DEVELOPING HIGH BIOFUEL COASTAL DOUGLAS-FIR FEEDSTOCKS BY GENETIC SELECTION

Author	ORGANIZATION
Keith Jayawickrama	Oregon State University
Glenn Howe	Oregon State University
Terrance Ye	Oregon State University
Matt Trappe	Oregon State University
Jennifer Kling	Oregon State University
Scott Kolpak	Oregon State University
Xiao Zhang	Washington State University
Scott Gelyense	Washington State University
Stephanie Guida	NCGR National Center for Genome Resources
Callum Bell	NCGR National Center for Genome Resources

COMPLETED 2017

TABLE OF CONTENTS

NARA is led by Washington State University and supported by the Agriculture and Food Research Initiative Competitive Grant no. 2011-68005-30416 from the USDA National Institute of Food and Agriculture.

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the U.S. Department of Agriculture.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

CR	call rate
DBH	diameter at breast height (1.4 meters)
FORK	incidences of stem forking
GEBV	genomic breeding values
HT	total tree height
HY	hydrolysis yield
GS	genomic selection
MVGS	multivariate genomic selection
NR	needle retention
OSU	Oregon State University
PA	prediction accuracy
PNW	Pacific Northwest
PY	pretreatment yield
RAMI	incidences of ramicorn branching
RE	relative efficiency
RF	recalcitrance factor
SG	wood specific gravity
SINU	stem sinuosity
SNP	single nucleotide polymorphism
TB	total biofuel product
VOL	stem volume index
UVGS	univariate genomic selection

EXECUTIVE SUMMARY

Narrow-sense heritabilities, genetic correlations between traits, and predicted genetic gains for pretreatment yield, pretreated holocellulose, enzymatic hydrolysis yield, and recalcitrance factor were predicted for 284 progeny trees, 30 woods-run (unimproved) trees, 28 crosses (between 6 and 12 progeny per cross) and 46 parents. Heritabilities ranged from 0.18 to 0.77, very comparable to many publications for other wood properties (jet-fuel related heritabilities have not be reported before). While specific gravity was favorably correlated with recalcitrance factor, the genetic correlation was not high enough to be a very reliable predictor (indirect selection trait). One of the forward selections had a 40.6% predicted gain in holocellulose yield and 34.7% predicted gain in recalcitrance factor. If instead we were to select existing seed producing parents, one parent had a 27.0% predicted gain in holocellulose yield and 21.5% predicted gain in recalcitrance factor.

The contract for building the array (50K SNPs) using the Affymetrix platform, and for genotyping 1,920 samples, was awarded to GeneSeek Inc., part of NeoGen Corp. We used the Affymetrix Axiom genotyping array to test 55,766 potential SNPs in Douglas-fir. Because the SNPs were derived from transcriptome sequence, the array targets SNPs in the expressed genes in the Douglas-fir genome. We collected 1920 needle samples from selected Douglas-fir trees at three progeny sites, four seed orchards, and one container nursery. DNA from all the samples were successfully extracted at the US Forest Service's NFGEL facility, and sent to GeneSeek by the end of February, 2015. We tested the array on these trees, and found that 22,126 SNPs could be genotyped with a call rate of 80%.

We planted out the first, or one of the first, genomic selection studies for coastal Douglas-fir on Roseburg Resources property near Elkton, Oregon, on March 6, 2015. The 1,420 one-year old seedlings were obtained from 25 full-sib crosses and one unimproved control. Individual trees are identified so that the seedlings from the full-sib crosses had DNA samples extracted from them as described above.

This study showed encouraging results of applying genomic selection in coastal Douglas-fir breeding programs. The predictive abilities of SNP markers were around 0.60 for growth and biofuel product, and 0.75 for branching / stem straightness in univariate models. For growth traits, accuracies remained high when using models generating at age 7 to predict phenotypes at age 12. Prediction using multivariate models were generally more accurate, but the increase of accuracy depends on the relationship among traits.

This project resulted in four significant conclusions. <u>First</u>, it is possible to measure biofuel production traits on Douglas-fir trees as part of an operational tree breeding program. We developed or implemented field, laboratory, and statistical methods for assessing the genetics of biofuel production traits. <u>Second</u>, we demonstrated that there is sufficient genetic variation and heritability to improve biofuel production traits in Douglas-fir. Thus, breeding programs aimed at improving biofuel production will be successful. <u>Third</u>, it is possible to integrate genomic evaluation into operational Douglas-fir breeding programs. We developed a high-density single nucleotide polymorphism (SNP) array for Douglas-fir, which will allow tree breeders to apply genomic techniques to the genetic improvement of Douglas-fir. <u>Fourth</u>, using the SNP genotyping array and an approach called genomic prediction, we demonstrated that we can use genomic evaluation to identify superior genotypes for biofuel production and growth traits. This is a significant achievement because SNP-based approaches can be used to speed the delivery of genetic gain from breeding programs. This will be particularly important for incorporating new traits, such as biofuel production, into breeding programs. Overall, we developed a complete roadmap for using traditional and molecular breeding approaches to improve Douglas-fir for biofuel production.

INTRODUCTION

Genetic selection and testing has been applied on timber species in the Pacific Northwest for over 50 years. One result of that work is data and genetic gain predictions for several traits from replicated, randomized progeny tests for over 30,000 families of Douglas-fir. A range of phenotypic variation, and some level of genetic control, has been demonstrated among families for every trait studied, so we expected variation and genetic control in traits pertaining to biofuel production. Another result is that over 150,000 timberland acres are reforested annually with seedlings from open-pollinated seed orchards, thus delivering real genetic gains (in whatever traits are selected for) to operational plantations in the PNW. Methods for measuring genetic control in commercially important traits (including growth rate and wood properties) are well developed, so we were able to apply them in this study.

Over the last decade, the cost of using genomic and marker-based tools to complement field-based breeding and testing has dropped rapidly in forest tree species. These tools have the potential to improve the efficiency, speed the delivery of genetic gain, especially given the long times needed for field-based breeding, and reduce costs. Genomic selection (GS) may transform tree breeding by allowing breeders to shorten the breeding cycle, reduce the costs of progeny testing, increase heritabilities, and select for mature traits such as wood properties at the seedling stage. After the first year of the project, NARA leadership decided on building a SNP chip for Douglas-fir as the main outcome of the genetics part of the project. We were able to collaborate with the Pacific Northwest Tree Improvement Research Cooperative at OSU, since that cooperative has identified GS as a main research focus for the near future.

Having enough markers is critical to the success of genomic selection. The number of markers needed for genomic selection varies based on the biology of the species, including the genome length (cM), number of QTL controlling the trait, and trait heritability. It also depends on the effective size of the breeding population (*Ne*), which is about equal to the number of parents used in the breeding cycle. In Douglas-fir, 2,000 to 40,000 SNPs may be needed for effective population sizes of 25 to 100 (Grattapaglia and Resende, 2011; Iwata et al., 2011; Resende et al., 2012a; Resende et al., 2012b). Because sublines in typical advanced generation breeding programs of Douglas-fir have effective population sizes closer to 30, the number of SNPs needed within such sublines is expected to be about 5,000 to 20,000, depending on the factors described above and the distribution of the markers in the genome (reviewed in Howe et al. 2013).

Recent advances by the Conifer Translational Genomics Network (a multi-institution project for major US conifers) were available to use in this project. Howe et al. (2013) used transcriptome sequencing to identify 278,979 potential SNPs in ~20,000 Douglas-fir genes, and then tested a subset of these SNPs (n=8067) using an Illumina Infinium genotyping array, resulting in 5847 successful SNPs (i.e., polymorphic SNPs that can be reliably measured). Although the Infinium array is highly robust, it is also expensive. Although costs have been decreasing, the cost at the time of purchase was about $120 per tree. Other, less-expensive genotyping arrays have become available more recently, most notably, the Axiom array manufactured by Affymetrix. Although the costs of this array vary widely based on the number of SNPs assayed and the number of trees genotyped, the entry point was about $75 per tree when we began this this project.

We undertook an expanded/strengthened Task 2 (Identify single nucleotide polymorphisms [SNP] genotypes) to use the power of both of these approaches in tandem, with a state-of-the-science genotyping array based on SNP technology for marker-based selection of phenotypes conducive to production of biofuels from woody residuals as a value added trait of trees selected for production of lumber and other products of saw logs.

TASK 1: COLLECT WOOD SAMPLES, OBTAIN WOOD CHEMISTRY DATA, COMBINE WITH EXISTING DATA ON GROWTH RATE

Task Objective

Some differences of genomic selection from progeny-testing-based selection, and potential advantages, are outlined by Luan et al. (2009). We therefore attempted to quantify the phenotypic variation in biofuel production potential in a subset of Douglas-fir families, pre-selected for commercially important traits such as rapid growth, adaptability, wood specific gravity and wood stiffness

Methodology

Our first step was prioritization of superior softwood (Douglas-fir) breeding stock. The progeny test populations most suitable for sampling should (1) have advanced-generation high-genetic gain germplasm, (2) have trees large enough to obtain amounts of wood needed for chemical analysis, (3) have good maps and accession information and (4) be available and accessible to OSU researchers and contractors. Two second-generation populations in Oregon, T96 (near Toledo) and CL98 (near Coos Bay), established by Plum Creek Timber Company in 1997 and 1999, respectively, were selected.

