Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of Processing: Effectiveness of Combined Untargeted and Targeted Approaches for Classification and Discrimination

Comprehensive Chemical Fingerprinting of High-Quality Cocoa at Early Stages of Processing: Effectiveness of Combined Untargeted and Targeted Approaches for Classification and Discrimination

Federico Magagna[1], Alessandro Guglielmetti[1], Erica Liberto[1], Stephen E. Reichenbach[2], Elena Allegrucci[3], Guido Gobino[3], Carlo Bicchi[1] and Chiara Cordero[1*]

Authors' affiliation:

[1] Dipartimento di Scienza e Tecnologia del Farmaco, Università di Torino, Turin, Italy

[2] Computer Science and Engineering Department, University of Nebraska, Lincoln NE, USA

[3] Guido Gobino Srl, Turin, Italy

* Address for correspondence:

Prof. Dr. Chiara Cordero - Dipartimento di Scienza e Tecnologia del Farmaco, Università di Torino, Via Pietro Giuria 9, I-10125 Torino, Italy – e-mail: chiara.cordero@unito.it ; phone: +39 011 6707662; fax: +39 011 2367662

Abstract

This study investigates chemical information in the volatile fractions of high-quality cocoa (*Theobroma Cacao* L. Malvaceae) from different origins (Mexico, Ecuador, Venezuela, Colombia, Java, Trinidad, and Sao Tomè) produced for fine chocolate. The study explores the evolution of the entire pattern of volatiles in relation to cocoa processing (raw, roasted, steamed, and ground beans). Advanced chemical fingerprinting (e.g., combined *U*ntargeted and *T*argeted (UT) *fingerprinting*) with comprehensive two-dimensional gas chromatography (GC×GC) coupled with mass spectrometry (MS) enables advanced pattern recognition for classification, discrimination, and sensory-quality characterizations. The entire data-set is analysed for 595 reliable 2D peak-regions, including 130 known analytes and 13 potent odorants. Multivariate analysis (MVA) with unsupervised exploration (principal component analysis (PCA)) and simple supervised discrimination methods (Fisher ratios and linear regression trees) reveal informative patterns of similarities and differences and locate characteristic compounds related to samples origin and manufacturing step.

Key-words

Theobroma Cacao L.; combined untargeted and targeted fingerprinting; comprehensive two-dimensional gas chromatography-mass spectrometry; classification and discrimination models; key-aroma compounds

Introduction

Cocoa, produced from cocoa beans (*Theobroma Cacao* L. *Malvaceae* family), is a crop of great economic relevance as the main raw ingredient for chocolate manufacturing.[1] Cocoa and chocolate are consumed worldwide and their popularity is primarily related to the pleasant sensory properties, although, recent evidence of several health benefits open new market perspectives and potential use in functionalized food(s).[2–7]

Theobroma cacao L. is a tree crop native to tropical forests of American continent. Recent studies, focused on cocoa germoplasm[8], defined 10 major genetic clusters, or groups named: Marañon, Curaray, Criollo, Iquitos, Nanay, Contamana, Amelonado, Purŭs, Nacional and Guiana. The new classification reflected accurately the genetic diversity available overcoming the traditional classification as Criollo, Forastero or Trinitario.

Cocoa quality and economic value are more strictly related to the unique and complex flavours. The sensory profile (aroma, taste, mouth feeling, and texture) is a key-factor in obtaining premium quality products suited to consumer preferences. Flavours develop from complex biochemical and chemical reactions occurring at post-harvesting and vary with genotype, geographical origin, farming practices, and technological processing.[9] Above all, post-harvest treatments and, in particular, fermentation [10,11] and roasting [12] are key steps in the formation of the characteristic cocoa aromas: in fact, the roasting of unfermented beans results in a product with a poor and unsatisfactory aroma profile.[13] Over the last few decades, several hundreds of volatiles have been identified in cocoa volatile fractions, including potent odorants whose particular distribution provides a diagnostic indicator for aroma qualification and products discrimination. The *molecular sensory science* approach, for example, was adopted to identify the *aroma blueprint* of different cocoa and chocolate products[14–16] while HS-SPME coupled with mono-dimensional (1D) GC-MS analysis is the technique of choice for cocoa volatile organic

compounds (VOCs) investigations.[17,18] Hyphenated techniques like in-line roasting in cooled injectors (ILR-CIS) and GC-MS were proposed for assessing process quality[19] and HS-SPME-GC-MS and direct MS-fingerprinting were combined to characterize cocoa volatiles.[20]

In this context, multidimensional analytical techniques, especially comprehensive two-dimensional gas chromatography (GC×GC) coupled with mass spectrometry (MS) are promising, powerful approaches for detailed characterization of the complex mixtures of cocoa volatiles as it has been proven for other foods[21,22]. GC×GC exploits the separation and detection potential of two separation dimensions providingincreased separation power, meaningful 2D chromatographic patterns with analytes structurally ordered in the chromatographic plane and enhanced sensitivity derived from the band focusing during modulation.[23–25] Compared to 1D platforms, GC×GC-MS improves the effectiveness of sample profiling, fingerprinting and, thereby, classification and discrimination.[21,24,26–29]

In the panorama of existing studies, only a few have exploited the full potential of GC×GC to explain the complex information in cocoa volatile fractions or proposed effective methods capable of replacing multiple, less-informative, 1D separation methods based on targeted analysis. In 2009, Humston and co-workers [30] developed and evaluated an analytical procedure combining HS-SPME and GC×GC with time-of-flight (TOF) MS, to study volatiles from cocoa beans of different geographical origin, at two storage conditions, and with low or high moisture content. Within the entire set of detectable analytes, they identified four compounds (i.e., *acetic acid*, *nonanal*, *tetramethylpyrazine*, and *trimethylpyrazine*) showing consistent quantitative changes depending on bean storage and not on cocoa origin.

More recently, Oliveira et al.[31] investigated the volatile fraction of cocoa nibs from Brazil and Ivory Coast by HS-SPME-GC×GC-MS and GC×GC-FID to select informative analytes for samples differentiation. First, they applied PCA on GC×GC-FID data to evaluate samples

clustering; then, selected samples were submitted to GC×GC-MS to identify the most informative compounds. Within 20 identified analytes, 15 were found to be present in different amounts in samples of the two origins under study.

The present study investigates the unique VOCs signatures from commercial grade, high-quality cocoa with a novel pattern recognition strategy that combines untargeted and targeted fingerprinting to GC×GC-MS data. Samples of interest for fine chocolate production, and from different geographical provenience (Mexico, Ecuador, Venezuela, Colombia, Java, Trinidad, and Sao Tomè) are studied along the early stages of industrial processing (raw, roasted, steamed, and nibs). The complex fraction of volatiles is extracted by automated HS-SPME sampling and subsequently analyzed by GC×GC-MS with thermal modulation. Advanced pattern recognition by UT fingerprinting strategy[32] is tested to validate its effectiveness to exploit chemical information encrypted in VOCs signatures. 2D data matrices are mined to explore different issues such as origin/process characteristics and sensory profile(s) differentiation.

Materials and methods

Reference compounds and cocoa samples

Pure reference standards for identity confirmation (key-aroma compounds and informative volatiles) of *acetic acid*, *3-methylbutanoic acid*, *3-methylbutanal*, *2-phenylethanol*, *2-heptanol*, *butanoic acid*, *2-methylbutanal*, *linalool*, *phenylacetaldehyde*, *2-ethyl-3,5-dimethylpyrazine*, *4-hydroxy-2,5-dimethyl-3(2H)-furanone*, *2-ethyl-3,6-dimethylpyrazine*, (*E,E*)-*2,4-nonadienal, dimethyl trisulfide*, *2-methylpropanoic acid*, *ethyl-2-methylbutanoate*, and *n*-alkanes (*n*-C9 to *n*-C25) for Linear Retention Index (I^T_S) determination were from Sigma-Aldrich (Milan, Italy).

Internal standards (ISTDs) for analyte response normalization were α- and β-*thujone* from Sigma Aldrich (Milan, Italy). A standard stock solution of ISTDs at 100 mg/L was prepared in *dibuthylphtalate* (Sigma-Aldrich, Milan, Italy) and stored in a sealed vial at -18°C.

High-quality cocoa samples (*Theobroma cacao* L.) of commercial grade were selected by confectionery experts on the basis of their peculiar sensory characteristics. Descriptive sensory analysis (data not shown) was performed by company internal panel to drive processing parameters toward a desirable sensory quality. Origins were: Ecuador, Venezuela, Colombia, Trinidad, Mexico from Chontalpa region of Tabasco, Java and Sao Tomè. Samples information is provided as supplementary material in Supplementary Table 1 - ST1. Chontalpa is a top-quality area for cocoa production that was recognized by the Slow Food Presidium in 2007 after severe floods destroyed most of the cocoa plantations.

All samples were harvested in 2014; they were analyzed at four different technological stages:raw, roasted, steamed nibs obtained after the removal of bean shells (4 processing steps).

Processing was by Guido Gobino srl (Turin, Italy) in three replicated batches using time and temperature protocols between 100 and 130°C for a timing from 20 up to 40 minutes. Processing was optimized for each origin and driven by a desirable flavour development. Hot-air roasting was conducted in a vertical roaster designed by Bühler AG (Uzwil, Switzerland).

Cocoa samples were freeze in liquid nitrogen immediately after each step of processing and then stored at -80°C. Before headspace analysis, samples were ground in a laboratory mill up to about 300 µm (Grindomix GM200, Retsch, Haan, Germany); particle size homogeneity was verified by visual inspection. The resulting cocoa powder was then precisely weighted (1.500 g) in headspace glass vials (20 mL) and submitted to automated HS-SPME sampling.

Automated Head Space Solid Phase Micro Extraction: sampling devices and conditions

Automated HS-SPME was performed using a MPS-2 multipurpose sampler (Gerstel, Mülheim a/d Ruhr, Germany) installed on the GC×GC-MS system. SPME fibers, Divinylbenzene/Carboxen/Polydimethyl siloxane (DVB/CAR/PDMS) d_f 50/30 µm - 2 cm were from Supelco (Bellefonte, PA, USA). Fibers were conditioned before use as recommended by the manufacturer. The standard-in-fiber procedure was adopted to pre-load the ISTDs (α- and β-thujone) onto the fiber before sampling. 5.0 µL of ISTDs solution were placed into a 20 mL glass vial and submitted to HS-SPME at 50°C for 10 min.

