Original research paper

JNIRS

Journal of Near Infrared Spectroscopy
2018, Vol. 26(2) 117–132

DOI: 10.1177/0967033518770211
journals.sagepub.com/home/jns
SAGE

Global near infrared spectroscopy models to predict wood chemical properties of *Eucalyptus*

Gary R Hodge[1], Juan Jose Acosta[1], Faride Unda[2], William C Woodbridge[1] and Shawn D Mansfield[2]

Abstract

Global near infrared spectroscopy models (multiple-species, multiple-sites) were developed to predict chemical properties of *Eucalyptus* wood. The sample data set included 186 samples from four data sets (five species) originating from six countries: *Eucalyptus urophylla* from Argentina, Colombia, Venezuela, and South Africa; *Eucalyptus dunnii* from Uruguay; *Eucalyptus globulus* and *Eucalyptus nitens* from Chile; and *Eucalyptus grandis* from Colombia. The 186 samples were all preselected from larger collections of 400 to nearly 1800 samples to represent the range of chemical and spectral variation in each data set. The chemical traits modeled were total lignin, insoluble lignin, soluble lignin, syringyl–guaiacyl ratio (S/G), glucose, xylose, galactose, arabinose, and mannose. Single-species models and global multiple-species models were developed for each chemical constituent. For the global model, the R^2_{cv} for total lignin, insoluble lignin and syringyl–guaiacyl ratio were 0.95, 0.96, and 0.86, respectively. An alternate expression of the syringyl–guaiacyl relationship (S/(S+G)) resulted in better near infrared calibrations (e.g., for the global model, $R^2_{cv} = 0.95$). The global models for sugar content were also very good, but were slightly inferior to those for the lignin related traits, with $R^2_{cv} = 0.74$ for glucose, 0.89 for xylose, and from 0.72 to 0.91 for the minor sugars. To investigate the utility of the global models to predict chemical traits for species not included in the calibration, three-species calibrations were used to predict each trait in a fourth species data set. The prediction fit statistics ranged from excellent to poor depending on the species and trait, but in general the predictions would be at least moderately useful for most species-trait combinations. For some species-trait combinations with poor initial predictions from the global model, the inclusion of 10 samples from the ''new'' species into the calibration global model improved the fit statistics substantially. The global calibrations will be useful in tree breeding programs to rank species, families, and clones for important wood chemical traits.

Keywords

Near infrared, global calibrations, lignin, S/G ratio, glucose, xylose, breeding, indirect selection, wood quality, *Eucalyptus*

Received 17 January 2018; accepted 21 March 2018

Introduction

Near infrared (NIR) spectroscopy can be used to provide rapid indirect assessments of chemical properties of various materials, including plant-derived materials such as grains and wood. As such, NIR is increasingly being utilized in the forest and forest products industry,[1] and in particular, in forest tree genetic improvement programs.[2–5] The primary use of NIR in tree breeding programs is to assess wood chemical properties, including pulp yield and wood lignin and cellulose content.[6–10] In addition, several studies report the utility of NIR to measure solid wood properties such as density, microfibril angle, modulus of elasticity, and nonrecoverable collapse.[11–19] Compared to traditional laboratory methods for measuring wood properties, NIR measurements offer the advantages of rapid speed, relatively low cost, easy and precise assessment of NIR spectra, nondestructive sampling, and perhaps most importantly, the ability to assess numerous traits with one analysis.[18] Most wood property traits are under a relatively high degree of genetic control, and genetic gains in wood property traits can have a significant impact on profitability.[20–22] However, breeders are generally interested in screening many

[1]Camcore, Department of Forestry & Environmental Resources, College of Natural Resources, North Carolina State University, Raleigh, NC, USA
[2]Department of Wood Science, University of British Columbia, Vancouver, Canada

Corresponding authors:
Gary R. Hodge, College of Natural Resources, North Carolina State University, 3221 Jordan Hall II, Raleigh, NC 27695, USA.
Email: grh@ncsu.edu
Juan Jose Acosta, College of Natural Resources, North Carolina State University, 3221 Jordan Hall II, Raleigh, NC 27695, USA.
Email: jjacosta@ncsu.edu

hundreds or thousands of selections candidates, so rapid, precise, and low cost assessments of wood traits is very important.

Camcore is an international university-industry research partnership working in the area of gene conservation and tree improvement (https://camcore.cnr.ncsu.edu). The program began in 1980, and currently includes 28 active and 4 associate members in 20 countries. Many of these organizations plant multiple species of pines and/or eucalypts, including tropical, subtropical and temperate species. Camcore has previously developed global NIR models (i.e., robust, multi-site, multi-species calibration models) for pines to predict lignin and cellulose content.[23,24] Comparable global NIR models to predict *Eucalyptus* wood lignin and cellulose contents, but also the quantity of xylan and lignin monomer ratio (often referred to as S:G ratio) would be of significant value, particularly for vertically integrated pulp and paper companies. The current study utilizes a data set of 186 wood samples (preselected from a larger set of 3901 samples) representing five commercially important eucalypt species originating from six countries to develop NIR models for important wood chemical traits. Camcore members would use these models to study the genetic control of wood chemical traits, and to screen breeding populations of the five species; in addition, the models would likely be used to screen populations of other species (and/or hybrids) not included in this study. The specific objectives of this study were:

1. To develop global NIR spectroscopy models to predict chemical composition of wood samples from different tropical, subtropical, and temperate eucalypt species grown in different locations around the world,
2. To examine how well a multi-species calibration could be extended to or extrapolated to an independent data set with samples from a different species, and
3. To examine how incorporation of a small number of samples from a new species into a multi-species calibration would improve the predictions for the new species.

Materials and methods

Four sets of samples were included in this study: 50 samples of *Eucalyptus urophylla* from various countries, 50 samples of *Eucalyptus dunnii* from Uruguay, 50 samples of *Eucalyptus globulus* and *Eucalyptus nitens* from Chile (41 and 9 samples, respectively), and 36 samples of *Eucalyptus grandis* from Colombia. In each case, the 50 samples were pre-selected from a larger set of samples ranging from 400 to 1795 (Table 1, and further described below).

All samples that were scanned with NIR were prepared and handled in the same way. Briefly, solid wood samples were chopped into smaller pieces using machetes, knives, or guillotines, and then ground into woodmeal using a Wiley mill to pass through a screen with 1.6 mm holes. The samples were dried at 50°C for 24 h, and then allowed to come to room temperature. For each sample, 4 g of wood meal was then scanned in a Foss NIRsystems 6500 NIR spectrometer using a spinning sample module. Reflectance readings were taken for NIR wavelengths from 1100 to 2500 nm, at 2 nm intervals, with a total of 32 scans averaged to produce a single reflectance spectrum for each sample.

Cell wall chemical composition of all 186 samples was assessed using the same 4-g woodmeal samples that were scanned with NIR. Initially, the woodmeal was Soxhlet extracted overnight in hot acetone to remove extractives, and the extractive-free material was used for all further analyses. Lignin and carbohydrate content was determined with a modified Klason method,[25] in which extracted ground stem tissue (0.2 g) was treated with 3 mL of 72% H_2SO_4 and stirred every 10 min for 2 h. Samples were then diluted with 112 mL deionized water and autoclaved for 1 h at 121°C. The

Table 1. Samples used for the development of global NIR calibration models for wood chemistry traits of *Eucalyptus*.

				Number of samples	
Species	Country	Age (years)	Wood samples	Prescreened	Wetlab analysis
Eucalyptus urophylla	Argentina, Colombia, Venezuela, S. Africa	4–13	Breast height 12 mm increment cores, pith-to-bark	1795	50
Eucalyptus dunnii	Uruguay	4	Breast height 12 mm increment cores, pith-to-bark	400	50
Eucalyptus globulus, Eucalyptus nitens[a]	Chile	8–25	Breast height wood shavings	480	50
Eucalyptus grandis	Colombia	6	Breast height wedges	1226	36
5 species	**6 countries**			**3901**	**186**

NIR: near infrared.
[a]The initial prescreen data set included 408 samples of *E. globulus* and 72 samples of *E. nitens*. The final wetlab data set included 41 samples of *E. globulus* and 9 samples of *E. nitens*.

acid-insoluble lignin fraction was determined gravimetrically by filtration through a preweighed medium coarseness sintered-glass crucible, while the acid-soluble lignin component was determined spectrophotometrically by absorbance at 205 nm. Carbohydrate content was determined by anion exchange high-performance liquid chromatography (Dx-600; Dionex, Sunnyvale, CA) equipped with an ion exchange PA1 (Dionex) column, a pulsed amperometric detector with a gold electrode, and a SpectraAS3500 auto injector (Spectra-Physics).

The lignin monomer composition (S:G ratio) was determined as per Robinson and Mansfield,[26] and were analyzed by gas chromatography on a Hewlett Packard 5890 series II instrument, equipped with an autosampler, splitless injector, flame ionizing detector, and a 30 m, 5% diphenyl/95% dimethyl polysiloxane coated RTX-5MS 0.25 mm ID capillary column.

In summary, the traits measured were: glucose, xylose, mannose, arabinose, and galactose content; soluble lignin, insoluble lignin, and total lignin content, and the ratio of syringyl lignin to guaiacyl lignin (S/G ratio). As an alternate expression of lignin composition, the S/G ratios were converted into percentage of syringyl lignin = S/(S+G) × 100%, where S = syringyl lignin content and G = guaiacyl lignin content. For convenience, throughout the rest of this manuscript this variable will be denoted simply as S/(S+G). This means that all of the chemical traits are expressed in percentage units, except for S/G ratio, which is unitless. Final NIR models were developed using these 10 wetlab values.

