University of Reading

Investigating the proteomic profile of cocoa beans for understanding the development of cocoa flavour

This dissertation is submitted to the School of Chemistry, Food and Pharmacy, University of Reading for the Degree of Doctor of Philosophy

Emanuele Scollo

November 2019

I. Abstract

Cocoa seed storage proteins play an important role in flavour development as aroma precursors are formed from their degradation during fermentation. Major proteins in the beans of *Theobroma cacao* are the storage proteins belonging to the vicilin and albumin classes. Although both these classes of proteins have been extensively characterised, there is still limited information on the expression and abundance of other proteins present in cocoa beans. This work is the first attempt to characterize the whole cocoa bean proteome by nano-LC-ESI MS/MS analysis using tryptic digests of cocoa bean protein extracts. The results of this analysis showed that over 1000 proteins could be identified using a species-specific *Theobroma cacao* database. The majority of the identified proteins were involved with metabolism and energy. Albumin and vicilin storage proteins showed the highest intensity values among all detected proteins. A comparison of MS/MS data searches carried out against larger non-specific databases confirmed that using a species-specific database can increase the number of identified proteins, and at the same time reduce the number of false positives.

The proteomic profiles of cocoa beans from four genotypes with different genetic background and flavour profiles have also been analysed employing a bottom-up label-free UHPLC-MS/MS approach. From a total of 430 identified proteins, 61 proteins were found significantly differentially abundant among the four cocoa genotypes analysed with a fold change of 2 or more. PCA analysis allowed a clear separation of the genotypes based on their proteomic profiles. Interestingly, proteases which degrade storage proteins during fermentation have been found differentially abundant in some of the genotypes analysed. These proteins are involved in the release of flavour precursors, and therefore might play a key role in the shaping of the final flavour profile. Different genotype-specific levels of other enzymes which generate volatiles compounds that could potentially lead to flavour-inducing compounds have also been detected. Overall, this study shows that UHPLC-MS/MS data can differentiate cocoa bean varieties, and thus might be linked to differences in their flavour profile.

Finally, a method to identify and quantify free peptides from fermented cocoa beans by UHPLC-MS/MS analysis has been developed. A total of 155 peptides could be identified and quantified in fermented cocoa beans using this approach. The vast majority of these peptides were associated to vicilin and a 21 kDa albumin, which are the most abundant proteins in cocoa beans. This methodology could be applied to assess the free peptides profiles of cocoa beans at different stage of fermentation.

II. Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

I confirm that I am the main author of the related publications listed in chapter V, and I have carried out all the experiments and data processing. Other authors listed in the articles have only marginally contributed with the design of the experiments and reviewing the contents of the manuscripts.

Signed:____

Emanuele Scollo

Date:____

III. Acknowledgements

I would like to thank Mondelēz International for giving me the possibility to undertake this PhD and covering research expenses and tuition fees.

I am very grateful to my internal supervisor Martine Trotin for believing in this project and pursuing it within the company so that the project could be approved. I want to thank Martine for all the help she has provided in organising project proposals, design of experiments, meeting with technical experts in the cocoa field and finalising the collaboration with the University of West Indies for sample collection.

I want to thank my former managers Ellen Norman and Laurence Darnell for agreeing to let me carry out this project, and for being supportive during the achievement of my PhD.

I am particularly grateful to my lab manager Alex Webbe for believing in this project and allowing me to work part-time to complete my PhD.

I want to express my gratitude to Peter Clarke for endorsing this project and his support during the course of my studies.

I am grateful to Jacinta George for the final approval of this PhD project.

I would like to thank my University supervisors Prof Rainer Cramer and Dr Maria Jose Oruna-Concha for all the help, advice and support they have offered during the course of my PhD.

I am very grateful to Pat Donahue for his pivotal role in proposing and finalising the collaboration with the University of West Indies, and for the help and support he has provided. I would like to thank other cocoa technical experts at Mondelēz Juan, Nick, Will, for always being available to help and advise on cocoa relating matters.

I am very grateful to Prof Uma, Darin and Rena from the University of West Indies for providing the samples analysed for my PhD, and for their support with technical questions related to cocoa.

I would like to thank David Neville for his support in organising experimental designs, reviewing project proposals and research articles.

I am very grateful to all team in ACHEM for being supportive over the course of my studies and sharing some of my workload while I was working for my PhD.

I would like to thank all members of the Cramer group present and past Cristian, Chhaya, Kanjana, Ryan, Oli, Evita along with Nick from the CAF, for the help provided during my PhD.

I am grateful to Trevor Duguid Farrant for the support he provided with the statistical treatment of the proteomics data.

I want to express my gratitude to my wife Ania for her support and patience during my studies, and my parents who have always encouraged me to pursue my career and showed me their support.

IV. Table of contents

V. Related publication

Sections of the data and text presented in this thesis have been compiled and submitted for publication in the form of peer-reviewed original research articles and peer-reviewed literature review articles.

1. **Characterization of the proteome of *Theobroma cacao* beans by nanoUHPLC-ESI MS/MS**

   Emanuele Scollo[1,2,] David 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

This manuscript was accepted for publication in the Journal "Proteomics" on 22nd December 2017.

2. **UHPLC-MS/MS analysis of the proteomic profiles of cocoa beans from different genotypes**

   Emanuele Scollo[1,2,] David 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

This manuscript was accepted for publication in the Journal "Food Chemistry" on 23rd July 2019.

VI. Abbreviations

1D	One dimensions
2D	Two dimensions
AGC	Acquisition gain control
BPC	Base peak current
BSA	Bovine serum albumin
CID	Collision induced dissociation
DC	Direct current
DDA	Data dependent acquisition
DIA	Data independent acquisition
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
emPAI	Exponentially modified protein abundance index
ESI	Electrospray ionisation
HCD	High energy collision dissociation
HPLC-UV	High performance liquid chromatography with UV detector
IAA	Iodoacetamide
ITRAQ	Isobaric tags for relative and absolute quantitation
LC	Liquid chromatography
MALDI	Matrix-assisted-laser-desorption-ionization
MANOVA	Multivariate analysis of variance
mRNA	Messenger ribonucleic acid
MS	Mass spectrometry
NAD(P)	Nicotinamide adenine dinucleotide phosphate
PCA	Principal component analysis
PVPP	Polyvinylpolypyrrolidone
RF	Radio frequency
RNA	Ribonucleic acid
RSD	Relative standard deviation
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SILAC	Stable Isotope Labeling with amino acids in cell culture
SPE	Solid phase extraction
TCA	Trichloroacetic acid
TFA	Trifluoroacetic acid
TOF	Time of flight mass spectrometer
UHPLC	Ultra high-performance liquid chromatography

VII. Amino acids chart

Name	One letter symbol	Three letters symbol
Alanine	A	Ala
Cysteine	C	Cys
Aspartic Acid	D	Asp
Glutamic Acid	E	Glu
Phenylalanine	F	Phe
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Lysine	K	Lys
Leucine	L	Leu
Methionine	M	Met
Asparagine	N	Asn
Proline	P	Pro
Glutamine	Q	Gln
Arginine	R	Arg
Serine	S	Ser
Threonine	T	Thr
Valine	V	Val
Tryptophan	W	Trp
Tyrosine	Y	Tyr

1 INTRODUCTION

1.1 General information on cocoa and chocolate

The cocoa tree, *Theobroma cacao* (family *Sterculiacae*), is native to the Amazon and Orinoco valleys and its natural habitats are the tropical areas of South and Central America [1]. The length of the cocoa tree can vary from 8 to 15 m. However trees subjected to an intensive cultivation are usually pruned to reduce their length to 2.5-3 m [2], see Figure 1.

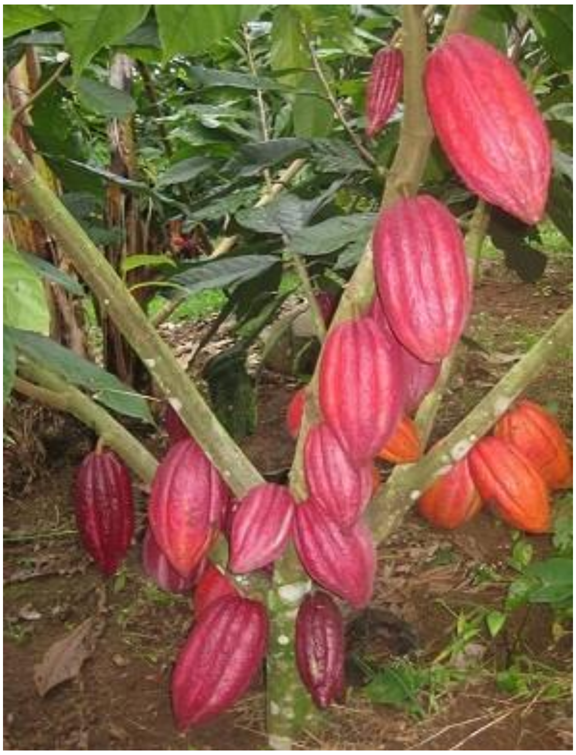


Figure 1. *Theobroma cacao* tree with mature pods

Mature fruits (pods) are thick-walled and contain 30-40 seeds, also called beans. Each bean consists of two cotyledons and an embryo (radicle) surrounded by a seed coat (testa), and is enveloped in a sweet, white mucilaginous pulp (see Figure 2). The name of the genus "Theobroma" was given by the Swedish botanist Linnaeus and is derived from the Greek words "Theo" (meaning god) and "Broma" (meaning food), referring to the Mayan and Aztec popular belief that chocolate was the food of the gods. It was Hernando Cortés who introduced this cultivar to Spain in 1520, and thereafter it spread to other European colonies in Africa and Asia, where it could find optimal habitat for its growth.

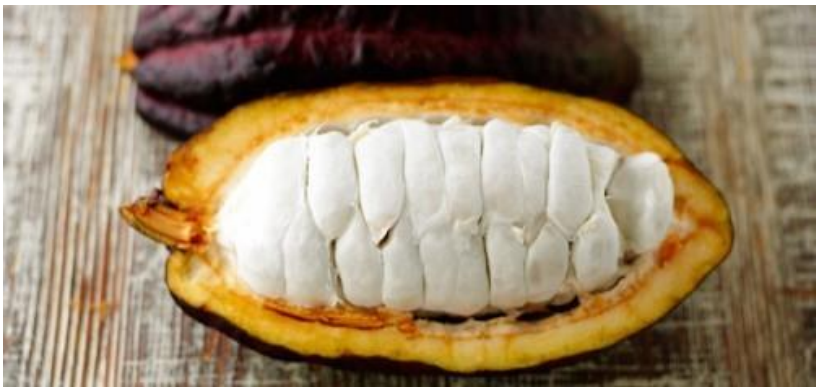


Figure 2. Internal view of a ripe cocoa pod with beans surrounded by pulp

Chocolate is made from the processing of cocoa beans. Traditionally, *Theobroma cacao* has been divided into three main genetic groups, Forastero, Criollo and Trinitario, which is a hybrid of the first two genetic groups. Two other cultivars have also been described: Amelonado which is considered a subvariety of Forastero mainly cultivated in West Africa, and Nacional, a cultivar native to Ecuador. However, this classification is quite broad as hybridization 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.* [3] who genotyped 1,241 cocoa samples from different geographic origin, has resulted in the identification of ten genetically distinct clusters. Based on these results the Forastero group has been differentiated into eight varieties: Amelonado, Contamana, Curaray, Guiana, Iquitos, Maranon, Nanay, and Purus. This new classification has not affected Criollo and Nacional varieties, as they have maintained their original terms.

Forastero trees have a very high yield and are relatively resistant to pest and diseases. Nowadays most populations cultivated in the world are locally adapted Forastero varieties, improved with the creation of related genotypes [2]. These varieties are regarded as "bulk cocoa in trade" and make up almost 95 % of the cocoa's total worldwide production [2]. Chocolate milk, cocoa butter and cocoa powder are manufactured using the beans of this variety. Forastero beans are flat, astringent and have a high content of anthocyanins which gives a purple colour to the mature pod [2].

The Criollo variety is derived from the native population indigenous to North, South and Central America. The beans of this variety show a white to ivory or a pale colour, which is caused by the presence of an anthocyanins inhibitor gene. The cultivation of Criollo is limited to few regions in Central America and Asia, as this population is susceptible to diseases and has a very low yield [2].

The Trinitario type is native to Trinidad and includes all hybridisation combinations of the Criollo and Forastero varieties. The colour of the beans is variable, although it is very rarely white, and this variety's susceptibility to diseases is intermediate between Forastero and Criollo. Both the Trinitario and the Criollo varieties produce the "fine flavour" cocoas, which account for less than 5% of the total cocoa's world production [2].These cocoas are used to make high quality dark chocolate [2].

The total cocoa's worldwide production as of 2017 was 4.552 million tonnes [4]. Africa is by far the largest cocoa-producing region in the word with an output of 3.37 million tonnes accounting for 74% of the total cocoa worldwide production [4]. Other important cocoa-producing region are the Americas, Asia and Oceania. Ivory Coast is the main cocoa-producing country in the world with a share of the total cocoa's worldwide production of 42 %, followed by Ghana and Nigeria, whose shares are 19 and 7%, respectively [4], as shown in Figure 3.

