Metabolomics fingerprint of coffee species determined by untargeted-profiling study using LC-HRMS

Florence Souard, Cédric Delporte, Piet Stoffelen, Etienne Thévenot, Nausicaa Noret, Bastien Dauvergne, Jean-Michel Kauffmann, Pierre van Antwerpen, Caroline Stevigny

Metabolomics fingerprint of coffee species determined by untargeted-profiling study using LC-HRMS

Florence Souard[a, b, 1], Cédric Delporte[c, 1], Piet Stoffelen[d], Etienne A. Thévenot[e], Nausicaa Noret[f], Bastien Dauvergne[b, g], Jean-Michel Kauffmann[g], Pierre Van Antwerpen[c], Caroline Stévigny[b]

[1] Equal contribution of these authors

[a] Département de Pharmacochimie Moléculaire, UMR 5063 CNRS, Université Grenoble Alpes, 470 rue de la chimie, 38400 Saint-Martin d'Hères, France.

[b] Laboratoire de Pharmacognosie, de Bromatologie et de Nutrition Humaine, Faculté de Pharmacie, Université Libre de Bruxelles, Campus Plaine, CP 205/09, 1050 Brussels, Belgium

[c] Plateforme analytique de la Faculté de Pharmacie et Laboratoire de Chimie Pharmaceutique Organique, Faculté de Pharmacie, Université Libre de Bruxelles, Campus Plaine, CP 205/5, 1050 Brussels, Belgium

[d] Botanic Garden Meise, Domein van Bouchout, Nieuwelaan 38, 1860 Meise, Belgium

[e] CEA, LIST, Laboratory for Data Analysis and Systems' Intelligence, MetaboHUB Gif-sur-Yvette, France

[f] Laboratoire d'Écologie végétale et Biogéochimie, Université Libre de Bruxelles, Campus Plaine, CP 244, 1050 Brussels, Belgium.

[g] Laboratoire de chimie analytique et instrumentale et bioélectrochimie, Faculté de Pharmacie, Université Libre de Bruxelles, Campus Plaine, CP 205/09, 1050 Brussels, Belgium

Corresponding author: Florence.souard@univ-grenoble-alpes.fr

Abstract

Coffee bean extracts are consumed all over the world as beverage and there is a growing interest in coffee leaf extracts as food supplements. The wild diversity in Coffea (Rubiaceae) genus is large and could offer new opportunities and challenges. In the present work, a metabolomics approach was implemented to examine leaf chemical composition of 9 Coffea species grown in the same environmental conditions. Leaves were analyzed by LC-HRMS and a comprehensive statistical workflow was designed. It served for univariate hypothesis testing and multivariate modeling by PCA and partial PLS-DA on the Workflow4Metabolomics infrastructure. The first two axes of PCA and PLS-DA describes more than 40% of variances with good values of explained variances. This strategy permitted to investigate the metabolomics data and their relation with botanic and genetic informations. Finally, the identification of several key metabolites for the discrimination between species was further characterized.

1. Introduction

Coffee is appreciated worldwide as a beverage due to its aroma, flavor and stimulant properties. Beverage quality is highly related to the chemical compounds in coffee beans. A complex combination of these chemicals determines all beverages or foods sensory characteristics (Ivamoto et al., 2017). Monitoring food quality is crucial and has to be a major concern in order to maintain and improve the standard of life. Both the quality and the origin of food could be established by monitoring target molecules being used as markers. Many different analytical tools are available to quantify those markers after complete isolation from the matrix. This targeted strategy is often annoying, time consuming and can be foiled by ill-intentioned persons aware of this approach. To the contrary, untargeted methods such as fingerprinting by techniques such as liquid chromatography coupled to high resolution mass spectrometry (LC-HRMS) detection rely on a global picture, a metabolic composition, and can thus highlight incoherent concentration ratios or matrix perturbations resulting from any adulteration.

Coffee seeds (*Coffea semen*) are successfully used for beverages but also in cosmetic and pharmaceutical industries (due to their caffeine and high polyphenol content). Nowadays, the two most cultivated and studied *Coffea* species are Arabica (*C. arabica* L.) and Robusta (*C. canephora* Pierre ex Froehner). A wide range of methods have shown promising results for the detection of adulterated or contaminated coffee beans or modifications to environmental conditions and agricultural practices. Isotope Ratio Mass Spectrometry (IRMS) (Rodrigues et al., 2009), direct infusion electrospray (ESI)-(HR)MS (Electron Spray Ionization Mass Spectrometry) (Garrett et al., 2013), gas chromatography-mass spectrometry (GC–MS) (Jumhawan, Putri, Yusianto, Bamba, & Fukusaki, 2015), Raman Spectroscopy (El-Abassy, Donfack, & Materny, 2011), Near Infrared Spectroscopy (NIRS) - (Zhang, Wang, Liu, & He,

2016), Nuclear Magnetic Resonance (NMR) (Defernez et al., 2017) and LC-UV (Craig, Fields, Liang, Kitts, & Erickson, 2016).

Even though many records are available on coffee in the scientific literature for Arabica and Robusta, most studies are not considering other species. The latter, however, could be very important for cultivation and consumption. Crop wild relatives do have an important potential for breeding programs or directly as alternative crops. The wild diversity in *Coffea* is large and could offer new opportunities and challenges for phytochemical and medical studies as well.

The genus *Coffea*, in its traditional and narrow circumscription, does consist of 103 coffee species, with a natural distribution restricted to the tropical and subtropical Africa, Mascarenes and Madagascar (Davis, Govaerts, Bridson, & Stoffelen, 2006) or consist of 124 species, in its broader circumscription, as proposed by Davis [i.e. *Coffea* s.s. plus the former *Psilanthus* species, which has a wider paleotropical distribution (Davis, Tosh, Ruch, & Fay, 2011)]. Most members of the Coffee family (Rubiaceae) characteristic is x = 11 chromosomes. Species of *Coffea,* as well as its former sister genus *Psilanthus* [which is now lumped with the genus *Coffea* (Davis et al., 2011)] are diploids with 2n = 22 chromosomes (Patay, Bencsik, & Papp, 2016), except for *Coffea arabica* which is an allotetraploid species. The majority of coffee taxa are self-incompatible with exception of *C. arabica*, *C. anthonyi* Stoff. & F.Anthony (Stoffelen, Noirot, Couturon, & Anthony, 2008) and *C. heterocalyx* Stoff. (Stoffelen, Robbrecht, & Smets, 1996). Moreover, polyploid *C. arabica* individuals, namely triploid (3n = 33), pentaploid (5n = 55), hexaploid (6n = 66) and octoploid (8n = 88) plants, have also been described and occasionally, haploid or dihaploid young plants with narrower leaves also appeared (Clifford, 2012).

While abundant literature describes the phytochemistry of coffee beans (green or roasted), few studies have described the metabolic compositions of coffee leaves. However, leaf

phytochemistry is important because it contributes to a better understanding of the synthetic pathways and metabolite reallocation from leaves to seeds. In addition, it should be pointed out that coffee leaves are used for medical purposes or as beverage similarly as with tealeaves. Furthermore, in Africa, leaves of Robusta are used for bleeding linked with abortion (Neuwinger, 2000). Dried Arabica leaves are also still used for preparation of a tea named "jeno, jenuai" in Ethiopia for headache (Patay et al., 2016) or "copi daon" in Indonesia (Patay et al., 2016). In Liberia, the leaves' infusion of *C. arabica* was consumed only for its taste, as a drink. This drink was sold in the UK markets but with no success, perhaps it has not the usual taste of "British tea" (Patay et al., 2016). Coffee leaves are mentioned for headache and stomach pains (as a decoction) in Nicaragua, as cough suppressant (as an infusion) in Peru, as well as for fever and stimulation of prolactin's production in Mexico (Ross, 2005).

Taking into account these food and medicinal consumptions, and considering that for some strange reasons, few researchers are interested in this easy organ to work with (easy to mill…), we investigated the metabolome of mature leaves of eight *Coffea s.s.* species and one subspecies of *Coffea* s.s. and 1 former *Psilanthus* species (namely *C. mannii*) over one year using a LC-HRMS based metabolomics approach. All plants have been grown in tropical greenhouses of the Botanic Garden Meise (Belgium), which allowed controlling the influence of environmental factors. Data processing and statistical analysis of the metabolic profiles were performed on the Workflow4Metabolomics online infrastructure (W4M; Giacomoni et al., 2015)). The full history (tools, parameters, input and output data files) is publicly available on W4M (W4M00007_Coffee-leaves; DOI: 10.15454/1.4985472277740251E12).

2. *Materials and methods*

2.1. Chemicals and reagents

MS quality acetonitrile, formic acid (FA) and trifluoroacetic acid (TFA), caffeine (99%), theobromine (>99%) and theophylline (>99%) were purchased from Sigma-Aldrich (Steinheim, Germany). Ultra-high purity water was prepared by filtration using a Milli-Q system from Millipore (Bedford, MA, USA).

