DOI 10.1007/s12155-016-9801-8

High-Throughput Profiling of the Fiber and Sugar Composition of Sugarcane Biomass

Nam V. Hoang [1,2] · Agnelo Furtado [1] · Laura Donnan [3] · Eloise C. Keeffe [3] · Frederik C. Botha [1,3] · Robert J. Henry [1,4]

Published online: 30 November 2016

Abstract Lignocellulosic biomass from sugarcane (*Saccharum* spp. hybrids) could potentially be a major feedstock for second-generation biofuel production. Consequently, selecting sugarcane varieties with favorable biomass characteristics, typically less enzymatic recalcitrance and better saccharification yield without sugar-yield penalty, will be important in sugarcane breeding. Economical and high-throughput techniques for profiling the major biomass components of this complex system will facilitate selection of clones with ideal lignocellulosic composition from large numbers of genotypes in breeding programs. We used a combined high-throughput profiling approach to evaluate the biomass composition of samples from a sugarcane germplasm collection. This employed near-infrared (NIR) spectroscopy for fiber characterization and high-performance liquid chromatography (HPLC) for determining the sugar content in juice. The results for 331 samples, from a diverse sugarcane population of 186 genotypes, derived from 143 parents of different genetic backgrounds, showed that high-quality NIR spectroscopic predictions were feasible for cellulose, hemicellulose, lignin, and extractives values in fiber, and sugars in juice were suitably analyzed by HPLC. The analysis of total biomass indicated that this NIR- and HPLC-based high-throughput method allowed a robust phenotypic assessment of a large number of samples for the key biomass traits in the sugarcane system, including total dry biomass, fiber, sugar content, and theoretical ethanol yields, and could potentially become the method of choice for sugarcane germplasm screening in breeding programs targeting the support of biofuel production.

Keywords Lignocellulosic biomass · Near-infrared spectroscopy · High-throughput · NIR prediction · Fiber/sugar ratio · Sugarcane · Biomass profiling

Introduction

Due to the depletion of nonrenewable liquid fossil fuels, production of biofuels from plant biomass has been emphasized in recent years. While first-generation biofuels use the food parts of the plants such as starch, sugar, and oil to produce ethanol, the emerging second-generation biofuels have been generated from the cell-wall-derived biomass (also referred to as lignocellulosic biomass) and other nonfood parts of the plants [1, 2]. Sugarcane is a very important industrial crop in tropical and subtropical countries. It has been selected as a candidate for biofuel production due to its exceptional capacity to produce biomass in a short time compared to other biomass-producing plants such as switch grass, *Miscanthus*, *Erianthus*, and woody candidates such as eucalypts reviewed in [3]. A dual-purpose sugarcane system, for sugar and lignocellulosic biomass production, stands out from other feedstock options because its breeding, cultivation, production and processing facilities have been well established and are already in place. Sugar and lignocellulosic biomass from sugarcane can be produced simultaneously, on the same land area [4] and

✉ Robert J. Henry
robert.henry@uq.edu.au

[1] Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, St. Lucia, QLD 4072, Australia

[2] College of Agriculture and Forestry, Hue University, Hue, Vietnam

[3] Sugar Research Australia, Indooroopilly, QLD 4068, Australia

[4] The University of Queensland, Level 2, The John Hay Building, Queensland Biosciences Precinct [#80], 306 Carmody Road, St Lucia, QLD 4072, Australia

with the same input costs. Approximately 1.9 billion tons of sugarcane biomass is produced each year [5], which contains about 580 million dry tons of lignocellulosic biomass that could be used as a feedstock for the production of second-generation biofuels [6].

To maximize the conversion efficiency of sugarcane biomass to biofuels, it is imperative to have sugarcane genotypes with improved total biomass: more cellulose and less lignin, resulting in less enzymatic recalcitrance and better saccharification yield. Achieving this requires that we first understand the genetic basis of the key biomass traits. Association studies have been emerging as a method of choice in dissecting complex traits, and could potentially allow better understanding of the genes underpinning biomass traits and relationships between them, as outlined in our review [3]. To carry out a large-scale association study, a population of sugarcane of genetic variability in the targeted traits is required for phenotyping and genotyping [7]. Phenotyping for association studies necessitates the costly assessment of hundreds to thousands of samples from a large number of accessions collected over several years and from several locations [8–10]. Hence, it is costly to generate a phenotypic dataset of sufficient size. The traits targeted for biofuel research are constituents of the lignocellulosic fraction (cellulose, hemicellulose, lignin) and sugars (mostly sucrose). A complete and accurate set of phenotypic data allows reliable dissection of the relationship between these traits and the genes controlling them.

Cellulose, hemicellulose and lignin are important constituents of the lignocellulosic biomass (or of the total fiber fraction), and therefore are normally characterized for biofuel research [4, 11, 12]. In fiber characterization, a high-throughput method to assess and screen a large number of lignocellulosic feedstocks is required to replace the expensive, time-consuming, and tedious chemical composition analysis techniques. To date, there are a range of available methods for assessment of lignocellulosic biomass reviewed in [13]. Among the available platforms, near-infrared (NIR) spectroscopy-based methods have been used widely for assessing biomass since they offer nondestructive analysis (reducing hazardous risks and allowing samples to be re-used for other purposes), a relatively low cost per sample and require minimal technical skill reviewed in [13–15]. To develop NIR spectroscopic models, paired spectra and reference values (obtained through traditional analytical methods) are combined using chemometric technique partial least squares (PLS) regression. Applying these models to an at-line analyzer allows rapid, high-throughput prediction of lignocellulosic constituents for new samples [16]. The prediction of biomass composition using NIR spectral data has been developed for many species, such as *Miscanthus* [17], switchgrass [18], poplar [19], sugarcane (on three selected genotypes) [11], and in broad-based multispecies models (including corn, sorghum, rice, and wheat) [16]. In the Australian sugarcane industry, NIR spectroscopic models have been developed for evaluation of many sugar mill products, such as prepared cane, bagasse, raw sugar, molasses, juice and syrups, massecuite, magma, and mud [4, 12, 20]. Additionally, fibrated cane (whole stalk) is analyzed in the field environment with SpectraCane (Biolab, Australia), an integrated at-line analyzer comprising a shredder to homogenize cane stalks into a fiber, conveyer, and sample presentation component, and an NIR spectroscopic instrument, as described in [21, 22]. High-performance liquid chromatography (HPLC) is a well-established system for quantifying and profiling sugars in sugarcane juice [23, 24]. Combination of the juice characterization, provided by HPLC, and fiber characterization, provided by NIR spectroscopy, gives a comprehensive profile of all phenotypic factors contributing to the total biomass for genetic studies. A complete data of sugar and fiber content in the biomass across a given population would allow to estimate the theoretical ethanol yield, and hence, the potential for biofuel production, as illustrated in [25].

The aims of this study were to (i) explore the potential and evaluate the robustness of a combined high-throughput approach based upon NIR spectroscopy for fiber characterization and HPLC for sugar analysis in assessing sugarcane biomass traits, (ii) systematically profile the biomass composition and ethanol yield potential of a sugarcane germplasm collection, and (iii) establish a method to determine the fiber/sugar ratio in sugarcane genotypes, as a phenotyping tool for sugarcane biofuel research.

Materials and Methods

Biomass Sample Collection

This study was based on 331 samples, derived from 186 sugarcane genotypes from a population with partial duplicates, triplicates, and three control genotypes (Q200, Q208, and KQ228) with six replicates each. These were grown in Sugar Research Australia (SRA) field trials at their Brandon Station in the Burdekin region of Queensland, Australia. Six millable stalks from each sugarcane sample were collected after leaves and green tops were removed. The following processing occurred within 1 h of sample collection. The stalks were fed into the SpectraCane (Biolab, Australia) NIR analysis system, which automatically shreds the cane and collects an NIR spectrum using a Bruker Matrix-F NIR spectrophotometer (Bruker, Billerica, MA, USA). This system provided a prediction for total fiber content, juice pol, juice brix, and commercial cane sugar (CCS), which was based on analysis of the collected spectral data using the calibration models previously developed for the whole stalk samples. For fiber composition prediction in this study, a second NIR spectrum was collected on the pressed fiber samples, using a different

NIR spectroscopic instrument, a Foss XDS, described in the next section. Following sample preparation and analysis by SpectraCane, 400 g of the well-mixed shredded cane of each sample was transferred to a canister and pressed using a Carver mechanical press auto M series (Carver, Inc., Wabash, IN, USA) at 20,000 lb per inch (psi) for 45 s to obtain the juice and a pressed fiber cake (referred to as pressed fiber). This pressed fiber was weighed and referred to as pressed fiber fresh weight in total biomass calculation. Approximately 50 ml of the juice was collected into a 50-ml Falcon tube, and the remainder was discarded. The collected juice and pressed fiber samples were stored at −20 °C until further analysis.

Fiber Sample Characterization

Sample Preparation and NIR Spectral Data Acquisition

Pressed fiber sample preparation and analysis was undertaken as follows, using protocols adapted from those developed by the US National Renewable Energy Laboratory (NREL) [26]. The pressed fiber was removed from the freezer and thawed at room temperature for 1 h. Once completely thawed, the sample was spread over a flat seedling tray lined with paper and mixed thoroughly. NIR spectral data (400 to 2500 nm with 0.5-nm data spacing) was collected using a Foss XDS with Rapid Content Analyzer (Foss NIRSystems, Hillerød, Denmark) and a moving solids module. Subsequently, the pressed fiber sample was dried overnight in an oven at 42 °C, until a constant mass was recorded. This value is referred to as pressed fiber dry weight. The dried samples were ground using a Retsch grinder (Retsch, Haan, Germany) with 1.5 mm sieve plate set in a reverse position to produce a fine and homogenous particle size. Each finely ground sample (25 g) was sieved in a shaker (Endecotts Ltd., London, England) for 20 min, and the 180–350-μm fraction was retained and stored at 4 °C until further analysis.

Selecting Samples for Compositional Analysis by Traditional Methods

Sample selection for chemical composition analysis was completed using the Kennard-Stone algorithm [27] in the Unscrambler X software, version 10.3 (Camo, Inc., Oslo, Norway). Of the original pool of 331 samples, 20 samples were selected based on the even distribution of their spectra in multivariate space [16]. An additional ten samples were selected manually from the pool by evaluating the predicted fiber content generated from the SpectraCane prediction, in combination with the sample dry weight data. This was to increase the number of selected samples specifically represent the population from low to high fiber genotypes. The 30 samples selected by these methods were subjected to compositional analysis and used to calibrate the NIR models.

