	52	0.0556	KB4×GW40
18	0.0670	GG1×KB20	53	0.0548	KB4×GW42
19	0.0666	KB4×KB33	54	0.0545	KB4×KB26
20	0.0665	KB4×KB20	55	0.0544	KB4×KB26
21	0.0662	KB4×KB5	56	0.0538	KB4×GW40
22	0.0657	KB4×KB5	57	0.0538	KB4×KB26
23	0.0656	KB4×KB33	58	0.0533	KB4×GW43
24	0.0653	KB4×KB20	59	0.0533	KB4×KB20
25	0.0652	GG1×GW76	60	0.0532	KB4×GW40
26	0.0650	GG1×GW84	61	0.0527	KB4×KB20
27	0.0647	KB4×GW69	62	0.0525	KB4×GW69
28	0.0645	KB4×GW40	63	0.0524	KB4×KB4
29	0.0644	GG1×KB20	64	0.0523	KB4×GW40
30	0.0632	KB4×KB26	65	0.0521	KB4×GG1
31	0.0631	KB4×GW44	66	0.0520	KB4×KB20
32	0.0629	KB4×GW40	67	0.0520	KB4×GW84
33	0.0625	KB4×CB3	68	0.0515	KB4×GW43
34	0.0620	KB4×GG1	69	0.0511	KB4×CB3
35	0.0619	KB4×CB3			

* CB, GG, GW and KB mean that female parent is plus tree from Chungbuk, Gyeonggi, Gangwon and Kyeongbuk provenance, respectively.

Table 3-9. Family composition of selected trees.

Family [a]	Combination[b]	No.of selected trees	Family	Combination	No. of selected trees
GG1×GW76	I×I	1	KB4×GW69	S×S	12
GG1 ×GW84	I×S	2	KB4×GW84	S×S	1
GG1 ×CB3	I×S	1	KB4×GG1	S×I	2
GG1×KB20	I×S	2	KB4×CB3	S×S	4
KB4×GW39	S×I	2	KB4×KB20	S×S	6
KB4×GW40	S×I	10	KB4×KB26	S×I	7
KB4×GW42	S×I	5	KB4×KB33	S×S	3
KB4×GW43	S×I	2	KB4×KB4	S×S	2
KB4×GW44	S×S	2	KB4×KB5	S×S	5

[a] CB, GG, GW and KB mean that female parent is plus tree from Chungbuk, Gyeonggi, Gangwon and Kyeongbuk provenance, respectively.

[b] S and I mean superior and inferior open-pollinated families according to GCA analysis on progeny trials (NIFoS, 2009).

3.5. Conclusion

In this chapter, the GS model was verified in a new population, and selection using GEBV was performed. The higher the relationship within the population was, the higher the prediction accuracy was. However, when the trained GS model was applied to the unrelated population, the genomic selection accuracy was as high as the results in other tree species. Therefore, it was concluded that the GS could be widely applied to the selection of Korean red pine in breeding populations and even wild populations. The result of this chapter is meaningful because it is one of the few studies on the verification step of GS.

References

Ahn, J. Y., Lee, J.W. and Hong, K.N. 2021. Genetic Diversity and Structure of *Pinus densiflora* Siebold & Zucc. Populations in Republic of Korea Based on Microsatellite Markers. Forests 12(6): 750.

Beaulieu, J., Doerksen, T., Clément, S., MacKay, J. and Bousquet, J. 2014a. Accuracy of genomic selection models in a large population of open-pollinated families in white spruce. Heredity 113(4): 343-352.

Beaulieu, J., Nadeau, S., Ding, C., Celedon, J. M., Azaiez, A., Ritland, C., ... and Bousquet, J. 2020. Genomic selection for resistance to spruce budworm in white spruce and relationships with growth and wood quality traits. Evolutionary applications 13(10): 2704-2722.

Chen, Z. Q., Baison, J., Pan, J., Karlsson, B., Andersson, B., Westin, J., ... and Wu, H. X. 2018. Accuracy of genomic selection for growth and wood quality traits in two control-pollinated progeny trials using exome capture as the genotyping platform in Norway spruce. BMC genomics 19(1): 1-16.

Daetwyler, H. D., Calus, M. P., Pong-Wong, R., de Los Campos, G. and Hickey, J. M. 2013. Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. Genetics, 193(2): 347-365.

Grattapaglia, D. 2017. Status and perspectives of genomic selection in forest tree breeding. In Genomic selection for crop improvement (pp. 199-249). Springer, Cham.

Heslot, N., Jannink, J. L. and Sorrells, M. E. 2015. Perspectives for genomic selection applications and research in plants. Crop Science 55(1): 1-12.

Lenz, P., Beaulieu, J., Mansfield, S. D., Clément, S., Desponts, M. and Bousquet, J. 2017. Factors affecting the accuracy of genomic selection for growth and wood quality traits in an advanced-breeding population of black spruce (Picea mariana). BMC genomics 18(1): 1-17.

Lenz, P., Nadeau, S., Azaiez, A., Gérardi, S., Deslauriers, M., Perron, M., ... and Bousquet, J. 2020. Genomic prediction for hastening and improving efficiency of forward selection in conifer polycross mating designs: an example from white spruce. Heredity 124(4): 562-578.

Lin, Z., Hayes, B. J. and Daetwyler, H. D. 2014. Genomic selection in crops, trees and forages: a review. *Crop and Pasture Science 65*(11): 1177-1191.

National Institute of Forest Science (NIFoS). 2009. Genetic test of timber species. Research Reports no. 09-12. Seoul, Korea.

Resende Jr, M. F. R., Munoz, P., Resende, M. D., Garrick, D. J., Fernando, R. L., Davis, J. M., ... and Kirst, M. 2012a. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). Genetics 190(4): 1503-1510.

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J. L. and Melchinger, A. E. 2013. Genomic predictability of interconnected biparental maize populations. Genetics 194(2): 493-503.

Ukrainetz, N. K. and Mansfield, S. D. 2020a. Prediction accuracy of single-step BLUP for growth and wood quality traits in the lodgepole pine breeding program in British Columbia. Tree Genetics & Genomes 16(5): 1-13.

Ukrainetz, N. K. and Mansfield, S. D. 2020b. Assessing the sensitivities of genomic selection for growth and wood quality traits in lodgepole pine using Bayesian models. *Tree Genetics & Genomes 16*(1): 1-19.

Van Eenennaam, A. L., Weigel, K. A., Young, A. E., Cleveland, M. A. and Dekkers, J. C. 2014. Applied animal genomics: results from the field. Annu. Rev. Anim. Biosci. 2(1): 105-139.

General conclusion

This study evaluated the efficiency of genomic selection (GS) as a selective breeding method for the growth of Korean red pine whose accelerated breeding is required due to high wood demand. To this end, a GS model was trained through the cross-validations in an open-pollinated progeny trial, and the model was verified with a control-pollinated progeny trial.