After ISTDs loading, the SPME device was exposed to the headspace of cocoa samples (1.500 g) for 40 min at 50°C. Extracted analytes were recovered by thermal desorption of the fiber into the split/splitless (S/SL) injection port of the GC×GC system at 250°C for 5 min. Each sample was analyzed in duplicate.

GC×GC-MS instrument set-up and analytical conditions

GC×GC analyses were performed on an Agilent 6890 GC unit coupled with an Agilent 5975C MS inert detector operating in the EI mode at 70 eV (Agilent, Little Falls, DE, USA). The transfer line was set at 270°C. An *Auto Tune* option was used and the scan range was set at *m/z* 40-240 with a scan rate of 12,500 amu/s to obtain a sampling frequency of 28 Hz.

The system was equipped with a two-stage KT 2004 loop thermal modulator (Zoex Corporation, Houston, TX) cooled with liquid nitrogen and controlled by Optimode™ V.2 (SRA Instruments, Cernusco sul Naviglio, MI, Italy). Hot jet pulse time was set at 250 ms, modulation time was 3s, and cold-jet total flow was progressively reduced with a linear function from 40% of Mass Flow Controller (MFC) at initial conditions to 8% at the end of the run. A deactivated fused silica capillary loop (1 m × 0.1 mm d_c) was used.

The column set was configured as follows: ^{1}D SolGel-Wax column (100% polyethylene glycol) (30 m × 0.25 mm d_c, 0.25 µm d_f) from SGE Analytical Science (Ringwood, Australia) coupled with a ^{2}D OV1701 column (86% polydimethylsiloxane, 7% phenyl, 7% cyanopropyl) (1 m × 0.1 mm d_c, 0.10 µm d_f), from J&W (Agilent, Little Falls, DE, USA).

SPME thermal desorption into the GC injector port was under the following conditions: split/splitless injector in split mode, split ratio 1:5. Carrier gas was helium at a constant flow of 1.2 mL/min. The oven temperature program was: from 40°C (1 min) to 200°C at 3°C/min and to 250°C at 10°C/min (5 min).

The *n*-alkanes liquid sample solution for I^{T}_{S} determination was analyzed under the following conditions: split/splitless injector in split mode, split ratio 1:50, injector temperature 250°C, and injection volume 2 µL.

Data acquisition and data elaboration

Data were acquired by Agilent MSD ChemStation *ver* D.02.00.275 and processed by GC Image® GC×GC Edition Software, Release 2.6 (GC Image, LLC Lincoln NE, USA). Statistical analysis was performed with XLstat (Addinsoft, New York, NY USA).

UT fingerprinting work-flow

*U*ntargeted and *T*argeted (UT) *fingerprinting* was carried out by the template matching approach, introduced by Reichenbach and co-workers in 2009[33] and following a work-flow previously validated for olive oil volatiles investigation.[32] The approach uses metadata collected from 2D peak patterns (retention times, MS fragmentation patterns, and single ions and/or total ions response) and establishes reliable correspondences between the same chemical entities

across multiple chromatograms. The output is a data matrix of aligned 2D peaks and peak-regions and their related metadata available for comparative purposes and further processing.

Targeted analysis focused on 130 compounds tentatively identified by matching their EI-MS fragmentation pattern (NIST MS Search algorithm, ver 2.0, National Institute of Standards and Technology, Gaithersburg, MD, USA, with Direct Matching threshold 900 and Reverse Matching threshold 950) with those collected in commercial (NIST2014 and Wiley 7n) and in-house databases. As a further check for identification, experimental Linear Retention Indices (I^T_S) were computed and compared to the tabulated indices.[34]

Untargeted analysis was based on peak-regions features[35,36] and was performed automatically by GC Image Investigator™ R 2.6 (GC-Image LLC, Lincoln NE, USA). The untargeted analysis included *all peak-regions* above the fixed peak response threshold of 5,000 counts together with all targeted peaks and related metadata. This process[32,35–39] aligned the feature template to each of the 168 chromatograms (7 cocoa origins × 4 technological steps × 3 technical batches × 2 analytical replicates) using a set of *registration peaks* that were reliably matched across all chromatograms. The resulting data matrix for untargeted and targeted reliable peak-regions was 168 × 595; column bleeding and SPME fiber interferent peaks were removed before chemometric analysis. Response data from all cross-aligned 2D peak-regions were used for multivariate analysis (MVA) and supervised discrimination approaches (Fisher ratio and regression trees).

Fisher ratios were used to measure class separation for individual features relative to the variance within classes. For the same number of observations in two classes, the square-root of the Fisher ratio is the t-value. For more than 20 samples (e.g., 21 samples at each of the four processing stages), a Fisher ratio of 1 has a p-value of 16%, a Fisher ratio of 1.77 exceeds 90% confidence, and a Fisher ratio of 6.45 exceeds 99% confidence. In this study, Fisher ratios (F value)

were calculated during the UT fingerprinting elaboration by the Image Investigator™ (GC Image v2.6) on normalized 2D peak-region volumes considering each class against the superset of all other classes (one *vs.* all).

Repeatability and intermediate precision results on retention times ($^{1}t_R$ and $^{2}t_R$) and on Normalized 2D volumes is reported as supplementary material. Repeatability was evaluated on single batch Chontalpa nibs replicate analyses over a three days time interval (three replicated samples) while intermediate precision was calculated on ISTDs (α- and β-thujone) 2D peaks from all nibs samples analyzed over the one-month period (42 runs). Data refers of good method precision[40] on both: (a) retention times, where RSD % ranges from 0.06 to 3.43 (average value 0.59) for ^{1}D and from 0.83 and 6.12 (average value 2.68) for the ^{2}D. Normalized 2D Volumes were always below 20% with an average RSD of 6.85%. 2D Normalized Volumes of the two ISTDs (α- and β-thujone), monitored over a wider period, never exceeded the 14% of RSD.

Results and discussion

This study exploits the power of GC×GC-MS for the detailed chemical profiling of complex samples, harnessing its intrinsic potential as a highly informative fingerprinting tool. Thanks to dedicated pattern recognition approaches, the large amount of (chemical) information encrypted in cocoa volatiles distributions, can be rationalized and mined to find compositional similarities/differences (fingerprinting) and to explain the informative role of single chemicals, whose distribution provides indications for origin traceability, effects of manufacturing processes, and aroma quality.

The following sections illustrate: (*a*) the chemical complexity of the volatile fraction of high-quality *Theobroma cacao* samples, as revealed by combining targeted and untargeted investigations; (*b*) the particular distributions of informative analytes (key-aroma compounds

214 and technological sensitive analytes) within samples and their evolution along processing steps,
215 (*c*) how simple supervised approaches could support the selection of informative chemicals to
216 discriminate samples.
217
218 **Information encrypted on cocoa volatiles distribution**
219 The high chemical complexity of cocoa volatile fractions results from many chemical
220 reactions, most of them catalyzed by specific enzymes (endogenous or exogenous from moulds,
221 yeasts and bacteria) and occurring at the different stages of its processing. Influential factors
222 have been extensively reviewed by Afoakwa et al[9,41] and include some of the variables considered
223 in our sampling design: roasting (time/temperature) and other physical and mechanical
224 treatments such as debacterization by steaming, and grinding.
225 Within the 595 detected VOCs by GC×GC-MS (peak-regions corresponding to detectable
226 analytes in at least two samples of the set), 130 analytes were tentatively identified and reported
227 in **Table 1**. Each analyte is characterized by absolute retention times (^1t_R - min and ^2t_R - sec),
228 experimental I^T_S, and odor descriptors as reported in reference literature.
229 Figure 1 visualizes a heat-map of the relative distributions (Normalized 2D Peak Volumes)
230 of 595 untargeted peak-regions, including the 130 known analytes. Columns follow processing
231 stages from raw to grinded beans after steaming. Analytes are ordered according to their inter-
232 class variance. Normalized peak volumes values were mean and centered before colorization.
233 Color scale varies between red (low abundance) to green (high abundance).
234 The evolution of the volatiles profile along the different steps of processing is illustrated
235 by changes in the heat-map colour spots (Fig. 1). In particular, after roasting and steaming, when
236 volatiles are developed from their non-volatile aroma precursors, dark spots predominate while

several analytes, already present in raw beans, increase their relative abundances (quantitative changes).

Potent odorants distribution within samples and their evolution along processing

Within the volatile fraction, the most significant changes occur for key-aroma and some technologically sensitive analytes (technological markers), in close accordance with reference studies.[1,42,41]

Cocoa key-aroma compounds, identified by Schieberle and co-workers,[14–16] deserve a detailed discussion, in that their distribution is fundamental for aroma properties. They include several chemical classes, especially alkyl pyrazines (*2,3,5-trimethylpyrazine*, *2-ethyl-3,5-dimethylpyrazine*, and *3,5-diethyl-2-methylpyrazine*) which impart characteristic earthy notes. Another important set of key-volatiles are short-chain and branched fatty acids: *acetic acid*, *butanoic acid*, *2-methylpropanoic acid*, and *3-methylbutanoic acid*, whose presence, at high concentrations, can impart off-flavours due to their rancid, sour, and sweaty notes. Strecker aldehydes (*2*- and *3-methylbutanal*), formed during fermentation and roasting, impress malty and buttery notes, and *phenylacetaldehyde*, derived from L-phenylalanine (L-Phe), is responsible for a pleasant honey-like note. Other key analytes are esters (*ethyl-2-methylbutanoate* – fruity, *2-phenylethyl acetate* – flowery), linear alcohols (*2-heptanol* – citrusy), phenyl propanoids derivatives (*2-phenylethanol* – flowery), and sulphurous derived compounds (*dimethyl trisulfide*).

Raw cocoa beans (just fermented) have specific distributions of potent odorants related to origin. Profiling data, in agreement with reference studies,[1,41] show that the volatile fraction of raw beans is dominated by short-chain fatty acids, especially *3-methylbutanoic* and *acetic acid*, that result from the enzymatic degradation of the pulp during fermentation. In particular, a*cetic acid* is the most abundant volatile and is present at high levels in unroasted beans (high Odour

Activity Value [16]), giving an intense vinegar-like perception which can affect cocoa aroma quality. However, during cocoa processing (roasting, above all) and later, during chocolate manufacturing (conching and refining), undesired volatiles with low boiling points, such as *acetic acid*, are removed resulting in a drastic decrease of its concentration (up to 70%).[16]

During the fermentation of raw beans, non-volatile aroma precursors obtained through the degradation of seeds storage proteins and carbohydrates react, mainly under enzymatic control, and generate odor-active volatiles (alcohols, esters, aldehydes, and organic acids). Bacteria and moulds are fundamental at this stage[9,41].

