IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

Targeted and non-targeted approaches for complex natural sample profiling by GCxGC-qMS

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/60828 since 2016-12-01T14:20:11Z

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

12 October 2024

UNIVERSITÀ DEGLI STUDI DI TORINO

This is the accepted version of the following article:

[Chiara Cordero1, Erica Liberto, Carlo Bicchi, Patrizia Rubiolo, Stephen E. Reichenbach2, Xue Tian and Qingping Tao.

Targeted and Non-Targeted Approach es for Complex Natural Sample Profiling by GC×GC–qMS. Journal of Separation Science Volume 48, Issue 4, pages 251-261, April 2010, DOI: doi: 10.1093/chromsci/48.4.251],

which has been published in final form at
[http://chromsci.oxfordjournals.org/content/48/4/251]

Targeted and non-targeted approaches for complex natural sample profiling by GCxGC-qMS.

Chiara Cordero[*1], Erica Liberto[1], Carlo Bicchi[1], Patrizia Rubiolo[1],
Stephen E. Reichenbach[*2], Xue Tian[2], Qingping Tao[3]

[1] Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy
[2] Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA
[3] GC Image, LLC, PO Box 57403, Lincoln NE 68505-7403, USA

[*] Address for correspondence:
Dr. Chiara Cordero - Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Via P. Giuria 9, I-10125 Torino, Italy – e-mail: chiara.cordero@unito.it ; phone: +39 011 670 7662; fax: +39 011 2367662
Prof. Stephen E. Reichenbach - Computer Science and Engineering Department, University of Nebraska – Lincoln, Lincoln NE 68588-0115, USA - e-mail: reich@unl.edu ; phone: 001.402.472.2401; fax: 001.402.472.7767

Abstract

The present study examined the ability of targeted and non-targeted methods to provide further, specific and complementary information on groups of samples on the basis of their component distribution on the GCxGC chromatographic plane. The volatile fraction of Arabica green and roasted coffee samples, differing for geographical origins and roasting treatments, and volatile secondary metabolites from juniper needles, sampled by Headspace-Solid Phase Micro Extraction (HS-SPME), were analyzed by GCxGC-q-MS and sample profiles interpreted through different methods. In the target analysis profiling, samples, submitted to different roasting cycles and/or differing for their origin and post-harvest treatment, are characterized on the basis of known constituents: botanical, technological and aroma markers. This approach provides highly reliable results on quali-quantitative compositional differences, because of the authentic standard confirmation, extending and improving the specificity of the comparative procedure to trace and minor components. On the other hand, non-targeted data-processing methods: direct image comparison and template-based fingerprinting include in the sample comparisons and correlations the whole volatiles offering an increased discrimination potential identifying compounds that are comparatively significant but are not known targets. Results demonstrates the ability of GCxGC to further explore the complexity of samples and emphasizes the advantages of a comprehensive and multidisciplinary approach to interpret the increased level of information provided by GCxGC separation.

Key-words: GCxGC, target analysis, template-based fingerprinting, direct image comparison, green and roasted coffee, juniper volatile fraction

Introduction

The volatile fractions of samples of vegetable origin (plants derivatives and food samples) have highly variable abundance of their components, which mainly consist of secondary metabolites deriving from specific biosynthetic pathways (e.g., mono- and sesquiterpenoids) and/or groups of chemically-correlated components, such as alcohols, carbonyl derivatives, acids, esters and heterocycles, mainly produced by known and unknown reactions induced by technological treatments. These compounds often show similar chromatographic retention behavior, due to their volatility and polarity, and are characterized by MS fragmentation patterns with several common isobaric ions (fragments) making their 1D-GC characterization and quantitation difficult.

Comprehensive two-dimensional gas chromatography (GCxGC) is a useful and powerful tool for in-depth analysis of such complex mixtures because of its high "practical" peak capacity and sensitivity, enabling trace and minor component investigations on sample comparisons and possibility to obtain specific and rationalized separation patterns for chemically correlated groups of substances characteristic of a sample. Under properly optimized conditions, the gain in analyte detectability, experimentally approximated to the LODs, of GCxGC when compared to 1D-GC, is in general from three to fivefold [1-3]. In addition, the number of separated peaks is larger with gain factors up to 10 resulting in a higher confidence level for analyte identification.

A direct consequence of this gain in separation power is that chromatograms, data files, and peak lists are highly complex. A GCxGC separation produces a large and complex dataset for each sample, consisting of bi-dimensional retention data, detector response and, for multi channel-detectors such as MS, the MS spectra, requiring suitable data mining methods to extract useful and consistent information from the dataset. These methods are a bridge between chromatographic data and knowledge of sample compositional characteristics.

Two general approaches are available to link raw data (i.e. separation data) with the chemical quali-quantitative composition of samples (and from there to correlate samples on the basis of their characteristics or technological treatment(s) and derive conclusions): targeted and non-targeted methods [4].

Targeted methods are based on the assumption that the overall chemical composition of the sample and/or the distribution of several target analytes (secondary metabolites, known key-aroma markers, technological markers, safety regulated components or geographical tracers) to establish sample comparisons and characterization are already known. On the other hand, non-targeted methods consider the entire multidimensional sample profile to: *(a)* provide a comprehensive survey of qualitative and quantitative differences in the chemical composition between samples as the basis of potential knowledge of important compositional characteristics and *(b)* support classification of

samples on the basis of degree of similarity of their 2D fingerprints. With non-targeted fingerprint analysis, chemometric techniques, such as multivariate-analysis (MVA), offer promising strategies to distill essential information from GCxGC datasets [4 and references cited therein].

Undoubtedly, as was discussed in a previous study [5], there are advantages in applying comprehensive and multidisciplinary approaches to interpret the increased level of information provided by GCxGC separation, in its full complexity.

This study evaluates advantages and limits of targeted and non-targeted approaches based on the bi-dimensionality of the separation (^{1}D and ^{2}D retention times, detector response, and MS spectrum) and specific to GCxGC in chemical speciation, differentiation, and correlation of complex matrices of natural origin. In particular, target-analysis characterization, direct image comparison, and template-based fingerprinting are evaluated. Each method is tested and applied to study the volatile fraction of green and roasted coffee samples and dried juniper needles, to evaluate its ability to differentiate samples on the basis of characteristics geographical origin, harvesting, technological and thermal treatments in the case of coffee. Coffee and Juniper samples, as representative examples for two different fields of application (i.e. processed food analysis and secondary metabolite profiling –metabolomic), were here chosen because of the peculiar composition of their volatile fraction and the challenging problem they offer in sample profiling. A set of Arabica coffee samples (*Coffea arabica*) of three different geographical origins: Colombia, Guatemala and Brazil, differently processed (i.e. washed and natural) and submitted to different roasting profiles and the volatile fraction of juniper needles (*Juniperus communis*), collected at different altitudes (sea level, 600, 900, 1100 and 1400 m) were investigated. The volatile fraction of these matrices was sampled by headspace-solid phase microextraction (HS-SPME), a technique that have shown to be effective for routinely characterizing the volatile fraction of vegetable matrices [6,7 and references cited therein].

EXPERIMENTAL

Reference Compounds and Solvents.

Pure standard samples of *n*-alkanes (from *n*-C9 to *n*-C25) and pure reference compounds adopted for the identity confirmation of target compounds were supplied by Sigma-Aldrich (Milan, Italy) except 2-Methyl-3-Propylpyrazine supplied by VWR International (Milan, Italy). Standard stock solutions at 1000 μg/mL were prepared in cyclohexane, stored at – 18°C and used to prepare standard working solutions whose concentration ranging from 50 to 5 μg/mL, likewise stored at –18°C.

Solvents (cyclohexane, *n*-hexane, dichloromethane) were all HPLC-grade from Riedel-de Haen

(Seelze, Germany).

Coffee.

Green beens of *Coffea arabica* (2008) from three different geographical origins: Colombia, Guatemala and Brazil were supplied by Lavazza SpA (Turin, Italy) and are listed in **Table 1** together with the post-harvest treatment and the roasting time/temperature profiles. Roasting was done in a Probat laboratory roasting device (Emmerich, Germany) and after process the roasted beans were hermetically sealed under vacuum in non-permeable packages (polypropylene/aluminum/polyethylene - PP/Al/PET) and stored at -20 °C, until required for chemical analysis.

Juniper.

Needles of *Juniperus communis* L from Norway were collected at different altitudes (5 replicates indicated with arabic numbering) sample A at 1400 m, B 1100 m, C 900 m and D at sea level in 2008 and dried and stored in paper bags until analyzed.

Headspace Solid Phase Microextraction (HS-SPME).

SPME device and fibers were from Supelco (Bellefonte, PA, USA). A Divinylbenzene/Carboxen/Polydimethylsiloxane (DVB/CAR/PDMS) df 50/30 µm, 2 cm length fiber was chosen and conditioned before use as recommended by the manufacturer. Material was left to reach ambient temperature before sampling. Roasted coffee was ground (0.5 g) and immediately sealed in a 12.5 mL vial and equilibrated for 10 min at 50°C.

Dried juniper needles (0.06 g) were ground and hermetically sealed in a 12.5 mL vial and equilibrated for 5 min at 50°C.

The SPME device was manually inserted into the sealed vial containing the sample prepared as described above, and the fiber was exposed to the matrix headspace, kept at 50°C, for 40 min for coffee samples and for 10 min for juniper needles during HS equilibration . The vial was vibrated for 10 s every 5 min with an electric engraver (Vibro-Graver V74, Burgess Vibrocrafters Inc, Brayslake, IL) to speed up the analyte equilibration process between headspace and fiber coating. Only that part of the vial in which the solid sample was present was heated, in order to keep the SPME fiber as cool as possible, to improve the vapor phase/fiber coating distribution coefficient. After sampling, the SPME device was immediately introduced into the GC injector for thermal desorption for 10 min at 250°C. Each experiment was carried out in triplicate and 2D-peak Area/Volume variability was ever below 15% of RSD.

GCxGC Instrumental set-up

GCxGC analyses were carried out on an Agilent 6890 GC unit coupled with an Agilent 5975 MS detector operating in EI mode at 70 eV (Agilent, Little Falls, DE, USA). The transfer line was set at 270°C. A Standard Tune option was used and the scan range was set at m/z 35-240 with the *fast scanning* option applied (10000 amu/s) to obtain a number of data points for each chromatographic peak suitable to make its identification and quantitation reliable.

The system was provided with a two-stage thermal modulator (KT 2004 loop modulator from Zoex Corporation, Houston, TX, USA) cooled with liquid nitrogen and with the hot jet pulse time set at 250 ms with a modulation time of 4 s adopted for all experiments. Fused silica capillary loop dimensions were 1.0 m length, 100 μm ID.

Column set adopted was configured as follows: ^{1}D SE52 column (95% polydimethylsiloxane, 5% phenyl) (30 m x 0.25 mm ID, 0.25 μm df) coupled with a ^{2}D OV1701 column (86% polydimethylsiloxane, 7% phenyl, 7% cyanopropyl) (1 m x 0.1 mm ID, 0.10 μm df); columns were from MEGA (Legnano (Milan)-Italy).