Preselection of wood samples for wetlab analysis

For each of the four species sample sets, preselection of the 50-sample (or 36-sample) subset for wetlab chemistry was done using the same approach. All samples in each large set were scanned with NIR, and a prior NIR model was used to make predictions of chemical traits. A principal component analysis (PCA) was also done on the spectral data set for the species. The 50 samples for wetlab chemistry were then selected to ensure good representation of the range variation for predicted chemical variation, and for variation of the first two principal components of the spectral data set. In all cases, the same wood sample used for NIR scanning was used for wetlab analysis.

Eucalyptus urophylla: Wood sampling and preselection. The total number of *E. urophylla* samples available for selection was 1795 taken across 16 provenance-progeny test sites in Argentina, Colombia, South Africa, and Venezuela. Trees ranged in age from 4 to 13 years. All specimens were collected as 12-mm diameter pith-to-bark increment cores taken at breast height (1.3 m). The prescreening was done with three independent NIR models for pulp yield developed for *E. urophylla*, *E. nitens*, and *E. grandis*. Each model was a proprietary (unpublished) model developed for one of Camcore's industry partners, with pulp yield assessed according to company protocols, which differed among all three companies. In general, the pulp yield models were moderately precise, with R^2 of calibration ranging from = 0.63 to 0.70, and standard errors of cross-validation (SECV) ranging from ±0.84% to ±2.20%. Interestingly, the pulp yield predictions for the 1795 *E. urophylla* samples from the three different models (*E. urophylla*, *E. nitens*, and *E. grandis*) were rather highly correlated (R = 0.83 to 0.86). The final 50 samples selected for wetlab assessment came from 50 families from 47 provenances (8, 13, 14, and 16 samples from Argentina, South Africa, Colombia, and Venezuela, respectively).

Eucalyptus dunnii: Wood sampling and preselection. The total number of *E. dunnii* samples was 400 trees in a series of four provenance—progeny test sites in Uruguay. The trees were 4 years old, and all samples were 12-mm diameter pith-to-bark increment cores taken at breast height (1.3 m). The prescreening was done with the NIR models developed on the *E. urophylla* samples discussed above, for the traits of glucose, xylose, insoluble lignin content, and S/(S+G). The 50 samples selected for wetlab analysis came from 34 different families and 12 different provenances.

Eucalyptus globulus and Eucalyptus nitens: Wood sampling and preselection. The total number of *E. globulus* and *E. nitens* samples was 480 trees sampled in Chiles (specifically, 408 *E. globulus* originating from nine plantations, and 72 *E. nitens* from three plantations). The trees ranged in age from 8 to 25 years old, and all samples were taken at breast height (1.3 m) on the stem with an auger drill. Bark was removed from both sides of the tree, and wood shavings were sampled through the entire stem. The prior NIR models used for prescreening were two-species NIR models developed for *E. urophylla* and *E. dunnii* for glucose, xylose, insoluble lignin content, and S/(S+G). The final 50 samples selected for wetlab analysis included 41 *E. globulus* and 9 *E. nitens*. For purposes of convenience, this data set was treated as a single species throughout the rest of the analysis.

Eucalyptus grandis: Wood sampling and preselection. The total number of *E. grandis* samples was 1226 trees sampled from three clonal genetic trials in Colombia. The trees were 9 years old and were sampled by taking discs at breast height (1.3 m), which were then sectioned into wedges, so that an approximate 22° arc of the disc was used as the wood sample. DBH measurements at ages 6 and 9 were used to guide removal of the outer wood representing the last 3 years of growth, so the final wood samples are best described as being 6 years old. The prescreening was done with three-species NIR models developed for *E. urophylla*, *E. dunnii,* and *E. globulus* and *E. nitens*. The final 36 samples selected for wetlab analysis represented 30 different clones, and were taken from three different sites.

R-NIR pipeline

For all preliminary modeling, a single transformation was selected and used for all models (specifically, Multiplicative Scatter Correction + Savitzy-Golay second derivative with window size of seven points and a second-order polynomial). This transformation has generally given good results with a variety of traits in both pines and eucalypts. For the development of the final models discussed in this manuscript, a systematic approach was used to examine a number of different mathematical transformations to ultimately identify an optimum model for each data set and wood chemical trait. In general, for a given data set, a large number of data transformations produced models with similar cross-validation fit statistics (R^2_{cv} and SECV). Typically, the transformation with the highest R^2_{cv} was selected as the best model, although occasionally a transformation with very similar fit, but fewer factors, was identified as optimum.

All model development was done using R software (environment version 3.3.2),[27] and a pipeline was written to conduct three separate phases of model development and prediction: transformation and outlier detection, model calibration and cross-validation, and prediction of new observations. The R-NIR pipeline will be described here.

R-NIR pipeline: Transformation and outlier detection. This module reads in the spectral data set and mathematical pretreatments are applied to the NIR spectra to remove the scattering of diffuse reflections associated with sample particle size and improve the subsequent regression. Spectra were transformed using standard normal variate (SNV), multiplicative scatter correction (MSC), detrend (DT), and second derivative Savitzky–Golay smoothing with two different window sizes of five and seven points (SG5 and SG7). Additionally, a number of paired transformations were used, with the scattering correction methods applied prior to spectral derivatives,[28] generating six additional transformed data sets: standard normal variate + Savitzky–Golay with five and seven points (SNV_SG5 and SNV_SG7); multiplicative scatter correction + Savitzky–Golay with five and seven points (MSC_SG5 and MSC_SG7); and detrend + Savitzky–Golay with five and seven points (DT_SG5 and DT_SG7). Pre-processing of our NIR spectral data was done using the R packages "pls"[29,30] and "Prospectr."[30]

To identify outliers, local outliers factors (LOF) were calculated for all observations on each spectral database.[31] Individuals with LOF values greater than 2 were excluded from the analysis, using a LOF algorithm implemented in the R package "DMwR."[32]

R-NIR pipeline: Model calibration, cross-validation, and prediction. The second part of the pipeline merges wetlab information with the transformed and outlier free databases, and develops NIR prediction models for all wood traits mentioned above. Partial least squares regression (PLS) was implemented in the R-package "pls,"[29] and model performance was evaluated using leave-one-out (LOO) cross-validation. Desirable PLS NIR models are those that (1) maximize the coefficient of determination (R^2_{cv}), (2) minimize the SECV, and (3) have a small number of latent variables (projection factors). Once the best model has been selected, it can be used to predict cell wall chemical attributes with the NIR spectra (transformed and outlier-free) for any other samples. The code for the R-NIR pipeline can be modified to input spectral data sets other than from the FOSS 6500, and to other mathematical transformations, and it is available upon request.

NIR model development

For each of the four single-species data sets (URO = *E. urophylla*, DUN = *E. dunnii*, GLN = *E. globulus* + *E. nitens*, and GRA = *E. grandis*), NIR models were developed for the aforementioned 10 wood chemical traits. A global model combining all four data sets was also developed.

To examine how well the global models might extrapolate to other species not included in the original model, we used models developed on three species data sets (e.g., data sets A, B, and C) to predict wetlab chemical values for a fourth independent data set (e.g., data set D). For each wood chemical trait, and for each of the four "new species extrapolations" (e.g., ABC→D), we used the data transformation selected as the best for the four-species global model for the three-species models and extrapolations. For example, the best mathematical transformation for the four-species global model for insoluble lignin was the Savitzky–Golay second derivative with five points in the convolution window, and six PLS factors. This transformation was used for all lignin model extrapolations (e.g., ABC→D, ABD→C, etc.).

Finally, a set of four species-trait scenarios was selected to examine the question of improving global NIR models for "new" species not included in the original calibration. For a given ABC→D scenario for a particular trait, 10 random observations from data set D were included with the ABC data set for model calibration, and then predictions were made for the remaining 40 observations from data set D. This was repeated five times, and average R^2_p and standard error of prediction (SEP) were calculated.

Results

Wetlab chemical analyses

Mean wetlab values for all chemical traits across all four species are presented in Table 2. Mean total lignin was 28.7%, and mean acid-insoluble lignin was 24.1%. Corresponding laboratory standard errors for a particular sample measurement of these traits were

approximately ± 0.33% (Table 2). Mean soluble lignin was 4.5% with a very low lab standard error (±0.07%). The two major sugar components were glucose with mean = 46.7% (lab standard error ±0.54%), and xylose with mean = 12.8% (lab standard error ±0.21%). Minor sugar components galactose and mannose content were approximately 1.25%, with lab standard errors around ±0.05%. Mean arabinose content was slightly lower at 0.32%, but the laboratory measurements were also very reproducible, with lab standard error of ±0.02%.

Single-species NIR models

Eucalyptus urophylla. NIR models for *E. urophylla* are presented in Table 3. The best models were found for lignin-related traits (total lignin, acid-insoluble lignin, S/G ratio and S/(S+G), with R^2_{cv} ranging from 0.87 to 0.90. Relatively good models were also found for the major sugars glucose and xylose, with $R^2_{cv} = 0.78$ and 0.83, respectively. The model for the minor sugar galactose was also good, with $R^2_{cv} = 0.72$. The models for the other minor sugars arabinose ($R^2_{cv} = 0.41$) and mannose ($R^2_{cv} = 0.44$), and for soluble lignin ($R^2_{cv} = 0.29$) were less satisfactory.

Table 2. Mean wetlab chemistry values for 189 samples from four species of eucalypts, and lab standard errors for a single sample wetlab measurement.

Trait %	Mean	Lab SE
Lignin	28.7	0.3445
Acid-insoluble lignin	24.1	0.3217
Soluble lignin	4.5	0.0776
S/G ratio	4.0	0.0757
S/(S+G)	78.4	0.2560
Glucose	46.7	0.5453
Xylose	12.8	0.2118
Galactose	1.17	0.0451
Arabinose	0.32	0.0168
Mannose	1.33	0.0647

Eucalyptus dunnii. NIR models for *E. dunnii* are presented in Table 4. The best NIR models for *E. dunnii* were found for total lignin, acid-insoluble lignin, glucose and xylose, with R^2_{cv} ranging from 0.60 to 0.76. Interestingly, the best model for *E. dunnii* was for the minor sugar galactose ($R^2_{cv} = 0.80$). For the traits S/G and S/(S+G), the models were moderate with $R^2_{cv} = 0.46$ and 0.56, respectively.