As a pilot study, wood cores were obtained from trees from a single half-sib family and shipped for analysis to WSU/Tri-cities (NARA researcher Xaio Zhang) for setting baseline carbohydrate, lignin, ash, and total extractives. Core samples (sampled at breast height) consisted of 18-20 grams fresh weight/sample for initial chemical analyses and evaluation. Various sampling tools (cordless drills, gas-powered drill, 5mm and 10mm manual corers) were evaluated and compared. The cordless and gas-powered drills were found adequate to obtain 5mm cores, but inadequate for taking 10mm cores. A modification was built to improve ease and efficiency of taking 10 mm cores with the manual corer. Fifty-five (55) different families were selected across the range of gains for growth rate from the T96 population, and then samples were obtained from a total of 700 trees from three sites. The cores were measured, weighed and shipped to the Zhang lab at WSU. Enough samples were obtained to provide 10g of dry wood for analysis. The final set of 150 cores were dried and ground in a Wiley mill at OSU to free-up time for the Zhang lab to expedite analysis.

We selected 30 more families and 3 woodsrun lots from the CL98 series as well, located and visited the Moon Creek progeny test site near Fairview, and collected a total of 360 samples. These samples were dried and ground at OSU and shipped to the Zhang lab at WSU for wood chemistry analysis.

Statistical Model and Analyses

Chemical analyses were described in Geleynse et al. (2016). For each trait, a univariate family model was used for estimating variance components and heritability. The following linear model was fitted using ASReml software:

$$y_{ijk} = \mu + F_i + M_j + (FM)_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the observation of the *k*th tree from the *i*th female and *j*th male parent, μ is the population mean, F_i is the random effect of *i*th female parent, M_j is the random effect of the *j*th male effect, $(FM)_{ij}$ is the random effect of the full-sib family (*i*th female x *j*th male), and ε_{ijk} is the random residual. Raw data were transformed by Y = Y × 100 prior to analyses to avoid losing precision. Narrow-sense individual-tree heritability (h_i^2)was estimated as the ratio of additive genetic variance (V_A) to the total phenotypic variance (V_P) among individual trees:

$$h_i^2 = \frac{V_A}{V_P} = \frac{2(\sigma_f^2 + \sigma_m^2)}{\sigma_f^2 + \sigma_m^2 + \sigma_\delta^2 + \sigma_e^2}$$

where σ_f^2 , σ_m^2, σ_δ^2, and σ_e^2 are the estimated variance components of female, male, female x male, and residual effects, respectively. Bivariate analyses were carried out to estimate genetic correlations between traits using a bivariate family model, expressed in matrix format:

$$\mathrm{y} = \mu + \mathrm{Z}_1\mathrm{f} + \mathrm{Z}_2\mathrm{m} + \mathrm{Z}_3\delta + \mathrm{e}$$

where $\mathrm{y} = [\mathrm{y}_1', \mathrm{y}_2']$, y_1 and y_2 are the vectors of individual tree observations for two traits; $\imath = [\mu_1', \mu_2']$, μ_1 and μ_2 are the vectors of fixed means of traits; $\mathrm{f} = [\mathrm{f}_1', \mathrm{f}_2']$, f_1 and f_2 are the vectors of random female effects; $\mathrm{m} = [\mathrm{m}_1', \mathrm{m}_2']$, m_1 and m_2 are the vectors of random male effects; $\delta = [\delta_1', \delta_2']$, δ_1 and δ_2 are the vectors of random female × male effects; $\mathrm{e} = [\mathrm{e}_1', \mathrm{e}_2']$, e_1 and e_2 are the vectors of random residuals; Z_1, Z_2, Z_3 are incidence matrices connecting the observations to female, male, and female x male effect, respectively. Variances and covariances were estimated using ASReml software, and genetic correlations (r_g) were calculated within ASReml according to the standard formulae [10]. The following individual-tree model was carried out for each trait to predict breeding values for individual trees and parents:

$$y_{ijk} = \mu + A_{ijk} + (FM)_{ij} + \varepsilon_{ijk}$$

where A_{ijk} is the random additive genetic value of the *k*th tree from *i*th female and *j*th male parents. This model incorporates the numerator relationship matrix in the analysis. The random effect solutions were obtained by solving the mixed model equations. Since the genetic covariance between relatives is provided by the supplied numerator relationship matrix, the predicted breeding values (*PBVs*) and the associated standard errors of prediction (*SEPs*) were computed for both parents and progeny simultaneously. *PBVs* for full-sib families were represented by their mid-parental *PBVs*. Genetic gains were predicted as the percentages of *PBV*s over the least-square mean of the test populations (woodsruns excluded). Narrow-sense heritabilities, genetic correlations between traits, and predicted genetic gains for pretreatment yield, pretreated holocellulose, enzymatic hydrolysis yield, and recalcitrance factor were predicted for 284 progeny trees, 28 crosses (between 6 and 12 progeny per cross) and 46 parents. 30 woodsrun (unimproved) trees.

Results

Heritabilities ranged from 0.18 to 0.77 (Table GS-1.1), very comparable to many published values for other wood properties (jet-fuel related heritabilities have not be reported before). While specific gravity was favorably correlated with recalcitrance factor, the genetic correlation was not high enough to be a very reliable predictor (indirect selection trait).

Table GS-1.1. Narrow-sense individual heritabilities and their standard errors for five wood traits in a Douglas-fir breeding population

	h^2_i	s.e.
Density (SG)	0.315	0.219
Pretreatment Yield (PY)	0.767	0.180
Pretreated Holocellulose (PH)	0.185	0.190
Hydrolysis Yield (HY)	0.496	0.142
Recalcitrance Factor (RF)	0.443	0.136

One of the forward selections had a 40.6% predicted gain in holocellulose yield and 34.7% predicted gain in recalcitrance factor. If instead we were to select existing seed producing parents, one parent had a 27.0% predicted gain in holocellulose yield and 21.5% predicted gain in recalcitrance factor (Table GS-1.2).

Table GS-1.2. Genetic correlation coefficients (lower triangle) & their standard errors (upper triangle).

	SG	PY	PH	HY	RF
SG		0.251	0.241	0.219	0.212
PY	0.048		0.272	0.189	0.253
PH	0.343	-0.021		0.259	0.273
HY	0.325	-0.497	-0.246		0.015
RF	0.402	-0.111	-0.159	0.972	

Conclusions/Discussion

The original plan was to assess a large number of families from multiple breeding populations, but given the costs of chemical analysis, this was not feasible. The estimates of heritability and predicted genetic gains show that it would be quite feasible to genetically select Douglas-fir for conversion to jet fuel. Given the sample sizes, these estimates should not be taken as the last word in genetic parameter estimates: we would typically want to sample from 100 families and 30 trees per family on at least three sites to increase our confidence in the estimates. However these results show a lot of promise.

From the CL98 test population, it would be possible to collect seed from a group of selected parents and start establishing high jet-fuel plantations in the near future. However for large-scale implementation into breeding programs in the Pacific Northwest, it would essential to either (1) identify indirect selection traits that are less expensive to measure or (2) find ways to simplify and accelerate the measurement of the wood chemistry traits so that we could (3) screen many more populations and trees.

TASK 2: IDENTIFY USEFUL SNP GENETIC MARKERS IN DOUGLAS-FIR THAT CAN BE USED TO ASSOCIATE WITH USEFUL PHENOTYPIC VARIATIONS IN BIOFUEL PRODUCTION POTENTIAL AND OTHER COMMERCIALLY IMPORTANT TRAITS

Task Objective

Establish Pilot GS study, build a new high-capacity Douglas-fir SNP chip and genotype trees selected from CL98 population and trees in the pilot GS study

Methodology

Design and Building of Genotyping Array

The contract for building the array (50K SNPs) using the Affymetrix platform, and for genotyping 1,920 samples, was awarded to GeneSeek Inc. (based in Lincoln, Nebraska), part of NeoGen Corp (http://www.neogen.com/ Genomics/). Due to the $ amount of the contract and OSU contracting rules, we needed to go through a long and time-consuming process including a Request for Proposals.