1 μL of the *n*-alkanes sample solution was automatically injected into the GC instrument with an Agilent ALS 7683B injection system under the following conditions: injector: split/splitless, mode: split, split ratio: 1/100, injector temperature: 280°C. The HS-SPME sampled analytes were recovered through thermal desorption of the fiber for 10 min into the GC injector under the following conditions: injector: split/splitless in split mode, split ratio: 1/50, injector temperature: 250°C. Carrier gas: helium at constant flow of 1.0 mL/min (initial head pressure 280 KPa). Temperature program for coffee analyses was: from 50°C (1 min) to 260°C (5 min) at 2°C/min. For the analysis of juniper needle samples it was: from 45°C (1 min) to 240°C (5 min) at 3°C/min; modulation period: 4 s

Data was acquired by Agilent – MSD Chem Station ver D.02.00.275 (Agilent Technologies, Little Falls, DE, USA) and processed using GC Image software, version 1.9b4 and pre-release version 2.0 (GC Image, LLC Lincoln NE, USA).

RESULT AND DISCUSSION

The study examined the ability of GCxGC to provide further and specific information on groups of samples on the basis of component distribution on the chromatographic plane. The first part of the study involved a target analysis profiling where samples, submitted to different technological treatments and/or differing for geographical origin, were characterized on the basis of known constituents. The second part examined specific non-targeted data-processing methods for GCxGC: direct image comparison and template-based fingerprinting to evaluate differences and similarities.

Target analysis: technological and Key-Aroma marker profiling of Arabica coffee samples by HS-SPME/GCxGC-qMS.

Coffee roasting induces several chemical reactions, whose control is fundamental to optimize flavor, color and texture. These reactions involve specific precursors following known and unknown pathways to originate a complex mixture of more than 20 different groups of substances, most of them contributing to the total flavor: furans, pyrazines, ketones, alcohols, aldehydes, esters, pyrroles, thiophenes, sulfur compounds, aromatic compounds, phenols, pyridines, thiazoles, oxazoles, lactones, alkanes, alkenes, and acids. Sample characterization was first run by selecting a suitable number of markers (targets) chosen in function of their significance for the purpose of describing botanical, technological and sensory characteristics of the samples under study [5,8-12]. **Table 2** reports the list of target analytes chosen for Arabica coffee samples, their ID numbers, chemical name, group classification, Retention Indices and ^{1}D and ^{2}D retention times. Each component was located in the 2D plot by its ^{1}D-^{2}D retention times, identified by both spectral library matching and authentic standard confirmation. The normalized peak volume of each component, i.e. absolute volume normalized versus the ISTD, was used to compare samples differing in geographical origin, post-harvesting treatment and roasting profiles. **Figures 1** and **2** report some results: histograms give pyrazines and aroma marker distribution. The separation power of GCxGC is here evident in particular for pyrazines, an important group of technological markers, comparison was in fact based on an large number of congeners that are difficult to detect, without a sample pre-concentration, and to separate with a one-dimensional GC system. For instance, **Figure 1** reports pyrazine 2D pattern of Arabica samples from Colombia, Guatemala and Brazil submitted to a standard roasting while **Table 3** reports in detail normalized peak volumes of the nineteen target analytes over the entire sample set. As expected, samples that have the same botanical origin (*Coffea arabica*), and thought to have a similar pyrazine precursor chemical distribution, showed similar quali-quantitative profiles. Some exceptions were evidenced in the Colombia standard

roasted coffee where, as a general consideration, the total peak volume corresponding to the nineteen selected pyrazines was lower, i.e 540 if compared to 740, the average value for Brazil and Guatemala, thus indicating a lower pyrazines concentration in this sample. In particular, 2-Ethyl-3-Methylpyrazine, 5-Ethyl-2,3-Dimethylpyrazine, 2,5-Diethylpyrazine, 2-Ethyl-6-Vinylpyrazine, 2-Acetyl-6-Methylpyrazine and 2,3-Diethyl-5-Methylpyrazine varied greatly with decrements ranging from 41% for 2,3-Diethyl-5-Methylpyrazine to 84% 2-Ethyl-6-Vinylpyrazine when compared to the Brazil and Guatemala samples at the corresponding degree of roasting. On the other hand, some other alkylpyrazines such as 2-Propylpyrazine, 2-Acetyl-6-Methylpyrazine, and in particular the two most odour active, i.e. 2-Ethyl-3,5-Dimethylpyrazine and 2,3-Diethyl-5-Methylpyrazine belonging to the *earthy* group of aroma marker, were well separated from the congeners and detect, their distribution is in fact very informative from both a technological and sensory point of view. Since the abundance of some markers is related to the extent of thermal treatment [8], and since samples submitted to a mild roasting treatment showed lower abundances (in terms of normalized peak volumes) than those of standard-roasted samples, the approach based on evaluating normalized peak volumes and/or areas of the technological markers in different structural groups (acids, aldehydes, ketones, furans, pyridines etc.) is also very illustrative as it was deeply discussed in a previous study [5]. The section dealing with a direct fingerprint comparison approach will discuss in greater detail advantages and some limits of quali-quantitative profiling performed by non-targeted techniques based on sample components distribution over the 2D plane.

A further interesting target characterization of the coffee sample set was done on a selection of 13, over the 28, key-aroma compounds indicated by Czerny *et al* [9-11]. These volatiles identified by aroma extract dilution analysis (AEDA) [10,12] and gas chromatography-olfactometry of headspace samples (HS-GC-O), showed a high odour potency and mainly contribute to the aroma of Arabica roasted coffee. However, their concentration in roasted samples varies greatly ranging from traces (ng/g) to several percent (g/100g), and, for a complete aroma profiling, sample pre-concentration is mandatory. Thanks to its high sensitivity, GCxGC enabled us to identify and semi-quantify, evaluating their relative abundance in the sample set, thirteen key-aroma compounds whose concentration in the original sample, expressed as mg/Kg and referred to the roasted material, ranges from: 130 mg/Kg of acetaldehyde, the most abundant, to 55 mg/Kg for 2-Methoxy-4-vinylphenol, 49.4 and 36.2 mg/Kg for 2,3-butanedione and 2,3-pentanedione respectively, 3.2 and 1.6 mg/Kg for 2-Methoxyphenol and 2-Methoxy-4-ethylphenol to 0.326 and 0.017 mg/Kg for 2-Ethyl-3,5-Dimethylpyrazine and 2,3-Diethyl-5-Methylpyrazine respectively [9]. Key-aroma compounds quali-quantitative distribution is visualized in the histogram of **Figure 2** for the three standard roasted Arabica samples under study. This profiling confirmed the relative/overall

homogeneity of the target distribution over the sample set, coherent with the botanical characteristics [8] and, in this case, with the extent of roasting. Higher distributional differences were detectable for highly volatile compounds such as 3-Methylbutanal, 2-Methylbutanal and 2,3-Pentanedione, among the others, while 2-Furanmethanethiol, responsible of the characteristic sulphurous/roasty note, was very low in the Guatemala sample as it was for the 2-Ethyl-3,5-Dimethylpyrazine.

Target analysis: sample profiling of Juniper volatile secondary metabolites by HS-SPME/GCxGC-qMS.

The genus *Juniperus* (*Cupressaceae*) consists of 68 species and 36 varieties mainly growing in the northern hemisphere. Common juniper, *Juniperus communis* L. is an aromatic and evergreen shrub and its berries are well known for their bioactivity [13,14] and as an ingredient in the production of juniper-based spirits, such as gin [15]. The juniper essential oils and/or extracts from needles, berries or wood have been the object of several studies [16].The results here reported are part of a systematic study on the variation of the composition of the volatile fraction of juniper (*Juniperus communis* L.) needles and berries, differing in their origin, age of the plant, and berries ripeness [15,17].

This application concerns the discrimination of the juniper needles collected in Norway at different altitudes as a further example of how informative can be GCxGC-qMS, adopted as target and non-target profiling method, in describing the distribution of specific secondary metabolites, mainly mono- and sesquiterpenoids, of a complex volatile fraction of vegetable origin.

The investigated Juniper needle samples are reported in the experimental section and the list of target analytes and their distribution in each sample in **Table 4**. Target analytes were semi-quantified as percent on total volatile areas and these values used to study the secondary metabolites distribution within the sample set. GCxGC-qMS data showed different profiles referable to main groups characterized by peculiar distribution of some target analytes and already described in literature [18,19] the α-pinene, the α-pinene/sabinene and sabinene/α-pinene types with some exceptions. Since a detailed discussion on the chemical composition of the volatile fraction of the complete set of samples is out of the scope of this paper, Juniper A_*1*, Juniper B_*4*, Juniper B_*5*, Juniper C_*4*, Juniper D_*1* selected on the basis of their peculiar composition, as representative samples, were submitted to further investigations. Their peculiarities were due to specific and unusual distributions of some markers. For instance: Juniper B_*4* showed the highest amount of α-pinene (59.3%), and a 0.8% of sabinene and a 0.1% of terpinen-4-ol, another informative marker. On the other hand, Juniper C_*4* and Juniper D_*1* were characterized by an intermediate amount of

α-pinene (36.7% and 33.5% respectively), and relatively low abundances of sabinene (0.5%), terpinen-4-ol (0.1%). In Juniper B_*5* α-pinene accounted for 12.8%, sabinene for 44.6% and terpinen-4-ol for 0.3%, and Juniper A_*1* contained 21.9%, of α-pinene 24.0% of sabinene and 0.4% of terpinen-4-ol. In conclusion, by extending the considerations on the distribution of all target analytes considered for juniper profiling, and listed in **Table 4**, results indicated the presence of two main groups of samples, distinguished by the % contents of α-pinene and sabinene suggesting their classification had to be studied more in depth independently of the sites where they were collected. On the other hand, the harvesting site seemed to condition the samples variability , although to a different extent. The monoterpenoid and sesquiterpenoid composition of the Juniper D_*n* sample series was very different from the others.

The high variability and complexity of the results obtained from the volatile fraction profiling of juniper samples suggested further approaches to simplify data interpretation and/or to correlate compositional variables. Besides the conventional approach based on 1D-GC as analysis combined with PCA and CA (Principal Component Analysis and Correlation Analysis), specific GCxGC data analysis methods can be applied to compare sample component distribution over the 2D-plane, abstracting or not from chemical sample speciation. On this basis, a non-targeted approach, template-based fingerprinting, was evaluated to differentiate juniper samples leading to useful, informative and consistent results.

Non-targeted analysis on Arabica coffee samples.

This section first describes a new approach for non-targeted comparative analysis of two-dimensional chromatographic data. The approach uses templates to generate chromatographic fingerprints and then builds lists of potentially significant minutiae (i.e., small features) in the fingerprints. As detailed below, the fingerprints are created with a comprehensive mesh of contiguous, non-overlapping polygonal panels that divide the retention-times plane into regions that separate chromatographic features. Within each chromatogram, the panels in the mesh are quantified individually and treated as fingerprint minutiae. Various rules can be used to identify potentially significant minutiae for a given sample set.

Reliable peak matching

The fingerprinting process begins with the task of matching corresponding peaks within a set of sample chromatograms. For a complex chromatogram, such as in **Figure 3**, there may be hundreds or thousands of peaks. Peaks in two or more chromatograms correspond if they result from the same

analyte. For comparative analysis, the matching problem is to identify which peaks in a pair (or a set) of chromatograms correspond.

