# *UHPLC-MS/MS analysis of cocoa bean proteomes from four different genotypes*

# UHPLC-MS/MS analysis of cocoa bean proteomes from four different genotypes

**Emanuele Scollo[1,2], David C. A. Neville[2], M. Jose Oruna-Concha[3], Martine Trotin[2] and Rainer Cramer[1*]**

[1] Department of Chemistry, University of Reading, Reading RG6 6AD, UK
[2] Mondelēz International, Reading Science Centre, Reading RG6 6LA, UK
[3] Department of Food and Nutritional Sciences, University of Reading, Reading RG6 6AP, UK

*Address correspondence to:
Prof Rainer Cramer, Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6AD, UK.
Tel.: +44-118-378-4550; e-mail: r.k.cramer@rdg.ac.uk

**Running title:** Comparative UHPLC-MS/MS analysis of cocoa bean proteomes

**Nonstandard abbreviations:** emPAI (exponentially modified protein abundance index); FDR (false discovery rate); BSA (bovine serum albumin)

**Keywords:** Theobroma cacao, cocoa beans, plant proteomics, storage proteins, cocoa bean proteome, cocoa flavour

**Abstract**

In this study the proteomic profiles of cocoa beans from four genotypes with different flavour profiles were analysed by bottom-up label-free UHPLC-MS/MS. From a total of 430 identified proteins, 61 proteins were found significantly differentially expressed among the four cocoa genotypes analysed with a fold change of ≥2. PCA analysis allowed clear separation of the genotypes based on their proteomic profiles. Genotype-specific abundances were recorded for proteases involved in the degradation of storage proteins and release of flavour precursors. Different genotype-specific levels of other enzymes, which generate volatiles compounds that could potentially lead to flavour-inducing compounds, were also detected. Overall, this study shows that UHPLC-MS/MS data can differentiate cocoa bean varieties.

## 1. Introduction

Chocolate as commonly sold and consumed is made from the beans of the cocoa tree *Theobroma cacao* (family *Sterculiacae*). This tree is native to the Amazon and Orinoco valley and requires hot and humid weather conditions to grow. Traditionally, *Theobroma cacao* has been divided into three main genetic groups that are of commercial interest, Forastero, Criollo and Trinitario, the latter being a hybrid of the first two genetic groups.

Most of the cocoa beans produced in the world comes from Forastero varieties, which are considered "bulk in trade" (Lima, Almeida, Nout, & Zwietering, 2011). Two other cultivars have also been described: Amelonado, which is considered a subvariety of Forastero and mainly cultivated in West Africa, and Nacional, a cultivar native to Ecuador. However, this general classification is broad as hybridisation has occurred over time, which has given rise to differentiation within the same genetic groups, especially Forastero.

A study carried out by Motamayor *et al* (Motamayor, Lachenaud, da Silva e Mota, Loor, Kuhn, Brown, et al., 2008) has resulted in the identification of ten genetically distinct clusters. Based on these results the Forastero group has been differentiated into eight subvarieties: Amelonado, Contamana, Curaray, Guiana, Iquitos, Maranon, Nanay, and Purus. This new classification has not affected the Criollo and Nacional varieties, as they have maintained their original terms. The term Trinitario is commonly used to describe hybrids of Forastero and Criollo.

Cocoa beans from Trinitario and Criollo generate the "fine cocoa flavour" and make up only 5% of cocoa's total worldwide production (Lima, Almeida, Nout, & Zwietering, 2011). Ivory Coast is the main cocoa-producing country in the world with a total worldwide production share of 42%, followed by Ghana and Nigeria whose shares are 19% and 7%, respectively ("Quarterly Bulletin of Cocoa Statistic," 2017).

In general, shortly after harvest cocoa beans undergo natural fermentation which results in the release of free peptides and amino acids from storage proteins (Voigt, Biehl, Heinrichs, Kamaruddin, Marsoner, & Hugi, 1994). These compounds are important flavour precursors which contribute to the generation of cocoa aroma during roasting, and that is why poorly fermented beans have a low amount of flavour precursors and do not generate the typical cocoa aroma upon roasting. The autolysis of storage proteins extracted from unfermented cocoa beans generate flavour precursors which produce the typical cocoa aroma when roasted in the presence of the cocoa butter and reducing sugars (Voigt, Biehl, Heinrichs, Kamaruddin, Marsoner, & Hugi, 1994). The LC-ESI MS/MS analysis of these autolysis products revealed the presence of mainly hydrophilic peptides whose sequence could be linked to cocoa globulins (Voigt, Janek, Textoris-Taube, Niewienda, & Wostemeyer, 2016). The proteome of *Theobroma cacao* beans has recently been characterised using ultrahigh performance liquid chromatography (UHPLC) coupled to electrospray ionisation (ESI) tandem mass spectrometry (MS/MS) (Scollo, Neville, Oruna-Concha, Trotin, & Cramer, 2018). The highest proportion of the identified proteins could be linked to 'metabolism and energy' and 'proteins and synthesis' functions (Scollo, Neville, Oruna-Concha, Trotin, & Cramer, 2018). The most abundant proteins were albumin and vicilins (Scollo, Neville, Oruna-Concha, Trotin, & Cramer, 2018).

The proteomic profile of cocoa beans during development was previously evaluated by LC-ESI MS/MS using a 'bottom-up shotgun' approach (Wang, Nagele, Doerfler, Fragner, Chaturvedi, Nukarinen, et al., 2016). Cell division, ATP synthesis, RNA processing, amino acid synthesis and activation, protein synthesis, sucrose transportation and degradation-associated proteins were upregulated in young beans compared to mature beans (Wang, et al., 2016). Proteins involved in defence and stress were present at a higher level in mature beans (Wang, et al., 2016).

The proteomic profiles of non-fermented cocoa beans from various origins and varieties have been characterised by 2D gel electrophoresis and subsequent analysis by MALDI-TOF MS/MS from a total of 49 2D gel spots (Kumari, Grimbs, D'Souza, Verma, Corno, Kuhnert, et al., 2018). The authors reported differences in terms of numbers and intensities of proteins between samples from different origins, and samples from the same varieties grown in different countries (Kumari, et al., 2018). According to the authors a vicilin subunit was specific to samples of CCN51 hybrids and the German Forastero variety CD03 (Kumari, et al., 2018). Two protein gel spots which revealed a degraded 17-kDa albumin subunit and an internal 15-kDa vicilin subunit showed significant differences among the samples analysed when selecting the geographical origin and protein intensities as variables in a MANOVA analysis (Kumari, et al., 2018). The authors stated that these

proteins could be used as markers to assess the geographical origin and variety (Kumari, et al., 2018).

Although there are cocoa varieties with different flavour characteristics, it is not fully understood whether there is a link between the proteomic profile and flavour development. In this work a UHPLC-MS/MS bottom-up label-free approach was employed to characterise qualitative and quantitative differences in the proteomic profiles of cocoa beans from four genotypes, which show differences in both genetic background and flavour characteristics.

## 2. Materials and methods

### 2.1. Chemicals

Petroleum ether 40-60 was obtained from Fisher Scientific, Loughborough, UK. All other chemicals were obtained from Sigma-Aldrich, Gillingham, UK, except where stated otherwise.

### 2.2. Plant materials

Cocoa beans were from four different genotypes of *Theobroma cacao*, namely ICS 1, ICS 39, SCA 6 and IMC 67 harvested at the Cocoa Research Centre of the University of West Indies, St. Augustine, Trinidad, see Supplementary Table 1. Cocoa pods were harvested from 6 different trees for each genotype. A total of six pods were harvested from each tree. Each pod of the same cocoa genotype was considered a biological replicate within the specified cocoa genotype.