In general, the *E. dunnii* NIR models had fit statistics inferior to the *E. urophylla* models (compare Tables 3 and 4). For some traits, this may be related to a smaller range of observed within-species variation for the chemical trait. For example, for *E. dunnii* and the trait acid-insoluble lignin, the range of observed wetlab values for model calibration was 4.4% and the NIR model had $R^2_{cv} = 0.64$. For *E. urophylla*, the range of observed acid-insoluble lignin values for model calibration was 15.2% and the NIR model had $R^2_{cv} = 0.87$. In fact, the SECV was lower for the *E. dunnii* acid-insoluble lignin model than for the *E. urophylla* model (SECV = ± 0.58% vs. ± 1.17%, respectively). There was a similar pattern observed for S/(S+G): for *E. dunnii*, range = 9.1%, $R^2_{cv} = 0.56$, and SECV = ± 1.60%, while for *E. urophylla*, range = 27.3%, $R^2_{cv} = 0.87$, and SECV = ± 1.75%.

Eucalyptus globulus–Eucalyptus nitens. The NIR models for *E. globulus* and *E. nitens* are presented in Table 5. In general, the models for this data set ranged from good to excellent, with R^2_{cv} ranging from 0.86 to 0.96 for lignin, acid-insoluble lignin, S/G, S/(S+G), glucose, xylose, and galactose. Even the minor sugars, arabinose and mannose, had good models with $R^2_{cv} = 0.70$ and

Table 3. Fit statistics for *Eucalyptus urophylla* single-species NIR calibration models for wood chemistry traits.

	Wetlab			NIR model					
Trait %	Mean	SD	Range	Transformation	Factors	R^2_c	SEC	R^2_{cv}	SECV	RPD_{cv}
Lignin	33.7	3.1	14.8	MSC_SG7	8	0.98	0.410	0.90	0.969	3.20
Acid-insoluble lignin	28.6	3.3	15.2	SNV_SG5	4	0.93	0.835	0.87	1.167	2.83
Soluble lignin	5.1	0.6	2.9	MSC_SG5	4	0.66	0.376	0.29	0.541	1.11
S/G ratio	3.0	0.7	3.3	SG7	9	0.98	0.091	0.90	0.212	3.30
S/(S+G)	74.3	4.9	27.3	SNV_SG7	10	0.99	0.444	0.87	1.745	2.81
Glucose	46.2	2.8	14.4	SNV_SG5	5	0.94	0.700	0.78	1.277	2.19
Xylose	12.0	1.1	4.6	SNV_SG7	6	0.93	0.286	0.83	0.453	2.43
Galactose	1.53	0.61	2.43	MSC_SG7	8	0.94	0.124	0.72	0.317	1.93
Arabinose	0.27	0.06	0.30	MSC_SG5	6	0.92	0.018	0.41	0.048	1.26
Mannose	0.99	0.39	1.81	MSC_SG7	6	0.81	0.172	0.44	0.292	1.34

MSG: multiplicative scatter correction; NIR: near infrared; SD: standard deviation; SEC: standard error of calibration; SECV: standard errors of cross-validation; SG: Savitzky–Golay; SNV: standard normal variate.

Table 4. Fit statistics for *Eucalyptus dunnii* single-species NIR calibration models for wood chemistry traits.

	Wetlab			NIR model						
Trait %	Mean	SD	Range	Transformation	Factors	R^2_c	SEC	R^2_{cv}	SECV	RPD_{cv}
Lignin	27.3	1.1	5.3	MSC_SG7	9	0.98	0.158	0.76	0.542	2.03
Acid-insoluble lignin	23.4	1.0	4.4	SG7	6	0.89	0.327	0.64	0.581	1.72
Soluble lignin	3.9	0.7	2.5	SG5	2	0.24	0.576	0.08	0.639	1.10
S/G ratio	4.5	0.7	2.9	SG7	7	0.89	0.249	0.46	0.547	1.28
S/(S+G)	81.6	2.4	9.1	SG7	7	0.91	0.729	0.56	1.604	1.50
Glucose	46.4	2.6	10.1	MSC_SG7	3	0.77	1.335	0.60	1.602	1.62
Xylose	15.1	1.4	5.7	SNV_SG5	3	0.83	0.596	0.69	0.796	1.76
Galactose	0.94	0.34	1.54	SNV_SG7	9	0.99	0.040	0.80	0.145	2.35
Arabinose	0.41	0.06	0.27	SG7	11	0.99	0.005	0.70	0.032	1.87
Mannose	1.41	0.49	2.05	SG5	4	0.79	0.224	0.38	0.384	1.28

MSG: multiplicative scatter correction; NIR: near infrared; SD: standard deviation; SEC: standard error of calibration; SECV: standard errors of cross-validation; SG: Savitzky–Golay; SNV: standard normal variate.

Table 5. Fit statistics for *Eucalyptus globulus* + *Eucalyptus nitens* NIR calibration models for wood chemistry traits.[a]

	Wetlab			NIR model						
Trait %	Mean	SD	Range	Transformation	Factors	R^2_c	SEC	R^2_{cv}	SECV	RPD_{cv}
Lignin	23.8	3.0	14.2	MSC_SG7	5	0.98	0.452	0.96	0.613	4.89
Acid-insoluble lignin	18.6	2.8	13.0	SG7	10	0.99	0.160	0.96	0.528	5.30
Soluble lignin	5.2	0.6	2.5	SNV_SG7	8	0.97	0.110	0.77	0.287	2.09
S/G ratio	5.6	1.4	5.2	MSC_SG7	8	0.98	0.195	0.86	0.515	2.72
S/(S+G)	84.1	3.7	13.8	SG7	12	0.99	0.146	0.92	1.047	3.53
Glucose	46.9	4.2	20.6	MSC_SG5	9	0.99	0.182	0.90	1.296	3.24
Xylose	11.7	2.6	10.7	SG7	5	0.98	0.386	0.96	0.517	5.03
Galactose	1.45	1.25	6.00	SNV_SG7	8	0.99	0.081	0.97	0.213	5.88
Arabinose	0.38	0.11	0.55	SG5	8	0.98	0.014	0.70	0.061	1.80
Mannose	1.50	0.70	3.12	SNV_SG7	8	0.98	0.094	0.74	0.341	2.05

MSG: multiplicative scatter correction; NIR: near infrared; SD: standard deviation; SEC: standard error of calibration; SECV: standard errors of cross-validation; SG: Savitzky–Golay; SNV: standard normal variate.

[a]The initial prescreen data set included 408 samples of *E. globulus* and 72 samples of *E. nitens*. The final wetlab data set included 41 samples of *E. globulus* and 9 samples of *E. nitens*.

0.74, respectively. For glucose, the range of *E. globulus–E. nitens* wetlab values was larger than for both *E. urophylla* and *E. dunnii* (20.6% vs. 14.4% and 10.1%, respectively), while the SECV for three species were roughly similar. Similarly, for xylose, the *E. globulus–E. nitens* range was 10.7%, vs. 4.6% and 5.7%, for *E. urophylla* and *E. dunnii*, respectively. This was broadly true for most of the other variables as well: the range in wetlab values was larger for *E. globulus–E. nitens* than in *E. urophylla* and *E. dunnii*, which contributes to higher R^2_{cv}.

Eucalyptus grandis. The NIR models for *E. grandis* are presented in Table 6. In general, the NIR models for *E. grandis* were moderate to good. The best models were found for S/G and S/(S+G), with $R^2_{cv} = 0.87$ and 0.93, respectively. Good models were also found for total lignin and acid-insoluble lignin, with $R^2_{cv} = 0.72$ and 0.78, respectively. Moderately good models were found for the major sugars, glucose and xylose, with $R^2_{cv} = 0.69$ and 0.68, respectively. Among the minor sugars (galactose, arabinose, mannose), the best model was found for galactose ($R^2_{cv} = 0.81$ vs. 0.50 and 0.55, respectively); this was consistent with the results from the other three species.

Comparisons among species for chemical composition

For each of the four species, the appropriate single-species NIR models were used to predict chemical traits for the entire population of samples (from which the small wetlab subpopulation was selected). For example, the *E. urophylla* NIR models (Table 3) were used to predict values for all 1672 *E. urophylla* (Table 1), and similarly for the other three species.

Table 6. Fit statistics for *Eucalyptus grandis* single-species NIR calibration models for wood chemistry traits.

	Wetlab			NIR model						
Trait %	Mean	SD	Range	Transformation	Factors	R^2_c	SEC	R^2_{cv}	SECV	RPD_{cv}
Lignin	30.3	2.2	10.7	SG7	9	0.98	0.311	0.72	1.128	1.95
Acid-insoluble lignin	26.6	2.3	10.5	SG7	9	0.99	0.273	0.78	1.041	2.21
Soluble lignin	3.7	0.6	2.3	SNV_SG7	10	0.99	0.036	0.74	0.290	2.07
S/G ratio	2.7	0.5	2.4	MSC_SG7	7	0.98	0.074	0.87	0.192	2.61
S/(S+G)	72.1	4.3	18.1	SNV_SG7	7	0.99	0.480	0.93	1.117	3.85
Glucose	47.6	3.3	15.5	MSC_SG7	2	0.77	1.538	0.69	1.788	1.85
Xylose	12.2	2.0	6.9	SNV	6	0.89	0.701	0.68	1.072	1.87
Galactose	0.64	0.41	1.85	SG7	3	0.76	0.200	0.81	0.275	1.49
Arabinose	0.20	0.02	0.10	SNV_SG7	7	0.91	0.007	0.50	0.016	1.25
Mannose	1.47	0.72	2.74	MSC_SG7	3	0.77	0.345	0.55	0.479	1.50

MSG: multiplicative scatter correction; NIR: near infrared; SD: standard deviation; SEC: standard error of calibration; SECV: standard errors of cross-validation; SG: Savitzky–Golay; SNV: standard normal variate.

