Journal of Chromatography A, 1217 (2010) 1963–1970

Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Quantitative assessment of moisture damage for cacao bean quality using two-dimensional gas chromatography combined with time-of-flight mass spectrometry and chemometrics

Elizabeth M. Humston[a], Joshua D. Knowles[b], Andrew McShea[c], Robert E. Synovec[a,*]

[a] Department of Chemistry, Box 351700, University of Washington, Seattle, WA 98195, USA
[b] School of Computer Science, Manchester Interdisciplinary Biocentre, The University of Manchester, Manchester M1 7ND, UK
[c] Theo Chocolate R & D, 3400 Phinney Avenue North, Seattle, WA 98103, USA

ARTICLE INFO

Article history:
Received 24 October 2009
Received in revised form 19 January 2010
Accepted 22 January 2010
Available online 29 January 2010

Keywords:
GC × GC–TOFMS
Chemometrics
Cacao
Quality control
Chocolate
Food safety

ABSTRACT

Quality control of cacao beans is a significant issue in the chocolate industry. In this report, we describe how moisture damage to cacao beans alters the volatile chemical signature of the beans in a way that can be tracked quantitatively over time. The chemical signature of the beans is monitored via sampling the headspace of the vapor above a given bean sample. Headspace vapor sampled with solid-phase microextraction (SPME) was detected and analyzed with comprehensive two-dimensional gas chromatography combined with time-of-flight mass spectrometry (GC × GC–TOFMS). Cacao beans from six geographical origins (Costa Rica, Ghana, Ivory Coast, Venezuela, Ecuador, and Panama) were analyzed. Twenty-nine analytes that change in concentration levels via the time-dependent moisture damage process were measured using chemometric software. Biomarker analytes that were independent of geographical origin were found. Furthermore, prediction algorithms were used to demonstrate that moisture damage could be verified before there were visible signs of mold by analyzing subsets of the 29 analytes. Thus, a quantitative approach to quality screening related to the identification of moisture damage in the absence of visible mold is presented.

1. Introduction

Quality control and food safety are critical concerns of food manufacturers, governments and consumers that have especially gained attention with a number of food borne illness outbreaks in recent years. Because of the processing in cacao bean manufacturing, including heat treatment and the removal of excess moisture, a food borne illness from chocolate products is comparatively less likely, but the cocoa, candy and chocolate industry still faces the same challenge of ensuring that raw materials, including cacao beans, are of high quality and safe. Cacao beans are spontaneously fermented seeds, and as such are subject to a high level of variability depending on growing conditions, genetics, postharvest fermentation and drying of the cocoa beans prior to shipment or handling. This variability can have a large impact on the finished chocolate product, so it is important to be able to determine if beans are properly fermented, of high quality, and lacking defects. In addition to trying to distinguish properly fermented beans and maintain a high quality raw material, there are also the food safety concerns. There is the potential for a variety of moisture damage related chemical processes occurring due to high humidity, etc., that can result in mold colonization and/or microbial growth during storage and transport, or the unintentional inclusion of other chemical alterations (or adulterants) to the raw material itself. While the presence of chemical degradation products due to moisture damage does not always indicate an unsafe product [1], it can often introduce unpleasant off-flavors in the finished chocolate product. Thus, moisture damage of cacao beans is both a food safety and quality control concern for this industry.

The current quality control methodology used for monitoring cacao is quite subjective and limited in its ability to quantify bean quality. Typically, a "cut-test" is performed on 50–100 (or more) representative beans. Beans are cut into half to expose the core for examination so that the color, which is an indicator of the state of fermentation, can be observed. Additionally, a small sample of beans is often roasted and ground for a taste test. Though cut-test results are generally reproducible with a significant margin of error, the results of taste tests are much harder to agree upon and communicate. Given that these methods are the current state-of-the-art, a more reliable and quantitative methodology would be quite useful in this industry. In particular, a method to ascertain the likely emergence and presence of chemical products due to

* Corresponding author. Tel.: +1 206 685 2328; fax: +1 206 685 8665.
E-mail address: synovec@chem.washington.edu (R.E. Synovec).

0021-9673/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2010.01.069

moisture damage could be very useful as this is a common spoilage mechanism. In addition, a method that could also identify toxic contaminants or adulterants during the same measurement procedure could improve product safety.

It is known that flavor and product quality are closely associated with the volatile profile of chocolate [2,3]. In order to improve the finished product, it would be useful to gain an understanding of how compounds in raw cacao beans are indicative of a high quality chocolate bar. Many chemical alteration processes, including mold, may also have a volatile profile that could possibly be detected. Because many of the compounds that do relate to the odor and taste of chocolate are volatile compounds, techniques that sample the headspace above a cacao bean sample are ideally suited for routine monitoring. Headspace solid-phase micro-extraction (HS-SPME) is a sampling technique that isolates volatile and semi-volatile compounds from a headspace gas sample by collecting the analytes onto a coated fiber [4]. We have recently reported implementations of this technique to cacao beans [5] and it has also been shown effective on cocoa powder [6]. The sampling method can be paired with comprehensive two-dimensional gas chromatography coupled to time-of-flight mass spectrometry (GC × GC–TOFMS) for separation and detection of the complex sample types (quantitative chemical signatures). This powerful analytical technique separates the complex sample into two dimensions by combining two columns with complementary stationary phases [7–12]. The combination of GC × GC with TOFMS detection has been successful at separating and detecting many complex sample types [13–26].

Implementation of GC × GC–TOFMS instrumental methodology converts complex sample types to complex raw data. Chemometric techniques are essential for extracting useful information from the complex data [27]. One common goal, in hypothesis driven studies such as this, is to identify features (i.e., analytes) that offer chemical selectivity to confidently distinguish the sample types. One such approach for identifying these analytes of interest is the Fisher Ratio (F-Ratio) algorithm [23]. This algorithm finds class-type distinguishing compounds, in the presence of biological variation, which become the focus for further investigation. Parallel Factor Analysis (PARAFAC) is another chemometric algorithm that can be used to mathematically resolve analyte(s) of interest from background noise and overlapping interferences (other compounds) thus providing quantitative information [28–31]. Further comparison of the samples can be made using pattern recognition and regression analysis techniques.