The GS in Korean red pine showed a high selection efficiency compared to phenotypic selection (PS) and family selection (FS). In this study, both the half-sib population and full-sib population were used as materials, and the test sites in 7 regions across South Korea were included. Also, it was revealed that in the verification stage, it is possible to predict genomic estimated breeding value (GEBV) between the half-sib population and the full-sib population who do not share their parents. Therefore, it was concluded that the GS model presented in this study could be utilized widely for the breeding population of Korean red pine in South Korea. However, as GS predictive models including deep learning are constantly being developed and the effort to take the interaction of genotype and environment into consideration has just started, further model research to improve the efficiency of GS in Korean red pine should be continued.

Currently, the development of single nucleotide polymorphism (SNP) chips for investigating genotypes that could be applied to the molecular breeding of Korean red pine is still in progress. In this study, the first implementation of the SNP chip made with genotype-by-sequencing (GBS) data of Korean red pine, the polymorphism of the genotypes was found to be lower than required for widespread use in practice. Because the genome of

conifer is large and repetitive, a multi-faceted approach for SNP array would be helpful. Therefore, for future various research on the genome of Korean red pine including GS, the SNP chip needs to be advanced by combining GBS which is efficient for handling large genomes, transcriptome sequencing containing expressed gene information, and re-sequencing suitable for obtaining accurate SNP data from genome-wide. In addition, including the populations with diverse genetic backgrounds as materials in the development of the SNP chip would improve its versatility.

Although this study focused on exploring the optimal GS model in Korean red pine and verifying its applicability, the final goal is to select individuals with high breeding values for the establishment of the advanced generation based on this. Therefore, future research on GS of Korean red pine should proceed in the direction of verifying whether the progenies of individuals selected by GEBV show excellent performances.

Appendix

Table S1. Open-pollinated families included in '87 progeny test

Family [a]	Number of trees		Family	Number of trees	
	phenotyped	genotyped		phenotyped	genotyped
GW92	53	17	KB72	142	122
GW99	103	84	KB75	151	82
GW109	84	65	KB77	135	108
GW119	86	71	KB78	110	60
GW122	18	7	KB80	84	63
GW124	139	114	KB81	80	37
GW139	86	49	KB82	96	54
GW140	132	87	KB83	82	64
GW141	90	48	KB86	66	47
GW149	26	15	KB87	29	19
GW151	138	111	KB88	74	49
GW154	96	73	KB89	10	8
GW155	137	83	KB92	77	61
GW156	64	46	KB95	90	80
GW157	82	70	KB96	99	73
GW158	119	74	KB97	114	74
GW160	141	83	KB98	72	56
KB48	22	12	KB99	78	51
KB53	20	11	KB100	90	68
KB66	11	4	KB101	124	74
KB67	88	67	KB102	81	63
KB68	89	65	KB103	112	72

Total	phenotyped	genotyped
	3,820	2,643

[a] GW and KB mean that female parent is plus tree from Gangwon and Kyeongbuk provenance, respectively.

Table S2. Control-pollinated families included in '10 progeny test

Family [a]	Number of trees		Family	Number of trees	
	phenotyped	genotyped		phenotyped	genotyped
GG1×GW39	13	12	KB4×GW39	9	9
GG1×GW40	7	7	KB4×GW40	86	84
GG1×GW42	6	6	KB4×GW42	49	49
GG1×GW43	1	1	KB4×GW43	13	14
GG1×GW44	6	5	KB4×GW44	10	10
GG1×GW69	4	4	KB4×GW69	31	31
GG1×GW76	26	26	KB4×GW84	20	20
GG1×GW84	18	17	KB4×GG1	27	25
GG1×GG1	7	7	KB4×CB3	20	20
GG1×CB3	14	14	KB4×KB20	74	73
GG1×KB20	14	14	KB4×KB26	46	43
GG1×KB26	18	17	KB4×KB33	90	88
GG1×KB33	4	4	KB4×KB4	17	17
GG1×KB4	8	9	KB4×KB5	56	56
GG1×KB5	9	9			

Total	phenotyped	genotyped
	703	691

[a] CB, GG, GW and KB mean that female parent is plus tree from Chungbuk, Gyeonggi, Gangwon and Kyeongbuk provenance, respectively.

Table S3. GBLUP accuracy and predictive ability according to the marker quality threshold

Trait	Site [a]	Accuracy			Predictive ability		
		loose	moderate	strict	loose	moderate	strict
DBH	T	0.29 (0.04)	0.28 (0.04)	0.22 (0.04)	0.11 (0.04)	0.05 (0.05)	0.02 (0.04)
	C	0.5 (0.04)	0.44 (0.03)	0.38 (0.04)	0.29 (0.05)	0.21 (0.04)	0.19 (0.05)
	G	0.33 (0.04)	0.31 (0.03)	0.24 (0.04)	0.18 (0.05)	0.17 (0.04)	0.11 (0.05)
	K	0.25 (0.06)	0.22 (0.08)	0.17 (0.07)	0.05 (0.05)	0.05 (0.08)	0.03 (0.08)
	N	0.44 (0.05)	0.36 (0.04)	0.37 (0.03)	0.32 (0.06)	0.25 (0.03)	0.24 (0.03)
	W	0.25 (0.05)	0.07 (0.05)	0.02 (0.06)	0.19 (0.06)	-0.03 (0.05)	-0.08 (0.06)
Height	T	0.5 (0.04)	0.33 (0.04)	0.3 (0.04)	0.44 (0.04)	0.16 (0.03)	0.15 (0.03)
	C	0.42 (0.01)	0.36 (0.03)	0.33 (0.02)	0.26 (0.03)	0.2 (0.04)	0.18 (0.03)
	G	0.48 (0.04)	0.37 (0.05)	0.35 (0.04)	0.41 (0.05)	0.26 (0.05)	0.26 (0.05)
	K	0.25 (0.06)	0.07 (0.07)	0.03 (0.06)	0.19 (0.07)	-0.02 (0.08)	-0.05 (0.07)
	N	0.44 (0.04)	0.43 (0.03)	0.31 (0.03)	0.32 (0.05)	0.3 (0.04)	0.21 (0.03)
	W	0.16 (0.04)	0.17 (0.05)	0.12 (0.06)	0.06 (0.06)	0.05 (0.04)	0.03 (0.05)
Straight-ness	T	0.32 (0.04)	0.25 (0.04)	0.2 (0.03)	0.1 (0.04)	0.01 (0.05)	0 (0.04)
	C	0.43 (0.03)	0.37 (0.03)	0.31 (0.03)	0.17 (0.03)	0.11 (0.03)	0.1 (0.04)
	G	0.34 (0.03)	0.28 (0.03)	0.22 (0.03)	0.21 (0.04)	0.16 (0.03)	0.11 (0.04)
	K	0.36 (0.05)	0.34 (0.04)	0.34 (0.05)	0.21 (0.05)	0.19 (0.05)	0.2 (0.05)
	N	0.29 (0.05)	0.21 (0.06)	0.16 (0.07)	0.18 (0.05)	0.11 (0.06)	0.06 (0.07)
	W	0.26 (0.06)	0.26 (0.05)	0.28 (0.05)	0.17 (0.05)	0.17 (0.04)	0.19 (0.05)
Volume	T	0.34 (0.05)	0.33 (0.04)	0.29 (0.04)	0.18 (0.05)	0.11 (0.05)	0.09 (0.04)
	C	0.49 (0.03)	0.43 (0.03)	0.37 (0.04)	0.3 (0.04)	0.21 (0.03)	0.19 (0.04)
	G	0.36 (0.03)	0.34 (0.04)	0.29 (0.05)	0.22 (0.05)	0.21 (0.05)	0.17 (0.06)
	K	0.24 (0.07)	0.24 (0.08)	0.17 (0.07)	0.02 (0.06)	0.05 (0.09)	0.02 (0.09)
	N	0.44 (0.05)	0.37 (0.04)	0.36 (0.03)	0.31 (0.05)	0.24 (0.03)	0.19 (0.03)
	W	0.23 (0.05)	0.07 (0.04)	-0.02 (0.05)	0.18 (0.06)	-0.01 (0.03)	-0.09 (0.05)