Compositional Analysis by Traditional Analytical Methods

Compositional analysis of the selected samples was conducted for cellulose, hemicellulose, lignin, ash, and extractives, using the NREL methods, modified for sugarcane fiber [26, 28]. Prior to analysis, the ground and sieved fiber was dried in an oven for at least 5 h to a constant weight at 105 °C, to remove the moisture absorbed during storage. The ground and sieved fiber was extracted sequentially with water and 100 % ethanol using an accelerated solvent extractor (ASE 350, Dionex, Sunnyvale, CA, USA). The settings for the ASE 350 were as follows: for water extraction (3 cycles): preheating 1 min, heating 5 min, pressure 1000 psi, temperature 80 °C, static 7 min, flush% 150 V, purge 120 s; ethanol extraction (3 cycles): preheating 1 min, heating 5 min, pressure 1000 psi, temperature 90 °C, static 7 min, flush% 150 V, purge 120 s. The total extractables in the sample were determined by subsequent quantitative recovery of the residue using a rotary evaporator at 40 °C. The extractives-free fiber was dried at 40 °C in a vacuum oven for at least 12 h to ensure evaporation of all residual solvent. Then, the dry extractives-free fiber was subjected to a two-stage sulfuric acid hydrolysis procedure (30 °C/72 % H_2SO_4/1 h) and (121 °C/4 % H_2SO_4/1 h). The first hydrolysis was completed at atmospheric pressure, while the second hydrolysis step was performed under high-pressure conditions at 20 psi in an autoclave. The hydrolysate was filtered by vacuum filtration and two fractions were obtained, filtrate and the solids. The acid-soluble lignin (ASL) component in the filtrate was determined by UV-visible spectroscopy at 320 nm. The acid-insoluble lignin (AIL) component was estimated from the total weight of dried solids (determined from drying at 105 °C) after subtracting the acid-insoluble ash content, which was determined by incinerating the solids at 575 °C. Total lignin was the sum of ASL and AIL. For structural cellulose and hemicellulose determination, the filtrate was neutralized with calcium carbonate and diluted before injecting into a Dionex HPLC system controlled by Chromeleon software (Dionex, Sunnyvale, CA, USA). Sugar standards, including glucose, arabinose, galactose, xylose, and mannose, were used to calibrate the sugars in the filtrate and fucose was used as an internal standard in all sample injections. The cellulose content was calculated from glucose and the hemicellulose content was calculated by adding individual results for arabinose, galactose, xylose, and mannose. Total ash content determination was performed by incinerating the dried sample using a muffle furnace at 575 °C, and the weight of the crucible containing the sample, before and after incineration, was used to calculate the ash content. All samples were analyzed in duplicate. The chemical compositional analysis results were quoted on a % dry mass

(%DM) basis or % extractives-free (%EF) basis in the pressed fiber fraction, in which the difference between these two expressed values was the amount of extractable materials remaining in the samples. For the NIR spectroscopic calibration model, the data on a %DM basis was used, as it is mentioned in [28] that models developed on this type of data typically showed an improved performance.

NIR Spectroscopic Model Development and Prediction

The analytical reference data and NIR spectroscopic data for the 30 selected samples were added to an existing database of sugarcane stalk and bagasse, previously reported in [28], to generate a calibration dataset of 300 samples. To maximize the number of samples available for developing the calibration models, all replicates in chemical composition analysis were treated as individual samples (approximately 600 individual samples). The spectral data was analyzed using first derivative or second derivative (Savitzky–Golay) transformation with 14-point smoothing to correct the variation in the baseline, and with standard normal variate (SNV) for scatter correction. Data was checked for outliers using principal component analysis (PCA). Partial least squares (PLS-1) regression models were developed for cellulose, hemicellulose, lignin, and extractives content using Unscrambler X 10.3. We applied two levels of model development in this study. At first, a calibration with an external validation set (of 152 samples, making a total of 452 validation samples) was performed to get an assessment of potential predictive performance of the models. This was followed by a cross-validation model using all available data, to maximize the variability captured in the dataset. The cross-validated model was used for the prediction of unknowns. To assess the predictive performance of the externally validated calibration models, a subset of each of the products (whole stalk, pressed fiber, and bagasse) were removed prior to model development and used as an independent validation set. The coefficient of determination (R^2), root mean standard error of calibration (RMSEC), and root mean standard error of prediction (RMSEP) were used to evaluate the performance of the models. Again, to maximize the number of samples in the calibration and improve the predictive performance, the models were recalculated using all of the available data, with only cross-validation. Experience has shown that this combination of model development will give the best possible predictive models while providing an indicative RMSEP, despite the low sample numbers. The cellulose, hemicellulose, lignin, and extractives content in the unknown samples were predicted using the cross-validated models.

Analysis of Soluble Sugars in Juice

The frozen juice samples were thawed at room temperature for 1 h and mixed thoroughly. A 2 mL aliquot of each juice sample was transferred to a 2-mL tube and heated on a water bath at 98 °C for 10 min. The sample was centrifuged at 15,000×*g* for 15 min and 50 μL of the supernatant, collected from 3 mm below the surface of the sample, was diluted 20 times by transferring to a new tube containing 950 μL Milli-Q water. This process was repeated three times, resulting in the original sample being diluted 8000 times. A further 1:1 dilution was undertaken with 500 μL of degassed Milli-Q water, making the final solution a dilution of 16,000 times to that of the original sample solution. The final diluted solution (1 mL) was transferred to a 96-well plate and subjected to analysis using a Dionex BioLC system DX 600 (Dionex, Sunnyvale, CA, USA). Sucrose, glucose, and fructose concentrations were measured at high pH (85 mM L^{-1} KOH), with a PA20 analytical anion exchange column and calibration with sugar standards as described by Wu and Birch [24]. Sugar standards used for calibration were in a dilution series ranging 1–150, 3–300, and 1–150 mM for sucrose, fructose, and glucose, respectively. The sugar concentration in the sample was adjusted for the dilution factor using Chromeleon Client software Version 6.8 SR8 build 2623 (Dionex, Sunnyvale, CA, USA). The sucrose, fructose, and glucose contents in juice samples were quoted in millimolar (mM) concentration and the values back-calculated to represent the amount of sugar in the total juice extracted from 400 g of shredded cane.

Total Biomass Composition and Data Analysis

All the data obtained from fiber characterization and juice sugar measurement were combined and used for back-calculation of the composition of 400-g shredded cane sample (referred to as total fresh biomass) originally collected (Fig. 1). The calculation for each sample is explained as follows:

- Moisture in pressed fiber = (pressed fiber fresh weight − pressed fiber dry weight).
- Cellulose = (predicted % cellulose × pressed fiber dry mass) = cellulose in 400 g total biomass.
- Hemicellulose = (predicted % hemicellulose × pressed fiber dry mass) = hemicellulose in 400 g total biomass.
- Lignin = (predicted % lignin × pressed fiber dry mass) = lignin in 400 g total biomass.
- Total fiber (in 400 g total biomass) = (cellulose + hemicellulose + lignin).
- Juice = (400 g − fresh weight of pressed fiber).
- Water in juice = (total juice − measured sugars − 3 % juice); *3 % of nonsugar solubles in juice [29].
- Water in 400 g total biomass = (moisture in pressed fiber + water in juice).
- Unextractable sugar in pressed fiber = moisture in pressed fiber × (sugar in juice / water in juice).
- Total sugar (in 400 g total biomass) = (sugar in juice + unextractable sugar in pressed fiber).

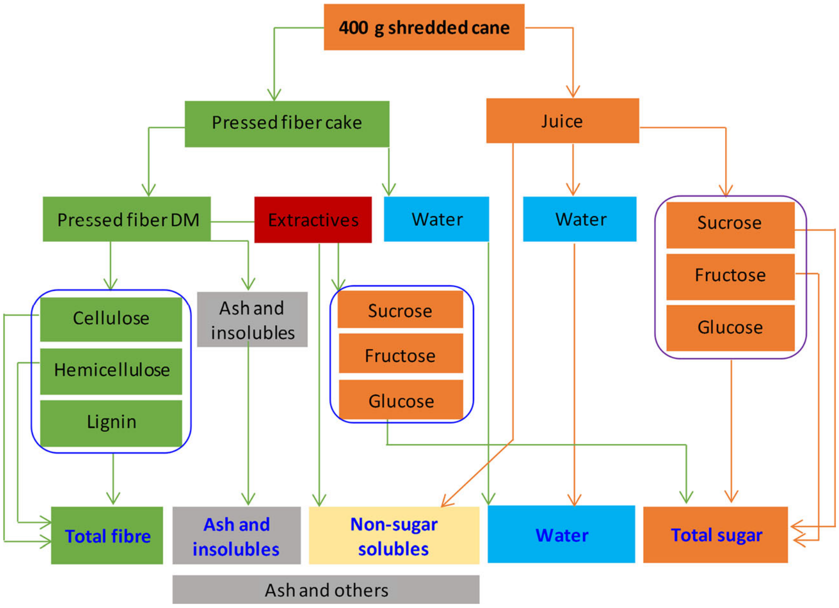

Fig. 1 The overview analysis strategies of sugarcane biomass profiling for samples collected from millable stalk, including fiber characterization and juice sugar assessment

- Total sucrose (in 400 g total biomass) = (sucrose in juice + unextractable sucrose in pressed fiber).
- Total glucose (in 400 g total biomass) = (glucose in juice + unextractable glucose in pressed fiber).
- Total fructose (in 400 g total biomass) = (fructose in juice + unextractable fructose in pressed fiber).
- Ash and insolubles in pressed fiber = (pressed fiber dry mass − total fiber − extractives).
- Nonsugar solubles in pressed fiber = (extractives in pressed fiber − unextractable sugar in pressed fiber).
- Total ash and others (in 400 g total biomass) = ash and insolubles in pressed fiber + nonsugar solubles in pressed fiber + 3 % juice.
- Total dry biomass = (fiber + total sugar + total ash and others).

All components were expressed on a percentage of total fresh biomass basis (calculated for the total 400 g) and a total dry biomass basis. Mass balance was used to check the accuracy of the biomass composition calculation. A Pearson correlation coefficient was computed pairwise for the major biomass components.

Prediction of Theoretical Ethanol Yields

Theoretical ethanol yields were computed separately for sugars, fiber, and then for both together, on a fresh and dry biomass basis, based on a stoichiometry calculation of prediction model described in [18, 30]. This used a conversion factor of 0.568 g/g for cellulose, 0.581 g/g for hemicellulose xylan, 0.537 g/g for sucrose, and 0.511 g/g for glucose and fructose. The equations for calculating theoretical ethanol yields are listed below (assuming 100 % conversion, and 1.267 mL is the volume of 1 g ethanol):

- Predicted sugar-based ethanol (L Mg^{-1}) = ((glucose + fructose) × 0.511) + (sucrose × 0.537)) × 1.267.