2.2. Coffee leaves samples

A total of 150 samples of leaves from 8 different *Coffea* species (*C. arabica ; C. anthonyi ; C. canephora ; C. charrieriana ; C. humilis ; C. kapakata ; C. mannii; C. liberica*) and one subspecies (*C. liberica var. liberica*) were collected in 2016 over a one-year period. All species grew around 10 years in the Botanic Garden Meise (Meise, Belgium). Leaves were collected between 10 am and 12 am on 5 days over 2016 (January, March, July, September and November). All plants were grown in the tropical greenhouses with the same environmental and edaphic conditions: natural daylight, substrate, watering regime, minimal temperature of 20 °C and relative humidity of the air.

The developmental stages of leaves were categorized as; (a) young leaves, (b) mature leaves, and (c) aged leaves. Young leaves were the most recently emerged and less than 1 cm long, mature leaves were fully developed, whereas aged leaves were dark green with often small brown necrosis on the leaf blade margins. Typically, they were the first, the second or third and the sixth leave on plagiotrophic branches, respectively (Ashihara & Crozier, 1999; Ashihara, Monteiro, Gillies, & Crozier, 1996). Only stage (b) mature leaves had been used for the present study. Unique accession codes of Botanic Garden Meise of trees collected are archived and

herbarium vouchers are deposited in the Herbarium of the Botanic Garden Meise (*C. arabica* 19073828*; C. anthonyi* 20070347-77*; C. canephora* 19800409 *; C. charrieriana* 20070349-79 *; C. humilis* 20110310-76 *; C. kapacata* 20110282-48*; C. liberica* 19391724 *; C. liberica var. liberica* 20110298-64*; Coffea mannii* 20091364-45 (Table 1).

2.3. Sample preparation

Samples were dried immediately after collecting by packing in sealed plastic bags filled with a large amount of silica gel. If necessary, the silica gel was replaced. Samples were dried during a minimum of 7 days. Dried leaves were ground using a mill and homogenized and extraction was subsequently carried out on batches of 15 mg of powdered leaves suspended in 1.5 mL of milliQ water for 5 min in a 55 kHz ultrasonic bath. Three sample extraction replicates were performed to check the repeatability of our procedure. Samples were filtered through a 0.2 µm cellulose acetate membrane and stored at -20 °C until analysis.

2.4. LC-HRMS analysis

Analyses were performed using a 1200 series rapid resolution LC (RRLC) system coupled to a 6520 series electrospray ionization (ESI)-quadrupole time-of-flight (QTOF) high-resolution mass spectrometer from Agilent Technologies (Waldbronn, Germany). Compound separation was performed using a Poroshell 120 EC-C18 column (2.7 µm, 100 mm × 2.1 mm) from Agilent. The column temperature was set at 55 °C. The mobile phases were composed of 0.025% of TFA and 0.075% of FA in water (solvent A) and in acetonitrile (solvent B). The applied gradient was as follows: 0 min, 0% B; 0–8 min, 0-10% B; 8–9 min, 10-12.5% B; 9–11 min, 12.5-15% B; 11–17 min, 15-80% B; 17–18 min, 80-100% B; 18–19 min, 100% B; 19–20 min, 100-0% B ; post-

run 8 min at 0.6 mL/min. ESI-QTOF parameters were as follows: positive mode, 2 GHz mode for resolution, mass range 100–1700 *m/z*, drying gas temperature and flow of 325 °C and 9 L/min respectively, nebulizer pressure 55 psi, and capillary voltage -4000 V. Nitrogen was used as the nebulizer gas. Data acquisition and LC-MS data analysis were carried out by MassHunter Acquisition® software for QTOF (Version B.04 SP3), MassHunter Qualitative Analysis® (Version B.06) software and MassHunter Quantitative Analysis® (Version B.04) software (Agilent Technologies). Batches were analyzed in random order. All samples were analyzed in one batch without any stopping or recalibration step. A same quality control (QC) sample (mix of all samples) was injected regularly throughout the run after every ten samples approximately. Finally, a target MS/MS approach was performed to get more details about metabolites of interest. *m/z* corresponding to these metabolites were isolated in the quadrupole (isolation width : -1 , +3 m/z) and fragmented in the collision cell with the collision energies: 0, 5, 10, 20, 30 and 40 V. Target MS/MS was performed in positive and negative modes on the corresponding $[M+H]^+$ or $[M-H]^-$ precursor ions. For negative mode, the same LC-MS/MS conditions were used except the polarity and the solvent. Solvent A was 20 mM ammonium formate pH 5.5 and solvent B ACN.

2.5. Data processing and statistical analysis

Agilent .d format data were converted to .mzXML format using the ProteoWizard MSConvert tools (Version 3.03.9393, 64-bit) with the Peak Picking filter option. Preprocessing of the data (automatic peak detection, integration, peak filtration, peak identification, peak grouping and smoothing, retention time correction, integration, annotation), normalization (batch correction), quality control (metabolites correlation analysis and determination of batch correction), and

statistical analyses (univariate testing and multivariate modeling) were conducted on the online and freely available Workflow4Metabolomics (W4M) platform; http://workflow4metabolomics.org). Detailed steps and parameters that were used for the different steps are shown in **Table S1** in supplementary data and are publicly available on the W4M workflow repository (W4M00007_Coffea-leaves; DOI: 10.15454/1.4985472277740251E12). Briefly, preprocessing was performed by using the implementation of the XCMS software (Smith et al., 2006) in W4M. The "*centWave"* algorithm (Tautenhahn, Böttcher, & Neumann, 2008) was used with the parameters adapted for an Agilent 6520 series LC-QTOF as defined in supplementary data (**Table S1**). Intensity drift correction was performed using a local quadratic (loess) model that represents the intensity variation along injection order using the QC sample (Dunn et al., 2011). Variables were then filtered to remove those with a mean intensity that was lower than twice the mean intensity in reagent blanks, or variables with a coefficient of variation in the QC samples above 30%. Finally, the intensities were $\log_{10}$ transformed.

Principal component analysis (PCA) was used for multivariate exploration of clusters and trends among the observations. Principal Components (PC) 1 to 4 (t1-t4) were selected as they capture 52% of the total variation (**Figure 2A** and for additional information, see Supporting Information). Differences of mean intensities between *Coffea* species were analyzed by multivariate analysis of variance (MANOVA) using the scores of the samples on the 4 PCs (followed by a Tukey post-hoc test for each PC; Statistica 7 software, Statsoft Inc). In parallel, univariate analysis of variance between *Coffea* species (ANOVA) was conducted with each of the original features (the False Discovery Rate threshold was set to 0.05, and the Tukey HSD post-hoc analysis was used). To determine whether the time of harvest influenced the metabolic

profiles, a Friedman's ANOVA (accounting for species effect) using the sample scores on the 4 PC (Statistica 7 software, Statsoft Inc) was realized. The loadings plot was used to identify variables accounting for the separation between the groups. Supervised partial least-squares-discriminant analysis (PLS-DA) models were also built (the significance of the Q^2Y prediction performance metric was assessed by comparison with 20 models built after random permutation of the response values). Hierarchical clustering of samples and variables (heatmap) was performed by using the 1-*cor* dissimilarity (where *cor* is the Spearman correlation) and the Ward's linkage method.

The variables that are significant for the classification performances between species (with either the PLS-DA, Random Forest or SVM approach) were selected with the Biosigner wrapper algorithm (Rinaudo, Boudah, Junot, & Thévenot, 2016). Pairwise comparisons of *Coffea* species for botanical, genomic or consumed interest were studied, including *C. arabica vs C. canephora* (ARA *vs* CAN)*; C. arabica vs C. anthonyi* (ARA *vs* ANTH)*; C. arabica vs C. mannii* (ARA *vs* PSI)*; C. arabica vs C. charrieriana* (ARA *vs* CHAR)*; C. mannii vs C. liberica* (PSI *vs* LIB)*; and C. liberica vs C. liberica var. liberica* (LIB *vs* LvL). Since Biosigner relies on an internal resampling approach, re-running of the module may result in slightly distinct signatures. Therefore, features (defined by their *m/z* and retention time values) that were present in at least two distinct Biosigner runs were selected. All statistical analyses were performed on the Workflow4Metabolomics infrastructure, unless otherwise specified.

2.6. Caffeine concentration

Caffeine concentration in the studied samples was determined by referring to a calibration curve drawn from 0.01 to 0.1 µg/mL using a caffeine standard. Mass Hunter Quantitative Analysis

software version B.04 (Agilent Technologies) was used for concentration determination in samples.

2.7. Characterization of the significant metabolite signatures for the discrimination between species

The metabolite signatures selected by statistical analysis were further characterized chemically by: determination of their fundamental composition with the MassHunter Qualitative Analysis Software (Agilent Technologies), and matching to the following databases: SciFinder (http://scifinder.cas.org), Kegg KEGG (http://www.kegg.jp/); Kanehisa and Goto, 2000) and ChemSpider (http://www.chemspider.com/). Furthermore, target MS/MS was performed when necessary to confirm metabolite (see above for MS parameters).