Mean (standard error) of accuracy and predictive ability from 10-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S4. GBLUP accuracy and predictive ability according to the number of randomly selected markers

Trait	Site [a]	No. of markers			
		2K	6K	10K	17K
Accuracy					
DBH	T	0.17 (0.03)	0.27 (0.03)	0.28 (0.03)	0.28 (0.04)
	C	0.35 (0.04)	0.47 (0.03)	0.48 (0.04)	0.5 (0.04)
	G	0.18 (0.05)	0.26 (0.05)	0.28 (0.04)	0.31 (0.04)
	K	0.16 (0.04)	0.23 (0.06)	0.25 (0.06)	0.25 (0.06)
	N	0.38 (0.07)	0.42 (0.06)	0.44 (0.05)	0.44 (0.05)
	W	0.23 (0.05)	0.22 (0.06)	0.25 (0.05)	0.26 (0.05)
Height	T	0.46 (0.03)	0.46 (0.04)	0.49 (0.04)	0.51 (0.03)
	C	0.34 (0.03)	0.38 (0.02)	0.42 (0.01)	0.42 (0.01)
	G	0.37 (0.06)	0.43 (0.04)	0.44 (0.06)	0.46 (0.04)
	K	0.11 (0.07)	0.18 (0.06)	0.24 (0.06)	0.24 (0.06)
	N	0.37 (0.06)	0.4 (0.03)	0.41 (0.05)	0.45 (0.04)
	W	0.09 (0.06)	0.18 (0.05)	0.19 (0.06)	0.22 (0.06)
Straight-ness	T	0.24 (0.04)	0.31 (0.04)	0.32 (0.04)	0.32 (0.04)
	C	0.32 (0.04)	0.37 (0.03)	0.41 (0.03)	0.43 (0.03)
	G	0.24 (0.04)	0.31 (0.03)	0.34 (0.03)	0.35 (0.04)
	K	0.25 (0.04)	0.31 (0.05)	0.32 (0.05)	0.34 (0.04)
	N	0.15 (0.03)	0.21 (0.05)	0.24 (0.05)	0.26 (0.05)
	W	0.23 (0.06)	0.29 (0.06)	0.29 (0.06)	0.26 (0.06)
Volume	T	0.24 (0.04)	0.3 (0.04)	0.32 (0.03)	0.33 (0.04)
	C	0.36 (0.04)	0.45 (0.03)	0.47 (0.03)	0.49 (0.03)
	G	0.23 (0.03)	0.31 (0.04)	0.32 (0.04)	0.35 (0.03)
	K	0.13 (0.05)	0.21 (0.07)	0.24 (0.07)	0.24 (0.07)
	N	0.35 (0.07)	0.43 (0.05)	0.44 (0.05)	0.45 (0.04)
	W	0.21 (0.04)	0.21 (0.05)	0.25 (0.05)	0.26 (0.05)
Predictive ability					
DBH	T	0 (0.03)	0.1 (0.03)	0.07 (0.03)	0.09 (0.03)
	C	0.21 (0.05)	0.28 (0.04)	0.3 (0.04)	0.31 (0.05)
	G	0.1 (0.05)	0.1 (0.07)	0.12 (0.05)	0.16 (0.05)
	K	0 (0.04)	0.05 (0.05)	0.04 (0.05)	0.05 (0.05)
	N	0.33 (0.06)	0.31 (0.06)	0.32 (0.05)	0.32 (0.06)
	W	0.18 (0.07)	0.18 (0.06)	0.19 (0.06)	0.2 (0.06)

Table S4. (Continued)

Trait	Site [a]	No. of markers			
		2K	6K	10K	17K
Height	T	0.43 (0.04)	0.42 (0.04)	0.44 (0.04)	0.46 (0.04)
	C	0.21 (0.02)	0.24 (0.03)	0.27 (0.03)	0.27 (0.03)
	G	0.33 (0.07)	0.35 (0.05)	0.37 (0.06)	0.4 (0.05)
	K	0.08 (0.06)	0.08 (0.07)	0.18 (0.07)	0.16 (0.07)
	N	0.27 (0.06)	0.28 (0.05)	0.28 (0.06)	0.33 (0.05)
	W	0 (0.06)	0.08 (0.06)	0.09 (0.07)	0.11 (0.07)
Straight-ness	T	0.06 (0.04)	0.15 (0.04)	0.11 (0.04)	0.11 (0.04)
	C	0.15 (0.04)	0.13 (0.02)	0.17 (0.03)	0.17 (0.03)
	G	0.14 (0.04)	0.2 (0.04)	0.21 (0.03)	0.21 (0.04)
	K	0.11 (0.05)	0.17 (0.05)	0.17 (0.05)	0.18 (0.05)
	N	0.05 (0.04)	0.11 (0.06)	0.12 (0.05)	0.14 (0.06)
	W	0.15 (0.06)	0.24 (0.05)	0.21 (0.05)	0.17 (0.06)
Volume	T	0.1 (0.04)	0.15 (0.04)	0.15 (0.04)	0.16 (0.05)
	C	0.21 (0.04)	0.28 (0.04)	0.31 (0.04)	0.32 (0.04)
	G	0.15 (0.05)	0.16 (0.06)	0.17 (0.05)	0.21 (0.05)
	K	-0.03 (0.05)	0.01 (0.07)	0.01 (0.07)	0.02 (0.06)
	N	0.27 (0.05)	0.31 (0.05)	0.31 (0.04)	0.32 (0.05)
	W	0.16 (0.06)	0.18 (0.06)	0.19 (0.06)	0.2 (0.05)

Mean (standard error) of accuracy and predictive ability from 10-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S5. GBLUP accuracy and predictive ability according to the marker selection based on minor allele frequency