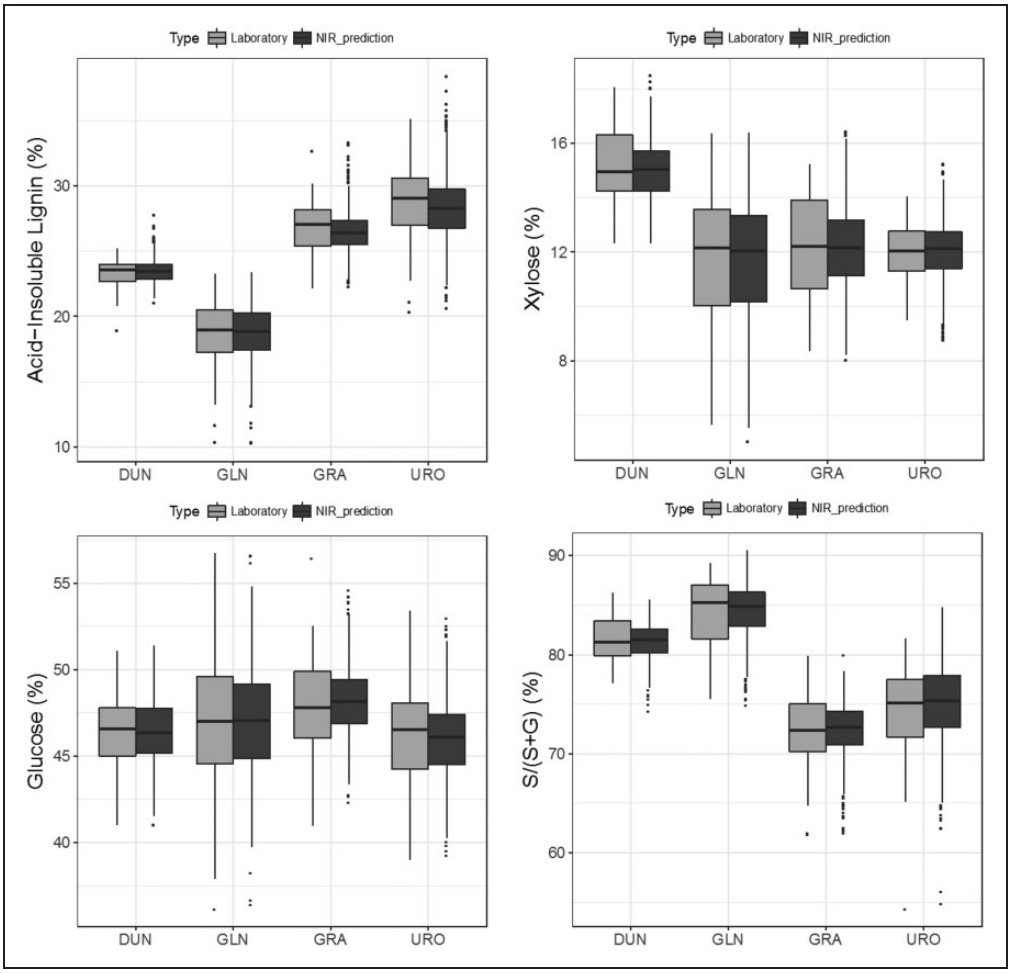

Figure 1. Boxplots for wood chemistry traits for four different *Eucalyptus* species data sets (URO = *Eucalyptus urophylla*; DUN = *Eucalyptus dunnii*; GLN = *Eucalyptus globulus* + *Eucalyptus nitens*, GRA = *Eucalyptus grandis*). For each species, the *left box plot* represents wetlab data for the small subset of samples selected by near infrared (NIR) pre-screening, and the *right box plot* represents NIR predictions for the entire population of samples, with predictions done using the appropriate single-species NIR calibration model developed in this study.

For the four species, boxplots of the chemical traits for the wetlab and full population were fairly similar (e.g., Figure 2), with the range of the 25th to 75th percentile slightly wider in the wetlab population, and the tails slightly wider in the full population. This would be expected, since the wetlab populations were small samples from the full population, which were intended to be somewhat uniformly sampled across the range.

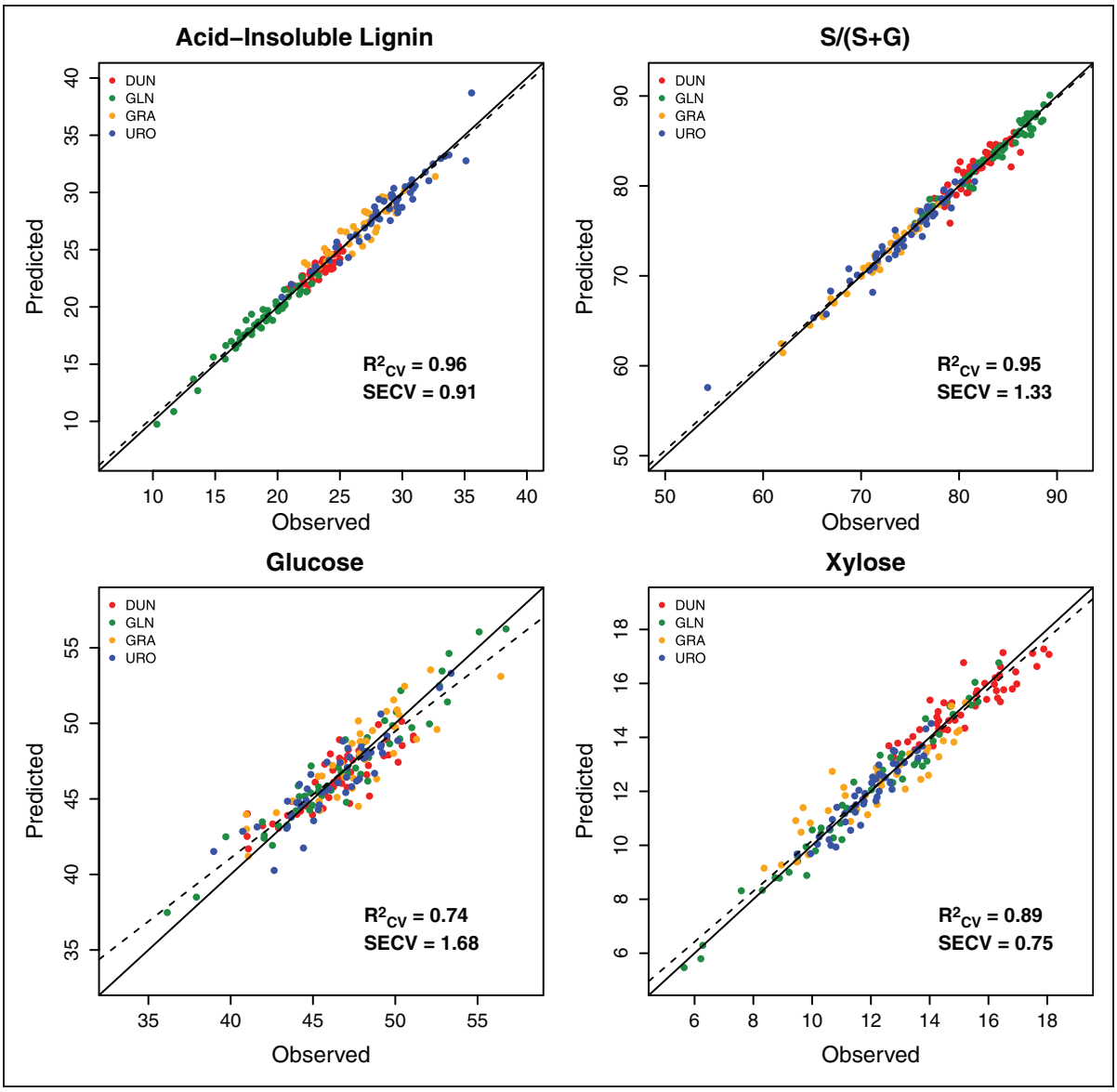

Figure 2. Global eucalyptus near infrared (NIR) cross-validation scatterplots for acid-insoluble lignin, S/(S+G), glucose, and xylose content. Laboratory-determined chemical content is on the *x*-axis, and NIR predicted value is on the *y*-axis.

Of more interest are some of the differences observed among the four species groups. Notably, the *E. globulus–E. nitens* populations had the lowest acid-insoluble lignin and highest S/(S+G) (18% and 85%), closely followed by *E. dunnii* with 24% acid-insoluble lignin and 82% S/(S+G). The *E. grandis* and *E. urophylla* had higher acid-insoluble lignin (26% and 28%, respectively) and lower S/(S+G) (73% and 75%, respectively), values which make these species less desirable for pulp production. The four species groups had similar distributions for glucose, but *E. dunnii* was unusual with its high xylose content, roughly 15% compared to 12% for the other three species groups.

Global calibration models–All data sets

Global NIR model statistics for a multiple-species model (all four data sets) are presented in Table 7. In general, the global models tended to include more factors than the single-species models; across all 10 traits the mean number of factors was 6.8 for the single-species models, and 9.7 for the global models. This could simply be a result of the larger data set (186 observations for the global model vs. 36 to 50 observations for the single-species models), or could be necessitated by the presence of different species requiring more complex models to account for the inherent variation in wood chemistry.