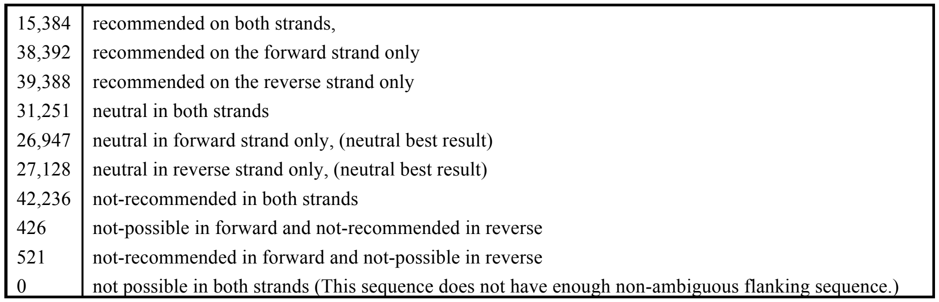
SNP resources

The potential SNPs chosen for the Axiom array were derived from transcriptome sequencing projects described by Muller et al. (2012 and Howe et al. (2013). We added the Muller SNPs to increase the number of genes that could be assayed, thereby increasing genome coverage for genomic selection. The Douglas-fir transcriptome (454 sequence data) and SNPs identified by Muller et al. (2012) were downloaded from http://www.treeversity.org by Stephanie Guida (National Center for Genome Resources). These data contained ~170,000 putative transcripts and ~188,000 SNPs. We used this information to identify 'new genes'—that is, genes that were absent from our transcriptome assembly—and then added the corresponding SNPs to our SNP database. To identify these new genes, NCGR compared the Muller transcripts to the Howe transcriptome assembly using BLAST and an e-value cutoff of 1e-10. Excluding singletons, 63,286 transcripts had no BLAST hits, and were classified as new genes. Muller et al. (2012) used three SNP detection programs (GSMapper, ssahaSNP, and bwa SAMtools) to identify 40,206 biallelic SNPs in the 63,286 unique transcripts described above. Of these 40,206 SNPs, 16,859 were detected by two or three SNP detection programs, and were the SNPs considered for inclusion on the genotyping array. These were added to our existing SNP database of 278,979 SNPs (Howe et al., 2013).

Axiom array design

Two steps were used to filter the combined SNP database described above. First, we removed SNPs that were highly repeated in the Douglas-fir genome. This was done by comparing the SNP sequences to a draft of the Douglas-fir genome (v0.5) provided by Jill Wegrzyn (University of Connecticut). Second, we removed SNPs that had flanking sequences that did not meet minimum Affymetrix criteria for inclusion on the array (Table GS-2.1).

Table GS-2.1. SNP quality for 221,674 SNPs first submitted to GeneSeek/Affymetrix.



15,384	recommended on both strands,
38,392	recommended on the forward strand only
39,388	recommended on the reverse strand only
31,251	neutral in both strands
26,947	neutral in forward strand only, (neutral best result)
27,128	neutral in reverse strand only, (neutral best result)
42,236	not-recommended in both strands
426	not-possible in forward and not-recommended in reverse
521	not-recommended in forward and not-possible in reverse
0	not possible in both strands (This sequence does not have enough non-ambiguous flanking sequence.)

After filtering, we submitted 111,648 SNPs in 21,659 genes to Affymetrix for the final array design: 108,299 SNPs in 19,336 genes came from the Howe SNP database, whereas 3,349 SNPs in 2,323 genes came from the Muller SNP database.

Because 111,648 SNPs exceeds the capacity of a 50K SNP array, we prioritized these SNPs for the final design phase. We ranked the SNPs sent to Affymetrix using various measures of SNP quality, giving high ranks to target SNPs that were successfully genotyped using the Infinium array, most likely to be true SNPs, and least likely to have other SNPs in their flanking sequences Howe et al. (2013). Affymetrix used our rankings and their proprietary 'p-convert' values to choose the final set of 55,766 SNPs representing 21,639 genes that were included on the array. The p-convert value reflects the probability that a SNP will be assayed reliable using the Axiom array system. The array also included a set of non-polymorphic 'control' probes that were used to judge array performance. Rich Cronn and Sanjuro Jogdeo developed these polymorphic sequences by identifying sequences that were identical between our Douglas-fir transcriptome and the loblolly pine genome. During processing, the control probes were used to calculate a quality control metric (DQC) that was used to identify and remove poor quality samples.

Establishment of Pilot Genomic Selection Study

We developed a 3-generation pilot genomic selection population with elite genetic material from a cooperative Douglas-fir breeding program, obtained consent from the breeding program to use the required seed from 3rd-cycle crosses, obtained greenhouse space to sow the study in 2014, and agreement by a large industrial landowner to outplant the study in 2015. The trial (one of the first, genomic selection studies for coastal Douglas-fir) was sown at the end of March 2014 (1,420 one-year old seedlings were sown from 25 full-sib crosses and one unimproved control) and 1,189 seedlings were planted out on Roseburg Resources property near Elkton, Oregon, on March 25, 2015. The test site was specially prepared and a grid put in for planting them. Individual trees are identified so that the seedlings from the full-sib crosses had DNA samples extracted from them as described above.

While phenotypic data from this study will be collected past the timeline of the NARA project, it will still be an important outcome for Douglas-fir improvement in the PNW.

DNA Extraction

We collected 1920 needle samples from selected Douglas-fir trees at three progeny sites, four seed orchards, and one container nursery (Table GS-2.2).

Table GS-2.2. Foliage samples were collected from the following sets of trees to be processed through the SNP genotyping array

No. of trees	Description
291	2nd-cycle CL98 progeny trees used in wood chemistry analysis or pilot genomic selection study
28	CL98 parents with wood chemistry data
46	other 1st generation parents or grandparents of 3rd cycle genomic selection crosses
26	2nd cycle parents of 3rd cycle genomic selection crosses
264	other 2nd cycle progeny, full-sibs of genomic selection study 2nd-cycle parents
1,141	3rd cycle progeny (genomic selection study selection population)
124	other parents of future 3rd cycle crosses

Each sample consisted of 5-10 green needles. Samples were placed in numbered 14-cm^3 vials and 10 cm^3 of crystalline silicate desiccant was added immediately to preserve DNA, and the vials were sealed. All samples were carefully tracked by spreadsheet.

Subsamples of three needles were taken from each vial, manually minced to 2-3mm lengths, and each sample was carefully loaded into a well in a 96-well DNA extraction plate (Qiagen DNeasy 96 Plant DNA kit). The location of each sample in each plate was carefully recorded. The loaded plates were transported to the USDA Forest Service National Forest Genetics Electrophoresis Laboratory (NFGEL) in Placerville, CA for extraction. The DNA extraction process followed the instructions in the Qiagen DNeasy kit. Extraction success was quantified using SYBR intercalating dye (Pico Green); any extraction producing less than 10ng DNA/µL was re-extracted. 1920 samples were successfully extracted at the NFGEL facility, dried down and shipped to Geneseek Corp., Lincoln, NE for SNP analysis.

Results

We measured 55,766 potential SNPs on 1,920 samples using the Axiom array. Of the 1,920 DNA samples submitted to GeneSeek, 1,866 passed DQC standards and 1,694 passed DQC, Plate QC and call rate QC rates (226 samples did not pass). Table GS-2.3 shows the number of SNPs falling into six SNP quality categories: PolyHighResolution, NoMinorHom, OTV, MonoHighRes, and CallRateBelowThreshold. The call rate (CR) is an important measure of SNP quality. CR is the proportion of trees that can be assigned a reliable genotype (called) relative to the total number of trees genotyped. The average call rate for the passing samples was 99.01%.

We worked with Affymetrix bioinformaticists to develop protocols to 'rescue' SNPs that previously did not pass the default Affymetrix quality control criteria (e.g., 97% call rate). For instance, lowering the call rate threshold from 97% to 60% using the new custom R scripts increased the number of successful SNPs from 16,177 to 24,192 in one population, and from 18,932 to 25,881 in another. We used a subset of 427 unrelated trees to calculate SNP population genetic statistics. Over a range of call rate thresholds (60% to 97%), the median call rate for SNPs in Hardy-Weinberg equilibrium ranged from 99.1% to 100.0%, and the median minor allele frequency ranged from 0.196 to 0.236. Based on a small number of samples, the successful SNPs also work well on Interior Douglas-fir. The Axiom genotyping array will serve as an excellent foundation for studying the population genomics of Douglas-fir and for implementing genomic selection.

Table GS-2.3. SNPs available to practice genomic selection in Douglas-fir. This table shows the number of SNPs that were classified into six SNP quality groups (PolyHighResolution, NoMinorHom, OTV, MonoHighRes, and CallRateBelowThreshold) using an Affymetrix Axiom genotyping array. For each SNP, the call rate (CR) is the proportion of trees that were assigned a genotype (called) relative to the total number of trees tested (n = 1,694).

	No. of SNPs with call rate (CR) of:	
Classification	**97%**	**Affymetrix abbreviation: description**
Polymorphic high resolution	16,177	PolyHighResolution: These are the very best SNPs because they vary among trees (are polymorphic) and can be measured very accurately (are high-resolution). These SNPs pass all thresholds (CR.cut >= 97; FLD.cut >= 3.6; HetSO.cut >= -0.1; HomRO2.cut >= 0.3; HomRO3.cut >= 0.9; nMinorAllele.cut >= 2).
No minor homozygote	4,786	NoMinorHom: Minor alleles were found, but no minor homozygotes. Many of these are probably true SNPs, but the MAF may be too low to be valuable for genomic selection.
Monomorphic high resolution	10,141	MonoHighResolution: These SNPs are high-resolution, but they did not vary among trees (not polymorphic). They may not be true SNPs or the minor allele frequency may be very low, and not valuable for genomic selection.
Converted	**31,104**	This number (PolyHighResolution + NoMinorHom + MonoHighResolution) is a good indication of the success of the SNP genotyping platform itself.
Off-target variant	1,170	OTV: OTVs usually indicate that the DNA hybridized poorly to the genotyping array, perhaps because of other unknown SNPs near the target SNP. It may be possible to measure these SNPs after using the OTV_Caller program to re-call the genotypes.
Other	18,817	Other: These SNPs did not pass various quality thresholds for various reasons.
Call rate below threshold	4,675	CallRateBelowThreshold: The SNP was below the 97% or 80% CR threshold, but the SNP passed all other thresholds except that the number of minor alleles was ignored. For genomic selection, a CR of 85% is probably more than sufficient (Rutkoski et al. 2013).
Not converted	**24,662**	OTV + Other + CallRateBelowThreshold
Total	**55,766**	Total number of SNPs attempted on the '50K' genotyping array.