Roasting has a larger impact on aroma: alkyl pyrazines and Strecker aldehydes (*3-methylbutanal* and, in some cases, *phenylacetaldehyde*) show a large increase after this stage. Roasting has only a minor impact on *3-methylbutanoic acid* and esters (rancid smelling), which were detected in similar amounts before and after this process. As general consideration, the differences in the volatile profiles between unroasted and roasted beans are quantitative rather than qualitative. Supplementary Figure 1 (SF1) illustrates GC×GC patterns and their evolution across stages for cocoa harvested in the Chontalpa region (Tabasco, Mexico). Relative distribution differences of some potent odorants, between raw and roasted beans, are visually shown, in logarithmic scale, on spider diagrams in Figure 2. Sample origins illustrated are Chontalpa, Mexico (2A); Venezuela (2B); and Sao Tomè (2C) and quantitative changes refer to raw (green lines) and roasted (brown lines) stages.

The Chontalpasample from Mexico (2A) average profile shows a remarkable increase for the Strecker aldehyde *3-methylbutanal*; alkyl pyrazines, with earthy and roasty notes; *2-heptanol*, with a citrusy smell; and *dimethyl trisulfide*, whereas the amounts of other characteristic odorants (*butanoic acid*, *3-methylbutanoic acid*, *ethyl-2-methylbutanoate*, *phenylethylalcohol*, etc.) remain rather similar, even after roasting.

285 The distribution profile of the Venezuela sample (2B) is characterised by a more significant
286 increase (compared to Chontalpa) of *3-methylbutanal* (malty odour), a significant increase for
287 *phenylacetaldehyde* (opposite the trend for Chotalpa), and no change for *2-heptanol*.
288 The Sao Tomè sample (2C) shows a different behaviour: even though some aroma and
289 technological markers increase after roasting (*3-methylbutanal* and pyrazines), the raw and
290 roasted cocoa present very similar patterns, as is the case for the Java sample (data not shown).
291
292
293 ***U*ntargeted and *T*argeted (UT) *Fingerprinting* results**
294 The distribution of all detected VOCs (known and unknown analytes) is a potentially
295 informative fingerprint for geographical origin and manufacturing stage differentiation.
296 Unsupervised multivariate analysis, i.e., PCA, was applied to map the natural conformation
297 (groups) of samples and to localize informative chemicals responsible for variations.
298 In the first step, PCA was performed on the matrix combining information from 130
299 targeted analytes in samples with different origins (CH-Chontalpa, VE-Venezuela, CO-Colombia,
300 EC-Ecuador, JA-Java, TR-Trinidad, ST-Sao Tomè), manufacturing stages, and processing batches
301 (3). Analytical replicates (2) were averaged. Auto-scaling was applied as pre-processing step and
302 baseline correction was performed on the 2D data by GC Image software.
303 Figures 3A-B show the scores plot for the first two principal components (F1-F2 plane) for
304 raw (Fig.3A) and roasted (Fig.3B) cocoa, based on the 84 × 130 matrix (samples × targets). The
305 variance explained by the first two components was similar in all elaborations (including those
306 based on steamed and nibs, not shown), ranging from a minimum of 48.83% for roasted samples
307 (30.22 % for F1, 18.61% for F2) to the 53.74 % of raw cocoa (30.00 % for F1 and 23.74 % for F2).
308 Origin dominates group conformation and grouping is maintained through manufacturing steps.

In particular, PCA clusters cocoa samples in three main sub-groups. In the first sub-group (highlighted with blue circles), cocoa from Ecuador, Venezuela and Colombia are close together at all stages. This outcome is consistent with their aroma profiles, considered relatively similar by confectionery experts. In the second sub-group (green circles), cocoa from Trinidad has a distinctive chemical fingerprint that yields independent clustering at all stages. In the third sub-group (red circles), Chontalpa,Java (and Sao Tomè show more similar chemical fingerprints, despite their different geographical provenience.

Cocoa clustering results from different variables (loadings plots not reported): for example, raw beans from Chontalpa and Sao Tomè had higher levels of some potent odorants such as a*cetic acid* (sour), *phenylacetaldehyde* (honey-like), *2-phenylethanol*, and other volatiles such as esters (*ethyl hexanoate*, *ethyl octanoate*, and *ethyl decanoate*) and organic acids. The volatiles signatures of South America cocoa (Ecuador, Colombia, Venezuela) is connoted by the presence of short chain primary alcohols (*1-butanol*, *1-pentanol*, *1-hexanol*, *2-ethyl-1-hexanol*, *2-hexanol*, and 2-*heptanol* (citrusy)) and *3-methylbutanoic acid* (rancid). These analytes (esters, alcohols, and acids) and some detectable linear aldehydes (*hexanal*, *octanal* and *nonanal*) are formed mostly during fermentation. The cluster of roasted samples from Chontalpa, Java, and Sao Tomè have a distinctive fingerprint of *alkyl pyrazines* (*2,3,5-trimethyl*, *2-ethyl-3,5-dimethyl*, *2-ethyl-5(6)-methyl* and *3,5-diethyl-2-methyl pyrazine*), important processing markers. Roasted beans of South American cocoa are connoted by higher amounts of aromatic ketones (*1-hydroxy-2-propanone*, *2,3-pentanedione* and *2,3-butanedione*) and other volatiles such as *1H-pyrrole-2-carboxaldehyde* and *2-furanmethanol*.

Fisher ratio values were therefore used for supervised ranking and selection of highly informative features characterising the chemical fingerprints of different sample sets. Fisher ratios (F value) were calculated automatically during the *UT fingerprinting* elaboration on

normalized 2D peak-region volumes considering each class against the superset of all other classes (one *vs.* all).

Figure 4 shows bar plots of F values for classes of three origins (Chontalpa-Mexico, Java, and Trinidad) and two processing steps (roasting and steaming), with an arbitrarily fixed cut-off of 30. As seen in this plot, several analytes are distinctive for origin independent of processing (those with paired cyan and orange bars). In most cases, cocoa origins are described by the same variables at roasted and/or steamed stage.

Chontalpa and Java, which clustered together with Sao Tomè in the PCA elaboration, have distinctive signatures: Java has a characteristic distribution of alkyl pyrazines (*tetramethyl-*, *2-ethyl-3,5-dimethyl-*, *2,3,5-trimethyl-* and *3,5-diethyl-2-methylpyrazine*) that is preserved after steaming, with most of those F values increasing (e.g., from 67 to 413 for *tetramethylpyrazine*, from 320 to 820 for *2-ethyl-3,5-dimethylpyrazine*, etc.) indicating a stronger diagnostic role.

Chontalpa from Mexico is characterized by esters, responsible for fruity notes, which probably derive from fermentation processes. The most significant ones are *hexyl acetate* (F value 1139 for roasted, but only 73 for steamed) and *1-butanol-3-methyl acetate* (F value 440 for roasted and 409 for steamed). Moreover, *3-hydroxy-2-butanone*, a technological marker influencing buttery perception, plays a less significant role for both roasted and steamed samples.

Cocoa from Trinidad, independently clustered at all stages of processing, is connoted by a distinctive signature of phenyl-propanoid derivatives (*benzaldehyde* and *2-phenylalcohol*), some process markers (*2,6* and *2,3-dimethylpyrazine*), and *trans-linalool oxide*. The highest F value (495) is observed in the roasted sample for *ethyl butanoate* (sweet, fruity), an analyte that does not keep its information potential after steaming.

To confirm the results obtained with the targeted fingerprinting and to evaluate if new informative markers could be revealed within the entire volatile fraction, the study was extended to all detected analytes, including unknowns. The set of 595 peak-regions, included the 130 target analytes (tentatively identified), was thereby used to validate targeted analysis results.

Figures 3C-D visualise PCA results with the 595 reliable peak-regions for raw (Fig. 3C) and roasted (Fig.3D) cocoa. Results are highly consistent with those from the targeted peaks distributions. Samples are clustered into three groups: Ecuador-Venezuela-Colombia (blue circles), Chontalpa-Sao Tomè-Java (red circles), and Trinidad (green circles). The total explained variability here ranges from 40.66% for roasted beans to 48.50% for steamed cocoa (data not shown).

Targeted peak-regions, included in the untargeted approach, cross-validate the classification based on previous PCAs: samples are described by almost the same variables and no additional informative roles of unknown features were hypothesized. This approach clearly highlights the strong accordance between targeted and untargeted fingerprinting for sample classification purposes suggesting that for some applications, untargeted fingerprinting is effective, efficient, and less time-consuming than targeted analysis.

Samples classification and discrimination: variables selection strategy

The classification and prediction potential of the proposed approach has a high risk of over-fittings due to the large number of analyte variables and the limited number of the samples under study. However, to demonstrate the flexibility of such comprehensive fingerprinting for pattern recognition, simple classification approaches have been adopted to define key-variables (explanatory quantitative variables) suitable to discriminate one sample, or a group of them, from others. This is illustrated by two following examples: (a) the identification of a univocal set

of processing variables capable of distinguishing raw from processed cocoa independently of origin and (b) the definition of origin-specific variables sensitive to thermal treatments (roasting and steaming).

The explanatory approach adopted was a regression tree analysis based on the CHAID algorithm.[43,44] The entire set of samples × target analyte variables was explored to find univocal variables indicating the effect of processing on the raw cocoa, independent of origin. The sample set was divided into estimation samples and validation samples. The validation set included the second of the three replicated batches of analyses (28 samples, then not included in the estimation set). The resulting regression tree correctly classified all samples from the estimation/training set (i.e., the confusion matrix for all processing steps had 100% true positives). In the validation test, the predictive model failed in classifying five steamed samples belonging to the nibs (3) and roasted (2) classes, but it was successful for all others (i.e., better than 82% correct). The most informative classification variable for discriminating raw from processed cocoas was *2,3-dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one*. Its formation, promoted by heating, is related to the presence of fructose and β-alanine in raw cocoa.[45] Variables with a secondary role for discriminating processing stages were: *2,6-dimethylpyrazine*, *2,3,5-trimethylpyrazine*, *2-ethyl-5-methylpyrazine*, and the potent odorant (*E)-2-phenyl-2-butenal* (intense chocolate note).

