Food Research International 140 (2021) 109983

Contents lists available at ScienceDirect

Food Research International

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

Cocoa origin classifiability through LC-MS data: A statistical approach for large and long-term datasets

Santhust Kumar [a,1,*], Roy N. D'Souza [a,1], Britta Behrends [a], Marcello Corno [b], Matthias S. Ullrich [a], Nikolai Kuhnert [a], Marc-Thorsten Hütt [a]

[a] Department of Life Sciences and Chemistry, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany
[b] Barry Callebaut AG, Westpark, Pfingstweidstrasse 60, Zurich 8005, Switzerland

ARTICLE INFO

Keywords:
Theobroma cacao
LC-MS
Principal component analysis (PCA)
Linear discriminant analysis (LDA)
Origin classification
Feature selection

ABSTRACT

Classification of food samples based upon their countries of origin is an important task in food industry for quality assurance and development of fine flavor products. Liquid chromatography – mass spectrometry (LC-MS) provides a fast technique for obtaining in-depth information about chemical composition of foods. However, in a large dataset that is gathered over a period of few years, multiple, incoherent and hard to avoid sources of variations e.g., experimental conditions, transportation, batch and instrumental effects, etc. pose technical challenges that make the study of origin classification a difficult problem. Here, we use a large dataset gathered over a period of four years containing 297 LC-MS profiles of cocoa sourced from 10 countries to demonstrate these challenges by using two popular multivariate analysis methods: principal component analysis (PCA) and linear discriminant analysis (LDA). We show that PCA provides a limited separation in bean origin, while LDA suffers from a strong non-linear dependence on the set of compounds. Further, we show for LDA that a compound selection criterion based on Gaussian distribution of intensities across samples dramatically enhances origin clustering of samples thereby suggesting possibilities for studying marker compounds in such a disparate dataset through this approach. In essence, we show and develop a new approach that maximizes, avoiding overfitting, the utility of multivariate analysis in a highly complex dataset.

1. Introduction

The classification of cocoa based on quality has been a major challenge in the chocolate industry, which is in industrial practice currently often based on a simple cut-test, rather not reflecting cocoa's vast chemical repertoire. Cocoa quality and geographic provenance define cocoa prices and quality. Assurance of cocoa authenticity protects manufacturers and consumers alike from commercial fraud.

Liquid chromatography – mass spectrometry (LC-MS) constitutes the most powerful technique for metabolomics analysis providing high resolution combined with high sensitivity. Cocoa products are among the most complex materials available to mankind with more than 10,000 individual peaks detectable in a single mass spectrum in processed cocoa (Milev, Patras, Dittmar, Vrancken, & Kuhnert, 2014). When comparing a multitude of samples, powerful chemometric algorithms must be used to extract meaningful information from such datasets. At the same time, large sample numbers are substantially challenging such algorithms.

Several research approaches have previously addressed classification of cocoa beans or cocoa products based on variety of cocoa-constituents classes (D'Souza et al., 2017; Kumari et al., 2018; Megías-Pérez, Grimbs, D'Souza, Bernaert, & Kuhnert, 2018; Sirbu, Grimbs, Corno, Ullrich, & Kuhnert, 2018) using information-rich techniques based on molecular fingerprints (Aculey et al., 2010; Magagna et al., 2017; Vázquez-Ovando, Molina-Freaner, Nuñez-Farfán, Betancur-Ancona, & Salvador-Figueroa, 2015), or low-resolution techniques based on sum parameters (Guehi, Zahouli, Ban-Koffi, Fae, & Nemlin, 2010). Multivariate statistical methods based on LC-MS datasets have been shown to assess the degree of fermentation of cocoa samples, but also to discriminate between its origin based on its metabolome fingerprint (Acierno et al., 2016, 2018; Bindereif, Brauer, Schubert, Schwarzinger, & Gebauer, 2019; D'Souza et al., 2017; Kumari et al., 2018; Marseglia et al., 2016; Oliveira et al., 2016). Differences in the cocoa metabolome have been shown to be a consequence of distinct agricultural practice, climate and soil influences (Adeniyi, de Clercq, & van Niekerk, 2019; Arévalo-

* Corresponding author at: Department of Life Sciences & Chemistry, Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany.
E-mail address: s.santhust@jacobs-university.de (S. Kumar).
[1] These authors contributed equally to this manuscript.

https://doi.org/10.1016/j.foodres.2020.109983
Received 4 August 2020; Received in revised form 26 November 2020; Accepted 8 December 2020
Available online 25 December 2020

Hernández, da Conceição Pinto, de Souza Júnior, de Queiroz Paiva, & Baligar, 2019; Asare, Asare, Asante, Markussen, & Ræbild, 2017; Ehiakpor, Danso-Abbeam, & Baah, 2016; Kongor et al., 2016, 2019) and are among other factors based on genetic variability found in cocoa tree populations in different origin countries as illustrated in various studies (Arevalo-Gardini et al., 2019; Gopaulchan et al., 2019; Lindo, Robinson, Tennant, Meinhardt, & Zhang, 2018; Zhang & Motilal, 2016).

Nevertheless, the size of dataset used in these studies is often small and does not deal with effects involved in data gathered over a prolonged period of time, for e.g., batch effects, temperature variation, changes in experimental condition, incoherent transportation influences. However, long-term studies, meta-studies combining individual datasets and large-scale, multi-laboratory investigations are becoming more prominent in food research and other application of LC-MS technology (Zabell, Stone, & Julian, 2017).

Here, we use the largest dataset, to the best of our knowledge, comprising of 297 LC-MS profiles of aqueous methanolic extracts rich in polyphenolics and peptides (positive and negative ion modes) of unfermented or fermented cocoa beans, as well as selected cocoa liquors, sourced from a total of 10 different countries. The data has been gathered over a period of four years, from 2014 to 2017, and thus possess multiple and difficult to control sources of variation. Selected parts of this dataset have been published already and is included here in a larger context. To begin with, we demonstrate the limitations of a popular unsupervised method of classification often used in food composition analytics, principal component analysis (PCA), in clustering of the cocoa samples on the basis of the country to which they belong. Next, we assess another popular but this time a supervised method of classification, the linear discriminant analysis (LDA), and show that its results depends strongly and nonlinearly on the number of compounds used while performing the analysis.