Conclusions/Discussion

SNPs classified as polymorphic and high-resolution (PolyHighResolution) are the ones that should work best for genomic selection. Using the default Affymetrix CR of 97%, 16,177 SNPs fell into this category (Table GS-2.3). However, for genomic selection, a CR of 85% is probably more than sufficient (Rutkoski et al. 2013). Therefore, we are now investigating genomic selection using lower CR thresholds and, thus, greater numbers of SNPs.

Two other categories of SNPs (NoMinorHom and MonoHighResolution) probably contain many true SNPs that can be measured reliably. However, their minor alleles may be too low for making them particularly valuable or genomic selection, at least in the populations we tested. Nonetheless, if we count all three categories of 'converted' SNPs (PolyHighResolution, NoMinorHom, and MonoHighResolution), we have about 30K SNPs that could contribute to the success of genomic selection. On the other hand, many of these may not be of sufficient quality, and we may need to exclude other SNPs in the PolyHighResolution category because of other issues, such as deviations from Hardy-Weinberg equilibrium. Balancing these considerations, and based on ongoing analyses, we conclude that we have between 20K and 30K SNPs that will allow us to practice genomic selection in Douglas-fir. This is probably more SNPs than are needed to practice effective genomic selection in NWTIC-type breeding programs.

TASK 3: MAKE SELECTIONS FOR INCREASED BIOFUEL PRODUCTION USING A COMBINATION OF PHENOTYPIC AND SNP GENETIC MARKER DATA.

Task Objective

In this study, we applied Genomic Selection (GS) to coastal Douglas-fir to investigate the accuracy and selection efficiency for the phenotypes of growth rate, biofuel product, wood chemistry properties, and branching characterless, by using SNP markers.

Methodology

This study was carried out using two series of coastal Douglas-fir full-sib progeny trials (previously described i.e. SCC and CL98) as training / validation populations. The overall objective was to explore the potential of accelerating breeding cycles of Douglas-fir through genomic selection. In this process, 640 trees were genotyped using an Axiom 55K SNP array with Call Rate (CR) ≥ 80%. All monomorphic SNP markers were excluded, but SNPs with rare alleles were retained. As a result, a total of 22,126 polymorphic SNPs were used.

The marker effects and, therefore, GEBVs were estimated using best linear unbiased prediction model (GBLUP). Our preliminary study indicated that the differences in PA between GBLUP and various Bayesian models (e.g., BL, BRR, BayesA, BayesB, and BayesC) were small for all the traits studied.

To assess prediction accuracy (PA) of GS, we used 10 replications of 10-fold cross-validation where 90% of the total population was used as a training population and 10% as the validation population. The PA was calculated as the mean Pearson correlation between the EBVs from pedigree-based models and the GEBVs from the GS models. The relative efficiency (RE) of GS to TS was estimated by comparing PAs from both schemes, assuming that the length of breeding cycle in GS is half of that in TS as a result of early selection.

For each of the 19 traits studied, we trained univariate genomic selection (UVGS) models with EBVs and validated GEBVs using the same (direct UVGS) or different (indirect UVGS) traits. In addition, we examined the consequences of including dominance variation in the UVGS models.

Since tree breeding programs normally deal with multiple trait selection, and some traits are difficult to evaluate, expensive, or they need a large sample size, we also evaluated and compared the accuracy of genomic predictions using multivariate genomic selection (MVGS) models. The following four scenarios were analyzed using MVGS models: (1) Training on TB, RF, SG, or VOL12; validated on TB, (2) Training on TB, RF, PY, or HY; validated on TB, (3) Training on HT12, HT7, DBH7, or VOL; validated on HT12, and (4) Training on HT12, HT7, DBH7, or VOL; validated on VOL12.

Results

The PAs from direct UVGS were relatively high for all the traits studied, ranging from 0.57 to 0.79 (Table GS-3.1). For example, the PA was 0.65 for age-12 height (HT12) and 0.64 for total biofuel product (TB) (Figures GS-3.1 and GS-3.2). The corresponding REs of GS to TS, assuming a conservative reduction of 50% in the length of the breeding cycle, were 1.79 and 1.92 respectively (Table GS-3.2, Figure GS-3.3), highlighting the increase in efficiency per unit time.

Table GS-3.1. Accuracy of genomic additive (A) and additive + dominance (AD) models for direct (i.e., same trait in training and validation) and indirect (i.e., different traits in training and validation) genomic predictions. HT, DBH, VOL, FORK, RAMI, SINU, and NRY are total height, diameter at breast height, volume index, number of incidents of forks, number of incidents of ramicorns, stem sinuosity score, and estimated years of needle retention, respectively. The trailing numbers refer to measurement ages. Age-17 wood chemistry traits include HY (hydrolysis yield), PH (pretreated holocellulose fraction), PY (pretreatment yield), RF (recalcitrance factor), and SG (specific gravity). TB is an index of total biofuel product calculated as VOL12 x SG x RF.

Model	Training trait	Validation trait																		
		DBH12	DBH7	FORK12	FORK7	HT12	HT7	HY	NRY7	PH	PY	RAMI12	RAMI7	RF	SG	SINU12	SINU7	TB	VOL12	VOL7
A	DBH12	0.586	0.588	0.285	0.333	0.256	0.229	-0.198	0.177	0.024	0.244	0.427	0.383	-0.147	-0.339	0.284	0.244	0.188	0.534	0.529
	DBH7	0.551	0.674	0.445	0.583	0.396	0.420		0.037			0.528	0.497			0.225	0.074		0.545	0.625
	FORK12	0.220	0.384	0.766	0.611	0.327	0.321	-0.079	0.132	0.153	-0.006	0.500	0.456	-0.067	0.089	0.257	0.105	0.252	0.297	0.405
	FORK7	0.264	0.492	0.615	0.765	0.363	0.323	-0.035	0.268	0.161	-0.091	0.529	0.499	-0.041	0.093	0.217	0.139	0.283	0.329	0.485
	HT12	0.225	0.363	0.367	0.409	0.650	0.567	-0.079	0.107	0.041	0.124	0.191	0.236	-0.053	0.019	0.199	0.151	0.194	0.371	0.463
	HT7	0.214	0.412	0.389	0.385	0.604	0.611	-0.035	0.006	0.039	0.102	0.201	0.271	0.001	0.014	0.199	0.131	0.222	0.355	0.499
	HY	-0.185		-0.128	-0.050	-0.087	-0.085	0.599		-0.205	-0.446	0.121	0.154	0.557	-0.066	-0.113	-0.113	0.125	-0.187	
	NRY7	0.212	0.069	0.181	0.376	0.114	0.012		0.790			0.069	0.156			0.055	0.015		0.178	0.073
	PH	0.019		0.141	0.170	0.086	0.104	-0.199		0.689	0.024	0.060	0.079	-0.119	0.188	-0.021	-0.021	-0.062	0.073	
	PY	0.263		0.011	-0.090	0.150	0.144	-0.423		0.019	0.633	-0.262	-0.276	-0.317	-0.023	-0.051	-0.047	-0.106	0.253	
	RAMI12	0.336	0.438	0.489	0.515	0.169	0.171	0.080	0.058	0.060	-0.230	0.784	0.707	0.032	-0.079	0.213	0.118	0.265	0.326	0.390
	RAMI7	0.309	0.428	0.461	0.503	0.215	0.232	0.124	0.128	0.080	-0.226	0.733	0.738	0.090	-0.096	0.223	0.130	0.253	0.315	0.403
	RF	-0.115		-0.110	-0.041	-0.029	-0.023	0.567		-0.114	-0.348	0.090	0.128	0.570	-0.067	-0.144	-0.142	0.120	-0.111	
	SG	-0.345		0.102	0.121	0.086	0.064	-0.065		0.201	-0.013	-0.130	-0.146	-0.065	0.582	-0.126	-0.133	-0.141	-0.256	
	SINU12	0.246	0.259	0.273	0.229	0.187	0.186	-0.071	0.096	-0.008	-0.051	0.238	0.239	-0.095	-0.082	0.727	0.669	0.143	0.269	0.251
	SINU7	0.206	0.117	0.116	0.148	0.141	0.123	-0.085	0.016	-0.010	-0.049	0.132	0.136	-0.112	-0.103	0.648	0.750	0.131	0.212	0.102
	TB	0.219		0.250	0.298	0.176	0.175	0.103		-0.068	-0.106	0.318	0.295	0.097	-0.134	0.102	0.119	0.642	0.282	
	VOL12	0.545	0.580	0.393	0.420	0.425	0.378	-0.188	0.172	0.055	0.220	0.420	0.394	-0.139	-0.294	0.319	0.262	0.273	0.569	0.581
	VOL7	0.518	0.639	0.488	0.588	0.523	0.521		0.057			0.477	0.476			0.221	0.065		0.568	0.644
AD	DBH12	0.586	0.581	0.290	0.335	0.250	0.222	-0.188	0.202	0.025	0.230	0.431	0.388	-0.141	-0.343	0.290	0.250	0.183	0.533	0.524
	DBH7	0.533	0.656	0.431	0.567	0.393	0.407		0.049			0.517	0.486			0.218	0.073		0.531	0.606
	FORK12	0.228	0.383	0.785	0.621	0.324	0.315	-0.094	0.138	0.146	-0.014	0.516	0.477	-0.083	0.068	0.278	0.131	0.263	0.302	0.403
	FORK7	0.267	0.493	0.621	0.769	0.366	0.322	-0.043	0.256	0.158	-0.075	0.529	0.499	-0.045	0.081	0.216	0.140	0.275	0.333	0.489
	HT12	0.229	0.358	0.372	0.409	0.650	0.568	-0.072	0.084	0.037	0.110	0.194	0.237	-0.045	0.013	0.195	0.149	0.198	0.376	0.460
	HT7	0.211	0.400	0.388	0.379	0.599	0.603	-0.060	0.008	0.037	0.106	0.200	0.271	-0.024	-0.008	0.204	0.138	0.213	0.350	0.482
	HY	-0.179		-0.137	-0.056	-0.085	-0.083	0.606		-0.212	-0.453	0.134	0.167	0.563	-0.074	-0.115	-0.116	0.108	-0.187	
	NRY7	0.209	0.057	0.196	0.393	0.108	-0.012		0.783			0.103	0.185			0.057	0.011		0.176	0.060
	PH	0.013		0.126	0.160	0.088	0.110	-0.193		0.685	0.017	0.062	0.081	-0.111	0.189	-0.026	-0.026	-0.065	0.071	
	PY	0.261		0.012	-0.093	0.140	0.136	-0.426		0.019	0.636	-0.273	-0.284	-0.319	-0.019	-0.034	-0.030	-0.096	0.246	
	RAMI12	0.339	0.435	0.501	0.523	0.169	0.165	0.088	0.041	0.078	-0.236	0.796	0.722	0.044	-0.095	0.229	0.132	0.263	0.324	0.380
	RAMI7	0.319	0.416	0.481	0.514	0.217	0.233	0.140	0.132	0.075	-0.247	0.753	0.759	0.101	-0.111	0.242	0.146	0.272	0.320	0.390
	RF	-0.123		-0.105	-0.034	-0.032	-0.024	0.580		-0.125	-0.344	0.074	0.116	0.585	-0.069	-0.161	-0.159	0.113	-0.119	
	SG	-0.355		0.092	0.117	0.063	0.036	-0.053		0.205	-0.020	-0.124	-0.139	-0.053	0.590	-0.126	-0.132	-0.143	-0.275	
	SINU12	0.248	0.265	0.294	0.243	0.181	0.183	-0.082	0.083	0.009	-0.049	0.254	0.259	-0.107	-0.111	0.742	0.672	0.130	0.266	0.255
	SINU7	0.206	0.103	0.126	0.153	0.128	0.113	-0.092	-0.005	0.007	-0.026	0.142	0.150	-0.117	-0.100	0.658	0.759	0.130	0.208	0.087
	TB	0.228		0.250	0.296	0.178	0.177	0.100		-0.077	-0.102	0.312	0.289	0.093	-0.150	0.100	0.117	0.639	0.289	
	VOL12	0.543	0.580	0.392	0.423	0.423	0.377	-0.177	0.157	0.045	0.208	0.423	0.400	-0.132	-0.299	0.318	0.260	0.264	0.567	0.583
	VOL7	0.506	0.628	0.481	0.584	0.508	0.511		0.060			0.476	0.477			0.217	0.062		0.555	0.638