Matching corresponding peaks enables direct comparison of analyte peak responses across samples and allows alignment of chromatograms for comprehensive comparisons. In previous research [20], template matching has been used to identify target analytes in two-dimensional chromatograms. Here, template matching is used for non-targeted analysis by attempting to match as many peaks as possible between chromatograms. First, the peaks detected in a source chromatogram are used to create a template that describes the pattern of expected peaks with their individual retention times. Next, given another chromatogram for comparison, the matching algorithm determines the geometric transformation in the retention-times plane that best fits the expected peak pattern in the template to detected peaks in the chromatogram. A correspondence is established if there is a detected peak within the retention-times window around a transformed template peak. Multi-type templates have geometric features, such as polygons that can delineate sets of peaks, and notational features, such as text labels and chemical-structure graphics to convey information visually. Such features are geometrically transformed with the peak pattern to maintain their relative positions. Smart templates [21,22] attach rules that constrain potential matches based on additional peak attributes, such as mass spectral match factor or fractional response.

In this step, template matching is used to generate a *consensus template* of non-targeted peaks that can be matched across all pairs within a set of chromatograms. Non-targeted peak matching is a difficult problem that can involve thousands of peaks for complex samples. With complex samples, correspondences cannot be reliably established for all peaks across multiple chromatograms. The matching problem is particularly difficult for co-eluting , small-intensity, and long-tailed peaks and is exacerbated by even small chromatographic variations. However, template matching can be used to establish a subset of peaks with reliable correspondences within a sample set.

The steps for establishing reliable peak correspondences across a set of chromatograms are:

1. Correct the baseline of each chromatogram [23].
2. Detect the peaks in each chromatogram. For explanatory purposes, consider a set of chromatograms denoted A, B, and C, in which the detected peaks in chromatogram A are denoted A(i) where i is a unique peak ID.
3. Create a template from each chromatogram. For each detected peak in a chromatogram, a peak is added to the chromatogram's template with expected retention times from the detected peak. For example, the template for chromatogram A will have an expected peak denoted a(i) at the retention times of the detected peak A(i).

4. For each template peak, add a rule to constrain matching based on mass spectrometric similarity. For readability of this sequence of steps, the details of this step are given below.
5. Successively match each template to the detected peaks in each other chromatogram. For example, when the template from chromatogram A is matched to the detected peaks in chromatogram B, template peak a(*i*) either will match some detected peak B(*j*) or will not be matched to any detected peak in B.
6. Find the set of matched peaks for each pair of chromatograms that are consistent across all pairs of chromatograms. If a(*i*) matches B(*j*) and b(*j*) matches A(*i*), then peaks A(*i*) and B(*j*) correspond. In this research, reliable correspondence is defined as consistent correspondences across all pairs within the set. So, for consistency within the set in this example, there also must be matches for a(*i*) and C(*k*), c(*k*) and A(*i*), b(*j*) and C(*k*), and c(*k*) and B(*j*). Other less-restrictive consistency rules could be used, e.g., sequential consistency, consistency for a majority of pairs, etc.

The steps to associate a rule for each template peak (Step 4 above) are:

1. For each peak, consider all other peaks in the source chromatogram for the template to determine the largest match factor (using the NIST MS Search method [24]). In other words, determine the match factor with the peak that has the most similar mass spectrum. If the largest match factor is less than the lower threshold (a parameter set to 500 in this research), use the lower threshold for the match factor test. If the largest match factor is greater than the upper threshold (a parameter set to 650 or larger in this research), use the upper threshold for the match factor test. Otherwise, use the largest match factor with the other peaks for the match factor test written in the following step.
2. For each peak, write a rule to constrain matching of the peak using CLIC™ [22] as:

   Match(“<ms>”)>match_factor

   where “<ms>” is the mass spectrum of the peak, the Match function computes the match factor between the template spectrum and the detected peak spectrum, and match_factor is the match factor determined in the previous step. During matching, the template peak can be matched to a detected peak in the chromatogram being compared only if the match factor between the template peak mass spectrum (from the source chromatogram) and the detected peak mass spectrum is greater than the match_factor value in the rule.

The set of peaks that are reliable are included in a consensus template. For each peak in the consensus template, the expected retention times are the averages of the retention times of the corresponding peaks in the set of individual templates, the mass spectrum is the average of the mass spectra, and the match factor value for the rule is the average of the match factor values. In the

example, if A(i), B(j), and C(k) are reliable peaks, then the consensus template denoted t is t($i.j.k$) = Average(a(i), b(j), c(k)). Alternatively, the match_factor value in the consensus template can be computed as described in Step 4a, but using all peaks (other than the corresponding peaks in all chromatograms in the set.

Figure 3 illustrates a two-dimensional chromatogram of standard-roast Colombian coffee, with the locations of all 1652 detected peaks in the chromatogram shown with black and yellow circles. The subset of 891 reliable peaks determined for a set of three chromatograms including this chromatogram and chromatograms of standard-roast Brazilian and Guatemalan coffees, with 1658 and 1700 detected peaks respectively, are indicated with black circles. The peaks indicated with yellow circles are not reliable for this set of three chromatograms. This figure makes clear the high complexity of the chromatogram.

The set of reliable peaks can be used to compare some of the peak responses across the chromatograms, but for such complex samples it is unavoidable that many peaks are not reliably matched. For this reason, the primary purpose of the reliable peak set is to provide a basis for aligning (or registering) the chromatograms, as described in the next subsection, rather than for comprehensive comparison. Here, the detection algorithm is applied to the entire chromatogram, including some regions with chromatographic artifacts. The user could decide either to include such regions in this and subsequent steps (e.g., to assess chromatographic changes) or to exclude such regions from analysis.

Chromatogram alignment and comparative visualization

For visual comparisons of two-dimensional chromatograms, the corresponding peaks should be aligned as well as possible and normalized in terms of response [25]. To align two-dimensional chromatograms for pairwise comparison, one of the chromatograms is transformed in the retention-times plane to minimize the mean-square misalignment of the reliable peaks. Affine transformations (with scaling, translation, and shearing) have been shown to account for large variations in chromatographic conditions [26] and so are used here to find the best fit between the peak pattern in a template and detected peaks in a chromatogram. In these experiments the chromatograms were intensity scaled to have the same total response after baseline correction.

After chromatograms are aligned and normalized for total response, they can be visually compared. **Figure 4** shows two pseudocolor comparisons [25] of standard-roast Brazilian and standard-roast Colombian coffees. The first image shows the colorized fuzzy difference, which uses the Hue-Intensity-Saturation (HIS) colorspace to color each pixel in the retention-times plane. The method first computes the difference at each datapoint. The pixel hue is set to green if the difference is

positive and red if the difference is negative. The pixel intensity is set to the larger of the two values. The pixel saturation is set to the magnitude of the difference between the datapoints. Peaks are visible because large-valued datapoints yield bright pixels and small-valued datapoints yield dark pixels. If the difference is large, the color is saturated with red or green (depending on which datapoint is larger); if the difference is small, the color saturation is low, producing a graylevel from black to white depending on intensity. So, peaks with large differences appear red or green and peaks with small differences appear white or gray. The fuzzy difference is computed as the difference between a datapoint and a small region of datapoints in the other chromatogram. The second image shows the colorized fuzzy ratio for the same two chromatograms. The pseudocolorization is the same as for colorized fuzzy difference except that the difference is divided by the larger of the two values in computing the saturation. So, the colors are saturated with red or green only where the relative difference (rather than the absolute difference) is large.

Differences are highlighted visually by colorization. For example, as seen in **Figure 4**, the two green peaks at approximately (7.9 min, 1.9 s) indicate larger responses in the chromatogram of the standard-roast Brazilian coffee. Similarly, the red peaks at (26 min, 3 s) and (43.3 min, 2.6 s) indicate larger responses in the chromatogram of the standard-roast Colombian coffee. However, visual comparisons are not quantitative, so it is difficult to "see" if these are the most significant differences. Also, with so many apparent differences, it is difficult to comprehensively catalog the visual differences. The lack of comprehensiveness with peak matching and the lack of quantification with visualization motivates a fingerprinting method that is both comprehensive and quantitative, as described in the next section.

Fingerprinting with meshes

The goal of chromatographic fingerprinting is to catalog features of a chromatogram comprehensively, quantitatively, and in a manner that can be compared across samples. The approach presented here is to comprehensively divide the chromatographic plane into regions that distinguish chromatographic features and then quantify the response in each region. The regions are incorporated into the *consensus template*. The positions of the regions in the retention-times plane are defined relative to the pattern of peaks in the *consensus template* and transformed with the template peaks during matching, so their relative positions are maintained when the *consensus template* peaks are matched to detected peaks in a chromatogram.

The comprehensive subdivision of the chromatographic plane is implemented with a new construct called a mesh — a polygon that is subdivided into non-overlapping polygonal panels. Currently, the subdivision is performed interactively with a set of convenient tools that make mesh editing simple

and fairly fast. The analyst outlines a region of the chromatogram and then draws subdividing polylines to delineate chromatographic features such as peaks or peak sets. Ongoing work will automate mesh subdivision.

In these experiments, the comprehensive mesh was created based on a cumulative chromatogram formed by summing all of the chromatograms in a set. The individual chromatograms can be aligned or not aligned before summing. Summing aligned chromatograms facilitates finer delineation in the mesh panels, whereas summing non-aligned chromatograms takes into account chromatographic variations in delineating the mesh panels. Here, the chromatographic variations were small and no alignment was performed in creating the cumulative chromatogram. The chromatograms were intensity-normalized before summing.

Figure 5 shows the cumulative chromatogram for three samples from standard-roast Brazilian, Colombian, and Guatemalan coffees and the meshes created for fingerprinting. In this analysis, 34 meshes covering the chromatographic features were divided into 1109 panels. The number of panels is on the order of, but less than, the number of peaks. Some panels were drawn to encompass more than one detected peak, e.g., along streaks of column bleed or for co-eluted peaks.

The pattern of reliable peaks in the consensus template is matched to the pattern of detected peaks in each chromatogram and the matched peaks in each chromatogram are labeled with the name of the matched template peak. For untargeted analysis, the chemical name is not known and so an ID number is used. The meshes in the consensus template are copied into each chromatogram with the least-squares-optimal retention-times transformation (geometric scaling and translation) determined from the peak matches. This maintains the positions of the mesh panels in the retention-times plane relative to the peaks.

The response in each panel can be computed in one of two ways: (1) treat the panel as a region (or area) of the chromatogram and sum the response at all datapoints in the panel or (2) treat the panel as a peak selector and sum the response of all peaks whose apex (the datapoint which has the largest value in a peak) is in the panel. In the results here, method (1), computing the response as the sum of datapoints in the panel, is used. This provides a quantitative measurement in each panel. The quantitative measurement of the response in each panel is one minutiae of the fingerprint. The set of all minutiae for a sample is its fingerprint.

Fingerprint analysis

Complex samples have extensive fingerprints. For example, the standard-roast coffee fingerprints from the mesh shown in **Figure 5** have 1109 minutiae. It is useful to list the most significant minutiae, but significance may depend on the goal of analysis. Sifting of the many minutiae can

indicate potentially significant chromatographic features. If many samples are available, then methods for dimensionality reduction, such as principal component analysis (PCA) or spectral clustering could be employed. Here, with three samples, the minutiae are sifted in various ways to generate tables of potentially significant features.