Pods were stored refrigerated for no longer than 3 days after being harvested. The beans were removed from the pods and the pulp manually removed with the aid of a scalpel. Depulped beans were stored at -20° C and subsequently freeze-dried for 24 hours. Following the freeze-drying step, the beans were stored at -20° C prior to shipping. The freeze-dried beans were air-freighted without temperature control to the University of Reading, UK. The shipment took less than 96 hours. Upon arrival, the beans were stored at -20° C prior to analysis. To obtain a representative sample for each cocoa genotype, approximately 2 g of beans from each biological replicate within the same genotype were combined, and the remainder of the beans were stored in their original container.

### 2.3. Fat and polyphenols removal

The freeze-dried beans were snap-frozen using liquid nitrogen and subsequently ground using a mortar and pestle. Fat from aliquots of approximately 160 mg were extracted with 3.5 ml of petroleum ether (boiling point 40-60º C) for 20 minutes in a vertical shaker. The suspensions were subsequently centrifuged at 3100 g for 5 minutes and the supernatants were discarded. The extraction was repeated twice and the precipitates were dried under a stream of nitrogen.

In order to prevent the formation of polyphenol-protein complexes during extraction (Voigt, Biehl, & Wazir, 1993), polyphenols were removed following a slight modification of a published method (Voigt, Wrann, Heinrichs, & Biehl, 1994). In brief, polyphenols were extracted from the defatted samples with 3.5 ml of a solution made up of cold (~4° C) aqueous acetone (80%; v/v), containing 5 mM sodium ascorbate. The suspensions were vortexed for 1 minute and centrifuged at 3100 g for 10 minutes at 4º C. The supernatant was discarded and the extraction repeated twice. Residual water was removed by extraction with 3.5 ml of cold acetone. The sample was then dried under a stream of nitrogen, resulting in acetone-dried powder (ACDP).

### 2.4. Protein extractions and Bradford assay

Proteins from the ACDP were extracted with 3.5 ml of a solution consisting of aqeous 7 M urea, 2 M thiourea and 20 mM dithiothreitol. The suspensions were placed on a vortexer for 1 minute and subsequently extracted for 1 hour at room temperature in a vertical shaker at 700 rpm. The suspension was subsequently centrifuged at 3100 g for 10 minutes at 20° C. The supernatant was removed and stored at -80° C prior to analysis. The protein concentration in each sample was assessed with the Bradford assay [4]. Bovine serum albumin (BSA) was used as reference standard for quantitation purposes.

### 2.5. Trypsin digestion

Aliquots of proteins extracts (35-47 µl) containing approximately 160 µg of proteins based on the Bradford assay were transferred into 0.5-ml microcentrifuge tubes and spiked with 30 µl of an aqueous 10 mg/l BSA solution. A volume of 20 µl of an aqueous 200-mM dithiothreitol (DTT) solution was then added to each tube, and the final concentration of DTT was adjusted to 10 mM by adding 290 µl of 77 mM ammonium bicarbonate. The solutions were incubated for 30 minutes at 37° C. A volume of 43 µl of an aqueous 200-mM iodoacetamide (IAA) solution was then added to each sample solution in order to obtain a final IAA concentration of 20 mM. By adding small aliquots of a 2-M urea solution the samples were adjusted to a final urea concentration of 0.6-0.7 M. To each sample tube, a volume of 20 µl of a 0.15-µg/µl trypsin (Promega, Southampton, UK) solution was added to obtain a 1:50 trypsin-to-protein ratio, and the solutions were incubated for approximately 16 hours at 37° C. After incubation the digestion was stopped by lowering the pH to below 3 with the addition of 20 µl of a 5% (v/v) solution of aqueous trifluoroacetic acid (TFA) to each sample tube.

### 2.6. Desalting of tryptic digests

The tryptic digest solutions were desalted with SOLAµ HRP 96 well plate 2 mg sorbent mass SPE cartridges (Thermo Scientific, Waltham, MA USA). The cartridges were initially conditioned with 0.2 ml of methanol and subsequently equilibrated with 0.2 ml of 0.2%

(v/v) TFA in 50 mM ammonium bicarbonate. After loading the sample solutions, the cartridges were washed with 0.2 ml of 0.2% TFA in water:methanol 97:3 (v/v), and then eluted with 3x 25 µl of 0.2% TFA in acetonitrile:water 50:50 (v/v) solution. The SPE eluates were diluted with 0.225 ml of 0.1% TFA in water and stored at -80° C prior to UHPLC-MS/MS analysis.

### 2.7. UHPLC-MS/MS analysis of tryptic digests

The desalted tryptic digests were analysed on a UHPLC-ESI MS/MS system consisting of an Orbitrap Q Exactive (Thermo Scientific) mass spectrometer coupled to a Dionex Ultimate 3000 (Thermo Scientific) UHPLC system. The injection volume was 15 µl. The UHPLC system was kept at 50° C and the column configuration included an Acquity Peptide CSH C18 150 mm × 0.1 mm ID, 1.7 µm particle size analytical column (Waters, Elstree, UK). The chromatographic separation of the digests was carried out under a linear gradient elution using 0.1 % (v/v) formic acid in water as mobile phase A and 0.1 % (v/v) formic acid in acetonitrile as mobile phase B with a flow rate of 0.1 ml/min. The gradient conditions were as follows: 2% B at 0-5 minutes, 30% B at 80 minutes, 60% B at 90 minutes, 90% B at 100-110 minutes, 2% B at 115-125 minutes. MS analysis was carried out in positive ion mode using the Orbitrap mass analyser, setting its resolution at 70,000 and its AGC (acquisition gain control) target at 1,000,000 with a maximum injection time of 200 ms. The MS scan covered a range between m/z 200 and 2400. For MS/MS analysis a data dependent experiment selecting the 10 most abundant precursor ions was performed, using the quadrupole mass analyser as the initial filter, and setting the isolation window width to m/z 2.0. For this experiment, the resolution of the Orbitrap was set to 17,500 with an AGC target of 20,000. The injection time for MS/MS acquisition was set to 300 ms. Fragmentation was performed by collision-induced dissociation (CID) with a normalised collision energy of 28%. Dynamic exclusion was enabled, setting the filter to 15 seconds. The threshold for triggering a data-dependent scan was set to '100,000' and only ions with a charge state between 2 and 5 were selected.

### 2.8. Data analysis

All MS/MS spectra were processed using Mascot Distiller software (Matrix Science Ltd, London, UK; Version 2.5.1.0) in order to convert the raw UHPLC-MS/MS data into peak lists suitable for database searching using the Mascot search routine. For the evaluation of the effect of harvest time and tree, Mascot Server Version 2.4.1 was used, while the analysis of the different cocoa genotypes was carried out employing the Mascot Server Version 2.6 (Matrix Science Ltd). Mascot searches were carried out against the Cacao Matina 1-6 Genome v1.1 *Theobroma cacao* database (http://www.cacaogenomedb.org/Tcacao_genome_v1.1#tripal_analysis-downloads-box;

downloaded on 31st May 2015; 59,577 sequences; 23,720,084 residues), and a custom-made contaminants database (70 sequences; 31,845 residues). Searches were performed using the following parameters: peptide mass tolerance, 10 ppm; MS/MS tolerance, 0.3 Da; peptide charge, +2, +3, +4; missed cleavages, 2; fixed modification, carbamidomethyl (C); variable modification, oxidation (M) and acetyl (N); enzyme, trypsin. The false discovery rate (FDR) for all searches was adjusted to 1%, which resulted in various significance thresholds for the different searches. However, the p-value was <0.05 for all searches. The amino acid sequence of BSA was added to the *Theobroma cacao* database. Functional annotation was carried out by matching the proteins' accession codes from the Cacao Matina 1-6 Genome v1.1 *Theobroma cacao* to the GoMapMan database (http://protein.gomapman.org). For each entry the highest hierarchical classification was used in this study. Label-free quantitation was carried out using replicate protocol with Mascot Distiller software. Normalisation of the proteins' intensities was carried out against BSA. Protein quantitation was performed by employing the median of the ion signal intensity ratios from all peptide for each protein, for which a minimum of two peptides were detected. For statistical analyses, JMP Pro 13.0 and XLSTAT 2108.5 software were used.