- Predicted fiber-based ethanol (L Mg^{-1}) = ((cellulose × 0.568) + (hemicellulose × 0.581)) × 1.267.

Data Analysis

All data analyses were done using the Data Analysis ToolPak in Microsoft Excel 2013 and RStudio ver.0.9.8/R ver.3.1.2 [31], including lattice [32], ggplot2 [33], reshape2 [34], and Hmisc [35] packages. Pairwise comparison panels of main biomass components were generated by using R built-in functions; code can be accessed here [36].

Results

Chemical Composition Analysis

The results for biomass compositional analysis on 30 selected pressed fiber samples are presented in Table 1. The results showed that, on a %DM basis of the pressed fiber, selected samples had 1–3 % total ash, 15–35 % extractives, 26–41 % cellulose, 16–26 % hemicellulose, and 12–20 % lignin. Within the EF fraction, the ranges for cellulose, hemicellulose, and lignin were 38–46, 23–30, and 18–22 %, respectively. The combined calibration dataset, derived from whole stalk, pressed fiber, and mill bagasse, is presented in Table 2. The amount of extractives in the samples mostly depends on the type of sample and the sample processing method. It is expected that there are less extractives in a pressed fiber sample

Table 1 Composition summary statistics for 30 selected pressed fiber samples

	% Total ash (DM)	% Extractives (DM)	% Glucose (DM)	% Arabinose (DM)	% Galactose (DM)	% Xylose (DM)	% Mannose (DM)	% Hemicellulose (DM)	% ASL (DM)	% AIL (DM)	% Lignin (DM)
Max	2.6	34.8	41.2	1.9	0.8	23.9	0.1	26.5	1.2	19.1	20.4
Min	1.1	15.1	26.2	0.7	0.4	14.2	0.0	15.7	0.6	11.3	12.1
Mean	1.7	23.8	33.0	1.3	0.5	18.0	0.1	19.8	0.8	14.5	15.3
SD	0.3	5.3	3.5	0.3	0.1	2.2	0.1	2.3	0.1	1.5	1.6
N	30	30	30	30	30	30	30	30	30	30	30
			% Glucose (EF)	% Arabinose (EF)	% Galactose (EF)	% Xylose (EF)	% Mannose (EF)	% Hemicellulose (EF)	% ASL (EF)	% AIL (EF)	% Lignin (EF)
		Max	46.4	2.3	0.9	26.9	0.1	29.7	1.3	21.2	22.2
		Min	37.8	0.9	0.6	20.8	0.0	22.6	0.9	16.7	17.8
		Mean	42.6	1.6	0.7	23.2	0.1	25.5	1.1	18.7	19.8
		SD	1.8	0.4	0.1	1.3	0.1	1.4	0.1	1.0	1.0
		N	30	30	30	30	30	30	30	30	30

DM dry mass, *EF* extractives-free, *ASL* acid-soluble lignin, *AIL* acid-insoluble lignin

than in a whole stalk sample and more than in a mill bagasse sample. Therefore, the calibration set derived from three types of samples should cover well the range of expected extractives in pressed fiber samples. We found various studies reported the composition of sugarcane lignocellulosic biomass with different extractives contents due to the use of different extraction methods. As a result, we reported data on an EF basis to allow comparison of our dataset with those reported in the literature. Our calibration set had 33–48 % cellulose, 21–30 % hemicellulose, and 16–23 % lignin. This final set exhibited a wide range of extractives content (5–68%DM), due to being derived from whole stalk samples (high extractives), pressed fiber (medium to high extractives), and bagasse samples (low extractives). In general, the range in the calibration set aligned well the values reported for cellulose, hemicellulose, lignin, and extractives content. This final set of data was suitable for modeling to predict sample composition in this study.

NIR Spectroscopic Model Development and Prediction

Table 3 summarizes the predictive performance of the calibration and validation of the developed models. A good correlation between the predicted values and the reference values from chemical composition analysis was observed in both calibration and validation (Fig. 2) of the four main predicted components. The R^2 value was greater than 0.9 in all cases. The RMSEC, RMSEP, number of principal components (factors), slope and offset values allowed comparison between calibration and validation in each model. The RMSEC and RMSEP values of the lignin model were lowest amongst the four models, while the extractives model had the highest values. All in all, the performance of calibration and validation illustrates a good predictive ability of the models suggesting that they can be applied to estimate the biomass fiber composition in the unknown samples. These models were used to predict the biomass composition of each sample from the spectra collected on the XDS NIR instrument.

Figure 3 presents predicted results using the developed models for the four biomass components. This data represents the potential sugar sources (mostly from hexose-glucose and pentose-xylose from cellulose and hemicellulose) in the sample to be used for biofuel production, as well as the recalcitrant challenge (from lignin content) of the biomass source [37, 38]. The extractives, on the other hand, indicate the total amount water- and ethanol-soluble substances in the pressed fiber samples, including juice sugars not extracted during pressing. Figure 3a and Table 4 show that within the four predicted ranges, the extractives has the widest variation compared to the other three ranges. The cellulose, hemicellulose, lignin, and extractives predicted in this population were at 27–42 % (range = 15), 16–25 % (range = 9), 12–20 % (range = 8), and 3–38 % (range = 35), on a %DM basis, respectively. All data fitted within the ranges of the calibration sample set. A

Table 2 Composition summary statistics for total ~300 calibration samples (whole stalk, pressed fiber, and bagasse), all values reported on an extractives-free basis, except the extractives expressed on %DM

	Cellulose (%)	Hemicellulose (%)	Lignin (%)	Extractives (%)
Max	48.3	29.7	23.0	68.2
Min	33.4	20.6	15.9	4.6
Range	14.9	9.1	7.1	63.6
Mean	39.8	25.0	20.0	30.9
Literature range[a]	38–46	22–27	19–31	–

[a] Data adapted from [28, 29]

Pearson correlation test (Fig. 3b) indicated that lignin content was more strongly correlated with hemicellulose content (R^2 = 0.92, $p < 0.001$) than it was with the cellulose content (R^2 = 0.83, $p < 0.001$). The coefficient of determination between cellulose and hemicellulose was 0.7 ($p < 0.001$). Extractives content were negatively correlated with the lignin ($R^2 = -0.98$, $p < 0.001$), hemicellulose ($R^2 = -0.94$, $p < 0.001$), and cellulose ($R^2 = -0.83$, $p < 0.001$) contents. All frequency histograms in the Fig. 3c–f show that the distributions of the four lignocellulosic traits resemble a normal distribution.

Sugar Profiling in Juice Samples

Sugars in the juice samples were profiled by HPLC using a set of equivalent sugar standards. Compared to fiber characterization, this procedure was straightforward. Sucrose, fructose, and glucose concentrations within each sample were determined against a set of equivalent sugar standards. Table 5 summarizes the statistics of sugar analysis across the juice samples. The total sugar content (sum of sucrose, glucose, and fructose), ranged from 240 to 844 mM, of which sucrose was the largest portion (70–99 %). There were wide ranges of sucrose, glucose, and fructose values across the population, ranging from 192 to 826, 0–112, and 3–123 mM, respectively. The ratio of sucrose content to reducing sugars ranged from 2 to 71. The primary correlation analysis of these sugars (on juice weight basis) follows the same pattern as in the analysis of combined data on the total biomass, and this is discussed in the Discussion section.

Analysis of Sugarcane Total Biomass, the Relationship Between Fiber and Sugar Content

When all data obtained from fiber characterization and sugar analysis was assembled, we were able to back-calculate the composition of the biomass collected from the millable sugarcane stalk, on both a fresh biomass basis and dry mass basis. Table 6 shows a summary of statistics of sugarcane biomass from 186 genotypes after averaging of replicate samples

Table 3 Descriptive statistics for performance of calibration and cross-validation of models developed for pressed fiber major composition

Constituent	Calibration performance							
	Pretreatment	Sample type	N[a]	Factors	RMSEC	R^2	Slope	Offset
Cellulose	1SNV	B + PC + WS	445	4	2.09	0.93	0.93	1.86
Hemicellulose	2SNV	B + PC + WS	444	5	1.28	0.93	0.93	1.25
Lignin	2SNV	B + PC + WS	444	5	0.98	0.94	0.94	0.83
Extractives	2SNV	B + PC + WS	446	5	3.12	0.98	0.98	0.76
Constituent	Validation performance							
	Pretreatment	Sample type	N[b]	Factors	RMSEP	R^2	Slope	Offset
Cellulose	1SNV	B + PC + WS	597	4	1.62	0.98	0.95	1.44
Hemicellulose	2SNV	B + PC + WS	596	5	0.97	0.98	0.97	0.47
Lignin	2SNV	B + PC + WS	596	8	0.66	0.99	1.00	−0.02
Extractives	2SNV	B + PC + WS	598	5	2.71	0.99	0.99	0.18

1SNV first derivative, standard normal variate, *2SNV* second derivative, standard normal variate, *B* bagasse, *PC* pressed cane, *WS* whole stalk