Table 5a lists the 15 minutiae with the largest average percent response, i.e., the response within the mesh panel divided by the response within the entire chromatogram. The logic of this sifting is that these minutiae indicate the regions of the chromatogram with the largest responses, presumably produced by the compounds that are the major constituents of the sample. The first column indicates the rank of the region's average percent response; the second and third columns list the average retention times of the region's apex; the third column lists the region's average percent response; the fourth through sixth columns provide the percent response in each chromatogram; and the last column refers to the marker compound list in **Table 2**. The largest percent response on each row is in bold and the smallest percent response is in italics. Eleven of the largest percent response minutiae are marker compounds. In **Figure 6a** the retention times of the minutiae listed in **Table 5a** are highlighted on the cumulative chromatogram of the set of standard-roast coffee samples.

Table 5b lists the 15 minutiae with the largest standard deviation in percent response for the set of chromatograms. The logic of this sifting is that these minutiae indicate the regions of the chromatogram in which the differences between the samples were quantitatively largest. Twelve of the 15 minutiae in the table are marker compounds. Eleven of the 15 minutiae in this table are among those with the largest percent response listed in **Table 5a**. It is not surprising that many of the major compounds exhibit the largest absolute difference in percent response. This table has two markers that are not among the major compounds listed in **Table 5a**: Marker 6 at Rank 5 and Marker 39 at Rank 12. In **Figure 6b**, the retention times of the minutiae listed in **Table 5b** are highlighted on the colorized fuzzy difference of the standard-roast Brazilian and standard-roast Colombian coffees.

Table 5c lists the 15 minutiae with the largest relative standard deviation in percent response, i.e., the standard deviation in percent response divided by the average percent response. Only minutiae with an average percent response of at least 0.01% (the denominator of the ratio) were ranked. The logic of this sifting is that small absolute quantitative differences still might be important if the relative difference is large. Three of these compounds are marker compounds, none of which are listed in **Tables 5a** or **5b**. In **Figure 6c**, the retention times of the minutiae listed in **Table 5c** are highlighted on the colorized fuzzy ratio of the standard-roast Brazilian and standard-roast Colombian coffees.

Non-targeted analysis: Template matching on Juniper samples.

Non-targeted fingerprint analysis was performed with a set of five chromatograms of Juniper samples (numbered A_1, B_4, B_5, C_4, and D_1). The cumulative chromatogram, created by summing the intensity-normalized sample chromatograms, as described in the previous section, is shown in **Figure 7.** The template used for analysis (overlayed on **Figure 7**) consists of a set of 109 consensus peaks for the set of five samples (retention times indicated by yellow circles in **Figure 7**) and 727 panels in six meshes (outlined in black in **Figure 7**). Template matching with the consensus peaks was performed on each chromatogram and the meshes were transformed consistent with the peak matching. The fingerprint of each chromatogram then was created as the list of all panels with the total response in each panel (just as was done for the coffee samples).

Table 6 lists potentially significant minutiae sifted with the same criteria used in the previous section: (6a) largest mean percent response, (6b) largest percent response standard deviation, and (6c) largest percent response standard deviation relative to mean percent response. Because the chromatograms for the Juniper samples had fewer peaks than the chromatograms for the coffee samples, the minimum mean percent response for inclusion by the third criterion was set to 0.1% rather than 0.01% for the standard-roast coffee analysis. Many of the minutiae selected by the first two criteria contain peaks listed in **Table 4**. The differences between chromatograms B_4 and B_5 are especially notable for minutiae selected by both criteria. The third criterion may be useful for identifying distinguishing trace constituents.

CONCLUSIONS

Fingerprint analysis can be highly useful for many purposes, including sample comparison and classification, but it is not a detailed assay of individual constituents. In particular, the specification of comprehensive mesh panels, whether interactive or automatic, at present may still delineate features incompletely (e.g., placing two important chromatographic features in the same panel) or incorrectly (e.g., splitting a chromatographic feature into two panels). Co-elutions and chromatographic variations still may cause problems, so as with targeted analysis selectivity, repeatability, and reproducibility are important. The fingerprints described here are defined by the total responses in the mesh panels, but spectral information could be extracted into the fingerprint, which could be especially useful for co-elutions.

Fingerprint analysis and targeted analysis can be combined to be more effective than either technique alone. Targeted analysis can better quantify individual compounds (especially when interactively performed by an expert analyst), but fingerprint analysis may identify compounds that

are comparatively significant but are not known targets. So, a productive analytical approach is to use fingerprinting to identify potentially significant chromatographic features and then use that information to inform the development of a list of compounds for targeted analysis.

Acknowledgments

The authors are indebted with Prof. Jan Karlsen (University of Oslo) for supplying Juniper samples and for helpful discussion and advice.

This research was carried out within the project entitled: “Sviluppo di metodologie innovative per l’analisi di prodotti agroalimentari” (FIRB Cod.: RBIP06SXMR_002) of the Ministero dell’Istruzione, dell’Università e della Ricerca (MIUR) (Italy).

REFERENCES

[1] J. Dalluge, J. Beens, U.A.Th. Brinkman. Comprehensive two-dimensional gas chromatography: a powerful and versatile analytical tool. *J. Chromatogr. A* **1000**: 69-108 (2003).

[2] M. Adahchour, J. Beens, R.J.J. Vreuls, U.A.Th. Brinkman. Recent developments in comprehensive two-dimensional gas chromatography (GCxGC) I. Introduction and instrumental set-up. *Trends Anal. Chem.* **25**: 438-454 (2006).

[3] M. Adahchour, J. Beens, U.A.Th. Brinkman. Recent developments in the application of comprehensive two-dimensional gas chromatography. *J. Chromatogr. A* **1186**: 67–108 (2008).

[4] K. M. Pierce, J. C. Hoggard, R. E. Mohler, R. E. Synovec. Recent advancements in comprehensive two-dimensional separations with chemometrics. *J. Chromatogr. A*, **1184**: 341–352 (2008).

[5] C. Cordero, C. Bicchi, P. Rubiolo. Group-Type and Fingerprint Analysis of Roasted Food Matrices (Coffee and Hazelnut Samples) by Comprehensive Two-Dimensional Gas Chromatography. *J Agr Food Chem.* **56:** 7655-7666 (2008).

[6] C. Bicchi, O. Panero, G. Pellegrino, A. Vanni. Characterization of Roasted Coffee and Coffee Beverages by Solid Phase Microextraction-Gas Chromatography and Principal Component Analysis. *J. Agric. Food Chem.* **45**: 4680-4686 (1997).

[7] C. Bicchi, C. Cordero, E. Liberto, B. Sgorbini, P. Rubiolo. Headspace sampling of the volatile fraction of vegetable matrices. *J. Chromatogr. A*, **1184**: 220–233 (2008).

[8] I. Flament. Coffee flavor chemistry. Chichester: John Wiley & Sons. 2001

[9] M. Czerny, F. Mayer, and W. Grosch. Sensory Study on the Character Impact Odorants of Roasted Arabica Coffee. *J. Agric. Food Chem.* **47**: 695-699 (1999)

[10] I. Blank, A. Sen, W. Grosch. Potent odorants of the roasted powder and brew of Arabica coffee. *Z. Lebensm.-Unters. Forsch.* **195**: 239-245 (1992)

[11] M. Czerny, R. Wagner, W.Grosch. Detection of odor-active ethenylalkylpyrazines in roasted coffee. *J. Agric. Food Chem.* **44**: 3268-3272 (1996)

[12] W. Holscher, O. Vitzthum, G. H. Steinhart. Identification and sensorial evaluation of aroma-impact compounds in roasted Colombian coffee. *Cafe, Cacao, The*, **34**: 205-212 (1990).

[13] B. Barjaktarovic´, M. Sovilj, Z. Knez. Chemical composition of Juniperus communis L. fruits supercritical CO_2 extracts: Dependence on pressure and extraction time. *J. Agric. Food Chem.*, **53**: 2630–2636(2005).

[14] H. Kallio, K. Junger-Mannermaa, Maritime influence on the volatile terpenes in the berries of different ecotypes of Juniper (Juniperus communis) in Finland. *J. Agric. Food Chem.* **37**: 1013–1016(1989).

[15] R. I. Aylott. Vodka, Gin and other flavoured spirits. In A. G. H. Lea & J. R. Piggott (Eds.), Fermented beverage production. New York: Kluwer Academic/Plenum Publishers 289–308 (2003).

[16] BM. Lawrence. Progress in essential oil. *Perfum. Flavor.* **26(4)**: 68-75 (2001).

[17] A. Angioni, A. Barra, M. T. Russo, V. Coroneo, P.Cabras. Chemical composition of the essential oils of Juniperus from ripe and unripe berries and leaves and their antimicrobial activity. *J. Agric. Food Chem.* **51**: 3073–3080 (2003).

[18] A. Baerheim Svendsen, J.J.C. Scheffer, A. Looman. A comaparative study of the composition of the essential needle oils of Norwegian lowlands juniper and high-mountains juniper. *Scientia Pharmaceutica***. 53:** 159-161 (1985).

[19] A. Looman, A. Baerheim Svendsen. The needle essential oil of norwegian mountain juniper, juniperus communis L. var. Saxatilis Pall. *Flavour Fragr. J.* **7**: 23–26 (1992)

[20] S. Reichenbach, M. Ni, V. Kottapalli,, and A. Visvanathan. Information Technologies for Comprehensive Two-Dimensional Gas Chromatography. *Chem. Intell. Lab. Syst.* **71(2)**:107–120 (2004).

[21] S. Reichenbach, P. Carr, D. Stoll, and Q. Tao. Smart Templates for Peak Pattern Matching with Comprehensive Two-Dimensional Liquid Chromatography. *J. Chromatogr. A*, http://dx.doi.org/10.1016/j.chroma.2008.09.058, 2008.

[22] S. Reichenbach, V. Kottapalli, M. Ni, and A. Visvanathan. Computer Language for Identifying Chemicals with Comprehensive Two-Dimensional Gas Chromatography and Mass Spectrometry (GCxGC-MS). *J. Chromatogr. A*, **1071**:263–269 (2005).

[23] S. Reichenbach, M. Ni, D. Zhang, and E. Ledford. Image Background Removal in Comprehensive Two-Dimensional Gas Chromatography. *J. Chromatogr. A*, **985**:47–56 (2003).

[24] S. Stein. NIST Mass Spectral Search Program (Version 2.0f). NIST Mass Spectrometry Data Center, 2008.

[25] B. Hollingsworth, S. Reichenbach, Q. Tao, and A. Visvanathan. Comparative Visualization for Comprehensive Two-Dimensional Gas Chromatography. *J. Chromatogr. A*, **1105**:51–58 (2006).

[26] M. Ni, S. Reichenbach, A. Visvanathan, J. TerMaat, and E. Ledford, Jr. Peak Pattern Variations Related to Comprehensive Two-Dimensional Gas Chromatography Acquisition. *J. Chromatogr. A*, **1086**:165–170 (2005).

Captions to Tables:

Table 1: *Coffea arabica* samples: sample acronyms, geographical origin, post-harvest treatment and roasting time/temperature profiles.