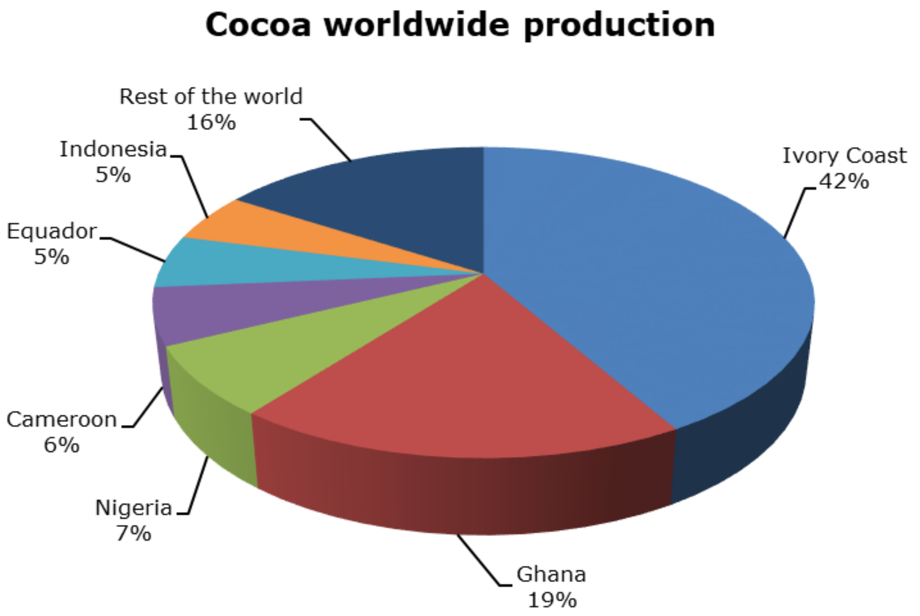


Figure 3. Cocoa's worldwide production in 2017 [4]**.**

1.2 Cocoa beans fermentation

Fermentation is an essential process for the development of cocoa flavour. Seeds inside the ripe pod are microbiologically sterile [5]. The pulp gets contaminated with microorganisms when the pods are opened with a knife. This microflora can originate from different sources such as the workers' hands, knives, bags used to transport the beans, residual mucilage left from previous fermentation on the wall of boxes or on the banana leaves used for heap fermentation [5]. Due to this microbial variability, differences in flavour characteristics between beans from the same cultivar fermented under similar conditions can be observed [6]. The pulp walls are broken down with the release of juice in a process called 'sweating'. The sugars from the pulp are fermented to ethanol and organic acids, causing the death of the bean. Storage proteins in the bean are degraded and reducing sugars are released from the hydrolysis of sucrose, leading to the formation of flavour precursors, which generate the characteristic cocoa aroma upon roasting [7]. Unfermented and under-fermented beans do not contain an appropriate amount of flavour precursors. Therefore, they do not generate cocoa flavour upon roasting. The duration of fermentation varies according to the cocoa variety, geographic origin and local practices [8]. Varying fermentation conditions can lead to differences in flavour profile of cocoa from the same varieties, confirming the crucial role of this process in the cocoa flavour generation [8].

Harvested pods are split in half usually with a machete, and the beans removed either manually or mechanically. Fermentation is started naturally and is carried out in heaps or boxes. The "heaps" method consists of piling up beans underneath plantain leaves, covering the surface and bottom of the pile, see Figure 4. The heaps can be 60-120 cm in diameter and are manually turned after 2-3 days to allow aeration of the mass and favour the growth of aerobic microorganisms.

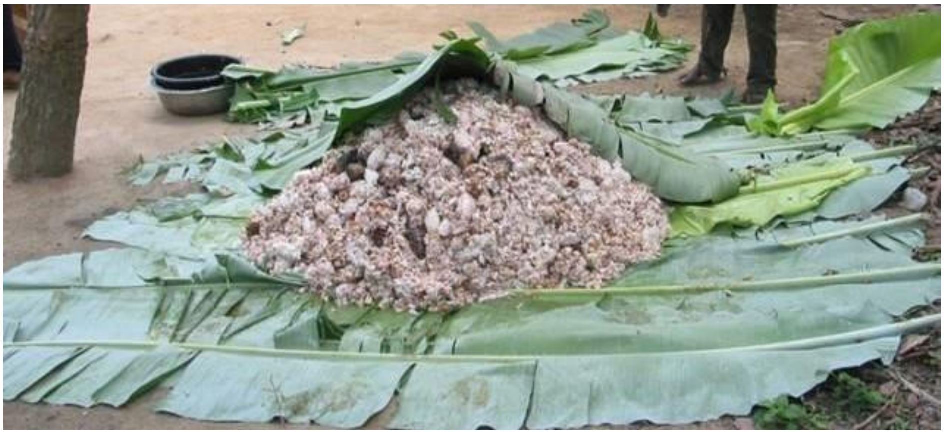


Figure 4. Beans piled up into "heaps"

The 'box' method involves fermentation of beans in large hardwood boxes holding up to 1.5 tonnes of beans, see Figure 5. These boxes have slatted bases or holes on the sides and base, which allow the sweating to drain away and aid access of air. Often these boxes are stacked in a descending order to allow easy transfer of beans to the box below and at the same time favour aeration of the cocoa mass. At the end of fermentation, the moisture content of the beans is reduced to less than 8% by drying. This process can be done naturally in the sun with regular turning, or with artificial dryers in closed rooms with temperatures not exceeding 60°C.

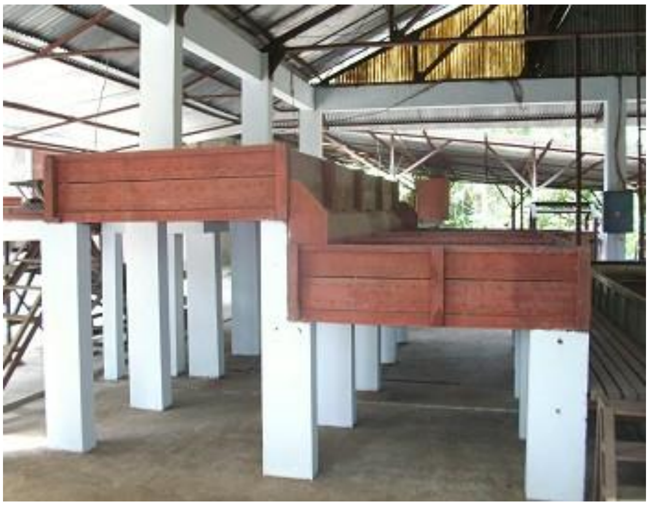


Figure 5. Boxes used for fermentation

1.2.1 Fermentation substrate and microbial succession

The fermentation substrate is the pulp surrounding the beans after removal from the pods. Cocoa pulp is a rich medium for microbial growth as it is characterised by a sugar content of around 9-13 % (w/w), high acidity due to the presence of diverse organic acids, and a protein content in the range of 0.4-0.6 % (w/w) [9]. The composition of the pulp for West African Amelonado cocoa is listed in Table 1, which may vary according to variety, origin and farming conditions.

Table 1. Composition of West African Amelonado cocoa pulp [9]

Component	% (w/w)
Water	82.60
Glucose and fructose	6.80
Sucrose	4.35
Plant and cell wall polymers	2.81
Citrate	1.31
Protein/peptides	0.57
Free amino acids	0.15
Fat	0.45
Metals	0.24
Vitamins	0.05

Microorganisms play a very important role in cocoa fermentation and up to 30 different species of bacteria have been found in fermented cocoa beans [10]. Fermentation carried out in aseptic conditions does not cause any significant biochemical changes in the cocoa mass, with no production of flavour precursors, leading to poor quality products, and thus, confirming the importance of the microflora in this process [10]. The acidic pH of the pulp (pH 3.6) and the low level of oxygen create a favourable environment for the growth of yeasts, as these microorganisms can metabolise the carbohydrates from the pulp under both aerobic and anaerobic conditions by mainly converting sugars into ethanol and CO_2 [5]. Yeasts release pectinolytic enzymes that break down the cement between the walls of the pulp cells resulting in the release of juice that drains away as "sweating" and formation of void spaces into which air percolates. Their growth is inhibited by an increase in pH due to citric acid metabolism and a high level of ethanol [5]. Lactic acid bacteria find a favourable environment in these conditions, and their growth reaches a peak around 36 hours after the fermentation has started [5]. These bacteria metabolise glucose with the release of lactic acid, ethanol, acetic acid, glycerol, mannitol and CO_2. As the oxygen level increases due to disappearance of the pulp and the temperature rises above 37° C, acetic acid bacteria become the main microorganism and they reach their maximum number at around 88 hours after the start of fermentation [5]. These bacteria oxidise ethanol to acetic acid and further oxidise this compound to CO_2 and water. The exothermic reactions of acetic acid bacteria cause the temperature of the fermenting mass to raise up to 50º C or more. The penetration of acetic acid into the beans kills the embryo and lowers the pH. The number of these bacteria starts to fall after 3 days of fermentation, and they completely disappear after 5 days [5]. Aerobic, spore-forming bacteria can be found during the first three days of fermentation, and they subsequently grow to become the dominant microorganisms in the fermenting mass, representing over 80% of the whole microflora [5]. Filamentous fungi are present at low levels throughout the fermentation, mostly on the surface of the fermenting mass where there is a greater air circulation and cooler temperatures. Under fermentation conditions compounds such as 2,3-butanediol, pyrazines, acetic and lactic acid are produced by these bacteria. It is thought that they may be responsible for the acidity and off-flavours of fermented cocoa beans [5]. A graph showing the microbial succession during cocoa bean fermentation is shown in Figure 6

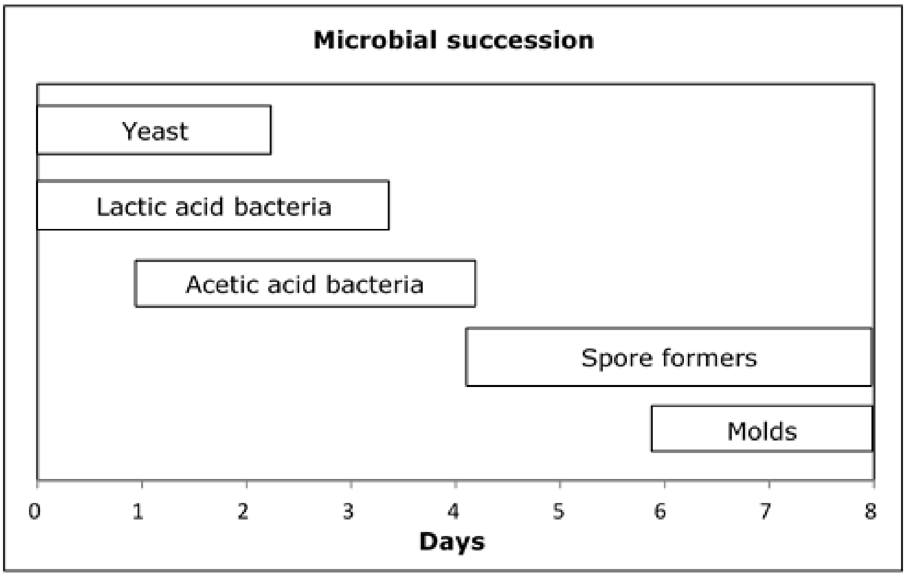


Figure 6. Microbial succession in cocoa bean fermentation [5]

The pulp amount dictates the duration of fermentation, as cocoa varieties with a low amount of pulp, such as Criollo, require a shorter fermentation which last usually 2-3 days, while Forastero varieties which have a high amount of pulp are fermented for 5-7 days. The amount of pulp can also vary during the main harvest seasons, and if it is too high can lead to over-fermented beans.

1.2.2 Flavour generation and chemical changes in the cocoa bean

Water and fat are the main components of unfermented cocoa beans. A significant amount of carbohydrates, proteins and polyphenols is also present. These values differ according to variety, season and farming conditions. The general composition of cocoa beans is shown in Table 2.

Table 2. Composition of unfermented cocoa beans [11]

Component	% w/w
Water	32.0-39.0
Sugars	2.0-3.0
Fat	30.0-32.0
Protein	7.0-10.0
Polyphenols	5.0-7.0
Starch	4.0-6.0
Acids	1.0
Cellulose	2.0-3.0
Theobromine	2.0-3.0
Caffeine	1.0

The major classes of seed proteins in unfermented cocoa beans are the vacuolar storage proteins albumins and globulins [12]. The globulins fraction is degraded by endogenous aspartic endoproteases and carboxypeptidases with the release of hydrophilic peptides and hydrophobic amino acids [7]. The albumins fraction acts as a protease inhibitor and therefore is only marginally degraded during fermentation [12]. Hydrophobic amino acids and hydrophilic peptides react with reducing sugars during roasting to generate cocoa aroma compounds. The pH plays a very important role in fermentation. At pH values of around 3.8 (the optimum for aspartic endopeptidases), the release of hydrophobic oligopeptides is favoured, while the production of free amino acids is negatively affected [5]. On the other hand, when the pH is close to 5.8 (the optimum for serine exopeptidases), an increase in hydrophilic oligopeptides and hydrophobic amino acids is observed. If the pH drops below 4.5 too soon, the production of flavour precursors is reduced and an over-acidic product is formed [5]. Sucrose is hydrolysed by an endogenous invertase to glucose and fructose. The amount of these reducing sugars can increase up to three-fold during fermentation [13]. The optimum pH for the activity of invertase is 4.5 [14], further confirming the importance of pH during fermentation.