## 3. Results

A total of four different genotypes (ICS 1, ICS 39, IMC 67 and SCA 6) were evaluated in this project. These cocoa genotypes were carefully selected in order to include varieties with differences in both genetic background and flavour profiles. The list of the selected cocoa genotypes with their genetic backgrounds and flavour profiles is shown in Supplementary Table 1.

In order to minimise variability of the protein expression due to external factors, the investigated cocoa varieties were grown in the ICGT (International Cocoa Genebank Trinidad) field in Trinidad under controlled conditions in terms of water intake, fertilisation, and soil structure. However, to evaluate the effect of a different location on the proteomic profile of cocoa beans, trees from the genotype IMC 67 were also grown in a different field called "Campus" located within 5 Km off the ICGT field.

### 3.1. Effect of harvest time and different trees

To evaluate the effect of different trees on the proteomic profile of cocoa beans, four pods of the cocoa genotype IMC 67 were harvested on the same day, one from each of four different trees grown in the ICGT field. Three preparative replicates were prepared for each of these four biological replicates, and each preparative replicate was analysed by a single UHPLC-MS/MS run. UHPLC-MS/MS reproducibility was previously checked and constantly monitored by quality control samples of the same standard cocoa bean protein

extract analysed alongside the preparative replicates. For each quantified protein, the mean of the intensities in the three preparative replicates for each biological replicate (preparative sample mean) was calculated, and subsequently the average of the preparative sample means of the four biological replicates was calculated (overall mean). For each biological replicate the fold increase/decrease from the overall mean expressed as the ratio between the sample mean and the overall mean was calculated.

A total of 511 proteins were detected in the four biological replicates and only six proteins (Thecc1EG042578t1: S-adenosyl-L-methionine-dependent methyltransferases; Thecc1EG025391t1: beta-amylase 6; Thecc1EG000326t1: salicylate O-methyltransferase; Thecc1EG026589t1: eukaryotic aspartyl protease; Thecc1EG027146t1: HSP20-like chaperones; Thecc1EG041163t1: glycosyl hydrolase family protein) showed a fold increase/decrease from the overall mean of >2 in at least one biological replicate, while none showed any increase/decrease from the overall mean of >2.7. Of these six proteins only beta-amylase 6 showed this increase/decrease for two biological replicates, while the other five proteins showed this differential abundance for exactly one biological replicate, covering the entire set of the biological replicates.

The harvest time in this study covered a period of six months. Therefore, to evaluate the effect of harvest time on the proteomic profile of cocoa beans, four pods from the same tree (genotype IMC 67; grown in the ICGT field) but harvested at different times (20$^{th}$ Dec 2016, 21$^{st}$ Feb 2017, 23$^{rd}$ March 2017, 17$^{th}$ May 2017) were analysed. As before three preparative replicates were prepared for each of the four biological replicates, and each preparative replicate was analysed by a single UHPLC-MS/MS run.

A total of 502 proteins were detected in the four biological replicates analysed. Among these proteins, only nine entries showed a fold increase/decrease from the overall mean of >2 in at least one biological replicate (see Supplementary Table 2). In this case, two proteins (Thecc1EG042149t1: serine carboxypeptidase-like 48; Thecc1EG047098t1: uncharacterised) fluctuated far more than any protein in the tree comparison experiment with a fold increase/decrease of >3 and up to 11. The number of proteins with a fold increase/decrease from the overall mean of >2 in each biological replicate was between three and five.

### 3.2. Investigation of proteome changes dependent on the genotype and field

Cocoa pods were harvested from six different trees for each genotype grown in the ICGT field and for the six IMC 67 trees that were grown in the Campus field. A total of six pods were collected from each tree. Pooled samples containing an equal amount of all of the biological replicates from the same cocoa variety (Campus and ICGT grown IMC 67 pods were pooled separately) were prepared as described in the Materials and Methods section.

To evaluate the proteome changes, which are dependent on the genotype, a UHPLC-MS/MS label-free proteomic analysis was carried out on each of the cocoa genotypes. A total of four preparative replicates were prepared for each genotype sample, and each preparative replicate was analysed by UHPLC-MS/MS. A reference sample was prepared by combining equal aliquots of all 20 preparative replicates. The Distiller software calculated the ratios of the intensities of the proteins in each preparative replicate against the same proteins in the reference sample. Only proteins which were identified and quantified in at least three preparative replicates of a cocoa genotype were selected for comparative label-free quantitative proteomic analysis. With this requirement a total of 430 proteins were identified and quantified (see Supplementary Table 3). The mean of the ratios for the preparative replicates of the same cocoa genotype was calculated for each quantified protein. The fold differences between the cocoa genotypes are reported as the ratio of the highest mean versus the lowest mean for each quantified protein.

Almost all of the 430 proteins were detected in all genotypes apart from a 60S acidic ribosomal protein (accession number Thecc1EG005040t1) that was not detected in the genotype SCA 6. However, the abundance of this protein was not significantly different in the other genotypes. From all other identified and quantified proteins, a total of 61 proteins showed a significant fold difference of >2 (p-value <0.05) within at least one pairwise comparison among the four cocoa genotypes. Among these proteins, those with a sum of the four sample-to-reference ion signal ratios that is outside the range of 75-125% from the theoretical value of 4 were further evaluated to assess their peptide ion signal intensities. In this case a total of four proteins showed a signal too weak for reliable quantitation, and therefore these proteins were not further investigate. A list of the differentially expressed proteins, which showed an acceptable ion signal intensity, including their biological process and function, is provided in Table 1. A graphical representation of the proteins' classification based on their biological processes and functions is provided in Figure 1. Biological processes, for which only one protein was identified and quantified, are labelled as "Others" in Figure 1.

303 **Table 1.** List of differentially abundant proteins with a fold difference of >2, obtained from the four cocoa genotypes analysed by label-free
304 LC-MS/MS.