[a] A few outliers were removed from the total 452 samples

[b] A few outliers were removed from the total 600 samples in each model

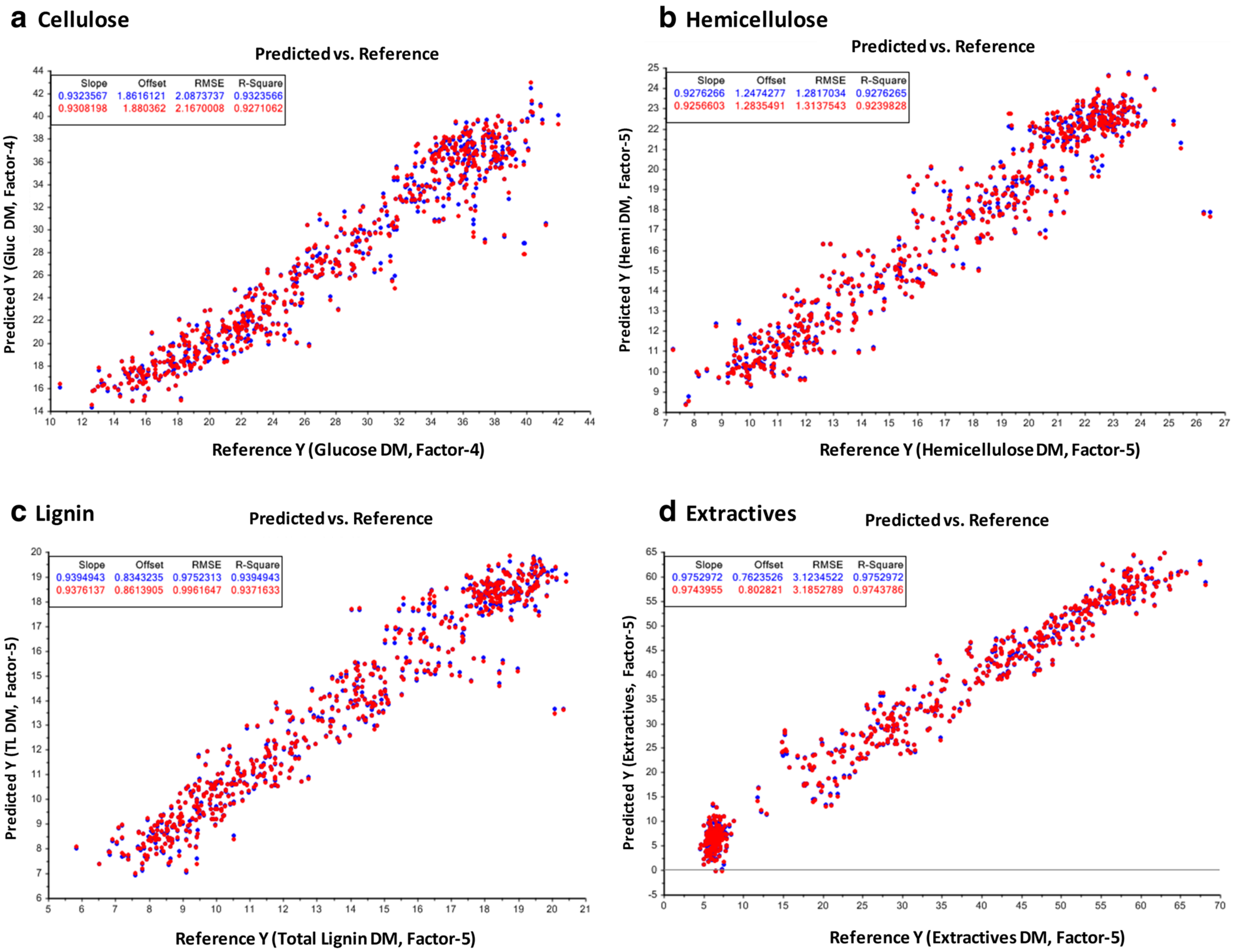

Fig. 2 Correlation of chemical composition analysis values versus predicted values of cellulose (**a**), hemicellulose (**b**), lignin (**c**), and extractives (**d**) in the cross-validation of the predictive models. The *x*-axis represents the values obtained from the chemical composition analysis while the *y*-axis represents the predicted values. The *red color* indicates calibration data points, whereas the *blue color* indicates validation data points. All values quoted on a %DM basis of pressed fiber samples (color figure online)

obtained from for each genotype. Overall, the total solids (fiber, sugars, ash, and other compounds), fiber content, and total sugar content were normally distributed (Fig. 4a, b). A general scheme of sugarcane biomass composition derived from the mature millable stalk is presented in Fig. 4c. On a total fresh biomass basis, the juice content accounted for 78–91 % of total biomass. The proportion of water and total solids was 61–78 and 22–39 %, respectively. The fiber fraction was composed of 4–9 % cellulose, 3–5 % hemicellulose, and 2–4 % lignin, and all made up 9–19 % of the total fresh biomass. Sugars were present at 6–23 %. Sucrose, being the primary sugar, was present between 6 and 22 %, whereas glucose and fructose were present at 0–1 %, each. Ash (inorganics), insolubles, and other nonsugar soluble compounds (i.e., wax, starch, phenolic compounds, protein, organic acids, etc.), from both pressed fiber and juice fractions, were estimated at 1–5 %. Figure 4d presents the biomass composition on a total dry biomass basis (total solid fraction) of the millable sugarcane stalk. The total fiber and total sugar accounted for 29–61 and 29–64 %, respectively, while an estimated fraction of ash and other compounds was 3–18 %.

A correlation analysis of the major components of total sugarcane dry biomass is presented in Fig. 5a. Sugar content showed a strong negative correlation with fiber and cellulose content (R^2 = −0.92, $p < 0.001$, for both), hemicellulose (R^2 = −0.89, $p < 0.001$), and lignin (R^2 = −0.89, $p < 0.001$). A tight correlation between total fiber and its three constituents (cellulose, hemicellulose, and lignin) (R^2 = 0.94–0.99, $p < 0.001$), and between glucose and fructose (R^2 = 0.96, $p < 0.001$), as observed previously in fiber characterization, in which predicted data were presented on a pressed fiber %DM basis, and in juice sugar analysis, respectively. There

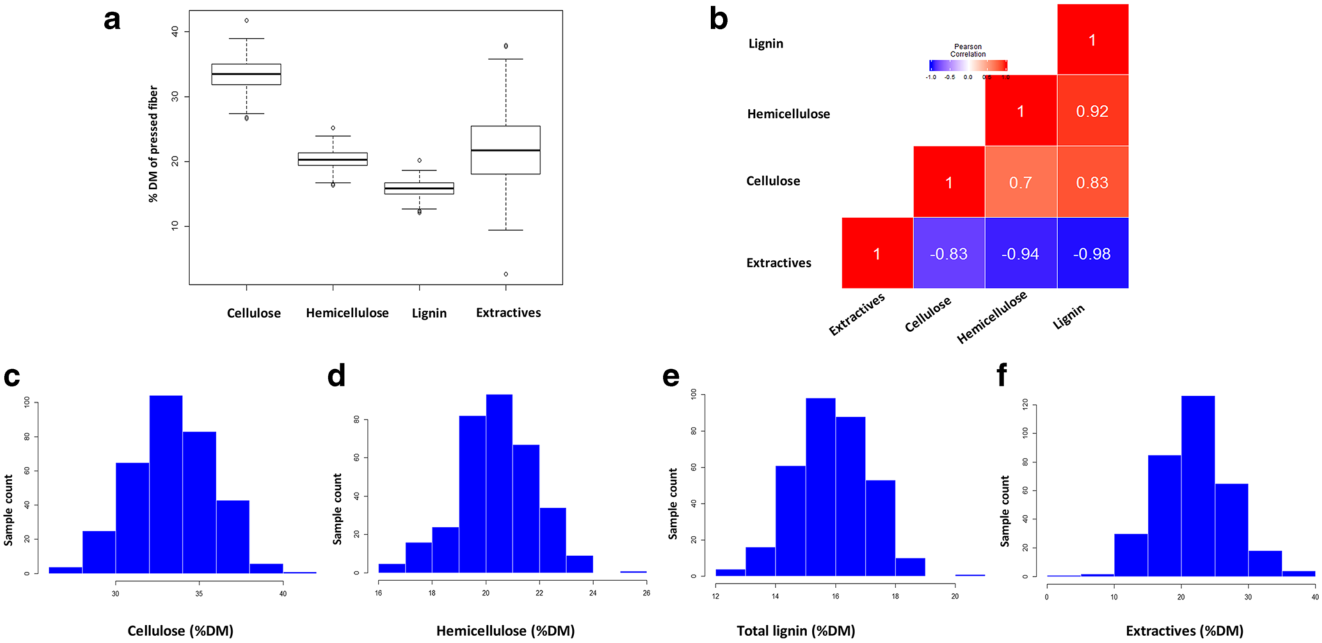

Fig. 3 Fiber characterization results of all samples from sugarcane population used in this study, predicted on the NIR spectral data of pressed fiber samples ($N = 331$). **a** Boxplot of cellulose, hemicellulose, lignin, and extractives predicted. **b** Pearson correlation analysis between predicted compositions, cellulose, hemicellulose, lignin, and extractives. The darker color is, the stronger the correlation (closer to either −1 or +1). The *red color* represents for a positive correlation while the *blue color* represents a negative correlation. **c–f** Frequency histograms for sugarcane lignocellulosic biomass composition on %DM basis of pressed fiber

was no significant correlation between sucrose and either reducing sugar (glucose or fructose), or between any of these reducing sugars and total fiber or any fiber components ($R^2 < 0.3$). The graph in Fig. 5b shows an increase in the total fiber/total sugar ratio (hereafter, referred to as fiber/sugar ratio), ranging from 0.4 to 2.2, when the fiber content increased. Based on the information obtained from this population, we identified two contrasting groups of genotypes based on fiber content and fiber/sugar ratio (highlighted in the graph). This will be important attribute and should be considered for further analysis to investigate the difference between these two groups, in terms of lignin and hemicellulose composition, and also to identify the genes associated with these differences. Figure 5c and Table 7 show a comparison between genotypes of low and high fiber content, quoted as a percentage on a total dry biomass basis. Even though there was an increase in the fiber fraction in the high fiber genotypes (mean = 55 %) compared to the low fiber genotypes (mean = 32 %), the proportion of cellulose, hemicellulose, and lignin within the fiber fraction remained the same.

Prediction of Theoretical Ethanol Yields

Calculations for 186 genotypes (presented in Table 8) showed that, for each ton (Mg) of total fresh biomass derived from millable sugarcane stalks, the theoretical fiber-based ethanol yield ranged from 51 to 107 L Mg^{-1}, while that of sugar-based ethanol was from 42 to 154 L Mg^{-1}. These added up to a combined ethanol yield from both fiber and sugars, ranging from 117 to 236 L Mg^{-1}. On a total dry biomass basis, 163–343 L of fiber-based ethanol and 193–438 L of

Table 4 Summary statistics of predicted values for cellulose, hemicellulose, lignin, and extractives, quoted on the %DM basis of pressed fiber sample

	Cellulose (%DM)	Hemicellulose (%DM)	Total lignin (%DM)	Extractives (%DM)
Max	41.7	25.2	20.2	37.9
Min	26.7	16.4	12.2	2.7
Range	15.1	8.9	8.0	35.2
Mean	33.4	20.4	15.9	21.9
SD	2.4	1.4	1.2	5.4
N	331	331	331	331

Table 5 Summary statistics for sugar analysis in juice samples

	Sucrose (mM)	Glucose (mM)	Fructose (mM)	Ratio of sucrose/ reducing sugars	Total sugar (mM)
Max	826.1	111.7	123.5	70.5	844.4
Min	192.2	0.4	2.5	2.4	239.7
Mean	548.9	19.0	18.4	21.9	586.3
SD	139.0	14.0	13.5	14.8	137.8
N	331	331	331	331	331

sugar-based ethanol could be produced from each ton of biomass. The combined ethanol yield from dry biomass was from 506 to 617 L Mg^{-1}.