Polyphenols in cocoa beans can be classified into three main groups: catechins or flavan-3-ols (ca. 37%), anthocyanins (ca. 4%) and proanthocyanidins (ca. 58%) [8]. Anthocyanidins are hydrolysed to cyanidins by the action of glycosidase enzymes, which results in bleaching of the purple colour and release of reducing sugars which are considered as important flavour precursors [2]. As the oxygen levels in the cotyledon increase, the endogenous polyphenol oxidase is activated resulting in the oxidation of polyphenols leading to the production of quinones, which can polymerise with other polyphenols or form complexes with amino acids and proteins [2]. As a result of these processes, the beans acquire a brown colour, and astringency and bitterness of the beans is reduced due to the formation of high-molecular-weight insoluble compounds.

Starch levels do not change during fermentation [15]. The concentration of citric acid decreases during this process, while the levels of lactic and malic acid are higher after fermentation and drying. Tetramethyl pyrazines, volatile alcohols, esters and aldehydes are also formed during drying. These compounds can be produced via microbial synthesis [2]. A schematic view of the chemical changes occurring in the beans during fermentation is shown in Figure 7:

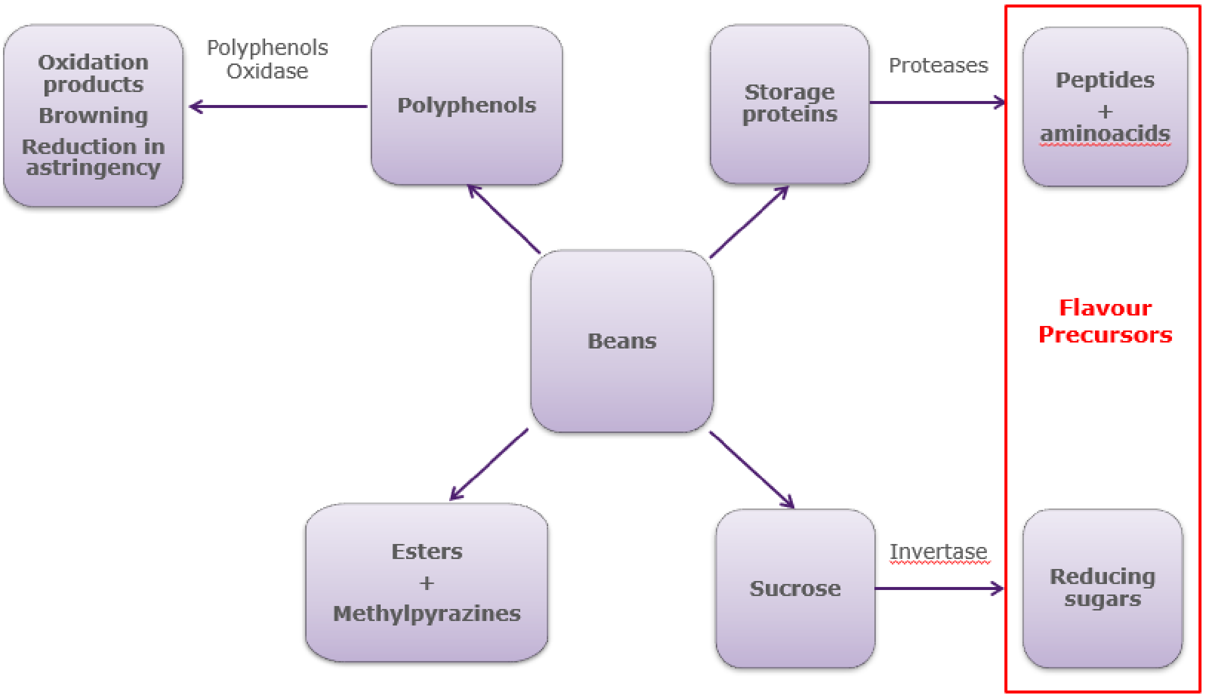


Figure 7. Chemical changes in cocoa beans during fermentation

1.2.3 Degradation of cocoa beans storage protein during fermentation

The levels of amino acids and peptides increase during fermentation due to partial degradation of cocoa bean storage proteins [16, 17]. During this process, a high amount of hydrophobic free amino acids such as leucine, phenylalanine, alanine and tyrosine are released, while the concentration of acidic amino acids is reduced [16-18]. Comparative 1D-SDS-PAGE analyses of fermented and unfermented beans have revealed that the amount of proteins is reduced during fermentation. The globulins are almost quantitatively degraded during fermentation, while the degradation of albumins is less pronounced during this process [12, 19, 20], as these proteins act as protease inhibitors. However, there is some discrepancy as to the extent of albumin degradation during fermentation reported in the literature [12, 19, 20]. Proteolysis of storage proteins occurs in the early stage of fermentation and after three days the majority of globulins are degraded [19]. After this initial phase, protein degradation is considerably reduced, probably due to the release of polyphenols and their subsequent complexation with the remaining proteins [19]. Proteins at 33, 17 and 11 kDa have shown increased levels during fermentation. However it has not been confirmed whether these peptides are synthesised during fermentation or are degradation products of larger polypeptides [19]. It has also been reported that proteins at 44.3, 46, 46.5 kDa may be formed from a 47.1 kDa protein in the globulins fraction during fermentation [20, 21]. The amino acid sequence of the protein at 44.3 kDa is 21 residues shorter than the sequence of the protein at 47.1 kDa, confirming that the protein at 44.3 kDa is a degradation product of the globulin protein at 47.1 kDa [20]. The proteins at 44.3 and 31 kDa undergo glycosylation during fermentation, while phosphorylation on the 47.1 kDa globulin protein has been observed [20]. The HPLC-UV profile from unfermented and fermented beans has shown that hydrophilic peptides are released from the proteolysis of cocoa storage proteins in the first three days of fermentation [22]. The peptide profiles of under-fermented cocoa beans from days 1 to 3 of fermentation revealed similar patterns [22], suggesting that the main proteolytic activity occurs at early stage of fermentation, confirming previous findings from Lerceteau *et al.* [19].

Cocoa-specific aroma precursors can be generated from autolysis at pH 5.2 of cocoa acetone-dry powder extracted from unfermented cocoa seeds [7]. The HPLC-UV of these extracts showed that mostly hydrophobic free amino acids and hydrophilic peptides are produced under these conditions. Two different methods of proteolysis led to the detection of cocoa and/or chocolate aroma from the roasting of the proteolysis products in the presence of reducing sugars [7]. The HPLC-UV chromatographic profile of *in-vitro* proteolysis products at pH 5.2 was similar to the chromatographic pattern of peptides extracted and partially purified from fermented cocoa seeds [7]. An endoprotease (an aspartic endoprotease with a pH optimum at 3.5) and a carboxypeptidase (pH optimum at 5.8) are present in ungerminated cocoa seeds. The pH plays an important role in the formation of cocoa flavour precursors, as there is no release of cocoa-specific flavour precursors from the incubation at pH 3.5 of acetone-dry powder extracted from unfermented cocoa bean [7]. Under these conditions many hydrophobic peptides were formed as a result of storage proteins degradation, while a small number of free amino acids were released. The digestion of these hydrophobic peptides with carboxypeptidase A from porcine pancreas generated mixtures of hydrophilic peptides and hydrophobic amino acids, which were similar to those obtained from the incubation at pH 5.2. As roasting of mixtures of amino acids with a distribution similar to the amino acids range present in fermented cocoa seeds did not generate any cocoa aroma, the authors concluded that “the essential cocoa-specific aroma precursors are among the hydrophilic oligopeptides” [7]. These results indicated that the endoprotease generates hydrophobic peptides which are turned into hydrophilic peptides by the action of the carboxypeptidase [7]. Therefore, cocoa aroma precursors are generated from the co-operative action of these two endogenous cocoa proteases.

In-vitro digestion by the aspartic endoprotease and the carboxypeptidase of ungerminated cocoa seeds of globular storage proteins from different crops has confirmed that the specific amino-acids sequence and structure of globulins present in cocoa beans determines the generation of cocoa-specific aroma precursors [23]. The flavour profiles of proteolysis products of globulins from hazelnut, sunflower and coconut roasted in the presence of sugars and deodorised butter evaluated by sensory analysis, showed significant differences from the typical aroma pattern generated by the proteolysis products of the globulins of cocoa seeds with the same proteases [23]. The amino acids and oligopeptides patterns resulting from the degradation of globulins from cocoa and different crops were considerably different [23]. Peptides with carboxyterminal arginine, lysine or proline are not easily cleaved by the cocoa seed carboxypeptidase [24]. Carboxyterminal hydrophobic amino acids are preferentially released by this enzyme, while the cleavage rate is significantly affected by the side chains of the neighbouring amino acids. A hydrophobic side chain of the adjacent amino acid residue favours the release of the carboxyterminal amino acid, as the degradation of serine-alanine is slower than the one of alanine-alanine [24]. This enzyme has a similar specificity to the one of carboxypeptidase A from porcine pancreas. Hydrophobic amino acids are preferentially released from both cocoa and porcine pancreas carboxypeptidases, and cleavage of carboxyterminal prolyl or basic amino acid residues cannot be performed by these enzymes. Strong acidification considerably reduces the cleavage rate of hydrophobic amino acid residues, while the release of acidic amino acids is favoured in these conditions, resulting in an unbalanced accumulation of the cocoa-specific aroma precursors [24].

Pepsin, chymotrypsin and an endogenous aspartic endoprotease can efficiently degrade vicilin extracted from cocoa seeds [25]. However, the reverse-phase HPLC-UV chromatographic pattern of the obtained proteolysis products showed some differences between the three endoproteases. More hydrophilic oligopeptides were obtained with chymotrypsin compared to the other two proteases. However, the action of aspartic endoprotease and pepsin on cocoa globulin resulted in different oligopeptide profiles [25]. Treatment of all peptide mixtures with carboxypeptidase A from porcine pancreas, released mainly hydrophobic amino acids. However, the amino acid profiles varied between the peptide mixtures digested by different proteases, as alanine and leucine were preferentially released from the cocoa globulins treated with pepsin and aspartic endoprotease, while a higher amount of aromatic amino acids were released from chymotryptic peptides [25]. Cocoa aroma was generated by the roasting of the proteolysis products of cocoa globulins treated with aspartic endoprotease and carboxypeptidase [25]. A less pronounced cocoa aroma was obtained by the roasting of peptic peptides post-treated with carboxypeptidase, while the roasting of peptides treated with chymotrypsin and carboxypeptidase did not generate cocoa aroma [25]. These results confirmed that the generation of cocoa aroma is dependent on the amino acid sequence of cocoa globulins and the specificity of an endogenous aspartic endoprotease [25].

The cleavage sites of the cocoa aspartic endoprotease have been identified by *in-vitro* digestion of cocoa vicilin in the presence of this protease and subsequent analysis of the proteolysis products by MALDI-TOF MS [26]. However, it is not possible to accurately predict the peptides generated by the action of this enzyme in fermented cocoa beans, as these peptides are modified by the endogenous carboxypeptidase [26]. The cocoa aspartic endoprotease is a protein complex that contains a 30.5 kDa aspartic proteinase polypeptide and an associated 20.5 kDa trypsin inhibitor polypeptide [27]. Two aspartic proteases encoded by two distinct genes (TcAP1 and TcAP2) have been identified [28]. These proteases are relatively different as they show a homology of only 73% in terms of amino acid sequences [28]. The protease encoded by the gene TcAP1 is observed at a higher level during germination, while developing seeds contain a higher proportion of the protease encoded by the gene TcAP2 [28]. It is not known whether these two enzymes are both involved in the proteolytic activity during fermentation [28]. Cocoa bean storage proteins are degraded by this protease into peptides that range in size from 65 amino acids down to di- and tripeptides [27].

The complex HPLC-UV oligopeptide patterns generated by autolysis of acetone dry powder did not differ across six cocoa genotypes [29]. The use of an aspartyl protease inhibitor (Pepstatin A) suppressed oligopeptides production, thus confirming the key role of the aspartic endoprotease. As HPLC-UV does not provide the molecular weight of the detected compounds, none of the oligopeptides were identified. Furthermore, the selectivity of this technique is limited by the fact that different compounds may co-elute and therefore cannot be distinguished.