Figure 5A shows the samples distribution as a function of two discriminating variables: *2,3-dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one* and *2-ethyl-5-methylpyrazine*. Raw cocoa (green markers in Fig. 5A) is clearly differentiated by processed derivatives independent of the origin; as those samples are closely clustered in the bottom-left of plot. Roasted cocoa is relatively well distinguished, but more dispersed along the *x*-axis, with *2,3-dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one* increasing from left to right. Steamed and ground samples are less

404 differentiated along the *y*-axis, representing the relative abundance of the earthy pyrazine (*2-*
405 *ethyl-5-methylpyrazine*).

406 The second model was developed to discriminate cocoa nibs (i.e., the last stage of
407 processing considered here) based on their origin. In this case, the model was effective with just
408 three variables: *2-pentylfuran*, *2,3,5-trimethylpyrazine*, and *linalool*. Figure 5B shows the
409 distribution of samples in three variables: *x*-axis *linalool*; *y*-axis*2,3,5-trimethylpyrazine* ; and
410 bubble-size *2-pentylfuran*). This model for nibs discrimination confirms what it was shown by
411 unsupervised approaches (PCA on targeted and on UT data, shown in Fig. 3). Samples from
412 Ecuador and Colombia are aligned along x axis (higher abundance of *linalool*) together with the
413 Venezuela samples. Java samples are connoted by a strong pyrazines signature (2*,3,5-*
414 *trimethylpyrazine* is one of the most origin sensitive), whereas Chontalpa, Sao Tomè, and
415 Venezuela samples are coherently positioned in the Cartesian space with lower amounts of both
416 chemicals.

References

(1) Aprotosoaie, A. C.; Luca, S. V.; Miron, A. Flavor Chemistry of Cocoa and Cocoa Products — An Overview. *Compr. Rev. Food Sci. Food Saf.* **2016**, *15*, 73–91.

(2) Andujar; Recio, M. C.; Giner, R. M.; Rios, J. R. Cocoa Polyphenols and Their Potential Benefits for Human Health. *Oxid. Med. Cell. Longev.* **2012**.

(3) Ackar, D.; Lendik, K. V.; Valek, M.; Subaric, D.; Milicevic, B.; Babic, J.; Nedic, I. Cocoa Polyphenols : Can We Consider Cocoa and Chocolate as Potential Functional Food ? *J. Chem.* **2013**.

(4) Araujo, Q. R. De; Gattward, J. N.; Almoosawi, S.; Conceição, G.; Costa, P.; Santana, P. A. De; Silva, M. D. G.; Dantas De Santana, A. P.; Araujo, Q. R. D. E.; Gattward, J. N. Cocoa and Human Health : From Head to Foot -A Review. *Crit. Rev. Food Sci. Nutr.* **2016**, *56*, 1–12.

(5) Hooper, L.; Kay, C.; Abdelhamid, A.; Kroon, P. A.; Cohn, J. S.; Rimm, E. B.; Cassidy, A. Effects of chocolate , cocoa , and flavan-3-ols on cardiovascular health: a systematic review and meta-analysis of randomized trials 1 – 3. *Am. J. Clin. Nutr.* **2012**, *95*, 740–751.

(6) Sokolov, A. N.; Pavlova, M. A.; Klosterhalfen, S.; Enck, P. Neuroscience and Biobehavioral Reviews Chocolate and the brain: Neurobiological impact of cocoa flavanols on cognition and behavior. *Neurosci. Biobehav. Rev.* **2013**, *37*, 2445–2453.

(7) Sarriá, B.; Martínez-lópez, S.; Sierra-cinos, J. L.; Garcia-diz, L.; Goya, L.; Mateos, R.; Bravo, L. Effects of bioactive constituents in functional cocoa products on cardiovascular health in humans. *Food Chem.* **2015**, *174*, 214–218.

(8) Motamayor, J. C.; Lachenaud, P.; da Silva e Mota, J. W.; Loor, R.; Kuhn, D. N.; Brown, J. S.; Schnell, R. J. Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PLoS One* **2008**, *3* (10).

(9) Kongor, E. J.; Hinneh, M.; Van De Walle, D.; Afoakwa, O. E.; Boeckx, P.; Dewettinck, K. Factors

influencing quality variation in cocoa (Theobroma cacao) bean flavour profile — A review. *Food Res. Int.* **2016**, *82*, 44–52.

(10) Camu, N.; De Winter, T.; Addo, S. K.; Takrama, J. S.; Bernaert, H.; De Vuyst, L. Fermentation of cocoa beans: influence of microbial activities and polyphenol concentrations on the flavour of chocolate. *J. Sci. Food Agric.* **2008**, *88*, 2288–2297.

(11) De Melo Pereira, G. V.; Soccol, V. T.; Soccol, C. R. Current state of research on cocoa and coffee fermentations. *Curr. Opin. Food Sci.* **2016**, *7*, 50–57.

(12) Ramli, N.; Hassan, O.; Said, M.; Samsudin, W.; Idris, N. A. Influence Of Roasting Conditions On Volatile Flavour Of Roasted Malaysian Cocoa Beans. *J. Food Process. Preserv.* **2006**, *30*, 280–298.

(13) Saltini, R.; Akkerman, R.; Frosch, S. Optimizing chocolate production through traceability: A review of the influence of farming practices on cocoa bean quality. *Food Control* **2013**, *29*, 167–187.

(14) Schnermann, P.; Schieberle, P. Evaluation of Key Odorants in Milk Chocolate and Cocoa Mass by Aroma Extract Dilution Analyses. *J. Agric. Food Chem.* **1997**, *45*, 867–872.

(15) Frauendorfer, F.; Schieberle, P. Identification of the Key Aroma Compounds in Cocoa Powder Based on Molecular Sensory Correlations. *J. Agric. Food Chem.* **2006**, *54*, 5521–5529.

(16) Frauendorfer, F.; Schieberle, P. Changes in Key Aroma Compounds of Criollo Cocoa Beans During Roasting. *J. Agric. Food Chem.* **2008**, *56*, 10244–10251.

(17) Ducki, S.; Miralles-Garcia, J.; Zumbè, A.; Tornero, A.; Storey, D. M. Evaluation of solid-phase micro-extraction coupled to gas chromatography – mass spectrometry for the headspace analysis of volatile compounds in cocoa products. *Talanta* **2008**, *74*, 1166–1174.

(18) Perego, P.; Fabiano, B.; Cavicchioli, M.; Del Borghi, M. Cocoa quality and processing a study by Solid-phase Microextraction and Gas Chromatography Analysis of Methylpyrazines. *Food*

Bioprod. Process. **2004**, *84*, 291–297.

(19) Van Durme, J.; Ingels, I.; De Winne, A. Inline roasting hyphenated with gas chromatography – mass spectrometry as an innovative approach for assessment of cocoa fermentation quality and aroma formation potential. *FOOD Chem.* **2016**, *205*, 66–72.

(20) Phuong, D. T.; Van De Walle, D.; De Clercq, N.; De Winne, A.; Kadow, D.; Lieberei, R.; Messens, K.; Tran, D. N.; Dewettinck, K.; Van Durme, J. Assessing cocoa aroma quality by multiple analytical approaches. *Food Res. Int.* **2015**, *77*, 657–669.

(21) Cordero, C.; Kiefl, J.; Schieberle, P.; Reichenbach, S. E.; Bicchi, C. Comprehensive two-dimensional gas chromatography and food sensory properties: Potential and challenges. *Analytical and Bioanalytical Chemistry*. 2015, pp 169–191.

(22) Cordero, C.; Schmarr, H.-G.; Reichenbach, S. E.; Bicchi, C. Current Developments in Analyzing Food Volatiles by Multidimensional Gas Chromatographic Techniques. *J. Agric. Food Chem.* **2017**, acs.jafc.6b04997.

(23) Cortes, H. J.; Winniford, B.; Luong, J.; Pursch, M. Comprehensive two dimensional gas chromatography review. *J. Sep. Sci.* **2009**, *32*, 883–904.

(24) Klee, M. S.; Cochran, J.; Merrick, M.; Blumberg, L. M. Evaluation of conditions of comprehensive two-dimensional gas chromatography that yield a near-theoretical maximum in peak capacity gain. *J. Chromatogr. A* **2015**, *1383*, 151–159.

(25) Adahchour, M.; Beens, J.; Brinkman, U. A. T. Recent developments in the application of comprehensive two-dimensional gas chromatography. *J. Chromatogr. A* **2008**, *1186* (1–2), 67–108.

(26) Cordero, C.; Bicchi, C.; Rubiolo, P. Group-type and fingerprint analysis of roasted food matrices (coffee and hazelnut samples) by comprehensive two-dimensional gas chromatography. *J. Agric. Food Chem.* **2008**, *56* (17), 7655–7666.

(27) Cordero, C.; Liberto, E.; Bicchi, C.; Rubiolo, P.; Schieberle, P.; Reichenbach, S. E.; Tao, Q. Profiling food volatiles by comprehensive two-dimensional gas chromatography coupled with mass spectrometry: Advanced fingerprinting approaches for comparative analysis of the volatile fraction of roasted hazelnuts (Corylus avellana L.) from different ori. *J. Chromatogr. A* **2010**, *1217* (37).

(28) Tranchida, P. Q.; Purcaro, G.; Dugo, P.; Mondello, L.; Purcaro, G. Modulators for comprehensive two-dimensional gas chromatography. *TrAC - Trends Anal. Chem.* **2011**, *30* (9), 1437–1461.

(29) Cordero, C.; Rubiolo, P.; Cobelli, L.; Stani, G.; Miliazza, A.; Giardina, M.; Firor, R.; Bicchi, C. Potential of the reversed-inject differential flow modulator for comprehensive two-dimensional gas chromatography in the quantitative profiling and fingerprinting of essential oils of different complexity. *J. Chromatogr. A* **2015**, *1417*, 79–95.

(30) Humston, E. M.; Zhang, Y.; Brabeck, G. F.; Mcshea, A.; Synovec, R. E. Development of a GC x GC – TOFMS method using SPME to determine volatile compounds in cacao beans. *J. Sep. Sci.* **2009**, *32*, 2289–2295.

(31) Oliveira, L. F.; Braga, S. C. G. N.; Augusto, F.; Hashimoto, J. C.; Efraim, P.; Poppi, R. J. Differentiation of cocoa nibs from distinct origins using comprehensive two-dimensional gas chromatography and multivariate analysis. *Food Res. Int.* **2016**, *90*, 133–138.