Further, we outline a statistics-based feature selection criterion for filtering compounds which is motivated from the grounds of (a) reducing noise in the dataset with regards to country information and (b) reducing arbitrariness in feature picking which is often seen in such analyses. We show that our approach, which we refer to as Gaussian Feature Stability (GFS) requirement, greatly improves the outcome of LDA for origin-based clustering in the disparate dataset at hand. Next, dividing the dataset into train and test subsets, we show that this procedure also improves the predictive capability of LDA.

Thus, our work not only clearly brings out some of the challenges encountered in applying two of the most popular multivariate analysis techniques, *viz* PCA and LDA, to a large and disparate LC-MS dataset, but also suggests a potential way for dealing with them. The improvement in clustering of LDA can potentially help in finding compounds which can differentiate between countries of origin.

2. Materials and methods

2.1. Details of LC-MS data

In the analysis we present here, a total of 297 LC-MS profiles of cocoa aqueous methanolic extracts rich in polyphenols and peptides were used. The LC-MS profiles are broadly categorized into two MS ion modes, three sample types and 10 origins (or countries). These LC-MS profiles have been obtained over a period of four years (2014–2017). The actual details of sample preparation, extraction, standard protocols, carrying out the LC-MS experiment and data collection are given in our earlier published reports (D'Souza et al., 2017, 2018). Thus the dataset used in this work augments on the smaller dataset used in earlier works from our group (D'Souza et al., 2017, 2018; Kumari et al., 2018). The number of LC-MS profiles with respect to ion modes, sample types and their countries of origin are provided in Table 1. Hybrid varieties of the samples, whenever available, and the year in which they were gathered are provided in Supporting Information S1.

Positive and negative ion mode data jointly provide a more holistic approach enabling an instant molecular snapshot of the coca bean chemical composition. Negative ion mode data show a bias towards polyphenolic compounds, acids and carbohydrates, whereas positive ion mode reveal a multitude of peptidic structures. Both modes are largely complementary with an overlap of identified compounds of around 30–40% depending on the sample type.

2.2. Data pre-processing and cleaning

We first processed the individual MS ion modes LC-MS data using MZMine (Pluskal, Castillo, Villar-Briones, & Orešič, 2010). This yields a retention time aligned peak area list indicating the *m/z* ratio, the *retention time* and the integrated peak areas of each sample. Then using an extensive precompiled list of compounds and their chemical formulas, corresponding *m/z* values were computed in four ion types for both modes: [M−H], [M−2H], [M−3H], [2M−H] for negative ion mode, and [M + H], [M + 2H], [M + 3H], [2M + H] for positive ion mode. Detailed assignment tables have been published previously (D'Souza et al., 2017, 2018). Mass errors were only accepted if below 5 ppm of the theoretical *m/z* value. It should be kept in mind that in the 297 different chromatographic runs mass error display a certain variability in-between runs. The instrument used (Bruker ESI-Q-TOF-Impact HD) operates at a specified mass accuracy of 0.9 ppm and most mass errors observed are in the range between 1 and 2 ppm. The details are further noted in Supporting Information S3. Using this information (D'Souza et al., 2017, 2018), the compounds detected in the LC-MS data were assigned to known structures, whenever possible based on authentic references, tandem MS data or literature precedents, else the compound was included, however, considered as 'Unknown_' suffixed with the *m/z* value (e.g. Unknown_865.1927). In cases where more than

Table 1
Information about the LC-MS dataset. Country-wise and process-wise distribution of number of LC-MS samples in the dataset used in this study.

	Positive ion mode				Negative ion mode			
	Unfermented	Fermented	Liquor	*Sum*	Unfermented	Fermented	Liquor	*Sum*
Brazil	4	4	0	**8**	4	4	0	**8**
Cameroon	3	3	6	**12**	3	3	7	**13**
Ecuador	8	12	3	**23**	8	12	5	**25**
Ghana	0	0	5	**5**	0	0	5	**5**
Indonesia	14	16	0	**30**	14	16	0	**30**
Ivory Coast	16	16	9	**41**	16	16	9	**41**
Madagascar	0	0	0	**0**	0	0	5	**5**
Malaysia	6	3	0	**9**	6	3	0	**9**
STAP[1]	0	0	0	**0**	0	0	5	**5**
Tanzania	3	9	0	**12**	3	9	4	**16**
Sum	**54**	**63**	**23**	**140**	**54**	**63**	**40**	**157**

[1] Sao Tome and Principe.

one compound matched an *m/z* value, a combined name, joined by 'or' was assigned. It is important to emphasize here that the assignment of names was performed for obtaining further insights about the detected compounds, when possible, and does not affect the PCA and LDA results discussed here. The processed data were then combined with metadata about each sample in a single data structure, which contained information about the sample type, origin and peak areas of various compounds. The sum of peak area values belonging to each sample was normalized to 100, so the peak area in the sample represent relative percentage amount of compound in the LC-MS profile. Henceforth, we refer to the percentage normalized peak area as the peak areas itself. Columns of compounds were sorted in descending order by their mean peak area across all samples, such that the compounds with higher mean peak areas are placed in initial columns, and those with lesser peak areas are placed in later columns.

2.3. Unsupervised and supervised learning methods

A couple of multivariate statistical analysis or 'machine learning' methods, both unsupervised (e.g., PCA – Principal Component Analysis) and supervised (e.g., LDA – Linear Discriminant Analysis, Random Forests, Support Vector Machines, Neural Nets) have been used in this study. These methods were applied using the popular scikit-learn module (Pedregosa et al., 2011) of Python programming language. Standard scaling was implemented while performing PCA. In case of Random Forests, Support Vector Machines, and Neural Nets, where it is possible to vary algorithm specific parameters, default parameter settings in the implementation in scikit-learn were used to avoid additional complexity of dealing with parameter specific algorithmic performance.