3. Results and discussion

3.1 Coffee plant

The two most studied and most widely cultivated coffee species, *C. arabica* and *C. canephora* (also named Robusta) were compared to some genetically close *Coffea* species with the aim of broadening the basis of the coffee economy. The recently published phylogeny of the genus *Coffea* (Hamon et al., 2017) is an interesting and useful framework to identify interesting taxa for this purpose. Africa is obviously the most interesting region to study genetic and phenotypic variations within the Coffee genus, as all the closely related species of the two principal crop species are native to this continent. For this model study, it was important to have access during one year to living collections of different coffee grown in the same environmental

conditions. In the Greenhouses of the Botanic Garden Meise, nine different *Coffea* taxa available were studied (**Table 1**).

Within this collection, we selected along *C. arabica* and *C. canephora,* species which are related to these two species namely: *C. liberica* a species with a wide central and West African distribution; *C. humilis*, a West African species (which is closely related to *C. liberica)*; *C. kapakata* a species from the wooded savannas of North-Western Angola and related to *C. liberica, C. canephora* and *C. anthonyi* a species (closely related to *C. eugenioides*, and therefore as well a potential ancestor species of *C. arabica*). We added two more distinctly related and ancestral Central African species to the sampling set in order to see more chemical variation within the sampling set namely: *C. charrieriana* a species from Cameroon with a position intermediary between *Coffea s.s.* and the former genus *Psilanthus* to which belonged *C. mannii*, a central African species, and from the former genus *Psilanthus,* both with caffeine free coffee bean. Two different accession of *C. liberica* were studied in order to see if there is intraspecific variation.

The leaf metabolomes of the coffee species were studied in a model context where leaves grown in a tropical greenhouse (Abdelsalam, Mahran, Chowdhury, Boroujerdi, & El-Bakry, 2017). All the plants were grown in the Botanical Garden Meise (Belgium, **Figure 1**), with the same environmental and edaphic conditions: daylight, substrate, watering daily regime, minimal temperature of 20 °C and relative humidity of the air. This permitted to infer that the influence of biotic and abiotic factors was the same and that the found differences were rather linked to genetic differences rather than to environmental ones. Mature leaves comprised the fully expanded second and third leaves from the apex (Ashihara & Crozier, 1999). The time period of the day for harvesting should also be considered since the level of some primary metabolites

vary throughout the daily cycle (De Vos et al., 2007). Samples should be prepared the same time period of the day for experiments that last several days, weeks or even months. For this reason, all samples were collected between 10 am and 12 am over the year 2016.

3.2 Coffee leaf metabolomics

As mentioned before, the literature is rich in coffee beans (green or roasted) phytochemical analyses but the composition of the coffee leaves is less described. Beverage quality of coffee bean extracts is of course highly related to the chemical compounds (Ivamoto et al., 2017). In addition to caffeine, other components of coffee beans like primary and secondary metabolites are important (Ivamoto et al., 2017). A complex combination of these chemicals determines beverage sensory characteristics.

In this context, metabolomics technologies have to be involved to examine the entire metabolome. Our interest was particularly devoted to untargeted metabolomics to examine the potential plasticity with phylogenetic evolution in the plant kingdom of *Coffea* species. Harvesting fresh plant material (**Figure 1**) is a crucial step in the analysis. Undesirable chemical or enzymatic reactions of metabolites can occur during harvesting and sample preparation (Kim, Choi, & Verpoorte, 2010). To avoid or reduce degradation of compounds a rapid drying of fresh leaves was undertaken. Normally, a rapid cooling of harvested samples is strongly recommended. In our case, we decided to use silica gel rather than cooling in liquid nitrogen. Indeed, this should make it possible to extend in the future some additional investigations of plants harvested in Africa in their biotope and pretreated as in this work. The use of silica gel is a praxis widely applied in order to preserve leaf samples for later DNA extraction for further

(phylo)-genetic studies. Another factor to be kept in mind is that leaves of different ages do have considerable differences in their metabolome (Kim et al., 2010). Particular attention has been given to collecting uniquely mature leaves of the same stage. After drying, a very simple extraction procedure has been undertaken. The extraction procedure is a crucial step for the detection of metabolites naturally occurring in the extracted tissues. Therefore, the extraction protocol should be simple enough to be reproducible and with high recovery and stability of most compounds, at least those of prime interest (De Vos et al., 2007). Moreover, our interest was to examine the metabolome in a sample commonly consumed (in water solution) and for this reason a water extract assisted by ultrasonication was performed. After sterile filtration, all extracts were stored frozen at -20 °C before analysis.

LC-HRMS using an LC-QTOF instrument is a common tool to obtain the metabolome fingerprint (**Figures 1** & **S4**) in pharmacognosy. Chromatographic separation prior to MS-analysis is particularly important in order to minimize ion suppression, maximize sensitivity as well as to separate isobaric and isomeric compounds. Reverse phase LC provides the most reliable and robust LC stationary phase for separation of the majority of the secondary metabolites at low concentration levels. As far as LC–MS interfaces are concerned, electrospray ionization (ESI) is the method of choice in most metabolomics applications (Millán et al., 2016). Sample analysis by LC-HRMS, was performed in one run list (one batch) and a single pooled sample was used as quality control (QC). The QC sample was processed as real samples to monitor the stability of the system. A random injection order was used to avoid confounding effects in case of signal drift during MS acquisition. The comprehensive data analysis workflow, including data preprocessing with XCMS (Smith et al., 2006), annotation with CAMERA (Kuhl, Tautenhahn, Böttcher, Larson, & Neumann, 2012), signal drift and batch effect correction,

univariate and multivariate statistical analyses, and feature selection with Biosigner (Rinaudo et al., 2016), were designed and performed on the Workflow4Metabolomics online platform, which provided a high-performance and user-friendly environment for computational analysis (Giacomoni et al., 2015). Data preprocessing resulted in the detection of 1,637 ion features. The QC pools were used for signal drift normalization (based on a loess type regression model) and quality control (QC coefficient of variation < 30%), as described in Dunn et al. (2011).

Multivariate analysis by PCA was first used to visualize groups, trends, and outliers among the observations (**Figures 2 A** & **S1**). The first 4 components capture 52% of the total variation. Nine clusters were detected and most of the taxa are well discriminated with the exception of LvL, KAP, and HUM that are clustered together (**Figures 2A** & **S1**). Most extraction replicates were clustered. Interestingly, clustering by collection period was observed within species clusters.

Supervised multivariate analysis was also performed using partial least-square analysis (PLS-DA; **Figure 2B**). The score plot from PLS-DA and the percentages of explained variation are similar to the PCA plots. These data indicate that it was possible to discriminate *Coffea* species and subspecies on the basis of the LC-HRMS metabolomics profiles. Furthermore, PCA and PLS-DA plots of the first predictive (t1) to the fourth predictive (t4) components are illustrated in **Figure S1** in supplementary data.

In the most recent phylogenetic study all the species studied here are positioned in the African subclade, except *C. mannii* (a former *Psilanthus* species), which is needed in Psilanthus clade, and sister to *Coffea* s.s. clade (all *Coffea* species except the species of the former genus *Psillanthus* and *C. Charrieriana*). *C. charrieriana* has an intermediary position between *Coffea* s.s. and the species of the former genus *Psilanthus* (incl *C. mannii*). Within the African subclade,

C. liberica (and *C. liberica* var. *liberica*) is closely but a little more distinctly related to *C. humilis*, these two species are closely related to *C. canephora* and *C. kapakata,* respectively. *C. anthonyi* does have a little more distinct position to *C. liberica, C. kapakata, C. canephora* and *C. humilis*. *C. arabica* is an allotetraploid species, the parents of *Coffea arabica* are closely related to *C. canephora* on the one hand and *C. anthonyi* on the other hand.

The significance of the separation between species along the principal components was assessed by MANOVA (Wilks test: F = 71.706; df: 32,116; p <0.001). For PC1, post-hoc tests distinguish several groups (**Figure 3**) with the two most different species being *C. mannii* and *C. liberica*. For PC2, two main groups were identified: one with *C. anthonyi, C. arabica* and C. *canephora*, and another including the 6 remaining taxa.

Metabolomics changes related to the harvesting period were shown with each of the four PCs by using Friedman's ANOVA (p value < 0.01). All species showed similar variations over time (Kendall's coefficient of concordance ranging from 0.51 to 0.60). Surprisingly, a decrease in the number of detected ions was observed in the November samples, ranging from 0.7% to 5.4% (ANTH 2.8%, ARA 5.4%, CAN 2.4%, CHAR 1%, HUM 0.7%, KAP 2.2%, LIB 3.5%, LvL 4%, PSI 2.1% - data not shown) as compared to the 1,637 features commonly present at all other time points. This reflected a lower metabolic diversity at this period of year. One explanation could be that an adaptation of the plant growing had occurred in the tropical greenhouse under natural photoperiod (Dunn, Bailey, & Johnson, 2005).