1.3 Post-fermentation processing and flavour formation

Cocoa beans may undergo alkalisation after fermentation and drying [1]. During this process an alkaline solution such as potassium or sodium carbonate solution is sprayed on the beans to raise the pH from 5.2-5.6 to near neutrality at 6.8–7.5. The aims of this process are primarily to change the colour and flavour of the cocoa powder or cocoa liquor, and at the same time provide an improved dispersibility or suspension of the cocoa solids in water [1]. Roasting of cocoa beans is carried out at temperatures between 120 and 150º C, for a period that varies from 5 to 20 minutes, depending on the nature of the beans and the required products. To remove the husks, winnowing is carried out on the beans. During this process the beans are poured into a machine which uses air and vacuum to separate the husks from the entire cotyledons. A series of chemical reactions also takes place in the roasting process resulting in further development of the original cocoa aroma [5]. Peptides and amino acids react with sugars during roasting through the Maillard reaction, generating a range of carbonyls which by fission or cyclisation can give rise to volatile compounds such as acetaldehyde, diacetyl and 5-hydroxymethyl furfural.

After roasting, a fluid paste (cocoa liquor) is obtained by grinding the nibs several times at high temperatures. On cooling, this paste solidifies into a cocoa mass, which is a dark material with astringent flavour deriving from polyphenols and tannins [5]. Cocoa butter and cocoa cake are obtained from pressing the cocoa mass after roasting. Cocoa butter is a pale-yellow fatty liquid with no cocoa flavour, while the cocoa cake is a strong-flavoured dark brown residue which is not palatable due to its high astringency and bitterness. This product is subsequently milled to obtain cocoa powder used by confectionary industries. Finished chocolate products are made by adding sugars, sweeteners, milk products, emulsifiers and cocoa butter to the cocoa mass. The amount of these added ingredients depends on the requirement of the final product [5]. In order to obtain high quality chocolate, conching is also required. This process is carried out at temperatures between 50 and 60° C for several hours, although it can last up to 5 days for specialist chocolate [5]. This process reduces the moisture level of the cocoa mass and removes certain undesirable flavour-active volatiles such as acetic acid, enhancing interactions between disperse and continuous phases [1]. Thanks to the prolonged mixing at elevated temperatures, conching promotes flavour development, resulting in a partly caramelised flavour in non-milk crumb chocolate. The particle size and the viscosity of refiner pastes are reduced throughout the process [1].

1.4 Proteomics

Proteins are large biomolecules which are composed of amino acid residues joined together by covalent bonds to form polymeric chains of varying length. The amino acid sequence of a protein is encoded in the genome of a cell or organism and is translated from messenger RNAs. Proteins play a crucial role in the life of an organism as these molecules participate in and regulate virtually every process within the cell. Each protein has its unique amino acid sequence which is a major determinant for the protein’s conformational structure and function.

The term proteome refers to the whole content of proteins, which are present in a sample (organism, cell, culture, tissue) at a specific time, under defined conditions. Proteomics is the science that studies a specific proteome at large scale, providing information on protein abundances, their structure and functions, post-translational modifications (PTMs) and variations. Performing any kind of proteomic analysis is a very challenging task, as the proteome varies between individuals of the same species and developmental stages, and in response to external factors. Besides, many proteins show some form of PTM [30]. These modifications can be permanent such as deamidation and oxidation, or temporary and reversible such as phosphorylation or methylation. PTMs cannot be predicted by the gene sequence but constitute a very important mechanism of functional regulation for living organisms or cells [30]. Proteins are biomolecules which show a very high degree of complexity, resulting from the different combinations of amino acid sequences within the protein structure. As a result, there is no universal buffer which can solubilise all proteins of a cell or organism.

1.5 Proteins separation techniques

1.5.1 SDS-PAGE

“The term electrophoresis refers to the movement of charged molecules in response to an electric field, resulting in their separation” [31]. In classical electrophoresis, when subjected to an electric field, proteins travel toward the electrode of opposite charge. Their migration rate (in units of cm^2/Vsec) depends on the physical characteristics of both the electrophoresis system and the proteins themselves. Protein migration in electrophoresis is affected by temperature, composition and concentration of the buffer, charge, size and shape of the proteins [31]. Polyacrylamide gels work as size-selective filters when electrophoresis is carried out, allowing smaller proteins to migrate faster than larger proteins, when subjected to an electric field. Small proteins and oligopeptides are best separated on gels with a high percentage of acrylamide, while the optimal separation conditions for larger proteins are achieved by decreasing the percentage of acrylamide in the gel. For mixtures of small and large proteins, gels with an acrylamide percentage gradient are also available. The gel is typically mounted between two buffer chambers in a vertical orientation, and the samples are applied on sample wells generated on the top of the gel, see Figure 8 [31].

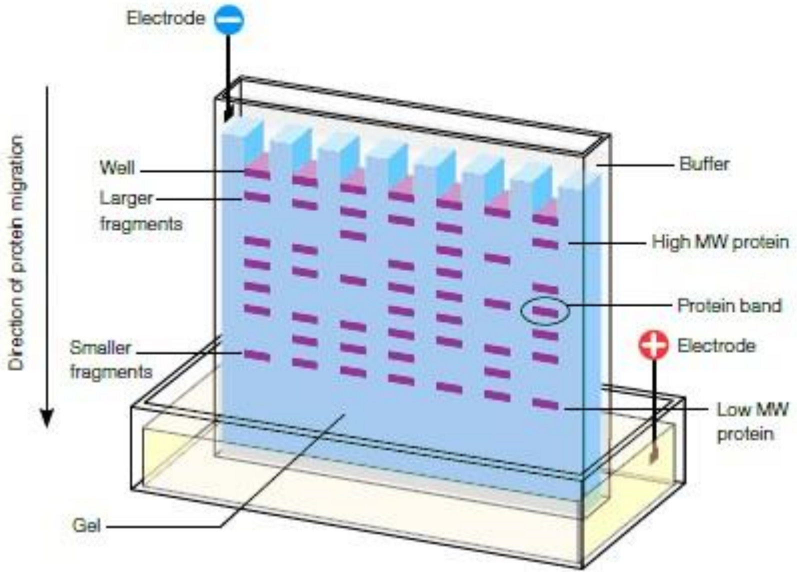


Figure 8. Schematic view of electrophoretic separation in a polyacrylamide gel [31]

In the presence of SDS and other denaturing agents, proteins become denatured and as a result lose their complex conformation (secondary and tertiary structures). Moreover, SDS binds non-covalently to proteins imparting an overall negative charge and a similar charge-to-mass-ratio to all proteins in a mixture [31]. SDS-bound proteins assume a rod-like shape instead of a complex tertiary conformation, see Figure 9. As a result, these proteins migrate in gels mainly according to their size, enabling molecular weight estimation by comparison with reference protein markers. After the electrophoretic run, gels are stained with dyes that bind to proteins so that protein bands can be visualised.

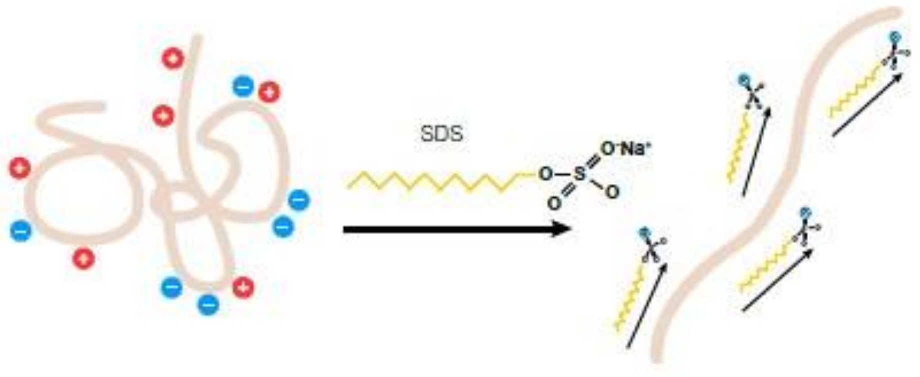


Figure 9. Effect of SDS on the conformation and charge of a protein [31]

1.5.2 Liquid chromatography

Chromatography is a physical separation method in which the compounds to be separated are distributed between two phases, one of which is fixed (stationary phase), while the other (mobile phase) moves in a defined direction [32].

The choice of the mobile phase is aimed at maximising the selective distribution of the compounds in a sample mixture between the mobile and stationary phase. Those components that strongly interact with the stationary phase move at a relatively slow rate through a chromatographic column. In contrast, components which have a weak affinity for the stationary phase spend more time in the mobile phase and as a result travel rapidly through the chromatographic column [32]. These differences in mobility result in the separation of sample components into discrete bands that can be analysed qualitatively and/or quantitatively (see Figure 10).

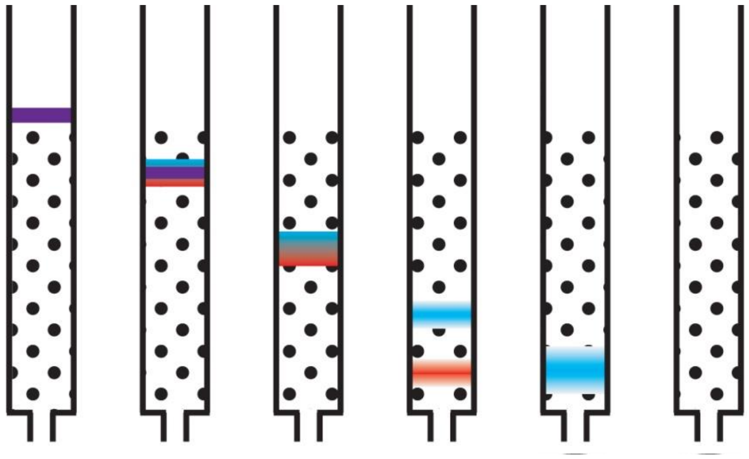


Figure 10. Chromatographic separation process for liquid chromatography

In liquid chromatography the mobile phase is a liquid whose composition can change (gradient) or remain constant (isocratic) over the analytical run. The stationary phase can be either liquid (partition chromatography) or solid (adsorption chromatography). The most common stationary phases used for proteomics analysis are C_{18} or C_8 alkyl chains bonded to the silica surface of the column, which constitute non-polar, "reverse phase" stationary phases. When using these stationary phases, the separation is mainly based on the hydrophobicity of the compounds analysed, with the most polar analytes eluting earlier than the less polar ones. Standard HPLC columns have internal diameters of 2.1-4.6 mm and flow rates of 0.2-1.0 ml/min. For proteomics applications, nano-LC columns with internal diameter of 75 µm are commonly employed. Thanks to the small internal diameter of these columns the analytes bands are more concentrated resulting in a dramatic increase in sensitivity. Ionisation in ESI is also much more efficient and sensitive at the typical low flow rates of nano-columns (150-300 nl/min).

1.6 Mass spectrometry

Mass spectrometry is a technique that "separates ions according to their mass-to-charge ratio (*m/z*) and detects them qualitatively and quantitatively by their respective *m/z* and abundance" [33]. The main components of a mass spectrometer are an ion source to generate gas-phase ions, a mass analyser to separate the gas-phase ion and a detector. The ion source can operate at atmospheric pressure or under vacuum, depending on the ionisation mechanism, while the mass analyser and the detector operate under high vacuum conditions. For this project, a Thermo Scientific Q Exactive and a Fusion orbitrap mass analysers have been used. Both instruments incorporate an orbitrap mass analyser, while the sample ionisation is carried out by ESI.

1.6.1 ESI

ESI is an ionisation technique which is part of a group of methodologies known as atmospheric pressure ionisation. In ESI, ions in a liquid eluent are sprayed/transferred into the gas phase prior to entering the mass analyser. For example, LC eluents enter a capillary metallic tube called "sprayer" which has an internal diameter of approximately 0.1 mm. A high voltage (usually between 2 to 6 kV) is applied to the capillary and as a result a potential difference between the sprayer and the sampling cone is formed [34]. The voltage creates an electrochemical cell in the interface which causes charge separation in the liquid droplets and accelerates the charged droplets from the sprayer towards the sampling cone. Based on the structure of the analyte of interest, positive and/or negative ions can be formed. The voltage applied to the capillary has typically the same polarity as the ionisation mode selected, while the sampling cone has an inverse polarity to the capillary, to attract the formed ions.

As the charge density in the charged droplets increases due to solvent evaporation, coulombic forces are also raised up. The point when these forces match the surface tension of the eluent is named the "Rayleigh Instability Limit". As the number of same-charge ions further increases, coulombic repulsion overcomes the Rayleigh limit and droplet fission occurs. According to Dole *et al.* [35], further coulombic explosions occur in the product droplets, which result in the formation of smaller and smaller droplets until individually charged analyte ions are formed. An alternative mechanism, the "ion evaporation model" [36] suggests that as the droplet reaches a certain radius the field strength at the surface of the droplet becomes large enough to assist the field desorption of gas phase molecular ions. These ions are then directed into the mass analyser through a series of focusing lenses.

ESI is a soft ionisation technique, which brings about only little fragmentation. Small molecules (<500 Da) are usually singly charged, while multiply charged ions are formed from larger molecules like proteins and peptides. A schematic view of the ESI process is shown in Figure 11.