| ID | Accession | Description | Biological process | Function | ICS 1 | ICS 39 | IMC 67 | SCA 6 | Fold diff. |
|---|---|---|---|---|---|---|---|---|---|
| **1** | Thecc1EG029400t1 | N-terminal nucleophile aminohydrolases | Protein degradation | ME | 4.80 | 4.02 | 0.70 | 0.69 | 6.95 |
| **2** | Thecc1EG029392t1 | Glutathione S-transferase family protein | Gluthatione S-transferase | ME | 0.47 | 0.48 | 2.46 | 0.85 | 5.28 |
| **3** | Thecc1EG025391t1 | Beta-amylase 6 | Carbohydrate metabolism | ME | 0.49 | 1.30 | 1.14 | 2.39 | 4.90 |
| **4** | Thecc1EG017184t1 | Sulfite oxidase | S-assimilation | ME | 0.96 | 4.08 | 2.12 | 1.06 | 4.26 |
| **5** | Thecc1EG038258t1 | Molybdenum cofactor sulfurase | Co-factor and vitamin metabolism | ME | 0.78 | 1.46 | 0.92 | 0.35 | 4.12 |
| **6** | Thecc1EG030320t1 | Ethylene-forming enzyme | Hormone metabolism | ME | 0.44 | 1.79 | 0.50 | 1.10 | 4.03 |
| **7** | Thecc1EG021639t1 | PEBP | Unspecified biological process | UN | 1.72 | 2.22 | 1.22 | 0.66 | 3.36 |
| **8** | Thecc1EG020604t1 | Primary amine oxidase | Oxidase | ME | 1.89 | 0.78 | 1.27 | 2.54 | 3.25 |
| **9** | Thecc1EG047098t1 | Uncharacterized protein | Unspecified biological process | UN | 0.93 | 0.89 | 0.61 | 1.94 | 3.18 |
| **10** | Thecc1EG036433t1 | HSP20-like chaperones protein | Stress | DFS | 1.19 | 0.38 | 1.11 | 1.00 | 3.15 |
| **11** | Thecc1EG026543t1 | Lipoxygenase 1 | Hormone metabolism | ME | 0.47 | 0.45 | 1.23 | 1.37 | 3.01 |
| **12** | Thecc1EG026589t1 | Eukaryotic aspartyl protease | Protein degradation | ME | 1.69 | 1.67 | 0.58 | 1.10 | 2.91 |
| **13** | Thecc1EG042578t1 | S-adenosyl-L-methionine-dependent methyltransferases protein | Hormone metabolism | ME | 0.66 | 1.53 | 0.54 | 1.53 | 2.85 |
| **14** | Thecc1EG019372t1 | Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin protein | Protease inhibitor/seed protein/lipid transfer | SP | 1.91 | 1.43 | 2.33 | 0.84 | 2.78 |
| **15** | Thecc1EG027146t1 | HSP20-like chaperones protein | Stress | DFS | 1.46 | 0.53 | 0.98 | 0.72 | 2.77 |
| **16** | Thecc1EG026193t1 | Threonine aldolase 1 | Amino acid metabolism | ME | 1.53 | 0.85 | 0.58 | 0.91 | 2.66 |
| **17** | Thecc1EG037345t1 | 17.6 kDa class II heat shock protein | Stress | DFS | 1.35 | 0.51 | 1.13 | 0.60 | 2.65 |
| **18** | Thecc1EG012673t1 | 21 kDa seed protein* | Stress | DFS | 0.83 | 1.56 | 0.59 | 1.16 | 2.65 |
| **19** | Thecc1EG025860t2 | Uncharacterized protein | Unspecified biological process | UN | 1.41 | 0.88 | 2.00 | 0.76 | 2.64 |
| **20** | Thecc1EG012662t1 | 21 kDa seed protein* | Stress | DFS | 0.79 | 1.56 | 0.59 | 1.16 | 2.63 |
| **21** | Thecc1EG038931t1 | Xyloglucan endotransglycosylase 6 | Cell wall degradation | ME | 0.58 | 0.79 | 0.59 | 1.48 | 2.57 |
| **22** | Thecc1EG006471t1 | Flavin-dependent monooxygenase 1 | Oxidase | ME | 1.14 | 0.62 | 1.57 | 0.81 | 2.55 |
| **23** | Thecc1EG030938t1 | Cc-nbs-lrr resistance protein | Unspecified biological process | UN | 0.51 | 1.28 | 1.22 | 0.53 | 2.53 |
| **24** | Thecc1EG036938t1 | Aldolase-type TIM barrel | Nucleotide metabolism | ME | 1.95 | 0.88 | 0.92 | 0.79 | 2.47 |
| **25** | Thecc1EG006154t1 | Glycinamide ribonucleotide synthetase | Nucleotide metabolism | ME | 1.06 | 0.68 | 0.43 | 0.84 | 2.45 |
| **26** | Thecc1EG041496t1 | Stress responsive A/B Barrel Domain | Unspecified biological process | UN | 1.27 | 1.75 | 0.76 | 0.72 | 2.44 |
| **27** | Thecc1EG030354t1 | Fumarylacetoacetase | Amino acid metabolism | ME | 1.06 | 0.88 | 0.75 | 1.80 | 2.40 |
| **28** | Thecc1EG019909t2 | Carrot EP3-3 chitinase | Stress | DFS | 0.77 | 0.52 | 0.76 | 1.25 | 2.39 |
| **29** | Thecc1EG040975t1 | Alpha/beta-Hydrolases protein• | Gluco.-gala.-mannosidase | ME | 0.64 | 0.81 | 0.98 | 1.52 | 2.37 |
| **30** | Thecc1EG020603t2 | Primary amine oxidase | Oxidase | ME | 0.91 | 0.52 | 1.02 | 1.22 | 2.36 |
| **31** | Thecc1EG016747t1 | Acyl-CoA-binding protein 6 | Lipid metabolism | ME | 0.94 | 0.80 | 0.49 | 1.13 | 2.33 |
| **32** | Thecc1EG022426t1 | Thioredoxin protein | Redox | ME | 1.31 | 0.99 | 0.56 | 1.25 | 2.33 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **33** | Thecc1EG008318t1 | Aldolase-type TIM barrel | Oligopeptide transport system permease protein | ME | 1.26 | 1.27 | 0.57 | 1.32 | 2.30 |
| **34** | Thecc1EG022506t1 | Monodehydroascorbate reductase seedling isozyme | Redox | ME | 1.25 | 1.03 | 1.44 | 0.63 | 2.28 |
| **35** | Thecc1EG047057t1 | Cystathionine beta-synthase | Unspecified biological process | UN | 2.21 | 1.29 | 1.78 | 0.97 | 2.27 |
| **36** | Thecc1EG029923t1 | Larreatricin hydroxylase | Unspecified biological process | UN | 1.36 | 2.26 | 1.21 | 1.02 | 2.21 |
| **37** | Thecc1EG000245t1 | Serine carboxypeptidase S28 | Protein degradation | ME | 1.47 | 3.01 | 1.37 | 1.51 | 2.20 |
| **38** | Thecc1EG021820t1 | Tau class glutathione transferase GSTU45 | Gluthatione S-transferase | ME | 1.75 | 0.96 | 1.08 | 0.79 | 2.20 |
| **39** | Thecc1EG025715t1 | Uncharacterized protein | Unspecified biological process | UN | 0.87 | 1.03 | 0.47 | 0.92 | 2.20 |
| **40** | Thecc1EG043707t1 | Anti-oxidant 1 | Metal handling | ME | 1.05 | 1.19 | 0.80 | 1.76 | 2.19 |
| **41** | Thecc1EG016386t1 | 6-Phosphogluconate dehydrogenase | Oligopeptide transport system permease protein | ME | 1.36 | 0.90 | 0.62 | 1.08 | 2.18 |
| **42** | Thecc1EG014591t1 | Malate synthase glyoxysomal | Gluconeogenesis | ME | 2.08 | 0.96 | 1.05 | 1.37 | 2.17 |
| **43** | Thecc1EG042584t1 | S-adenosyl-L-methionine-dependent methyltransferases protein | Hormone metabolism | ME | 0.90 | 1.34 | 0.63 | 1.36 | 2.16 |
| **44** | Thecc1EG034339t1 | Dehydrin 2 | Stress | DFS | 1.37 | 0.91 | 0.72 | 0.63 | 2.16 |
| **45** | Thecc1EG035433t1 | Alcohol dehydrogenase 1$^{+}$ | Fermentation | ME | 1.14 | 0.90 | 0.63 | 1.37 | 2.16 |
| **46** | Thecc1EG010364t2 | Carbonic anhydrase 2 CA2 | TCA/organic transformation | ME | 1.20 | 0.98 | 0.79 | 0.56 | 2.13 |
| **47** | Thecc1EG006694t2 | Triosephosphate isomerase | Photosynthesis | ME | 1.40 | 1.35 | 0.71 | 1.51 | 2.11 |
| **48** | Thecc1EG006498t1 | Basic chitinase | Stress | DFS | 0.58 | 0.92 | 1.15 | 1.22 | 2.10 |
| **49** | Thecc1EG026326t2 | Pathogenesis-related protein P2 | Stress | DFS | 1.33 | 1.49 | 0.71 | 0.87 | 2.10 |
| **50** | Thecc1EG029913t1 | Alpha/beta-Hydrolases protein$^{\bullet}$ | Gluco.-gala.-mannosidase | ME | 1.55 | 1.16 | 0.75 | 0.81 | 2.07 |
| **51** | Thecc1EG036604t1 | Secretory laccase | Secondary metabolism | ME | 0.75 | 1.55 | 1.10 | 0.83 | 2.05 |
| **52** | Thecc1EG015253t1 | RNA binding Plectin/S10 domain-containing protein | Protein synthesis | PSP | 1.02 | 1.35 | 0.72 | 1.48 | 2.05 |
| **53** | Thecc1EG005533t1 | Transketolase | Photosynthesis | ME | 1.36 | 1.14 | 0.81 | 1.65 | 2.05 |
| **54** | Thecc1EG000770t1 | Acetamidase/Formamidase | Photosynthesis | ME | 2.12 | 1.52 | 1.04 | 1.49 | 2.03 |
| **55** | Thecc1EG014683t1 | Hydroxysteroid dehydrogenase 1 | Dehydrogenase | ME | 1.30 | 0.73 | 1.08 | 0.65 | 2.01 |
| **56** | Thecc1EG001447t1 | Alcohol dehydrogenase 1$^{+}$ | Fermentation | ME | 1.32 | 1.11 | 0.70 | 1.41 | 2.01 |
| **57** | Thecc1EG001141t1 | Lipase/lipooxygenase PLAT/LH2 | Unspecified biological process | UN | 1.61 | 1.11 | 1.02 | 0.81 | 2.00 |