To further study the grouping of samples and variables, hierarchical clustering was performed (**Figure 4**). As expected, the two groups of species (KAP-LvL-LIB-HUM and CAN-PSI-CHAR-ANTH-ARA) previously evidenced on the PCA score plot, were also observed (sample clusters 5-8, and 1-4 and 9, respectively). Furthermore, groups of variables were shown

to have increased or decreased concentrations in specific species or couple of species. First, within the first group of species (KAP-LvL-LIB-HUM), several features were less concentrated in *C. humilis* (HUM; **Figure 4**, green box). In particular, the lower intensities of the metabolites from cluster 4 were closer to the second group of species (CAN-PSI-CHAR-ANTH-ARA). Second, within this second group of species, variables from cluster 12 had higher concentrations in *C. mannii* (PSI) and *C. charrieriana* (CHAR; **Figure 4**, blue box). This variable cluster was shown to contain several *ent*-kaurane diterpenoid derivatives. In addition, *C. anthonyi* (ANTH) and *C. arabica* (ARA) were shown to have increased concentrations of variables from cluster 11 (**Figure 4**; red box). In this cluster, the concentration of the *m/z* = 247.0598 feature (rt = 399 s) was higher for ANTH and ARA compared to all other species. The $C_{13}H_{10}O_5$ putative formula was determined for this ion. Further experiments are required to identify the structure of this metabolite. Finally, higher intensities of variables from cluster 5, which contains caffeine, were observed in *C. arabica* (ARA) compared to the other species from the second group (**Figure 4;** white box). Finally, it was observed that most of the samples harvested in November have lower intensities for the majority of the metabolites than the other harvested periods as shown in **Figure 4** where these columns have mostly blue-purple colors. This corroborates the MANOVA observations (**Figure 3**).

3.3 Identification of significant metabolites for species prediction

The objective was then to identify metabolites that significantly contribute to classification between the *Coffea* species. Based on the results of the multivariate analyses, we investigated the variables that influenced most the statistical models. Over all 1,637 variables analyzed, 92% (1,505) were significantly different over all taxa.

Features with significant value for the performances of classifiers between species of botanical or phylogenetic interest were selected with the Biosigner software tool (Rinaudo et al., 2016; **Table 2**). Biosigner performs recursive elimination of features, which do not significantly account for the prediction performances of binary classifiers (either PLS-DA, Random Forest, or Support Vector Machine). It has been reported that it is useful to select from omics data sets a (minimal) signature for predictive diagnosis.

When the two well-known *C. arabica* (ARA) and *C. canephora* (CAN) were compared, a feature with a *m/z* = 195.0870 that corresponds the $[M+H]^+$ of caffeine came out. The injection of caffeine standard confirmed the retention time. Caffeine concentration was found approximately 800 times higher in *C. arabica* leaves compared to *C. canephora* (see the discussion about caffeine below). Another feature observed with a *m/z* value of 247.0598 at retention time 399 s has much higher intensities in *C. arabica* than in *C. canephora* (Table 2). Identification of this feature will require further experiments.

Caffeine was also found to significantly account for the discrimination between *C. anthonyi* (ANTH) and *C. arabica* (ARA), in addition to another ion with a *m/z* value of 561.3617 (**Table 2**).

Interestingly, when *C. mannii* (PSI) and *C. liberica* (LIB) were compared, caffeine was not one of the most significant features. Those two species highly differed along PC 1 in the PCA analysis and the two main features that discriminate both species were *m/z* = 299.1997 and *m/z* = 868.4599 (**Table 2**). A single putative composition could be determined for the 868.4599 ion. The other feature, *m/z* 299.1997, has been identified as a possible derivative of a *ent*-kaurane diterpenoid derivative like a methyl-atractylgenin or the aglycone of 20-nor-cofaryloside I which

has been described before in coffee but in others organs (Chu et al., 2016; Kučera, Papoušek, Kurka, Barták, & Bednář, 2016). It is interesting to highlight that after a search of the 299.1997 *m/z* in the sample data, several peaks were observed and always in parallel to other *m/z* values that included one or two H_2O (317.2093, 335.2209, respectively) and one or two hexoses (497.2742 and 659.3270, respectively). It is common for these compounds to have a neutral loss in source as hexose and likewise water can be readily lost, and was confirmed by MS/MS analysis (**Figure S5**). It is therefore difficult to determine if the 299.1997 feature was an in-source-fragment or in the dehydrated-aglycone form of the diterpenoid derivatives (Garrett et al., 2013; Kučera et al., 2016).

For *C. liberica* (LIB) vs *C. mannii* (PSI) comparison, both features were selected: *m/z* = 335.2209 and *m/z* = 785.4229 (**Table 2**). The first one was certainly the di-hydrated form of the previously described ion with *m/z* = 299.1997, a diterpenoid.

Interestingly, the algorithm highlighted three major features when comparing *C. arabica* (ARA) vs *C. charrieriana* (CHAR) (**Table 2**). Those two species were compared because *C. charrieriana* is a somewhat enigmatic species as it is combining morphological characteristics of *Coffea* s.s. and the former genus *Psilanthus* (Stoffelen et al., 2008) and it is the most recent in the phylogeny (Hamon et al. 2017). It has an intermediate position between the *Coffea* clade and the *Psillanthus* clade and it has so an interest for future analysis and comparison with well-studied and developed species, i.e. *C. arabica*. The three selected features are derivatives of the described diterpenoids with *m/ z* 514.3005 being the ammonium adduct of the mono-hexose form, and the *m/z* 335.2209 and 299.1997 being the ones described above . These features were present in *C. charrieriana* but absent in *C. arabica.*

Finally, when *C. liberica* (LIB) and *C. liberica var liberica* (LvL) were compared, two features were selected but could not be identified (**Table 2**).

3.4 Untargeted metabolomics part summary

Metabolomics is a powerful tool to investigate the plant metabolome. In the current LC-HRMS study on mature leaves of various coffee species, differences were clearly detected. Metabolomics analyses applied to simple water extracts have proved to be particularly efficient on consumed beverage. They can provide useful information (about botanical origin, sampling period, etc …). Metabolomics analyses have highlighted important markers of botanical species even in the case of LIB, LvL, HUM and KAP, were 95% confidence ellipse were partially superposed.

3.5 Study of key metabolite in *Coffea* leaves: caffeine, theobromine and theophylline and *ent*-kaurane diterpenoids.

Regarding the endogenous caffeine, we focus on its concentration in mature leaves. First, we searched for the caffeine ion in the results of the univariate analysis. In mature leaves of *C. arabica*, the concentration of caffeine was clearly higher than in the others species (**Figure S2 A**). Surprisingly, the quantitative approach of caffeine determination allowed us to show that caffeine was present only in mature leaves of *C. arabica* and not in the other species (**Figure S2 B**). The values observed in the univariate analyses for other species than *C. arabica* were close to the background level. Indeed, no caffeine was detected and if present, the concentration in the extracts was below the lowest point of the curve, namely below 0.01 µg/mL.

It is well described that caffeine biosynthesis occurs in both fruits and leaves of coffee. In fact, caffeine had been described in at least 80 species in 13 orders within the plant kingdom (Ashihara, Mizuno, Yokota, & Crozier, 2017). In general in *Coffea* species, caffeine concentration is often described in beans and it is known that distribution is mainly in leaves and cotyledons of coffee seedlings with small amounts in stems and roots in particular in *C. arabica*. Caffeine biosynthesis is especially active in young leaves of *C. arabica* and declines with the leaf age. Caffeine accumulates in *C. arabica* due to extremely slow catabolism to theophylline (1,7-dimethylxanthine) (Ashihara & Crozier, 1999). Various hypotheses have been proposed to explain the role of the high concentration of caffeine in coffee and a few other plant species.

The « chemical defense theory » proposes that caffeine and other methylxanthines particularly in young leaves are able to act as a pesticide against herbivores such as insects or gastropods. Caffeine concentration is even increased after insect herbivore attack in *C. arabica*. Another hypothesis is the « allelopathic theory ». The purpose is that caffeine can contribute as allelopathic and auto-toxic compound in old coffee plantations. Caffeine would be released in vicinal soil over the years and accumulated (in monoculture) an explanation of a low productivity over the long-term. A last hypothesis has been proposed to clarify why caffeine is present in flowers. Caffeine would have the ability to encourage efficient foraging thank to a learned olfactory capacity of honeybee (Wright et al., 2013) with behavioral consequences described by Couvillon et al. (2015).

Young expanding leaves of *C. arabica* plants are known to contain caffeine and traces of theobromine and interactions with polyphenols are well described (Ashihara et al., 2017). More exactly, a complex made of purine alkaloids and chlorogenic acids in the vacuoles of coffee leaves has been suggested.

Curiously, the other endogenous purine alkaloids detected in any of the leaf extracts was theobromine/theophylline, but no correlation of their concentration has been put forward with caffeine concentration in *C. arabica.*

According to the univariate results regarding the mix of theophylline and/or theobromine (**Figure S2**), there are only 2 species where the signal emerges from the noise, namely *C. canephora* (CAN) and *C. humilis* (HUM). After checking with standards under our condition of analysis, the peak detected in these species corresponded to theobromine and no theophylline was detected. No clear explanation has been highlighted but metabolic changes might be involved. Caffeine is ~65 more soluble in water than theobromine (they have the same pK_a) and theophylline is generally intermediate (in function of pH). Our extraction method was not optimized to extract methylxanthine alkaloids and possibly all purines were not extracted; it seemed however curious that theobromine was more extracted in mature leaves if detectable quantities are present in tissue. Concerning xanthine alkaloids biosynthesis, experiments have revealed that the incorporation of [8-^{14}C]adenine into theobromine and caffeine was found in small, young leaves of *C. arabica* but it disappeared in fully developed leaves (Ashihara et al., 2017), which is in agreement with our observations.