Very good to excellent models were obtained for total lignin, acid-insoluble lignin, S/G, S/(S+G), xylose and galactose, with R^2_{cv} ranging from 0.89 to 0.96. The models for glucose, arabinose, and mannose were moderate to good, with R^2_{cv} ranging from 0.72 to 0.75 for the three traits. SECV for the global models was similar to but slightly larger than the average

Table 7. Fit statistics for global *Eucalyptus* NIR calibration models for wood chemistry traits.[a]

	Wetlab			NIR model						
Trait %	Mean	SD	Range	Transformation	Factors	R^2_c	SEC	R^2_{cv}	SECV	RPD_{cv}
Lignin	28.7	4.6	26.0	SG5	7	0.97	0.846	0.95	1.055	4.36
Acid-insoluble lignin	24.1	4.6	25.2	SNV + SG5	6	0.97	0.751	0.96	0.905	5.08
Soluble lignin	4.5	0.9	3.9	SG7	10	0.81	0.404	0.65	0.545	1.70
S/G ratio	4.0	1.5	7.1	MSC + SG7	9	0.91	0.454	0.86	0.551	2.71
S/(S+G)	78.4	6.3	34.9	MSC + SG7	11	0.98	0.909	0.95	1.334	4.69
Glucose	46.7	3.3	20.6	MSC + SG7	9	0.84	1.312	0.74	1.680	1.95
Xylose	12.8	2.3	12.4	SNV + SG7	9	0.94	0.580	0.89	0.754	3.09
Galactose	1.17	0.82	6.07	SG7	13	0.97	0.150	0.91	0.252	3.27
Arabinose	0.32	0.11	0.63	SG7	8	0.81	0.048	0.72	0.058	1.91
Mannose	1.33	0.61	3.29	SG7	15	0.94	0.148	0.75	0.307	2.00

MSG: multiplicative scatter correction; NIR: near infrared; SD: standard deviation; SEC: standard error of calibration; SECV: standard errors of cross-validation; SG: Savitzky–Golay; SNV: standard normal variate.
[a]The data set includes samples of *Eucalyptus urophylla* ($n = 50$), *Eucalyptus dunnii* ($n = 50$), *Eucalyptus globulus* ($n = 41$), *Eucalyptus nitens* ($n = 9$), and *Eucalyptus grandis* ($n = 36$). Mean, SD, and range are for the wetlab data, fit statistics are for calibration and leave-one-out cross-validation.

SECV in the four single-species models. For example, for acid-insoluble lignin, the average single-species SECV was ±0.829%, while the global SECV was ±0.905%. Similarly, for glucose, the average single-species SECV was ±1.491%, while the global SECV was ±1.680%. The only exception to this trend was for the trait S/(S+G), where the average single-species SECV was ± 1.378%, while the global SECV was ±1.334%.

For many breeders interested in pulp and paper breeding objectives, the four most important wood chemical traits are glucose, xylose and acid-insoluble lignin content, and S/(S+G). Glucose is the primary component of cellulose, which is the target of the pulping process, and xylose is the other major sugar in wood, and is negatively associated with glucose content (i.e., wood with more xylose will typically have lower glucose). Acid-insoluble lignin is important because is the primary target for removal in the chemical pulping process. Regarding lignin composition (syringyl and guaiacyl lignin content), a high S/G ratio (or a high S/(S+G)) is beneficial in the pulping process in terms of reduction in chemical costs, as syringyl-rich lignins are inherently rich in cleavable β-ethers, making them less recalcitrant to the chemical pulping process.[33] Moreover, the greater the inherent incorporation of syringyl monomers into the lignin, the more linear the polymer (bond limitations imposed by the monomers largely only participating in β-0-4 linkages) and hence a greater ease of extraction during chemical processing. NIR model calibrations were better (higher R^2_{cv} and lower SECV) for the alternate expression of syringyl content, S/(S+G) than for the standard expression S/G for three of the four species and for the global calibration. For this reason, the trait S/(S+G) will be highlighted throughout the remainder of this manuscript.

Extrapolation of models to independent data sets (ABC→D)

It seems clear from the above results that the single-species and global models in this study could be used to predict values for samples from the same or similar populations. To examine the question of how well these models could be extended or extrapolated to independent data sets with samples from different species, all possible three-species models were be used to predict chemical traits for the fourth species data set.

Extrapolations were investigated for four important chemical traits (acid-insoluble lignin content, S/(S+G), glucose, and xylose). For acid-insoluble lignin and with the four-species global data set, the selected model used the paired SNV-SG5 transformation (Table 7). This transformation was then used for all four extrapolations, that is, ABC→D, ABD→C, etc., in other words, the calibration model was built using three species data sets (ABC), and then predictions for acid-insoluble lignin were made for data set D, etc. Similarly, investigation of S/(S+G) extrapolations was done using the SNV_SG7 transformation, for glucose the SNV_SG5 transformation was used, and for xylose the SNV_SG7 transformation was used. To examine how well the extrapolations functioned, model fit statistics for the predictions ($R^2_{p,}$ and SEP) were compared to the model fit statistics of the single-species cross-validation ($R^2_{cv,}$ SECV). The results of the ABC→D extrapolations are presented in complete detail in Table 8, but will be summarized here.

Across the 16 scenarios examined (four different traits × four different species extrapolations), there was a wide range in the quality of the extrapolation predictions from very good to poor. Comparing the four traits, the best extrapolations were found for the lignin traits (acid-insoluble lignin and S/(S+G)), and

Table 8. Fit statistics for extrapolations of a three-species NIR model to predict *Eucalyptus* wood chemistry for an independent fourth species of *Eucalyptus*.

	URO cross-validation					URO prediction with DUN, GLN, GRA					
Trait	Mean	SD	Range	R^2_{cv}	SECV	Mean	SD	Range	R^2_p	SEP	RPD_p
Acid-insoluble lignin	28.6	3.3	15.2	**0.87**	**1.17**	26.5	2.8	15.3	**0.90**	**1.08**	3.03
S/(S+G)	74.3	4.9	27.3	**0.87**	**1.75**	77.4	4.9	27.6	**0.45**	**3.98**	1.23
Glucose	46.2	2.8	14.4	**0.78**	**1.28**	42.7	5.0	25.2	**0.58**	**3.04**	0.92
Xylose	12.0	1.1	4.6	**0.83**	**0.45**	11.0	1.4	5.8	**0.68**	**0.78**	1.42
	DUN cross-validation					DUN prediction with URO, GLN, GRA					
Trait	Mean	SD	Range	R^2_{cv}	SECV	Mean	SD	Range	R^2_p	SEP	RPD_p
Acid-insoluble lignin	23.4	1.0	4.4	**0.64**	**0.58**	23.4	1.0	3.9	**0.40**	**0.83**	1.32
S/(S+G)	81.6	2.4	9.1	**0.56**	**1.60**	83.9	2.7	12.7	**0.64**	**1.66**	1.47
Glucose	46.4	2.6	10.1	**0.60**	**1.60**	45.4	3.3	15.6	**0.17**	**3.24**	0.80
Xylose	15.1	1.4	5.7	**0.69**	**0.80**	13.7	1.4	6.5	**0.41**	**1.22**	1.19
	GLN cross-validation					GLN prediction with URO, DUN, GRA					
Trait	Mean	SD	Range	R^2_{cv}	SECV	Mean	SD	Range	R^2_p	SEP	RPD_p
Acid-insoluble lignin	18.6	2.8	13.0	**0.96**	**0.53**	19.0	2.9	14.2	**0.93**	**0.80**	3.45
S/(S+G)	84.1	3.7	13.8	**0.92**	**1.05**	82.0	3.5	14.1	**0.90**	**1.18**	3.14
Glucose	46.9	4.2	20.6	**0.90**	**1.30**	54.2	6.0	30.8	**0.85**	**2.67**	1.55
Xylose	11.7	2.6	10.7	**0.96**	**0.52**	12.3	3.2	15.1	**0.93**	**1.01**	2.54
	GRA cross-validation					GRA prediction with URO, DUN, GLN					
Trait	Mean	SD	Range	R^2_{cv}	SECV	Mean	SD	Range	R^2_p	SEP	RPD_p
Acid-insoluble lignin	26.6	2.3	10.5	**0.78**	**1.04**	27.3	1.9	7.7	**0.80**	**1.04**	2.20
S/(S+G)	72.1	4.3	18.1	**0.93**	**1.12**	71.7	4.4	17.8	**0.97**	**0.81**	5.26
Glucose	47.6	3.3	15.5	**0.69**	**1.79**	47.5	2.8	11.6	**0.51**	**2.35**	1.39
Xylose	12.2	2.0	6.9	**0.68**	**1.07**	12.4	1.6	6.9	**0.67**	**1.13**	1.73

Note: Results of single-species wetlab values and cross-validation (left side of the table) are compared to trait predictions and fit statistics from a three-species NIR model to predict that target species (right side of the table). SD: standard deviation; SECV: standard errors of cross-validation; SEP: standard error of prediction; URO: *Eucalyptus urophylla*; DUN: *Eucalyptus dunnii*; GLN: *Eucalyptus globulus* + *E. nitens*; GRA: *Eucalyptus grandis*.

the poorest extrapolations were for glucose (Table 8). Comparing the species, the best extrapolations were found for *E. globulus–E. nitens*, with good to excellent predictions for all four traits. For this data set, across the four traits, the single-species R^2_{cv} ranged from 0.90 to 0.96, and the extrapolation R^2_p ranged from 0.85 to 0.93. In comparison, for the *E. dunnii* extrapolations, the predictions were moderately good for S/(S+G) ($R^2_p = 0.64$), but moderately poor for acid-insoluble lignin and xylose ($R^2_p = 0.40$ and 0.41), and very poor for glucose ($R^2_p = 0.17$).