DFS, defence and stress; ME, metabolism and energy; PSP, protein synthesis and processing; SP, storage proteins; UN, unclassified. In the genotype columns the average abundance ratio values relative to the reference sample of the preparative replicates are reported.*These proteins entries have a 99.5% homology and can therefore be considered to be proteoforms of the same gene. $^{+}$These protein entries have a 87% homology. $^{\bullet}$These protein entries have a 35% homology.

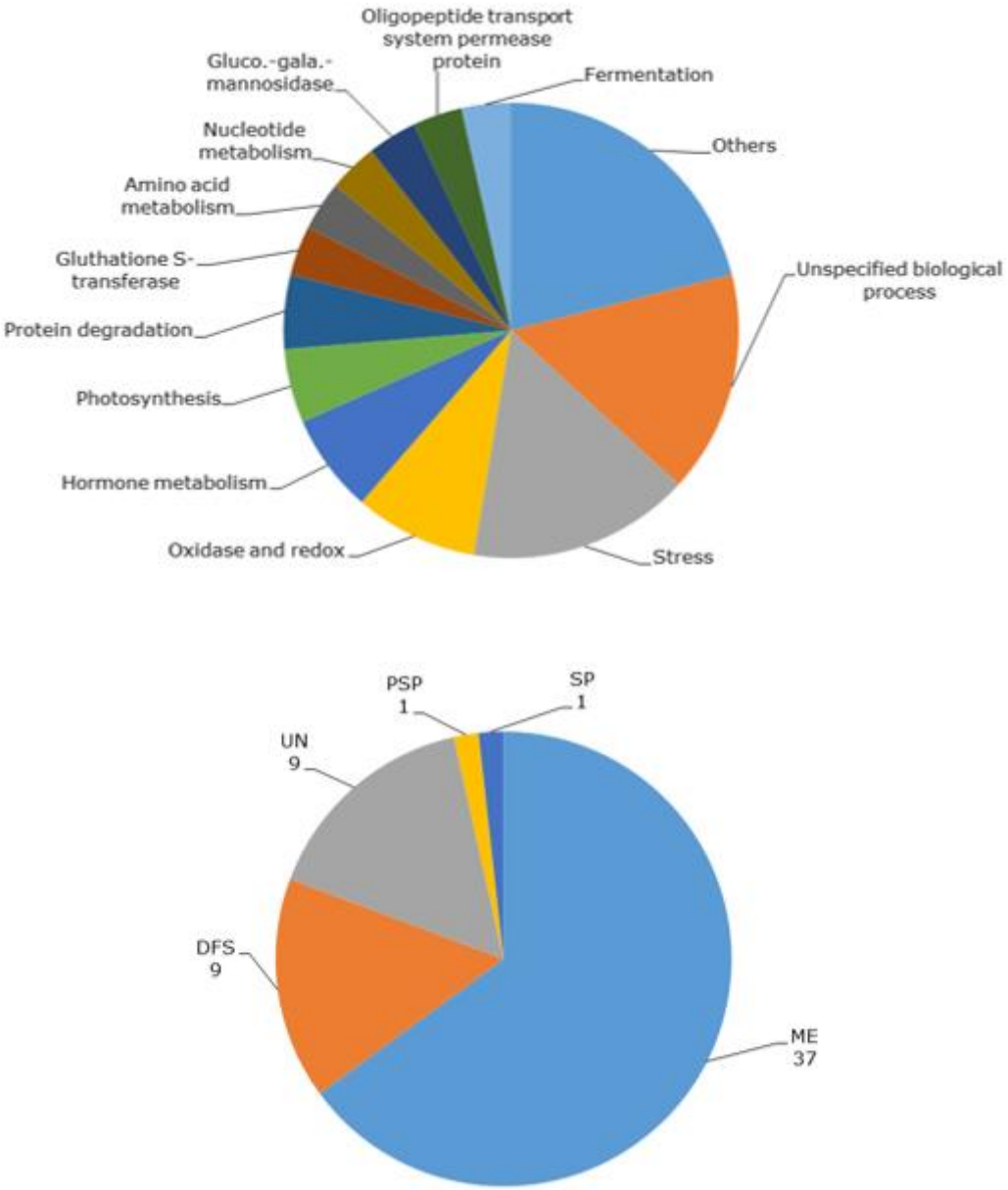


**Figure 1.** Classification of the differentially abundant proteins listed in Table 1 based on their biological process (upper pie chart) and their function (lower pie chart). 'Others' in the upper pie chart refers to all biological process, for which only one protein was found. The function group labels are as follows: DFS, defence and stress; ME, metabolism and energy; PSP, protein synthesis and processing; SP, storage proteins; UN, unclassified.

For each genotype, the number of proteins listed in Table 1 whose intensity was highest and lowest compared to the other genotypes are graphically represented in a histogram in Figure 2.

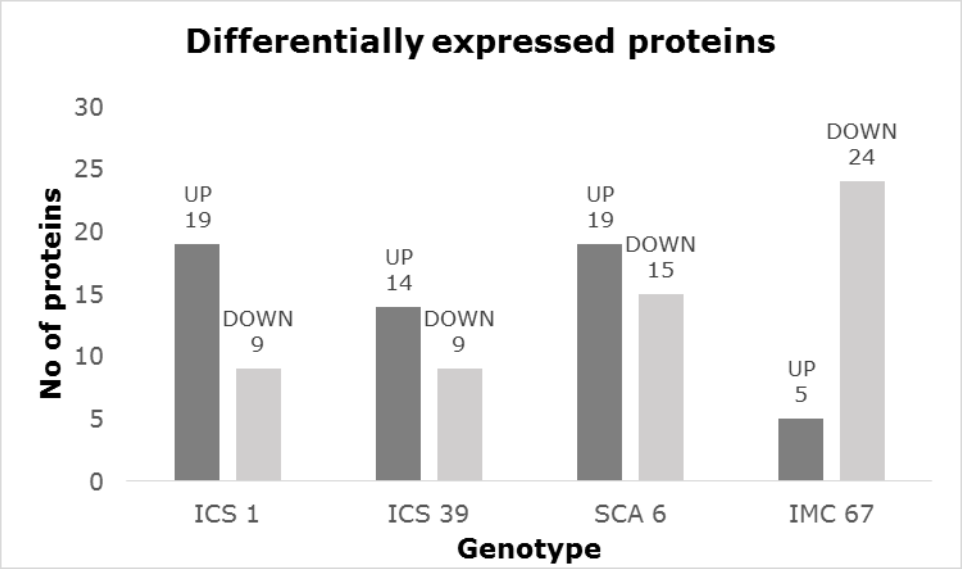


**Figure 2.** Number of differentially expressed proteins (fold difference of >2) detected at the higher (UP) and lower (DOWN) level for each genotype (see text for further details).