Table 2: List of markers adopted for the target characterization approach of *Coffea arabica* samples: ID number, compound name, group name, Retention Indices (RI), ^{1}D and ^{2}D retention times. Markers were identified on the basis of their linear retention indices and MS-EI spectra compared with those of authentic standards (indicated with *ref*) or tentatively identified through their MS-EI fragmentation patterns and retention indices.

Table 3: normalized peak volumes (mean of 3 replicates – RSD% below 15) distribution of the nineteen pyrazine derivatives identified in the Guatemala, Brazil and Colombia coffee samples submitted to standard and mild roasting conditions.

Table 4: Marker compounds for target characterization of *Juniper communis* L. samples: compound name, ^{1}D and ^{2}D retention times, relative abundance on selected samples. Markers were identified on the basis of their MS-EI spectra compared with those of authentic standards or tentatively identified through their MS-EI fragmentation patterns.

Table 5: Selected standard-roast coffee minutiae. The first part (5a) lists the 15 minutiae with the largest mean percent response, first and second dimension retention times (^{1}D and ^{2}D) and target analytes ID (for target ID refer to Table 2). The second part (5b) lists the 15 minutiae with the largest percent response standard deviation. The third part (5c) lists the 15 minutiae with the largest relative percent response standard deviation.

Table 6: Selected juniper fingerprint minutiae. The first part (6a)lists the 15 minutiae with the largest mean percent response, first and second dimension retention times (^{1}D and ^{2}D) and target analytes ID (for target ID refer to Table 4). The second part (6b) lists the 15 minutiae with the largest percent response standard deviation. The third part (6c) lists the 15 minutiae with the largest relative percent response standard deviation

Captions to Figures

Figure 1: pyrazine 2D pattern of Arabica samples from Guatemala, Brazil and Colombia submitted to a standard roasting. Results are reported as normalized 2D-Peak Volume over the ISTD, for analyte ID (*x*-axis) see Table 2.

Figure 2: key-aroma marker 2D pattern of Arabica samples from Guatemala, Brazil and Colombia submitted to a standard roasting. Results are reported as normalized 2D-Peak Volume over the ISTD. For analyte ID (*x*-axis) see Table 2.

Figure 3: GCxGC-MS chromatogram for standard-roast Colombian coffee. Circles indicate the retention times of 1652 peaks. Black circles indicate the subset of 891 reliable peaks that were consistently matched for a set of three chromatograms (including this chromatogram and chromatograms of standard-roast Brazilian and Guatemalan coffees) and yellow circles indicate unreliable peaks that were not matched consistently for the set.

Figure 4: Pseudocolor comparisons of chromatograms from standard-roast Brazilian and Colombian coffees. **4a** shows the colorized fuzzy difference and **4b** shows the colorized fuzzy ratio. In both, green indicates a larger response for the Brazilian sample and red indicates a larger response for the Colombian sample.

Figure 5: Mesh panels (shown as black polygons) for analysis of the set of three chromatograms from standard-roast coffee overlaid on the cumulative chromatogram. There are 34 meshes covering the chromatographic features divided into 1109 panels. The *consensus template* also contains the reliable peaks shown in Figure 3.

Figure 6: The first image (**6a**) indicates the apex retention times of the 15 minutiae with the largest mean percent response. The second image (**6b**) indicates the apex retention times of the 15 minutiae with the largest percent response standard deviation. The third image (**6c**) indicates the apex retention times of the 15 minutiae with the largest relative percent response standard deviation.

Figure 7: Cumulative chromatogram for a set of six Juniper samples with a template of reliable peaks indicated by yellow circles and mesh panels indicated by black outlines.

Table 1: *Coffea arabica* samples: sample acronyms, geographical origin, post-harvest treatment and roasting time/temperature profiles.

Sample acronym	Geographical origin	Post-harvest treatment	Degree of Roasting	Roasting conditions	
				Time (min)	Temp. (°C)
Colombia Green	Colombia	washed	No roasting	-	-
Colombia Mild	Colombia	washed	Mild	10.29	194
Colombia Standard	Colombia	washed	Standard	10.21	203
Guatemala Green	Guatemala	washed	No roasting	-	-
Guatemala Mild	Guatemala	washed	Mild	9.46	187
Guatemala Standard	Guatemala	washed	Standard	10.28	194
Brazil Green	Brazil	natural	No roasting	-	-
Brazil Mild	Brazil	natural	Mild	9.24	202
Brazil Standard	Brazil	natural	Standard	10.15	185

Table 2: List of markers adopted for the target characterization approach of *Coffea arabica* samples: ID number, compound name, group name, Retention Indices (RI), ^{1}D and ^{2}D retention times. Markers were identified on the basis of their linear retention indices and MS-EI spectra compared with those of authentic standards (indicated with *ref*) or tentatively identified through their MS-EI fragmentation patterns and retention indices.

ID	Compound name	Group Name	Identification	RI	^{1}D (min)	^{2}D (s)
1	Acetaldehyde	key aroma	*ref*	546	3.42	2.48
2	2-Propanone	target group	*ref*	550	3.69	1.64
3	Formic acid	target group	*tentatively*	552	3.82	0.51
4	2,3-Butanedione	key aroma	*ref*	559	4.22	4.21
5	Acetic acid	target group	*ref*	560	4.29	0.51
6	2-Methylfuran	target group	*ref*	761	4.35	2.02
7	3-Methylbutanal	key aroma	*ref*	772	5.02	4.38
8	2-Methylbutanal	key aroma	*ref*	772	5.02	4.38
9	1-Hydroxy-2-Propanone	target group	*ref*	772	5.02	0.67
10	Propanoic acid	target group	*ref*	778	5.35	1.35
11	2,3-Pentanedione	key aroma	*ref*	779	5.42	4.50
12	3-Hydroxy-2-Butanone	target group	*ref*	786	5.82	0.80
13	Butanoic acid	target group	*ref*	813	7.42	2.36
14	2,3-Butanediol	target group	*ref*	823	8.02	1.94
15	Methylpyrazine	target group	*ref*	841	9.09	5.09
16	2-Furancarboxaldehyde	target group	*ref*	846	9.42	1.73
17	3-Methylbutanoic acid	target group	*ref*	856	10.02	2.86
18	2-Furanmethanol	target group	*ref*	862	10.35	2.69
19	3-Methyl-4-Heptanone	target group	*ref*	887	11.89	1.26
20	2-Furfuryl formate	target group	*tentatively*	902	12.75	1.64
21	3-(methylthio)-Propanal	key aroma	*ref*	904	12.89	1.94
22	2-Acetylfurane	target group	*ref*	906	13.02	1.94
23	2-Furanmethanethiol	key aroma	*ref*	908	13.09	1.43
24	2,5-Dimethylpyrazine	target group	*ref*	910	13.22	1.39
25	Ethylpyrazine	target group	*ref*	914	13.49	1.39
26	Ethenylpyrazine	target group	*tentatively*	929	14.35	1.52
27	5-Methylfurfural	target group	*ref*	957	16.02	2.61
28	Benzaldehyde	target group	*ref*	959	16.15	1.94
29	Hexanoic acid	target group	*ref*	972	16.95	3.79
30	2-Furfuryl acetate	target group	*ref*	988	17.89	2.02
31	2-Ethyl-6-Methylpyrazine	target group	*ref*	996	18.35	1.68
32	2-Ethyl-3-Methylpyrazine	target group	*ref*	999	18.55	1.64
33	Propylpyrazine	target group	*ref*	1008	19.15	1.64
34	2-Ethenyl-6-Methylpyrazine	target group	*tentatively*	1015	19.62	1.81
35	2-Acetypyrazine	target group	*ref*	1019	19.89	2.19
36	2-Ethenyl-5-Methylpyrazine	target group	*ref*	1019	19.89	1.77
37	Benzeneacetaldehyde	target group	*ref*	1041	21.35	2.44
38	Furaneol	key aroma	*ref*	1054	22.29	4.04
39	5-Ethyl-2,3-Dimethylpyrazine	target group	*tentatively*	1074	23.62	1.73
40	2-Ethyl-3,5-Dimethylpyrazine	key aroma	*tentatively*	1080	24.02	1.73
41	2-Methoxyphenol	key aroma	*ref*	1083	24.22	2.69
42	3-Ethyl-2,5-Dimethylpyrazine	target group	*tentatively*	1083	24.22	1.77
43	2,5-Diethylpyrazine	target group	*ref*	1089	24.62	1.77
44	2-Ethyl-6-Vinylpyrazine	target group	*ref*	1111	26.15	2.40
45	2-Acetyl-6-Methylpyrazine	target group	*tentatively*	1117	26.55	2.31
46	2,3-Diethyl-5-Methylpyrazine	key aroma	*ref*	1148	28.75	1.81
47	3,5-Diethyl-2-Methylpyrazine	target group	*ref*	1152	29.02	1.81
48	2-Allyl-5-Methylpyrazine	target group	*tentatively*	1159	29.49	2.06
49	4-Ethyl-2-Methoxyphenol	key aroma	*ref*	1272	37.22	2.78
50	2-Methoxy-4-Vinylphenol	key aroma	*ref*	1308	39.69	3.20
51	β-*E*-Damascenone	key aroma	*ref*	1377	44.15	2.53

Table 3: normalized peak volumes (mean of 3 replicates – RSD% below 15) distribution of the nineteen pyrazine derivatives identified in the Guatemala, Brazil and Colombia coffee samples submitted to standard and mild roasting conditions.

ID	Compound name	^{1}D (min)	^{2}D (s)	Normalized 2D-peak volumes					
				Guatemala		Brazil		Colombia	
				mild	standard	mild	standard	mild	standard
15	Methylpyrazine	9.09	1.09	168	194	182	210	137	141
24	2,5-Dimethylpyrazine	13.22	1.39	177	165	171	205	128	133
25	Ethylpyrazine	13.49	1.39	52	63	41	50	49	60
26	Ethenylpyrazine	14.35	1.52	9	13	11	13	7	11
31	2-Ethyl-6-Methylpyrazine	18.35	1.68	54	66	54	58	40	43
32	2-Ethyl-3-Methylpyrazine	18.55	1.64	90	81	89	97	67	51
33	Propylpyrazine	19.15	1.64	5	7	5	7	3	5
34	2-Ethenyl-6-Methylpyrazine	19.62	1.81	18	13	13	14	9	16
35	2-Acetypyrazine	19.89	2.19	16	17	18	18	13	13
36	2-Ethenyl-5-Methylpyrazine	19.89	1.77	8	8	8	10	7	7
39	5-Ethyl-2,3-Dimethylpyrazine	23.62	1.73	52	39	40	53	33	21
40	2-Ethyl-3,5-Dimethylpyrazine	24.02	1.73	6	6	8	13	6	7
42	3-Ethyl-2,5-Dimethylpyrazine	24.22	1.77	13	13	3	4	10	9
43	2,5-Diethylpyrazine	24.62	1.77	3	3	3	5	2	2
44	2-Ethyl-6-Vinylpyrazine	26.15	2.40	13	13	14	14	2	2
45	2-Acetyl-6-Methylpyrazine	26.55	2.31	26	25	28	31	10	10
46	2,3-Diethyl-5-Methylpyrazine	28.75	1.81	3	3	1	2	2	2
47	3,5-Diethyl-2-Methylpyrazine	29.02	1.81	5	6	6	9	5	5
48	2-Allyl-5-Methylpyrazine	29.49	2.06	10	10	8	9	5	8
			SUM	**727**	**747**	**702**	**821**	**536**	**545**

Table 4: Marker compounds for target characterization of *Juniper communis* L. samples: compound name, ^{1}D and ^{2}D retention times, relative abundance on selected samples. Markers were identified on the basis of their MS-EI spectra compared with those of authentic standards or tentatively identified through their MS-EI fragmentation patterns.