For four species-trait scenarios, the extrapolation predictions equaled or exceeded the quality of the single-species cross-validations. For the *E. urophylla* extrapolation and the trait acid-insoluble lignin, the prediction fit was slightly better than the cross-validation fit, with $R^2_p = 0.90$ and SEP = 1.08, versus $R^2_{cv} = 0.87$ and SECV = 1.17. Similarly, for the *E. grandis* extrapolation for the trait acid-insoluble lignin, R^2_p was slightly greater than R^2_{cv} (0.80 vs. 0.78, respectively). Also for the *E. grandis* extrapolation for the trait S/(S+G), the prediction fit was slightly superior to the single-species cross-validation fit: $R^2_p = 0.97$ and SEP = 0.81, versus $R^2_{cv} = 0.93$ and SECV = 1.12. Finally, for *E. dunnii* and the trait S/(S+G), R^2_p was slightly better than R^2_{cv}, and SEP was slightly worse than SECV.

For at least five of the species-trait scenarios, the extrapolation predictions were excellent, relative to the single-species cross-validations: For the *E. globulus–E. nitens* data set, extrapolations for all four traits had R^2_p only slightly less than R^2_{cv} (Table 8), and for the *E. grandis*–xylose scenario, R^2_p 0.67 compared to $R^2_{cv} = 0.68$ for the single-species cross-validation. For the *E. urophylla*–xylose scenario, $R^2_p = 0.68$ from the extrapolation was substantially less than the single-species $R^2_{cv} = 0.83$, but it was still high enough to be useful in a breeding program.

For the remaining 6 out of 16 trait-species scenarios, the extrapolation fit statistics were clearly worse than the single-species cross-validation fit statistics, with average $R^2_p = 0.42$, while average $R^2_{cv} = 0.71$.

Improvement of global models

We wished to examine the question of how global calibration models could be improved in terms of prediction fit statistics for a "new" species not in the current models. Specifically, we wanted to examine the impact of incorporating a small number of wetlab observations from the "new" species into the calibration model. Four species-trait extrapolation scenarios were chosen:

- *Eucalyptus urophylla* extrapolation, glucose, and S/(S+G)
- *Eucalyptus dunnii* extrapolation, glucose, and acid-insoluble lignin

These four scenarios each represent cases where the prediction fit statistics were significantly worse than the single-species cross-validation fit statistics (Table 8). The two *E. urophylla* scenarios represent cases where there is a relatively wide range of variation in the wetlab values for the trait to be predicted (e.g., the largest range in S/(S+G) among all four species). In contrast, the two *E. dunnii* scenarios represent cases where there is more narrow range of variation in the wetlab values (for both traits, the smallest range of all four species).

The process of examining global model improvement will be illustrated for the scenario for *E. dunnii* and the trait glucose. The *E. dunnii* data set contained 50 wetlab observations, and these were divided randomly into five sets of 10 observations. One set of 10 *E. dunnii* observations was incorporated into the global calibration model containing 50 *E. urophylla* + 36 *E. grandis* + 50 *E. globulus–E. nitens* observations. The calibration model was developed as described above, and used to predict the remaining 40 *E. dunnii* glucose observations. This was repeated five times for each of the five sets of 10 observations, and average R^2_p and SEP were calculated. This same process was repeated for the other three species-trait scenarios. For a particular ABC→D extrapolation, the single-species R^2_{cv} and SECV for species D provide an upper baseline for comparison, while the R^2_p and SEP for the initial predictions using the ABC calibration are the lower baseline. These values can be compared to the R^2_p and SEP from the model developed with the 136 observations from the A, B, and C species plus 10 observations from the new species D.

The inclusion of 10 samples of a "new" species into a global calibration made a significant improvement in the prediction fit statistics (Figure 3). For *E. urophylla* S/(S+G), single-species $R^2_{cv} = 0.87$ and the extrapolation prediction $R^2_p = 0.45$. With 10 observations from *E. urophylla* included in the calibration, the mean R^2_p for the remaining *E. urophylla* observations was $R^2_p = 0.82$, with a substantial reduction in SEP. For *E. urophylla* glucose, single-species $R^2_{cv} = 0.78$ and the extrapolation prediction $R^2_p = 0.58$. Adding 10 observations from *E. urophylla* to the calibration improved the R^2_p to 0.71. Similarly, for *E. dunnii* acid-insoluble lignin, single-species $R^2_{cv} = 0.78$, $R^2_p = 0.58$, and the improved-R^2_p was 0.52. For the most challenging species-trait scenario, *E. dunnii*–glucose, the single-species R^2_{cv} was 0.60, and the extrapolation was very poor, $R^2_p = 0.17$). Adding 10 observations of *E. dunnii* into the calibration improved the R^2_p to 0.40, which is still only moderate, but much closer to the upper baseline of the single-species R^2_{cv} of 0.60.

Discussion

NIR models to predict wood chemistry traits

In this study, very good multi-species models were obtained for total lignin and acid-insoluble lignin content and lignin composition (S/G, S/(S+G)), and xylose and galactose content. Models for glucose, arabinose, and mannose were moderate to good. Zhou et al.[34] reported on models developed for *E. dunnii* with a 70-sample data set that also examined both lignin and sugar contents. The highest R^2 values were for lignin content ($R^2_{cv} = 0.87$), with all models for sugar content having slightly worse fit (cellulose $R^2_{cv} = 0.80$, glucose $R^2_{cv} = 0.84$, xylose $R^2_{cv} = 0.82$, arabinose $R^2_{cv} = 0.71$, mannose $R^2_{cv} = 0.63$). Perhaps there is a general trend that lignin-related traits can be predicted slightly more accurately and more precisely than glucose-related traits (e.g., pulp yield, cellulose composition). This agrees with findings for a mixed-species pine model for lignin and cellulose content.[24]

Other authors have reported good NIR calibrations for lignin content, S/G, cellulose content, and kraft pulp yield for a number of eucalypt species. For the trait S/G in *E. urophylla* (178 samples), Alves et al.[35] reported $R^2_{cv} = 0.97$ and $R^2_p = 0.96$, similar to but higher than the value obtained in this study ($R^2_{cv} = 0.90$, Table 3). Also for S/G in mixed species data set (120 samples total, *E. camaldulensis, E. grandis*, *E. urophylla*, *E. pellita*, and *E. alba*), Ramadevi et al.[36] reported $R^2_{cv} = 0.83$. This is very similar to the $R^2_{cv} = 0.86$ for the global model in the current study (Table 7). For the traits S/G and acid-insoluble lignin in a *E. urophylla* × *E. grandis* hybrid data set (193 samples), Baillères et al.[37] reported $R^2_{cv} = 0.87$ and 0.90, respectively, similar to the values reported for *E. urophylla* (Table 3) and *E. grandis* (Table 6) in this study. Regarding glucose-related traits, Downes et al.[38] reported good models for kraft pulp yield and cellulose using a very large mixed-species data set (44 species and over 1300 samples): for pulp yield, $R^2_{cv} = 0.90$ and $R^2_p = 0.80$, while for cellulose content $R^2_{cv} = 0.90$ and $R^2_p = 0.85$. For cellulose content in *E. nitens*, Schimleck et al.[2] reported on three single-site NIR models (60 trees for calibration and 108–128 trees for validation) with average $R^2_{cal} = 0.74$ and $R^2_p = 0.67$.

S/G versus S/(S+G)

S/G ratio measures the relative abundance of syringyl and guaiacyl lignin monomers, and it is an important

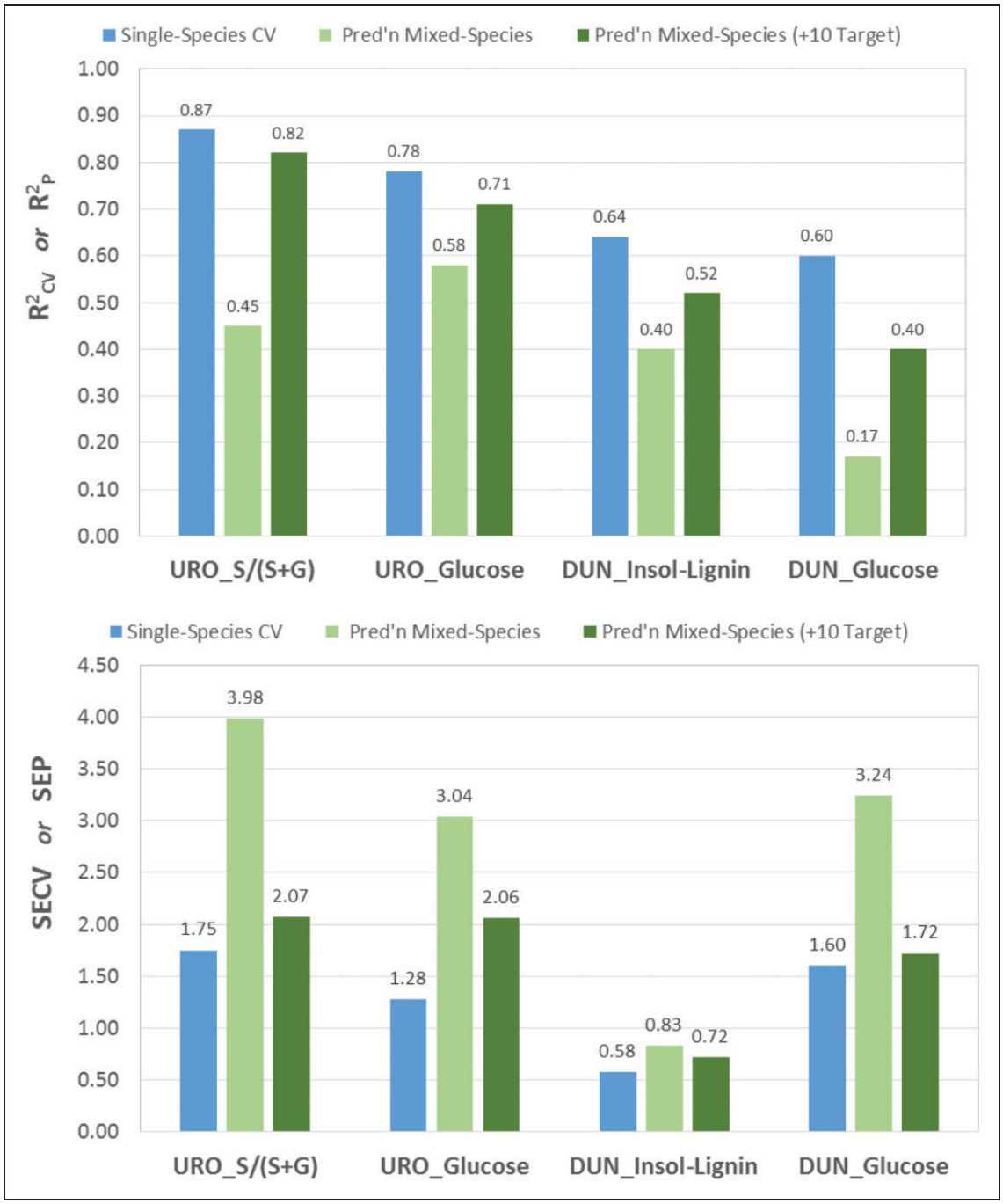

Figure 3. Prediction fit statistics for single-species near infrared (NIR) calibration, a mixed-species NIR calibration (minus the target species), and a mixed-species NIR calibration (+10 samples of the target species). Four species-trait scenarios were selected for study, all of which had poor mixed-species predictions compared to the single-species cross-validations: two traits from *Eucalyptus urophylla* with a large range of the wetlab value to be predicted, and two traits from *Eucalyptus dunnii* with a small range of the wetlab value to be predicted.