A pairwise comparison between the genotypes for each protein listed in Table 1 was also carried out, using the non-parametric Whitney Mann test to assess the significance of the differential expression ($p < 0.05$). The result of this comparison can be found in Supplementary Table 4.

To evaluate whether the proteomic data would allow a graphical differentiation of the four cocoa genotypes analysed, PCA analysis loading the ratios of the differentially expressed proteins listed in Table 1 as variables and the genotypes as observations was performed. In this case the data from all analytical replicates were used. The PCA score plot of the first two components clearly separates the four cocoa genotypes (see Figure 3). Each point in this graph represents a preparative replicate, and the replicates from the same genotype are displayed with the same colour. In order to assess which proteins were positively correlated to each genotype, a PCA loading plot of the differentially expressed proteins listed in Table 1 is also shown in Figure 3 (lower plot). Using this plot, variables should be positively correlated to observations which are located in similar regions of the score plot. For instance, the proteins with the ID 2, 19 and 22 are closest to the region in the score plot where the genotype IMC 67 is located and are greatly more abundant in the same genotype.

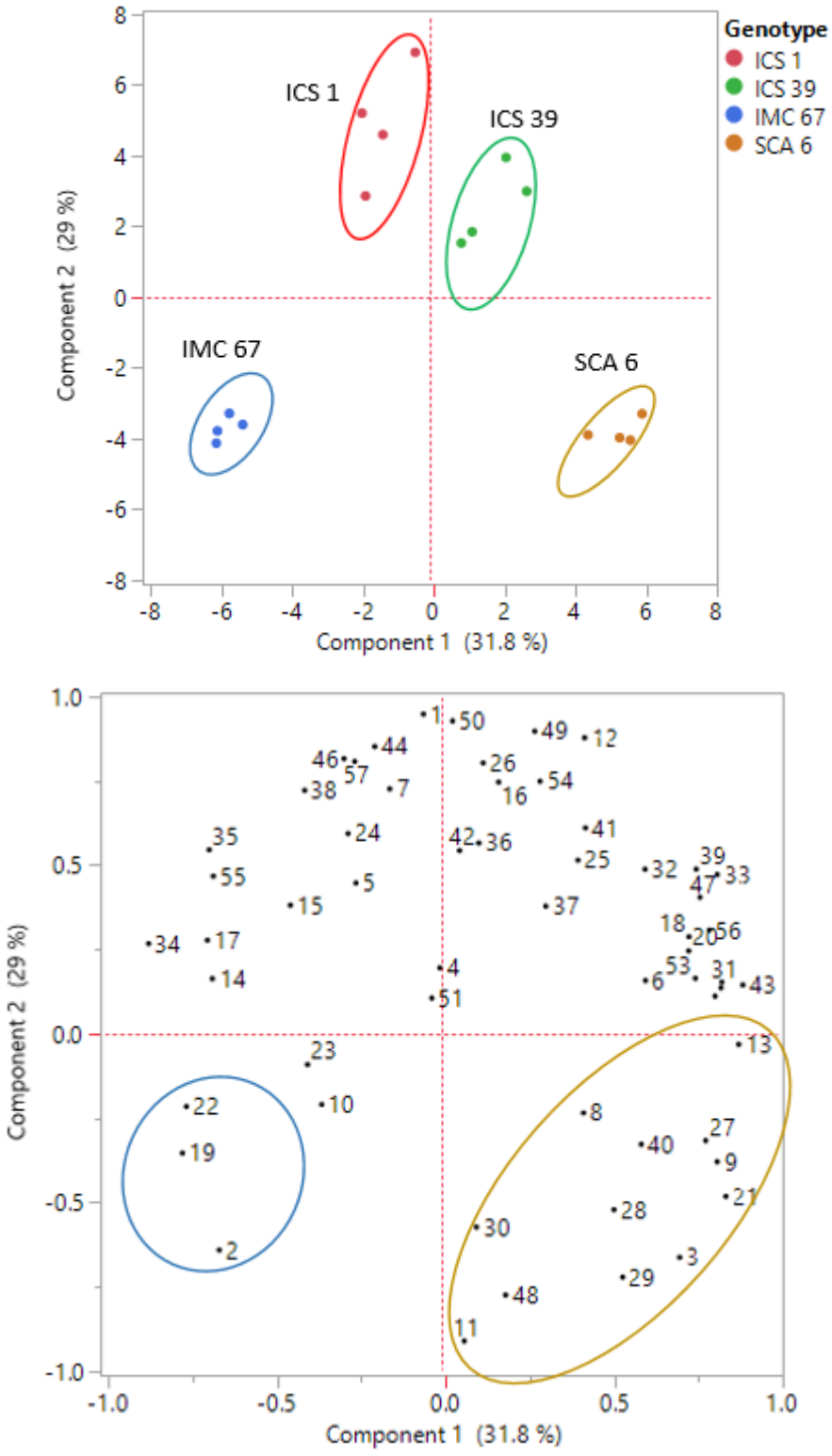


**Figure 3.** PCA score plot (upper plot) of the 57 differentially abundant proteins listed in Table 1. Preparative replicates of the same genotype are displayed with the same colour. The lower plot shows the PCA loading plot. Each variable is labelled with the corresponding ID number as listed in Table 1. The blue and yellow oval in the loading plot indicate clusters related to IMC 67 and SCA 6, respectively**.**

Comparing the proteomic profiles of the cocoa genotype IMC 67 grown in two different fields allowed the identification and quantitation of 430 proteins in the two biological replicates. Among these proteins, only four proteins were significantly different with a fold change of >2 between the two samples, while a ribosomal protein and a secretory laccase were detected only in the IMC 67 genotype grown in the ICGT field (see Supplementary Table 5). The latter two proteins were detected at low levels while the others had a fold change of <3.4.

Data supporting the results of this work are available in the PRIDE (Proteomics Identifications Database) partner repository at the European Bioinformatics Institute, PXD011984 (http://www.ebi.ac.uk/pride/).

## 4. Discussion

In the experiments carried out to assess the tree-to-tree, harvest time and field-to-field effects, the same genotype was used and only a few proteins, i.e. less than ten in each case, were detected with a fold increase/decrease of >2. Amongst these only three showed a fold increase/decrease of >3 and two were solely detected in one biological replicate but at a low level. Thus, given the total number of identified and quantified proteins, these results indicate that the variability in the detected proteome is very low between the biological replicates analysed to assess these effects.

In contrast, the analysis of the proteomic difference with respect to genotype revealed a high variability with more than 60 proteins showing a significant fold change of >2 for at least one pairwise genotype comparison. The overall highest fold difference in this comparison was found for an aminohydrolase (see Table 1). This protein was detected at significantly higher levels in both ICS genotypes, while it was found at much lower abundance in the genotypes IMC 67 and SCA 6. A blast search of the amino acid sequence of this protein returned a 100% match to a 20S proteasome alpha subunit which is part of the N-terminal nucleophile hydrolase superfamily. This class of proteins are involved in the hydrolysis of the amide bonds in either proteins or small molecules (Marchler-Bauer, Bo, Han, He, Lanczycki, Lu, et al., 2017). The active site is the N-terminal amino group which accepts a proton during the hydrolysis activating as a result either the nucleophilic hydroxyl in a Ser or Thr residue or the nucleophilic thiol in a Cys residue (Marchler-Bauer, et al., 2017).

The next highest fold change was recorded for a glutathione S-transferase (GST) family protein which was found at a much higher level in the genotype IMC 67 compared to all other genotypes. GST family proteins catalyse the conjugation of a variety of substrates to the reduced form of glutathione and therefore are involved in detoxification processes (Armstrong, 1997).