2.4. Gaussian Feature Stability criterion

The Gaussian Feature Stability (GFS) criterion is based on the test whether a set of values is distributed normally, or not. If the values in the peak area list corresponding to a compound in the LC-MS profiles, under a given sample type and belonging to a given origin, are found to be normally distributed, we say that the compound satisfies the Gaussian Feature Stability requirement. For testing normality of a set of values, we use the Shapiro-Wilk (Shapiro & Wilk, 1965) test implemented in the popular Scipy module (Jones, Oliphant, & Peterson, 2001) of Python programming language. We test normality at a p-value threshold of 0.05.

2.5. Null model

First, we find out how many compounds (say n) satisfy GFS criterion under a group of samples. Here, a group is comprised of samples belonging to a given country (say X), under a given sample-type (say Y). Then we randomly choose n compounds from the same group of origin (X) and sample type (Y). Using these n randomly chosen compounds, we calculate prediction from LDA. This procedure is repeated a number of times to obtain result statistics. The result so obtained is referred to as result from null model.

3. Results and discussion

For this study we have selected 297 samples representing well the world of industrial cocoa, including high production origins from Africa, South America and Asia over several years of harvest. The dataset includes as well cocoa beans of different processing stages reflecting chemical changes along the processing chain. In wine chemistry for example a memory effect with respect to origin was recently proposed allowing improved origin prediction following ageing (Roullier-Gall, Boutegrabet, Gougeon, & Schmitt-Kopplin, 2014). In the main text we report results, primarily, about the negative ion mode data. When appropriate or needed, corresponding plots, results, etc., for the positive ion mode are reported in the Supporting Information S2.

3.1. Classification of cocoa using PCA

Principal component analysis (PCA), an unsupervised method of classification (James, Witten, Hastie, & Tibshirani, 2013), is often the first choice in food chemistry analyses for exploring and studying grouping relationships in samples (Aculey et al., 2010; Cordella, 2012; Granato, Santos, Escher, Ferreira, & Maggio, 2018). However, PCA has been found of limited success in the studying country wise classification of cocoa beans (D'Souza et al., 2017; Kumari et al., 2018). In Fig. 1, we show PCA score plots for the negative ion mode data using unfermented, fermented and liquor samples of cocoa in our analysis (using top 2000 compounds; see Supporting Information S2 for positive mode data and plots with different numbers of compounds). It can be seen that samples belonging to same country (depicted by dots with same color in the PCA plots) tend to be present close by. However, one can also clearly witness the mixing of samples belonging to different countries. We note that this separation or mixing of samples further varies with (a) the number, and (b) the actual set, of compounds (actually a RT/*m/z* pair corresponding to a chemical compound) used as features in the analysis, see Supporting Information S2. This points to a need for an alternative approach for identifying marker compounds for distinction between cocoa samples belonging to different countries.

3.2. Classification of cocoa using LDA

We next applied linear discriminant analysis (LDA) (James et al., 2013) to the cocoa samples analyzed by LC-MS. LDA, as a supervised method of classification, uses the available class information of the samples (in our case, countries of origin) in order to find out axes, which give best possible grouping of samples belonging to the same class (i.e., country, in our case). It achieves this by simultaneously minimizing the within-class variation and maximizing between-class variation. This is in contrast with PCA, which determines axes with most variation independent of the class the samples. In this way, LDA is better suited for determining compounds, which are good differentiators of samples belonging to different origins and compounds, which makes samples belonging to the same country similar.

In Fig. 2, we show the LDA of unfermented, fermented, and liquor samples. For each sample type, the analysis is performed under four divisions, using: (a) all compounds, (b) the first 500 most abundant compounds, (c) the first 100 most abundant compounds, and (d) only the first 10 most abundant compounds. It becomes apparent that the grouping of samples varies with number of abundant compounds (or features) included prior to performing the LDA. A marked improvement in the grouping of samples belonging to same origin country is observed when only including the top 30–50 most abundant compounds for the LDA. However, this soon disappears as the number of compounds are further decreased.

The clustering of samples on a given feature may be quantified by calculating the LDA score. One can see that the "good" clustering corresponds to the high LDA score and "bad" clustering to the low score. While in originality the LDA score is a measure of prediction power of the trained LDA model upon the test dataset, it works well as a measure of clustering of the datapoints (or the *classifiability* of the data), when the test dataset is kept the same as the training dataset. In Fig. 2B, we show the variation of LDA score with the number of compounds used for the analysis. The classifiability from LDA changes with the number of compounds in a non-linear manner. In the Supporting Information S2, we provide results on positive ion mode and the corresponding plots with the number of compounds displayed on a linear (not logarithmic) scale. This clearly points to the criticality of the number and set of compounds used as features in performing the LDA, and how to select them.

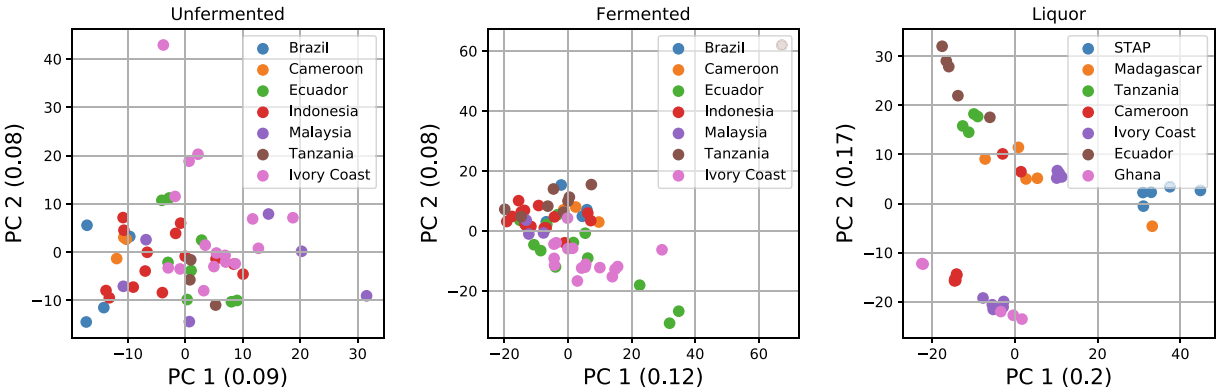

Fig. 1. PCA of Unfermented, Fermented and Liquor samples obtained from Cocoa processing pipeline. The PCA gives a limited amount of grouping of samples belonging to same country. (STAP: Sao Tome and Principe).

