Bioenerg. Res. (2015) 8:953–963
DOI 10.1007/s12155-015-9578-1

High-Throughput Prediction of *Acacia* and Eucalypt Lignin Syringyl/Guaiacyl Content Using FT-Raman Spectroscopy and Partial Least Squares Modeling

Jason S. Lupoi · Adam Healey · Seema Singh · Robert Sykes · Mark Davis · David J. Lee · Merv Shepherd · Blake A. Simmons · Robert J. Henry

Published online: 16 January 2015

Abstract High-throughput techniques are necessary to efficiently screen potential lignocellulosic feedstocks for the production of renewable fuels, chemicals, and bio-based materials, thereby reducing experimental time and expense while supplanting tedious, destructive methods. The ratio of lignin syringyl (S) to guaiacyl (G) monomers has been routinely quantified as a way to probe biomass recalcitrance. Mid-infrared and Raman spectroscopy have been demonstrated to produce robust partial least squares models for the prediction of lignin S/G ratios in a diverse group of *Acacia* and eucalypt trees. The most accurate Raman model has now been used to predict the S/G ratio from 269 unknown *Acacia* and eucalypt feedstocks. This study demonstrates the application of a partial least squares model composed of Raman spectral data and lignin S/G ratios measured using pyrolysis/molecular beam mass spectrometry (pyMBMS) for the prediction of S/G ratios in an unknown data set. The predicted S/G ratios calculated by the model were averaged according to plant species, and the means were not found to differ from the pyMBMS ratios when evaluating the mean values of each method within the 95 % confidence interval. Pairwise comparisons within each data set were employed to assess statistical differences between each biomass species. While some pairwise appraisals failed to differentiate between species, *Acacias*, in both data sets, clearly display significant differences in their S/G composition which distinguish them from

J. S. Lupoi (✉) · A. Healey · B. A. Simmons · R. J. Henry
Queensland Alliance for Agriculture and Food Innovation, University of Queensland, 306 Carmody Road, St. Lucia, Queensland 4072, Australia
e-mail: jslupoi@lbl.gov

A. Healey
e-mail: adam.healey@uq.net.au

B. A. Simmons
e-mail: basimmons@lbl.gov

R. J. Henry
e-mail: robert.henry@uq.edu.au

J. S. Lupoi · S. Singh · B. A. Simmons
Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, 5885 Hollis Street, Emeryville, CA 94608, USA

S. Singh
e-mail: seesing@sandia.gov

S. Singh · B. A. Simmons
Biological and Materials Science Center, Sandia National Laboratories, 7011 East Avenue, Livermore, CA 94551, USA

R. Sykes · M. Davis
BioEnergy Science Center, Oak Ridge National Laboratory, 1 Bethel Valley Rd, Oak Ridge, TN 37831, USA

R. Sykes
e-mail: Robert.Sykes@nrel.gov

M. Davis
e-mail: Mark.Davis@nrel.gov

R. Sykes · M. Davis
National Bioenergy Center, National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401, USA

D. J. Lee
Forest Industries Research Centre, University of the Sunshine Coast and Queensland Department of Agriculture, Fisheries and Forestry, Locked Bag 4, Maroochydore DC, Queensland 4558, Australia
e-mail: dlee@usc.edu.au

M. Shepherd
Southern Cross Plant Science, Southern Cross University, Military Road, East Lismore, NSW 2480, Australia
e-mail: mervyn.shepherd@scu.edu.au

eucalypts. This research shows the power of using Raman spectroscopy to supplant tedious, destructive methods for the evaluation of the lignin S/G ratio of diverse plant biomass materials.

Keywords Lignocellulose · Raman spectroscopy · High-throughput · Multivariate analysis · Lignin S/G · *Eucalyptus* · *Corymbia* · *Acacia*

Abbreviations

ANOVA	Analysis of Variance
CCC	*Corymbia citriodora* subspecies *citriodora*
CCV	*Corymbia citriodora* subspecies *variegata*
CV	Cross-validation
EMSC	Extended multiplicative scatter correction
MIR	Mid-infrared spectroscopy
MVA	Multivariate analysis
NIR	Near-infrared spectroscopy
PLS	Partial least squares regression
pyMBMS	Pyrolysis molecular beam mass spectrometry
r	Coefficient of correlation for validation set
R^2Cal	Coefficient of determination for calibration set
R^2CV	Coefficient of determination for full cross-validation
RMSECV	Root mean standard Error of cross-validation
RMSEP	Root mean standard error of prediction
S/G	Syringyl to guaiacyl ratio

Background

The ratio of lignin syringyl (S) to guaiacyl (G) moieties has been characteristically quantified as one method to evaluate biomass recalcitrance [1–7]. While higher S/G ratios have resulted in increased lignin pulping reactivity [2, 7], a clear trend linking S/G ratio to the enzymatic degradation of plant cell walls has not been established. Some reports have indicated high S/G to correlate with increased sugar release [4, 6], while other studies have concluded that ratio reductions are optimal [1]. Regardless of the exact effects S/G ratios have on the saccharification of biomass, this parameter has proven to be important in developing a better understanding of lignin structure and degradation.

In a previous study, mid-infrared (MIR), near-infrared (NIR), and FT-Raman spectra were coupled with lignin S/G ratios obtained using pyrolysis/molecular beam mass spectrometry (pyMBMS) for the construction of multivariate analysis (MVA) models [8]. Various iterations were performed to determine the spectral processing techniques that provided the most robust calibration models. The models were vigorously assessed using statistical metrics including root mean standard error (RMSE) or Scree plots for determining the appropriate number of factors, and the RMSE values and calculated coefficients of correlation (r) and determination (R^2) after using a full cross-validation and a 50 sample validation data set. These parameters illustrated increased accuracy from using MIR and Raman spectroscopy for the development of models capable of predicting lignin S/G ratios. For example, MIR and Raman spectroscopy partial least squares (PLS) models resulted in a root mean standard error of prediction (RMSEP) of 0.13 to 0.15 and 0.13 to 0.16, respectively, while the RMSEP measured using NIR spectra was slightly more erroneous (0.18 to 0.21). While the construction of these models illustrated the potential of MVA and vibrational spectroscopy to screen biomass based on lignin S/G ratios, the complete evaluation of the MVA models required the prediction of the S/G ratios for an unknown data set. The execution of this step is integral for the determination of a model's practicality for assessing future samples.

The motivations for conducting this study were twofold. One was to evaluate the most robust FT-Raman model for the prediction of lignin S/G ratios from 269 trees from three genera (*Acacia*, *Corymbia*, and *Eucalyptus*) encompassing 17 diverse species. The analysis of the S/G predictions calculated from this model displayed an accuracy correlative to the pyMBMS reference results, highlighting the use of non-destructive Raman spectroscopy to reduce experimental time and expense. The second rationale behind this study was to determine which plants (whether measured directly using pyMBMS or predicted using the Raman model) had the lowest and highest lignin S/G ratios. Evaluations between species were conducted using pairwise comparisons within measured and predicted S/G data sets to ensure that any statistical differences found within the modeled data could be verified against a widely accepted chemical analysis technique.