Discussion

Sugarcane is considered an efficient biomass accumulator for sugar and biofuel production. Its sugar has been used for first-generation biofuels in the last 10 to 20 years, while its lignocellulosic biomass has only recently been exploited for second-generation biofuels. Sugarcane lignocellulosic biomass is derived mostly from its millable stalk and partially from leaves and green tops, estimated to be two thirds of the total sugarcane biomass produced [39–41]. Hence, developing sugarcane varieties with a desirable fiber/sugar ratio within this biomass fraction, from the current diverse germplasm collections, is essential and should be advanced in parallel

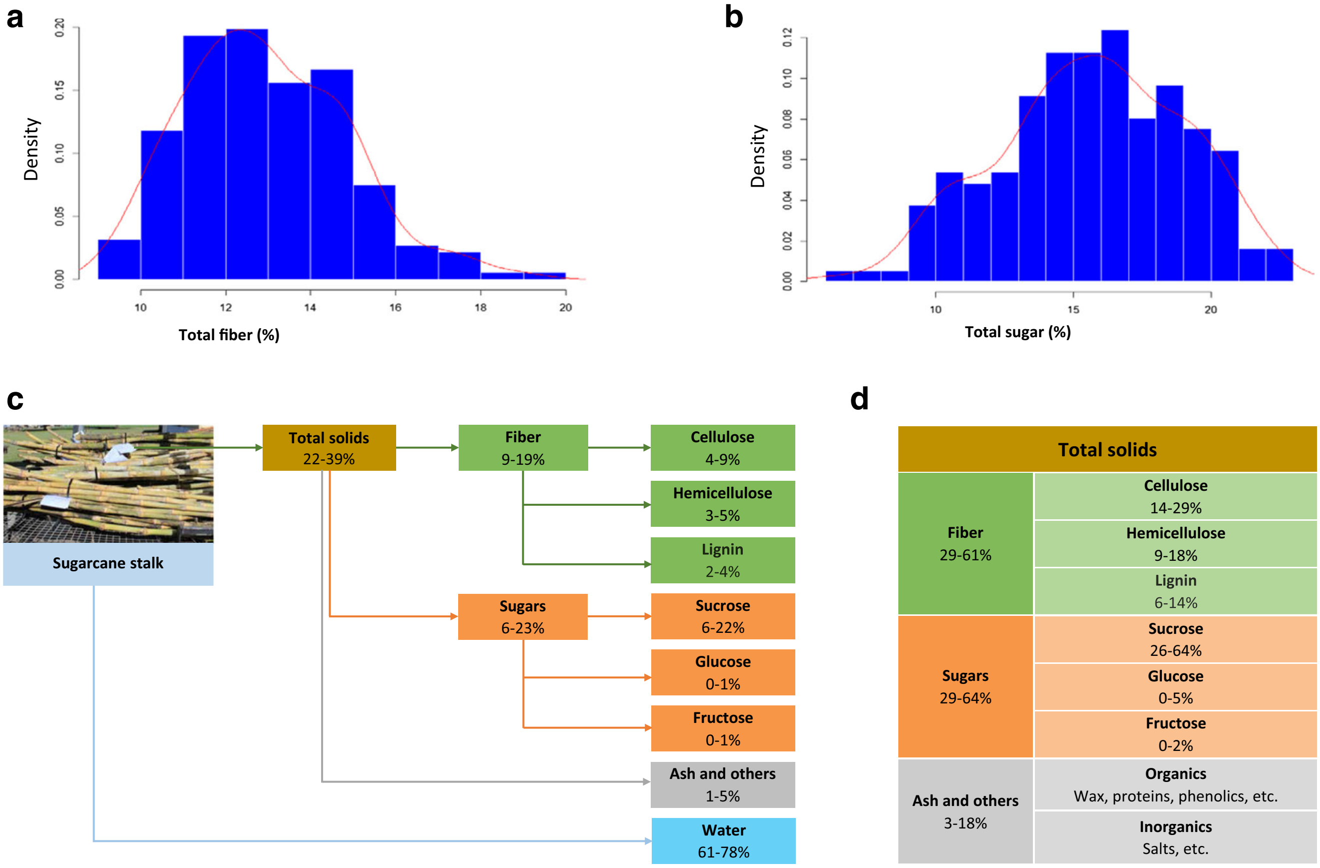

Fig. 4 Summary statistics of general composition of sugarcane total biomass from millable stalk. **a**, **b** Distribution of total fiber and total sugar, quoted on percentage of total fresh biomass. **c** General composition of sugarcane biomass, quoted on the total fresh biomass basis. **d** General composition of sugarcane biomass quoted on total dry biomass (solid fraction)

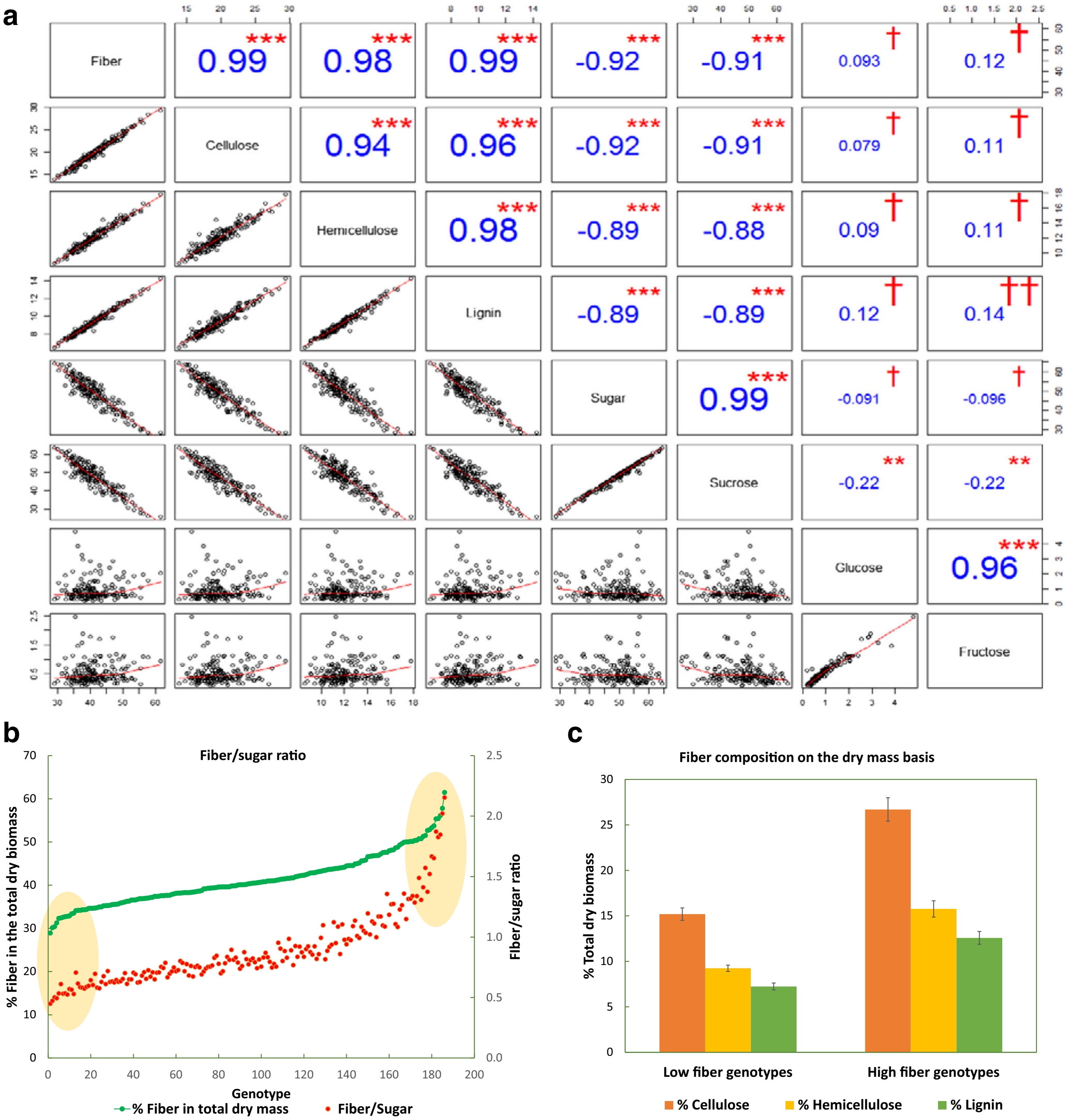

Fig. 5 Relationship between major compositions in sugarcane total dry biomass from millable stalk. **a** Correlation analyses of fiber and sugar composition. Correlation coefficient text size is scaled based on its values. $*p < 0.05$; $**p < 0.01$; $***p < 0.001$; $^{\dagger}p < 0.1$. All the values on the axis are expressed on percentage of total dry biomass. **b** Fiber/sugar ratio across the population, sorted as fiber content from low to high. Two groups of low fiber and high fiber genotypes are highlighted in the plot. **c** Comparison of fiber composition between contrasting genotypes, low and high fiber. Data of each group was averaged of ten highest and ten lowest genotypes

with biomass pretreatment and conversion technologies. We investigated a total of 331 sugarcane samples, from a population of 186 commercial hybrid and introgressed genotypes, deriving from parents of different genetic backgrounds, including *Saccharum officinarum*, *Saccharum spontaneum*, *Erianthus arundinaceus*, and *Erianthus procerus*. This study set out to analyze the sugarcane total biomass in a systematic way, in order to identify suitable germplasm to be included in genotyping, association studies, and breeding programs. The NIR- and HPLC-based methods in this study provided a

Table 6 Summary statistics of total fresh biomass composition, including juice, total solids, fiber and sugar, ash and other compounds

	% Juice	% Total solids	% Fiber	% Cellulose	% Hemicellulose	% Lignin	% Sugars	% Sucrose	% Glucose	% Fructose	% Ash and others
Max	90.9	39.0	19.0	9.4	5.4	4.3	22.7	22.5	1.2	0.6	5.28
Min	78.0	21.8	9.2	4.1	2.7	2.1	6.2	5.6	0.1	0.0	1.03
Range	12.8	17.2	9.8	5.3	2.7	2.2	16.4	16.9	1.2	0.6	4.25
Mean	85.8	31.6	12.9	6.2	3.8	2.9	15.7	15.2	0.3	0.2	2.99
SD	2.2	2.9	1.9	0.9	0.5	0.4	3.3	3.4	0.2	0.1	0.89
N	186	186	186	186	186	186	186	186	186	186	186

semidestructive, fast, and high-throughput way to collect and process the biomass data and screen large populations.