	Compound name	**^{1}D (min)**	**^{2}D (s)**	**Relative Abundance (%)**				
ID				**Juniper A_*1***	**Juniper B_*4***	**Juniper B_*5***	**Juniper C_*4***	**Juniper D_*1***
1	acetic acid	4.27	1.26	n.d.	n.d.	n.d.	n.d.	tr
2	hexanal	8.53	1.39	0.1	tr	tr	0.1	0.1
3	*E*-2-hexenal	10.67	1.70	0.1	0.1	0.1	0.1	1.2
4	tricyclene	13.47	1.34	tr	0.2	tr	0.1	0.1
5	α-thuyene	13.60	1.35	2.2	0.2	4.1	0.1	0.3
6	α-pinene	14.13	1.39	21.9	59.3	12.8	33.5	36.7
7	camphene	14.80	1.44	0.2	0.4	0.1	0.2	0.2
8	benzaldehyde	15.40	2.09	tr	0.1	tr	tr	0.2
9	sabinene	16.00	1.52	24.0	0.8	44.6	0.5	0.5
10	β-pinene	16.20	1.48	2.3	2.4	1.1	1.9	1.7
11	myrcene	16.60	1.52	8.8	6.0	7.9	9.3	8.3
12	Δ-2-carene	17.13	1.48	0.3	0.1	0.2	0.7	0.6
13	Δ–3-carene	17.67	1.52	4.0	6.6	1.9	8.7	7.1
14	α-terpinene	18.00	1.52	0.4	tr	1.0	0.1	0.1
15	p-cymene	18.36	1.65	0.7	0.2	0.4	1.4	2.0
16	limonene	18.63	1.56	16.1	6.3	4.7	19.2	3.4
17	β-phellandrene	18.76	1.61	3.6	0.8	0.7	11.6	7.4
18	β-ocimene	19.33	1.57	0.1	2.7	0.1	0.7	0.1
19	γ-terpinene	20.07	1.61	1.3	0.1	2.4	0.1	0.1
20	*cis* sabinene hydrate	20.70	1.96	n.d.	0.2	n.d.	0.1	n.d.
21	α-terpinolene	21.49	1.65	3.4	1.9	4.5	1.9	2.0
22	*p*-cymenene	21.60	1.80	tr	tr	tr	0.1	0.1
23	*trans*-sabinene hydrate	22.23	2.08	0.2	tr	0.1	tr	n.d
24	terpinen-4-ol	26.10	2.09	0.4	0.1	0.3	0.1	0.1
25	α-terpineol	26.77	2.17	tr	tr	0.1	0.4	0.3
26	bornyl acetate	30.94	2.11	0.1	0.1	0.2	0.4	0.4
27	terpinyl acetate	31.40	2.11	n.d.	0.1	n.d.	n.d.	tr
28	α-cubebene	33.80	1.83	0.1	0.1	tr	0.1	0.1
29	α-copaene	35.13	1.87	0.1	tr	tr	0.2	n.d
30	β-bourbonene	35.40	1.94	n.d.	tr	tr	n.d.	0.1
31	myrtanol acetate	35.53	1.90	n.d.	n.d.	n.d.	n.d.	tr
32	β-elemene	35.73	1.96	1	1.5	0.9	1.5	1.3
33	γ-elemene	37.43	2.00	0.6	0.1	0.2	0.1	0.3
34	α-humulene	38.64	2.09	0.6	0.8	2.0	0.5	5.0
35	γ-muurolene	39.40	2.01	0.1	0.1	tr	0.1	0.1
36	germacrene D	39.67	2.13	1.5	4.6	2.2	1.6	4.9
37	β-selinene	40.07	2.07	0.1	0.1	0.1	0.1	0.4
38	bicyclogermacrene	40.27	2.09	0.2	0.8	0.3	0.2	0.7
39	γ-cadinene	40.93	2.09	0.1	0.1	0.1	0.2	0.2
40	δ-cadinene	41.07	2.04	1.3	0.9	0.5	1.5	0.8
41	α-cadinene	41.87	2.01	0.1	0.1	0.1	0.1	0.1
42	germacrene B	42.87	2.13	1.6	0.1	0.5	0.2	0.9
43	caryophyllene oxide	43.70	2.48	n.d.	n.d.	n.d.	n.d.	0.1

Table 5: Selected standard-roast coffee minutiae. The first part (5a) lists the 15 minutiae with the largest mean percent response, first and second dimension retention times (^{1}D and ^{2}D) and target analytes ID (for target ID refer to Table 2). The second part (5b) lists the 15 minutiae with the largest percent response standard deviation. The third part (5c) lists the 15 minutiae with the largest relative percent response standard deviation.

Table 5a

Rank	^{1}D (min)	^{2}D (s)	Std Dev	Brazil	Colombia	Guatemala	Target ID
1	10.35	2.64	13.6919	***14.0311***	*13.0304*	14.0142	18
2	16.04	2.64	6.6054	*6.0931*	**7.1828**	6.5402	27
3	9.42	1.74	4.8270	*3.8194*	**5.9078**	4.7537	16
4	17.93	1.99	3.5881	**3.6550**	3.5565	*3.5528*	30
5	9.13	1.04	3.4910	**3.9582**	*2.9041*	3.6107	15
6	4.29	0.51	3.4191	3.2690	*3.1775*	**3.8108**	5
7	10.80	2.23	3.1921	3.1846	**3.2474**	*3.1444*	
8	6.64	0.72	2.9283	**3.0356**	*2.8312*	2.9182	
9	13.22	1.44	2.6059	**2.8419**	*2.3215*	2.6543	24
10	18.62	1.66	1.4742	**1.6840**	*1.1619*	1.5769	32
11	13.15	3.80	1.3769	**1.4239**	*1.3435*	1.3634	
12	13.49	1.40	1.2316	**1.5180**	1.1284	*1.0485*	25
13	26.00	3.31	1.2006	**1.2356**	1.1996	*1.1666*	
14	18.38	1.67	1.0378	1.1287	*0.7820*	**1.2025**	31
15	39.73	3.14	1.0140	*0.8873*	**1.2593**	0.8954	50

Table 5b

Rank	^{1}D (min)	^{2}D (s)	Std Dev	Brazil	Colombia	Guatemala	Target ID
1	9.42	1.74	1.0462	*3.8194*	**5.9078**	4.7537	16
2	10.35	2.64	0.5729	**14.0311**	*13.0304*	14.0142	18
3	16.04	2.64	0.5478	*6.0931*	**7.1828**	6.5402	27
4	9.13	1.04	0.5371	**3.9582**	*2.9041*	3.6107	15
5	4.33	1.91	0.3644	0.9195	**1.3296**	*0.6028*	6
6	4.29	0.51	0.3423	3.2690	*3.1775*	**3.8108**	5
7	18.62	1.66	0.2758	**1.6840**	*1.1619*	1.5769	32
8	13.22	1.44	0.2635	**2.8419**	*2.3215*	2.6543	24
9	13.49	1.40	0.2512	**1.5180**	1.1284	*1.0485*	25
10	18.38	1.67	0.2245	1.1287	*0.7820*	**1.2025**	31
11	39.73	3.14	0.2125	*0.8873*	**1.2593**	0.8954	50
12	23.64	1.74	0.1822	0.7643	*0.4589*	**0.7837**	39
13	5.00	2.93	0.1371	*0.0383*	**0.2854**	0.0589	
14	5.09	0.32	0.1191	0.4693	*0.2730*	**0.4880**	
15	6.64	0.72	0.1026	**3.0356**	*2.8312*	2.9182	

Table 5c

Rank	^{1}D (min)	^{2}D (s)	Rel Std Dev	Brazil	Colombia	Guatemala	Target ID
1	19.42	3.96	1.4512	0.0029	**0.0411**	*0.0021*	
2	8.02	0.52	1.3599	**0.0924**	*0.0077*	*0.0077*	
3	28.75	1.87	1.2854	0.0037	*0.0029*	**0.0318**	46
4	30.11	2.40	1.2368	0.0045	**0.0323**	*0.0031*	
5	24.02	1.74	1.1218	0.0211	*0.0031*	**0.0733**	40
6	5.00	2.93	1.0753	*0.0383*	**0.2854**	0.0589	
7	7.66	1.84	1.0120	**0.1444**	0.0297	*0.0257*	
8	23.98	2.09	0.9775	0.0290	*0.0172*	**0.1114**	
9	8.00	1.94	0.9370	**0.1668**	*0.0346*	0.0390	14
10	20.53	3.96	0.8679	*0.0047*	**0.0330**	0.0127	
11	29.35	1.81	0.8400	**0.0234**	0.0074	*0.0050*	
12	20.62	3.21	0.8362	*0.0068*	**0.0298**	0.0090	
13	14.35	2.48	0.7784	**0.0254**	0.0084	*0.0065*	
14	30.93	3.91	0.7432	0.0094	**0.0214**	*0.0044*	
15	25.04	3.91	0.7387	0.0165	**0.0415**	*0.0098*	

Table 6: Selected juniper fingerprint minutiae. The first part (6a)lists the 15 minutiae with the largest mean percent response, first and second dimension retention times (^{1}D and ^{2}D) and target analytes ID (for target ID refer to Table 4). The second part (6b) lists the 15 minutiae with the largest percent response standard deviation. The third part (6c) lists the 15 minutiae with the largest relative percent response standard deviation

Table 6a

Rank	^{1}D (min)	^{2}D (s)	Average	Juniper A_*1*	Juniper B_*4*	Juniper B_*5*	Juniper C_*4*	Juniper D_*1*	Target ID
1	14.26	1.44	12.9336	7.7678	**20.0186**	*7.7527*	14.9357	14.1935	10
2	18.67	1.65	9.1767	10.9812	6.8474	*5.5728*	**15.8854**	6.5965	16
3	16.62	1.59	6.5296	4.5154	4.7077	**16.0891**	3.9084	*3.4273*	11
4	39.71	2.15	5.5815	4.1242	**9.1019**	*3.6288*	4.2073	6.8452	36
5	17.57	1.62	4.8975	4.3377	4.6155	*3.0374*	**7.0728**	5.4240	12
6	37.11	2.08	4.4388	2.3308	3.7380	4.6715	*2.0895*	**9.3642**	33
7	35.71	2.04	4.1955	4.4434	**5.3720**	*2.6057*	4.7771	3.7793	32
8	16.18	1.58	3.7113	2.6565	2.3350	**10.3134**	1.9002	*1.3515*	10
9	38.63	2.12	3.2559	2.1526	2.8471	3.1141	*2.1487*	**6.0171**	34
10	15.93	1.59	3.2469	**8.7470**	*0.0287*	5.9328	0.9833	0.5423	9
11	40.27	2.17	2.7069	2.6394	**3.4917**	*1.4583*	2.6000	3.3450	38
12	37.41	2.04	2.5815	**4.0114**	1.9980	1.9147	*1.6661*	3.3174	33
13	21.45	1.73	2.4276	2.9508	2.0259	**3.4667**	1.9892	*1.7054*	21
14	41.07	2.11	1.3339	1.6744	1.4762	*0.6602*	**1.8962**	0.9627	40
15	42.83	2.19	1.3327	**2.6352**	*0.6225*	0.9893	0.8942	1.5221	42