Methods

Wood Samples

The sampling techniques used for the acquisition of the wood samples used in this study have been described in a previous manuscript [8]. In addition to the 245 samples used for the construction of the PLS model, 269 diverse *Acacia* and eucalypt samples comprised the unknown sample matrix.

Fourier Transform Raman Spectroscopy

The FT-Raman spectral collection parameters have been described in a previous manuscript [8].

Pyrolysis/Molecular Beam Mass Spectrometry

The pyMBMS instrumental and spectral processing methodologies have been previously described [9].

Multivariate Analysis

All modeling was conducted using the Unscrambler X software package (Camo, Inc., Oslo, Norway). Samples used for calibrating and validating the original models were united, creating a 245 sample calibration matrix composed of *Acacia*, *Corymbia*, and *Eucalyptus* trees. The model was evaluated employing a full cross-validation (CV) before predicting the S/G ratios of 269 unknown samples. Overfitting of the data was gauged by analyzing the RMSE or Scree plot and by studying the effects of using non-optimal numbers of factors on the predictive capacity of the model. The most influential variables utilized for model construction were identified from the regression coefficients plot. The model was recalculated using solely these vibrational modes, thereby diminishing spectral noise and subsequently increasing the calibration and prediction accuracy.

Statistical Analysis

The predicted Raman and measured pyMBMS lignin S/G ratios were compared to assess whether there were statistical differences between the samples. A non-parametric Kruskal-Wallis test (χ^2=155.99, p value<2.2×10^{-16}) was used to evaluate differences between taxa for the Raman S/G predicted values [10]. Post hoc comparisons between taxa were carried out using Mann-Whitney U tests with a Holm adjustment for multiple comparisons [11]. Pyrolysis S/G ratios were analyzed with a standard one-factor analysis of variance (ANOVA) (F(18,182)=16.82, p value<2×10^{-16}). Tukey's honestly significant differences (HSD) protocol was performed as a post hoc comparison between taxa. To determine if there were any significant differences between the predicted and reference S/G values for each species (Table 2), a Mann-Whitney U test was conducted. Analyses were performed using R Studio (R Studio, version 3.0.2, Boston, MA, USA).

Results and Discussion

In a previous study, PLS models employing first derivative Raman spectra and an extended multiplicative scatter correction (EMSC) provided the highest accuracy and robustness [8]. These models, developed to predict lignin S/G ratios in *Acacias* and eucalypts, contained randomly generated calibration and validation sets encompassing 195 and 50 samples, respectively. Figure 1 provides a comparison between first derivative, EMSC-transformed Raman spectra of *Acacia microbotrya* (black) and *Eucalyptus globulus* subspecies *globulus* (red) trees. These two specific samples represent the extremes encompassed in the pyMBMS measurements, and the Raman spectra were analyzed to attempt the elucidation the spectral differences correlative to this range. While spectral differences near 781, 1037, 1150, 1259, 1332, 1603, and 1627 cm^{-1} can be identified between the two samples, lignin and its derivatives have vibrational modes at these locations corresponding to both S and G moieties, as well as lignin skeletal and phenyl ring vibrations, making the assignment of these bands challenging (see Table 1). Cellulose can further complicate the assignment of some of these bands, as it has known Raman peaks near 1119 and 1150 cm^{-1} [12]. The complexity of the Raman spectra of heterogeneous biomass samples obscures sample-to-sample qualitative comparisons. Figure 1 also illustrates the deficiency of striking spectral disparities, thereby highlighting the proficiency of employing MVA to hone in on previously obfuscated sample variance.

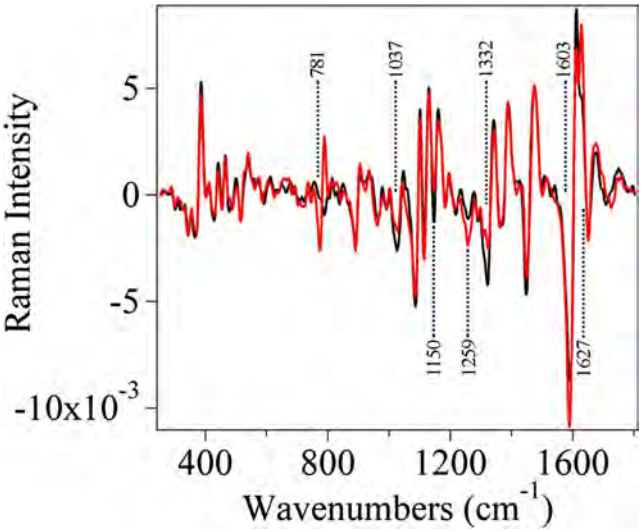

Fig. 1 Comparison of the first derivative+EMSC Raman spectra of low S/G *Acacia microbotrya* (black spectrum, S/G=1.2) and higher S/G *Eucalyptus globulus* subspecies *maidenii* (red spectrum, S/G=3.0), as measured by pyrolysis/molecular beam mass spectrometry. The *x*-axis is in wavenumbers, while the *y*-axis represents the Raman intensity (*EMSC* extended multiplicative scatter correction, *S/G* syringyl to guaiacyl ratio). Vertical dashed lines have been added to illustrate spectral differences

To predict the S/G ratio for the unknown data set consisting of 269 *Acacia* and eucalypt samples, the validation set was combined into the calibration matrix, providing a new, more vigorous model. This resulted in a five-factor calibration model with a calibration R^2 of 0.848 and a validation R^2 of 0.824, following a full CV, generating an RMSECV of 0.13. It should be noted that these five factors do not represent individual biomass constituents but rather represent sources of spectral variance being drawn out by the model. The score plot produced by the model is shown in Fig. 2 and represents the

Table 1 Raman vibrational modes identified from regression coefficient plot and spectral assignments corresponding to lignin and/or lignin monomers