Finally, the last class of key metabolites, which caught our attention, concerned *ent*-kaurane diterpenoids that were described as key metabolites to discriminate the species (see section 3.3 above). Coffee beans are known for containing nearly 90 diterpenoids (Chu et al., 2016). An overview of the literature on the possible structures led us to assign the discriminant metabolites with m/z 299.1997, 335.2209, 514.3005 to (*O*-heteroside-)diterpenoid derivatives with an *ent*-kaurane skeleton. Furthermore, this class of compounds includes atractyl-, nor-cofaryl-, steviol-, kaurenolide-, stevane-, diketoatractyli-genine, atractylitriol, hydroatractylitriol derivates, *etc.*

which exhibit various biological activities(Chu et al., 2016; Lang et al., 2013). Several of these compounds are known to have therapeutic or toxic effects. Toxicity of sulfoglucosides of the norditerpenoid atractyligenin from *Atractylis gummifera* (L.) is well known (Daniele et al., 2005) by inhibition of mitochondrialoxidative phosphorylation through ANT blockage. Nevertheless the activity of a carboxy-derivates presented in raw coffee seeds was about three times lower than that of the well-known toxic atractyloside from *Atractylis gummifera* (Lang et al., 2013). However, as these derivatives could also be present in leaves of several species of *Coffea*, the use of leave infusions as a "tea drink" might be toxic. This implies the necessity of analytically controlling the levels of *ent*-kaurane in raw leaf coffee extracts when used as food products. The full characterization of this family of derivatives should be further investigated to determine the exact compounds and if for example, they are atractyloside derivatives that are mainly known as possible toxic compounds. Further fragmentation analyses (**Figures S5** & **S6**) by tandem mass spectrometry (MS/MS) have not enabled us to discriminate a particular atractyloside or other *ent*-kaurane (like20-nor-cofaryloside I) derivatives as the fragments and the neutral losses are highly similar. The fragmentation pattern includes the loss of H_2O (−18 Da) from a hydroxyl group, of acetate (−60 Da) or CH2=CO (−42 Da) from an acetate group, of CO (−28 Da) from the ring and CH_2O (−30 Da) (**Figures S5** & **S6**). It is interesting to highlight that *C. arabica*, *C. canephora* and *C. anthonyi* species that we analyzed do not contain these derivatives. Nevertheless, these compounds have to be monitored either in leaves and in (roasted-) beans of coffee before further large human consumption as already mentioned (Chu et al., 2016; Lang et al., 2013).

Conclusions

Previous metabolomics studies of *Coffea* have mostly focused on green and roasted coffee beans of the two major species, namely *C. arabica* and *C. canephora*. Furthermore, only few studies have been performed on leaves although coffee leaves are also used as either infusion, like a tea, or for medicinal purposes. We so undertook metabolomics investigations on leaves aqueous extracts of studied species of *Coffea* that grew in a greenhouse where no environmental aspect (less noise) might cause interspecific differences in the metabolomes.

The goal of this metabolic fingerprinting study was to determine the species differences between the metabolomes. PCA and PLS-DA of approximately 150 samples reflected the variation in the data. All nine clusters of each species studied were observed on both PCA and PLS-DA score plots, with good discrimination between the eight *Coffea* species and one subspecies: species that are known to be genetically close. PCA suggested that *C. arabica*, *C. canephora* and *C. anthonyi* have similar metabolomics profiles in our analytical conditions. However, several interesting results arose when specific metabolites were analyzed. Indeed, caffeine was only detected in leaf aqueous extracts of *C. arabica*. This was surprising (Perrois et al., 2015) and if several hypotheses could be developed, more investigations should be undertaken in the future to understand the absence of caffeine in the other species like genetics/genomics investigation. Another important observation was the detection of *ent*-kaurane diterpenoids (C20) derivatives in several species. The latter might be toxic (Stewart & Steenkamp, 2000) or medicinal (Chu et al., 2016) and should be monitored in any *Coffea* leaf extracts that would be used for human consumption. Furthermore, seasonal effects were observed with changes in the metabolomes over the collection times in 2016. As perspective of this work, it has been planned to study other coffee leaf extracts, and other wild plants with African origin to increase metabolomics knowledge on *Coffea* species.

Acknowledgments

FS is particularly thankful to UGA (Grenoble) for the scientific delegation conceded at Faculty of Pharmacy ULB (Brussels) and in particular C. Ribuot and E. Peyrin. FS thanks ULB analytical platform for facilities employing HRMS and good scientific exchanges. FS thanks FNRS subvention (scientific mission). FS thanks CS for her welcome. FS and CS thank M. Faes & O. Vaillant for their skillful assistance.

This study was supported by the Belgian National Fund for Scientific Research (FRS, N° 3.4553.08 and T.0136.13 PDR), the Université Libre de Bruxelles (FER-207). CD is a postdoctoral researcher founded by the Belgian National Fund for Scientific Research.

The IRD (Institut pour la Recherche et le Developpement, France) is acknowledged as they donated several of the studied living plants to the Botanic Garden Meise. The staff of the Botanic Garden Meise is acknowledge for granting access to the collections.

The Workflow4Metabolomics infrastructure is supported by the French Institute of Bioinformatics (IFB; ANR-11-INBS-0013) and by the French Infrastructure for Metabolomics and Fluxomics (MetaboHUB; ANR-11-INBS-0010).

References

Abdelsalam, A., Mahran, E., Chowdhury, K., Boroujerdi, A., & El-Bakry, A. (2017). NMR-based metabolomic analysis of wild, greenhouse, and in vitro regenerated shoots of Cymbopogon schoenanthus subsp. proximus with GC–MS assessment of proximadiol. *Physiology and Molecular Biology of Plants*, *23*(2), 369–383. https://doi.org/10.1007/s12298-017-0432-0

Ashihara, H., & Crozier, A. (1999). Biosynthesis and catabolism of caffeine in low-caffeine-containing species of Coffea. *Journal of Agricultural and Food Chemistry*, *47*(8), 3425–3431.

Ashihara, H., & Crozier, A. (2001). Caffeine: a well known but little mentioned compound in plant science. *Trends in Plant Science*, *6*(9), 407–413. https://doi.org/10.1016/S1360-1385(01)02055-6

Ashihara, H., Mizuno, K., Yokota, T., & Crozier, A. (2017). Xanthine Alkaloids: Occurrence, Biosynthesis, and Function in Plants. *Progress in the Chemistry of Organic Natural Products*, *105*, 1–88. https://doi.org/10.1007/978-3-319-49712-9_1

Ashihara, H., Monteiro, A. M., Gillies, F. M., & Crozier, A. (1996). Biosynthesis of Caffeine in Leaves of Coffee. *Plant Physiology*, *111*(3), 747–753.

Chu, R., Wan, L.-S., Peng, X.-R., Yu, M.-Y., Zhang, Z.-R., Zhou, L., … Qiu, M.-H. (2016). Characterization of New Ent-kaurane Diterpenoids of Yunnan Arabica Coffee Beans. *Natural Products and Bioprospecting*, *6*(4), 217–223. https://doi.org/10.1007/s13659-016-0099-1

Clifford, M. N. (2012). *Coffee: Botany, Biochemistry and Production of Beans and Beverage*. Springer Science & Business Media.

Couvillon, M. J., Al Toufailia, H., Butterfield, T. M., Schrell, F., Ratnieks, F. L. W., & Schürch, R. (2015). Caffeinated Forage Tricks Honeybees into Increasing Foraging and Recruitment Behaviors. *Current Biology*, *25*(21), 2815–2818. https://doi.org/10.1016/j.cub.2015.08.052

Craig, A. P., Fields, C., Liang, N., Kitts, D., & Erickson, A. (2016). Performance review of a fast HPLC-UV method for the quantification of chlorogenic acids in green coffee bean extracts. *Talanta*, *154*, 481–485. https://doi.org/10.1016/j.talanta.2016.03.101

Daniele, C., Dahamna, S., Firuzi, O., Sekfali, N., Saso, L., & Mazzanti, G. (2005). Atractylis gummifera L. poisoning: an ethnopharmacological review. *Journal of Ethnopharmacology*, *97*(2), 175–181. https://doi.org/10.1016/j.jep.2004.11.025

Davis, A. P., Govaerts, R., Bridson, D. M., & Stoffelen, P. (2006). An annotated taxonomic conspectus of the genus Coffea (Rubiaceae). *Botanical Journal of the Linnean Society*, *152*(4), 465–512. https://doi.org/10.1111/j.1095-8339.2006.00584.x

Davis, A. P., Tosh, J., Ruch, N., & Fay, M. F. (2011). Growing coffee: Psilanthus (Rubiaceae) subsumed on the basis of molecular and morphological data; implications for the size, morphology, distribution and evolutionary history of Coffea. *Botanical Journal of the Linnean Society*, *167*(4), 357–377. https://doi.org/10.1111/j.1095-8339.2011.01177.x