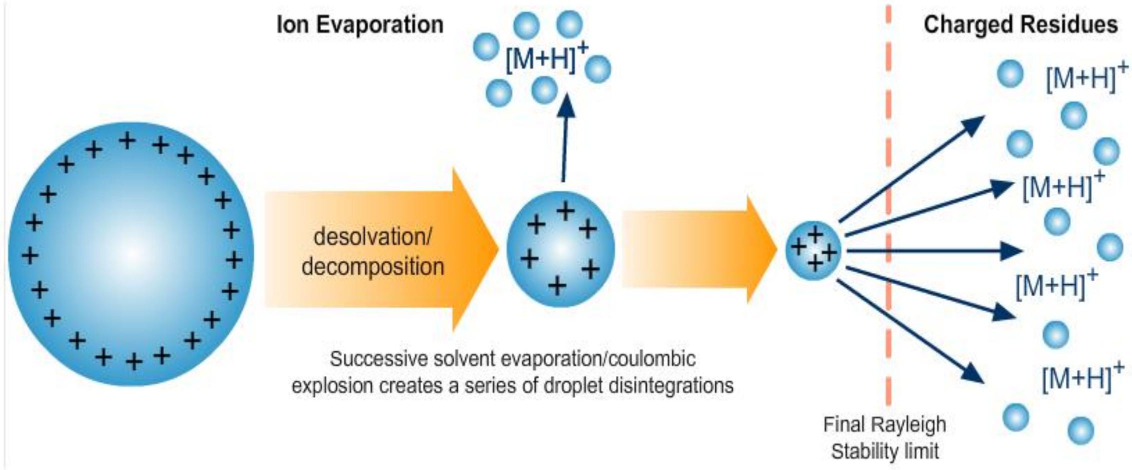


Figure 11. Schematic view of the ESI process [34]

1.6.2 Full scan mode data acquisition (MS)

A mass spectrum shows the intensity of the ions on the ordinate versus the *m/z* on the abscissa. The height of the ions or peaks is directly proportional to their abundance. In a full scan acquisition mode the mass spectra are continuously acquired over a defined *m/z* range for a pre-set period of time, and a number of data points are acquired for each *m/z*. Instruments which can achieve a unit-mass resolution require a low number of data points per *m/z*, while high-resolution instruments acquire far more data points per *m/z* in order to provide a high resolution and accuracy. The data can also be acquired in centroid mode where only the centre of the peak is saved, to significantly reduce the size of the data file.

The term *mass accuracy* refers to "how close the mass measured by the mass spectrometer comes to the calculated exact mass of an ion" [33] and is calculated using the following equation:

Equation 1:

$$Mass\ accuracy\ (ppm) = \frac{measured\ exact\ mass - calculated\ exact\ mass}{calculated\ exact\ mass} \times 10^6$$

The mass accuracy which can be achieved depends on the resolution of the instrument which is defined as follows: $R = \frac{M}{\Delta M}$ where M is the mass of an observed ion and ΔM is the peak width of the ion, and it refers to the ability of an instrument to separate narrow mass spectral peaks.

The software can calculate possible elemental compositions for an ion based on its accurate mass. As the mass accuracy increase the number of theoretical formulae which can be assigned to an ion decrease rapidly. As the mass of ions increases, a higher mass resolution is required to separate two ions with very similar masses. The number of hits can be reduced by considering the isotopic pattern of the ions, as the relative abundance of ions with *m/z* of M+1, M+2, M+3 can indicate the presence and number of specific elements such as Cl, S and Br. The number of C in a molecule can also be estimated from the relative abundance of the M+1 peak.

1.6.3 Tandem mass spectrometry (MS/MS)

Tandem mass spectrometry can provide further information on the elucidation of the molecular structure of a compound. With this technique the ions of interest (precursors ions) are isolated and fragmented, to produce fragment (product) ions. An MS/MS spectrum ideally shows the precursor ions together with the product ions. The most common dissociation mechanism is called CID. With this technique ions accelerated through the application of an electrical potential are collided with an inert gas. The high kinetic energy of the ions resulting from the collision with the gas is converted into internal energy which can break chemical bonds generating fragmentation of the molecular ions. HCD is a variation of CID as it uses a higher RF voltage to keep the fragment ions in a C-trap and is specific to Orbitrap mass analysers. An advantage of HCD is that this fragmentation technique allows detection of low *m/z* ions, which are usually not detected when using CID [37]. The fragmentation ions generated are recorded and an MS/MS spectrum is obtained.

The fragmentation pattern of a molecule reflects its molecular structure, as chemical bonds which require lower energy are more easily broken, and fragments which are more thermodynamically stable are preferentially formed. As a result, the interpretation of the MS/MS spectrum provides useful information for the confirmation and identification of the analytes of interest. This is especially useful for peptides, as these compound are cleaved at preferential sites, see Figure 12 [38], which allows the determination of their amino acid sequences. The main fragment ions in a CID MS/MS spectrum are usually b-type ions which retain the charge on the N-terminus, or y-type ions which retain the charge on the C-terminus, see Figure 12 for nomenclature/definition. However, the peptides fragmentation pattern depends on the fragmentation technique, the mass analyser, and the peptide structure.

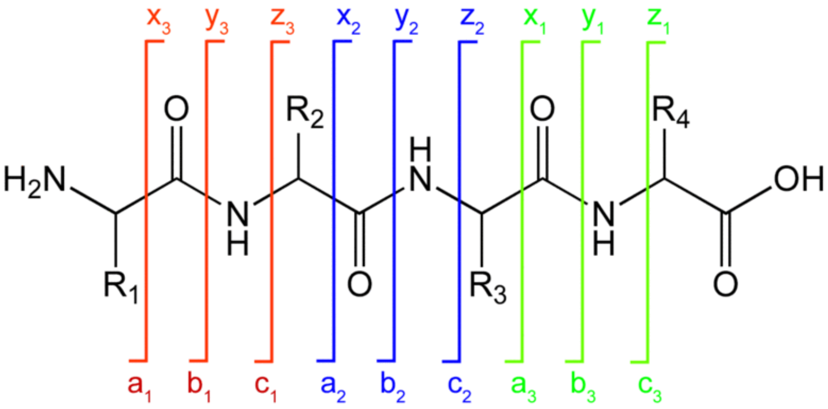


Figure 12. Peptide backbone fragmentation sites and nomenclature of the resulting fragment ions [38]

1.6.4 Orbitrap mass analyser

The orbitrap mass analyser was invented by Makarov and is based on a 'Knight-style' Kingdon trap with specially shaped inner and outer electrodes [39]. The orbitrap consists of two electrodes: an outer electrode, which is split in half by an insulating ceramic ring and a central electrode called the "spindle" around which ions are forced to move in a spiral trajectory [40], see Figure 13.

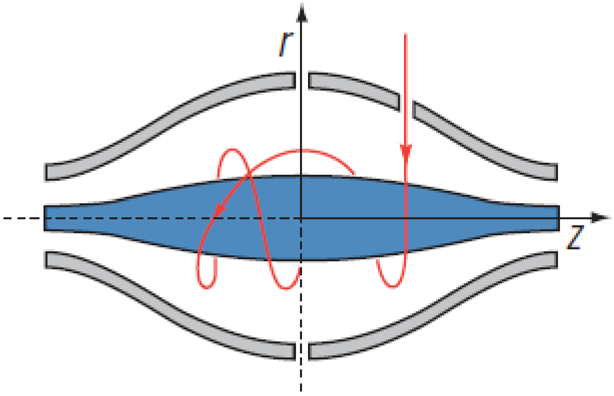


Figure 13. Orbitrap cell and ion trajectories [40]

The orbitrap operates in a pulsing mode, therefore a C-Trap is required to store ions generated from a continuous source such as a liquid chromatography. Ions produced in the API source are transferred into the C-Trap through a series of lenses which can focus the ion beam thanks to the application of RF and DC voltages [40]. Ions trapped in the C-Trap are pushed through a slot in the inner electrode which is located orthogonally to the curved axis [40]. Ion ejection is achieved by ramping up the RF voltage and applying voltage pulses to the electrodes.

Ions enter the orbitrap as short packets at a position offset from its equator, and they start oscillating coherently around the central electrode. The frequency ω of the oscillations of the ions along the z axis is not dependent on initial energy, angles and positions, but only on the mass-to-charge ratio *m/z* and the instrumental constant k according to the following equation:

Equation 2: $\omega = \sqrt{\frac{\mathbf{z}}{\mathbf{m}}} \times \mathbf{k}$ [40]

The oscillating ions generate an image current on the two split halves of the outer electrode, encapsulating the orbitrap analyser, which can be detected. The oscillation frequencies for ions with different masses are translated into high-resolution and accurate readings of *m/z* by the application of a Fourier transform.

A schematic view of a Q Exactive mass analyser is shown in Figure 14.

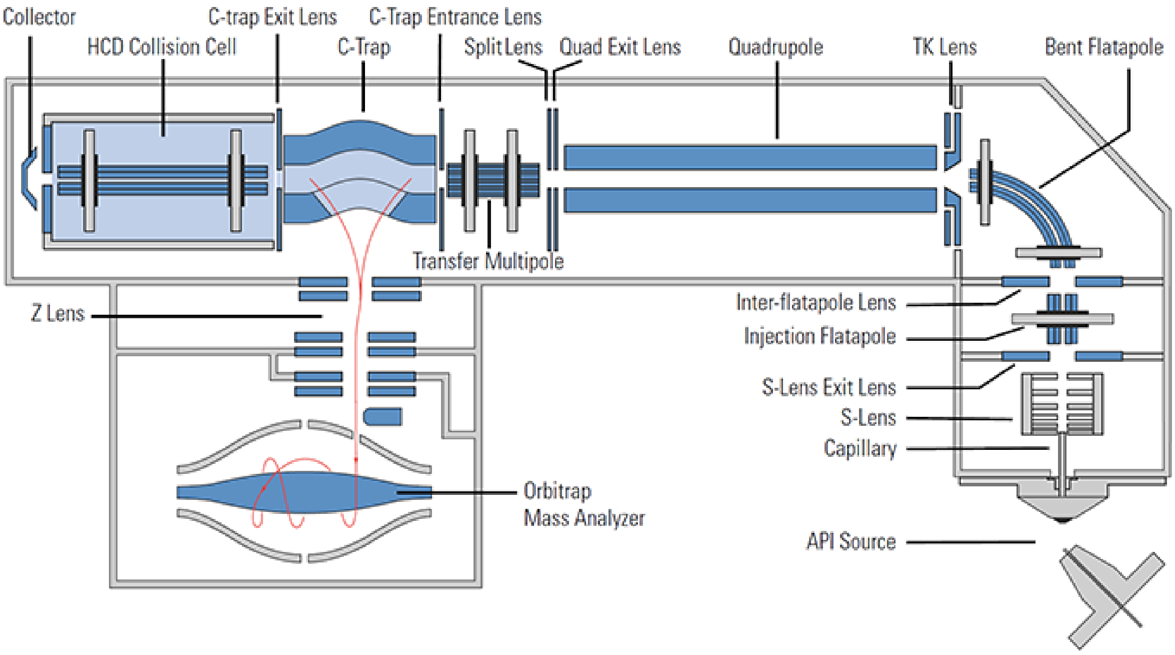


Figure 14. Schematic view of an orbitrap mass analyser [40]

Both the Q Exactive and Fusion orbitrap can measure the *m/z* value of an ion with an accuracy below 1 part per million (ppm). The Q Exactive can reach a resolution of up to 140,000 while the Fusion can achieve a resolution of up to 500,000. The high resolution and mass accuracy achieved by these mass analysers allows identification of compounds based on their accurate masses.

In order to carry out MS/MS analysis on the Q Exactive and Fusion orbitrap, the ions stored on the C-Trap can be transferred into an HCD cell and fragmented through an HCD mechanism. In this case helium gas is introduced in the cell and the ions are accelerated toward the gas by applying an electrical potential, which cause fragmentation of the molecular ions as described in paragraph 1.6.3.

1.6.5 Time of Flight (TOF) mass analysers

The working principle of TOF mass analysers is relatively simple when compared to other mass analysis devices. In this case ions generated from the ionisation source are directed towards an empty tube of known length and accelerated by the application of a voltage. Ions fly through the field-free path of the tube, and the time required to cover the whole length of the tube is dependent on their mass (m) and charge (z). A representation of the scheme of a linear TOF instrument is shown in Figure 15.

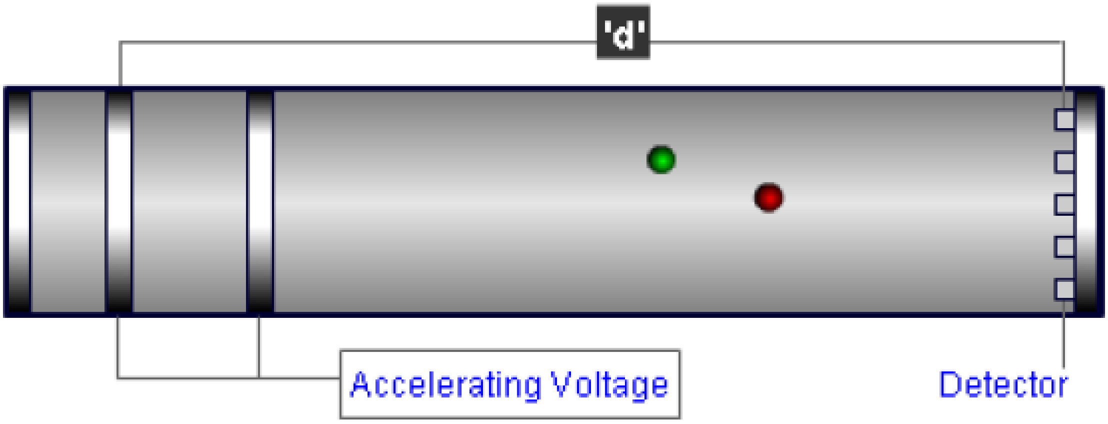


Figure 15. Linear TOF mass analyser [41]

For singly charged ions, the higher the mass the slower is the speed at which the ions travel through the tube to reach the detector. It is necessary that the ions reach the mass analyser as "packets", so that they can start flying at the same time so as to calculate the differences in arrival time.