Vibrational mode from regression coefficient plot	S/G/H vibrational mode and spectral assignment(s)
322–376	369 (S), 357, 370 (G) [22]
378–403	370–399 (S) [23]
453–476	Skeletal deformation of lignin [24] 468 (G, H) [22]
527–536 574–577 592–598	529, 564, 582 (S), 541, 559, 590 (G) [22]
717–725	711 (S) [25] 712 (G), 701 (H) [22]
733–739	741 (S) [25] 741 (H) [22]
748–771	761 (G) [25]
773–800	781–820 (S) [23] 784 (G) [25] 793 (G) [22]
773–800 835–839	819–864 (H) [23] 810 (S) [25] 799 (S), 823 (H) [22]
889–930	920 (G)[25] 907 (S), 921 (G) [22]
1001–1035 1041–1074	1024 (G) [25] 1043 (S), 1036 (G) [22]
1090–1093 1099–1137	1108 (S), 1124 (G), 1094 (H) [25] 1116 (S), 1122 (G), 1105 (H) [22]
1099–1137 1138–1155 1188–1192	1154 (S), 1158 (G), 1168 (H), 1170 (Lignin) [26] 1138–1160 (S), 1162–1188 (G), 1163–1179 (H) [23] 1148 (S), 1186 (G), 1164 (H) [25] 1152, 1187 (S), 1155, 1186 (G), 1173, 1199 (H) [22]
1215–1219	1200 (H) [26] 1213–1218 (H) [23] 1228 (S), 1215 (H) [25] 1214, 1241 (S), 1208, 1241 (G), 1216 (H) [22]
1232–1263 1281–1324 1329–1348	1337 (S), 1263 (H), 1270 (Lignin) [26] 1262–1275 (G), 1318–1332, 1331–1338 (S), 1286–1299 (H) [23] 1331 (S), 1270–1285 (G), 1338 H [25] 1331 (S), 1272, 1288 (G), 1298, 1331 (H) [22]
1433–1450	1454–1460 (S), 1452–1465 (G), 1452–1459 (H) [23] 1452 (S), 1455 (G), 1455 (H) [22]
1488	1488 (H) [25]
1583–1594	1594 (S), 1589, 1588, 1606 (H), 1591 [26] 1588 (S) [25]
1624–1629	1634 (S), 1633 (G), 1632 (H), 1634 (Lignin) [26]
1649–1657	Coniferyl (G) and sinapyl (S) alcohol [24]
1662–1708	1661–1664 (G, coniferyl alcohol, coniferaldehyde) [24] 1662–1695 C=O conjugated monomers [25]

S, G, or H refers to syringyl, guaiacyl, or *p*-coumaryl lignin, respectively

classification of samples based on second and third principal components (PCs) or factors. In this plot, the blue squares, red circles, and green triangles represent the *Acacia*, *Corymbia*, and *Eucalyptus* genera, respectively. Three distinct groups can be identified, although the *Corymbia* and *Eucalyptus* groups have some overlap. This is expected, as the pyMBMS measured lignin S/G ratios of these genera are more comparable, juxtaposed to the *Acacia* samples. Also, as anticipated, *Corymbia* and *Eucalyptus* trees measured to contain lower S/G ratios using pyMBMS, such as all *Eucalyptus crebra* samples (S/G=1.6±0.4) or a *Corymbia torelliana* heartwood sample (S/G=1.5), were closest to the *Acacias* (*A. microbotrya* S/G=1.3±0.1, *Acacia saligna* S/G=1.7±0.2). Since the bottom left quadrant contains the plants with lower S/G ratios, on average, the top right quadrant was expected to reveal an opposite trend. Indeed, samples located at the farthest corner of this quadrant show increased pyMBMS lignin S/G ratios (*Corymbia citriodora* subspecies *variegate* (CCV)=2.6, *Eucalyptus cladocalyx*=2.6, *Eucalyptus dunnii*=2.8, *E. globulus*=2.8, and *Eucalyptus moluccana*=2.5). Given the lack of statistical differences between many of these higher S/G samples (see Table 3), however, the classifications are much less defined, contrasted to the *Acacia* cluster. The loading plots for the first three factors are provided in Fig. 3a–c. Loadings plots represent which vibrational modes are important in composing a specific factor. The vibrational modes of polymeric lignin and its individual phenylpropanoid constituents have similar spectral signatures, complicating the analysis of the loading plots. While specific peaks indicative of G, S, and polymeric lignin can be identified in the loading plots of the first three factors, there is no discernible trend aligning a specific factor with an unambiguous lignin moiety. Rather, each of the loadings contributes G, S, and lignin spectral features to the overall classification. This can be exemplified in Fig. 2, where, as previously discussed, the lower left quadrant contains samples with the lowest S/G ratios, while, in general,

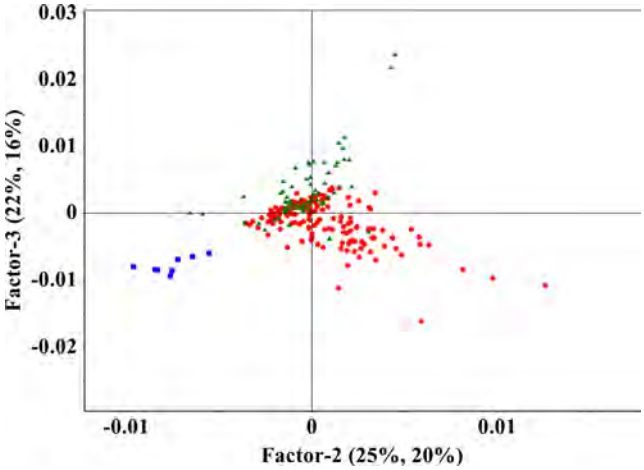

Fig. 2 Scores plot showing the classification of samples by genus. The *x*- and *y*-axes represent the second and third factors, respectively. The *blue squares* represent the genus *Acacia*; the *red circles* depict the genus *Corymbia*; and the *green triangles* show the genus *Eucalyptus*

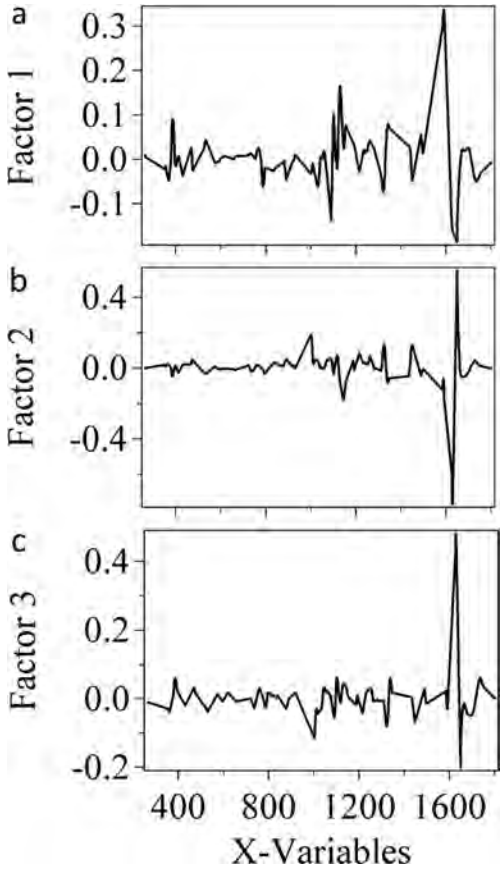

Fig. 3 Graphical representations of the **a** first PC loadings, **b** second PC loadings, and **c** third PC loadings used in the classification of the plant samples by genus

higher ratios can be identified along a diagonal path to the upper right quadrant. This suggests that both factors 2 and 3 are being employed to develop the classification of the trees based on S/G ratios.