(32) Magagna, F.; Valverde-Som, L.; Ruíz-Samblás, C.; Cuadros-Rodríguez, L.; Reichenbach, S. E.; Bicchi, C.; Cordero, C. Combined Untargeted and Targeted fingerprinting with comprehensive two-dimensional chromatography for volatiles and ripening indicators in olive oil. *Anal. Chim. Acta* **2016**, *936*.

(33) Reichenback, S. E.; Carr, P. W.; Stoll, D. R.; Tao, Q. Smart Templates for Peak Pattern Matching with Comprehensive Two-Dimensional Liquid Chromatography. *J. Chromatogr. A* **2009**, *1216*

(16), 3458–3466.

(34) Adams, R. P. *Identification of Essential Oil Components by Gas Chromatography—Mass Spectroscopy*; Allured Publishing: New York, 1995.

(35) Reichenbach, S. E.; Tian, X.; Boateng, A. A.; Mullen, C. A.; Cordero, C.; Tao, Q. Reliable peak selection for multisample analysis with comprehensive two-dimensional chromatography. *Anal. Chem.* **2013**, *85* (10), 4974–4981.

(36) Reichenbach, S. E.; Tian, X.; Tao, Q.; Ledford, E. B.; Wu, Z.; Fiehn, O. Informatics for cross-sample analysis with comprehensive two-dimensional gas chromatography and high-resolution mass spectrometry (GCxGC-HRMS). *Talanta* **2011**, *83* (4), 1279–1288.

(37) Bressanello, D.; Liberto, E.; Collino, M.; Reichenbach, S. E.; Benetti, E.; Chiazza, F.; Bicchi, C.; Cordero, C. Urinary metabolic fingerprinting of mice with diet-induced metabolic derangements by parallel dual secondary column-dual detection two-dimensional comprehensive gas chromatography. *J. Chromatogr. A* **2014**, *1361*, 265–276.

(38) Reichenbach, S. E.; Tian, X.; Cordero, C.; Tao, Q. Features for non-targeted cross-sample analysis with comprehensive two-dimensional chromatography. *J. Chromatogr. A* **2012**, *1226*, 140–148.

(39) Reichenbach, S. E.; Rempe, D. W.; Tao, Q.; Bressanello, D.; Liberto, E.; Bicchi, C.; Balducci, S.; Cordero, C. Alignment for Comprehensive Two-Dimensional Gas Chromatography with Dual Secondary Columns and Detectors. *Anal. Chem.* **2015**, *87* (19), 10056–10063.

(40) Citac; Eurachem. Quantifying Uncertainty in Analytical Measurement. *English* **2000**, *2nd*, 126.

(41) Afoakwa, E. O.; Paterson, A.; Fowler, M.; Ryan, A. Flavor Formation and Character in Cocoa and Chocolate : A Critical Review. *Crit. Rev. Food Sci. Nutr.* **2008**, *48*, 840–857.

(42) Owusu, M.; Petersen, M. A.; Heimdal, H. EFFECT OF FERMENTATION METHOD, ROASTING AND CONCHING CONDITIONS ON THE AROMA VOLATILES OF DARK CHOCOLATE. *J. Food*

Process. Preserv. **2012**, *36*, 446–456.

(43) Loh, W.-Y. Fifty Years of Classification and Regression Trees. *Int. Stat. Rev.* **2014**, *82* (3), 329–348.

(44) Wilkinson, L. Tree structured data analysis: AID, CHAID and CART. *Proc. Sawtooth Softw. …* **1992**, 1–10.

(45) Nishibori, S.; Kawakishi, S. Formation of 2,3-Dihydro-3,5-dihydroxy-6-methyl-4(H)-pyran-4-one from fructose and b-Alanine under conditions used for baking. *J. Agric. Food Chem.* **1994**, *42*, 1080–1084.

(46) Rychlik, M.; Schieberle, P.; Grosch, W. *Compilation of Odor Thresholds, Odor Qualities and Retention Indices of Key Food Odorants*; Deutsche Forschungsanstalt für Lebensmittelchemie and Institut für Lebensmittelchemie der Technischen Universit€at München, Garching: Germany, 1998.

(47) Owusu, M.; Petersen, M. A.; Heimdal, H. Relationship of sensory and instrumental aroma measurements of dark chocolate as influenced by fermentation method, roasting and conching conditions. *J. Food Sci. Technol.* **2013**, *50* (5), 909–917.

Figure Captions

Figure 1: heat map showing *UT fingerprinting* results (untargeted and targeted) on 595 reliable peak-regions detected in the headspace of cocoa samples of seven origins and analyzed at each step of technological processing (raw, roasted, steamed and nibs). The heat-map quantitative descriptors (Normalized 2D Peak Volumes) are colorized according to a linear scale. Colour intensity goes from red (minimum) to light green (maximum).

Figure 2: Distribution of key-aroma compounds in raw/fermented (green line) and roasted (brown line) cocoa from Chontalpa-Mexico, Venezuela and Sao Tomè. Relative abundances reported in logarithmic scale refer to normalized 2D Peak Volumes.

Figure 3: Scores plots on the first two principal components (F1-F2 plane), based on the targeted fingerprinting of (3A) raw/fermented cocoa beans and (3B) roasted cocoa beans of all origins (Chontalpa, Mexico (CH), Ecuador (EC), Venezuela (VE), Colombia (CO), Java (JA), Sao Tomè (ST), Trinidad (TR). The complete set of untargeted + targeted peak-regions (i.e., 595 peak-regions above the fixed threshold of 5,000 counts) resulted in the distribution shown in 3C for raw and in 3D for roasted cocoa samples.

Each origin is represented by three processing batches while the two analytical replicates have been averaged before statistical analysis.

Figure 4: histograms with most significant Fisher Ratio values obtained with *one-vs-all* comparison; (**a**) roasted and (**b**) steamed Chontalpa, (**c**) roasted and (**d**) steamed Java, (**e**) roasted and (**f**) steamed Trinidad, (**g**) roasted and (**h**) steamed Ecuador. F values were selected by above the fixed threshold of 30.

Figure 5: dispersion graphs illustrating the discrimination potential of: (5A) *2,3-dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one* and *2-ethyl-5-methylpyrazine* on processed cocoa; and (5B) *2,3,5-trimethylpyrazine, linalool* and *2-pentylfuran* on cocoa nibs from different origin.

Table 1: list of targeted volatiles together with their absolute retention times (^1t_R min and ^2t_R sec), experimental I^T_S, informative role and odour descriptors as reported in the reference literature [14–16,46,47].

ID	Compound Name	^1t_R (min)	^2t_R (sec)	Exp I^T_S	Compound Confirmation	Informative Role	Odour descriptor
1	2-Methylpropanal	4.19	0.35	833	a	-	Green, pungent
2	Methyl acetate	4.59	0.52	853	a	-	-
3	2-Methyl tetrahydrofuran	4.94	0.69	870	b	-	-
4	Ethyl Acetate	5.09	0.59	878	a	-	Fruity, aromatic
5	2-Methylbutanal	5.44	0.64	895	a	-	Malty
6	3-Methylbutanal	5.50	0.65	898	a	Key-aroma marker	Malty
7	Ethanol	5.79	0.41	913	a	-	Ethanol-like
8	Ethyl propanoate	6.24	0.79	935	b	-	-
9	Ethyl-2-methylpropanoate	6.59	1.03	953	b	-	Fruity
10	2,3-Butanedione	6.64	0.55	955	a	Technological marker	Buttery
11	2-Pentanone	6.69	0.79	958	a	-	Fruity
12	Pentanal	6.84	0.79	965	a	-	Almond-like, pungent, malt
13	1-Methylpropyl acetate	6.84	1.03	966	b	-	-
14	2-Methylpropyl acetate	7.49	1.00	998	a	-	-
15	2-Butanol	7.60	0.55	1001	a	-	Winey
16	α-Pinene	7.69	1.93	1005	a	-	Harsh, terpene-like, minty
17	2-Ethyl-5-methyl-furan	7.89	1.00	1012	b	-	-
18	Ethyl butanoate	7.99	1.21	1016	b	-	Sweet, fruity
19	2-Methyl-3-Buten-2-ol	8.19	1.24	1022	b	-	-
20	Ethyl-2-methylbutanoate	8.54	1.41	1035	b	Key-aroma marker	Fruity
21	2,3-Pentandione	8.74	0.76	1041	a	Technological marker	Caramel
22	Ethyl-3-methylbutanoate	8.99	1.38	1050	b	-	Fruity
23	Dimethyl disulfide	9.13	0.83	1055	a	Technological marker	Sulfurous
24	2-Pentyl acetate	9.15	1.38	1056	b	-	-
25	Butyl acetate	9.19	1.14	1057	a	-	Fruity, herbaceous
26	Hexanal	9.54	1.14	1069	a	-	Tallowy, leaf-like
27	2-Methyl-1-propanol	9.69	0.59	1074	a	-	-
28	2-Methyl-2-butenal	10.04	0.86	1086	a	-	-
29	2-Pentanol	10.74	0.69	1108	a	-	Light, seedy, sharp
30	3-Methylbut-1-yl acetate	10.89	1.38	1112	a	-	-
31	Ethyl pentanoate	11.01	1.39	1115	b	-	Fruity, sweet
32	Butyl-2-methylpropanoate	11.19	1.83	1120	b	-	Fruity, sweet
33	4-Methyl-3-penten-2-one	11.29	1.00	1123	b	-	-
34	1-Butanol	11.64	0.59	1132	a	-	Winey
35	β-Myrcene	12.24	1.83	1148	a	-	-
36	1-Pentylacetate	12.79	1.41	1163	a	-	Fruity, metallic, green
37	2-Heptanone	13.19	1.38	1173	a	-	Sweet, fruity
38	2-Ethylhexanal	13.39	1.76	1179	a	-	-
39	Limonene	13.74	1.93	1188	a	-	Citrus, mint
40	2-Methyl-1-butanol	14.19	0.66	1200	a	-	Fermented, fatty
41	Pyrazine	14.39	0.75	1205	a	-	Earthy
42	Butyl butanoate	14.64	1.86	1211	b	-	Fruity, flowery, sweet