and well known trait in the pulping literature.[33] In this study, an alternate formulation of the lignin composition was examined, S/(S+G). In three of the four single-species data sets, the NIR models had better fit statistics for S/(S+G) than for S/G. This was true also for the global model, where the model for S/(S+G) had $R^2_{cv} = 0.95$ and $RPD_{cv} = 4.69$, versus $R^2_{cv} = 0.86$ and $RPD_{cv} = 2.71$ for S/G.

Analyses of genetic trials of *E. dunnii* and *E. grandis* indicate that the S/(S+G) variable has higher heritability than the S/G variable,[39,40] and therefore breeders may wish to use S/(S+G) for NIR models and genetic rankings, then convert back to S/G for use of the data by pulp researchers. The conversion between the two variable formulations is quite simple:

For S/G = X

$$\frac{X}{X+1} = S/(S+G)$$

For S/(S+G) = Y

$$\frac{Y}{(1-Y)} = S/G$$

Glucose, lignin, and pulp yield

In this study, the NIR models were developed for sugar and lignin content and composition. For many pulp and paper producers, the most important breeding objective would actually be pulp yield, and in fact, much work has been done to develop NIR models for pulp yield in eucalypts.[8,9,38,41–44] Measuring pulp yield is typically slow and expensive, and requires large sample sizes, typically from 1 to 5 kg of oven-dried woodchips. Furthermore, pulp yield estimates are specific to the particular cooking conditions (time, temperature, chemicals) and often to a specific kappa number. For a given organization, it may be possible to develop laboratory conditions to mimic the large-scale mill conditions that apply. However, in this case, the NIR models were being developed for a large number of mills, each with their own specific processes. Relative to assessing pulp yield, assessing wood chemical composition is relatively inexpensive, and it can be done with smaller amounts of wood. Many of these traits have been shown to highly correlated to pulp yield.[38,45,46] For example, in a multiple-species data set, Downes et al.[38] found a correlation of $R^2 = 0.85$ between laboratory pulp yield and laboratory cellulose content, and a similar correlation between NIR predictions of pulp yield and cellulose content ($R^2 = 0.88$). Kien et al.[46] found a similar correlation ($R^2 = 0.83$) between pulp yield and cellulose content for a single-species data set of *E. urophylla*. In an *E. camaldulensis* data set, Ramadevi et al.[36] found a strong relationship between S/G ratio and kraft pulp yield ($R^2 = 0.71$), and between S/G and alkali consumption ($R^2 = 0.91$). It seems reasonable that pulp researchers could identify which of the sugar and lignin traits that can be predicted by the current models have the strongest impact on pulp yield (and pulping costs) in their own organization.

Utility of the models

Sandak et al.[47] provide some guidelines to assess the quality and utility of NIR models for use in forestry. In general, models should have high R^2, low SECV and SEP, a low bias, and a slope near one. NIR models can be employed for the purposes of quality control in a laboratory or production facility, in which case the standards for a "good" model may be quite high, for example, $R^2 \geq 0.90$ and RPD ≥ 5.0. Tree breeders are primarily interested in accurate ranks or comparison between genotypes (a "screening" application), so the criteria for a "good" model may be lower, for example, $R^2 \geq 0.80$ and RPD ≥ 2.0. The global models in this study (Table 7) meet or exceed the latter criteria for "good" screening models for 7 of the 10 variables (total lignin, acid-insoluble lignin, S/G, S/(S+G), xylose, galactose, and mannose). The global models for glucose ($R^2_{cv} = 0.74$ and $RPD_{cv} = 1.95$) and arabinose ($R^2_{cv} = 0.72$ and $RPD_{cv} = 1.91$) are just below the thresholds for "good" screening models.

The assessment of NIR model quality and utility is somewhat subjective,[48] and researchers must examine the fit statistics of the available NIR models to determine if they can be used to meet their specific objective. To evaluate the current models for the purposes of tree breeding, it is important to remember that the R^2, RPD and standard errors in this study apply to the measurement of a single 4g sample of woodmeal. Breeders are ultimately interested ranking genotypes for their underlying breeding value or clonal genetic value for a particular trait of interest, and this is done by sampling multiple offspring from a given parent or full-sib family to provide estimates of parental or family genetic value, or of multiple ramets of a given clone to estimate clonal genetic value. Thus, a selection decision would not be made on the basis of a single sample NIR prediction, rather, multiple observations from all related genotypes are considered, typically through the use of a mixed-model genetic analysis to calculate best linear unbiased predictions (BLUPs) of genetic values. Breeders also use genetic parameter estimates to evaluate the precision of those BLUPs.

Relevant to the current study are the results reported for *Eucalyptus nitens* by Schimleck et al.[2] and *E. pellita* by Hung et al.[5] for NIR-predicted traits using models with fit statistics comparable to the current models (i.e., $R^2 = 0.80$ to 0.85). For *E. nitens*, the authors found high heritability ($h^2 = 0.60$–0.70) and low levels of genotype × environment interaction (Type G genetic correlation,[49] $r_{Bg} = 0.85$–0.95) for both laboratory and NIR-determined cellulose content. Similarly, estimated genetic parameters for kraft pulp yield for *E. pellita*[5] (predicted using NIR models from Meder et al.[50]) indicate an average single-site heritability of $h^2 = 0.37$ and a Type B genetic correlation $r_{Bg} = 0.85$. If these genetic parameter are typical of chemical traits for most eucalypt species, sampling 10 trees per family on each of two sites would result in family heritability between 0.80 and 0.90, which would produce very precise rankings of parental genotypes. Similarly, for clonal selection, breeders would typically assess a number of ramets from the same clone, and clonal heritabilities of $H^2 \geq 0.90$ would be likely.

Age variation among samples

In many eucalypt breeding programs, growth is measured at half-rotation age, and breeders will generally measure wood traits at that same age, or 1 or 2 years later. Rotation ages for tropical and subtropical eucalypts grown for pulp are often around 8 years, while for temperate eucalypts, pulp rotations might be as long

as 12 years. Under these scenarios, wood samples would often be taken at ages ranging from 4 to 8 years old. The models in this study included samples ranging in age from 4 years (all of the *E. dunnii* samples) to 13 years (for some of the *E. urophylla* samples) to 25 years old (for some of the *E. globulus*-*E. nitens* samples. These age differences should be kept in mind when comparing among species for chemical properties in this study. However, since the global models cover a very broad range of values for all traits, and the precision of the predictions seem very uniform across the range, we do not believe the age differences among the calibration samples in this study will have any impact on the precision of rankings (i.e., comparisons of genotypes) in future studies.

Utility for extrapolation to new species

While the current global model could be used to make predictions for new data sets of the species included in this study (*E. urophylla*, *E. dunnii*, *E. grandis*, *E. globulus*, and *E. nitens*), we may wish to use these models on other eucalypt species not included in the calibration data set. In this situation, a breeder would expect that, in general, extrapolation predictions would be at least moderately useful for most species and most traits. A better approach would be to incrementally improve the model by adding a few samples of the new species into the calibration data set. It has been demonstrated that the addition of just a few samples from a new data set into the calibration model can greatly improve the predictions for the remainder of that data set,[51] and the current results suggest that the addition of 10 samples of the ''new'' species into the calibration data set makes substantial improvement in the prediction fit statistics. In this study, the 10 samples were randomly selected, but in fact the 50 samples of a given species (from which the 10 random samples were chosen) had already been preselected to cover a range of chemical and spectral variation. Thus, it might be wise to screen a few hundred samples of the new species, and use the current NIR models to ensure that the 10 samples selected for wetlab analysis and inclusion in the calibration equation do indeed cover some of the range of variation for important wood properties.