Figure 4 shows the linearity of this pyMBMS/Raman model for both the calibration and full CV data sets. The reference and cross-validated lignin S/G ratios deviate from the linear trendline at higher S/G values. This overcompensation of S lignin is likely due to the fact that syringyl units are being preferentially released during the chemical degradation of lignin [13, 14]. Regardless of this deviation, the calibration (blue) and full CV (red) trendlines display a strong correlation. Plotting the regression coefficients (Fig. 5) allowed the isolation of integral spectral regions used for constructing the model. Table 1 lists the shaded wavenumber sections identified in Fig. 5, and characteristic lignin and lignin monomer vibrational modes potentially corresponding to these regions, as previously assigned in the literature. It should be noted that given the complex nature of biomass, there may be overlap between the vibrational modes of lignin and lignin monomers, with other cell wall constituents.

The model successfully identified and extracted the lignin spectral regions, including the regions of significant variance ascertained from Fig. 1. The regression coefficient plot was evaluated for specific monomeric trends; however, no distinct pattern emerged regarding the relationship between S or G moieties being predominantly positively or negatively correlated (Table 1). Despite some overlap between the Raman spectral assignments, there is a general consensus amid the references regarding peak location and their classification as bonds indicative of lignin and lignin monomers. It should be noted that differences in instrumental configurations can result in variation of vibrational mode peak locations. The strongest vibrational modes of cellulose occur at 1091 and 1117 cm^{-1}

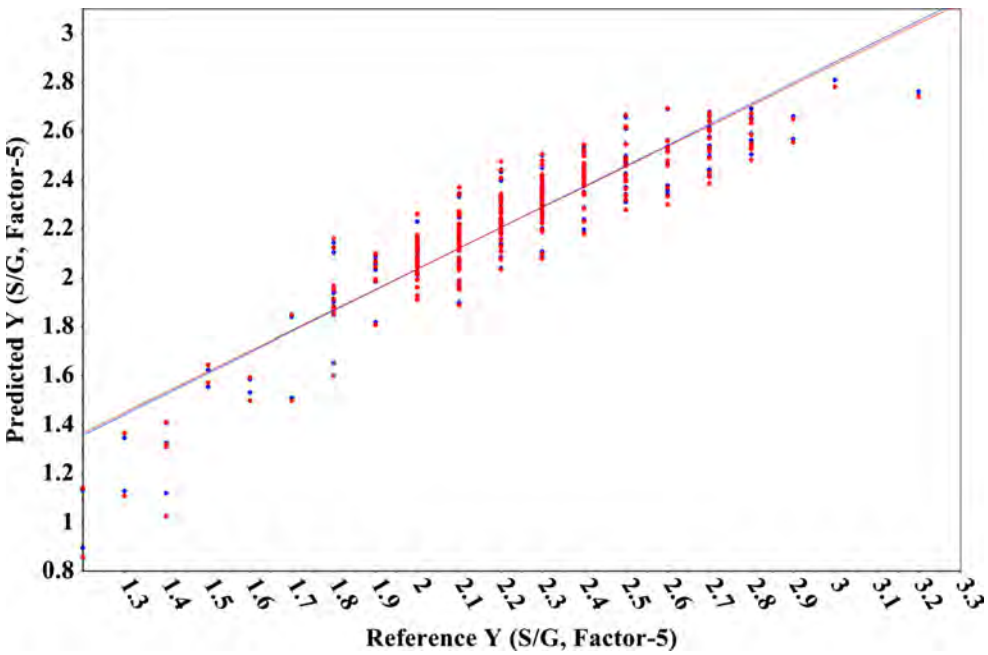

Fig. 4 Plot of the predicted lignin S/G ratio using a model built from first derivative, EMSC-transformed FT-Raman spectra and pyMBMS reference data. The *blue* and *red lines* signify the linearity of the calibration and prediction data sets, following a full cross-validation. The *x*-axis depicts the pyMBMS measured ratio, while the *y*-axis indicates the Raman predicted ratio. *S/G* syringyl to guaiacyl ratio, *FT* Fourier transform, *pyMBMS* pyrolysis molecular beam mass spectrometry, *EMSC* extended multiplicative scatter correction

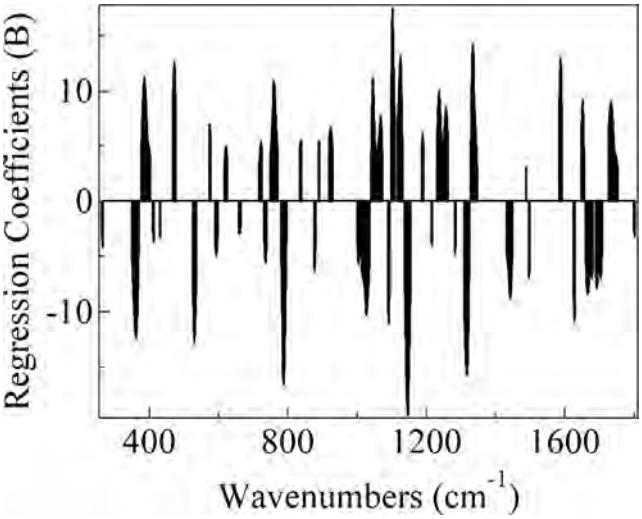

Fig. 5 Regression coefficients plot illustrating the spectral regions denoted as integral to the model calculation. The *black shaded spectral regions* illustrate the vibrational modes used in producing the model, while the *blue shaded areas* signify spectral noise excluded from the model construction. The *x*-axis is in wavenumbers, while the *y*-axis is the calculated regression coefficient values. *FT* Fourier transform, *PLS* partial least squares

[12]. These and weaker cellulose peaks are encompassed by the spectral regions identified in the Raman regression coefficients plots. Further analysis revealed that the cellulose or polysaccharide vibrational modes were either negatively correlated (Fig. 5, 1091 cm^{-1}), not identified as important to the model construction (1268 cm^{-1}), or positively correlated due to spectral overlap with lignin vibrational modes (e.g., 896, 1117, and 1338 cm^{-1}). Other potential sources of spectral overlap include xylan and extractive material such as proteins, lipids, etc. [12, 15–17]. Interestingly, three spectral regions identified in the regression coefficient plot correlated with postulated H lignin markers, as listed in Table 1. These occur between 833–838 and 1176–1178 cm^{-1}, and at 1488 cm^{-1}. The distinctive bands of S and G lignin, as well as cellulose, can be eliminated as correlative to these vibrational modes. Although its content in hardwoods is often minute, *Acacia* and *Eucalyptus* species have been determined to contain potentially 2–9 % H lignin, depending on the age of the trees [18–21]. Further chemical or instrumental analysis, such as thioacidolysis or 2D nuclear magnetic resonance, is required to determine if these vibrational modes correspond to H lignin.