The kinetic energy of each ion may be expressed as: $E_{kin}=mv^2/2=zeV_1=ezU$ [33], where *m* is the mass of the ion, *v* the final velocity, *e* the charge of an electron, *z* the number of charges on the ion. The velocity of an ion (v) can be calculated based on the length of the path (*l*) and the time (*t*) taken to cover such a distance using the following equation: $v=l/t$. Therefore, the *m/z* ratio can be measured as follows:

$m/z=(2eU/l^2)t^2$ [33]

The distribution of kinetic energy is not homogenous among all ions generated in the source, therefore ions with the same *m/z* but different kinetic energies will fly at different speeds, affecting negatively the resolution of the mass analyser. To overcome the loss of resolution due to the uneven distribution of kinetic energy of the ions, a reflectron is employed. This device consists of a series of rings at increasing potential, which create a retarding field with the same polarity of the generated ions, placed either before the ions are accelerated in the field-free tube or at the end of the ions path. Ions with higher kinetic energy will penetrate the reflectron deeper than ions with lower kinetic energy, allowing focusing of isobaric ions with different kinetic energy which would result in a much-improved resolution. A graphical view of a TOF with a reflectron is shown in Figure 16:

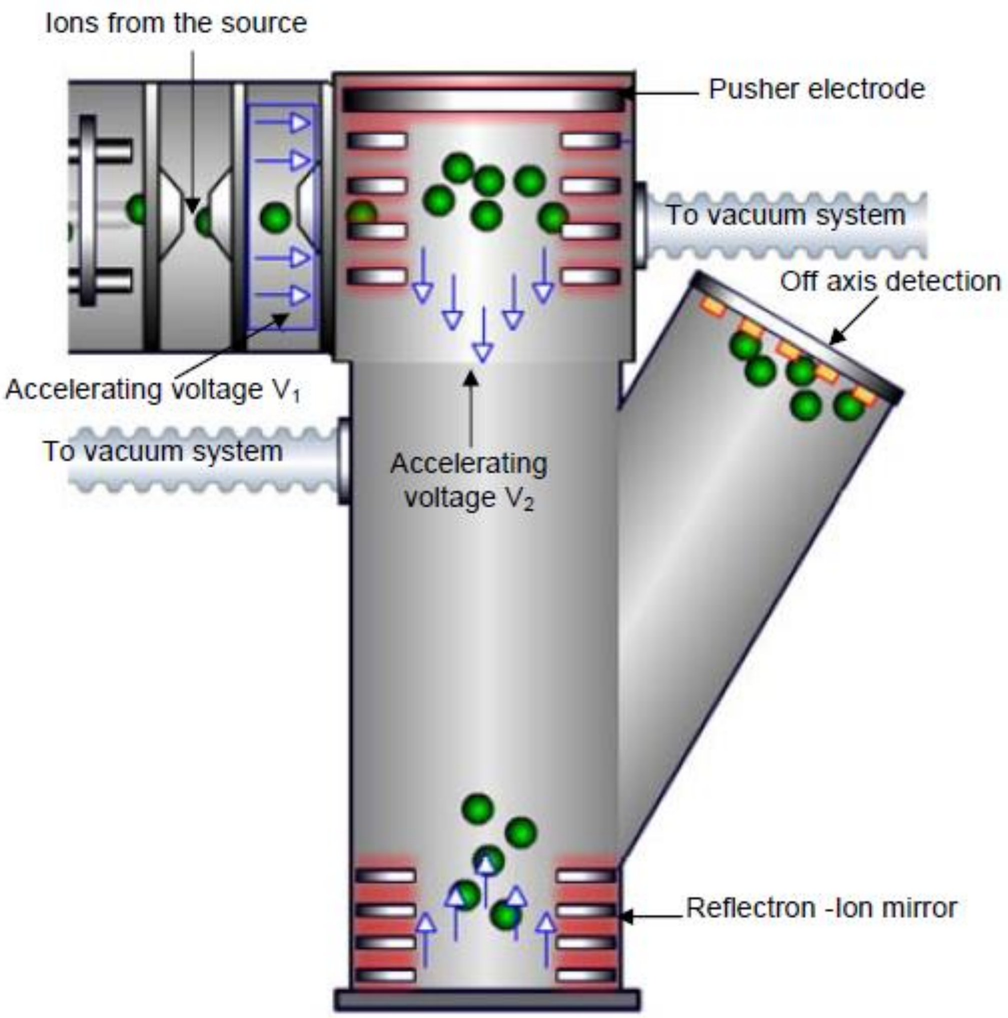


Figure 16. Reflectron TOF mass analyser [41]

TOF mass analysers can separate ions with m/z ratio up to 100,000 and can reach resolution of up to 12,000 [41].

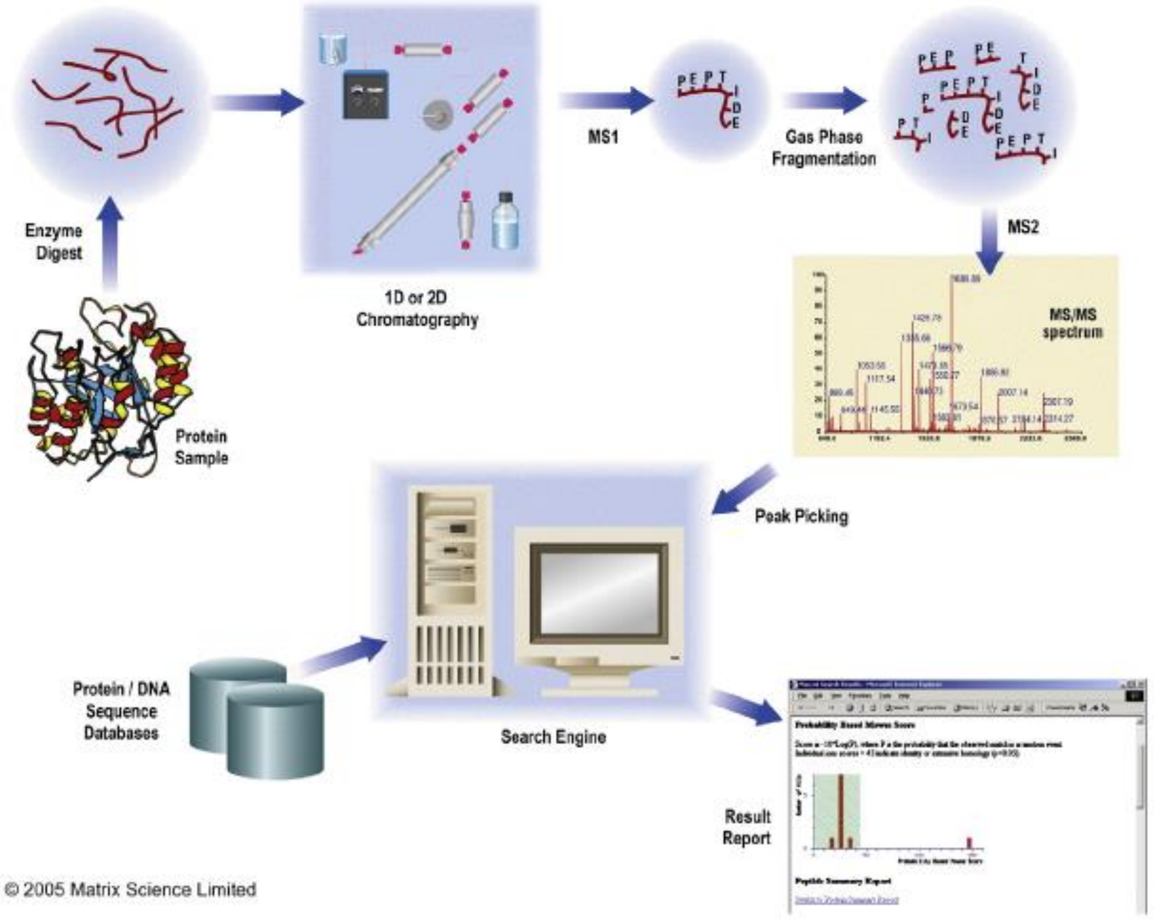
1.6.6 Mass spectrometry applied to proteomics

Mass spectrometry is a key analytical technique in proteomics. The main approaches employed in MS-based proteomics are bottom-up and top-down. In the bottom-up approach, proteins initially extracted from a sample or purified from a complex biological matrix, are digested with a highly specific enzyme, and the proteolysis products are chromatographically separated and analysed by mass spectrometry [42]. The initial task in this approach is to break the disulphide bonds within the protein chains to unfold the protein and aid denaturation. Once the protein is unfolded the protease can more easily access the entire chain of amino acids. Disulphide bonds are reduced by employing DTT, while the subsequent addition of IAA prevents the reformation of these bonds. The most common protease used for protein digestion is trypsin due to its specificity and low cost. This protease is highly effective on both native and denatured proteins, generates peptides of optimal size for mass spectrometry analysis, can work well in the presence of medium level of denaturing agents and detergents, which are the most common chemicals used to extract proteins. Trypsin is a serine protease, which cleaves the peptide bond at the C-terminal side of lysine and arginine amino acid residues. However, the cleavage will not occur if a proline residue is on the carboxyl side of the cleavage site.

After proteolysis, each peptide is then isolated and fragmented, and an MS/MS spectrum is acquired. When analysing samples with large mixtures of unknown proteins, peptides are usually analysed by MS/MS using a DDA method. In this case, the instrument performs an initial full MS scan, and subsequently peptide ions that are detected above a set threshold are isolated and fragmented. The most abundant ions are selected for DDA experiment, and the precursors ions are fragmented in order of decreasing intensity. A dynamic exclusion window is also employed to avoid redundant selection of ions which have already been selected for MS/MS fragmentation. Using this approach, the detection of low abundant ions co-eluting with other high intensity compounds is made possible. The duration of the windows is subjective to the width of the peak, as broader peaks require a larger dynamic exclusion time. Employing a DDA approach presents some limitations, as due to the stochastic nature of this acquisition the reproducibility of the detection of low abundant ions can be negatively affected. Next, MS/MS fragment ions information is recorded and uploaded into a software [38]. The masses of the peptides resulting from the *in silico* digestion of each entry in the database are calculated by the software, based on the specificity of the selected enzyme [38].

As previously mentioned in paragraph 1.6.3, peptides sequences can be easily determined from their matches to data obtained from protein database entries based on their specific fragmentation pattern, see Figure 12.

If there is a match between a calculated peptide and a detected one, the software calculates the masses of the expected fragment ions of the *in silico* peptide and compares these values to the experimental ones [38]. Since each MS/MS spectrum is related to one single peptide, it makes no difference whether the analysed sample is a single protein or a mixture. However, a peptide can be matched to more than one protein, unless the peptide is unique to a specific protein. Peptides shared by more than one protein are ranked according to their score which is based on the probability that the identified peptide is not a random event. Usually a minimum of two matched peptides is required to identify a protein with enough confidence. The bottom-up approach is preferable for complex samples whose protein composition is not known. A typical workflow for bottom-up identification of proteins by mass spectrometry is shown in Figure 17:



© 2005 Matrix Science Limited

Figure 17. Typical workflow for protein identification and characterisation using MS/MS data [38]

In top-down proteomics, intact protein ions or large protein fragments are ionised and fragmented in the gas-phase for mass spectrometry analysis directly with no prior enzymatic digestion [42]. A chromatographic or electrophoretic separation stage may also be involved at protein level. The masses of the protein fragment ions are compared to expected masses calculated from database entries.

1.6.7 Quantitative proteomics

Carrying out quantitative analyses in proteomics poses some challenges. The number of proteins in a plant or a biological sample is very large, and the type of the proteins present in these mixtures is not always known. As a result, quantitation with external calibration using reference proteins standards for large mixtures of unknown proteins it is not feasible, due to the complexity of the sample and the lack of reference standards for each of the proteins present in the sample. Using a single reference standard to perform absolute quantitation of all proteins in a mixture is also not appropriate, as the response factor would vary for each different protein. In order to overcome these constraints, computational methods have been developed to carry out protein quantitation by mass spectrometry. In this study, label-free protein quantitation was performed. This methodology can be divided into two types. The first is based on the changes in the intensity of the ion signals, while the second is based on the spectral count of identified peptides after LC-MS/MS analysis [43].

In the method based on the ion signal intensity, the areas of each detected peptide ion peak from the extracted ion current chromatograms are measured, and the intensity of a protein is measured as the sum of the areas of the peptides within that protein. The relative concentration of a proteins is determined as the ratio of the protein intensity between different samples.