A 60S acidic ribosomal protein was not detected in any of the preparative replicates of the genotype SCA 6, while it was found in all other genotypes without any significant abundance differences. This class of proteins regulates the translation of mRNA in protein synthesis (Remacha, JimenezDiaz, Santos, Briones, Zambrano, Gabriel, et al., 1995).

With respect to the 57 proteins in Table 1, the highest number of less abundant proteins was found in the genotype IMC 67, and only 5 proteins were detected in this genotype at a higher level compared to the other genotypes (see Figure 2). The highest relative number of more abundant proteins compared to less abundant proteins (19 vs 9) was found for the genotype ICS 1.

The PCA score plot of the differentially expressed proteins for the four genotypes analysed shows that the individual genotypes are located in different quadrants of the plot and can be clearly separated from each other (see Figure 3). Both ICS 1 and ICS 39 belong to the same genetic group Trinitario, which originates from hybridizations between Criollo and Forastero. Therefore, the positive correlation of these genotypes in the PCA score plot could result from their closer genetic background compared to the other genotypes. IMC 67 and SCA 6 are genotypes from the genetically distant varieties Forastero and Contamana, respectively, of which both have a different genetic background from Trinitario (Motamayor, et al., 2008). Therefore, the separation pattern observed on the PCA score plot reflects the differences in genetic background among the four genotypes evaluated. Based on these findings, the PCA score plot of the differentially expressed proteins could be used as a tool to differentiate cocoa genotypes.

Loading the differentially expressed proteins as variables on a PCA loading plot allows a graphical visualisation of the proteins positively correlated to each genotype. The majority of the proteins more abundant in IMC 67 and SCA 6 form respective clusters in the bottom left and bottom right corner of the PCA loading plot (see Figure 3, lower plot), reflecting the separation of these genotypes observable in the PCA score plot. The genotypes ICS 1 and ICS 39 show a high degree of correlation in the PCA score plot. Therefore, the proteins found at a higher level in each of these genotypes cannot be separated in the PCA loading plot and form a single large cluster located at the top centre of the PCA loading plot. The location of this cluster is consistent with the position of these genotypes in the PCA score plot.

The highest number of differentially expressed proteins can be associated to metabolism and energy. This function class generally encompasses the majority of the proteins expressed in cocoa beans as shown in a previous study (Scollo, Neville, Oruna-Concha, Trotin, & Cramer, 2018), and includes two primary amine oxidases (ID 30 and ID 8 in Table 1) and two alcohol dehydrogenases identifications (ID 45 and ID 56 in Table 1), of which the latter are highly homologous (87% homology). Primary amine oxidases catalyse

the oxidation of alkylamines to aldehydes with the release of ammonia and hydrogen peroxide (Conklin, Prough, & Bhatanagar, 2007), while alcohol dehydrogenases catalyse the oxidation of primary and secondary alcohols to the corresponding aldehydes and ketones (Svensson, Hoog, Schneider, & Sandalova, 2000). It has been reported that both aldehydes and ketones are formed during roasting of fermented cocoa beans as a result of the Maillard reaction and Strecker degradation, and both classes of compounds contribute to the cocoa flavour (Aprotosoaie, Luca, & Miron, 2016). These reactions are endothermic as they require high temperatures to be activated and are not catalysed by enzymes. In theory, aldehydes and ketones could also be produced from oxidation of amines and alcohols during fermentation catalysed by amine oxidases and alcohol dehydrogenases. However, it is not known whether these enzymes are activated during this process, and whether there is a relation between their concentration and the generation of cocoa flavour. The primary amine oxidase (ID 30) was significantly more abundant in the genotype SCA 6 compared to ICS 39, while the other primary amine oxidase was significantly higher in the genotype SCA 6 versus ICS 39, and in the genotype ICS 1 versus ICS 39. Both alcohol dehydrogenase identifications IDs 45 and 56 were significantly more expressed in the genotype SCA 6 compared to IMC 67, reflecting their high homology and indicating that two proteoforms of the same gene were detected.

A total of 9 proteins involved in stress response were differentially expressed. Four of these proteins (ID 10, 15, and 17 in Table 1) were heat shock proteins by name which are linked to the response of the plant to stress conditions (Al-Whaibi, 2011). There are no significant differences in the abundances of these proteins in ICS 1 versus IMC 67 and SCA 6 versus IMC 67 but they were significantly more abundant in ICS 1 compared to ICS 39. Both these genotypes belong to the Trinitario variety which is originally from Trinidad and includes all hybridisation combinations of the Criollo and Forastero varieties. Criollo varieties are more susceptible to disease and adverse environmental factors. The genotype ICS 39 has a stronger Criollo ancestry compared to ICS 1, which could explain why heat shock proteins are more abundant in ICS 1 compared to ICS 39.

A eukaryotic aspartyl protease (ID 12 in Table 1) was significantly more abundant in the genotypes ICS 1 and ICS 39 compared to IMC 67 (fold difference of 2.9). Eukaryotic aspartyl protease is a cocoa endogenous protease which has an optimum pH of around 3.8 and is active during early stage of fermentation, cleaving internal peptides bonds with the release of mainly hydrophobic peptides (Voigt, Biehl, Heinrichs, Kamaruddin, Marsoner, & Hugi, 1994). The abundance of this protease was not consistent in the biological replicates of IMC 67 harvested from different trees on the same day. Therefore, the low amount found in the pooled sample may be due to natural variations amongst biological tree replicates.

A serine carboxypeptidase (ID 37 in Table 1) was detected at a significant higher level in ICS 39 compared to the other genotypes. Carboxypeptidase is an exopeptidase which cleaves off C-terminal amino acids from mainly hydrophobic oligopeptides formed by the action of aspartyl protease during fermentation with the preferential release of hydrophobic amino acids and hydrophilic peptides (Bytof, Biehl, Heinrichs, & Voigt, 1995). These compounds are important flavour precursors which react with sugars during roasting to form volatiles compounds which contribute to the cocoa aroma. A higher amount of aspartyl protease and carboxypeptidase could result in an increase in the generation of flavour precursors during fermentation, which could lead to changes in the flavour profiles of roasted cocoa beans.

A beta-amylase was detected at a significantly higher level in the genotype SCA 6 compared to ICS 1 and IMC 67 (ID 3 in Table 1). Beta-amylases are part of the glycoside hydrolase family, which are a group of enzymes catalysing the cleavage of the glycosidic bond in polysaccharides with release of maltose units (Rejzek, Stevenson, Southard, Stanley, Denyer, Smith, et al., 2011). This disaccharide can react with nitro compounds such as amino acids and peptides during roasting through the Maillard reaction which results in the generation of volatile compounds (Kramholler, Pischetsrieder, & Severin, 1993). Therefore, the release of maltose can be affected by the levels of beta-amylase present in cocoa beans, which in turn could have an effect on the flavour profile of roasted cocoa beans. However, the abundance of this specific beta-amylase was not consistent in the biological replicates of IMC 67 harvested from different trees on the same day. Therefore, the low amount found in the pooled sample may be due to natural variations amongst biological tree replicates.

Two 21-kDa seed albumin identifications were obtained at a significant higher level in the genotype ICS 39 compared to IMC 67 (ID 18 and 20 in Table 1). These albumins are storage proteins with endopeptidase inhibitor activity, which contain 219 amino acids residues, and can be considered to originate from the same gene as they are 99.5% homologous. The main 21-kDa seed albumin in cocoa beans is a protein with 221 residues which shares a homology of 80% with the albumins ID 19 and 21 listed in Table 1. The 221-residues albumin was not differentially expressed in the cocoa genotypes analysed (see Supplementary Table 3). LC-MS/MS identification of free peptides released from this protein during fermentation have been reported by several authors (Caligiani, Marseglia, Prandi, Palla, & Sforza, 2016; D'Souza, Grimbs, Grimbs, Behrends, & Corno, 2018; Marseglia, Sforza, Faccini, Bencivenni, Palla, & Caligiani, 2014). However, so far there is no evidence that the 219-residues albumins are also degraded during this process. As a result, the shorter chain albumins may not play a role in the generation of cocoa flavour.