De Vos, R. C., Moco, S., Lommen, A., Keurentjes, J. J., Bino, R. J., & Hall, R. D. (2007). Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *2*(4), 778–791. https://doi.org/10.1038/nprot.2007.95

Defernez, M., Wren, E., Watson, A. D., Gunning, Y., Colquhoun, I. J., Le Gall, G., … Kemsley, E. K. (2017). Low-field 1H NMR spectroscopy for distinguishing between arabica and robusta ground roast coffees. *Food Chemistry*, *216*, 106–113. https://doi.org/10.1016/j.foodchem.2016.08.028

Dunn, W. B., Bailey, N. J. C., & Johnson, H. E. (2005). Measuring the metabolome: current analytical technologies. *Analyst*, *130*(5), 606–625. https://doi.org/10.1039/B418288J

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., … Consortium, T. H. S. M. (HUSERMET). (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, *6*(7), 1060–1083. https://doi.org/10.1038/nprot.2011.335

El-Abassy, R. M., Donfack, P., & Materny, A. (2011). Discrimination between Arabica and Robusta green coffee using visible micro Raman spectroscopy and chemometric analysis. *Food Chemistry*, *126*(3), 1443–1448. https://doi.org/10.1016/j.foodchem.2010.11.132

Garrett, R., Schmidt, E. M., Pereira, L. F. P., Kitzberger, C. S. G., Scholz, M. B. S., Eberlin, M. N., & Rezende, C. M. (2013). Discrimination of arabica coffee cultivars by electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry and chemometrics. *LWT - Food Science and Technology*, *50*(2), 496–502. https://doi.org/10.1016/j.lwt.2012.08.016

Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., … Caron, C. (2015). Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. *Bioinformatics*, *31*(9), 1493–1495. https://doi.org/10.1093/bioinformatics/btu813

Hamon, P., Grover, C. E., Davis, A. P., Rakotomalala, J.-J., Raharimalala, N. E., Albert, V. A., … Guyot, R. (2017). Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content. *Molecular Phylogenetics and Evolution*, *109*, 351–361. https://doi.org/10.1016/j.ympev.2017.02.009

Ivamoto, S. T., Sakuray, L. M., Ferreira, L. P., Kitzberger, C. S. G., Scholz, M. B. S., Pot, D., … Pereira, L. F. P. (2017). Diterpenes biochemical profile and transcriptional analysis of cytochrome P450s genes in leaves, roots, flowers, and during Coffea arabica L. fruit development. *Plant Physiology and Biochemistry*, *111*, 340–347. https://doi.org/10.1016/j.plaphy.2016.12.004

Jumhawan, U., Putri, S. P., Yusianto, null, Bamba, T., & Fukusaki, E. (2015). Application of gas chromatography/flame ionization detector-based metabolite fingerprinting for authentication of Asian palm civet coffee (Kopi Luwak). *Journal of Bioscience and Bioengineering*, *120*(5), 555–561. https://doi.org/10.1016/j.jbiosc.2015.03.005

Kim, H. K., Choi, Y. H., & Verpoorte, R. (2010). NMR-based metabolomic analysis of plants. *Nature Protocols*, *5*(3), 536–549. https://doi.org/10.1038/nprot.2009.237

Kučera, L., Papoušek, R., Kurka, O., Barták, P., & Bednář, P. (2016). Study of composition of espresso coffee prepared from various roast degrees of Coffea arabica L. coffee beans. *Food Chemistry*, *199*, 727–735. https://doi.org/10.1016/j.foodchem.2015.12.080

Kuhl, C., Tautenhahn, R., Böttcher, C., Larson, T. R., & Neumann, S. (2012). CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical Chemistry*, *84*(1), 283–289. https://doi.org/10.1021/ac202450g

Lang, R., Fromme, T., Beusch, A., Wahl, A., Klingenspor, M., & Hofmann, T. (2013). 2-O-β-d-Glucopyranosyl-carboxyatractyligenin from Coffea L. inhibits adenine nucleotide translocase in isolated mitochondria but is quantitatively degraded during coffee roasting. *Phytochemistry*, *93*, 124–135. https://doi.org/10.1016/j.phytochem.2013.03.022

Millán, L., Sampedro, M. C., Sánchez, A., Delporte, C., Van Antwerpen, P., Goicolea, M. A., & Barrio, R. J. (2016). Liquid chromatography-quadrupole time of flight tandem mass spectrometry-based targeted metabolomic study for varietal discrimination of grapes according to plant sterols content. *Journal of Chromatography. A*, *1454*, 67–77. https://doi.org/10.1016/j.chroma.2016.05.081

Neuwinger, H. D. (2000). *African traditional medicine: A dictionary of plant use and applications with supplement : search system for diseases*. Stuttgart: Medpharm Scientific Publishers.

Patay, É. B., Bencsik, T., & Papp, N. (2016). Phytochemical overview and medicinal importance of Coffea species from the past until now. *Asian Pacific Journal of Tropical Medicine*, *9*(12), 1127–1135. https://doi.org/10.1016/j.apjtm.2016.11.008

Perrois, C., Strickler, S. R., Mathieu, G., Lepelley, M., Bedon, L., Michaux, S., … Privat, I. (2015). Differential regulation of caffeine metabolism in Coffeaarabica (Arabica) and Coffea canephora (Robusta). *Planta*, *241*(1), 179–191. https://doi.org/10.1007/s00425-014-2170-7

Rinaudo, P., Boudah, S., Junot, C., & Thévenot, E. A. (2016). biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences*, *3*, 26. https://doi.org/10.3389/fmolb.2016.00026

Rodrigues, C. I., Maia, R., Miranda, M., Ribeirinho, M., Nogueira, J. M. F., & Máguas, C. (2009). Stable isotope analysis for green coffee bean: A possible method for geographic origin discrimination. *Journal of Food Composition and Analysis*, *22*(5), 463–471. https://doi.org/10.1016/j.jfca.2008.06.010

Ross, I. A. (2005). *Medicinal Plants of the World, Volume 3: Chemical Constituents, Traditional and Modern Medicinal Uses* (2005 edition). Totowa, NJ: Humana Press.

Stewart, M. J., & Steenkamp, V. (2000). The biochemistry and toxicity of atractyloside: a review. *Therapeutic Drug Monitoring*, *22*(6), 641–649.

Stoffelen, P., Noirot, M., Couturon, E., & Anthony, F. (2008). A new caffeine-free coffee from Cameroon. *Botanical Journal of the Linnean Society*, *158*(1), 67–72. https://doi.org/10.1111/j.1095-8339.2008.00845.x

Stoffelen, P., Robbrecht, E., & Smets, E. (1996). Coffea (Rubiaceae) in Cameroon: A new species and a nomen recognized as species. *Belgian Journal of Botany*, *129*(1), 71–76.

Tautenhahn, R., Böttcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics*, *9*, 504. https://doi.org/10.1186/1471-2105-9-504

Wright, G. A., Baker, D. D., Palmer, M. J., Stabler, D., Mustard, J. A., Power, E. F., … Stevenson, P. C. (2013). Caffeine in Floral Nectar Enhances a Pollinator's Memory of Reward. *Science*, *339*(6124), 1202–1204. https://doi.org/10.1126/science.1228806

Zhang, C., Wang, C., Liu, F., & He, Y. (2016). Mid-Infrared Spectroscopy for Coffee Variety Identification: Comparison of Pattern Recognition Methods. *Journal of Spectroscopy*, *2016*, 1–7. https://doi.org/10.1155/2016/7927286