43	2-Hexanol	14.69	0.83	1212	a	-	Mushroom, green
44	Ethyl hexanoate	14.94	1.74	1219	a	-	Fruity
45	2-Pentylfuran	15.09	1.52	1222	b	-	Buttery, green bean-like
46	(*E*)-2-methyl-2-butenoate	15.39	1.31	1229	b	-	-
47	1-Pentanol	15.89	0.69	1241	a	-	Sweet, pungent
48	2,4-Dimethyl-3-pentanol	16.34	0.66	1252	b	-	-
49	Methylpyrazine	16.69	0.79	1260	a	Technological marker	Earthy
50	Hexyl acetate	16.89	1.62	1265	a	-	Fruity
51	3-Hydroxy-2-butanone	17.24	0.62	1274	a	Technological marker	Buttery
52	2-Octanone	17.49	1.55	1280	a	-	Mould, green
53	Octanal	17.64	1.59	1283	a	-	Fatty, sharp
54	1-Hydroxy-2-propanone	17.94	0.52	1290	a	Technological marker	Buttery
55	2-Methyl-1-pentanol	17.99	0.76	1292	a	-	-
56	2-Ethyl-(*E*)-2-hexenal	18.04	1.59	1293	a	-	-
57	3-Hepten-2-one	18.09	1.56	1294	b	-	-
58	2-Heptanol	18.94	0.90	1314	a	Key-aroma marker	Citrusy
59	2,3-Octanedione	19.14	1.28	1319	b	Technological marker	-
60	2,5-Dimethylpyrazine	19.24	0.88	1321	a	Technological marker	Earthy
61	2,6-Dimethylpyrazine	19.34	0.90	1324	b	Technological marker	Earthy
62	Ethylpyrazine	19.54	0.89	1328	b	Technological marker	Earthy
63	6-Methyl-5-hepten-2-one	19.64	1.28	1331	a	-	Pungent, green
64	2,3-Dimethylpyrazine	20.09	0.93	1341	b	Technological marker	Earthy
65	1-Hexanol	20.29	0.79	1346	a	-	
66	4-Hydroxy-4-methyl-2-pentanone	20.54	0.83	1352	b	-	Fruity, banana, soft
67	Dimethyl trisulfide	21.24	1.03	1368	a	Key-aroma marker	sulfury, cabbage
68	2-Ethyl-6-methylpyrazine	21.74	1.07	1380	a	Technological marker	Earthy
69	2-Nonanone	21.94	1.72	1385	a	-	-
70	2-Ethyl-5-methylpyrazine	22.04	1.07	1387	a	Technological marker	Earthy
71	Nonanal	22.14	1.72	1389	a	-	Fatty, waxy, pungent
72	2,3,5-Trimethylpyrazine	22.64	1.03	1401	a	Key-aroma marker	Earthy
73	**α-Thujone**	**23.19**	**1.79**	**1414**	**a**	**ISTD**	**-**
74	2-Octanol	23.49	1.07	1422	a	-	Mushroom, fatty, creamy
75	Ethyl octanoate	23.89	2.00	1431	a	-	-
76	1-Octen-3-ol	23.94	0.64	1433	a	-	Mould, earthy
77	Acetic acid	23.99	0.57	1434	a	Key-aroma marker	Sour, vinegary
78	2-Ethyl-3,6-dimethylpyrazine	24.29	1.20	1441	a	Technological marker	Earthy
79	Furfural	24.88	0.69	1455	a	Technological marker	Sweet, bread-like
80	1-Acetyloxy-2-propanone	24.89	1.31	1455	b	Technological marker	-
81	2-Ethyl-3,5-dimethylpyrazine	24.94	1.17	1457	a	Key-aroma marker	Earthy
82	*Trans*-linalool oxide	25.29	1.21	1465	a	-	Sweet floral, citrus, fruity
83	2,6-Dimethyl-4-heptanol	25.44	1.40	1469	b	-	-
84	Tetramethylpyrazine	25.54	1.14	1471	a	Technological marker	Earthy
85	2,3-Butanediol diacetate	25.99	1.14	1482	b	-	-
86	2-Ethyl-1-hexanol	26.04	0.97	1483	a	-	-
87	Decanal	26.54	1.86	1495	a	-	Penetrating, sweet, waxy
88	2-Acetylfuran	26.64	0.72	1498	b	-	-
89	3,5-Diethyl-2-methylpyrazine	27.09	0.55	1509	b	Key-aroma marker	Earthy
90	Benzaldehyde	27.34	0.79	1515	a	-	Almond, burnt sugar
91	2,3-Butanediol diacetate	27.49	1.10	1519	b	-	-
92	Furfuryl acetate	27.79	0.79	1526	a	-	-

93	2-Nonanol	27.83	0.94	1527	a	-	-
94	2,3-Butanediol	27.99	0.55	1531	a	-	-
95	Propanoic acid	27.99	0.40	1531	a	-	Fruity, pungent
96	Linalool	28.44	1.00	1542	a	-	Citrus
97	1-Octanol	28.84	0.93	1552	a	-	Moss, nut, mushroom
98	2-Methylpropanoic acid	29.09	0.48	1558	b	Key-aroma marker	Rancid
99	2,3-Butanediol	29.44	0.52	1567	a	-	-
100	Dihydro-2(3H)-furanone	31.34	0.76	1614	b	-	-
101	Butanoic acid	31.54	0.48	1619	a	Key-aroma marker	Sweaty, rancid
102	Phenylacetaldehyde	32.04	0.83	1632	a	Key-aroma marker	Honey-like
103	Ethyl decanoate	32.14	2.31	1635	a	-	Fruity
104	Acetophenone	32.34	0.86	1640	a	-	-
105	2-Furanmethanol	32.59	0.52	1646	a	Technological marker	Burned
106	Ethyl benzoate	33.04	1.49	1658	a	-	-
107	3-Methylbutanoic acid	33.09	0.52	1659	b	Key-aroma marker	Rancid
108	Dodecanal	34.94	2.01	1707	a	-	Fatty, citrus-like
109	Pentanoic acid	35.94	0.46	1734	a	-	Sweaty
110	4-Ethylphenyl acetate	37.39	1.03	1773	b	-	-
111	4-Methylpentanoic acid	38.19	0.52	1795	b	-	-
112	1-Phenylethanol	38.24	0.63	1796	b	-	-
113	2-Phenylethyl acetate	38.39	0.62	1800	b	Key-aroma marker	Flowery
114	Hexanoic acid	39.69	0.52	1837	a	-	Rancid
115	Ethyl dodecanoate	39.74	2.48	1838	a	-	-
116	Guaiacol	39.84	0.75	1841	a	-	Spicy
117	2-Methyl propyl benzoate	40.19	1.21	1851	b	-	-
118	Benzyl alcohol	40.44	0.59	1858	a	-	Sweet, fruity
119	Phenylethylalcohol	41.64	0.66	1892	a	Key-aroma marker	Honey-like
120	(*E*)-2-Phenyl-2-butenal	42.44	0.97	1915	b	Technological marker	-
121	Acetyl pyrrole	43.64	0.59	1950	a	-	Popcorn-like
122	Phenol	44.74	0.52	1982	a	-	-
123	1H-Pyrrole-2-carboxaldehyde	45.34	0.52	2000	a	-	-
124	4-Hydroxy-2,5-dimethyl-3(2H)-furanone	45.64	0.59	2009	a	Technological marker	Caramel-like
125	Octanoic acid	46.94	0.55	2049	a	-	Sweaty
126	5-Methyl-2-phenyl-2-(*Z*)-hexenal	47.24	1.21	2058	b	-	-
127	Nonanoic acid	50.34	0.57	2156	a	-	Sweaty, waxy
128	2,3-Dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one	52.74	1.41	2231	b	-	-
129	Decanoic acid	53.49	0.66	2259	a	-	Soap-like, fatty
130	2-Phenylacetic acid	59.44	0.79	2549	b	-	Honey-like

[a]: targets identified by means of authentic standards

[b]: targets tentatively identified on MS fragmentation patterns and Linear Retention Indices available in commercial libraries

Associated content

Supplementary Table 1 (ST1): Samples characteristics

Supplementary Table 2 (ST2): Validation data. Repeatability and intermediate precision on retention times and 2D peaks quantitative descriptors (Normalized 2D Volumes).

Supplementary Figure 1 (SF1):2D patterns of volatiles from cocoa samples harvested in the Chontalpa region (Tabasco - Mexico) from raw (SF1A), to roasted (SF1B) than steamed (SF1C) and at the end to nibs (SF1D). Light blue circles indicate the positions of targeted peaks, pink the untargeted and yellow circles the ISTDs peaks (α- and β-thujones).