Table 6b

Rank	^{1}D (min)	^{2}D (s)	Std Dev	Juniper A_*1*	Juniper B_*4*	Juniper B_*5*	Juniper C_*4*	Juniper D_*1*	Target ID
1	16.62	1.59	5.3679	4.5154	4.7077	**16.0891**	3.9084	*3.4273*	11
2	14.26	1.44	5.2278	7.7678	**20.0186**	*7.7527*	14.9357	14.1935	6
3	18.67	1.65	4.2818	10.9812	6.8474	*5.5728*	**15.8854**	6.5965	16
4	15.93	1.59	3.8814	**8.7470**	*0.0287*	5.9328	0.9833	0.5423	9
5	16.18	1.58	3.7230	2.6565	2.3350	**10.3134**	1.9002	*1.3515*	10
6	37.11	2.08	2.9482	2.3308	3.7380	4.6715	*2.0895*	**9.3642**	33
7	39.71	2.15	2.3353	4.1242	**9.1019**	*3.6288*	4.2073	6.8452	36
8	38.63	2.12	1.6011	2.1526	2.8471	3.1141	*2.1487*	**6.0171**	34
9	17.57	1.62	1.4885	4.3377	4.6155	*3.0374*	**7.0728**	5.4240	13
10	13.58	1.42	1.3793	2.0003	*0.0072*	**3.1584**	0.3038	0.2380	**5**
11	19.35	1.69	1.1003	*0.1525*	**2.7059**	0.2527	1.1356	0.1585	**18**
12	20.05	1.66	1.0910	1.6347	0.1966	**2.4762**	0.1424	*0.0822*	**19**
13	35.71	2.04	1.0588	4.4434	**5.3720**	*2.6057*	4.7771	3.7793	32
14	37.41	2.04	1.0258	**4.0114**	1.9980	1.9147	*1.6661*	3.3174	33
15	43.49	2.46	0.8982	2.0657	0.1725	*0.5593*	**2.1147**	0.7560	**43**

Table 6c

Rank	^{1}D (min)	^{2}D (s)	Rel Std Dev	Juniper A_*1*	Juniper B_*4*	Juniper B_*5*	Juniper C_*4*	Juniper D_*1*	Target ID
1	16.29	3.80	2.1837	0.0028	0.0034	**1.6915**	*0.0027*	0.0236	
2	29.33	2.11	2.1270	*0.0057*	0.0129	**1.8950**	0.0205	0.0380	
3	21.55	1.47	2.1168	0.0079	**0.5973**	0.0057	0.0091	*0.0039*	
4	16.45	1.26	2.0740	0.0071	**0.6760**	0.0282	*0.0033*	0.0034	
5	40.55	2.27	2.0401	0.0135	*0.0031*	0.0136	**0.5399**	0.0105	
6	10.46	1.84	1.9510	*0.0087*	0.0093	0.0406	0.0215	**0.6999**	
7	43.43	2.36	1.8927	0.0131	**1.3346**	0.1291	0.0409	*0.0080*	
8	14.82	1.33	1.8229	0.0242	**0.5288**	0.0705	0.0005	*0.0003*	**7**
9	41.79	2.01	1.7324	0.0169	**0.5875**	0.0231	0.0880	*0.0056*	
10	33.70	2.14	1.5932	*0.0053*	0.0164	0.5542	**1.7239**	0.0265	
11	43.74	2.50	1.4931	0.0696	*0.0433*	0.0810	0.1113	**0.8371**	
12	40.55	2.06	1.4207	**0.5168**	*0.0231*	0.0357	0.0769	0.0811	
13	31.43	2.16	1.4201	*0.0027*	0.0381	**0.8293**	0.3011	0.0525	
14	16.02	3.35	1.3672	0.0165	0.0250	0.1255	*0.0077*	**0.3394**	
15	40.01	2.14	1.3602	0.4553	0.0082	*0.0042*	0.0072	**0.6373**	

Figure 1: pyrazine 2D pattern of Arabica samples from Guatemala, Brazil and Colombia submitted to a standard roasting. Results are reported as normalized 2D-Peak Volume over the ISTD, for analyte ID (*x*-axis) see Table 2.

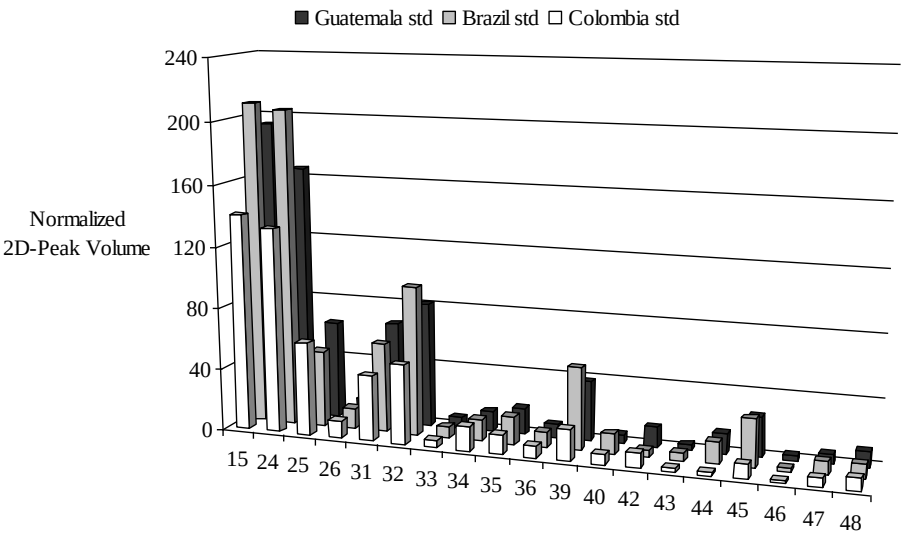


Figure 2: key-aroma marker 2D pattern of Arabica samples from Guatemala, Brazil and Colombia submitted to a standard roasting. Results are reported as normalized 2D-Peak Volume over the ISTD. For analyte ID (*x*-axis) see Table 2.

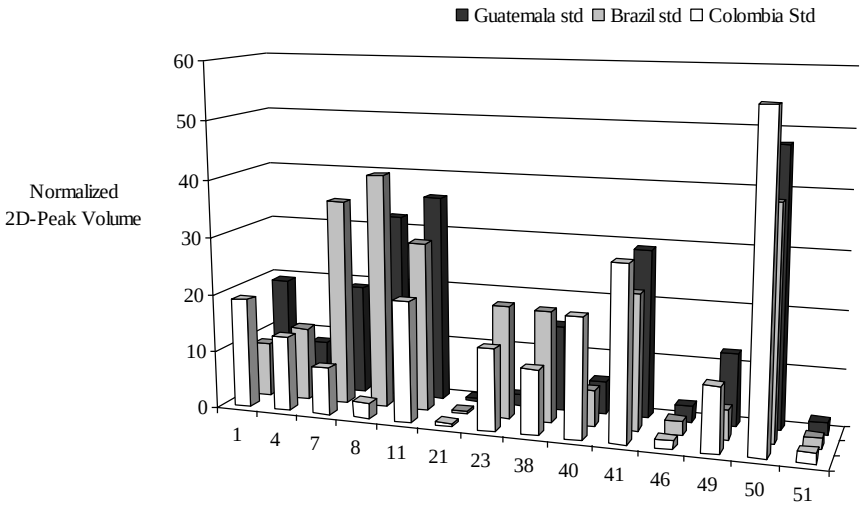


Figure 3: GCxGC-MS chromatogram for standard-roast Colombian coffee. Circles indicate the retention times of 1652 peaks. Black circles indicate the subset of 891 reliable peaks that were consistently matched for a set of three chromatograms (including this chromatogram and chromatograms of standard-roast Brazilian and Guatemalan coffees) and yellow circles indicate unreliable peaks that were not matched consistently for the set.

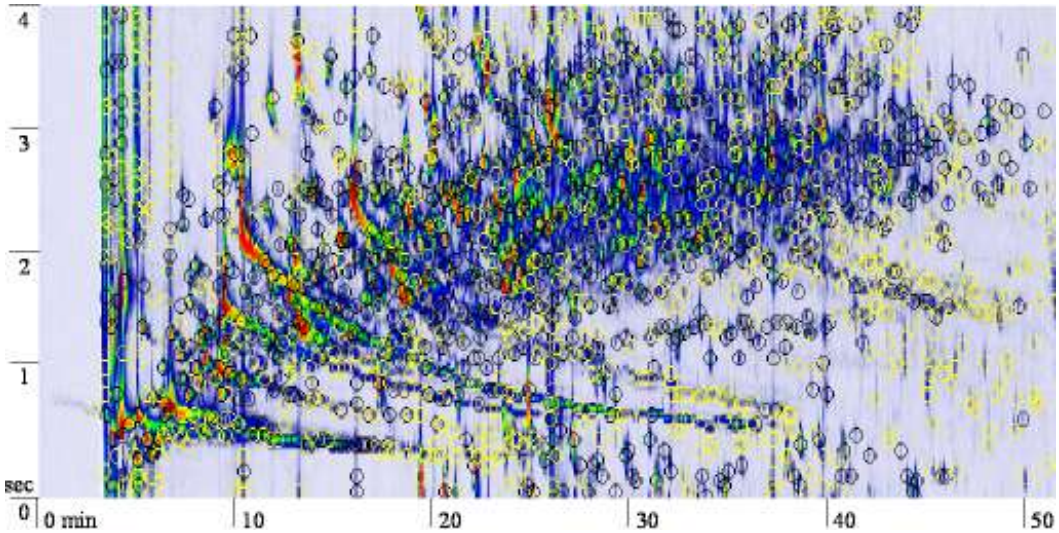
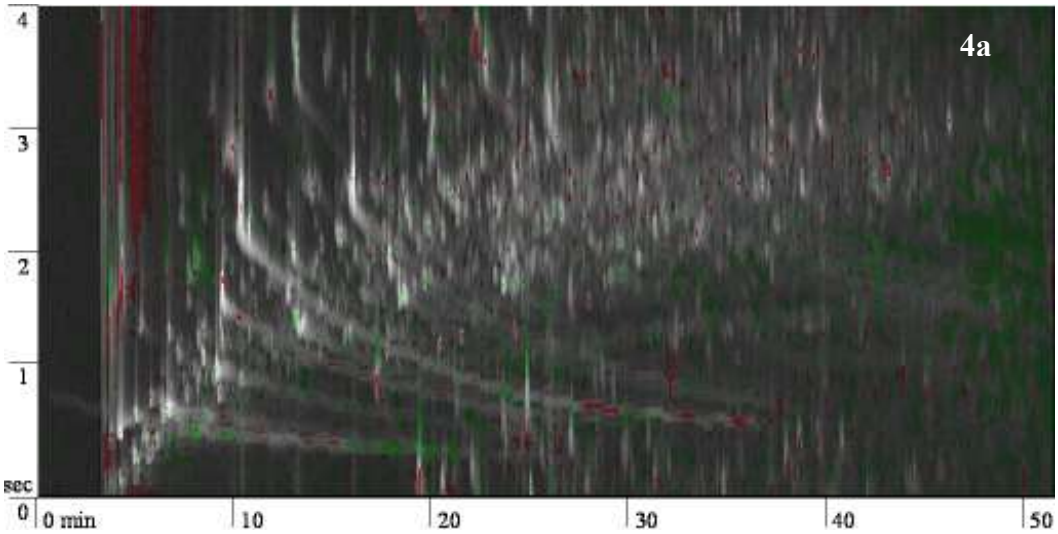


Figure 4: Pseudocolor comparisons of chromatograms from standard-roast Brazilian and Colombian coffees. **4a** shows the colorized fuzzy difference and **4b** shows the colorized fuzzy ratio. In both, green indicates a larger response for the Brazilian sample and red indicates a larger response for the Colombian sample.