Trait	Site [a]	Minor allele frequency			
		maf≥0.25	0.25>maf ≥0.05	0.05>maf ≥0.0005	0.0005>maf >0
Accuracy					
DBH	T	0.27 (0.04)	0.29 (0.04)	0.28 (0.04)	0.28 (0.04)
	C	0.48 (0.03)	0.5 (0.04)	0.5 (0.04)	0.5 (0.04)
	G	0.3 (0.04)	0.33 (0.04)	0.31 (0.04)	0.31 (0.04)
	K	0.21 (0.07)	0.25 (0.06)	0.25 (0.06)	0.25 (0.06)
	N	0.41 (0.06)	0.44 (0.05)	0.44 (0.05)	0.44 (0.05)
	W	0.25 (0.05)	0.25 (0.05)	0.26 (0.05)	0.26 (0.05)
Height	T	0.43 (0.03)	0.5 (0.04)	0.51 (0.03)	0.51 (0.03)
	C	0.39 (0.02)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)
	G	0.43 (0.05)	0.48 (0.04)	0.46 (0.04)	0.46 (0.04)
	K	0.14 (0.07)	0.25 (0.06)	0.24 (0.06)	0.24 (0.06)
	N	0.4 (0.04)	0.44 (0.04)	0.45 (0.04)	0.45 (0.04)
	W	0.15 (0.05)	0.16 (0.04)	0.22 (0.06)	0.22 (0.06)
Straight -ness	T	0.31 (0.03)	0.32 (0.04)	0.32 (0.04)	0.32 (0.04)
	C	0.42 (0.03)	0.43 (0.03)	0.43 (0.03)	0.43 (0.03)
	G	0.32 (0.04)	0.34 (0.03)	0.35 (0.04)	0.35 (0.04)
	K	0.32 (0.05)	0.36 (0.05)	0.34 (0.04)	0.34 (0.04)
	N	0.24 (0.06)	0.29 (0.05)	0.26 (0.05)	0.26 (0.05)
	W	0.21 (0.06)	0.26 (0.06)	0.26 (0.06)	0.26 (0.06)
Volume	T	0.31 (0.04)	0.34 (0.05)	0.33 (0.04)	0.33 (0.04)
	C	0.47 (0.03)	0.49 (0.03)	0.49 (0.03)	0.49 (0.03)
	G	0.33 (0.04)	0.36 (0.03)	0.35 (0.03)	0.35 (0.03)
	K	0.22 (0.07)	0.24 (0.07)	0.24 (0.07)	0.24 (0.07)
	N	0.41 (0.05)	0.44 (0.05)	0.45 (0.04)	0.45 (0.04)
	W	0.22 (0.05)	0.23 (0.05)	0.26 (0.05)	0.26 (0.05)
Predictive ability					
DBH	T	0.08 (0.03)	0.11 (0.04)	0.09 (0.03)	0.09 (0.03)
	C	0.27 (0.05)	0.29 (0.05)	0.31 (0.05)	0.31 (0.05)
	G	0.16 (0.05)	0.18 (0.05)	0.16 (0.05)	0.16 (0.05)
	K	-0.01 (0.06)	0.05 (0.05)	0.05 (0.05)	0.05 (0.05)
	N	0.3 (0.06)	0.32 (0.06)	0.32 (0.06)	0.32 (0.06)
	W	0.17 (0.06)	0.19 (0.06)	0.2 (0.06)	0.2 (0.06)

Table S5. (Continued)

Trait	Site [a]	Minor allele frequency			
		maf≥0.25	0.25>maf ≥0.05	0.05>maf ≥0.0005	0.0005>maf >0
Height	T	0.36 (0.04)	0.44 (0.04)	0.46 (0.04)	0.46 (0.04)
	C	0.23 (0.03)	0.26 (0.03)	0.27 (0.03)	0.27 (0.03)
	G	0.34 (0.06)	0.41 (0.05)	0.4 (0.05)	0.4 (0.05)
	K	0 (0.08)	0.19 (0.07)	0.16 (0.07)	0.16 (0.07)
	N	0.25 (0.05)	0.32 (0.05)	0.33 (0.05)	0.33 (0.05)
	W	0.04 (0.06)	0.06 (0.06)	0.11 (0.07)	0.11 (0.07)
Straight -ness	T	0.11 (0.04)	0.1 (0.04)	0.11 (0.04)	0.11 (0.04)
	C	0.15 (0.04)	0.17 (0.03)	0.17 (0.03)	0.17 (0.03)
	G	0.19 (0.04)	0.21 (0.04)	0.21 (0.04)	0.21 (0.04)
	K	0.17 (0.05)	0.21 (0.05)	0.18 (0.05)	0.18 (0.05)
	N	0.12 (0.06)	0.18 (0.05)	0.14 (0.06)	0.14 (0.06)
	W	0.13 (0.05)	0.17 (0.05)	0.17 (0.06)	0.17 (0.06)
Volume	T	0.13 (0.04)	0.18 (0.05)	0.16 (0.05)	0.16 (0.05)
	C	0.28 (0.05)	0.3 (0.04)	0.32 (0.04)	0.32 (0.04)
	G	0.18 (0.06)	0.22 (0.05)	0.21 (0.05)	0.21 (0.05)
	K	-0.02 (0.07)	0.02 (0.06)	0.02 (0.06)	0.02 (0.06)
	N	0.28 (0.05)	0.31 (0.05)	0.32 (0.05)	0.32 (0.05)
	W	0.15 (0.05)	0.18 (0.06)	0.2 (0.05)	0.2 (0.05)

Mean (standard error) of accuracy and predictive ability from 10-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S6. Accuracy and predictive ability according to predictive models