3.3. Filtering data for relevant features

In general, feature selection is a factor that needs to be decided before performing a multivariate analysis. Before performing LDA upon a given sample-type, we find out the set of compounds which follow a Gaussian distribution under each country under the given sample type. We refer to this filter as Gaussian Feature Stability (GFS) requirement (see section 2. Materials and Methods). The ground reasoning is as follows. It is fair to expect that the distribution of intensities for a compound belonging to a given country under a said process category is centered on some stable value, which might be characteristic for this subset of samples. Otherwise, if the amount of the compound in the samples belonging to a given country under a process category is not stable enough, it cannot be considered as a reliable feature enabling classification for the samples. The reason for instability can be varied: ranging from error in faithful detection of the compounds to the influence of other factors such as subsamples being procured in different batches over a period of time (as in our case the data is gathered over a period of few years). For each of the sample type categories (i.e., unfermented, fermented and liquors), we obtain the common set of compounds that satisfied GFS criterion for all the countries within the said category. After obtaining the list of compounds adhering to the GFS criterion, we use them as features for performing the LDA. Fig. 3 shows the resulting improvement in the LDA brought through this approach. Essentially, the GFS criterion seems to reduce conflicting information from samples belonging to same country.

It should be pointed out that once the country clusters become clearly separated by the LDA, the LDA score used above will no longer discriminate the amount of cluster separation (giving a value close to 1 for all cases). The higher quality of the reduction of compounds based on GFS will become clear in the following, when the LDA result is used for predictive purposes.

A glimpse into how the concentration of compounds obtained from GFS criterion is distributed for unfermented, fermented, and liquors, can be obtained through boxplots in Fig. 4. The compounds have differing country-wise distribution, for e.g., *p-Coumaroyl aspartate [M−H]* serves as marker within the liquor category. Identifying compounds with stable and differing distributions can be helpful in identification of origin specific characteristic compounds. Fig. 4 shows some selected examples of relative quantities of potentially marker compounds within the three stages of cocoa processing investigated.

3.4. Effect of Gaussian Feature Stability on prediction through LDA

Next, we study the effect of Gaussian Feature Stability (GFS) criterion on *prediction* of sample's country through LDA. First, for each sample-type category, the list of compounds satisfying GFS criterion was found. The sample dataset belonging to each category was divided into training and test sets in ratio 3:1. After that, the LDA model was trained through the training dataset, and its predictive power was assessed through test dataset by calculating the LDA score. The whole procedure, from data splitting to prediction score, was repeated several times (100) to achieve result statistics. To put the outcome into perspective, a comparison was made with an appropriate null model. In this null model, the same number of compounds as obtained previously by GFS criterion, were randomly chosen from all compounds under a given category and used to obtain the prediction statistics (see 2. Materials and Methods). This procedure was also repeated 100 times and was done for each category. The result is plotted in Fig. 5, with blue bars representing prediction power of LDA when compounds satisfying GFS requirement were used, and orange bars represent the prediction power when same number of randomly chosen compounds were used. It can be seen that higher predictive ability of LDA was achieved, when the compounds chosen through GFS criterion were used as compared to the case when same number of randomly chosen compounds were used (2. Materials and Methods, Null model). The trend of increasing prediction score in the negative mode data from unfermented to fermented to liquor stage suggests better origin prediction capability down a typical cocoa processing pipeline. While in the positive, the prediction score increases from unfermented to fermented stage, the drop in score for liquor disturbs this trend. Here it is important to mention that the liquor category has the least number of samples in our dataset, so it would be an interesting future task to investigate the suggested trend with more liquor samples in the data. In order to put the prediction statistics from LDA into perspective, we provide the prediction statistics of some popular machine learning algorithms in Supporting Information S2. As scope of this study is not to test plethora of machine learning algorithms, whenever an algorithm provided various possibilities for setting its parameters, the default parameter values of their implementation in the scikit-learn machine learning library in Python were used.

4. Conclusion

Classification of country of origin using LC-MS profiles of cocoa samples has remained a challenging task, especially for a large dataset gathered over a prolonged period of time. This study employs the biggest dataset of the LC-MS chemical profile of cocoa available to date for studying the feasibility of cocoa origin classification. Data from both positive and negative ion modes have been analyzed in order to obtain comprehensive insights.

Our work clearly shows some limitations of two popular multivariate analysis methods generally employed in such studies, PCA and LDA, in dealing with such a large and potentially disparate dataset. Further, we suggest a statistical approach motivated from the ground of noise reduction to infer stable features across the dataset thereby reducing arbitrariness which is often encountered in feature selection before