We previously observed dramatic quantifiable differences in the volatile profile of cacao beans depending on moisture damage relating to the presence or absence of surface mold [5]. By the time mold is visible, however, a simple visual inspection would make it quite clear that the beans have been compromised without a need for sampling the headspace analytes. For routine quality screening, it would be more useful to quantitatively detect changes in the headspace analytes that indicate moisture damage prior to visible mold. As mold is both a food safety and quality control concern, this would be an important capability. In this study, we have monitored the changes that occur in the headspace analytes as beans deteriorate from no visible mold to essentially a 100% surface mold coverage. This time course data was determined for cacao beans from six geographical origins in order to identify consistent chemical changes related to food safety and bean quality that may be origin independent (i.e., independent of bean variety). The specific compounds that relate to the origin-independent chemical changes could potentially be used as routine biomarkers for moisture damage. Cacao beans from six origins were intentionally subjected to moisture damage and sampled over the course of approximately 1 week. Essentially, the chemical composition of the bean surface changes due to the moisture damage process, and the volatile and semi-volatile components provide a chemical signature that can be readily sampled using HS-SPME. Thus, HS-SPME was used in conjunction with GC × GC–TOFMS for data collection. Class distinguishing analytes were located with F-Ratio analysis and then quantified with PARAFAC. These results were further compared with principal component analysis (PCA) and various regression techniques (e.g., CART and random forests) in order to demonstrate that moisture damage could be detected prior to visible mold growth, hence to provide predictive capability in a timely fashion.

2. Experimental

2.1. Sample preparation

Cacao beans from six geographical origins (Costa Rica, Ghana, Ivory Coast, Venezuela, Ecuador, and Panama) were acquired by Theo Chocolate (Seattle, WA, USA). A stock sample of raw beans from each origin was stored under cool and dry conditions in order to preserve the bean quality prior to assessing the impact of moisture damage, defined as 0% mold coverage. Filtered water was added to a subset of these stock samples and the beans were allowed to mold to what visually appeared to be total external coverage (100% coverage). This provided for reference, stock bean samples at 0% and 100% coverage. These qualitative definitions are indicative of what can be observed by eye without microscopic magnification (as in a field assessment), thus not representative of microscopic mold spores that may be present. For each geographical origin, 18 representative beans were taken from the 0% coverage stock sample and placed in sealable plastic bag to which 10 ml of filtered water was added. To ensure that the total time with added moisture was consistent for each sample at the time of analysis, the addition of water was staggered by 1 h for each origin to compensate for analysis time. The beans were stored at room temperature and samples, i.e., three representative beans per sample, were analyzed from the 0% coverage stock and then from the sealable bag for a total of 7 time points (0, 1, 2, 3, 4, 5, and 6 days) over the course of the moisture damage process. One additional replicate from the 100% coverage stock samples (at ~1 month) was also collected for beans from each origin.

2.2. Solid-phase micro-extraction (SPME)

The SPME procedure previously described was used for this study [5]. Briefly, a 65 μm PDMS/DVB SPME fiber (Supelco, PA, USA) served to preconcentrate headspace analytes above a cacao bean sample. The fiber was conditioned at 250 °C for 30 min prior to each sample extraction. At the given time (0, 1, 2, 3, 4, 5, and 6 days), three cacao beans were removed from the sealable bag of a particular origin and sealed together in a new 15 ml SPME vial. The origins were sampled in the same order that water was added so each bean sample was analyzed at the same total time since the water had been added. For the sample preparation via HS-SPME, each sample was heated in a water bath to 60 °C for 15 min, after which the SPME fiber was exposed to the headspace for 10 min. After being extracted, beans were not returned to the bag as the extraction alters the bean.

2.3. GC instrument parameters

GC instrument parameters were also maintained as previously described [5]. A GC × GC–TOFMS consisting of an Agilent 6890N GC (Agilent Technologies, CA, USA) and a thermal modulator (4D upgrade, LECO, St. Joseph, MI, USA) paired with a Pegasus III TOFMS (LECO, St. Joseph, MI, USA) was used to separate the HS-SPME sampled analytes. The SPME fiber was introduced to the GC inlet, maintained at 250 °C with a constant He flow of 1 ml/min, for

5 min as the injection. A GC × GC column arrangement of non-polar (20 m × 250 μm i.d. × 0.5 μm RTX-5MS (Restek, PA, USA)) to polar (2 m × 180 μm i.d. × 0.2 μm RTX-200MS (Restek, PA, USA)) was implemented. For the first 5 min, during injection, the first column was held at 40 °C and the second column at 50 °C. Both columns then followed a temperature program that ramped at a rate of 8 °C/min from 40 °C to 140 °C for the first column and 50 °C to 150 °C for the second column. The rate was then increased to 30 °C/min to a final temperature of 250 °C for the first column and 260 °C for the second, where the columns were held constant for an additional min. The modulator temperature was maintained 20 °C higher than the temperature of column one and transferred the effluent every 1.5 s. The transfer line was held at 280 °C and the TOFMS ion source at 250 °C. Mass channels 40–250 m/z were collected and stored at a rate of 100 spectra/s. Three beans were combined together and analyzed at each time point (7 plus an additional 100% coverage stock sample) for each origin (6 total) for a total of 48 chromatographic injections. Based on our prior study [5], replicates at each time point and bean origin were not deemed necessary, as will be further discussed.

2.4. Data analysis

Raw chromatographic data were collected using LECO ChromaTOF software v 3.32 (LECO, St. Joseph, MI, USA). Data were exported to Matlab for Fisher Ratio (F-Ratio) calculations. The data collected from time 0 days (i.e., 0% coverage) and from the 100% coverage stock samples from each origin were used as the two sample class-types for F-Ratio analysis [23] in order to find the compounds that were up or down regulated due to the moisture damage process. F-Ratios were calculated both by weighting to chromatographic intensity (weighted) and without weighting (unweighted). At the top locations identified through F-Ratio analysis, preliminary analyte identification was determined with ChromaTOF software via searching the mass spectra from the chromatograms against the National Institute of Standards and Technology (NIST) library. An in-house developed target-analyte Parallel Factor Analysis (PARAFAC) Graphical User Interface (GUI) was used to mathematically resolve the pure peak profile and mass spectra for quantitative purposes [30]. Quantitative information was obtained across the complete time course for all bean origins.