Figure 1

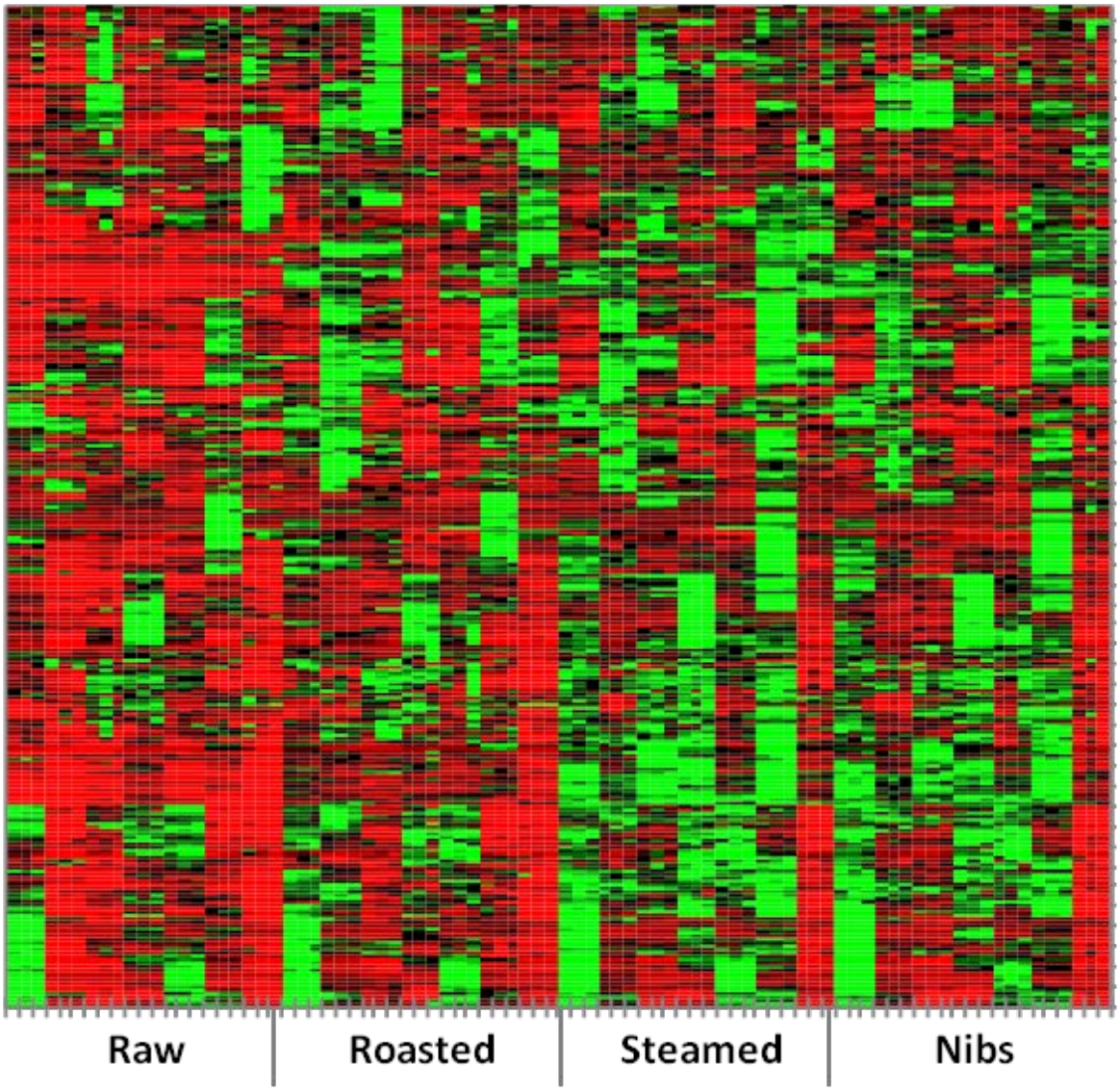
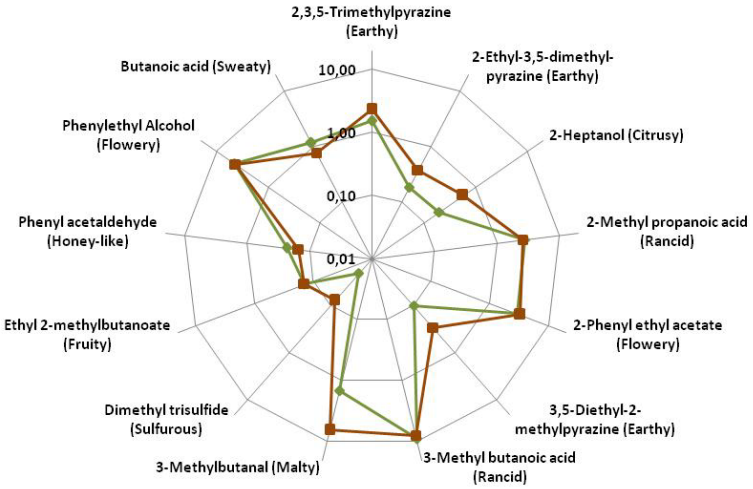
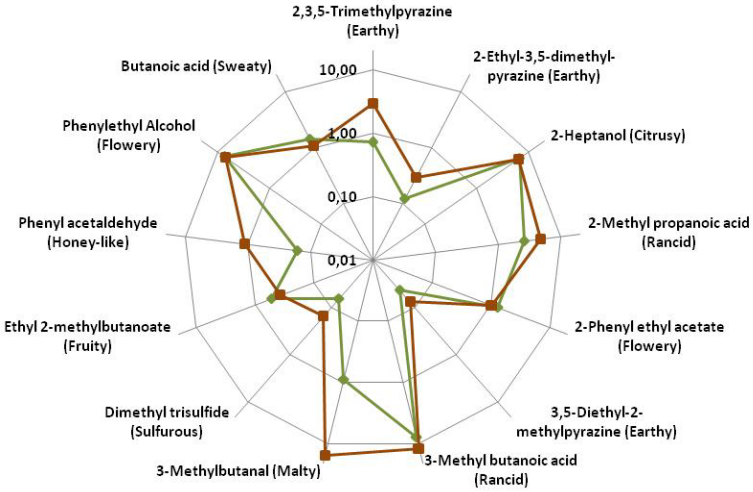


Figure 2

A
Chontalpa (Trinitario)
Raw *vs* Roasted



B
Venezuela (Trinitario)
Raw *vs* Roasted



C
Sao Tomè (Forastero)
Raw *vs* Roasted

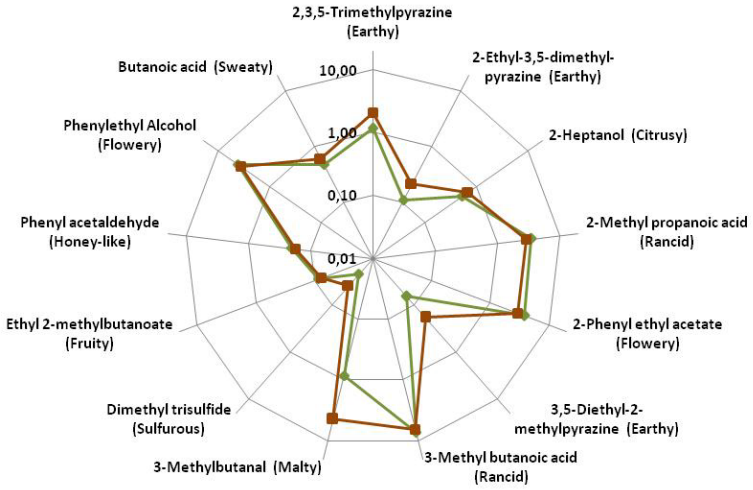


Figure 3

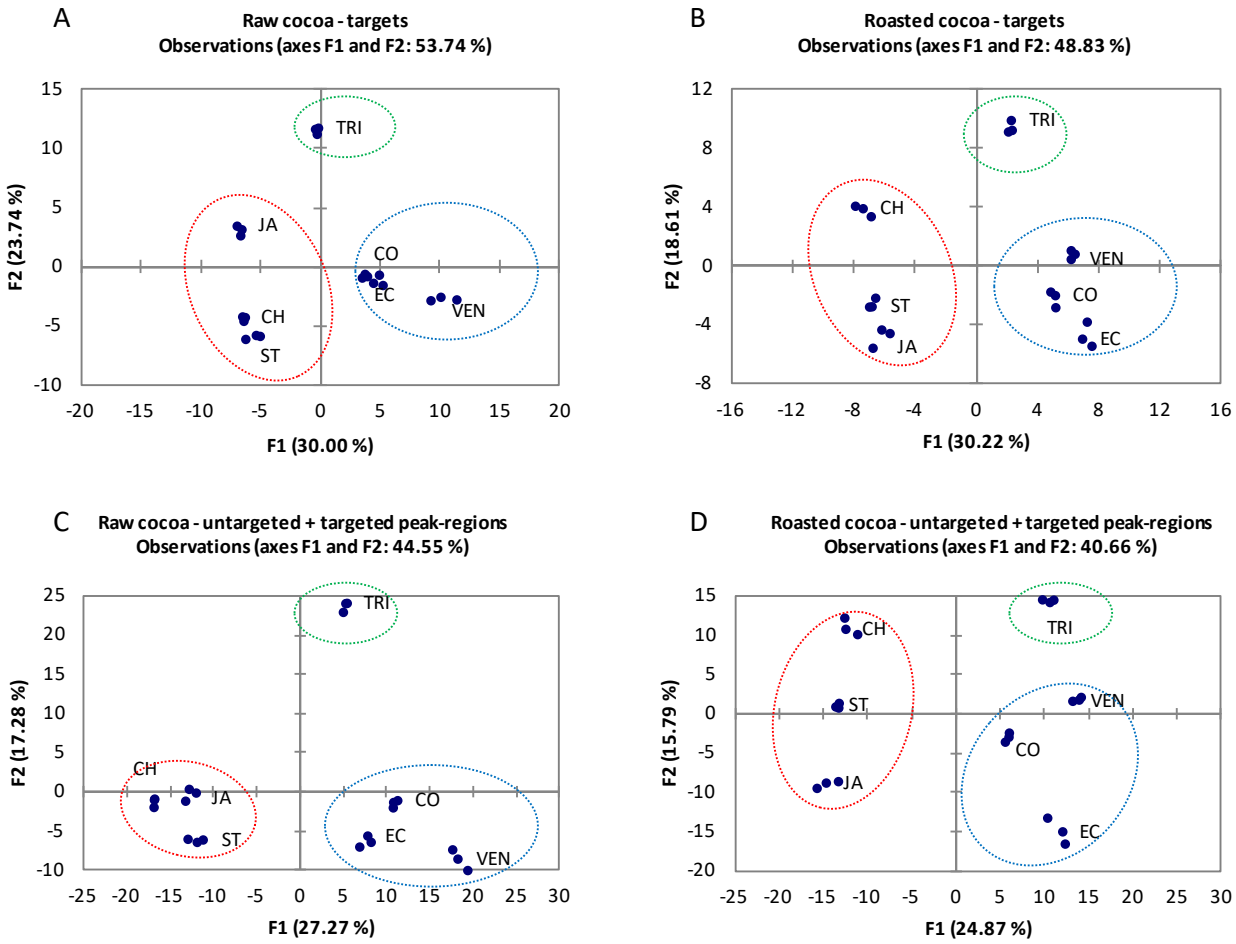

A
Raw cocoa - targets
Observations (axes F1 and F2: 53.74 %)
F2 (23.74 %)
F1 (30.00 %)
TRI
JA
CO
EC
CH
ST
VEN
B
Roasted cocoa - targets
Observations (axes F1 and F2: 48.83 %)
F2 (18.61 %)
F1 (30.22 %)
TRI
CH
VEN
ST
CO
JA
EC
C
Raw cocoa - untargeted + targeted peak-regions
Observations (axes F1 and F2: 44.55 %)
F2 (17.28 %)
F1 (27.27 %)
TRI
CH
JA
CO
ST
EC
VEN
D
Roasted cocoa - untargeted + targeted peak-regions
Observations (axes F1 and F2: 40.66 %)
F2 (15.79 %)
F1 (24.87 %)
CH
TRI
ST
VEN
CO
JA
EC