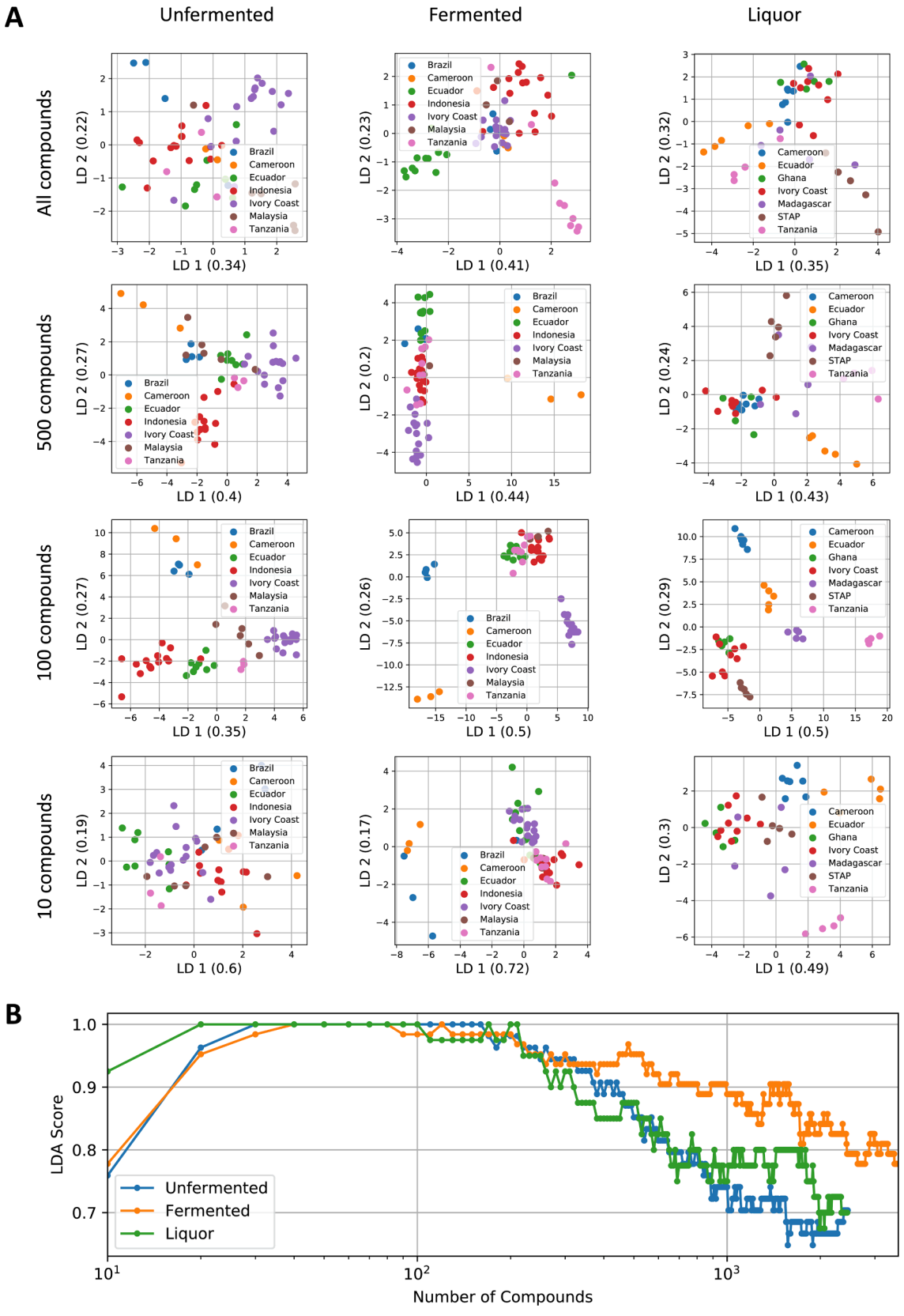

Fig. 2. LDA for different sample-types using different number of compounds (features). (A) The scatter plots show that the quality of grouping of samples depends on the set of compounds used in the analysis. (B) Non-linear variation of LDA-score (an estimate of clustering of same samples) as a function of number of compounds employed in the LDA. (STAP: Sao Tome and Principe).

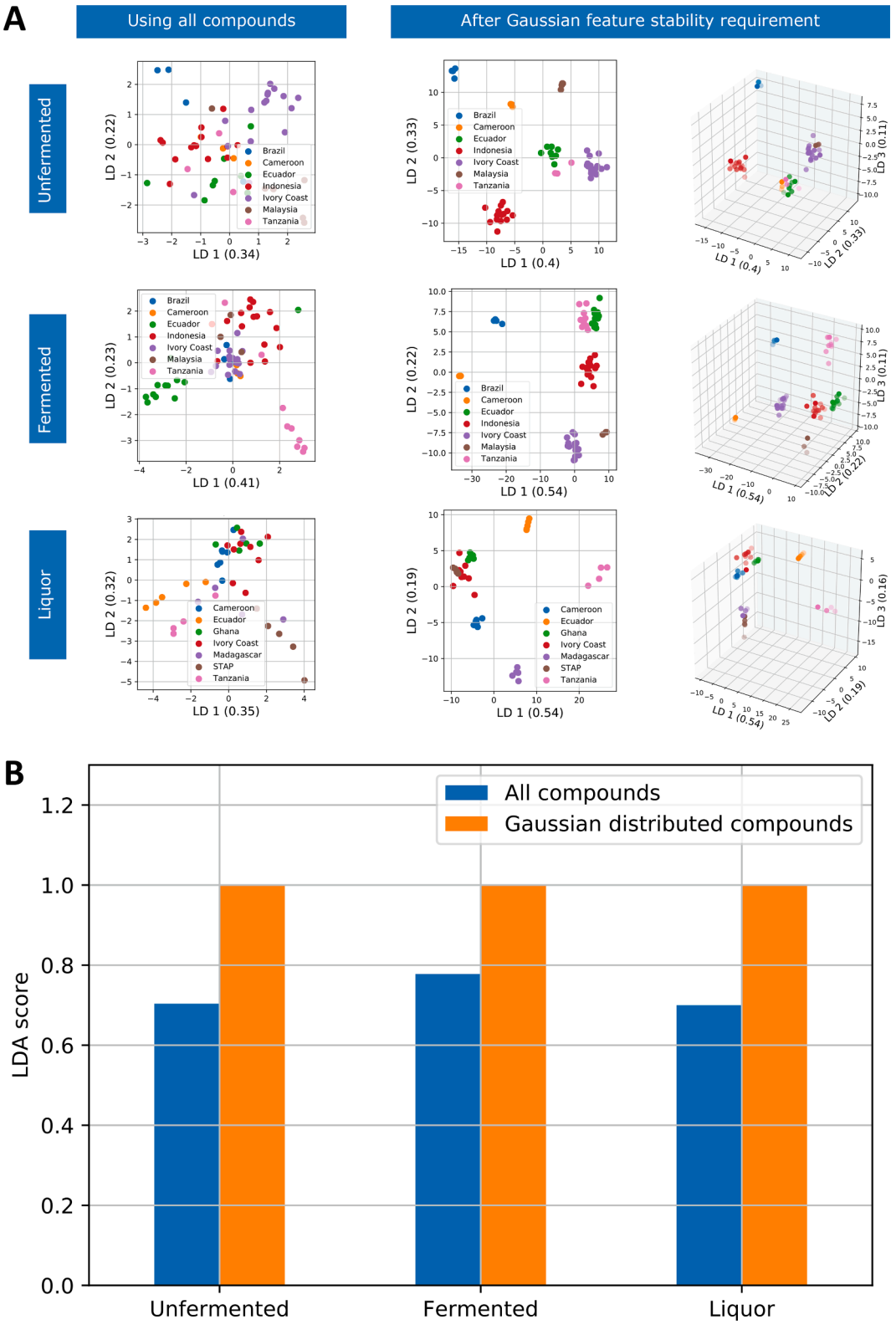

Fig. 3. LDA clustering improvement after Gaussian Feature Stability (GFS) requirement. A: Visual representation of clustering in LDA before and after application of GFS. B: Quantitative comparison of improvement in clustering in LDA after application of GFS. (STAP: Sao Tome and Principe).

performing such multivariate analyses.