2.5. Data interpretation

Principal Component Analysis (PCA) was employed as a data comparison tool as we have previously demonstrated with metabolomics data [19,22]. Each analyte was loaded as a sample with the time course information across each origin as the variables. PCA was then calculated using preprocessing of mean centered data. Additional modeling software was used for regression methods. For CART and random forests, the R statistical software package was employed [32]. The standard implementations in WEKA version 3.5.7 [33] were applied for all other regression methods. For CART, the rpart library [34] was used and, following standard procedures, a complexity parameter of cp = 0.05 was selected. For random forests, the random forest library [35] was used, and models with the default of 500 trees were trained. 10-fold cross-validation [36] for these two methods was coded in-house and results of the mean and standard deviation from repeating the whole validation 10 times are reported herein. For the methods run in WEKA, default parameter settings were used in all cases except for the Radial Basis Function Network. For that model, eight basis functions were selected (instead of the default two), one to represent each time data point. WEKA performs 10-fold cross-validation as a default testing method. The coefficient of determination (R^2 value) is a measure of the fraction of variation in a dependent variable that is explained by a model [37]. In all cases here, it was calculated as the square of the Pearson correlation coefficient between the prediction by the models and the actual predicted variable value.

Fig. 1. This picture indicates beans at various stages of moisture damage, visibly indicated by mold, from 0 days to 100% coverage.

3. Results and discussion

Beans from various geographical origins were studied: Costa Rica, Ghana, Ivory Coast, Venezuela, Panama, and Ecuador. Beans were sampled and monitored 7 times over the course of approximately 1 week (0 days to 6 days by a 1-day interval), as well as one additional measurement from a stock molded sample at ~1 month of moisture damage (i.e., 100% coverage) for a total of 8 data points across the time course moisture damage process for each origin class. Again, the expression of the moisture damage was visualized by the appearance of mold, while other less visible chemical degradation processes were also occurring. There was some variability in the appearance and rate of mold growth, but all beans reached what appeared to be 100% coverage by the end of the 6-day period. Beans did not convert from an absence of visible mold to 100% coverage from 1 day to the next. Instead, mold appeared in a localized spot and gradually grew to cover the entire bean. Fig. 1 illustrates a recreation of the moisture damage process via mold expression. As there was some variability, the most representative beans were sampled at each time point. We have previously observed only small bean-to-bean variability for a given origin [5]. We evaluated the reproducibility of the extraction and injection of 5 separate beans from the same origin for 8 samples (4 origins at 2 moisture conditions each) and observed an average %RSD of only 16.2% in the TIC (acceptable for biological variation). However, to compensate for the small amount of variability that may be present between beans, three beans were selected and sampled together for each analysis. In essence, the samples were averaged prior to analysis and information on the average headspace of three beans was provided at each time point.

SPME sampling coupled with GC × GC–TOFMS separation and detection effectively converted these complex sample types into complex data. Representative 2D TIC chromatograms of the beginning and the end of the moisture damage process are shown in Fig. 2, in which wrap around can be observed. This could be corrected by making changes to the modulation period, however the wrap around more fully utilizes the 2D peak capacity, and is not problematic for software analysis. It is possible to visibly identify some of the chemical signature differences between the beginning and the end of the moisture damage process as shown in the 2D TIC separations in Fig. 2. However, chemometric techniques are more useful than visualization to more thoroughly probe this complex data for useful information. The F-Ratio algorithm can be used to find analytes that differ between classes; in this case between the unmolded samples and those with 100% coverage, across all bean origin classes. All of the origins were included for each sample class so that any origin differences would be counted as within class variation and the differences identified would likely be origin independent. The F-Ratio algorithm can be calculated in both

Table 1
F-Ratio results including the combined list of the top 20 analytes found by weighted and unweighted F-Ratio analysis. The calculated F-Ratios are provided in the columns F-Ratio W (weighted) and F-Ratio U (unweighted) and the order of each as Rank W and Rank U. The analyte identification (determined through mass spectral matching) is listed with observed retention times in s (tR1 and tR2) and match value to library spectra.

Rank W	Rank U	F-Ratio W	F-Ratio U	tR1	tR2	Analyte	MV
1	7	1.60E+08	4.00E+03	156	0.91	Acetic acid	970
2	2	1.40E+08	7.60E+03	898.5	0.58	Tetramethyl-pyrazine	901
3		7.30E+07		94.5	1.45	Carbon dioxide	992
4	14	1.50E+07	2.40E+03	151.5	0.61	Methyl butenol	899
5		1.30E+07		1125	0.10	Unknown	
6	1	1.20E+07	9.30E+03	1089	0.46	Nonanoic acid	844
7	6	1.20E+07	5.20E+03	754.5	0.74	4-Hydroxy-benzenesulfonic acid	928
8	5	1.00E+07	5.20E+03	543	1.02	3-Methyl-butanoic acid	895
9		9.60E+06		1156.5	0.05	2,6,10-Trimethyl-dodecane	900
10	4	9.20E+06	5.60E+03	751.5	0.83	Hexanoic acid	913
11		9.00E+06		423	0.15	2,3-Butanediol	932
12		8.70E+06		441	0.05	2,3-Butanediol, [S-(R*,R*)]	940
13		7.20E+06		916.5	1.02	Nonanal	801
14		5.70E+06		820.5	0.67	2-Ethyl-1-hexanol	921
15	3	5.70E+06	5.70E+03	873	0.76	Heptanoic acid	876
16		5.60E+06		1095	0.19	2,3,7-Trimethyl-octane	894
17	11	5.20E+06	2.60E+03	787.5	0.74	Trimethyl-pyrazine	920
18		4.40E+06		1167	0.04	Hexadecane	924
19		3.70E+06		1203	1.50	Mercaptoacetic acid	750
20		3.50E+06		1114.5	0.11	Pentadecane	919
	8		3.80E+03	1090.5	0.82	2-Decenal	879
	9		3.00E+03	562.5	0.85	2-Methyl-butanoic acid	785
	10		2.80E+03	1068	0.57	Unknown	
	12		2.50E+03	1089	0.67	2-Phenylethyl ester acetic acid	887
	13		2.50E+03	1056	0.65	4-(Prop-2-enoyloxy) octane	877
	15		2.20E+03	897	0.76	α,α-Dimethyl-benzenemethanol	817
	16		2.20E+03	436.5	1.10	Butanoic acid	860
	17		2.20E+03	988.5	0.71	Octanoic acid	855
	18		2.10E+03	1075.5	0.54	2-Ethyl-2,3,3-trimethyl-butanoic acid	734
	19		1.90E+03	1318.5	0.35	Isobutyl phthalate	806
	20		1.80E+03	699	0.89	4-Methyl-pentanoic acid	822