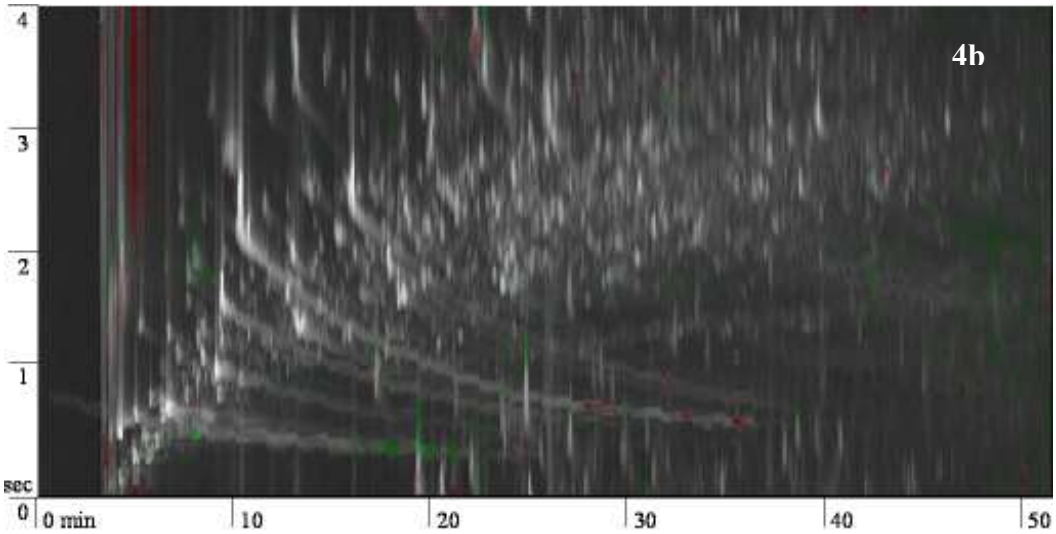


Figure 5: Mesh panels (shown as black polygons) for analysis of the set of three chromatograms from standard-roast coffee overlaid on the cumulative chromatogram. There are 34 meshes covering the chromatographic features divided into 1109 panels. The *consensus template* also contains the reliable peaks shown in Figure 3.

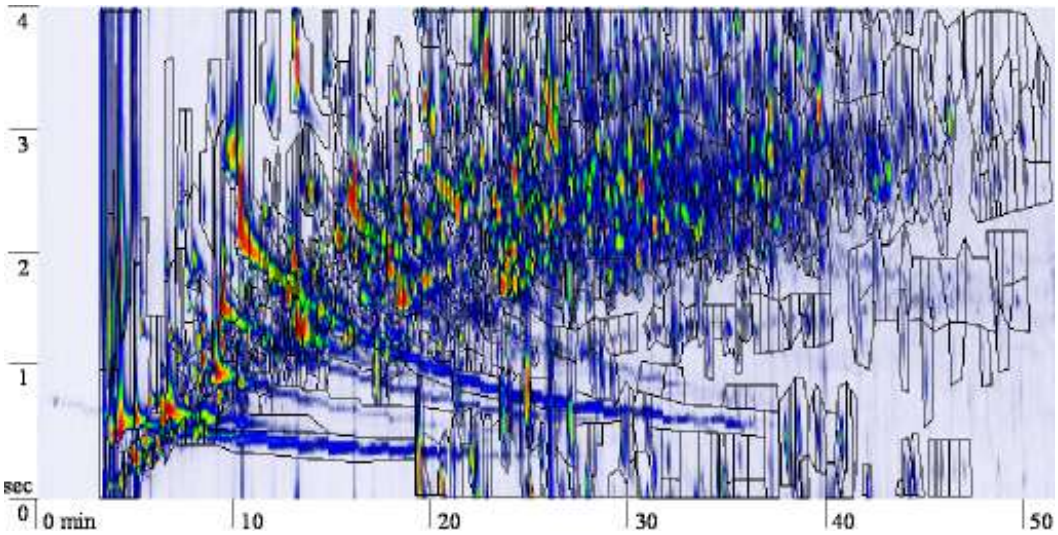


Figure 6: The first image (6a) indicates the apex retention times of the 15 minutiae with the largest mean percent response. The second image (6b) indicates the apex retention times of the 15 minutiae with the largest percent response standard deviation. The third image (6c) indicates the apex retention times of the 15 minutiae with the largest relative percent response standard deviation.

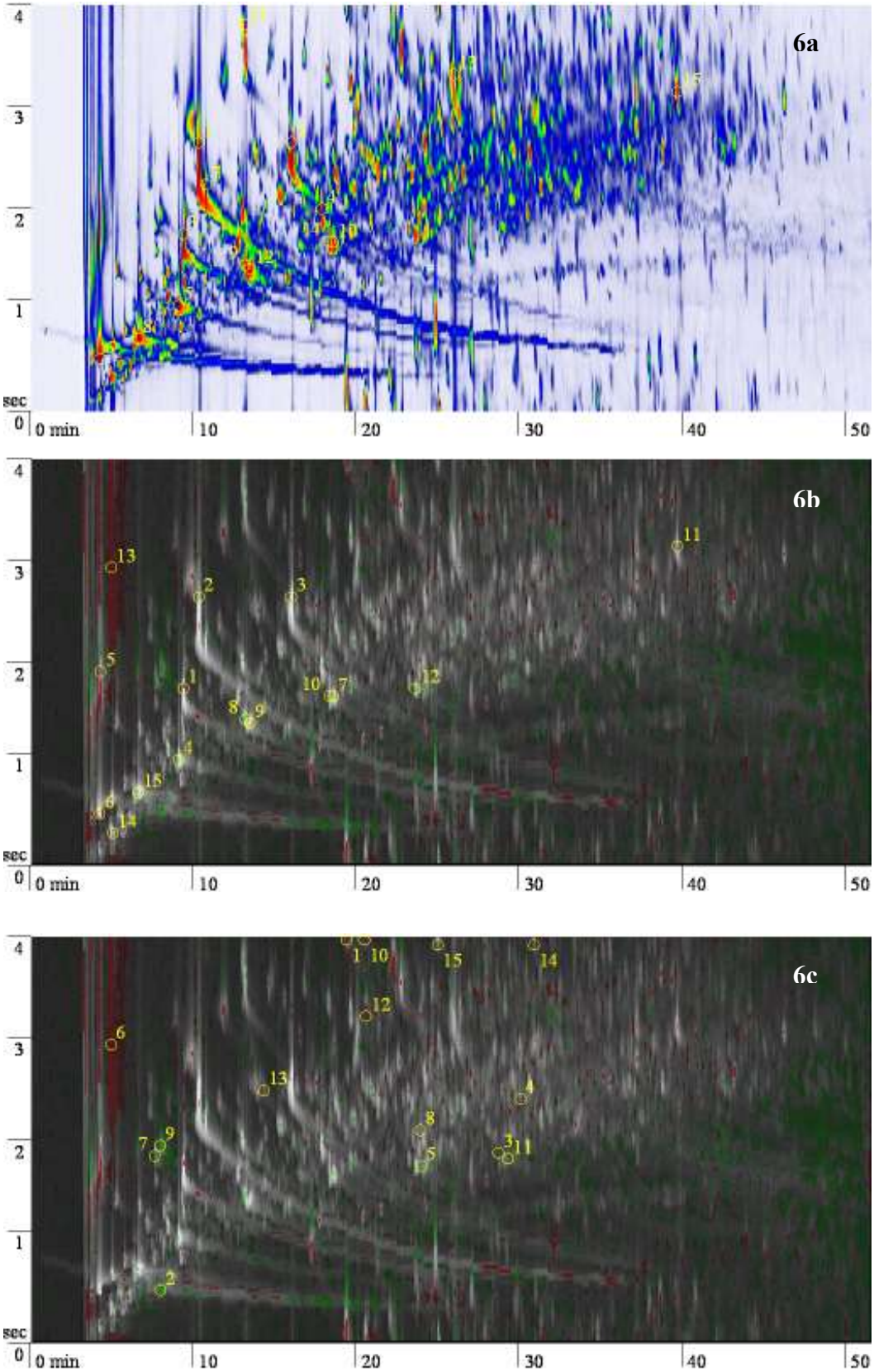
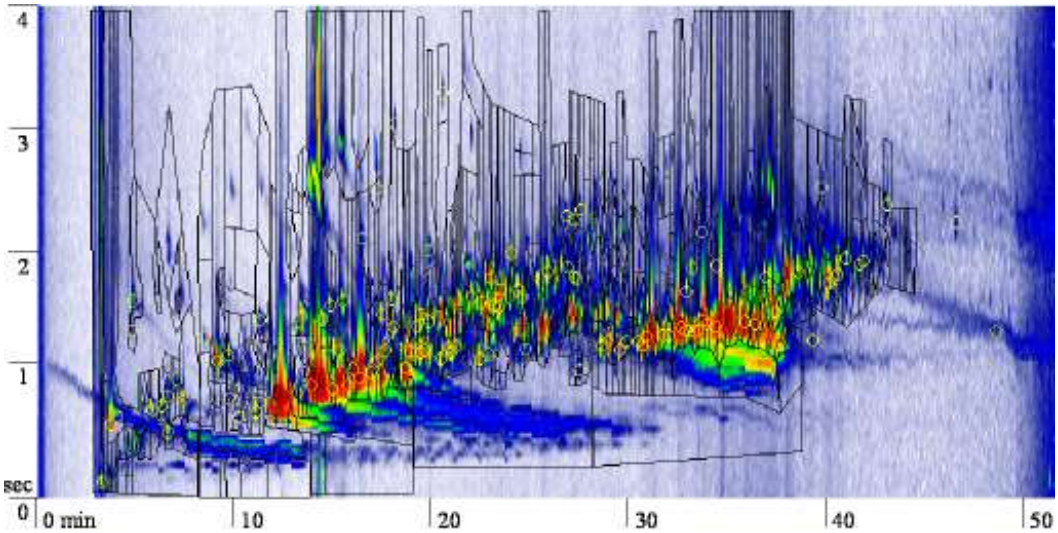



Figure 7: Cumulative chromatogram for a set of six Juniper samples with a template of reliable peaks indicated by yellow circles and mesh panels indicated by black outlines.