Through the improved results of LDA, not only we show a marked improvement in classification of origin of cocoa samples, but our work also suggests possibilities of extending current state-of-art of multivariate analysis to large datasets not necessarily obtained at the same time. Thus, our work makes advancements towards big data analysis in terms of LC-MS profiles cocoa in particular and of food items in general.

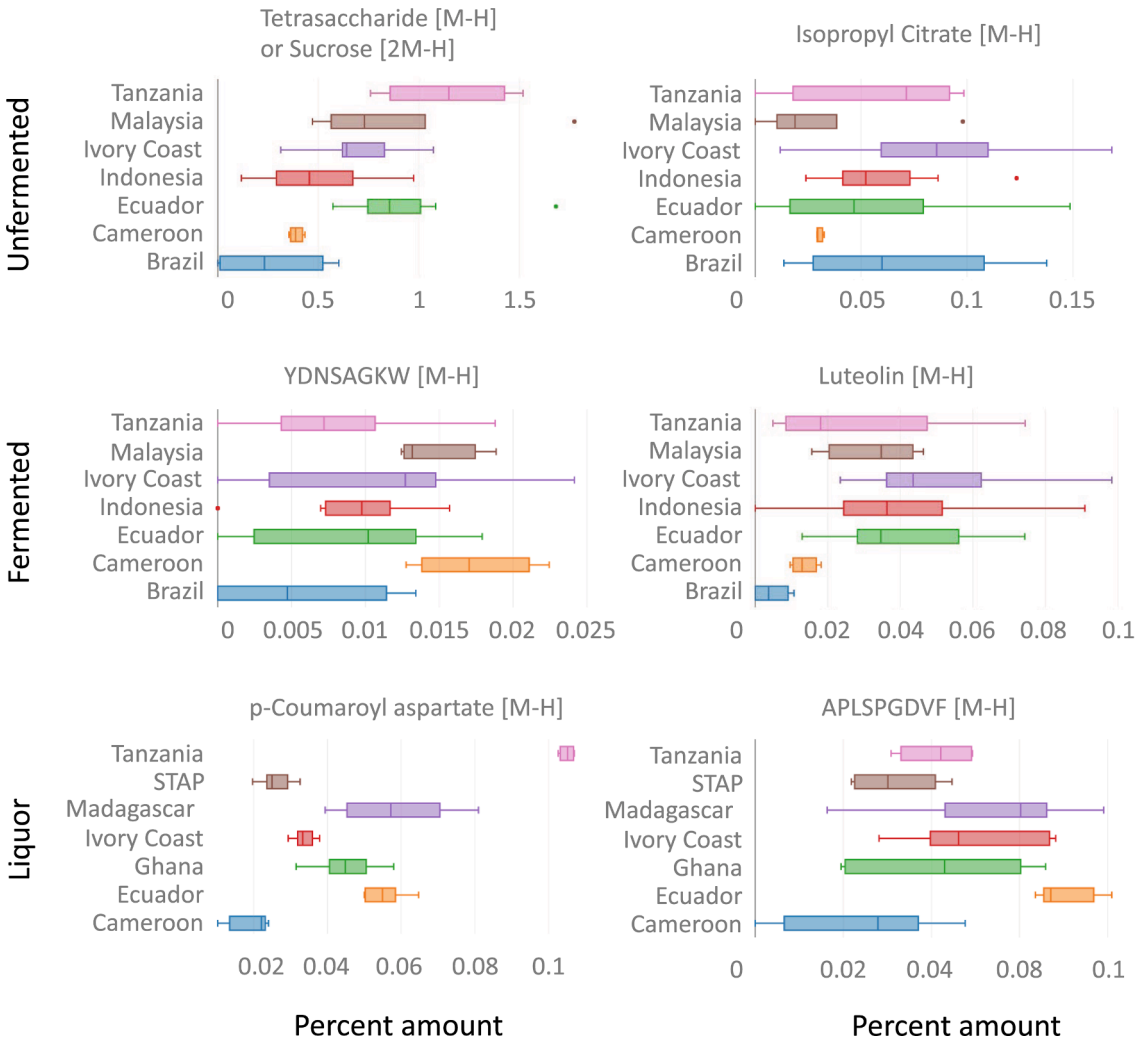

Fig. 4. Profiles of compound concentration (percentage). The Figure shows profiles of variation of concentration of compounds belonging to different countries for some compounds obtained after the application of Gaussian Feature Stability requirement. It can be seen that the countries have differing distributions of concentration of compounds. (STAP: Sao Tome and Principe).