a weighted and unweighted mode. In the weighted mode, the calculated F-Ratio is scaled with the analyte signal. This is a benefit in that false positive noise hits can be reduced, but it may also over emphasize the analytes that have the largest intensities. The unweighted mode does no analyte scaling so small peaks with large differences are not under represented in the results. As a long term goal is to select analytes for routine quantitative analysis that are indicators of food quality and safety, it may be beneficial to screen for more intense, thus easier to detect, analytes with the weighted approach. However, at this discovery stage, the inclusion of less intense analytes is also worthwhile as part of the chemical fingerprint for the determination of pattern changes, so F-Ratios were calculated with both the weighted and unweighted approaches. A combined list was compiled of the top analytes determined by each approach, with results provided in Table 1. Both the weighted and unweighted F-Ratio ranks are listed. There is overlap between the two lists with nearly half (nine analytes) of the analytes on each list also found with the other method. Analyte identification was determined through matching to library standards with match values provided to indicate the confidence in the identification. Analytes that matched below 700 are listed as unknowns; each approach found one unknown analyte in the top 20. The combined list of the top 20 analytes resulted in a total of 29 identified (through mass spectral matching) analytes that were quantified for further investigation. Many of these analyte compounds are routinely observed in the sampling of cacao beans [2,3]. Since the primary goal of this study is to determine if a set of analytes can be used as a chemical fingerprint to predict moisture damage, it was not deemed necessary at this stage to more rigorously identify the specific compounds via retention time matching with standards. As these analytes have potential of being origin-independent markers of moisture damage, retention time verification for a more confident identification in conjunction with the mass spectral matching may be warranted for routine screening.

Quantitative information was determined for each of the 29 analytes across the entire time course by utilizing the target PARAFAC algorithm [30]. This allowed for the determination of relative changes over the moisture damage process per analyte. For ease of visualization, the PARAFAC signal volumes were normalized to the mean for each particular analyte and plotted on a single heat map, shown in Fig. 3. The plot contains data for the entire time course study as well as the additional 100% coverage data point. The time course trends for many of these analytes track with the appearance of mold and most of the trends appeared to be origin independent. This suggests that these changes are due to the moisture damage rather than random bean-to-bean or origin to origin variability. The additional 100% coverage data point (~1 month) follows closely with the final time point (6 days) in the time course data providing an additional indication of consistency. Further normalization could be done, if needed, to account for extraction and injection variability with the introduction of an internal standard. The sample mass (sum of all three beans) could also be corrected for with normalization, if necessary. However, the average mass of the three beans for all 48 injections in this study was sufficiently similar for this bioanalytical study (3.97 g with 21% RSD.) If this were not the case, normalization should be performed to account for the variability.

As shown in Fig. 3, several analytes track with the appearance of mold. Some analytes increase after moisture damage and others decrease. There appears to be a complex relationship between a large set of analytes with many correlated trends. We have previously found that PCA can be employed as a data comparison tool to summarize the relationship between specific analytes and the samples [19,22]. In this study, PCA is being applied in a supervised mode, using PARAFAC signal volume data identified by the F-Ratio analysis. The PARAFAC signal volume data was loaded for PCA analysis with each analyte as a sample and the time course and origin information contained as the variables. This approach

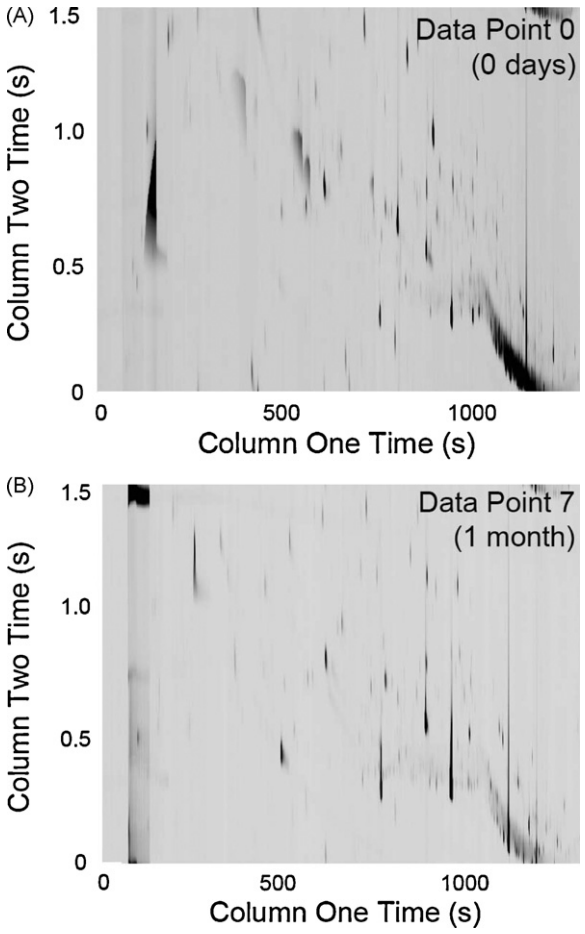

Fig. 2. Raw data from the Costa Rica sample. Differences can be identified visually between data point 0 (0 days) and data point 7 (1 month, defined as 100% coverage).

to PCA provides information on which analytes are most similar to one another in the context of the time course in the scores plot and in the loadings, which time points are similar to each other in the context of these 29 analytes. Initially PCA was performed on each origin independently as can be seen in Fig. 4A–F. In all instances, the first 2 PCs capture at least 91.1% of the variance and as much as 98.6%. The variables in this case relate to a time course process, so it is not surprising that the loadings form a continuum from one end to the other. For most origins, there is a fairly distinct transition in the middle of the continuum that roughly coincides with a qualitative observation of significant mold growth. For example, for the E sample (Ecuador) the transition point was 3 days.