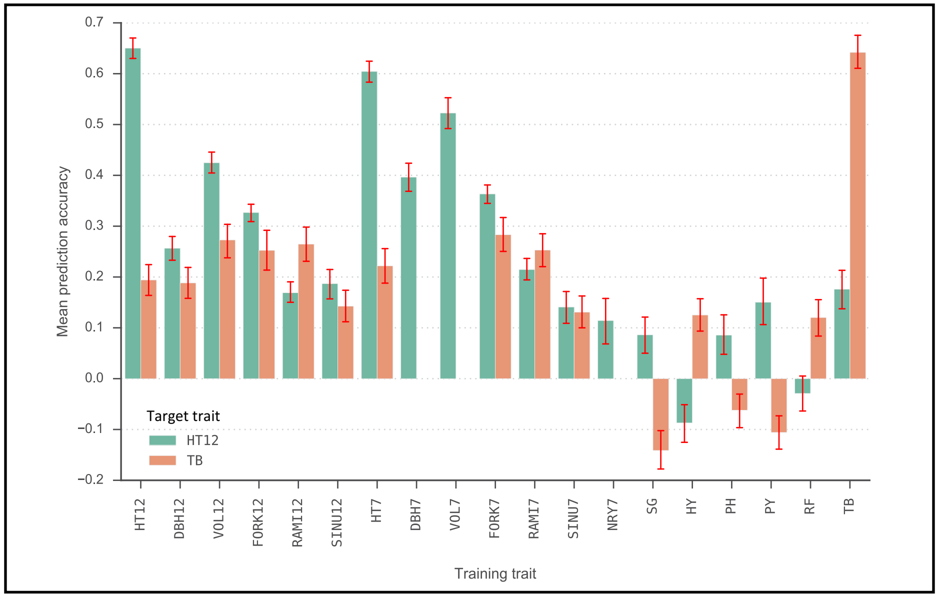


Figure GS-3.1. Mean prediction accuracy (PA) for age-12 height (HT12) and total biofuel product (TB).

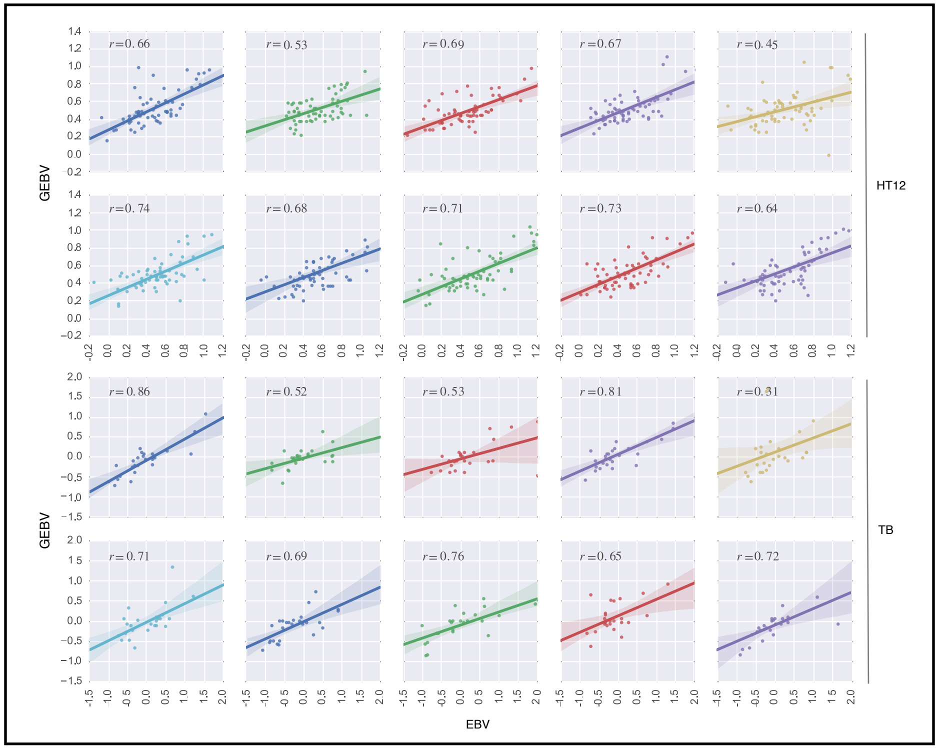


Figure GS-3.2. GEBVs vs. EBVs from a 10-fold cross-validation in the direct genomic selection for age-12 height (HT12) and total biofuel product (TB).

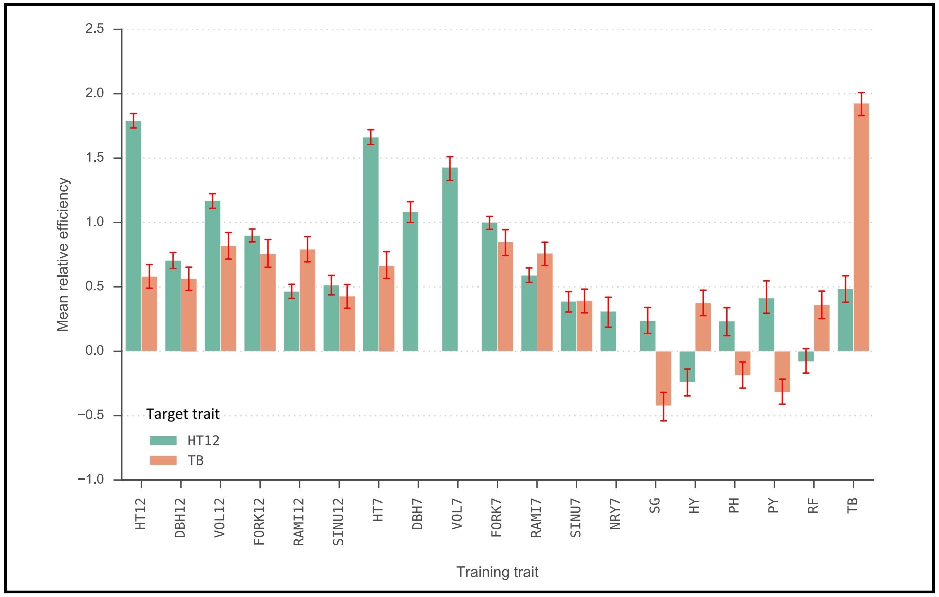


Figure GS-3.3. Mean relative efficiency of genomic selection to traditional selection for age-12 height (HT12) and total biofuel product (TB).