Near-infrared spectral data (400–2500 nm) was acquired for all biomass samples using the Foss XDS instrument. The peaks in the NIR spectrum provides information about all organic substances, and those associated with them, present within the biomass, which allows the quantitative assessment of biomass components of interest [42]. However, it has been suggested that the spectral regions between 400–1100 nm and 2300–2500 nm are not suited to reflectance measurement and contain high levels of noise that could interfere with the analysis of biomass compositions [16, 43]. These were removed prior to the pretreatment, and only the range of 1100–2300 nm was used for further analysis. NIR spectral data requires different pretreatment in comparison with data obtained from other platforms such as Raman or NMR [44], to remove scatter effects while maintaining a correlation between the signal and analyte concentration. The pretreatment can be first derivative, second derivative, multiplicative scatter correction (MSC), or SNV transforms [16, 45]. In this study, we focused on predicting the major composition, including cellulose, hemicellulose, lignin, and extractives, in the sugarcane biomass. Ash and other minor organic contents (accounted for 1–3 % total biomass [29]) were not predicted nor measured, with the assumption that they present at relatively constant, low levels across the samples being analyzed. These fractions can be confirmed after the information of other major constituents such as water, total dry biomass, fiber, and sugars is obtained.

In NIR spectroscopic model development, the number of samples in the calibration set is an important factor in determining the success of the prediction, and is normally

Table 7 Data of two contrasting groups of ten low and ten high fiber genotypes, data expressed on percentage of total dry biomass basis

Genotype	% Fiber	% Cellulose	% Hemicellulose	% Lignin	% Sugar	% Sucrose	% Glucose	% Fructose	% Ash and others	Fiber/sugar
Low fiber genotypes										
QN05-237	28.9	13.9	8.6	6.5	64.4	64.0	0.3	0.2	6.6	0.4
KQ08-2926	30.2	14.7	8.8	6.7	63.8	63.1	0.4	0.3	6.0	0.5
KQ08-2552	30.4	14.3	9.1	7.0	60.6	58.4	1.4	0.8	8.9	0.5
KQ08-2744	31.3	15.0	9.1	7.2	63.3	61.9	0.9	0.5	5.4	0.5
QN05-503	32.3	15.3	9.7	7.4	60.5	59.1	0.9	0.5	7.1	0.5
QC02-402	32.4	15.4	9.4	7.6	53.1	50.1	1.9	1.1	14.5	0.6
QN05-307	32.6	15.7	9.5	7.5	61.3	60.8	0.3	0.2	6.1	0.5
KQB07-24815	32.7	15.6	9.5	7.6	61.2	59.2	1.3	0.7	6.1	0.5
-2859	32.8	16.3	9.2	7.3	62.6	61.3	0.7	0.5	4.6	0.5
QS99-2014	32.9	15.7	9.6	7.6	57.7	56.7	0.6	0.4	9.4	0.6
Mean	31.7	15.2	9.2	7.2	60.9	59.5	0.9	0.5	7.5	0.5
High fiber genotypes										
KQB09-20432	51.5	25.4	14.5	11.7	32.8	30.6	1.4	0.8	15.7	1.6
QBYC05-20706	52.7	25.7	15.2	11.8	38.3	37.4	0.5	0.4	9.0	1.4
KQB07-24423	52.9	25.5	15.1	12.2	34.7	33.7	0.7	0.3	12.4	1.5
QBYN04-26171	53.3	25.8	15.2	12.3	32.0	29.3	1.8	0.9	14.7	1.7
KQB07-23162	53.8	25.9	15.7	12.1	32.5	31.7	0.5	0.3	13.7	1.7
QBYN04-26050	55.4	27.4	15.5	12.5	29.6	27.8	1.1	0.7	15.1	1.9
KQB08-22526	55.4	26.9	15.7	12.8	30.4	29.8	0.4	0.1	14.2	1.8
QB01-10021	56.1	26.5	16.6	13.0	30.4	27.3	1.9	1.1	13.5	1.8
QB01-10003	57.8	28.4	16.4	13.0	28.6	26.0	1.6	1.1	13.6	2.0
KQ08-2628	61.5	29.5	17.8	14.2	28.6	25.6	2.0	0.9	10.0	2.2
Mean	55.0	26.7	15.8	12.6	31.8	29.9	1.2	0.7	13.2	1.8

Table 8 Predicted yields of sugars, fiber, and ethanol from sugarcane total biomass, data expressed per ton of fresh biomass and dry biomass

	Cellulose (kgMg^{-1})	Hemicellulose (kgMg^{-1})	Fiber-based ethanol (LMg^{-1})	Soluble sugars (kgMg^{-1})	Sugar-based ethanol (LMg^{-1})	Combined ethanol (LMg^{-1})
On fresh biomass basis						
Max	93.6	53.9	107.0	226.7	154.2	235.8
Min	40.6	27.2	50.6	62.3	42.2	117.0
Average	62.1	37.9	72.6	156.6	106.4	179.0
N	186	186	186	186	186	186
On dry biomass basis						
Max	294.6	177.9	342.9	644.5	438.3	617.3
Min	139.0	85.9	163.2	285.5	193.3	506.2
Average	197.6	120.6	230.9	489.9	332.9	563.8
N	186	186	186	186	186	186

influenced by the natural variation of the traits. Hundreds, if not thousands, of samples are normally expected to be used for model development of key lignocellulosic components due to the complex agricultural matrix of the samples. For instance, in modeling of the cell wall composition, 232 samples were used in [16], 130 samples in [46], 160–208 samples in [28], and 228 samples in [11]. In this study, we analyzed a total collection of ~300 samples, from three sample types: whole stalk, pressed fiber, and bagasse. We treated all duplicate analyses as individual samples to maximize the number of samples, resulting in a final ~600-sample calibration set. Our results suggest that this was a reasonable number to develop good prediction models.

In evaluating the predictive performance of models, R^2 validation (R^2val) and RMSEP are compared. The R^2val value shows how well the predicted values fit with the reference values derived from chemical composition analysis, while the RMSEP is an index indicating how well the prediction model predicts the unknown samples [16, 47]. Normally, a low RMSEP and high R^2val values (close to 1), are desired. The R^2val values of the four models in this study were 0.98–0.99, which is higher than that in a corn model (0.42–0.85) [48], and a sorghum model (0.90–0.94) [46]. The RMSEP of the cellulose model was 1.62 (1.96 for corn, 1.45 for sorghum); the RMSEP of the hemicellulose model was 0.97 (1.33 for corn, 0.81 for sorghum), lignin model was 0.66 (1.49 for corn, 0.82 for sorghum), and extractives model was 2.71 (2.33 for corn model, 2.33 for sorghum model). In the two earlier studies of sugarcane stalk derived biomass, the R^2val was 0.45–0.77 and RMSEP values were 2.57, 1.99, and 1.88 (for cellulose, hemicellulose, and lignin, respectively) [11], while R^2val was 0.86–0.96, but RMSEP were not provided [28]. A higher R^2val value in this study indicates that the predicted values were well correlated with the chemical composition analysis; a good predictive performance is expected. While the cellulose and hemicellulose models had an RMSEP in a reasonable range, the lignin model had the lowest RMSEP of all models being compared, which could indicate the best predictive performance expected from it. The high RMSEP for the extractives model is likely due to the mixture of materials (bagasse, pressed fiber, and whole stalk) being used to develop the calibration models. It could also be due to the calibration set containing a large number of bagasse samples, which could weigh the models slightly. Improvements could be made by increasing the proportion of pressed fiber and whole stalk in the calibration models. The uncertainty associated with the prediction has been discussed in [46, 47, 49] which could prevent the nonsensical prediction from being reported correctly. The average deviation for our cellulose, hemicellulose, lignin, and extractives models were 1.87%DM, 1.31%DM, 1.13%DM, and 3.60%DM, respectively. As expected, the extractives model had the highest deviation, and overall, the obtained values were consistent with results reported earlier in [46, 47].

The models in this study were based on NIR spectral and chemical composition analysis data collected from different products derived from sugarcane stalk, corresponding to 3 years of sampling. This allows the models to be used in biomass composition prediction for samples with different extractives presentation, ranging from whole stalk and pressed fiber to bagasse. The population studied included wide variation ranges in cellulose, hemicellulose, and lignin content on a %DM basis of the pressed fiber. The highest cellulose, hemicellulose, and lignin contents reported in this population were 42, 25, and 20 % of the pressed fiber dry mass, which are close to the data reported in [50, 51] for energy cane (type I) which has 43 % cellulose, 24 % hemicellulose, and 22 % lignin. On the other hand, the lowest figures were 27, 16, and 12 %, respectively, which are lower than the typical composition in a commercial variety [29]. It is important to keep in mind that, in this comparison, the values on pressed fiber samples were used to compare with those in literature, since most studies reported for either pressed fiber, bagasse, or extractives-free fiber samples. For the values on a total dry biomass basis, see Fig. 4d. It is also important to clarify here that the predicted cellulose content was based on the total glucose released in the acid hydrolysis, and it was reported that the glucose (in xyloglucan) makes up about 15 % of the hemicellulose fraction in sugarcane bagasse [52] (this could be ~3 % of the total fiber). Since cellulose and hemicellulose were determined from the same hydrolysis, this result likely overestimated the cellulose content in the fiber by including the amount of hemicellulose glucose. On a total fresh biomass basis, normal distributions of major biomass components in Fig. 4a, b indicate that the calibration models and sugar assessment performed well on this sample collection. The data summarized in Fig. 4c was in agreement with the previous observations reported in literature for sugarcane total fresh biomass. The total solids, fiber, and sugars were 22–39, 9–19, and 6–23 % of total fresh biomass, respectively, while they were 18–25, 10–16, and 15–24 % in Mutton (2008) cited in [29], and 24–27, 11–16, and 10–16 % in Irvine (1977) cited in [53]. The ratio of fiber/sugar ranged from lower than that of a typical commercial variety (around 1) [51, 54] to value reported for high fiber energy cane varieties (around 2) [50]. The wide ranges of solid content and fiber/sugar ratio may be a result of the population, which was generated from diverse genetic backgrounds, including different accessions of sugarcane parental species and two *Erianthus* species.

While fermentable sugars indicate the potentials of the biomass sources for biofuels, the lignin content is considered the center of lignocellulosic biomass recalcitrance [55–57]. Lignin inhibits the acid pretreatment and enzymatic fermentation of biomass [58]. The use of low-lignin, high-fiber biomass sources are preferred in biofuel production, not only to reduce the cost of pretreatment but also to gain higher sugar yield [59]. Typical lignin content reported for some lignocellulosic biomass sources are as follows: eucalypts (27.5–32.7 %) [60], poplar (21–29 %) reviewed in [61], and sugarcane (19.1–31.4 % for extractives-free bagasse reviewed in [28], 13.2 % dry biomass basis in [62]). Our data reported lignin content ranging 12–20 % on a pressed fiber dry mass basis, 2–4 % of total fresh biomass, and 6–14 % of total dry biomass. The wide range of lignin content in this population would make a valuable source for breeding programs, since lignin content would be one of the most important selection criteria when screening germplasm to include in breeding programs for biofuels.