In addition to looking at PCA results for each origin independently, PCA was also performed on all origins simultaneously. The results of PCA on all origins combined are provided in Fig. 5. In Fig. 5A, the first PC is plot against the sample number. The samples are first ordered by origin and then over the 8 data points (time course, plus 100% coverage from stock molded.) PC1 captures 53.9% of the variance and is primarily an indicator of the differences between the beans when they are unmolded. Nearly all of the time points correlated at or approaching 100% coverage have low PC1 loadings. Fig. 5B shows the second PC plot against the sample numbers. PC2 captures 41.0% of the total variance and appears to correlate to the variation that occurs after the samples have visible mold growth. In this case, nearly all of the initial time points have very low loadings on PC2 while the later time points have larger PC2 loadings. The first 2 PCs combine to capture nearly 95% of the total variance in the data, which is similar to the variance captured when each origin was calculated independently, and are plotted against each other in Fig. 5C. The differences identified between beans with PCA are related to the presence or absence of mold and not the origin of the bean. Crossing from positive to negative on PC1 primarily coincides with the presence or absence of visible mold with PC1 showing the variations between the initial time points and PC2 showing the variations between the later time points. The time point at which mold first became visible by eye is indicated in the Fig. 5C scores plot and occurs close to the (0, 0) coordinate for most origins. This demonstrates that by using PCA with this set of analytes, it is possible to track the appearance of moisture damage.

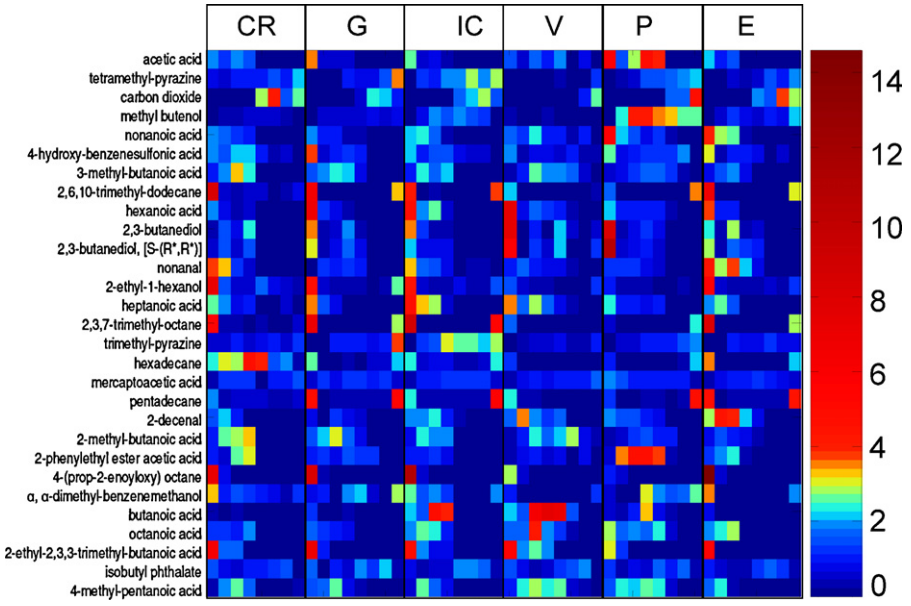

Fig. 3. PARAFAC signal volumes for all 29 analytes. Time course information can be discerned. The columns are organized from data point 0 (0 days) through data point 7 (1 month, i.e., 100% coverage) for each origin. CR: Costa Rica, G: Ghana, IC: Ivory Coast, V: Venezuela, P: Panama, and E: Ecuador. Each row represents an analyte (labeled) in the same order as Table 1.