Table GS-3.2. Relative efficiency of genomic selection (GS) to traditional selection (TS) based on genomic additive (A) and additive + dominance (AD) models for direct (i.e., same trait in training and validation) and indirect (i.e., different traits in training and validation) genomic predictions, assuming that the length of breeding cycle in GS is half of that in TS. HT, DBH, VOL, FORK, RAMI, SINU, and NRY are total height, diameter at breast height, volume index, number of incidents of forks, number of incidents of ramicorns, stem sinuosity score, and estimated years of needle retention, respectively. The trailing numbers refer to measurement ages. Age-17 wood chemistry traits include HY (hydrolysis yield), PH (pretreated holocellulose fraction), PY (pretreatment yield), RF (recalcitrance factor), and SG (specific gravity). TB is an index of total biofuel product calculated as VOL12 x SG x RF.

Model	Training trait	Validation trait																		
		DBH12	DBH7	FORK12	FORK7	HT12	HT7	HY	NRY7	PH	PY	RAMI12	RAMI7	RF	SG	SINU12	SINU7	TB	VOL12	VOL7
A	DBH12	1.738	1.681	1.017	1.322	0.705	0.632	-0.519	0.544	0.088	0.577	1.456	1.242	-0.438	-1.004	0.855	0.746	0.563	1.550	1.490
	DBH7	1.611	1.927	1.564	2.279	1.082	1.148		0.114			1.762	1.581			0.665	0.222		1.565	1.761
	FORK12	0.653	1.098	2.735	2.424	0.899	0.886	-0.209	0.403	0.569	-0.015	1.705	1.480	-0.201	0.264	0.773	0.321	0.756	0.863	1.140
	FORK7	0.782	1.406	2.196	3.033	0.999	0.891	-0.093	0.821	0.602	-0.215	1.804	1.621	-0.123	0.277	0.653	0.424	0.848	0.955	1.367
	HT12	0.667	1.036	1.312	1.623	1.788	1.568	-0.208	0.326	0.154	0.294	0.651	0.767	-0.159	0.056	0.598	0.461	0.581	1.078	1.303
	HT7	0.634	1.177	1.387	1.527	1.662	1.688	-0.092	0.018	0.146	0.242	0.683	0.879	0.002	0.042	0.597	0.399	0.664	1.031	1.403
	HY	-0.551		-0.459	-0.194	-0.239	-0.236	1.568		-0.768	-1.057	0.416	0.506	1.660	-0.195	-0.339	-0.346	0.375	-0.545	
	NRY7	0.609	0.199	0.606	1.396	0.307	0.034		2.429			0.220	0.481			0.158	0.043		0.502	0.207
	PH	0.056		0.507	0.681	0.234	0.286	-0.522		2.569	0.057	0.210	0.261	-0.355	0.555	-0.066	-0.067	-0.186	0.212	
	PY	0.785		0.042	-0.357	0.413	0.398	-1.109		0.071	1.498	-0.903	-0.905	-0.945	-0.068	-0.156	-0.145	-0.317	0.738	
	RAMI12	0.997	1.251	1.745	2.043	0.465	0.472	0.210	0.180	0.222	-0.543	2.671	2.294	0.096	-0.233	0.640	0.359	0.792	0.947	1.097
	RAMI7	0.917	1.223	1.647	1.995	0.590	0.642	0.324	0.392	0.298	-0.537	2.498	2.396	0.269	-0.284	0.672	0.397	0.758	0.915	1.136
	RF	-0.342		-0.402	-0.168	-0.079	-0.062	1.485		-0.425	-0.824	0.313	0.421	1.696	-0.197	-0.436	-0.438	0.359	-0.321	
	SG	-1.028		0.367	0.485	0.236	0.178	-0.172		0.750	-0.030	-0.446	-0.478	-0.193	1.726	-0.382	-0.410	-0.423	-0.745	
	SINU12	0.729	0.740	0.977	0.910	0.514	0.514	-0.187	0.296	-0.028	-0.121	0.813	0.775	-0.281	-0.243	2.186	2.044	0.428	0.781	0.706
	SINU7	0.612	0.335	0.415	0.588	0.388	0.341	-0.224	0.048	-0.037	-0.114	0.450	0.443	-0.333	-0.304	1.948	2.292	0.392	0.617	0.289
	TB	0.655		0.900	1.190	0.484	0.486	0.270		-0.254	-0.251	1.096	0.967	0.288	-0.397	0.308	0.367	1.922	0.823	
	VOL12	1.616	1.657	1.405	1.669	1.168	1.046	-0.493	0.531	0.204	0.522	1.430	1.279	-0.415	-0.872	0.957	0.799	0.818	1.652	1.638
	VOL7	1.515	1.827	1.713	2.302	1.426	1.427		0.176			1.590	1.512			0.648	0.191		1.631	1.815
AD	DBH12	1.738	1.659	1.036	1.330	0.688	0.615	-0.493	0.622	0.094	0.545	1.470	1.260	-0.419	-1.017	0.872	0.765	0.547	1.549	1.473
	DBH7	1.559	1.877	1.515	2.217	1.074	1.114		0.150			1.724	1.543			0.642	0.216		1.525	1.707
	FORK12	0.675	1.097	2.804	2.464	0.891	0.872	-0.245	0.424	0.544	-0.033	1.758	1.548	-0.249	0.202	0.836	0.403	0.787	0.877	1.135
	FORK7	0.793	1.410	2.219	3.051	1.006	0.889	-0.114	0.788	0.590	-0.178	1.802	1.620	-0.136	0.241	0.650	0.427	0.822	0.967	1.376
	HT12	0.678	1.024	1.328	1.619	1.789	1.570	-0.189	0.255	0.137	0.260	0.660	0.769	-0.133	0.039	0.584	0.454	0.593	1.091	1.295
	HT7	0.624	1.142	1.384	1.504	1.648	1.666	-0.158	0.022	0.136	0.251	0.680	0.879	-0.071	-0.024	0.612	0.419	0.638	1.016	1.357
	HY	-0.532		-0.492	-0.222	-0.231	-0.229	1.587		-0.793	-1.074	0.460	0.547	1.675	-0.221	-0.348	-0.357	0.321	-0.542	
	NRY7	0.603	0.163	0.660	1.459	0.288	-0.032		2.407			0.336	0.572			0.166	0.031		0.498	0.170
	PH	0.039		0.453	0.638	0.241	0.303	-0.506		2.557	0.042	0.218	0.270	-0.332	0.560	-0.077	-0.078	-0.194	0.206	
	PY	0.778		0.043	-0.372	0.385	0.378	-1.116		0.069	1.506	-0.941	-0.935	-0.949	-0.056	-0.104	-0.094	-0.288	0.717	
	RAMI12	1.005	1.244	1.789	2.076	0.465	0.456	0.231	0.130	0.288	-0.558	2.714	2.344	0.132	-0.281	0.689	0.402	0.787	0.942	1.070
	RAMI7	0.947	1.189	1.718	2.040	0.596	0.644	0.368	0.406	0.281	-0.585	2.565	2.466	0.302	-0.327	0.729	0.448	0.815	0.930	1.098
	RF	-0.365		-0.378	-0.135	-0.086	-0.065	1.519		-0.465	-0.815	0.255	0.379	1.740	-0.203	-0.486	-0.486	0.338	-0.344	
	SG	-1.059		0.327	0.464	0.169	0.096	-0.139		0.766	-0.047	-0.429	-0.458	-0.157	1.748	-0.383	-0.409	-0.427	-0.803	
	SINU12	0.736	0.759	1.051	0.963	0.498	0.505	-0.214	0.253	0.035	-0.116	0.866	0.841	-0.319	-0.328	2.233	2.056	0.388	0.774	0.717
	SINU7	0.612	0.295	0.448	0.605	0.351	0.313	-0.240	-0.018	0.027	-0.062	0.482	0.485	-0.350	-0.297	1.978	2.321	0.387	0.602	0.245
	TB	0.681		0.901	1.187	0.489	0.490	0.261		-0.288	-0.241	1.076	0.949	0.276	-0.444	0.303	0.360	1.912	0.843	
	VOL12	1.611	1.659	1.398	1.676	1.164	1.043	-0.464	0.485	0.170	0.493	1.442	1.298	-0.393	-0.885	0.956	0.793	0.789	1.646	1.642
	VOL7	1.481	1.796	1.691	2.284	1.387	1.397		0.181			1.588	1.516			0.640	0.182		1.593	1.795

The indirect UVGS revealed interesting patterns. For height and volume growth at age 12, the models developed at age 7 and age 12 performed equally well in predicting the growth at age 12 (Figure GS-3.4). For example, the PA was 0.6 for the model trained on HT7 and validated on HT12. This number was almost the same as the PA (=0.61) from the direct GS on HT12. For wood chemistry and biofuel traits, however, PAs from indirect GS were generally much lower than that from the direct GS.