As previously mentioned, the first motivation for conducting this research was to evaluate whether MVA models produced using Raman spectra and pyMBMS S/G ratios could accurately predict the S/G ratios in an unknown sample set, diminishing the need to destructively pyrolyze of all samples. The pyMBMS reference S/G ratios averaged for each plant species, including the number of samples for each tree, and the range of S/G ratios contained in each data set were previously reported [8]. Figure 6 illustrates the mean pyMBMS ratios using the 95 % confidence interval. Table 2 reveals the Raman predicted S/G ratios for the plant species in the unknown data matrix. A comparison between the pyMBMS and Raman predicted S/G values for each species

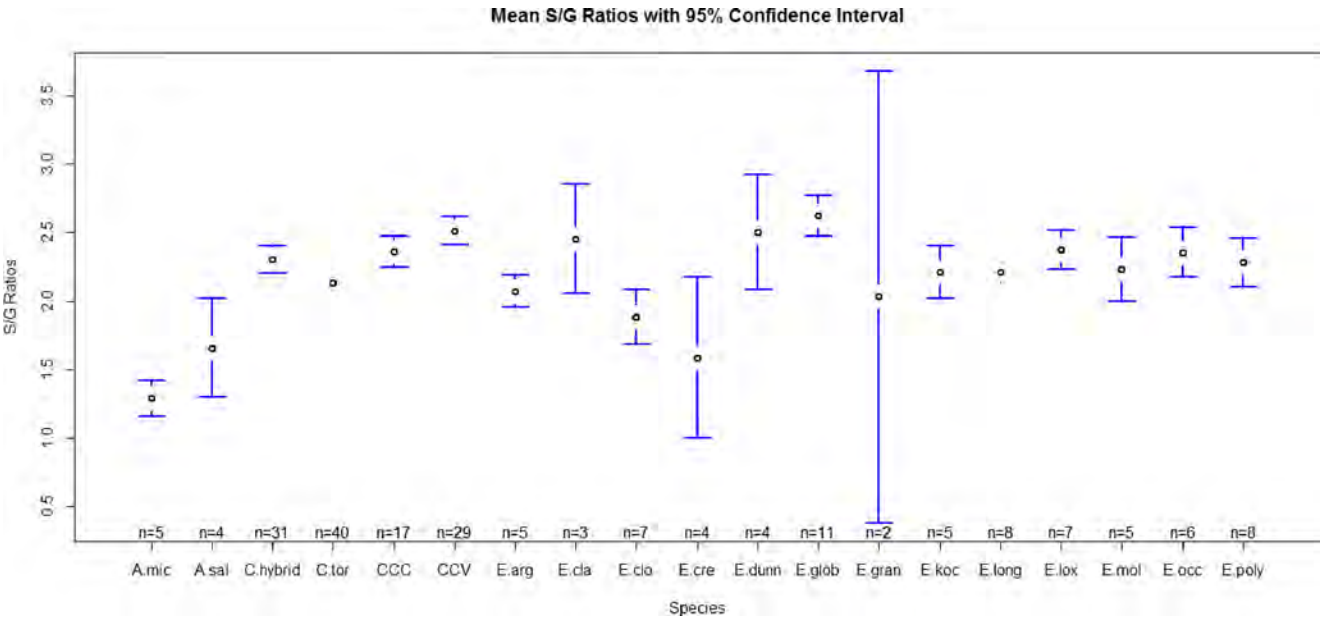

Fig. 6 Plot of species mean S/G ratios as determined by pyrolysis/molecular beam mass spectroscopy (pyMBMS). The 95 % confidence interval bars for each species is shown in *blue*. The number of samples for each species measured is located above the *x*-axis. Species labeled from left to right: *E. kochii*, *A. microbotrya*, *A. saligna*, *Corymbia* hybrids, *C. torelliana*, *C. citriodora* subsp. *citriodora*, *C. Citriodora* subsp. *variegata*, *E. argophloia*, *E. cladocalyx*, *E. cloeziana*, *E. crebra*, *E. dunnii*, *E. globulus* (subspecies *globulus* and *maidenii*), *E. grandis*, *E. longirostrata*, *E. loxophleba*, *E. moluccana*, *E. occidentalis*, and *E. polybractea*. *S/G* syringyl to guaiacyl lignin ratio

Table 2 Prediction matrix sample characteristics, S/G averages, and comparisons with the pyMBMS measured ratios

Plant species	No. of samples	Prediction range	Raman predicted S/G average	pyMBMS vs Raman comparisons (*p* values)
A. microbotrya	10	0.9–1.5	1.3±0.2	0.8
A. saligna	11	1.2–2.0	1.7±0.2	0.7
C. citriodora subsp. *citriodora*	44	2.0–2.7	2.3±0.1	0.6
Corymbia hybrids	–	–	–	NA
C. torelliana	–	–	–	NA
C. citriodora subsp.*variegata*	61	2.2–2.7	2.5±0.1	0.6
E. argophloia	5	1.7–2.0	1.8±0.1	*0.03*
E. cladocalyx	2	2.1, 2.4	2.2±0.2	0.4
E. cloeziana	15	1.7–2.4	2.1±0.2	0.3
E. crebra	6	1.2–2.1	1.8±0.3	0.6
E. dunnii	11	2.2–2.5	2.4±0.1	0.4
E. globulus[a]	19	2.0–2.8	2.5±0.2	0.2
E. grandis	13	1.9–2.4	2.2±0.1	0.2
E. kochii	10	1.7–2.5	2.2±0.2	0.8
E. longirostrata	7	2.0–2.3	2.2±0.1	0.9
E. loxophleba	23	2.2–2.7	2.4±0.1	0.2
E. moluccana	11	1.8–2.5	2.2±0.2	0.9
E. occidentalis	9	2.2–2.6	2.4±0.1	0.6
E. polybractea	12	1.9–2.5	2.2±0.1	0.4

The number of pyMBMS calibration samples, calibration ranges, and S/G averages for the plant species has been previously reported in the Additional Information of reference [8]

pyMBMS pyrolysis molecular beam mass spectrometry, *S/G* syringyl to guaiacyl lignin ratio

The italicized value is the only *p*-value < 0.05, which indicates a statistical difference between thepyMBMS and Raman mean S/G values

[a] *E. globulus* includes subspecies *globulus* and *maidenii*

using a non-parametric Mann-Whitney *U* test shows no significant differences (p value<0.05), with the exception of *Eucalyptus argophloia* (W=2, p value=0.03). This significant difference could be attributed to a small reference population (n=5), different genetic backgrounds, and/or environmental microsite variation, but the statistical comparison clearly illustrates the power and advantages of using robust, high-throughput multivariate modeling to predict lignin monomeric content, as the predicted lignin S/G ratios exhibit strong correlation with the pyMBMS range and average.