Trait	Site [a]	Predictive model [b]						
		ABLUP	GBLUP	BRR	BL	Bayes A	Bayes B	Bayes C
Accuracy								
DBH	T	0.68 (0.02)	0.29 (0.04)	0.28 (0.04)	0.3 (0.04)	0.28 (0.04)	0.28 (0.04)	0.28 (0.04)
	C	0.64 (0.02)	0.5 (0.04)	0.49 (0.04)	0.5 (0.04)	0.49 (0.04)	0.49 (0.04)	0.49 (0.04)
	G	0.51 (0.05)	0.33 (0.04)	0.33 (0.03)	0.32 (0.04)	0.33 (0.03)	0.33 (0.04)	0.33 (0.03)
	K	0.57 (0.05)	0.25 (0.06)	0.24 (0.05)	0.25 (0.06)	0.24 (0.06)	0.23 (0.06)	0.24 (0.06)
	N	0.55 (0.05)	0.44 (0.05)	0.44 (0.05)	0.45 (0.05)	0.44 (0.05)	0.44 (0.06)	0.44 (0.05)
	W	0.49 (0.04)	0.25 (0.05)	0.23 (0.05)	0.24 (0.05)	0.23 (0.04)	0.23 (0.05)	0.24 (0.05)
Height	T	0.59 (0.03)	0.5 (0.04)	0.5 (0.04)	0.5 (0.04)	0.5 (0.04)	0.5 (0.04)	0.5 (0.04)
	C	0.57 (0.03)	0.42 (0.01)	0.41 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)	0.42 (0.01)
	G	0.53 (0.05)	0.48 (0.04)	0.49 (0.04)	0.49 (0.04)	0.49 (0.04)	0.5 (0.04)	0.49 (0.04)
	K	0.32 (0.07)	0.25 (0.06)	0.24 (0.06)	0.24 (0.06)	0.24 (0.06)	0.25 (0.06)	0.24 (0.06)
	N	0.5 (0.03)	0.44 (0.04)	0.45 (0.04)	0.46 (0.04)	0.45 (0.04)	0.46 (0.04)	0.46 (0.04)
	W	0.42 (0.09)	0.16 (0.04)	0.17 (0.06)	0.16 (0.06)	0.16 (0.06)	0.15 (0.06)	0.17 (0.05)
Straight-ness	T	0.62 (0.02)	0.32 (0.04)	0.31 (0.03)	0.31 (0.03)	0.32 (0.03)	0.31 (0.03)	0.31 (0.04)
	C	0.72 (0.02)	0.43 (0.03)	0.41 (0.03)	0.42 (0.03)	0.42 (0.03)	0.4 (0.04)	0.41 (0.03)
	G	0.61 (0.02)	0.34 (0.03)	0.33 (0.04)	0.34 (0.03)	0.33 (0.03)	0.33 (0.03)	0.34 (0.03)
	K	0.56 (0.03)	0.36 (0.05)	0.35 (0.04)	0.36 (0.05)	0.36 (0.05)	0.36 (0.05)	0.35 (0.05)
	N	0.48 (0.04)	0.29 (0.05)	0.29 (0.05)	0.27 (0.05)	0.29 (0.05)	0.3 (0.05)	0.29 (0.05)
	W	0.52 (0.05)	0.26 (0.06)	0.27 (0.06)	0.27 (0.06)	0.26 (0.06)	0.26 (0.06)	0.26 (0.06)

Table S6. (Continued)

Trait	Site	Predictive model *						
		ABLUP	GBLUP	BRR	BL	Bayes A	Bayes B	Bayes C
Volume	T	0.68 (0.01)	0.34 (0.05)	0.33 (0.04)	0.33 (0.04)	0.33 (0.04)	0.32 (0.05)	0.33 (0.05)
	C	0.65 (0.02)	0.49 (0.03)	0.48 (0.03)	0.49 (0.03)	0.48 (0.03)	0.47 (0.03)	0.48 (0.03)
	G	0.52 (0.04)	0.36 (0.03)	0.36 (0.03)	0.35 (0.03)	0.36 (0.03)	0.35 (0.03)	0.36 (0.03)
	K	0.55 (0.05)	0.24 (0.07)	0.24 (0.06)	0.26 (0.06)	0.24 (0.06)	0.24 (0.06)	0.25 (0.06)
	N	0.56 (0.04)	0.44 (0.05)	0.43 (0.05)	0.45 (0.04)	0.44 (0.05)	0.43 (0.05)	0.43 (0.05)
	W	0.44 (0.04)	0.23 (0.05)	0.2 (0.04)	0.22 (0.05)	0.22 (0.04)	0.22 (0.05)	0.21 (0.04)
Predictive ability								
DBH	T	0.11 (0.04)	0.11 (0.04)	0.12 (0.04)	0.11 (0.04)	0.12 (0.04)	0.11 (0.04)	0.11 (0.04)
	C	0.15 (0.03)	0.29 (0.05)	0.28 (0.05)	0.29 (0.05)	0.29 (0.05)	0.28 (0.05)	0.29 (0.05)
	G	0.21 (0.07)	0.18 (0.05)	0.19 (0.05)	0.17 (0.05)	0.19 (0.05)	0.18 (0.05)	0.18 (0.05)
	K	0.04 (0.07)	0.05 (0.05)	0.04 (0.05)	0.05 (0.05)	0.03 (0.05)	0.02 (0.05)	0.04 (0.05)
	N	0.08 (0.06)	0.32 (0.06)	0.33 (0.06)	0.32 (0.06)	0.32 (0.06)	0.32 (0.06)	0.32 (0.06)
	W	0.06 (0.06)	0.19 (0.06)	0.15 (0.05)	0.17 (0.06)	0.15 (0.05)	0.15 (0.05)	0.16 (0.05)
Height	T	0.17 (0.05)	0.44 (0.04)	0.44 (0.04)	0.44 (0.04)	0.44 (0.04)	0.44 (0.04)	0.44 (0.04)
	C	0.22 (0.04)	0.26 (0.03)	0.26 (0.03)	0.27 (0.03)	0.27 (0.03)	0.27 (0.02)	0.26 (0.03)
	G	0.25 (0.06)	0.41 (0.05)	0.41 (0.05)	0.42 (0.05)	0.42 (0.04)	0.43 (0.04)	0.42 (0.05)
	K	-0.18 (0.07)	0.19 (0.07)	0.19 (0.07)	0.18 (0.07)	0.19 (0.07)	0.19 (0.07)	0.19 (0.07)
	N	0.07 (0.03)	0.32 (0.05)	0.32 (0.05)	0.33 (0.05)	0.32 (0.05)	0.33 (0.05)	0.33 (0.05)
	W	0.07 (0.09)	0.06 (0.06)	0.05 (0.07)	0.06 (0.06)	0.06 (0.07)	0.04 (0.07)	0.06 (0.06)

Table S6. (Continued)

Trait	Site	Predictive model *						
		ABLUP	GBLUP	BRR	BL	Bayes A	Bayes B	Bayes C
Straight-ness	T	0.05 (0.04)	0.1 (0.04)	0.09 (0.03)	0.1 (0.04)	0.1 (0.04)	0.09 (0.03)	0.1 (0.04)
	C	0.06 (0.03)	0.17 (0.03)	0.17 (0.03)	0.16 (0.03)	0.17 (0.03)	0.16 (0.03)	0.16 (0.03)
	G	0.2 (0.03)	0.21 (0.04)	0.2 (0.04)	0.21 (0.04)	0.21 (0.04)	0.2 (0.04)	0.2 (0.04)
	K	0.22 (0.05)	0.21 (0.05)	0.21 (0.05)	0.22 (0.05)	0.21 (0.05)	0.21 (0.05)	0.21 (0.05)
	N	0.18 (0.05)	0.18 (0.05)	0.19 (0.05)	0.16 (0.05)	0.2 (0.05)	0.2 (0.05)	0.2 (0.05)
	W	0.04 (0.07)	0.17 (0.05)	0.17 (0.05)	0.17 (0.06)	0.16 (0.05)	0.17 (0.04)	0.17 (0.05)
Volume	T	0.14 (0.04)	0.18 (0.05)	0.18 (0.05)	0.18 (0.05)	0.18 (0.05)	0.17 (0.05)	0.18 (0.05)
	C	0.17 (0.03)	0.3 (0.04)	0.29 (0.04)	0.3 (0.05)	0.29 (0.05)	0.28 (0.04)	0.29 (0.04)
	G	0.22 (0.06)	0.22 (0.05)	0.23 (0.04)	0.21 (0.05)	0.23 (0.05)	0.22 (0.05)	0.22 (0.04)
	K	-0.02 (0.07)	0.02 (0.06)	0.02 (0.07)	0.03 (0.06)	0.01 (0.07)	0.01 (0.06)	0.02 (0.06)
	N	0.06 (0.05)	0.31 (0.05)	0.31 (0.05)	0.31 (0.04)	0.31 (0.05)	0.3 (0.05)	0.3 (0.05)
	W	0.09 (0.06)	0.18 (0.06)	0.12 (0.05)	0.16 (0.05)	0.15 (0.05)	0.16 (0.05)	0.14 (0.05)