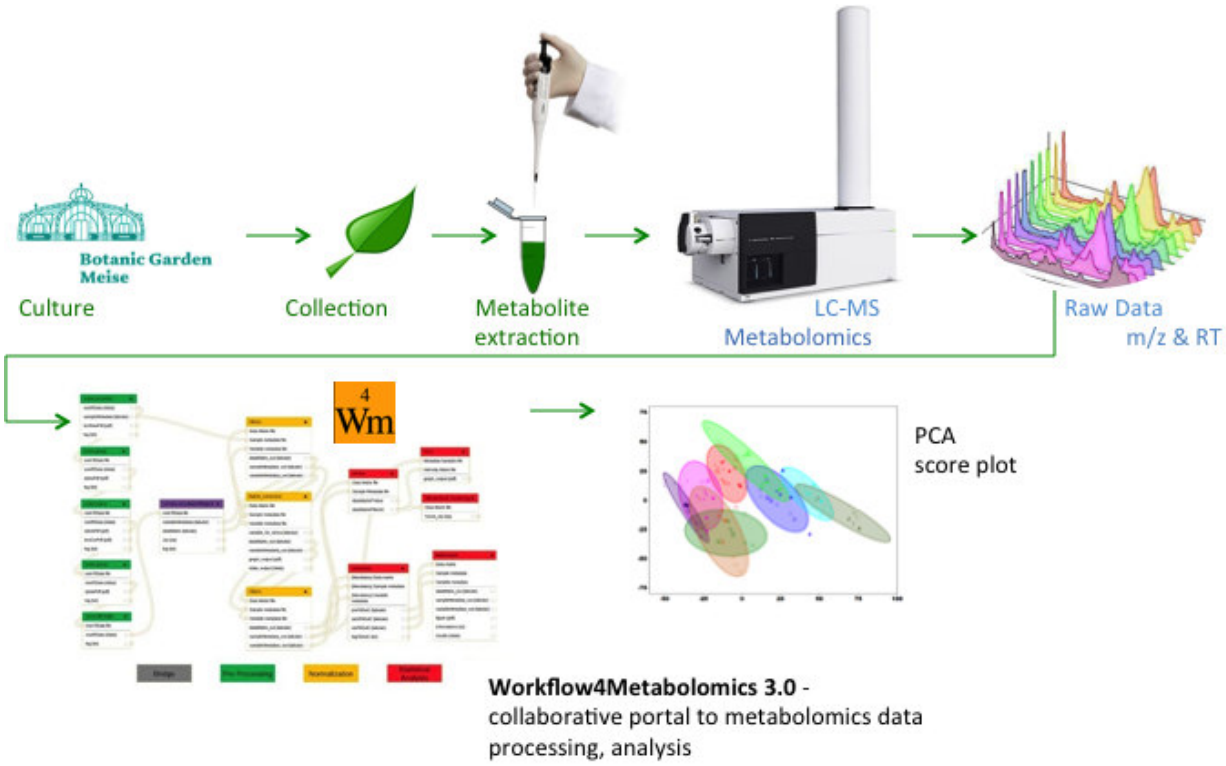

Figure 1: Overview of the applied metabolomics workflow. After cultivation of coffee trees in tropical greenhouses, leaves were collected and the metabolites were extracted. An LC-HRMS analysis was performed and raw data were collected before data analysis (preprocessing, statistics, and annotation) on the Workflow4Metabolomics online platform.

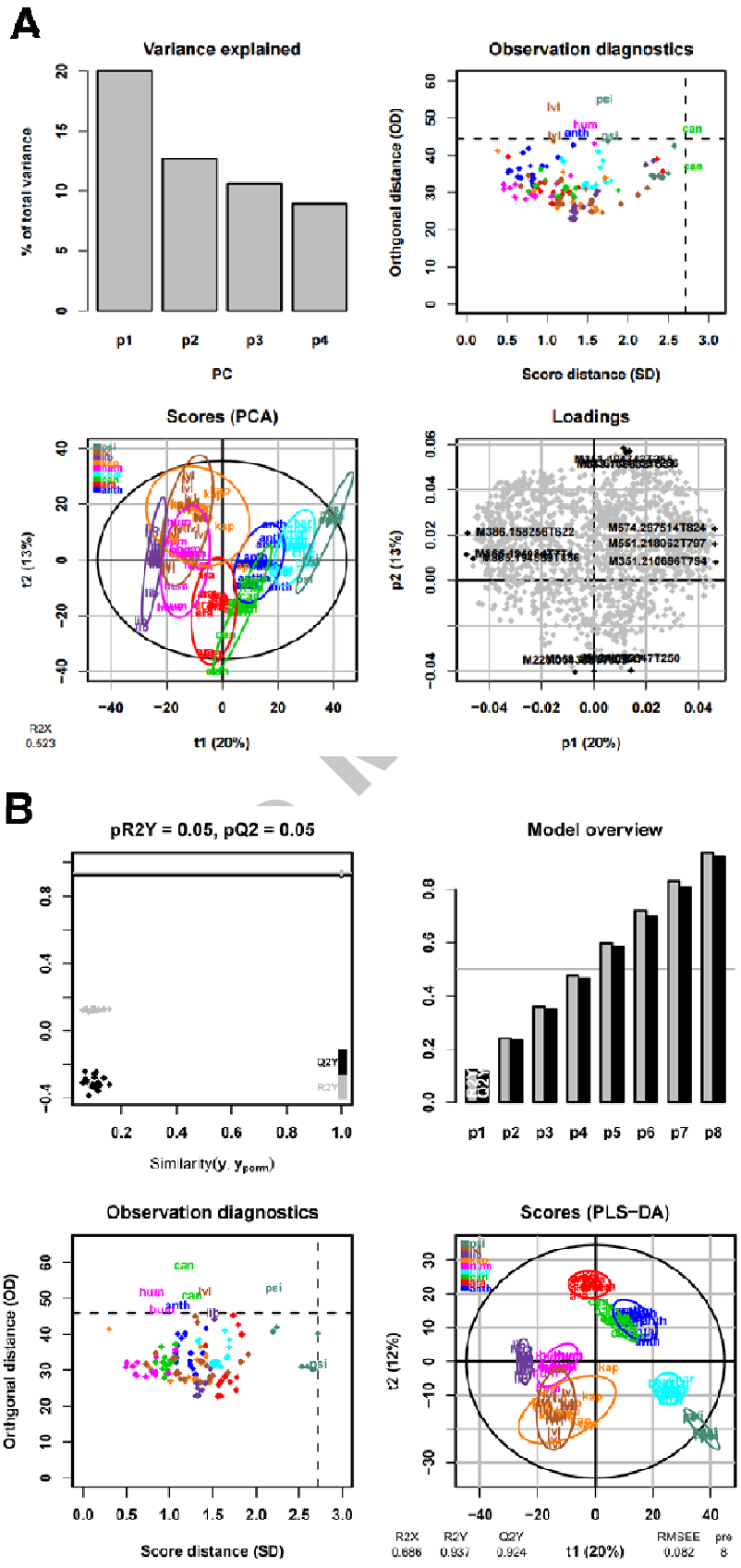

Figure 2: Multivariate modeling of variations using (A) PCA or (B) PLS-DA of the 8 *Coffea* species and one sub-species of *C. liberica*. Score plots of the first predictive (t1) and the second predictive (t2) components are illustrated. On the score plots, the percentage of total variation explained by the component is indicated in parentheses. The black (respectively colored) ellipses include 95% of the multivariate normal distribution of all (respectively the specific groups of) samples. On the loading plots, the names of the 6 variables with most extreme values in each direction, is indicated. The observation diagnostic plot shows the distances within and orthogonal to the selected score plane (Engelen, Hubert, & Branden, 2016). For PLS modeling, an additional diagnostic plot (top left) shows the Q^2Y (and R^2Y) values from the model (horizontal lines) compared to the values from the models obtained after random permutations of the y response (dots).