Once the Raman/pyMBMS model was verified to exhibit high accuracy, the ensuing question to explore was which samples exhibited the most significant variance in S/G ratios (i.e., which samples were at the measured or predicted S/G extremes). Evaluation of statistical differences within the pyMBMS and Raman data sets exposes some unique trends between the S/G ratios of the species measured. Pairwise comparisons between species in each data set were evaluated using Tukey's honestly significant differences for the pyMBMS data and Holm corrected Mann-Whitney *U* tests for the Raman data. These statistical analyses were selected since the pyMBMS data roughly followed a normal distribution, whereas the Raman predictions did not (negatively skewed). A p value lower than 0.05 indicates that the S/G ratios of the two species being compared are statistically different, while p values at or above 0.05 indicate analogous S/G ratios. *A. microbotrya*, *A. saligna*, and *E. crebra* are among the lowest S/G ratios in both the reference [8] and predicted data sets (Table 2). The Raman predictions for both *Acacia* species show significant differences from each *Corymbia* species (*Corymbia citriodora*subspecies *citriodora* (CCC) and CCV) and eight *Eucalyptus* species (Table 4). This result was confirmed in the pyMBMS data set where the *Acacias* displayed statistical differencesfrom 13 other eucalypts (Table 3). Although the S/G ratios of the *Acacia* trees were similar to *E. crebra* values in the Raman data set, *E. crebra* only showed significant disparity with the *Corymbia* samples and *Eucalyptus loxophleba*. This could potentially be attributed to the small number (n=6) of *E. crebra* samples used to generate the Raman prediction model, the fact that *E. crebra* was from a limited number of provenances sampled at only one site, or because of the greater number of *Corymbia* (CCV, n=61; CCC, n=44, sampled across multiple sites and provenances) and *E. loxophleba* contained in the model (n=23, multiple provenances), thereby increasing the predictive capabilities of the model for those samples. The need for larger sample sizes to predict significant variance between species is further illustrated by *E. cladocalyx* (n=3), which showed similarity with all other species (Table 4). CCV S/G ratios, predicted by Raman spectroscopy, reveal more statistical differences than those of CCC, when compared to the other plant species. *E. dunnii*, *Eucalyptus kochii*, *Eucalyptus longirostrata*, *E. moluccana*, and *Eucalyptus occidentalis* show similarities within the genus, but significant variance when juxtaposed with both *Acacias* (*E. kochii* differs only from *A. microbotrya*) and CCC. The prediction of the S/G ratio of unknown *E. globulus* subspecies *maidenii* shows no statistical dissimilarity to other eucalypts, with the exception of *Eucalyptus cloeziana*, *Eucalyptus grandis*, and *Eucalyptus polybractea.*

An assessment of the Raman predicted S/G ratios with the pyMBMS data further exemplifies the predictive capabilities of Raman spectroscopic modeling. Using Tables 3 and 4, a total of 136 evaluations were made to investigate which statistical dissimilarities were found in both predicted and reference data sets. Pairwise comparisons between species in the Raman data set identified 40 statistical differences (shown in bold), whereas 54 significant differences (shown in bold) were detected for the same species within the pyMBMS data set

Table 3 Statistical comparison between pyMBMS measured lignin S/G ratios using calculated *p* values from post hoc analysis of a single-factor ANOVA ($F(18,182)=16.82$, p value$<2\times10^{-16}$)

Plant species	*A.microbotrya*	*A. saligna*	CCC	CCV	*Corymbia* hybrids	*C.torelliana*	*E.argophloia*	*E.cladocalyx*	*E.cloeziana*
A. saligna	0.5								
C. citriodora subsp. *citriodora* (CCC)	**$<1\times10^{-7}$***	**3×10^{-6}***							
C. citriodora variegata-(CCV)	**$<1\times10^{-7}$***	**$<1\times10^{-7}$***	0.07						
Corymbia hybrids	**$<1\times10^{-7}$**	**1×10^{-5}**	1	**0.03**					
C. torelliana	**$<1\times10^{-7}$**	**0.006**	**0.03**	**$<1\times10^{-7}$**	0.09				
E. argophloia	**6×10^{-6}**	0.3	0.4	**0.005***	0.7	1			
E. cladocalyx	**$<1\times10^{-7}$**	**4×10^{-4}**	1	1	1	0.6	0.6		
E. cloeziana	**8×10^{-4}***	1	**2×10^{-4}***	**$<1\times10^{-7}$***	**8×10^{-4}**	0.3	1	**0.02**	
E. crebra	0.8	1	**1×10^{-7}***	**$<1\times10^{-7}$***	**4×10^{-7}**	**4×10^{-4}**	0.09	**6×10^{-5}**	0.8
E. dunnii	**$<1\times10^{-7}$***	**2×10^{-5}***	1	1	1	0.1	0.2	1	**0.001**
E. globulus[a]	**$<1\times10^{-7}$***	**$<1\times10^{-7}$***	0.1	1	**0.004**	**$<1\times10^{-7}$**	**6×10^{-4}**	1	**$<1\times10^{-7}$***
E. grandis	**0.008***	0.9	0.8	0.2	1	1	1	0.8	1
E. kochii	**$<1\times10^{-7}$***	**0.02**	1	0.3	1	1	1	1	0.5
E. longirostrata	**$<1\times10^{-7}$**	**0.006**	1	0.06	1	1	1	1	0.3
E. loxophleba	**$<1\times10^{-7}$***	**5×10^{-5}***	1	1	1	0.4	0.6	1	**0.004***
E. moluccana	**$<1\times10^{-7}$***	**0.01***	1	0.4	1	1	1	1	0.3
E. occidentalis	**$<1\times10^{-7}$***	**2×10^{-4}***	1	1	1	0.6	0.8	1	**0.01**
E. polybractea	**$<1\times10^{-7}$***	**6×10^{-4}***	1	0.4	1	1	1	1	**0.04**