Figure 4

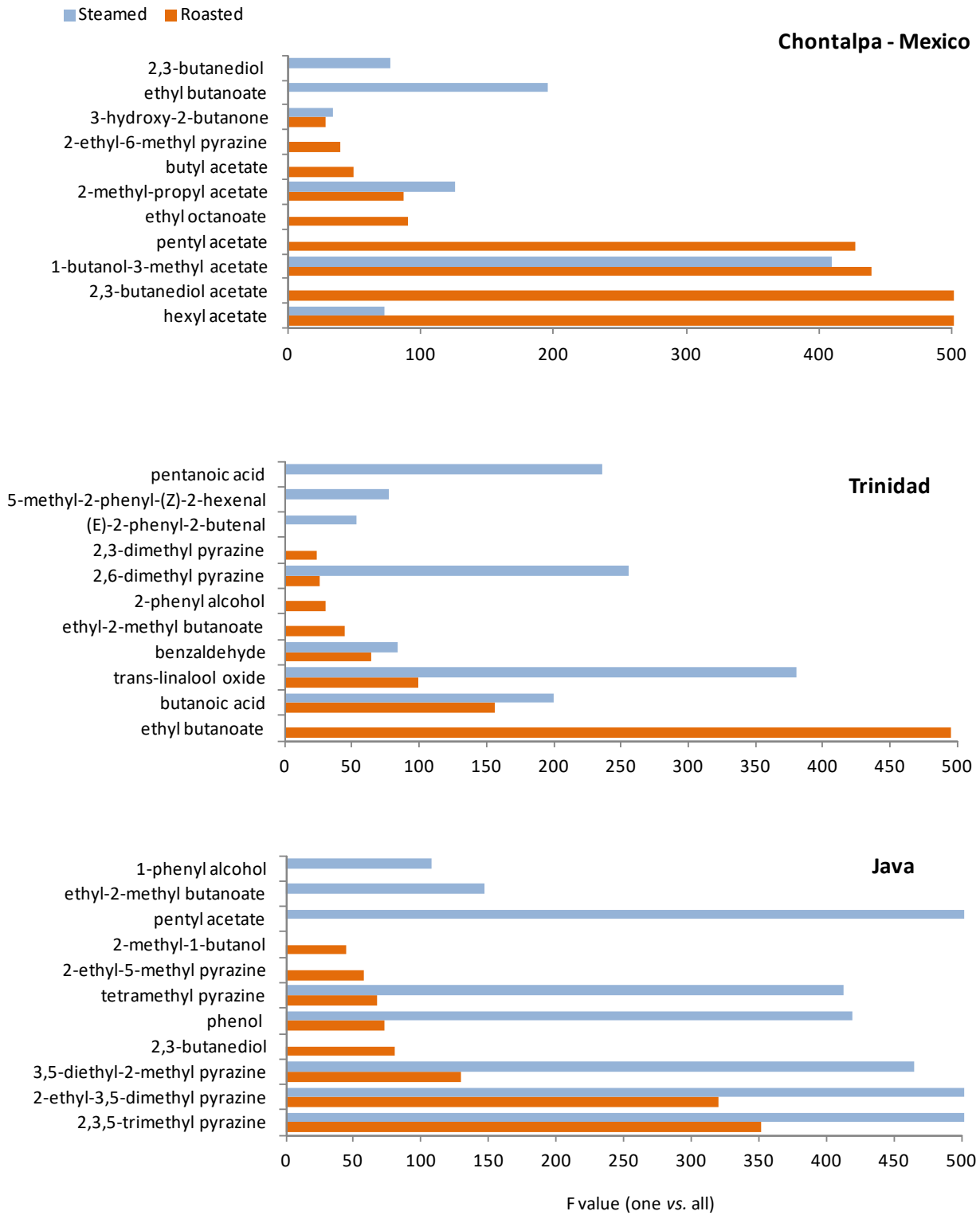
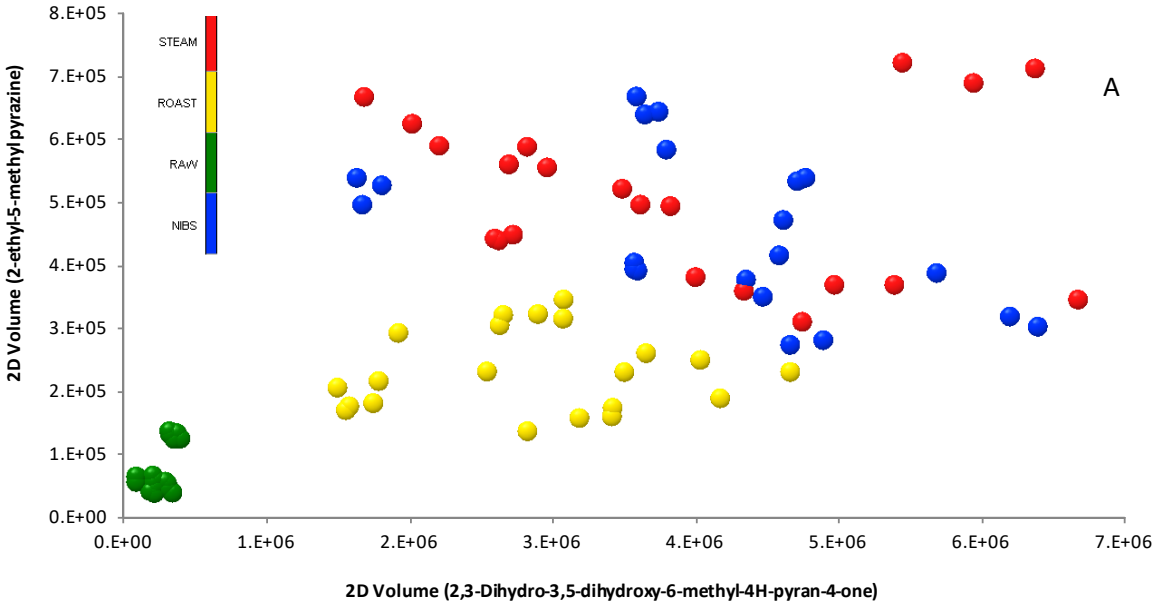
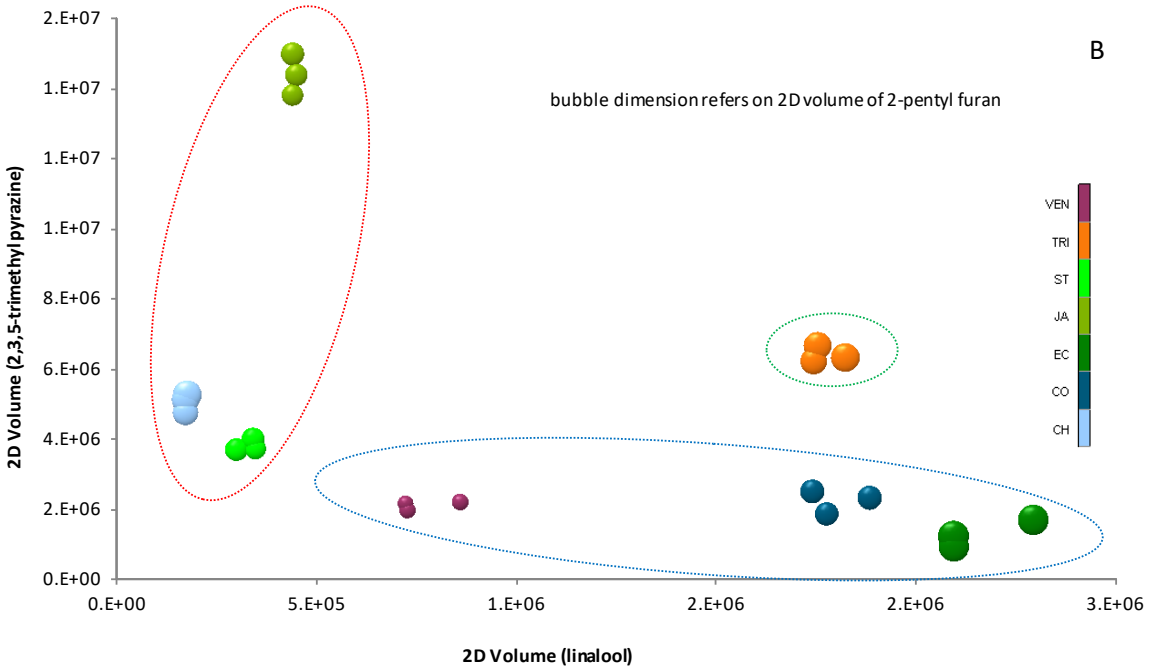

Steamed
Roasted
Chontalpa - Mexico
2,3-butanediol
ethyl butanoate
3-hydroxy-2-butanone
2-ethyl-6-methyl pyrazine
butyl acetate
2-methyl-propyl acetate
ethyl octanoate
pentyl acetate
1-butanol-3-methyl acetate
2,3-butanediol acetate
hexyl acetate
0
100
200
300
400
500
Trinidad
pentanoic acid
5-methyl-2-phenyl-(Z)-2-hexenal
(E)-2-phenyl-2-butenal
2,3-dimethyl pyrazine
2,6-dimethyl pyrazine
2-phenyl alcohol
ethyl-2-methyl butanoate
benzaldehyde
trans-linalool oxide
butanoic acid
ethyl butanoate
0
50
100
150
200
250
300
350
400
450
500
Java
1-phenyl alcohol
ethyl-2-methyl butanoate
pentyl acetate
2-methyl-1-butanol
2-ethyl-5-methyl pyrazine
tetramethyl pyrazine
phenol
2,3-butanediol
3,5-diethyl-2-methyl pyrazine
2-ethyl-3,5-dimethyl pyrazine
2,3,5-trimethyl pyrazine
0
50
100
150
200
250
300
350
400
450
500
F value (one *vs.* all)

Figure 5


A
STEAM
ROAST
RAW
NIBS
2D Volume (2-ethyl-5-methyl pyrazine)
2D Volume (2,3-Dihydro-3,5-dihydroxy-6-methyl-4H-pyran-4-one)
8.E+05
7.E+05
6.E+05
5.E+05
4.E+05
3.E+05
2.E+05
1.E+05
0.E+00
0.E+00
1.E+06
2.E+06
3.E+06
4.E+06
5.E+06
6.E+06
7.E+06


B
bubble dimension refers on 2D volume of 2-pentyl furan
VEN
TRI
ST
JA
EC
CO
CH
2D Volume (2,3,5-trimethyl pyrazine)
2D Volume (linalool)
2.E+07
1.E+07
1.E+07
1.E+07
8.E+06
6.E+06
4.E+06
2.E+06
0.E+00
0.E+00
5.E+05
1.E+06
2.E+06
2.E+06
3.E+06

TOC graphics