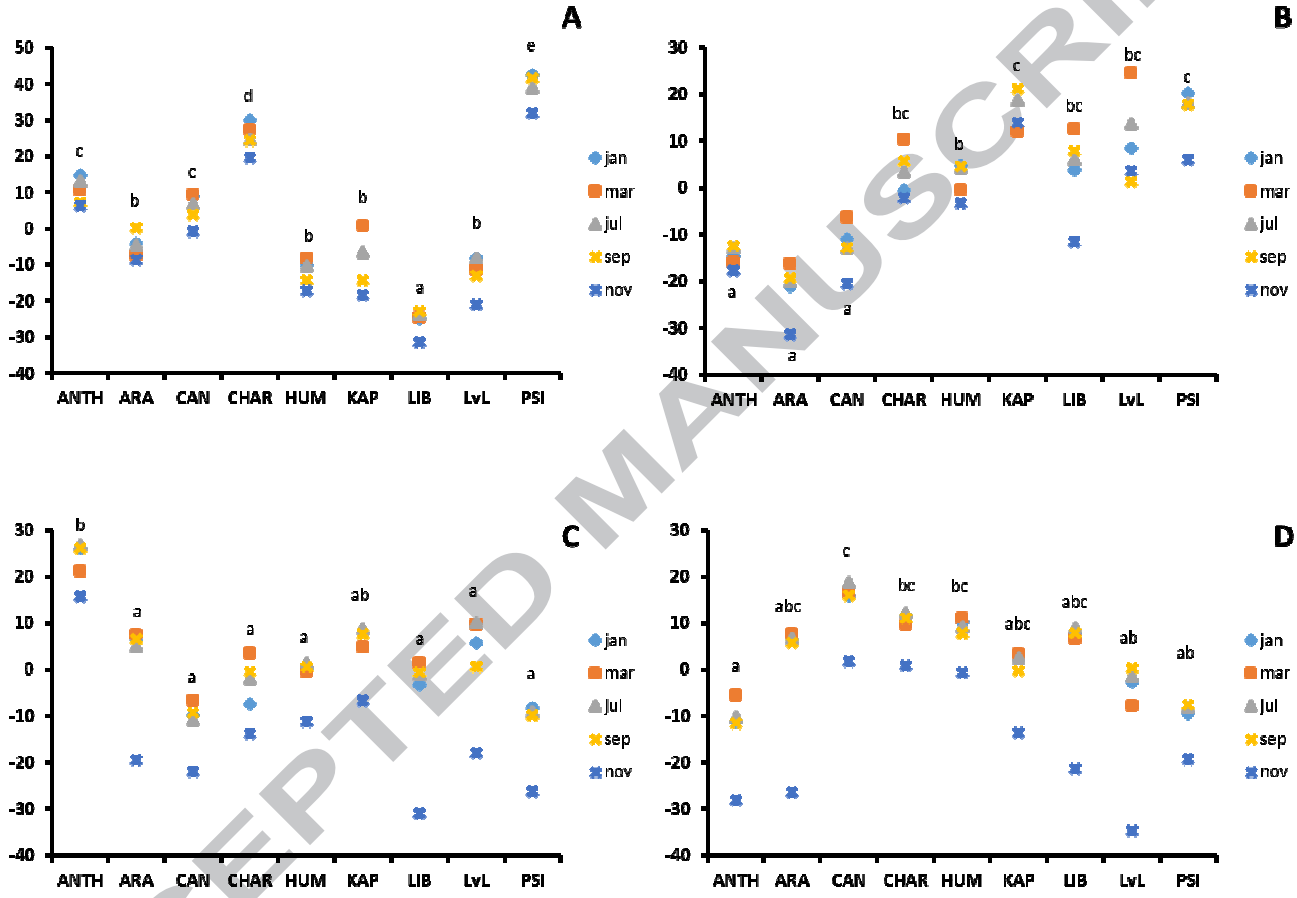

Figure 3: PCA scores (y-axis) of the samples of each *Coffea* species (x-axis) for each of the 4 first PCA axes (A, PC1; B, PC2; C, PC3; D, PC4). For each of the 4 graphs, results of Tukey post-hoc tests are shown as letters above the symbols; different letters for two compared species means that these two have significantly different metabolic profiles regarding the PC axis considered. For each species, PCA scores for each harvest date is represented, showing that samples collected in November are often different from the others. Species abbreviations: ANTH, *C. anthonyi*; ARA, *C. arabica*; CAN, *C. canephora*; CHAR, *C. charrieriana*; HUM, *C. humilis*; KAP, *C. kapakata*; PSI: *C. manii*; LIB, *C. liberica*; LvL, *C. liberica* var. *liberica.*

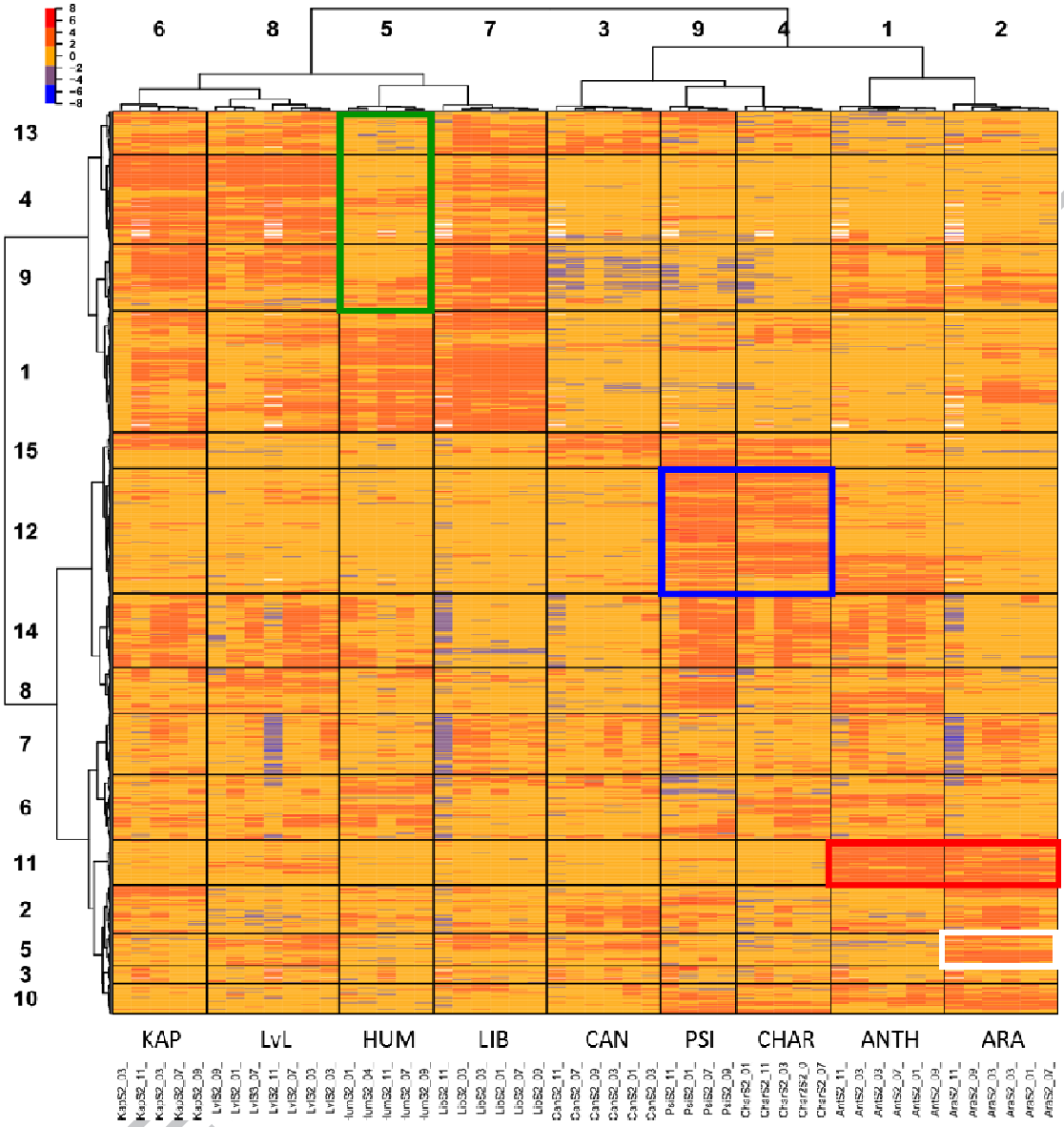

Figure 4: Heatmap visualization constructed based on the differential metabolites of importance for *Coffea* species (on one replicate per species and harvested times). Heatmap represents unsupervised hierarchical clustering of samples (columns) and variables (rows). To increase the contrast, intensities of each variable have been scaled to unit variance. The lower intensities of the samples harvested in November compared to other periods can be observed by blue-purple colors in the corresponding columns of the heatmap. Color boxes highlight clusters of metabolites which specifically characterize the species (or couple of species) within the two main groups of species (KAP-LvL-LIB-HUM and CAN-PSI-CHAR-ANTH-ARA; see the text for the details).

Table1: Set of coffee leaves samples

Name	Code	Reference	Origin	Months of collection
C. anthonyi	ANTH	20070347-77	Cameroon	01-03-06-09-11
C. arabica	ARA	19073828	Ethiopia	01-03-06-09-11
C. canephora	CAN	19800409	Central Africa	01-03-06-09-11
C. charrieriana	CHAR	20070349-79	Cameroon	01-03-06-09-11
C. humilis	HUM	20110310-76	Ivory Coast	01-03-06-09-11
C. kapakata	KAP	20110282-48	North-Western Angola	03-06-09-11
C. liberica	LIB	19391724	Central Africa	01-03-06-09-11
C. liberica var liberica	LvL	20110298-64	Ivory Coast	01-03-06-09-11
C. mannii	PSI	20091364-45	Cameroon	01-06-09-11

Table 2: List of significant biomarkers for the performance of classifiers between species.

Compared species	*m/z* of discriminant features	RT (s)	Ratio	Ion type	Elemental composition (hypothesis)	Δppm
ARA/CAN	195.0870	519	794	$[M+H]^+$	$C_8H_{10}N_4O_2$	3.34
	247.0598	399	63	$[M+H]^+$	$C_{13}H_{10}O_5$	1.21
ANTH/ARA	195.0870	519	0.0010	$[M+H]^+$	$C_8H_{10}N_4O_2$	3.34
	561.3617	915	0.0032	$[M+H]^+$	$C_{29}H_{52}O_{10}$	2.89
ARA/PSI	299.1997	899	0.0020	$[M+H]^+$	$C_{20}H_{26}O_2$	2.86
	868.4599	952	0.0020	$[M+H]^+$	*	-
LIB/PSI	335.2209	899	0.0010	$[M+H]^+$	$C_{20}H_{30}O_4$	2.34
	785.4229	945	0.00040	$[M+H]^+$	*	-
ARA/CHAR	514.3005	893	0.0010	$[M+NH_4]^+$ of 497.2742	$C_{26}H_{40}O_9$	1.09
	299.1997	899	0.0016	$[M+H]^+$	$C_{20}H_{26}O_4$	2.86
	335.2209	899	0.0016	[M+H]+	$C_{20}H_{30}O_4$	2.34
LIB/LvL	439.1555	700	20	$[M+H]^+$	*	-
	722.3207	455	16	$[M+H]^+$	*	-

* Various elemental compositions are proposed and are described in supporting information in Table S2

Research highlights

- 9 species of Coffee leaves metabolomics analysis has been undertaken by LC-HRMS
- Data processing and statistical analysis were performed on Workflow4Metabolomics
- Metabolomics data have been put in relation with botanic and genetic informations
- Some key metabolites for the discrimination between species has been characterized