Mean (standard error) of accuracy and predictive ability from 10-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

[b] ABLUP, additive best linear unbiased prediction; GBLUP, genomic BLUP; BRR, Bayesian ridge regression; BL, Bayesian LASSO

Table S7. Accuracy and predictive ability of ABLUP, HBLUP, and GBLUP

Trait	Site [a]	Accuracy			Predictive ability		
		ABLUP	HBLUP	GBLUP	ABLUP	HBLUP	GBLUP
DBH	T	0.7 (0.02)	0.65 (0.02)	0.26 (0.04)	0.12 (0.04)	0.13 (0.03)	0.12 (0.04)
	C	0.63 (0.03)	0.65 (0.03)	0.5 (0.03)	0.12 (0.04)	0.15 (0.04)	0.25 (0.03)
	G	0.51 (0.03)	0.52 (0.03)	0.34 (0.04)	0.22 (0.05)	0.24 (0.04)	0.19 (0.04)
	K	0.56 (0.03)	0.52 (0.03)	0.2 (0.05)	-0.03 (0.06)	0.01 (0.04)	-0.03 (0.04)
	N	0.58 (0.04)	0.57 (0.04)	0.31 (0.04)	0.2 (0.08)	0.2 (0.08)	0.23 (0.05)
	W	0.5 (0.04)	0.39 (0.04)	0.24 (0.08)	0.07 (0.06)	0.02 (0.05)	0.19 (0.06)
Height	T	0.57 (0.03)	0.58 (0.03)	0.49 (0.03)	0.14 (0.04)	0.15 (0.04)	0.42 (0.03)
	C	0.55 (0.03)	0.55 (0.03)	0.39 (0.04)	0.19 (0.04)	0.21 (0.04)	0.22 (0.04)
	G	0.53 (0.03)	0.53 (0.03)	0.48 (0.05)	0.23 (0.05)	0.23 (0.05)	0.39 (0.06)
	K	0.28 (0.05)	0.23 (0.07)	0.23 (0.05)	-0.23 (0.05)	-0.24 (0.06)	0.19 (0.06)
	N	0.54 (0.04)	0.56 (0.03)	0.37 (0.06)	0.13 (0.06)	0.17 (0.05)	0.3 (0.04)
	W	0.38 (0.06)	0.36 (0.07)	0.14 (0.05)	0.05 (0.07)	0.07 (0.07)	0.06 (0.05)
Straight-ness	T	0.58 (0.04)	0.57 (0.03)	0.29 (0.04)	0.06 (0.03)	0.1 (0.04)	0.12 (0.05)
	C	0.73 (0.02)	0.69 (0.02)	0.43 (0.03)	0.01 (0.04)	0.03 (0.05)	0.19 (0.03)
	G	0.58 (0.02)	0.55 (0.03)	0.31 (0.05)	0.16 (0.03)	0.15 (0.03)	0.17 (0.06)
	K	0.59 (0.02)	0.59 (0.02)	0.43 (0.03)	0.2 (0.03)	0.22 (0.02)	0.27 (0.03)
	N	0.52 (0.06)	0.55 (0.05)	0.29 (0.08)	0.28 (0.09)	0.33 (0.07)	0.2 (0.08)
	W	0.49 (0.04)	0.45 (0.02)	0.24 (0.08)	0.01 (0.05)	0.02 (0.04)	0.16 (0.08)
Volume	T	0.69 (0.02)	0.65 (0.02)	0.29 (0.03)	0.12 (0.04)	0.12 (0.03)	0.16 (0.05)
	C	0.65 (0.04)	0.66 (0.04)	0.5 (0.03)	0.15 (0.04)	0.18 (0.04)	0.27 (0.03)
	G	0.51 (0.03)	0.53 (0.02)	0.35 (0.05)	0.2 (0.04)	0.23 (0.03)	0.21 (0.05)
	K	0.55 (0.04)	0.51 (0.04)	0.2 (0.06)	-0.07 (0.05)	-0.02 (0.04)	-0.03 (0.04)
	N	0.56 (0.03)	0.54 (0.03)	0.29 (0.03)	0.14 (0.05)	0.15 (0.06)	0.19 (0.04)
	W	0.44 (0.05)	0.37 (0.03)	0.24 (0.07)	0.09 (0.07)	0.06 (0.06)	0.19 (0.06)

Mean (standard error) of accuracy and predictive ability of 10 replication

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S8. GBLUP accuracy and predictive ability by cross-validation fold number.

Trait	Site [a]	Cross-validation folds			
		CV3	CV5	CV10	CV20
Accuracy					
DBH	T	0.3 (0.02)	0.2 (0.04)	0.29 (0.04)	0.3 (0.05)
	C	0.47 (0.04)	0.49 (0.03)	0.5 (0.04)	0.5 (0.03)
	G	0.33 (0.05)	0.35 (0.07)	0.33 (0.04)	0.37 (0.04)
	K	0.18 (0.05)	0.25 (0.07)	0.25 (0.06)	0.27 (0.05)
	N	0.36 (0.06)	0.39 (0.06)	0.44 (0.05)	0.44 (0.06)
	W	0.2 (0.09)	0.19 (0.04)	0.25 (0.05)	0.23 (0.07)
Height	T	0.49 (0.04)	0.51 (0.03)	0.5 (0.04)	0.49 (0.03)
	C	0.41 (0.05)	0.41 (0.04)	0.42 (0.01)	0.43 (0.02)
	G	0.48 (0.01)	0.47 (0.02)	0.48 (0.04)	0.5 (0.03)
	K	0.21 (0.05)	0.19 (0.08)	0.25 (0.06)	0.22 (0.05)
	N	0.36 (0.03)	0.41 (0.03)	0.44 (0.04)	0.41 (0.04)
	W	0.14 (0.07)	0.18 (0.06)	0.16 (0.04)	0.21 (0.06)
Straight-ness	T	0.3 (0.01)	0.29 (0.04)	0.32 (0.04)	0.3 (0.05)
	C	0.39 (0.05)	0.42 (0.03)	0.43 (0.03)	0.42 (0.03)
	G	0.33 (0.04)	0.32 (0.04)	0.34 (0.03)	0.34 (0.03)
	K	0.37 (0.02)	0.34 (0.04)	0.36 (0.05)	0.37 (0.05)
	N	0.22 (0.05)	0.3 (0.03)	0.29 (0.05)	0.31 (0.06)
	W	0.2 (0.07)	0.23 (0.07)	0.26 (0.06)	0.26 (0.06)
Volume	T	0.33 (0.01)	0.27 (0.05)	0.34 (0.05)	0.33 (0.05)
	C	0.47 (0.04)	0.47 (0.02)	0.49 (0.03)	0.49 (0.03)
	G	0.36 (0.04)	0.35 (0.06)	0.36 (0.03)	0.39 (0.04)
	K	0.16 (0.03)	0.24 (0.08)	0.24 (0.07)	0.25 (0.06)
	N	0.35 (0.07)	0.4 (0.04)	0.44 (0.05)	0.46 (0.05)
	W	0.17 (0.1)	0.18 (0.04)	0.23 (0.05)	0.23 (0.07)
Predictive ability					
DBH	T	0.13 (0.03)	0.03 (0.02)	0.11 (0.04)	0.13 (0.04)
	C	0.27 (0.03)	0.27 (0.03)	0.29 (0.05)	0.29 (0.04)
	G	0.18 (0.06)	0.21 (0.07)	0.18 (0.05)	0.21 (0.05)
	K	-0.04 (0.03)	0.04 (0.05)	0.05 (0.05)	0.07 (0.06)
	N	0.25 (0.06)	0.28 (0.04)	0.32 (0.06)	0.32 (0.05)
	W	0.13 (0.12)	0.14 (0.05)	0.19 (0.06)	0.18 (0.06)