A 2S albumin was significantly more abundant in the genotypes ICS 1 and IMC 67 compared to SCA 6 (see ID 14 in Table 1). This albumin is a seed storage protein with protease inhibitor activity which is also involved in the transfer of phospholipids and fatty acids through the cell membrane (Kader, 1996). Degradation of this protein during fermentation has not been reported in the literature.

In addition to cocoa endogenous enzymes, the amount of pulp in the cocoa pod and the surrounding microflora can also play a role in the generation of cocoa flavour.

## 5. Conclusions

This work has shown that UHPLC-MS/MS can be employed to characterise qualitative and quantitative differences in the proteomic profiles of cocoa beans from various genotypes. The PCA analysis has allowed separation of the cocoa genotypes from different varieties and has shown a correlation between close genotypes and their genetic background. Using this approach, it was possible to graphically visualise proteins positively correlated with each genotype, and assess which proteins contribute most to the separation of the genotypes in the PCA plot. This methodology could be employed as a platform to build larger datasets of proteins which could allow traceability of cocoa beans from different varieties. Proteases which degrade storage proteins during fermentation with the release of flavour precursors have been found differentially expressed in some of the genotypes analysed. Changes in the amount of these proteases could be related to variation in the flavour profiles of cocoa varieties. Different genotype-specific levels of other enzymes that could potentially lead to flavour-inducing compounds have also been detected. Thus, further experiments could be performed to assess whether the different amounts of these enzymes, present during fermentation, affect the final flavour profiles obtained

## Acknowledgements

The authors are grateful to the University of West Indies Cocoa Research Centre for providing the cocoa bean samples. This work has been financially supported by Mondelēz International.

**References**


Al-Whaibi, M. H. (2011). Plant heat-shock proteins: A mini review. *Journal of King Saud University Science, 23*(2), 139-150.

Aprotosoaie, A. C., Luca, S. V., & Miron, A. (2016). Flavor Chemistry of Cocoa and Cocoa Products-An Overview. *Comprehensive Reviews in Food Science and Food Safety, 15*(1), 73-91.

Armstrong, R. N. (1997). Structure, catalytic mechanism, and evolution of the glutathione transferases. *Chemical Research in Toxicology, 10*(1), 2-18.

Bytof, G., Biehl, B., Heinrichs, H., & Voigt, J. (1995). SPECIFICITY AND STABILITY OF THE CARBOXYPEPTIDASE ACTIVITY IN RIPE, UNGERMINATED SEEDS OF THEOBROMA-CACAO L. *Food Chemistry, 54*(1), 15-21.

Caligiani, A., Marseglia, A., Prandi, B., Palla, G., & Sforza, S. (2016). Influence of fermentation level and geographical origin on cocoa bean oligopeptide pattern. *Food Chemistry, 211*, 431-439.

Conklin, D., Prough, R., & Bhatanagar, A. (2007). Aldehyde metabolism in the cardiovascular system. *Molecular Biosystems, 3*(2), 136-150.

D'Souza, R. N., Grimbs, A., Grimbs, S., Behrends, B., & Corno, M. (2018). Degradation of cocoa proteins into oligopeptides during spontaneous fermentation of cocoa beans. *Food Research International, 109*, 516-516.

Kader, J.-C. (1996). Lipid-transfer proteins in plants. *Annual Review of Plant Physiology and Plant Molecular Biology, 47*, 627-654.

Kramholler, B., Pischetsrieder, M., & Severin, T. (1993). MAILLARD REACTIONS OF LACTOSE AND MALTOSE. *Journal of Agricultural and Food Chemistry, 41*(3), 347-351.

Kumari, N., Grimbs, A., D'Souza, R. N., Verma, S. K., Corno, M., Kuhnert, N., & Ullrich, M. S. (2018). Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans. *Food research international (Ottawa, Ont.), 111*, 137-147.

Lima, L. J. R., Almeida, M. H., Nout, M. J. R., & Zwietering, M. H. (2011). *Theobroma cacao L.*, "The Food of the Gods": quality determinants of commercial cocoa beans, with particular reference to the impact of fermentation. *Critical Reviews in Food Science and Nutrition, 51*(8), 731-761.

Marchler-Bauer, A., Bo, Y., Han, L. Y., He, J. E., Lanczycki, C. J., Lu, S. N., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z. X., Yamashita, R. A., Zhang, D. C., Zheng, C. J., Geer, L. Y., & Bryant, S. H. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research, 45*(D1), D200-D203.

Marseglia, A., Sforza, S., Faccini, A., Bencivenni, M., Palla, G., & Caligiani, A. (2014). Extraction, identification and semi-quantification of oligopeptides in cocoa beans. *Food Research International, 63*, 382-389.

Motamayor, J. C., Lachenaud, P., da Silva e Mota, J. W., Loor, R., Kuhn, D. N., Brown, J. S., & Schnell, R. J. (2008). Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L). *Plos One, 3*(10).

Quarterly Bulletin of Cocoa Statistic. (2017). *International Cocoa Organization, XLIII*(1).

Rejzek, M., Stevenson, C. E., Southard, A. M., Stanley, D., Denyer, K., Smith, A. M., Naldrett, M. J., Lawson, D. M., & Field, R. A. (2011). Chemical genetics and cereal starch metabolism: structural basis of the non-covalent and covalent inhibition of barley beta-amylase. *Molecular Biosystems, 7*(3), 718-730.

Remacha, M., JimenezDiaz, A., Santos, C., Briones, E., Zambrano, R., Gabriel, M. A. R., Guarinos, E., & Ballesta, J. P. G. (1995). Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects. *Biochemistry and Cell Biology, 73*(11-12), 959-968.

Scollo, E., Neville, D., Oruna-Concha, M. J., Trotin, M., & Cramer, R. (2018). Characterization of the Proteome of Theobroma cacao Beans by Nano-UHPLC-ESI MS/MS. *Proteomics, 18*(3-4).

Svensson, S., Hoog, J. O., Schneider, G., & Sandalova, T. (2000). Crystal structures of mouse class II alcohol dehydrogenase reveal determinants of substrate specificity and catalytic efficiency. *Journal of Molecular Biology, 302*(2), 441-453.

Voigt, J., Biehl, B., Heinrichs, H., Kamaruddin, S., Marsoner, G. G., & Hugi, A. (1994). In-vitro formation of cocoa-specific aroma precursors: aroma-related peptides generated from cocoa-seed proteins by co-operation of an aspartic endoprotease and a carboxypeptidase. *Food Chemistry, 49*(2), 173-180.

Voigt, J., Biehl, B., & Wazir, S. K. S. (1993). The major seed proteins of *Theobroma cacao L. Food Chemistry, 47*(2), 145-151.

Voigt, J., Janek, K., Textoris-Taube, K., Niewienda, A., & Wostemeyer, J. (2016). Partial purification and characterisation of the peptide precursors of the cocoa-specific aroma components. *Food Chemistry, 192*, 706-713.

Voigt, J., Wrann, D., Heinrichs, H., & Biehl, B. (1994). The proteolytic formation of essential cocoa-specific aroma precursors depends on the particular chemical structures of the vicilin-class globulin of the cocoa seeds lacking in the globular storage proteins of coconuts, hazelnuts and sunflower seeds. *Food Chemistry, 51*(2), 197-205.

Wang, L., Nagele, T., Doerfler, H., Fragner, L., Chaturvedi, P., Nukarinen, E., Bellaire, A., Huber, W., Weiszmann, J., Engelmeier, D., Ramsak, Z., Gruden, K., & Weckwerth, W. (2016). System level analysis of cacao seed ripening reveals a sequential interplay of primary and secondary metabolism leading to polyphenol accumulation and preparation of stress resistance. *Plant Journal, 87*(3), 318-332.