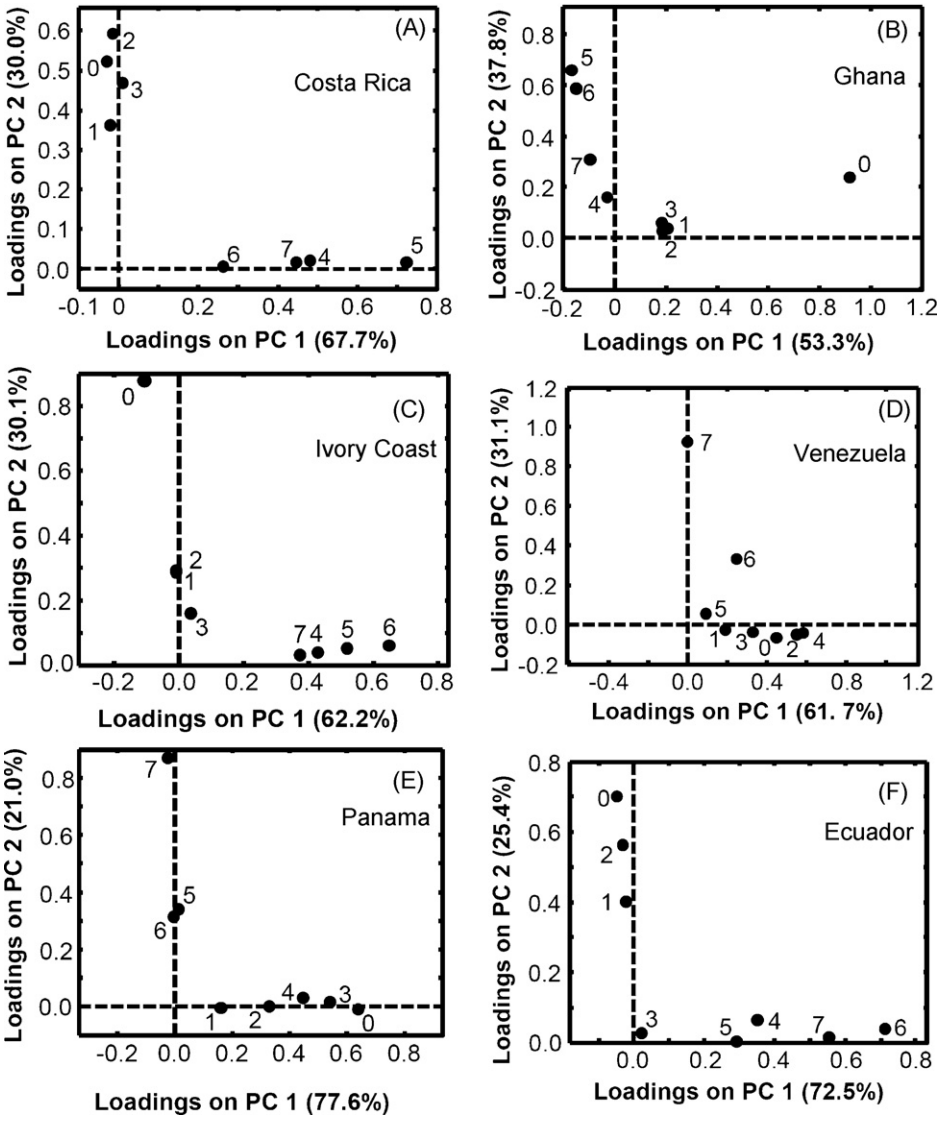

Fig. 4. PCA on each origin independently. (A) Costa Rica, (B) Ghana, (C) Ivory Coast, (D) Venezuela, (E) Panama, and (F) Ecuador. The days, 0–6, are labeled as such, and the 1-month time point is labeled as data point 7 (defined as 100% coverage).

While in Fig. 5A–C it is possible to determine whether or not a bean is visibly molded, the precise order of the time points is not always discernable with PCA in either the individual origins or the combined origins. For this reason, machine-learning techniques were employed to determine if the data could be modeled so as to predict the amount of time since moisture damage, which could allow for the detection of moisture damage prior to visible mold growth and aid in screening for bean quality. In order to quantify our ability to detect moisture damage from measurements of the volatile headspace analytes alone, a regression analysis was performed on the data. We regress on the data point (number of days after moisture damage 0–6 and the additional 100% coverage point) as a function of 28 of the 29 analytes listed in Table 1. The carbon dioxide level was not included as a predictor variable in the regression analysis because it is not regarded as a suitable candidate for in-field detection systems as it may be affected by numerous external conditions.

A selection of ten machine-learning regression methods were applied to the complete training data, and 10-fold cross-validation was used to estimate generalization performance with a summary of the results shown in Table 2. Based on these data, linear regression using the complete set of variables was found to be the least accurate model. This is almost certainly due to significant multi-colinearity in the data, which confounds this technique. The colinearity arises because many of the analyte levels are not independent, but vary together due to their chemical-reactive and/or metabolic relationships. PLS is better suited to handle this problem and does much better, but it still does not reach the performance of other models. This is likely because its linear model cannot ade-

Table 2
Coefficients of determination are listed for a selection of machine-learning regression methods.

Regression model	R^2 from 10-fold cross-validation. Mean (SD) from 10 runs
Linear regression	0.499 (0.12)
Partial least squares regression	0.608 (0.10)
Gaussian processes	0.661 (0.11)
Radial basis function network	0.671 (0.08)
1-Nearest neighbour	0.737 (0.09)
SMO-support vector regression	0.743 (0.12)
Random subspaces	0.702 (0.07)
Random forest	0.861 (0.11)
M5 regression tree	0.705 (0.10)
CART	0.702 (0.11)

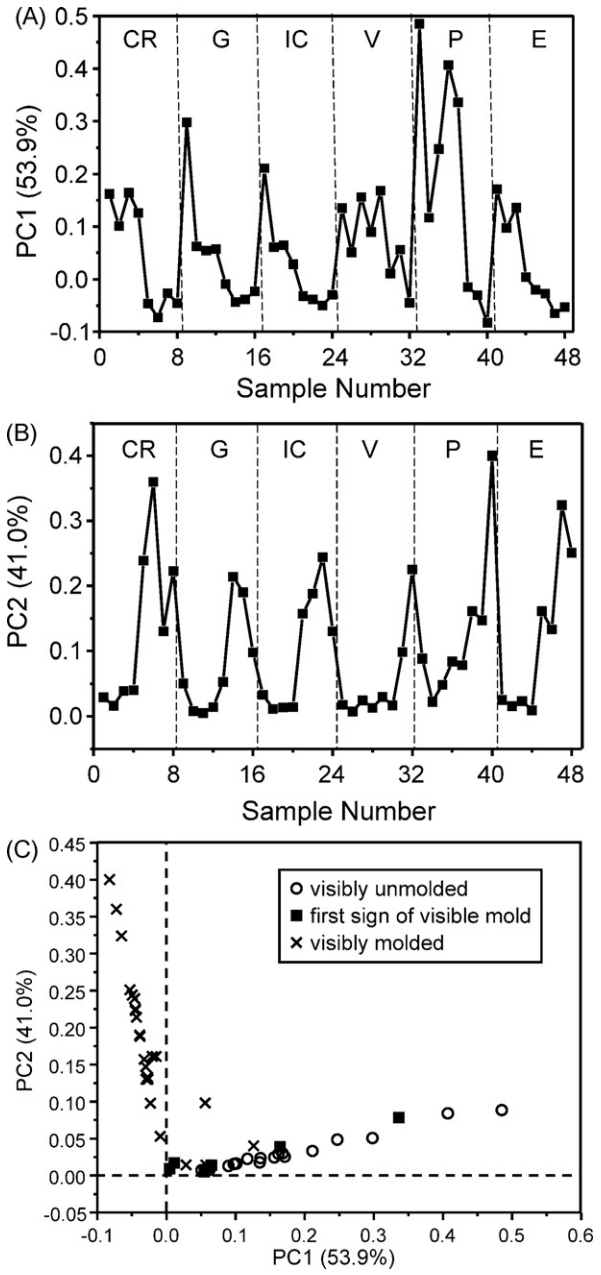

Fig. 5. PCA on all origins combined. (A) PC1 is plotted versus sample number with the samples in the same order as Fig. 3. (B) PC2 is plotted versus sample number. Within each origin, the data points are organized from data point 0 (0 days) through data point 7 (1 month, i.e., 100% coverage). (C) PC1 is plotted versus PC2. CR: Costa Rica, G: Ghana, IC: Ivory Coast, V: Venezuela, P: Panama, and E: Ecuador.