Understanding the relationships between components in the biomass, especially between the total solids-fiber-sugars, would potentially help to improve the sugarcane system for fiber and/or sugar production. The negative correlation between fiber and sugar, positive correlation among the fiber components, as well as the proportion of fiber components between the low and high fiber samples, are in agreement with the current knowledge about the highly regulated process of carbon partition in the sugarcane plant [63–69]. The extractives obtained in this study represented for the ethanol- and water-soluble compounds, including sugars from the unextracted juice in the pressed fiber. The negative correlation between extractives and fiber content could likely be due to the different ability to retain juice of the low and the high fiber samples. For the accuracy of fiber characterization, the extractives need to be removed prior to composition analysis [38]. The extractive content cannot be used as a trait for assessing biomass for genetic studies, since it depends on the juice extraction method used to generate pressed fiber samples.

In terms of potentials for biofuels, our calculation showed that 117–236 and 506–617 L ethanol could be produced from each fresh and dry ton of sugarcane total biomass, respectively. This result is consistent with the realistic yield from sugarcane fresh biomass (~120 L Mg^{-1}) in [70] and predicted yield for sweet sorghum fresh biomass (81–138 L Mg^{-1}) [25]. It was higher than the estimated figures reported for sugarcane dry bagasse (422 L Mg^{-1}) (http://www.afdc.energy.gov/fuels/ethanol_feedstocks.html), and switchgrass dry biomass (406.3–427.8 L Mg^{-1}) [18], both based on the US Department of Energy prediction model, which assumed that 1.7 kg hexose-equivalent sugars and 2.65 kg fiber were required to produced 1 L of sugar-based and fiber-based ethanol, respectively. The prediction model used in this study assumed that an efficient pentose-fermenting organism available, and ethanol yield was calculated separately for glucan, xylan, sucrose, glucose, and fructose in the biomass, with different conversion factors accordingly (see Materials and Methods). It is noteworthy that the fiber-based and sugar-based predicted ethanol in this study were computed from the corrected fiber and sugar fractions, in which the juice sugars remained in the pressed fiber were deducted and added back to the total sugar. Sugar-based ethanol yield represents the potential for first-generation biofuel production, while fiber-based ethanol yield represents potential second-generation biofuel production, from sugarcane system. The combined ethanol yield indicates the potential to integrate first-generation production into second-generation biofuel production, to make biofuel production more feasible on a large scale and would allow full use of the infrastructure already well-established in sugarcane production for biofuels [70]. The data provides another indicator of the value of screening and selecting sugarcane genotypes from a large collection for use in biofuel production.

Conclusions

In order to develop a suitable collection of sugarcane varieties for biofuel production, it is imperative that the key biomass traits be evaluated. The results for the analysis of 331 samples from a sugarcane population demonstrate that it is possible to use the NIR spectroscopic methods to predict sugarcane biomass composition, allowing high-throughput assessment of sugarcane germplasm, with acceptable accuracy compared to conventional wet chemistry. The employment of an NIR spectroscopic method in combination with HPLC can enable the semidestructive, rapid profiling of sugarcane biomass samples. This method is suitable for screening large populations for biomass key traits, to be included in breeding programs, generating more efficient sugarcane varieties for biofuel production.

Overall, based on the biomass compositional analyses of 186 genotypes, it is concluded that this population possesses a wide variation in total solids, fiber, and sugar content. This is a valuable resource for research on sugarcane biomass for biofuels. The identification of the contrasting genotypes, such as lowest and highest in the total sugar, total fiber, cellulose content, hemicellulose individual components, lignin S/G ratio, and saccharification yield, would allow further investigation into this type of germplasm, for example, for association studies, to discover the candidate genes that control the traits of interest. The genetic information obtained would facilitate the selection of the sugarcane varieties and modification of sugarcane biomass for biofuel production [71]. Having sugarcane biomass with less recalcitrant components would reduce the cost of pre-treatment and enzymes; and make producing biofuels from biomass more feasible on a large scale.

ASL, Acid soluble lignin; AIL, Acid-insoluble lignin; CV, Cross-validation; DM, Dry mass; EF, Extractives free; HPLC, High-performance liquid chromatography; NIR, Near-infrared spectroscopy; PLS, Partial least squares regression; psi, Pounds per square inch; R^2, Coefficient of correlation; R^2val, R^2 validation; RMSEC, Root mean standard error of calibration; RMSECV, Root mean standard error of cross-validation; RMSEP, Root mean standard error of prediction; SNV, Standard normal variate

Acknowledgments We are grateful to the Australian Agency for International Development (AusAID) for financial support through an Australian Awards Scholarship to NVH; to the Queensland Government for funding this research; and to Sugar Research Australia (SRA) for providing the sugarcane population used in this study, providing access to the calibration data set for sugarcane whole stalk and milled bagasse, and facilities for the chemical composition analysis. We thank SRA staff at the Indooroopilly laboratory and in Brandon station (Burdekin), Queensland, Australia, for helping with the sample collecting and processing; Ravi Nirmal for helping us in sample collection and transport; and Luguang Wu for assistance with sugar analysis of the juice samples.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Authors' Contributions RJH, FCB, AF, and NVH conceived and designed the experiments. AF and NVH collected the samples. NVH, LD, and ECK processed the samples and analyzed chemical composition and NIR spectral data. NVH, AF, ECK, FCB, and RJH analyzed the data. NVH prepared the first draft. All authors discussed, edited the manuscript, and approved the final manuscript.

Reference

1. Schubert C (2006) Can biofuels finally take center stage? Nat Biotech 24(7):777–784
2. Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN Jr (2008) Plants to power: bioenergy to fuel the future. Trends Plant Sci 13(8):421–429. doi:10.1016/j.tplants.2008.06.001
3. Hoang NV, Furtado A, Botha FC, Simmons BA, Henry RJ (2015) Potential for genetic improvement of sugarcane as a source of biomass for biofuels. Frontiers in Bioengineering and Biotechnology 3. doi:10.3389/fbioe.2015.00182
4. Oxley J, Fong C, Sant GG, O'Shea MG Accelerating the characterisation of sugarcane biomass using near-infrared (NIR) spectroscopic techniques. In: Proceedings of the 34th Conference of the Australian Society of Sugar Cane Technologists held at Cairns, Queensland, Australia, 1–4 May 2012, 2012. Australian Society of Sugar Cane Technologists,
5. Faostat (2015). http://faostat3fao.org/home/E. Accessed on 25 May 2015
6. van der Weijde T, Alvim Kamei CL, Torres AF, Vermerris W, Dolstra O, Visser RG, Trindade LM (2013) The potential of C4 grasses for cellulosic biofuel production. Front Plant Sci 4:107. doi:10.3389/fpls.2013.00107

7. Huang E, Aitken K, George A (2010) Association studies. In: genetics, genomics and breeding of sugarcane. Genetics, genomics and breeding of crop plants. Science Publishers, Enfield, NH:43–68. doi:10.1201/EBK1578086849-5

8 Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants. The New Phytologist 189(4):909–922. doi:10.1111/j.1469-8137.2010.03593.x

9 Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Gen 1(1):5–20. doi:10.3835/plantgenome2008.02.0089

10 Hall D, Tegstrom C, Ingvarsson PK (2010) Using association mapping to dissect the genetic basis of complex traits in plants. Briefings in Functional Genomics 9(2):157–165. doi:10.1093/bfgp/elp048

11 Sabatier D, Thuries L, Bastianelli D, Dardenne P (2012) Rapid prediction of the lignocellulosic compounds of sugarcane biomass by near infrared reflectance spectroscopy: comparing classical and independent cross-validation. J Near Infrared Spectrosc 20(3):371–385. doi:10.1255/jnirs.999

12 O'Shea M, Donald D, Simpson J, Keeffe E, Staunton S (2014) Development of laboratory near infrared (NIR) spectroscopy systems for sugar mill and refinery applications. Int Sugar J 116(1390): 748–753

13 Lupoi J, Singh S, Simmons B, Henry R (2013) Assessment of lignocellulosic biomass using analytical spectroscopy: an evolution to high-throughput techniques. Bioenerg Res:1–23. doi:10.1007/s12155-013-9352-1

14 Xiao L, Wei H, Himmel ME, Jameel H, Kelley SS (2014) NIR and Py-mbms coupled with multivariate data analysis as a high-throughput biomass characterization technique: a review. Front Plant Sci:5. doi:10.3389/fpls.2014.00388

15 Tsuchikawa S, Kobori H (2015) A review of recent application of near infrared spectroscopy to wood science and technology. J Wood Sci 61(3):213–220. doi:10.1007/s10086-015-1467-x

16 Payne CE, Wolfrum EJ (2015) Rapid analysis of composition and reactivity in cellulosic biomass feedstocks with near-infrared spectroscopy. Biotechnol Biofuels 8:43. doi:10.1186/s13068-015-0222-2

17 Huang J, Xia T, Li A, Yu B, Li Q, Tu Y, Zhang W, Yi Z, Peng L (2012) A rapid and consistent near infrared spectroscopic assay for biomass enzymatic digestibility upon various physical and chemical pretreatments in Miscanthus. Bioresour Technol 121:274–281. doi:10.1016/j.biortech.2012.06.015

18 Vogel K, Dien B, Jung H, Casler M, Masterson S, Mitchell R (2011) Quantifying actual and theoretical ethanol yields for switchgrass strains using NIRS analyses. Bioenerg Res 4(2):96–110. doi:10.1007/s12155-010-9104-4

19 Hou S, Li L (2011) Rapid characterization of woody biomass digestibility and chemical composition using near-infrared spectroscopy. J Integr Plant Biol 53(2):166–175

20 O'shea M, Staunton S, Donald D, Simpson J (2011) Proceedings of the Australian Society of Sugar Cane Technologists. In: Developing laboratory near infra-red (NIR) instruments for the analysis of sugar factory products, pp. 1–8

21 Berding N, Brotherton G, Le Brocq D, Skinner J (1991) Near infrared reflectance spectroscopy for analysis of sugarcane from clonal evaluation trials: I. Fibrated cane. Crop Sci 31(4):1017–1023