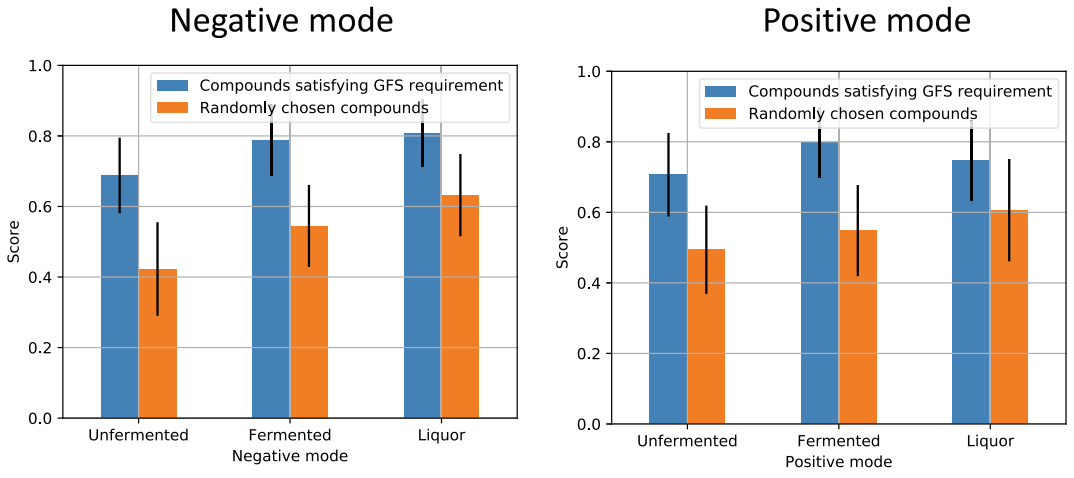

Fig. 5. Improvement in LDA predictive capability upon usage of compounds satisfying GFS requirement. Using the set of compounds obtained from GFS for predicting the country to which a sample belongs gives higher predictive power (blue bars) than expected by chance (orange bars). The chance predictability was calculated by randomly choosing the as many compounds as obtained by GFS requirement from the list of available compounds in respective positive/negative ion mode datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Santhust Kumar: Conceptualization, Data curation, Methodology, Formal analysis, Visualization, Writing - original draft, Writing - review & editing, Investigation, Software, Validation. **Roy N. D'Souza:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation. **Britta Behrends:** Methodology, Investigation. **Marcello Corno:** Project administration, Writing - review & editing. **Matthias S. Ullrich:** Conceptualization, Writing - review & editing, Funding acquisition, Project administration. **Nikolai Kuhnert:** Conceptualization, Writing - review & editing, Funding acquisition, Project administration. **Marc-Thorsten Hütt:** Conceptualization, Formal analysis, Investigation, Methodology, Supervision, Validation, Writing - review & editing, Funding acquisition, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Nina Böttcher excellent technical support in sample logistics and preparation. This work was funded by the COMETA project, which is financially supported by Barry Callebaut AG. Barry Callebaut also provided samples for analysis.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2020.109983.

References

Acierno, V., Alewijn, M., Zomer, P., & van Ruth, S. M. (2018). Making cocoa origin traceable: Fingerprints of chocolates using Flow Infusion - Electro Spray Ionization - Mass Spectrometry. *Food Control, 85*, 245–252. https://doi.org/10.1016/j.foodcont.2017.10.002.

Acierno, V., Yener, S., Alewijn, M., Biasioli, F., & van Ruth, S. (2016). Factors contributing to the variation in the volatile composition of chocolate: Botanical and geographical origins of the cocoa beans, and brand-related formulation and processing. *Food Research International, 84*, 86–95. https://doi.org/10.1016/j.foodres.2016.03.022.

Aculey, P. C., Snitkjaer, P., Owusu, M., Bassompiere, M., Takrama, J., Nørgaard, L., … Nielsen, D. S. (2010). Ghanaian Cocoa Bean fermentation characterized by spectroscopic and chromatographic methods and chemometrics. *Journal of Food Science, 75*(6), S300–S307. https://doi.org/10.1111/j.1750-3841.2010.01710.x.

Adeniyi, S. A., de Clercq, W. P., & van Niekerk, A. (2019). Assessing the relationship between soil quality parameters of Nigerian alfisols and cocoa yield. *Agroforestry Systems, 93*(4), 1235–1250. https://doi.org/10.1007/s10457-018-0238-2.

Arevalo-Gardini, E., Meinhardt, L. W., Zuñiga, L. C., Arévalo-Gardni, J., Motilal, L., & Zhang, D. (2019). Genetic identity and origin of "Piura Porcelana"—A fine-flavored traditional variety of cacao (Theoborma cacao) from the Peruvian Amazon. *Tree Genetics & Genomes, 15*(1), 11. https://doi.org/10.1007/s11295-019-1316-y.

Arévalo-Hernández, C. O., da Conceição Pinto, F., de Souza Júnior, J. O., de Queiroz Paiva, A., & Baligar, V. C. (2019). Variability and correlation of physical attributes of soils cultivated with cacao trees in two climate zones in Southern Bahia, Brazil. *Agroforestry Systems, 93*(3), 793–802. https://doi.org/10.1007/s10457-017-0176-4.

Asare, R., Asare, R. A., Asante, W. A., Markussen, B., & Ræbild, A. (2017). Influences of shading and fertilization on on-farm yield of cocoa in Ghana. *Experimental Agriculture, 53*(3), 416–431. https://doi.org/10.1017/S0014479716000466.

Bindereif, S. G., Brauer, F., Schubert, J.-M., Schwarzinger, S., & Gebauer, G. (2019). Complementary use of 1H NMR and multi-element IRMS in association with chemometrics enables effective origin analysis of cocoa beans (Theobroma cacao L.). *Food Chemistry, 299*, Article 125105. https://doi.org/10.1016/j.foodchem.2019.125105.

Cordella, C. B. Y. (2012). PCA: The Basic Building Block of Chemometrics. *Analytical Chemistry*. https://doi.org/10.5772/51429.

D'Souza, R. N., Grimbs, A., Grimbs, S., Behrends, B., Corno, M., Ullrich, M. S., & Kuhnert, N. (2018). Degradation of cocoa proteins into oligopeptides during spontaneous fermentation of cocoa beans. *Food Research International, 109*, 506–516. https://doi.org/10.1016/j.foodres.2018.04.068.

D'Souza, R. N., Grimbs, S., Behrends, B., Bernaert, H., Ullrich, M. S., & Kuhnert, N. (2017). Origin-based polyphenolic fingerprinting of Theobroma cacao in unfermented and fermented beans. *Food Research International, 99*, 550–559. https://doi.org/10.1016/j.foodres.2017.06.007.

Ehiakpor, D. S., Danso-Abbeam, G., & Baah, J. E. (2016). Cocoa farmer's perception on climate variability and its effects on adaptation strategies in the Suaman district of western region, Ghana. *Cogent Food & Agriculture, 2*(1), 1210557. https://doi.org/10.1080/23311932.2016.1210557.