In the spectral counting approach, the relative quantitation of proteins is carried out by comparing the number of identified MS/MS spectra from the same protein in each of the samples analysed by LC-MS/MS. This is made possible because as the protein concentration increases, a higher number of peptides belonging to that protein are detected, which then results in more spectra recorded for the same protein. Label-free protein quantitation methods can also be used to determine absolute protein abundance. One such approach is based on the protein abundance index or "PAI", which is defined *as:*

$$\mathbf{PAI} = \frac{\mathbf{N_{obs}}}{\mathbf{N_{theo}}} \qquad \textbf{Equation 3}$$

where N_{obs} is the number of detected peptides and N_{theo} is the number of theoretical peptides for a given protein [44]. The PAI index can be converted to the emPAI value defined as:

Equation 4

$$emPAI = 10^{PAI} - 1$$

Using this approach Ishihama *et al.* [44], could determine the absolute amount of 46 proteins in a mouse cell lysate. However, this quantitative method presents also some limitations, as although the number of identified sequences for a certain protein increases with his concentration, there may be proteins which are present at different levels but show the same number of identified peptides.

Other methods for quantitative proteomics include in vivo metabolic labelling such as SILAC or employing isobaric tags such as ITRAQ. SILAC methodology requires that cell cultures or plant of a specific group are grown in a metabolic medium that contains $^{13}C_6$-lysine and/or $^{13}C_6$-arginine, as trypsin cleaves at the C-terminus of these amino acids, resulting in all tryptic peptides from cultures grown in SILAC medium having at least one labelled amino acid [30]. Therefore, a constant mass increment will be observed in labelled samples versus non-labelled ones. This approach reduces considerably the error related to sample preparation and instrumentation, however, it cannot be applied to cultures, plant or other living organisms which cannot be grown in a medium.

Isobaric tag can be applied to samples which are not amenable to metabolic labelling, as these reagents are added to the samples after extraction and digestion. Tag include a mass reporter (tag) that has a unique number of ^{13}C substitutions and a mass normalizer that has a unique mass that balances the mass of the tag to make all of the tags equal in mass [30]. Isobaric mass tags also have a reactive moiety that crosslinks to primary amines or cysteines (depending on the product used). These tags are designed so that the mass tag is cleaved at a specific linker region upon high-energy CID (HCD), yielding the different sized tags that are then quantitated by LC-MS/MS. Using this methodology, a higher sensitivity and reproducibility of the quantitative results can be achieved. However, reagents are expensive and additional steps in sample preparation are required when compared to label-free quantitation.

1.7 Biomolecular analysis of cocoa beans

This chapter provides a review of the proteomics techniques applied to characterise the cocoa beans proteome, and references to proteomic analysis of other parts of the cocoa tree such as cocoa husk and embryo.

1.7.1 DNA sequencing of storage proteins and characterisation by electrophoretic techniques

During their development, plant seeds accumulate large amounts of storage proteins that serve as a source of nitrogen, sulphur and carbon compounds during seed germination. The major cocoa seed proteins are albumins (water soluble) and globulins (salt soluble). According to Voigt *et al.* [12], albumins and globulins fractions represent 52 and 43% of the total cocoa seed proteins. Other authors, however, have stated that the globulins storage proteins represent 23% of the soluble seed proteins, and the albumins 14.1 % [19]. The globulins fraction includes several polypeptides with different molecular weights [12, 19]. The predominant components of this fraction are polypeptides with apparent molecular weights of 47 kDa, 31 kDa and 14.5 kDa, as shown by SDS-PAGE analysis of the cocoa beans' globulins fraction [12, 19, 20]. It has also been reported that globulins prepared in the absence of an aspartyl protease inhibitor, pepstatin A, degrade to form two additional polypeptides with apparent molecular sizes of 28 kDa and 16 kDa, respectively [12]. However, there is a debate whether these two polypeptides are degradation products formed during extraction or genuine components of the globulins fraction, as Lerceteau *et al..* [19] have stated that there is no evidence that the polypeptide of 16 kDa is cleaved off from globulins during extraction. Therefore, this polypeptide may be an authentic component of the globulins fraction. "The polypeptides of 47 and 31 kDa are derived from a common cDNA precursor that translates to give a 566 amino acids polypeptide of 65 kDa in size" [45]. The N-terminus of this precursor contains a hydrophobic sequence with a site of cleavage which is predicted to be located 20 amino acids after the start. Next to this site, a high hydrophilic domain of ~110 amino acids is located, which is predicted to be cleaved off leaving a domain of approx. 47 kDa, consistent with the molecular weight of the polypeptide observed in the cocoa globulins fraction. The polypeptide sequence shows homology to globulins storage proteins in legumes and cotton [45]. Polypeptides at 47 and 31 kDa are also present in the globulins fraction of *Theobroma bicolor* and *grandiflorum* analysed by SDS-PAGE [46]. The intensity of both these bands was higher in *Theobroma bicolor* compared to *Theobroma cacao.* The light globulin chain at 31 kDa was the predominant polypeptide of the globulins fraction of *Theobroma grandiflorum* [46]. The globulins band at 47 kDa was only slightly detected in this species, whereas this band represented most of the globulins fraction of *Theobroma bicolor* and *cacao*. It has been suggested that these findings indicate a lower potential to produce flavour precursors for *Theobroma grandiflorum* and a higher potential for the generation of flavour precursors in *Theobroma bicolor*, compared to *Theobroma cacao* [46].

The 1D and 2D gel electrophoresis profiles of globulins from cocoa cotyledons belonging to genetically distant varieties and genotypes which had been reported to produce genotype-specific flavour characteristics have not shown significant visual differences [47]. Similar results were obtained from the 1D SDS-PAGE analysis of the globulins fractions of 43 cocoa seeds from different origins and genotypes [48]. According to the analytical methods used in these studies, the cotyledon storage proteins from various genetically different cocoa varieties are the same. It has been suggested that rather than the genetic background, other factors such the pulp composition could be related to aroma differences after fermentation [47]. However, although total protein contents of various cocoa genotypes have not shown major differences, the relative amount of each polypeptide in the samples analysed is not reported. In addition, gel electrophoresis shows only the apparent molecular weight of a protein and does not provide additional information on the amino acids sequence or post-translational modifications. This information can be obtained by mass spectrometry.

A polypeptide with apparent molecular weight of approximately 21 kDa has been shown to be the main component of the cocoa albumins fraction analysed by 1D SDS-PAGE [12, 19]. This polypeptide is derived from a cDNA precursor that translates to give a 221-amino-acid polypeptide of 24 kDa [49]. A hydrophobic signal sequence of 26 amino acids is located before the mature start of this precursor, and the molecular weight of the mature polypeptide would be 21 kDa [49]. The polypeptide sequence shows homology with sequences of the Kunitz protease and the α-amylase inhibitor family, and it is thought that its main function is to inhibit degradation of cocoa seed storage proteins from digestive enzymes of invading pests [49]. The albumin band at 20 kDa has also been detected in the 1D SDS-PAGE analysis of albumins from *Theobroma bicolor* and *grandiflorum*, although this band was less intense in *Theobroma bicolor* and its apparent molecular weight was also slightly lower in this species [46].

1.7.2 Enzyme activities in *Theobroma cacao* and related species

Cocoa endoproteases are stable during fermentation as only 50% of their activity is lost during this process [14], indicating a continuous proteolytic activity of these enzymes throughout fermentation. These findings provide an explanation as to the proteolytic activity observed in over-fermented beans. Aminopeptidases are considerably inactivated by fermentation, as their activity is reduced to only 5 % of the initial value after 2 days of fermentation. These enzymes are stable to sun and artificial drying. Carboxypeptidases remain active after sun and artificial drying. However, approximately 50 and 85% of their activity is lost after 3 and 4 days of fermentation, respectively. Cotyledon invertases are almost completely inactivated after 2 days of fermentation. The accumulation of sugars during fermentation might be limited by the low invertase activity and stability. The activity of polyphenol oxidases is reduced considerably during fermentation. These enzymes are also inactivated by sun and artificial drying. β-galactosidases, α-arabinosidases and α-mannosidases are not inactivated during fermentation or sunlight exposure and artificial drying [14].

Aspartic endopeptidases and carboxypeptidases in *Theobroma bicolor* and *Theobroma grandiflorum* have similar activities to those of *Theobroma cacao* [46]. Carboxypeptidases have a specificity for hydrophobic amino acids which does not significantly differ across the three species. However, a lower optimal pH for these enzymes is found in *Theobroma bicolor* compared to the other two species.

Although the enzyme activities among certain genotypes differ significantly, there is not a clear link between the key enzyme activities of under-fermented beans and their flavour potentials [50]. Therefore, the formation of flavour precursors during fermentation is not limited by the level of enzyme activities present in unfermented beans [50].

1.7.3 Characterisation of cocoa proteins by MS

An analysis of the proteomic profile of *Theobroma cacao* pod husk was carried out by initial separation of intact proteins using 2D gel electrophoresis and subsequent *de novo* sequencing of 4-sulfophenyl isothiocyanate-derivatized tryptic peptides of the excised gel bands using MALDI-TOF/TOF MS/MS [51]. Most of the identified proteins could be related to metabolism and energy, and a significant proportion was linked to pod growth and development processes [51]. A similar procedure was employed to perform proteomic analysis of *Theobroma cacao* embryos [52]. In this case the majority of the identified proteins were involved in genetic information processing, carbohydrate metabolism and stress response [52].

Two-dimensional electrophoresis is a technique which separates mixtures of proteins based on two distinct properties of proteins. In the first dimension, proteins are separated based on their isoelectric point, while in the second dimension the separation is achieved according to the molecular size of the proteins. This methodology provides a higher resolution when compared to one-dimensional electrophoresis, however, the whole protocol is considerably longer as two separation mechanisms are required, and each sample must be run on a dedicated gel, while multiple samples can be analysed on the same gel for one-dimensional electrophoresis. The use of specific software to spot differences between samples is also required for two-dimensional gel electrophoresis.

More than 1300 proteins were identified in white and translucent somatic embryos of *Theobroma cacao* analysed by LC-ESI MS/MS using a bottom-up shotgun approach [53]. A total of 25 proteins, among which β-glucosidase, NAD(P)-linked oxidoreductase and electron transfer flavoprotein were detected at a higher level in the white somatic embryos, whereas 35 proteins including cytochrome P450 and pathogenesis-related proteins were upregulated in the translucent somatic embryos [53]. A similar approach was employed to evaluate the proteomic profile of cocoa beans during development [54]. The authors reported a total of 887 identified proteins, although it is not clear whether this value refers to the combined list of proteins identified in all samples analysed. 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. Proteins involved in defence and stress were present at a higher level in mature seeds. Although a list of all detected proteins is provided in the supplementary file, the classification of the cocoa seeds proteins based on their function and abundance has not been carried out [54].

The amino acid sequence of the cocoa globulin subunit polypeptides detected by 1D SDS-PAGE at apparent molecular masses of 47 kDa, 31 kDa and 15 kDa could be localised on their common 66-kDa precursor sequence [55]. The characterisation of these polypeptides was carried out by MALDI-TOF MS analysis of tryptic digests of the bands excised from the gel. A similar procedure was used by Kumari *et al..* [20] to assess the amino acid sequences of vicilin subunits. The amino acid sequence of cocoa albumin has been characterised by MALDI-TOF MS and was found to be nine amino acid residues shorter than expected from its encoded DNA [56]. The molecular weight of this protein (20,234 Da) was identical in seven different genotypes representing the four cocoa varieties Criollo, Forastero, Nacional and Trinitario [56]. A polypeptide with a molecular weight of 8,515 Da was identified by LC-MS in unfermented cocoa beans [57]. The sequence of this polypeptide closely matched the internal sequence of the 2s albumin precursor, suggesting that the 9 kDa polypeptide is an albumin (fragment) from *Theobroma cacao*. Some degree of homology was found between this polypeptide and albumins of cotton, Brazil nut and sweet protein [57].

A large number of proteins with molecular weights ranging from 8 to 13 kDa and a cluster of peaks centred at 21 kDa, which was attributed to albumin, were observed in the mass spectra of the protein profile of the seed of *Theobroma cacao* obtained by MALDI-TOF MS [58]. The protein MS profiles of different cocoa varieties were similar. No proteins were detected in the husks, suggesting that this part of the bean does not contain water-soluble proteins. Albumin was present in both apical and cortical parts, at a higher level in the former. An increase in the average molecular weight of these proteins was observed during roasting, which could be a result of sugar addition through the Maillard reaction. The protein extraction in this study was carried out using a low ionic strength buffer, which does not solubilise membrane proteins and globulins, therefore the protein MS profile obtained for the sample analysed is limited to water-soluble proteins only. The findings of this study focus on qualitative protein profiles only, furthermore, the low resolving power of the mass spectrometer did not allow a clear separation of the glycated proteins.