quately capture the non-linear and even non-monotonic changes observed in the analyte levels over the time course. There are two main groups of learning models that do substantially better at modeling this data. The nearest-neighbor and radial basis function methods both use all variables without weighting them, and the predictions are based largely on the most proximal data points in this unweighted space. Such methods are not hindered by non-linearity, a large number of variables, or multi-colinearity. The random forest, random subspaces, CART and M5 regression tree methods all use rules arranged in tree structures, which also enables them to model non-linear data more effectively. Also, random subspaces and random forests pool a large number of trees together, allowing them to make use of all the variables without over-fitting, while CART and M5 use only a selection of variables.

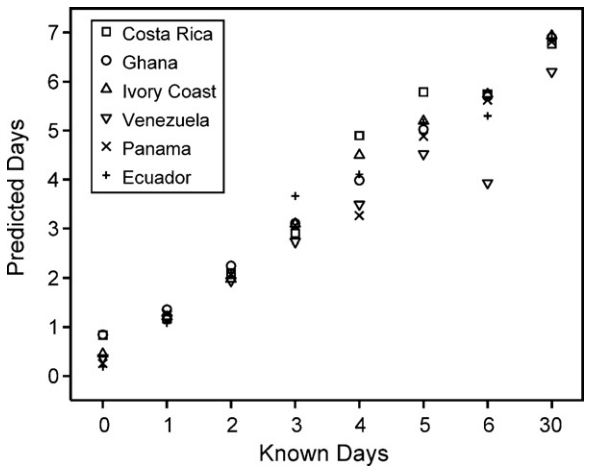

Fig. 6. Predictions of a random forest regression model. The *x*-axis displays the actual time points, from 0 days (data point 0) through 1 month, i.e., 100% coverage (data point 7).

Overall, the high R^2 values of the best models indicate that the headspace data contain sufficient information to detect not just the presence of moisture damage, but also to determine *when* the contamination occurred. A plot of the predictions made by a random forest model, which proved to be the best prediction method, is shown in Fig. 6. We can see from this plot that it is particularly good at accurately differentiating between no contamination and each of the first 2 or 3 days of contamination. This is the time-frame where there is no visible mold and visual inspection of cacao beans will not detect moisture damage. The ability to quantitatively distinguish between these time points indicates a clear potential for the development of early-warning systems related to food safety and quality issues.

The plot in Fig. 6 also shows that no bean variety has consistently higher or lower predictions than the main trend, suggesting that the varieties behave largely similarly. This implies that these changes are origin independent and is consistent with what was observed in the PARAFAC signal volumes shown in Fig. 3 and the combined origin PCA results shown in Fig. 5. We further tested this hypothesis by retraining the random forest model on five of the bean varieties and testing on the sixth, arguably the one with the largest apparent differences, the Venezuela variety. The R^2 value obtained on this test was still high and averaged around 0.7 (10 runs). The experiment was then repeated using the nearest-neighbor classifier (1 run because it is deterministic) which gave an R^2 value of 0.788. These values are less reliable than those quoted in Table 2, because they are based on a test set of just 8 data points, but there is no indication from our data that different bean varieties cannot be treated similarly.

All the regression methods reported above were trained on all variables. However, one might be interested in using only a subset of the variables. This would be especially the case for the development of more cost-effective measurement and detection systems for in-field applications. The CART and M5 decision tree methods implicitly perform a selection of variables during the training process, and the final models use only these variables. Fig. 7 shows a single CART model trained in the same way as those that were cross-validated and reported in Table 2. It uses only six of the available variables to make the predictions and performs at a relatively high level compared to the much larger nearest-neighbor and random forest models (see Table 2). Additionally, there seem to be numerous options in which subsets of variables can be used to still arrive at accurate predictions. Models trained on ten different random selections of *k* variables were assessed, at a variety of *k* values for the random forest method. Certain combinations of variables

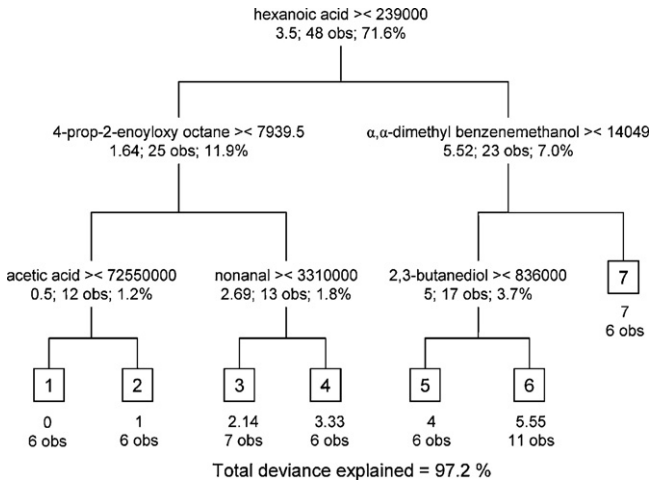

Fig. 7. CART tree that predicts the number of days since moisture damage. At each node, the decision rule displayed (top line) indicates which branch to follow (left or right) based on the value of the variable named. The % of total deviance explained by the rule is shown as the third value on the second line. The number of observations entering a node and their mean value are the first two values on the second line, respectively. The leaf nodes indicate the mean value of the observations there, and their number. The model shown splits the data into seven categories based on six explanatory variables (analytes). It does not manage to separate days 5 and 6, grouping eleven observations together, but makes few other errors. See Table 2 for the cross-validated R^2 performance.

Table 3
R^2 values (under 10-fold cross-validation) are given for random forest models, based on selections of k variables, with k ranging from 10 down to 2. Ten independent and entirely random selections at each value of k were done; the table summarizes the distribution of results obtained.