Gopaulchan, D., Motilal, L. A., Bekele, F. L., Clause, S., Ariko, J. O., Ejang, H. P., & Umaharan, P. (2019). Morphological and genetic diversity of cacao (Theobroma cacao L.) in Uganda. *Physiology and Molecular Biology of Plants, 25*(2), 361–375. https://doi.org/10.1007/s12298-018-0632-2.

Granato, D., Santos, J. S., Escher, G. B., Ferreira, B. L., & Maggio, R. M. (2018). Use of principal component analysis (PCA) and hierarchical cluster analysis (HCA) for multivariate association between bioactive compounds and functional properties in foods: A critical perspective. *Trends in Food Science & Technology, 72*, 83–90. https://doi.org/10.1016/j.tifs.2017.12.006.

Guehi, T. S., Zahouli, I. B., Ban-Koffi, L., Fae, M. A., & Nemlin, J. G. (2010). Performance of different drying methods and their effects on the chemical quality attributes of raw cocoa material. *International Journal of Food Science & Technology, 45*(8), 1564–1571. https://doi.org/10.1111/j.1365-2621.2010.02302.x.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Introduction to Statistical Learning. http://www-bcf.usc.edu/~gareth/ISL/.

Jones, E., Oliphant, T., Peterson, P., & others. (2001). SciPy: Open source scientific tools for Python. http://www.scipy.org/.

Kongor, J. E., Boeckx, P., Vermeir, P., Van de Walle, D., Baert, G., Afoakwa, E. O., & Dewettinck, K. (2019). Assessment of soil fertility and quality for improved cocoa production in six cocoa growing regions in Ghana. *Agroforestry Systems, 93*(4), 1455–1467. https://doi.org/10.1007/s10457-018-0253-3.

Kongor, John Edem, Hinneh, M., de Walle, D. V., Afoakwa, E. O., Boeckx, P., & Dewettinck, K. (2016). Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile—A review. *Food Research International, 82*, 44–52. https://doi.org/10.1016/j.foodres.2016.01.012.

Kumari, N., Grimbs, A., D'Souza, R. N., Verma, S. K., Corno, M., Kuhnert, N., & Ullrich, M. S. (2018). Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans. *Food Research International, 111*, 137–147. https://doi.org/10.1016/j.foodres.2018.05.010.

Lindo, A. A., Robinson, D. E., Tennant, P. F., Meinhardt, L. W., & Zhang, D. (2018). Molecular characterization of Cacao (Theobroma cacao) Germplasm from Jamaica Using Single Nucleotide Polymorphism (SNP) Markers. *Tropical Plant Biology, 11*(3), 93–106. https://doi.org/10.1007/s12042-018-9203-5.

Magagna, F., Guglielmetti, A., Liberto, E., Reichenbach, S. E., Allegrucci, E., Gobino, G., … Cordero, C. (2017). Comprehensive chemical fingerprinting of high-quality cocoa at early stages of processing: Effectiveness of combined untargeted and targeted approaches for classification and discrimination. *Journal of Agricultural and Food Chemistry, 65*(30), 6329–6341. https://doi.org/10.1021/acs.jafc.7b02167.

Marseglia, A., Acquotti, D., Consonni, R., Cagliani, L. R., Palla, G., & Caligiani, A. (2016). HR MAS 1H NMR and chemometrics as useful tool to assess the geographical origin of cocoa beans – Comparison with HR 1H NMR. *Food Research International, 85*, 273–281. https://doi.org/10.1016/j.foodres.2016.05.001.

Megías-Pérez, R., Grimbs, S., D'Souza, R. N., Bernaert, H., & Kuhnert, N. (2018). Profiling, quantification and classification of cocoa beans based on chemometric analysis of carbohydrates using hydrophilic interaction liquid chromatography coupled to mass spectrometry. *Food Chemistry, 258*, 284–294. https://doi.org/10.1016/j.foodchem.2018.03.026.

Milev, B. P., Patras, M. A., Dittmar, T., Vrancken, G., & Kuhnert, N. (2014). Fourier transform ion cyclotron resonance mass spectrometrical analysis of raw fermented cocoa beans of Cameroon and Ivory Coast origin. *Food Research International, 64*, 958–961. https://doi.org/10.1016/j.foodres.2014.07.012.

Oliveira, L. F., Braga, S. C. G. N., Augusto, F., Hashimoto, J. C., Efraim, P., & Poppi, R. J. (2016). Differentiation of cocoa nibs from distinct origins using comprehensive two-dimensional gas chromatography and multivariate analysis. *Food Research International, 90*, 133–138. https://doi.org/10.1016/j.foodres.2016.10.047.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pluskal, T., Castillo, S., Villar-Briones, A., & Orešič, M. (2010). MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics, 11*(1), 395. https://doi.org/10.1186/1471-2105-11-395.

Roullier-Gall, C., Boutegrabet, L., Gougeon, R. D., & Schmitt-Kopplin, P. (2014). A grape and wine chemodiversity comparison of different appellations in Burgundy: Vintage vs terroir effects. *Food Chemistry, 152*, 100–107. https://doi.org/10.1016/j.foodchem.2013.11.056.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). Biometrika, 52(3/4), 591–611. JSTOR. 10.2307/2333709.

Sirbu, D., Grimbs, A., Corno, M., Ullrich, M. S., & Kuhnert, N. (2018). Variation of triacylglycerol profiles in unfermented and dried fermented cocoa beans of different origins. *Food Research International, 111*, 361–370. https://doi.org/10.1016/j.foodres.2018.05.025.

Vázquez-Ovando, A., Molina-Freaner, F., Nuñez-Farfán, J., Betancur-Ancona, D., & Salvador-Figueroa, M. (2015). Classification of cacao beans (Theobroma cacao L.) of southern Mexico based on chemometric analysis with multivariate approach.

European Food Research and Technology, 240(6), 1117–1128. https://doi.org/10.1007/s00217-015-2415-0.

Zabell, A., Stone, J., & Julian, R. (2017). Using big data for LC-MS/MS quality analysis. *Clinical Laboratory News, 43*(5), 30–31.

Zhang, D., & Motilal, L. (2016). Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity. In B. A. Bailey, & L. W. Meinhardt (Eds.), *Cacao Diseases: A History of Old Enemies and New Encounters* (pp. 3–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-24789-2_1.