22 Brotherton G, Berding N At-line analysis of mill-prepared cane using near infrared spectroscopy. In: Proceedings-Australian Society of Sugar Cane Technologists, 1998. Watson Ferguson and Company, pp 34–42

23 Wu L, Birch RG (2007) Doubled sugar content in sugarcane plants modified to produce a sucrose isomer. Plant Biotechnol J 5(1):109–117. doi:10.1111/j.1467-7652.2006.00224.x

24 Wu L, Birch RG (2004) Characterization of Pantoea dispersa UQ68J: producer of a highly efficient sucrose isomerase for isomaltulose biosynthesis. J Appl Microbiol 97(1):93–103. doi:10.1111/j.1365-2672.2004.02274.x

25 Tew TL, Cobill RM, Richard EP (2008) Evaluation of sweet sorghum and sorghum × Sudangrass hybrids as feedstocks for ethanol production. Bioenerg Res 1(2):147–152. doi:10.1007/s12155-008-9013-y

26 NREL (2015) Standard Procedures for Biomass Compositional Analysis. http://www.nrel.gov/biomass/analytical_proceduresht ml Accessed on 05 June 2015

27 Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11(1):137–148. doi:10.2307/1266770

28 O'shea MG, Keeffe EC, Burns EM, Staunton SP (2013) An analysis of the lignocellulosic diversity of sugarcane. Proc Int Soc Sugar Cane Technol:28

29 Canilha L, Kumar Chandel A, dos Santos Milessi TS, Fernandes Antunes FA, da Costa Freitas WL, das Gracas Almeida Felipe M, da Silva SS (2012) Bioconversion of sugarcane biomass into ethanol: an overview about composition, pretreatment methods, detoxification of hydrolysates, enzymatic saccharification, and ethanol fermentation. J Biomed Biotechnol 2012:989572. doi:10.1155/2012/989572

30 Dien BS (2010) Mass balances and analytical methods for biomass pretreatment experiments. Biomass to Biofuels. Blackwell Publishing Ltd., In, pp. 213–231. doi:10.1002/9780470750025.ch11

31 Team RC (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2013 ISBN 3-900051-07-0

32 Sarkar D (2008) Lattice: multivariate data visualization with R. Springer Science & Business Media,

33 Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer Science & Business Media,

34 Wickham H (2007) Reshaping data with the reshape package. J Stat Softw 21(12):1–20

35 Harrell FE Jr (2008) Hmisc: harrell miscellaneous. R Package Version 3:4–4

36 Recology (2016) Five ways to visualize your pairwise comparisons. http://recology.info/ Accessed on 01 April 2016

37 Jung H-JG, Valdez FR, Hatfield RD, Blanchette RA (1992) Cell wall composition and degradability of forage stems following chemical and biological delignification. J Sci Food Agric 58(3):347–355. doi:10.1002/jsfa.2740580309

38 Thammasouk K, Tandjo D, Penner MH (1997) Influence of extractives on the analysis of herbaceous biomass. J Agric Food Chem 45(2):437–443. doi:10.1021/jf960401r

39 Soccol CR, Vandenberghe LPD, Medeiros ABP, Karp SG, Buckeridge M, Ramos LP, Pitarelo AP, Ferreira-Leitao V, Gottschalk LMF, Ferrara MA, Bon EPD, de Moraes LMP, Araujo JD, Torres FAG (2010) Bioethanol from lignocelluloses: status and perspectives in Brazil. Bioresour Technol 101(13):4820–4825. doi:10.1016/j.biortech.2009.11.067

40 Macrelli S, Mogensen J, Zacchi G (2012) Techno-economic evaluation of 2nd generation bioethanol production from sugar cane bagasse and leaves integrated with the sugar-based ethanol process. Biotechnology for Biofuels 5:22. doi:10.1186/1754-6834-5-22

41 Alexander AG (1985) The energy cane alternative. Elsevier Science Publishers BV,

42 Manley M (2014) Near-infrared spectroscopy and hyperspectral imaging: non-destructive analysis of biological materials. Chem Soc Rev 43(24):8200–8214. doi:10.1039/c4cs00062e

43 Xie S, Xiang B, Yu L, Deng H (2009) Tailoring noise frequency spectrum to improve NIR determinations. Talanta 80(2):895–902. doi:10.1016/j.talanta.2009.08.010

44 Lupoi JS, Gjersing E, Davis MF (2015) Evaluating lignocellulosic biomass, its derivatives, and downstream products

with Raman spectroscopy. Front Bioeng Biotechnol 3. doi:10.3389/fbioe.2015.00050

45 Laurens LML, Wolfrum EJ (2013) High-throughput quantitative biochemical characterization of algal biomass by NIR spectroscopy; multiple linear regression and multivariate linear regression analysis. J Agric Food Chem 61(50):12307–12314. doi:10.1021/jf403086f

46 Wolfrum E, Payne C, Stefaniak T, Rooney W, Dighe N, Bean B, Dahlberg J (2013) Multivariate calibration models for sorghum composition using near-infrared spectroscopy. Contract 303:275–3000

47 Templeton DW, Scarlata CJ, Sluiter JB, Wolfrum EJ (2010) Compositional analysis of lignocellulosic feedstocks. 2. Method uncertainties. J Agric Food Chem 58(16):9054–9062. doi:10.1021/jf100807b

48 Wolfrum EJ, Sluiter AD (2009) Improved multivariate calibration models for corn stover feedstock and dilute-acid pretreated corn stover. Cellulose 16(4):567–576. doi:10.1007/s10570-009-9320-2

49 Bu D, Wan B, McGeorge G (2013) A discussion on the use of prediction uncertainty estimation of NIR data in partial least squares for quantitative pharmaceutical tablet assay methods. Chemom Intell Lab Syst 120:84–91. doi:10.1016/j.chemolab.2012.11.005

50 Kim M, Day DF (2010) Composition of sugar cane, energy cane, and sweet sorghum suitable for ethanol production at Louisiana sugar mills. J Ind Microbiol Biotechnol 38(7):803–807. doi:10.1007/s10295-010-0812-8

51 Tew T, Cobill R (2008) Genetic improvement of sugarcane (Saccharum spp.) as an energy crop. In: Vermerris W (ed) Genetic improvement of bioenergy crops. Springer, New York, pp. 273–294. doi:10.1007/978-0-387-70805-8_9

52 Canilha L, Rodrigues RdCLB, Antunes FAF, Chandel AK, Milessi TSdS, Felipe MdGA, Silva SSd (2013) Bioconversion of hemicellulose from sugarcane biomass into sustainable products. Sustainable degradation of lignocellulosic biomass—techniques, applications and commercialization. doi:44267

53 Clarke MA, Godshall MA (1988) Chemistry and processing of sugarbeet and sugarcane: proceedings of the Symposium on the Chemistry and Processing of Sugarbeet, Denver, Colorado, April 6, 1987 and the Symposium on the Chemistry and Processing of Sugarcane, New Orleans, Louisiana, September 3–4, 1987. Elsevier,

54 Botha F (2009) Energy yield and cost in a sugarcane biomass system. Proc Aust Soc Sugarcane Tech:1–9

55 Keating JD, Panganiban C, Mansfield SD (2006) Tolerance and adaptation of ethanologenic yeasts to lignocellulosic inhibitory compounds. Biotechnol Bioeng 93(6):1196–1206. doi:10.1002/bit.20838

56 Torres AF, van der Weijde T, Dolstra O, Visser RGF, Trindade LM (2013) Effect of maize biomass composition on the optimization of dilute-acid pretreatments and enzymatic saccharification. BioEnergy Res 6. doi:10.1007/s12155-013-9337-0

57 Himmel M, Ding S, Johnson D, Adney W, Nimlos M, Brady J, Foust T (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315:804–807

58 Chen F, Dixon RA (2007) Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25(7):759–761. doi:10.1038/nbt1316

59 Bose SK, Francis RC, Govender M, Bush T, Spark A (2009) Lignin content versus syringyl to guaiacyl ratio amongst poplars. Bioresour Technol 100(4):1628–1633. doi:10.1016/j.biortech.2008.08.046

60 Gomide JL, Colodette JL, Oliveira RC, Silva CM (2005) Technological characterization of the new generation of eucalyptus clones in Brazil for Kraft pulp production. Revista Árvore 29(1):129–137

61 Sannigrahi P, Ragauskas AJ, Tuskan GA (2010) Poplar as a feedstock for biofuels: a review of compositional characteristics. Biofuels Bioprod Biorefin 4(2):209–226

62 Pereira SC, Maehara L, Machado CMM, Farinas CS (2015) 2G ethanol from the whole sugarcane lignocellulosic biomass. Biotechnol Biofuels 8:44. doi:10.1186/s13068-015-0224-0

63 Gibeaut DM (2000) Nucleotide sugars and glycosyltransferases for synthesis of cell wall matrix polysaccharides. Plant Physiol Biochem 38(1–2):69–80. doi:10.1016/S0981-9428(00)00167-4

64 Joshi CP, Mansfield SD (2007) The cellulose paradox–simple molecule, complex biosynthesis. Curr Opin Plant Biol 10(3):220–226. doi:10.1016/j.pbi.2007.04.013

65 Botha FC, Black KG (2000) Sucrose phosphate synthase and sucrose synthase activity during maturation of internodal tissue in sugarcane. Funct Plant Biol 27(1):81–85

66 Botha F, Whittaker A, Vorster D, Black K (1996) Sucrose accumulation rate, carbon partitioning and expression of key enzyme activities in sugarcane stem tissue. Sugarcane: Research Towards Efficient and Sustainable Production:98–101

67 Whittaker A, Botha FC (1997) Carbon partitioning during sucrose accumulation in sugarcane internodal tissue. Plant Physiol 115(4):1651–1659

68 Carpita NC (1996) Structure and biogenesis of the cell walls of grasses. Annu Rev Plant Physiol Plant Mol Biol 47:445–476. doi:10.1146/annurev.arplant.47.1.445

69 Lin F, Waters CL, Mallinson RG, Lobban LL, Bartley LE (2015) Relationships between biomass composition and liquid products formed via pyrolysis. Frontiers in Energy Research: 3. doi:10.3389/fenrg.2015.00045

70 Benjamin Y, García-Aparicio MP, Görgens JF (2014) Impact of cultivar selection and process optimization on ethanol yield from different varieties of sugarcane. Biotechnology for Biofuels 7(1):1–17. doi:10.1186/1754-6834-7-60

71 Furtado A, Lupoi JS, Hoang NV, Healey A, Singh S, Simmons BA, Henry RJ (2014) Modifying plants for biofuel and biomaterial production. Plant Biotechnol J 12(9):1246–1258. doi:10.1111/pbi.12300