Plant species	*E. crebra*	*E.dunnii*	*E.globulus*	*E.grandis*	*E.kochii*	*E.longirostrata*	*E.loxophleba*	*E.moluccana*	*E.occidentalis*
A. saligna									
C. citriodora subsp. *citriodora* (CCC)									
C. citriodora variegata-(CCV)									
Corymbia hybrids									
C. torelliana									
E. argophloia									
E. cladocalyx									
E. cloeziana									
E. crebra									
E. dunnii	**1×10^{-6}**								
E. globulus[a]	**$<1\times10^{-7}$**	1							
E. grandis	0.6	0.5	**0.04***						
E. kochii	**0.003**	0.9	0.05	1					
E. longirostrata	**6×10^{-4}**	0.7	**0.006**	1	1				
E. loxophleba	**4×10^{-6}***	1	0.6	0.9	1	1			
E. moluccana	**0.002**	0.9	0.08	1	1	1	1		
E. occidentalis	**2×10^{-5}**	1	0.6	0.9	1	1	1	1	
E. polybractea	**6×10^{-5}**	1	0.07	1	1	1	1	1	1

Values in bold indicate statistically different pairwise comparisons, while the asterisk (*) denotes statistically different comparisons found in both the pyMBMS and Raman data sets

pyMBMS pyrolysis molecular beam mass spectrometry, *S/G* lignin syringyl to guaiacyl ratio, *ANOVA* analysis of variance

[a] *E. globulus* includes subspecies *globulus* and *maidenii*

Table 4 Statistical comparison between Raman spectroscopy predicted lignin S/G ratios using calculated *p* values from Kruskal-Wallis test (χ^2=155.99, *p* value<2.2×10^{-16}) and post hoc pairwise Mann-Whitney *U* tests

Plant species	*A.microbotrya*	*A.saligna*	CCC	CCV	*E.argophloia*	*E.cladocalyx*	*E.cloeziana*	*E.crebra*
A. saligna	0.2							
C. citriodora subsp. *citriodora* (CCC)	**1×10^{-4}***	**6×10^{-5}***						
C. citriodora variegata (CCV)	**6×10^{-5}***	**2×10^{-5}***	**8×10^{-4}**					
E. argophloia	0.2	1	**0.03**	**0.02***				
E. cladocalyx	1	1	1	1	1			
E. cloeziana	**0.004***	**0.02**	**0.03***	**7×10^{-5}***	1	1		
E. crebra	1	1	**0.01***	**0.007***	1	1	1	
E. dunnii	**0.01***	**0.009***	1	1	0.2	1	0.1	0.09
E. globulus[a]	**0.002***	**0.001***	0.05	1	0.07	1	**0.01***	0.05
E. grandis	**0.008***	**0.006**	0.2	**2×10^{-4}**	0.2	1	1	0.2
E. kochii	**0.02***	0.05	1	**0.01**	1	1	1	1
E. longirostrata	0.07	0.05	1	**0.02**	0.4	1	1	0.4
E. loxophleba	**9×10^{-4}***	**4×10^{-4}***	0.5	1	0.06	1	**0.008***	**0.02***
E. moluccana	**0.01***	**0.03***	1	**0.04**	0.9	1	1	1
E. occidentalis	**0.03***	**0.02***	1	1	0.2	1	0.21	0.14
E. polybractea	**0.01***	**0.008***	0.1	**5×10^{-4}**	0.2	1	1	0.2

Plant species	*E.dunnii*	*E.globulus*	*E.grandis*	*E.kochii*	*E.longirostrata*	*E.loxophleba*	*E.moluccana*	*E.occidentalis*
A. saligna								
C. citriodora subsp. *citriodora* (CCC)								
C. citriodora variegata (CCV)								
E. argophloia								
E. cladocalyx								
E. cloeziana								
E. crebra								
E. dunnii								
E. globulus[a]	0.8							
E. grandis	0.09	**0.01***						
E. kochii	1	0.06	1					
E. longirostrata	0.2	0.2	1	1				
E. loxophleba	1	1	**0.01**	0.1	0.1			
E. moluccana	1	0.11	1	1	1	0.4		
E. occidentalis	1	1	0.4	1	1	1	1	
E. polybractea	0.2	**0.02**	1	1	1	**0.01**	1	0.4

Values in bold indicate a statistically different pairwise comparison, while the asterisk (*) denotes statistically different comparisons found in both the pyMBMS and Raman data sets

S/G lignin syringyl to guaiacyl ratio

[a] *E. globulus* includes subspecies *globulus* and *maidenii*

(*C. torelliana* and *Corymbia* hybrids were excluded for comparison as they were only tested using the reference method). Of those, 28 evaluations between species were found to be significant in both data sets (denoted with an asterisk). These results clearly illustrate the lower S/G values of the *Acacia*s when contrasted with the eucalypt samples. The other significant variations in S/G ratios, discovered between species within one data set, but not both, are likely due to the input of small sample sizes into the tests (*E. grandis* (reference), $n=2$; *E. cladocalyx* (predicted), $n=2$), or the use of a single factor analysis of variance (ANOVA) and post hoc testing with the pyMBMS data set, which is more robust at elucidating significant discrepancies than non-parametric tests. The data sets clearly illustrate differences in monomeric lignin composition between a diverse group of *Acacia* and eucalypt wood samples. Further investigation into additional sources of variance, such as age and site effects, will provide understanding into the biological context of wood formation for these important forestry species.

Conclusions

There are greater than 900 diverse species of both *Acacias* and eucalypts, the latter including *Corymbia* and *Eucalyptus*. In order to isolate which trees may be the most advantageous for developing biofuels and bio-based chemicals, phenotypic traits that correlate to plant cell wall structure and recalcitrance must be evaluated, such that suitable deconstruction strategies can be postulated. Many of the standard techniques for measuring monomeric content and ratio are laborious, destructive, toxic, and may require complex data analysis, making these methods unsuitable for screening large populations. The employment of Raman spectroscopy can enable the rapid, non-destructive, screening of potential feedstocks, such as *Acacias* and eucalypts, for traits deemed important for biofuel and/or bio-based chemical production, and most attuned to the needs of biorefineries. The construction of a robust, multivariate, high-throughput Raman model has been previously established. The current study examined the actual practicality of using this model to gauge the lignin S/G ratio in a large unknown data setof *Acacias* and eucalypts. The means of the predicted *Acacias* and eucalypts S/G ratios were not statistically different from those measured using pyMBMS, with the exception of *E. argophloia*, which could be due to the small sample size analyzed, genetic variations, and/or environmental microsite variations. This research shows the potential of using Raman spectroscopy to supplant tedious, destructive methods for the evaluation of the lignin S/G ratios of different biomass.

Acknowledgments This manuscript was supported as part of a collaboration between the Queensland Alliance for Agriculture and Food Innovation and the Joint BioEnergy Institute. The work conducted by the Joint BioEnergy Institute was supported by the Office of Science, Office of Biological and Environmental Research, of the US Department of Energy under contract no. DE-AC02-05CH11231. The BioEnergy Science Center is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. The authors would like to thank Erica Gjersing at the National Renewable Energy Lab, for assistance and guidance with respect to the high-throughput pyMBMS pipeline, and John Bartle, Western Australian Department of Environment and Conservation, for the collecting and processing of some the wood samples and information regarding the environmental specifications of the growing site. The material from the Queensland and New South Wales sites was accessed from Queensland Department of Agriculture, Fisheries and Forestry trials.