Acknowledgements

The authors would like to acknowledge the contributions of Smurfit Kappa Colombia, Smurfit Kappa Venezuela, Weyerhaeuser Company, CMPC Forestal Mininco, Sappi Forests, Mondi Forests, and Arauco Argentina for collection of wood samples used in this project. Thanks also to all Camcore members for their continuing support of our work in gene conservation and tree breeding, as well as to the Department of Forestry & Environmental Resources at NC State University, and to the Department of Wood Science at the University of British Columbia. Finally, many thanks to Juan Pedro Posse, Martha Salas, and Romeo Jump for all their hard work in the field and the lab.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. So CL, Via BK, Groom LH, et al. Near infrared (NIR) spectroscopy in the forest products industry. *For Prod J* 2004; 54: 8–18.
2. Schimleck LR, Kube PD and Raymond CA. Genetic improvement of kraft pulp yield in *Eucalyptus nitens* using cellulose content determined by near infrared spectroscopy. *Can J For Res* 2004; 34: 2363–2370.
3. Sykes R, Li B, Isik F, et al. Genetic variation and genotype by environment interactions of juvenile wood chemical properties in *Pinus taeda* L. *Ann For Sci* 2006; 63: 897–904.
4. Stackpole DJ, Vaillancourt RE, Downes GM, et al. Genetic control of kraft pulp yield in *Eucalyptus globulus*. *Can J For Res* 2010; 40: 917–927.
5. Hung TD, Brawner JT, Meder R, et al. Estimates of genetic parameters for growth and wood properties in *Eucalyptus pellita* F. Muell. to support tree breeding in Vietnam. *Ann For Sci* 2015; 72: 205–217.
6. Wright JA, Birkett MD and Gambino MJT. Prediction of pulp yield and cellulose content from wood samples using near infrared reflectance spectroscopy. *Tappi J* 1990; 73: 164–166.
7. Garbutt DCF, Donkin MJ and Meyer JH. Near infra-red reflectance analysis of cellulose and lignin in wood. *Pap South Afr* 1992; April: 45–48.
8. Michell AJ. Pulpwood quality estimation by near-infrared spectroscopic measurements on eucalypt woods. *Appita J* 1995; 48: 425.
9. Michell AJ and Schimleck LR. Developing a method for the rapid assessment of pulp yield of plantation eucalypt trees beyond the year 2000. *Appita J* 1998; 51: 428–432.
10. Raymond CA and Schimleck LR. Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in *Eucalyptus globulus*. *Can J For Res* 2002; 32: 170–176.
11. Schimleck L, Evans R and Ilic J. Application of near infrared spectroscopy to a diverse range of species demonstrating wide density and stiffness variation. *IAWA J* 2001; 22: 415–429.
12. Schimleck LR and Evans R. Estimation of microfibril angle of increment cores by near infrared spectroscopy. *IAWA J* 2002; 23: 225–234.
13. Schimleck LR and Evans R. Estimation of wood stiffness of increment cores by near infrared spectroscopy: the development and application of calibrations based on selected cores. *IAWA J* 2002; 23: 217–224.

14. Kelley SS, Rials TG, Snell R, et al. Use of near infrared spectroscopy to measure the chemical and mechanical properties of solid wood. *Wood Sci Technol* 2004; 38: 257–276.
15. Kelley SS, Rials TG, Groom LR, et al. Use of near infrared spectroscopy to predict the mechanical properties of six softwoods. *Holzforschung* 2004; 58: 252–260.
16. Jones PD, Schimleck LR, Peter GF, et al. Nondestructive estimation of *Pinus taeda* L. wood properties for samples from a wide range of sites in Georgia. *Can J For Res* 2005; 35: 85.
17. Jones PD, Schimleck LR, Daniels RF, et al. Comparison of *Pinus taeda* L. whole-tree wood property calibrations using diffuse reflectance near infrared spectra obtained using a variety of sampling options. *Wood Sci Technol* 2008; 42: 385–400.
18. Schimleck LR. Near infrared spectroscopy: a rapid, non-destructive method for measuring wood properties and its application to tree breeding. *New Zeal J For Sci* 2008; 38: 14.
19. Downes GM, Touza M, Harwood C, et al. NIR detection of non-recoverable collapse in sawn boards of *Eucalyptus globulus*. *Eur J Wood Prod* 2014; 72: 563–570.
20. Borrahlo NMG, Cotterill PP and Kanowski PJ. Breeding objectives for pulp production of *Eucalyptus globulus* under different industrial cost structures. *Can J For Res* 1993; 23: 648–656.
21. Schmidt EA. A practical model relating kraft pulping costs to hardwood chemical properties and morphology. *Appita J* 2005; 58: 218–224.
22. Lopez J, Gomide JL and Phillips R. Influence of eucalyptus wood properties on the financial performance of a modeled Brazilian pulp mill. *O Papel* 2009; 70: 53–71.
23. Hodge GR and Woodbridge WC. Use of near infrared spectroscopy to predict lignin content in tropical and subtropical pines. *J Near Infrared Spectrosc* 2004; 12: 381.
24. Hodge GR and Woodbridge WC. Global near infrared models to predict lignin and cellulose content of pine wood. *J Near Infrared Spectrosc* 2010; 18: 367–380.
25. Coleman HD, Canam T, Kang KY, et al. Over-expression of UDP-glucose pyrophosphorylase in hybrid poplar affects carbon allocation. *J Exp Bot* 2007; 58: 4257–4268.
26. Robinson AR and Mansfield SD. Rapid analysis of poplar lignin monomer composition by a streamlined thioacidolysis procedure and near-infrared reflectance-based prediction modeling. *Plant J* 2009; 58: 706–714.
27. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, www.R-project.org/ (2016) (accessed 15 May 2016).
28. Rinnan A, van den Berg F and Engelsen S. Review of the most common pre-processing techniques for near-infrared spectra. *Trends Anal Chem* 2009; 28: 10.
29. Mevik BH and Wehrens R. The pls package: principal component and partial least squares regression in R. *J Stat Softw* 2007; 18: 1–22.
30. Stevens A and Ramirez-Lopez L. An introduction to the prospectr package. R package Vignette R package version 0.1.3. 2013. Available at: ftp://200.236.31.2/CRAN/web/packages/prospectr/vignettes/prospectr-intro.pdf.
31. Breunig M, Hans-Peter K, Raymond T, et al. LOF: identifying density-based local outliers. In ACM Sigmod Record 2000; 29: 93–104.
32. Torgo L. Data mining with R, learning with case studies. Chapman and Hall/CRC. www.dcc.fc.up.pt/∼ltorgo/DataMiningWithR (2010) (accessed 15 May 2016).
33. Stewart JJ, Kadla JF and Mansfield SD. The influence of lignin chemistry and ultrastructure on the pulping efficiency of clonal aspen (*Populus tremuloides* Michx.). *Holzforschung* 2006; 60: 111–122.
34. Zhou C, Jiang W, Via BK, et al. Monitoring the chemistry and monosaccharide ratio of *Eucalyptus dunnii* wood by near infrared spectroscopy. *J Near Infrared Spectrosc* 2016; 24: 537–548.
35. Alves A, Simões R, Stackpole DJ, et al. Determination of the syringyl/guaiacyl ratio of *Eucalyptus globulus* wood lignin by near infrared-based partial least squares regression models using analytical pyrolysis as the reference method. *J Near Infrared Spectrosc* 2011; 19: 343–348.
36. Ramadevi P, Hegde DV, Varghese M, et al. Evaluation of lignin syringyl/guaiacyl ratio in *Eucalyptus camaldulensis* across three diverse sites based on near infrared spectroscopic calibration modelling with five *Eucalyptus* species and its impact on kraft pulp yield. *J Near Infrared Spectrosc* 2016; 24: 529–536.
37. Baillères H, Davrieux F and Ham-Pichavant F. Near infrared analysis as a tool for rapid screening of some major wood characteristics in a eucalyptus breeding program. *Ann For Sci* 2002; 59: 479–490.
38. Downes GM, Meder R, Bond H, et al. Measurement of cellulose content, kraft pulp yield and basic density in eucalypt woodmeal using multisite and multispecies near infra-red spectroscopic calibrations. *South For* 2011; 73: 181–186.
39. Salas M. Clonal variation of *Eucalyptus grandis* W. Hill ex Maiden in Colombia. MS Thesis, North Carolina State University, Raleigh NC, USA, 2018.
40. Posse JP. Genetic parameters for growth and wood traits in a *Eucalyptus dunnii* Maiden population planted in Uruguay. PhD dissertation, North Carolina State University, Raleigh, NC, USA, 2018.
41. Greaves BL, Borrahlo NMG and Raymond CA. Breeding objective for plantation eucalypts grown for production of kraft pulp. *For Sci* 1997; 43: 465–472.
42. Schimleck LR and Michell AJ. Determination of within-tree variation of kraft pulp yield using near-infrared spectroscopy. *Tappi J* 1998; 81: 229–223.
43. Raymond CA, Schimleck LR, Muneri A, et al. Non-destructive sampling of *Eucalyptus globulus* and *E. nitens* for wood properties. III. Predicted pulp yield using near infrared reflectance analysis. *Wood Sci Technol* 2001; 35: 203–215.
44. Downes GM, Catela F and Meder R. Developing and evaluating a multisite and multispecies NIR calibration for the prediction of kraft pulp yield in eucalypts. *South For* 2009; 71: 155.
45. Kube PD and Raymond CA. Prediction of whole-tree basic density and pulp yield using wood core samples in *Eucalyptus nitens*. *Appita J* 2002; 55: 43–48.

46. Kien ND, Quang TH, Jansson G, et al. Cellulose content as a selection trait in breeding for kraft pulp yield in *Eucalyptus urophylla*. *Ann For Sci* 2009; 66: 711.
47. Sandak J, Sandak A and Meder R. Assessing trees, wood and derived products with near infrared spectroscopy: hints and tips. *J Near Infrared Spectrosc* 2016; 24: 485–505.
48. Esbensen KH, Geladi P and Larsen A. The RPD myth . . . *NIR News*, 24–28 August 2014; 25: 24–28.
49. Burdon RD. Genetic correlation as a concept for studying genotype–environment interaction in forest tree breeding. *Silvae Genet* 1977; 26: 168–175.
50. Meder R, Marston D, Ebdon N, et al. Spatially-resolved radial scanning of tree increment cores for near infrared prediction of microfibril angle and chemical composition. *J Near Infrared Spec* 2011; 18: 499–505.
51. Schimleck LR, Kube PD, Raymond CA, et al. Extending near infrared reflectance (NIR) pulp yield calibrations to new sites and species. *J Wood Chem Technol* 2006; 26: 299–311.