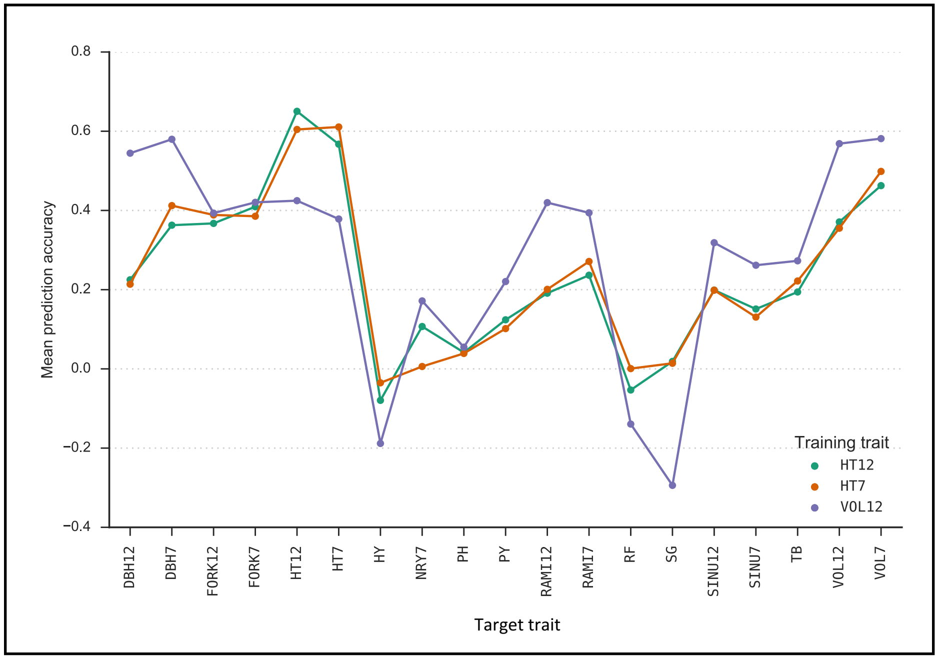


Figure GS-3.4. Mean prediction accuracy (PA) for all traits when training on age-7 height (HT12), and age-12 volume (VOL12).

Results indicated that the additive model (A) and the combined additive and dominance model (AD) produced similar predictive abilities for all traits (Figure GS-3.5), despite the fact that dominance variation did contribute some genetic variance in some traits. This suggests that there is little merit of including genomic dominance effects in the GBLUP prediction models.

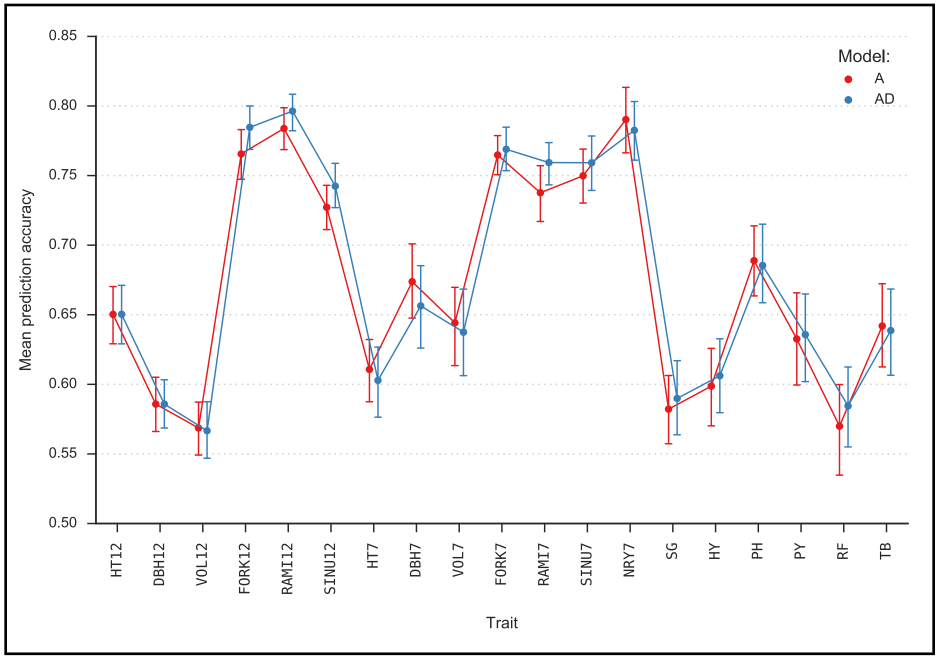


Figure GS-3.5. Mean prediction accuracy (PA) in direct GS: additive (A) vs. Additive + dominance (AD) models.

MVGS provided higher PA and RE in each scenario (Figures GS-3.6 and GS-3.7); all were higher than their respective cross-validated UVGS results. It appears that MVGS exploits even weak trait correlations, and provided improved accuracy in a time and cost manner thus increasing genetic gain from selection among untested genotypes.

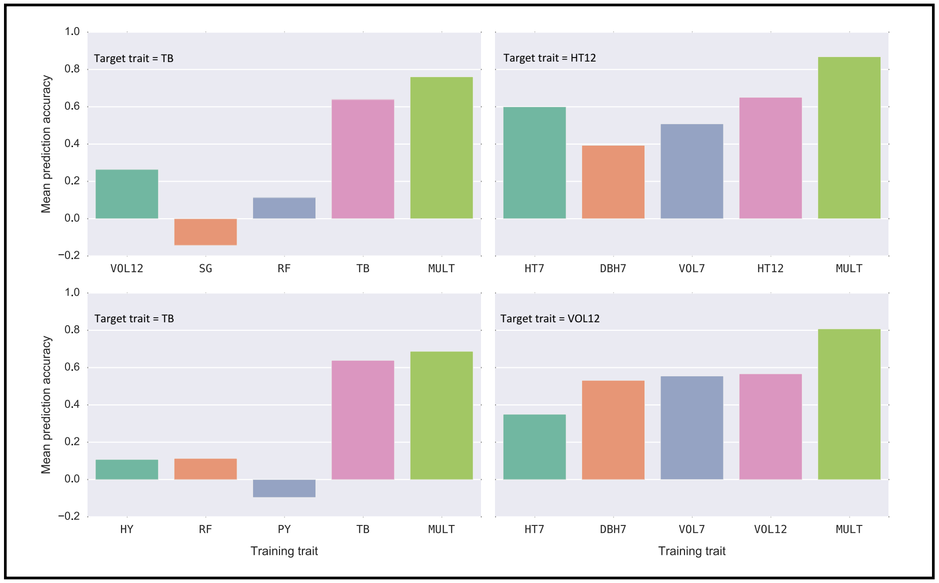


Figure GS-3.6. Comparisons of mean prediction accuracy in univariate / multivariate analyses.

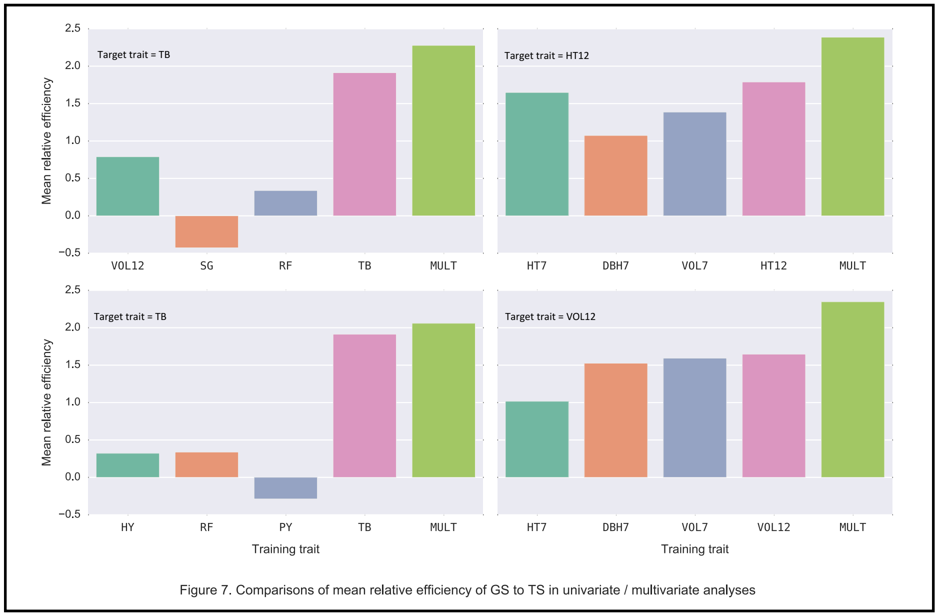


Figure 7. Comparisons of mean relative efficiency of GS to TS in univariate / multivariate analyses

Figure GS-3.7. Comparisons of mean relative efficiency of GS to TS in univariate / multivariate analyses.