Conflict of Interest The authors declare that they have no competing interests.

References

1. Davison BH, Drescher SR, Tuskan GA, Davis MF, Nghiem NP (2006) Variation of S/G ratio and lignin content in a *Populus* family influences the release of xylose by dilute acid hydrolysis. Appl Biochem Biotechnol 129–132:427–435. doi:10.1385/abab:130:1:427
2. del Rio JC, Gutierrez A, Hernando M, Landin P, Romero J, Martinez AT (2005) Determining the influence of eucalypt lignin composition in paper pulp yield using Py-GC/MS. J Anal Appl Pyrolysis 74(1–2): 110–115. doi:10.1016/j.jaap.2004.10.010
3. Grabber JH, Hatfield RD, Ralph J (2003) Apoplastic pH and monolignol addition rate effects on lignin formation and cell wall degradability in maize. J Agric Food Chem 51(17):4984–4989. doi: 10.1021/jf030027c
4. Li X, Ximenes E, Kim Y, Slininger M, Meilan R, Ladisch M, Chapple C (2010) Lignin monomer composition affects *Arabidopsis* cell-wall degradability after liquid hot water pretreatment. Biotechnol Biofuels 3:27–33. doi:10.1186/1754-6834-3-27
5. Mechin V, Argillier O, Menanteau V, Barriere Y, Mila I, Pollet B, Lapierre C (2000) Relationship of cell wall composition to in vitro cell wall digestibility of maize inbred line stems. J Sci Food Agric 80(5):574–580
6. Studer MH, DeMartini JD, Davis MF, Sykes RW, Davison B, Keller M, Tuskan GA, Wyman CE (2011) Lignin content in natural *Populus* variants affects sugar release. Proc Natl Acad Sci U S A 108(15): 6300–6305. doi:10.1073/pnas.1009252108, S6300/6301-S6300/6306
7. Tsutsumi Y, Kondo R, Sakai K, Imamura H (1995) The difference of reactivity between syringyl lignin and guaiacyl lignin in alkaline systems. Holzforschung 49(5):423–428. doi:10.1515/hfsg.1995.49.5.423
8. Lupoi JS, Singh S, Davis M, Lee DJ, Shepherd M, Simmons BA, Henry RJ (2014) High-throughput prediction of eucalypt lignin syringyl/guaiacyl content using multivariate analysis: a comparison between mid-infrared, near-infrared, and Raman spectroscopies for model development. Biotechnol Biofuels in print 7:93-106
9. Sykes R, Yung M, Novaes E, Kirst M, Peter G, Davis M (2009) High-throughput screening of plant cell-wall composition using

pyrolysis molecular beam mass spectroscopy. Methods Mol Biol 581:169–183
10. Kruskal W, Wallis W (1952) Use of ranks in one-criterion variance analysis. J Am Stat Assoc 47(260):583–621
11. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6(2):65–70
12. Wiley JH, Atalla RH (1987) Band assignments in the Raman spectra of celluloses. Carbohydr Res 160:113–129. doi:10.1016/0008-6215(87)80306-3
13. Sarkanen KV, Ludwig CH (eds) (1971) Lignins: occurrence and formation, structure, chemical and macromolecular properties, and utilization. Wiley, New York
14. Sykes R, Kodrzycki B, Tuskan G, Foutz K, Davis M (2008) Within tree variability of lignin composition in Populus. Wood Sci Technol 42(8):649–661. doi:10.1007/s00226-008-0199-0
15. Kacurikova M, Wellner N, Ebringerova A, Hromidkova Z, Wilson RH, Belton PS (1998) Characterization of xylan-type polysaccharides and associated cell wall components by FT-IR and FT-raman spectroscopies. Food Hydrocoll 13(1):35–41
16. Schenzel K, Fischer S (2001) NIR FT Raman spectroscopy—a rapid analytical tool for detecting the transformation of cellulose polymorphs. Cellulose (Dordrecht, Netherlands) 8(1):49–57
17. Schulz H, Baranska M (2007) Identification and quantification of valuable plant substances by IR and Raman spectroscopy. Vib Spectrosc 43(1):13–25. doi:10.1016/j.vibspec.2006.06.001
18. Lourenco A, Gominho J, Marques AV, Pereira H (2013) Comparison of Py-GC/FID and wet chemistry analysis for lignin determination in wood and pulps from *Eucalyptus globulus*. BioResources 8(2):2967–2980
19. Pinto PC, Evtuguin DV, Pascoal Neto C (2005) Effect of structural features of wood biopolymers on hardwood pulping and bleaching performance. Ind Eng Chem Res 44(26):9777–9784. doi:10.1021/ie050760o
20. Pinto PC, Evtuguin DV, Pascoal Neto C (2005) Chemical composition and structural features of the macromolecular components of plantation *Acacia mangium* wood. J Agric Food Chem 53(20):7856–7862. doi:10.1021/jf058081b
21. Rencoret J, Gutierrez A, Nieto L, Jimenez-Barbero J, Faulds CB, Kim H, Ralph J, Martinez AT, del Rio JC (2011) Lignin composition and structure in young versus adult *Eucalyptus globulus* plants. Plant Physiol 155(2):667–682. doi:10.1104/pp.110.167254
22. Larsen KL, Barsberg S (2010) Theoretical and Raman spectroscopic Sstudies of phenolic lignin model monomers. J Phys Chem B 114(23):8009–8021. doi:10.1021/jp1028239
23. Sun L, Varanasi P, Yang F, Loque D, Simmons BA, Singh S (2012) Rapid determination of syringyl:guaiacyl ratios using FT-Raman spectroscopy. Biotechnol Bioeng 109(3):647–656. doi:10.1002/bit.24348
24. Agarwal UP, McSweeny JD, Ralph SA (2011) FT-Raman investigation of milled-wood lignins: softwood, hardwood, and chemically modified black spruce lignins. J Wood Chem Technol 31(4):324–344. doi:10.1080/02773813.2011.562338
25. Saariaho A-M, Argyropoulos DS, Jaeaeskelaeinen A-S, Vuorinen T (2005) Development of the partial least squares models for the interpretation of the UV resonance Raman spectra of lignin model compounds. Vib Spectrosc 37(1):111–121. doi:10.1016/j.vibspec.2004.08.001
26. Meyer MW, Lupoi JS, Smith EA (2011) 1064 nm dispersive multichannel Raman spectroscopy for the analysis of plant lignin. Anal Chim Acta 706(1):164–170. doi:10.1016/j.aca.2011.08.031