Table S8. (Continued)

Trait	Site [a]	Cross-validation folds			
		CV3	CV5	CV10	CV20
Height	T	0.42 (0.02)	0.44 (0.05)	0.44 (0.04)	0.43 (0.03)
	C	0.27 (0.06)	0.26 (0.04)	0.26 (0.03)	0.27 (0.03)
	G	0.39 (0.02)	0.38 (0.02)	0.41 (0.05)	0.43 (0.04)
	K	0.18 (0.03)	0.13 (0.08)	0.19 (0.07)	0.18 (0.06)
	N	0.27 (0.06)	0.29 (0.04)	0.32 (0.05)	0.32 (0.04)
	W	0.06 (0.06)	0.1 (0.06)	0.06 (0.06)	0.11 (0.07)
Straight-ness	T	0.11 (0.02)	0.11 (0.05)	0.1 (0.04)	0.09 (0.04)
	C	0.14 (0.05)	0.16 (0.02)	0.17 (0.03)	0.17 (0.03)
	G	0.19 (0.03)	0.19 (0.03)	0.21 (0.04)	0.2 (0.04)
	K	0.23 (0.04)	0.19 (0.03)	0.21 (0.05)	0.22 (0.05)
	N	0.1 (0.03)	0.18 (0.03)	0.18 (0.05)	0.21 (0.07)
	W	0.15 (0.01)	0.15 (0.07)	0.17 (0.05)	0.17 (0.06)
Volume	T	0.18 (0.04)	0.11 (0.03)	0.18 (0.05)	0.18 (0.05)
	C	0.28 (0.04)	0.27 (0.03)	0.3 (0.04)	0.3 (0.04)
	G	0.23 (0.07)	0.22 (0.06)	0.22 (0.05)	0.24 (0.05)
	K	-0.07 (0)	0.01 (0.05)	0.02 (0.06)	0.02 (0.06)
	N	0.23 (0.07)	0.29 (0.03)	0.31 (0.05)	0.35 (0.05)
	W	0.1 (0.11)	0.13 (0.05)	0.18 (0.06)	0.17 (0.06)

Mean (standard error) of accuracy and predictive ability from 3, 5, 10, 20-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S9. GBLUP accuracy and predictive ability according to the environment of training and test population

Trait	Site [a]	Accuracy		Predictive ability	
		within	between	within	between
DBH	T	0.29 (0.04)	0.42 (0.02)	0.11 (0.04)	0.05 (0.03)
	C	0.5 (0.04)	0.45 (0.03)	0.29 (0.05)	0.17 (0.03)
	G	0.33 (0.04)	0.39 (0.05)	0.18 (0.05)	0.09 (0.04)
	K	0.25 (0.06)	0.49 (0.05)	0.05 (0.05)	0.1 (0.07)
	N	0.44 (0.05)	0.41 (0.05)	0.32 (0.06)	0.1 (0.07)
	W	0.25 (0.05)	0.34 (0.06)	0.19 (0.06)	0.08 (0.03)
	Combined	0.38 (0.02)		0.12 (0.02)	
Height	T	0.5 (0.04)	0.46 (0.04)	0.44 (0.04)	0.16 (0.04)
	C	0.42 (0.01)	0.46 (0.04)	0.26 (0.03)	0.19 (0.04)
	G	0.48 (0.04)	0.43 (0.04)	0.41 (0.05)	0.18 (0.04)
	K	0.25 (0.06)	0.44 (0.04)	0.19 (0.07)	0.17 (0.05)
	N	0.44 (0.04)	0.37 (0.05)	0.32 (0.05)	0.11 (0.06)
	W	0.16 (0.04)	0.45 (0.05)	0.06 (0.06)	0.24 (0.06)
	Combined	0.48 (0.02)		0.09 (0.02)	
Straight-ness	T	0.32 (0.04)	0.4 (0.03)	0.1 (0.04)	0.09 (0.06)
	C	0.43 (0.03)	0.38 (0.02)	0.17 (0.03)	0.1 (0.03)
	G	0.34 (0.03)	0.52 (0.04)	0.21 (0.04)	0.23 (0.06)
	K	0.36 (0.05)	0.39 (0.04)	0.21 (0.05)	0.1 (0.06)
	N	0.29 (0.05)	0.36 (0.07)	0.18 (0.05)	0.09 (0.07)
	W	0.26 (0.06)	0.41 (0.05)	0.17 (0.05)	0.13 (0.03)
	Combined	0.45 (0.02)		0.18 (0.03)	
Volume	T	0.34 (0.05)	0.46 (0.02)	0.18 (0.05)	0.1 (0.02)
	C	0.49 (0.03)	0.47 (0.03)	0.3 (0.04)	0.19 (0.03)
	G	0.36 (0.03)	0.42 (0.05)	0.22 (0.05)	0.13 (0.05)
	K	0.24 (0.07)	0.48 (0.05)	0.02 (0.06)	0.12 (0.06)
	N	0.44 (0.05)	0.41 (0.07)	0.31 (0.05)	0.11 (0.05)
	W	0.23 (0.05)	0.34 (0.06)	0.18 (0.06)	0.13 (0.05)
	Combined	0.42 (0.02)		0.07 (0.01)	

Mean (standard error) of accuracy and predictive ability from 10-fold cross-validation

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Table S10. Predictive accuracy of GBLUP using every single region as training and test population