Conclusions/Discussion

In conclusion, this study showed encouraging results of applying genomic selection in coastal Douglas-fir. Remarkable gain can be achieved by incorporating genomic selection in breeding programs. The predictive abilities of SNP markers were around 0.60 for growth and biofuel product, and 0.75 for branching / stem straightness in univariate models. They are comparable to the accuracies estimated in the pedigree-based TS. For example 0.710 for HT12 (compared to 0.710 by pedigree-based selection), and .62 vs. .69 for VOL12.

For growth traits, accuracies remained high when using models generating at age 7 to predict phenotypes at age 12. For age-12 growth and branching traits, genomic selection models trained at age-7 had similar predictive abilities as models trained at age-12. Prediction using multivariate models were generally more accurate, but the increase of accuracy depends on the relationship among traits.

Assuming that the length of breeding cycle in genomic selection is half of that in field-based selection, the relative efficiency of genomic selection to field-based ≈ 200%. Prediction accuracies from some other studies in forestry species were as follows: Loblolly pine (Resende Jr et al., 2012): 0.63 – 0.74 for HT6, 0.65 – 0.75 for DBH6; Eucalyptus (Resende et al., 2012): 0.73 – 0.79 for HT3, 0.65 – 0.78 for SG4; Maritime pine (Isik et al. 2016): 0.47 for HT12, 0.43 for DBH12; Loblolly pine (Resende Jr et al. 2012): 0.39 for HT, 0.46 for DBH; Interior spruce (Ratcliffe et al., 2015): 0.37 – 0.47 for HT (ages 3 - 40).

We have tried to optimize prediction procedures in genomic selection in the following ways:

- *Compare different statistical approaches*: GBLUP vs. Bayesian methods: GBLUP method performed equally well as Bayesian methods in general.
- *Add non-additive component to the additive genomic selection model*: Including non-additive component in the genomic selection model did not improve prediction accuracy for most traits. For FORK12, SINU12 and SINU7, Adding dominance effect into the genomic selection model boosted prediction accuracy by 13 – 31%.
- *Use multiple-trait models to make use of among-trait correlations*: Multiple-trait models are better than single-trait models even when the among-trait correlations were weak. However, multiple-trait models show no benefit for predicting new individuals without any phenotypic information.
- *Use a subset of SNP markers to reduce genotyping cost*: It appears that similar predictive ability can be reached by using only a subset of SNP markers (~3K).

The results from this study should motivate implementation of genomic selection in Douglas-fir cooperative breeding programs.

There are several outstanding issues for genomic selection in Douglas-fir:

- *What is the optimal size / age / type of reference population*? The efficiency of genomic selection largely depends on the design of the reference population.
- *Can different breeding zones or regions share the same genomic selection model*? Our data are only relevant to a single breeding zone. A study in loblolly pine also showed that prediction accuracy remained high across sites as long as they were used within the same breeding zones.
- *How many generations does a genomic selection model need to be retrained*? Results from dairy cattle breeding suggested that prediction accuracy eroded quickly with generations.

What is the cost-benefit analysis (genomic selection vs. TS)?

The genotyping cost was $75 / tree, the DNA extraction probably added $5-10 more per tree. In contrast, growing, planting, measuring a Douglas-fir progeny tree is about $10-20 / tree. However, relative benefits of genomic selection for Douglas-fir may be higher than other important conifer species (e.g., radiata pine, southern pines, and eucalypts). The testing cycle is longer for Douglas-fir, and testing costs much higher (fencing is needed). The crucial breakthrough would be decreasing genotyping costs (e.g., fewer SNPs, larger volume, etc.).

NARA OUTPUTS

Geleynse S., Alvarez-Vasco C., Garcia, K., Jayawickrama K., Trappe M. and Zhang X. 2014. "A Multi-Level Analysis Approach to Measuring Variations in Biomass Recalcitrance of Douglas fir," BioEnergy Research DOI: 10.1007/s12155-014-9483-z.

Geleynse S., Jayawickrama K., Trappe M., Ye T., Zhang X. 2016. Genetic Parameters of Factors Affecting the Biomass Recalcitrance of Douglas fir Trees BioEnergy Research DOI: 10.1007/s12155-016-9718-z.

Geleynse, S., Alvarez-Vasco, C., Jayawickrama, K., Trappe, M, Garcia, K. and Zhang, X. 2013. Phenotypic variations of biomass recalcitrance in Douglas-fir families. Poster presented at 35th Symposium on Biotechnology for Fuels and Chemicals, April 29- May 2, 2013. Hilton Portland, Portland, Oregon.

Geleynse, S., K. Jayawickrama, K. Garcia, M. Trappe and X. Zhang. Improving Douglas-Fir Feedstocks by Screening Families for Biomass Recalcitrance. Poster presentation at the NARA Annual Meeting, Corvallis, OR, September 10, 2013.

Jayawickrama, K.J.S., G. Howe, S. Guida and C.J. Bell. 2013. SNP chip development for Coastal Douglas-fir. Poster presentation at the NARA Annual Meeting, Corvallis, OR, September 10, 2013.

Jayawickrama, KJS. 2013. Overview of Feedstock Development: NARA Years 1-3. Presentation at 2nd NARA Annual Meeting, September 12, 2013, Oregon State University, Corvallis, OR

NARA OUTCOMES

The estimates of heritability and predicted genetic gain show that it would be quite feasible to genetically select Douglas-fir for conversion to jet fuel. Given the sample sizes, these estimates should not be taken as the last word in genetic parameter estimates: we would typically want to sample from 100 families and 30 trees per family on at least three sites to increase our confidence in the estimates. However these results show a lot of promise.

From the CL98 test population, it would be possible to collect seed from a group of selected parents and start establishing high jet-fuel plantations in the near future. However for large-scale implementation into breeding programs in the Pacific Northwest it would essential to either (1) identify indirect selection traits that are less expensive to measure or (2) find ways to simplify and accelerate the measurement of the wood chemistry traits so that we could (3) screen many more populations and trees.

This study sets the stage for the application of high-density genotyping and genomic selection in coastal Douglas-fir in the Pacific Northwest. The results from this study was very promising, since a 50% increase in selection efficiency by shifting to GS would substantially increase the rate of delivering genetic gain to Douglas-fir breeding programs. There would need to be reductions in the cost of genotyping, however, since GS is not necessarily less expensive than progeny-test based breeding.

FUTURE DEVELOPMENT

In the future, we plan to optimize the prediction procedures in GS in terms of population sampling strategy, cost-effective genotyping strategy, and consideration of G x E effect (e.g., GS at very early stage, across wide range of test sites, etc.). We will explore the possibility of replacing the individual-tree model used since 2003 with single-step model by combining genotypes, phenotypes, and pedigree. We also plan to conduct cost analysis for incorporating GS into Douglas-fir breeding programs.

LIST OF REFERENCES

Geleynse, S., Jayawickrama, K., Trappe, M., Ye,T. & Zhang, X. (2016) Genetic Parameters of Factors Affecting the Biomass Recalcitrance of Douglas-Fir Trees. *BioEnerg. Res*., 9(3), 731-739. doi: 10.1007/s12155-016-9718-2

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

Howe, G.T., Yu, J.B., Knaus, B., Cronn, R., Kolpak, S., Dolan, P., Lorenz, W.W. & Dean, J.F.D. (2013). A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. *BMC Genomics*, 14, 137.

Iwata, H., Hayashi, T., & Tsumura, Y. (2011). Prospects for genomic selection in conifer breeding: a simulation study of Cryptomeria japonica. *Tree Genetics & Genomes, 7(4), 747-758.*

Isik, F., Bartholomé, J., Farjat, A., Chancerel, E., Raffin, A., Sanchez, L., Plomion, C. & Bouffier, L. (2016). Genomic selection in maritime pine. *Plant Science,* 242, 108–119.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M. & Meuwissen, T.H.E. (2009). The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation *Genetics*,183(3), 1119-1126. doi: 10.1534/genetics.109.107391.

Muller, T., Ensminger, I. & Schmid, K.J. (2012). A catalogue of putative unique transcripts from Douglas-fir (Pseudotsuga menziesii) based on 454 transcriptome sequencing of genetically diverse, drought stressed seedlings. *BMC Genomics,* 13.

Resende, M.D.V., Resende, M.F.R., Sansaloni, C.P., Petroli, C.D., Missiaggia, A.A., Aguiar, A.M., Abad, J.M., Takahashi, E.K., Rosado, A.M., Faria, D.A., Pappas, G.J., Kilian, A. & Grattapaglia, D. (2012a). 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, M.F.R., Munoz, P., Acosta, J.J., Peter, G.F., Davis, J.M., Grattapaglia, D., Resende, M.D.V. & 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.

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

Resende Jr, M.F., Munoz, P., Acosta, J.J., Peter, G.F., Davis, J.M., Grattapaglia, D., Resende, M.D. & Kirst, M. (2012). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol*, 193, 617–624.

Ratcliffe, B., El-Dien, O.G., Klápště**, J.,** Porth, I., Chen, C., Jaquish, B. & 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, 547-555.

Rutkoski, J.E., Poland, J., Jannink, J.L. & Sorrells, M.E. (2013). Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. *G3-Genes Genomes Genetics*, 3(3), 427-439.