		Number of variables, k				
		10	5	4	3	2
R^2	Min	0.67	0.64	0.62	0.5	0.03
	Median	0.85	0.76	0.7	0.7	0.5
	Max	0.9	0.79	0.85	0.8	0.71

give better performance than others, but accurate models based on only three or four variable combinations exist and are not difficult to find, as shown in Table 3. We have also run more direct variable-selection techniques, but these do not give a consistent list of top predictor variables, such that they could be presented as definitive. Rather, it is found that many different subsets of the variables are informative enough to make accurate (cross-validated) predictions. As seen in Table 3, more than half the random forest models based on only three variables give R^2 values of 0.7 or better.

The success of the various prediction algorithms with all of the data combined, or with just a subset of the analytes, shows that it is possible to determine whether moisture damage has occurred before there are visible signs of mold. The precise time since damage may differ due to many environmental factors (i.e., humidity, exposure to elements, etc.), which would be important to further investigate prior to any implementation as a field device. However, there seem to be numerous origin-independent markers that indicate whether or not the moisture damage has occurred.

4. Conclusions

We have shown that moisture damage to cacao beans alters the volatile chemical signature in a way that can be tracked over time. These headspace vapor changes can be sampled with HS-SPME and detected and analyzed with GC × GC–TOFMS. A number of analytes that change in concentration levels via the moisture damage process were determined using the F-Ratio algorithm and quantified with the PARAFAC algorithm. It is possible to use prediction algorithms to determine whether moisture damage has occurred before there are visible signs of mold by analyzing subsets of the analytes.

Acknowledgments

We thank the Washington Technology Center (WTC) for their financial support. JDK thanks the University of Manchester for supporting his visit to the University of Washington as a Visiting Scholar.

References

[1] R. Dand, The International Cocoa Trade, Woodhead Publishing, 1999.
[2] F. Frauendorfer, P. Schieberle, J. Agric. Food Chem. 65 (2006) 5521.
[3] T. Stark, S. Bareuther, T. Hofmann, J. Agric. Food Chem. 54 (2006) 5530.
[4] Z. Zhang, J. Pawliszyn, Anal. Chem. 65 (1993) 1843.
[5] E.M. Humston, Y. Zhang, G.F. Brabeck, A. McShea, R.E. Synovec, J. Sep. Sci. 32 (2009) 2289.
[6] S. Ducki, J. Miralles-Garcia, A. Zumbe, A. Torenero, D.M. Storey, Talanta 74 (2008) 1166.
[7] J. Beens, M. Adahchour, R.J.J. Vreuls, K. van Altena, U.A.Th. Brinkman, J. Chromatogr. A 919 (2001) 127.
[8] C.A. Bruckner, B.J. Prazen, R.E. Synovec, Anal. Chem. 70 (1998) 2796.
[9] R.M. Kinghorn, P.J. Marriott, J. High Resolut. Chromatogr. 21 (1998) 620.
[10] Z. Liu, J.B. Phillips, J. Chromatogr. Sci. 29 (1991) 227.
[11] J.V. Seeley, F. Kramp, C.J. Hicks, Anal. Chem. 72 (2000) 4346.
[12] R. Shellie, L. Mondello, P. Marriott, G. Dugo, J. Chromatogr. A 970 (2002) 225.
[13] J.L. Hope, B.J. Prazen, E.J. Nilsson, M.E. Lidstrom, R.E. Synovec, Talanta 65 (2005) 380.
[14] R. Shellie, P. Marriott, P. Morrison, Anal. Chem. 73 (2001) 1336.
[15] A.E. Sinha, B.J. Prazen, C.G. Fraga, R.E. Synovec, J. Chromatogr. A 1019 (2003) 79.
[16] M. van Deursen, J. Beens, J. Reijenga, P. Lipman, C. Cramers, J. Blomberg, J. High Resolut. Chromatogr. 23 (2000) 507.
[17] J. Dalluge, M. van Rijn, J. Beens, R.J. Vreuls, U.A.Th. Brinkman, J. Chromatogr. A 965 (2002) 207.
[18] J. Dalluge, R.J. Vreuls, J. Beens, U.A.Th. Brinkman, J. Sep. Sci. 25 (2002) 201.
[19] E.M. Humston, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 80 (2008) 8002.
[20] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756.
[21] R.E. Mohler, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, Anal. Chem. 78 (2006) 2700.
[22] R.E. Mohler, B.P. Tu, K.M. Dombek, J.C. Hoggard, E.T. Young, R.E. Synovec, J. Chromatogr. A 1186 (2008) 401.
[23] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.M. Wright, R.E. Synovec, Anal. Chem. 78 (2006) 5068.
[24] R. Shellie, W. Welthagen, J. Zrostlikova, J. Spranger, M. Ristow, O. Fiehn, R. Zimmermann, J. Chromatogr. A 1086 (2005) 83.
[25] A.E. Sinha, J.L. Hope, B.J. Prazen, E.J. Nilsson, R.M. Jack, R.E. Synovec, J. Chromatogr. A 1058 (2004) 209.
[26] W. Welthagen, R. Shellie, J. Spranger, M. Ristow, R. Zimmermann, O. Fiehn, Metabolomics 1 (2005) 65.
[27] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341.
[28] A.E. Sinha, J.L. Hope, B.J. Prazen, C.G. Fraga, E.J. Nilsson, R.E. Synovec, J. Chromatogr. A 1056 (2004) 145.
[29] R. Bro, Chemometr. Intell. Lab. Syst. 38 (1997) 149.
[30] J.C. Hoggard, R.E. Synovec, Anal. Chem. 79 (2007) 1611.
[31] A.E. Sinha, C.G. Fraga, B.J. Prazen, R.E. Synovec, J. Chromatogr. A 1027 (2004) 269.
[32] R.D.C. Team, R: A Language and Environment for Statistical Computing, 2008.
[33] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, 2005.
[34] T.M. Therneau, B. Atkinson, B. Ripley, R Package Version 3.1-41, 2008.
[35] A. Liaw, M. Wiener, R News 2/3 (2002) 18.
[36] R. Kohavi, Proc. Int. Joint Conf. Artif. Intell. 14 (1995) 1137.
[37] C.R. Rao, Probability and Mathematical Statistics, Wiley, 1973.