Training Pop. [a]	Test Pop.	Accuracy				Predictive ability			
		DBH	Height	Straight-ness	Volume	DBH	Height	Straight-ness	Volume
T	C	0.19	0.19	0.282	0.177	0.113	0.076	0.067	0.127
	G	0.066	0.179	0.328	0.037	-0.05	0.103	0.085	-0.01
	K	0.178	0.195	0.297	0.147	-0.03	0.165	0.04	0.01
	N	0.116	0.116	0.309	0.071	0.057	0.007	0.105	0.054
	W	0.181	0.277	0.173	0.15	0.059	0.191	0.005	0.088
C	T	0.355	0.36	0.245	0.357	0.094	0.055	0.079	0.11
	G	0.382	0.373	0.33	0.393	0.1	0.107	0.14	0.113
	K	0.428	0.399	0.208	0.42	0.11	0.118	0.035	0.11
	N	0.397	0.474	0.195	0.417	0.056	0.283	0.066	0.125
	W	0.297	0.4	0.255	0.294	0.025	0.135	0.037	0.077
G	T	0.211	0.211	0.356	0.244	-0.04	0.018	0.07	-0.02
	C	0.303	0.251	0.329	0.337	0.108	0.044	0.106	0.116
	K	0.271	0.267	0.422	0.306	0.14	0.057	0.199	0.138
	N	0.177	0.229	0.356	0.242	0.025	0.094	0.084	0.074
	W	0.268	0.298	0.435	0.274	0.1	0.197	0.215	0.123
K	T	0.172	0.242	0.207	0.215	-0.01	0.156	0.043	-0.01
	C	0.377	0.307	0.179	0.416	0.142	0.143	0.029	0.147
	G	0.285	0.233	0.23	0.339	0.128	0.127	0.151	0.163
	N	0.386	0.105	0.229	0.431	0.198	-0.05	0.092	0.185
	W	0.127	0.181	0.108	0.209	-0.04	0.13	-0.08	0.064
N	T	0.089	0.258	0.252	0.14	0.027	0.024	0.06	0.026
	C	0.156	0.361	0.264	0.228	0.048	0.177	0.024	0.103
	G	0.115	0.362	0.284	0.206	0.01	0.135	0.061	0.062
	K	0.158	0.148	0.215	0.18	0.153	-0.06	0.085	0.12
	W	0.041	0.091	0.15	0.03	0.035	-0.04	0.105	-0.02
W	T	0.223	0.253	0.134	0.275	0.017	0.162	0.022	0.057
	C	0.308	0.253	0.078	0.345	0.057	0.109	0.007	0.086
	G	0.168	0.218	0.164	0.217	0.06	0.157	0.147	0.09
	K	0.203	0.276	0.059	0.245	0.016	0.125	-0.06	0.06
	N	0.164	0.07	0.186	0.161	0.047	-0.05	0.126	-0

Shaded cells mean prediction within genetic zone.

[a] T, Taean; C, Chuncheon; G, Gongju; K, Kyeongju; N, Naju; W, Wanju

Abstract in Korean

소나무는 동아시아 지역에 분포하는 우리나라의 자생종으로, 목재로서 가치가 높아 국내 조림 수요가 많은 산림 수종이다. 목재 생산성과 직결되는 소나무의 생장형질을 개량하기 위해서 전통적으로 차대검정을 통한 선발육종을 수행하여 세대를 진전시켜 왔다. 그러나 이러한 전통적인 방법은 시간이 오래 소요되어 연간 개량효과가 제한적이라는 한계가 있다. 유전체 선발은 분자 마커를 이용하여 얻은 유전체 정보를 통해 개체의 유전형 기반 육종가를 추정하는 방법으로 육종 과정에서 차대검정을 대신할 수 있다. 따라서 본 연구는 소나무의 육종 세대를 단축하는 것을 목적으로 소나무 육종 집단에 유전체 선발 기법을 도입하여 선발 효율 및 적용성을 평가하였다. 먼저 유전체 선발 연구 집단의 특성을 파악하고자 표현형을 통계적으로 분석하고 유전력과 유전상관 등의 유전모수를 추정하였다. 연구 집단인 소나무 풍매 및 인공교배 차대검정림의 표현형은 시험지 및 가계에 따라 평균에 차이가 있었다. 또한 시험지 별로 우수한 가계의 순위가 달랐으며 시험지 간 유전상관은 대부분 낮은 것으로 나타났다. 이에 따라 연구 집단은 유전형과 환경의 상호작용이 큰 것으로 판단되었으며, 유전체 선발 시 여러 지역을 대상으로 하는 경우에는 표현형의 보정이 필요한 것으로 여겨졌다. 다음으로 소나무 유전체 선발 모형을 훈련하는 단계로서 풍매 차대검정림에서 마커 선발 방법, 예측 모형의 종류, 훈련 데이터 구성 등 여러 조건에서의 교차검정을 수행하고 모형 최적화를 위하여 예측 정확도를 비교하였다. 또한 유전체 선발의 효율을 평가하기 위하여 유전적 개량효과를 전통적인 선발 방법과 비교하였다. 그 결과 예측 모형 또는 교차검정 배수는 유전체 선발의 예측 정확도에 영향을 주지 않았으며, 유전력과 마커 선발 방법, 훈련집단과 검정집단의 환경과 가계 구성은 예측 정확도에 영향을 미쳤다. 이를 통해 소수

대립유전자 빈도가 0.05 이상인 마커를 사용하여 유전체 기반의 최적선형불편예측법(GBLUP)으로 여러 환경의 개체를 포함하도록 유전체 선발 모형을 훈련하는 것이 효율적인 것으로 판단되었다. 또한 유전체 선발을 통하여 기존에 실시되던 표현형 선발과 가계 선발에 비해 높은 연간 개량효과를 얻을 수 있는 것으로 분석되었다. 마지막으로, 소나무 유전체 선발의 활용성을 검증하기 위하여 훈련된 유전체 선발 모형으로 훈련집단과 유연관계가 없는 인공교배 차대검정림에서의 예측 정확도를 평가하였다. 대상 집단의 유연관계에 따른 예측 정확도 비교 결과, 반형매 집단보다 유연관계가 서로 높은 전형매 집단에서 유전체 선발의 예측 정확도가 높게 나타났다. 그리고 훈련된 유전체 선발 모형으로 인공교배 차대검정림의 유전체 육종가 예측이 가능했다. 따라서 본 연구의 소나무 유전체 선발 모형을 가계 구성 및 환경이 다른 육종 집단에 적용하는 것이 가능하다고 판단되었다. 본 연구를 통해서 소나무에서 유전체 선발은 기존 선발 방법을 대체할 수 있는 선발 효율을 가진 것으로 판단되었으며, 향후 유전체 선발을 통한 선발목의 진전 세대 검정을 통하여 소나무 가속육종의 기반을 다질 수 있을 것으로 기대된다.

주요어: 소나무, 차대검정, 유전체 선발, 가속육종, 육종가, 개량효과

학 번: 2018-35767