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 of a total of 49 spots identified on the 2D gel [59]. The authors reported differences in the proteomic profiles of samples from different varieties, and samples from the same varieties grown in different countries. According to the authors, a vicilin subunit was specific to samples of CCN51 hybrids and the German Forastero cacao variety CD03 [59]. Two protein spots, which were identified as a degraded 17 kDa albumin subunit and the internal 15 kDa vicilin subunit, allowed differentiation of samples by MANOVA analysis based on geographical origin and variety. The authors stated that these proteins could be used as markers to assess the geographical origins and the different varieties [59]. This study evaluated only a limited number of proteins among the different samples, therefore did not provide a comprehensive characterisation of the proteomic profiles of the various samples analysed. The differences observed in the proteomic profiles of cocoa samples from different origins could also be due to environmental factors and agricultural practices used in different countries.

1.7.4 MS characterisation of peptides formed during fermentation

Buyukpamukcu *et al.* [60] claimed that two peptides with *m/z* values of 902 and 621 are formed during fermentation of cocoa beans. These peptides were detected by LC-MS analysis of fermented cocoa extracts and gave matches to sequences in cocoa globulins. The ion at *m/z* 621 was reported to be a hexapeptide (sequence SPGDVF) formed during fermentation from the ion at *m/z* 902, which was identified as a nonapeptide with the sequence APLSPGDVF. One of the limitations of the analytical method used by the authors is that the separation mechanism was not optimal for the separation of amino acids and peptides with slight differences in molecular size. As a result, products formed during fermentation might not have been visualised due to co-elution with other compounds.

Other short chain peptides formed during fermentation have also been identified by LC-MS [61]. The levels of these peptides were reduced after roasting, suggesting that these compounds might have reacted with sugars through the Maillard reaction, and undergone a series of further rearrangements ultimating in the production of volatile compounds. Among these peptides, 18 could be linked to sequences derived from both globulin and albumin, while 25 had sequences which matched albumin only. The authors claimed that these compounds are flavour precursors formed during fermentation, which could be added to food products to reproduce chocolate aroma. However, this claim has not been confirmed with a panel test or PAC (principal aroma compounds) analysis.

Further, peptides formed during fermentation have been identified and semi-quantified in cocoa beans of different geographic origin analysed at various fermentation stages [62]. The samples were analysed by reverse phase LC-ESI MS. Among these peptides, 25 could be matched to the sequence of globulins and 14 to albumins. Peptides which were not related to either globulins or albumins were also detected. This study showed that the level of peptides varied among the same cocoa varieties grown in different geographic locations, and among different varieties grown in the same geographic location. The quantitative data, however, are only semi-quantitative and were obtained by comparison to a dipeptide internal standard. Moreover, it is not clear whether the cocoa pods were grown and processed under controlled conditions. Therefore, the differences detected may be due to external parameters rather than genetic variations among the analysed genotypes.

A similar approach was employed by Caligiani *et al.* [63] to assess the effect of fermentation level and geographical origin on the distribution of peptides in fermented and unfermented cocoa beans from different varieties and geographical origins. Low amounts of peptides were found in slaty and under-fermented beans [63]. The authors stated that the ratio of vicilin to albumin peptides is higher in partially fermented beans compared to fully fermented beans, and therefore this ratio could be used as a fermentation marker to assess the quality of commercial beans. In this work, however, only a limited list of 35 peptides were evaluated in all samples analysed and a comprehensive characterisation of the peptide profiles of the fermented beans was not provided.

Recently, a procedure for the fractionation and concentration of aroma precursor extracts from well-fermented cocoa beans has been developed by Voigt *et al.* [64]. Cocoa aroma was produced when these extracts were roasted in the presence of sugars and deodorised butter, confirming that the extracts contained aroma precursors. MALDI-TOF MS and LC-MS analysis of the fractions producing cocoa aroma have revealed the presence of several peptides, whose amino acid sequences could be linked to cocoa globulins [64]. These peptides were mostly hydrophilic confirming the finding of previous works, which had reported that hydrophilic peptides were important cocoa flavour precursors [7]. The same authors have shown that the pH can affect significantly the type of peptides released from cocoa vicilin, as proteolysis carried out at pH 5.2 produced peptides with longer residues than at pH 4.8 [65].

Analysis of free peptides by LC-MS/MS of cocoa bean samples at different fermentation stages have revealed that peptides generated from the degradation of cocoa vicilin during fermentation are localised in different regions of the amino acids sequence of this protein [20]. The majority of the peptides were released during early stages of fermentation, and proteolytic activity could be observed up to 72 h from the start of fermentation. Several peptides which shared the same N-terminus but showed a different C-terminal were observed, confirming activity of endogenous carboxypeptidases [20].

Oligopeptides formed from the degradation of cocoa proteins during spontaneous fermentation have been extensively characterised by Souza *et al.* [66] employing LC-MS/MS. The results showed that during the early stage of fermentation longer peptides were predominantly released and subsequently degraded to shorter peptides as the fermentation progressed. The identified peptides could be linked to the action of both endo- and exopeptidases degrading mostly albumin and vicilin at both protein termini. The authors claimed to have identified over 800 peptides when combining the results of all samples analysed at different fermentation stages. This work focused on one variety of cocoa beans only and included di and tri-peptides as well.

A similar methodology was used to assess differences in terms of peptide profiles between 25 samples of different geographic origin and various degree of fermentation [59]. The authors stated that the number of identified peptides was correlated to the fermentation stages, as poorly fermented beans showed a lower number of peptides compared to fully fermented beans. The degree of fermentation was the factor that led to the main differences in the peptide profiles of the samples analysed, while no significant differences were found when taking into account the geographic origin only [59]. However, the authors did not report specific peptides markers that could be used to assess the quality of the fermented beans. In addition, there is no information as to the protocol used for fermentation of the samples analysed and treatment of the cocoa trees bearing the cocoa pod prior to fermentation. Therefore, the differences found could also be due to environmental factors or variation in fermentation practices among the different origins of the samples analysed.

Lab-based fermentation of cocoa beans carried out in sterile glass bottles has very recently allowed the identification of 449 peptides by LC-MS, ranging from 4 to 23 amino acid residues [67]. A total of 9 peptides derived from proteolysis of cocoa vicilin and formed only in the late fermentation stages showed a significant loss after roasting, which would suggest according to the authors that these peptides may be responsible for the generation of cocoa aroma [67].

1.8 Aims and objectives

The main aims of this project are to characterise the cocoa bean proteome and understand whether differences in flavour characteristics between cocoa beans from different varieties are also reflected in their proteomic profiles.

The initial phase of the project was based on the development of a methodology that would allow a comprehensive characterisation of the proteomic profile of cocoa beans using mass spectrometry-based techniques.

There are cocoa varieties, which produce different flavour profiles, however, it is not known whether differences in flavour profiles are also reflected in the proteomic profile. Therefore, the second phase of the project focused on the analysis of cocoa varieties with contrasting flavour characteristics and assessed whether qualitative and quantitative differences in the proteomic profiles of these varieties can be found. To minimise variability of the protein expression due to external factors, the cocoa varieties to be analysed would have to be grown under controlled conditions in terms of water intake, fertilisation, sun exposure and soil structure. A collaboration with the University of West Indies (UWI) was sought as this institution has the facilities to grow cocoa trees under controlled conditions. A total of four different genotypes were selected for the second phase of this project. The choice of genotypes was made based on the following criteria:

- Varieties available at UWI
- Varieties belonging to different genetic groups
- Varieties which show differences in flavour profile

- Varieties which cover the range of raw material used by commercial chocolate producers

Another objective of this project was to understand how proteins are degraded during fermentation, and what proteolysis products are formed during this process. As a result, a method was developed to extract and characterise free peptides from fermented cocoa beans by LC-MS/MS analysis. This methodology can be used for future work aimed at carrying out the characterisation and quantitation of free peptides released from cocoa beans proteins during fermentation.

2 MATERIALS AND METHOD

2.1 Chemicals and solutions

Petroleum ether 40-60 was obtained from Fisher Scientific, Loughborough, UK. All other chemicals and solvents were obtained from Sigma-Aldrich, Gillingham, UK, except where stated otherwise. Buffer 1 was an aqueous solution containing 5 mM sodium ascorbate, 2 mM EDTA and 10 mM Tris-HCl adjusted to pH 7.5±0.2 with the addition of aqueous 1 M NaOH (Fisher Scientific). Buffer 2 consisted of a 0.5-M NaCl solution containing 5 mM sodium ascorbate, 2 mM EDTA and 10 mM Tris-HCl adjusted to pH 7.5±0.2 with the addition of aqueous 1 M NaOH. The solubilisation solution consisted of aqueous 7 M urea, 2 M thiourea and 20 mM dithiothreitol. The wash solution was made up of cold (~4° C) aqueous acetone (80%; v/v) containing 5 mM sodium ascorbate.

2.2 Plant materials

2.2.1 Method development and characterisation of cocoa proteome

Cocoa seeds were from West African Amelonado ripe pods harvested at the Cocoa Research Centre of the University of West Indies, St. Augustine, Trinidad. The pods were stored at room temperature after harvest and air-freighted within four days to Reading, UK. The temperature of the pods was not controlled during shipping. Upon arrival, seeds were removed from pods, washed with sand and water to remove the pulp and stored at -80° C prior to analysis. Approximately 240 beans from 6 pods were combined.

2.2.2 Proteomics analysis of cocoa genotypes

Cocoa seeds were from four different genotypes of *Theobroma cacao* and a variety of *Theobroma speciosum,* harvested at the Cocoa Research Centre of the University of West Indies, St. Augustine, Trinidad, as described in Table 3:

Table 3. Selected genotypes for the initial phase of the project

Accession code	Genetic group	Flavour attributes
ICS 1	Trinitario	Fresh and brown fruity [68, 69]
ICS 39	With a strong Criollo ancestry	Nutty and caramel [70]
IMC 67	Forastero of Iquitos origin	Fruity [68, 69]
SCA 6	Contamana	Floral and fruity [68, 69]
Negative control	*Theobroma speciosum*	No cocoa flavour

To assess the effect of a different location on the proteomic profile of cocoa beans, samples from the genotype IMC 67 grown in two different fields were also provided. Samples from *Theobroma speciosum* were also selected to be used as negative control, since this species belongs to the same genus as *Theobroma cacao,* but its beans do not generate cocoa aroma. The main phenotypic traits of the *Theobroma cacao* genotypes selected for this project are listed in Table 4.

Table 4. Main phenotypic traits of selected genotypes

Accession	Origin	Yield	Bean	Pod
ICS 1	Trinidad	Low-medium	Purple colour with an elliptical shape. Average dry weight 1.27 g. Seed index* 58 **[71].**	Elliptical shape, moderate rugosity and moderate anthocyanin colour. Pod index* 19.6 **[71]**.
ICS 39	Trinidad	High	Purple colour with an ovate shape. Average dry weight 1.16 g. Seed index 68 **[71].**	Angoleta shape, moderate rugosity and anthocyanin absent colour. Pod index 21.3 **[71].**
IMC 67	Peru	Low	Dark purple colour with an ovate shape. Average dry weight 1.04 g. Seed index 109 **[71].**	Obovate shape, moderate rugosity and anthocyanin absent colour. Pod index 24.9 **[71].**
SCA 6	Upper Amazonian Forest	Very low	Purple colour with an oblong shape. Average dry weight 0.51 g. Seed index 133 **[71]**.	Angoleta shape, moderate rugosity and anthocyanin absent colour. Pod index 35.2 **[71].**

*The seed index is the number of seeds required to produce 100 g of dried beans, while the pod index is the number of pods required to produce 1 Kg of dried beans.

Cocoa pods were harvested from 6 different trees for each genotype. A detailed list of the number of trees and pods for each genotype is provided in Table 5:

Table 5. List of biological replicates provided for each cocoa genotype

Genotype	Tree	No. of pods	Genotype	Tree	No. of pods
SCA 6	T1	6	IMC 67 ICGT	T2	6
	T5	6		T3	6
	T10	6		T5	6
	T12	6		T6	6
	T14	7		T10	6
	T15	5		T13	6
Total		36	**Total**		36

Genotype	Tree	No. of pods	Genotype	Tree	No. of pods
IMC 67 CAMPUS	TREE A	6	ICS 1	T3	6
	TREE B	6		T5	6
	TREE D	6		T7	6
	TREE F	5		T8	6
	TREE I	6		T11	6
	TREE J	6		T12	6
Total		35	**Total**		36

Genotype	Tree	No. of pods	Genotype	Tree	No. of pods
ICS 39	T1	6	*Theobroma Speciosum*	TREE C	3
	T2	6		TREE D	3
	T3	6		TREE E	3
	T4	6		TREE H	6
	T5	5		TREE I	6
	T7	6		TREE J	6
Total		35	**Total**		27

Each pod of the same cocoa genotype was considered a biological replicate within the specified cocoa variety. 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. After the freeze-drying step, the beans were stored at -20° C prior to shipping. The freeze-dried beans were air-freighted to Reading with no control of the temperature during the shipment. Upon arrival, the beans were stored at -20° C prior to analysis. In order to obtain a representative sample for each cocoa variety, approximately 2 g of beans from each biological replicate within the same genotype were combined, and the remainder of the beans were left in their original container. The harvest time for the collected biological replicates spanned over a period of six months from November 2016 to May 2017 as shown in Figure 18.

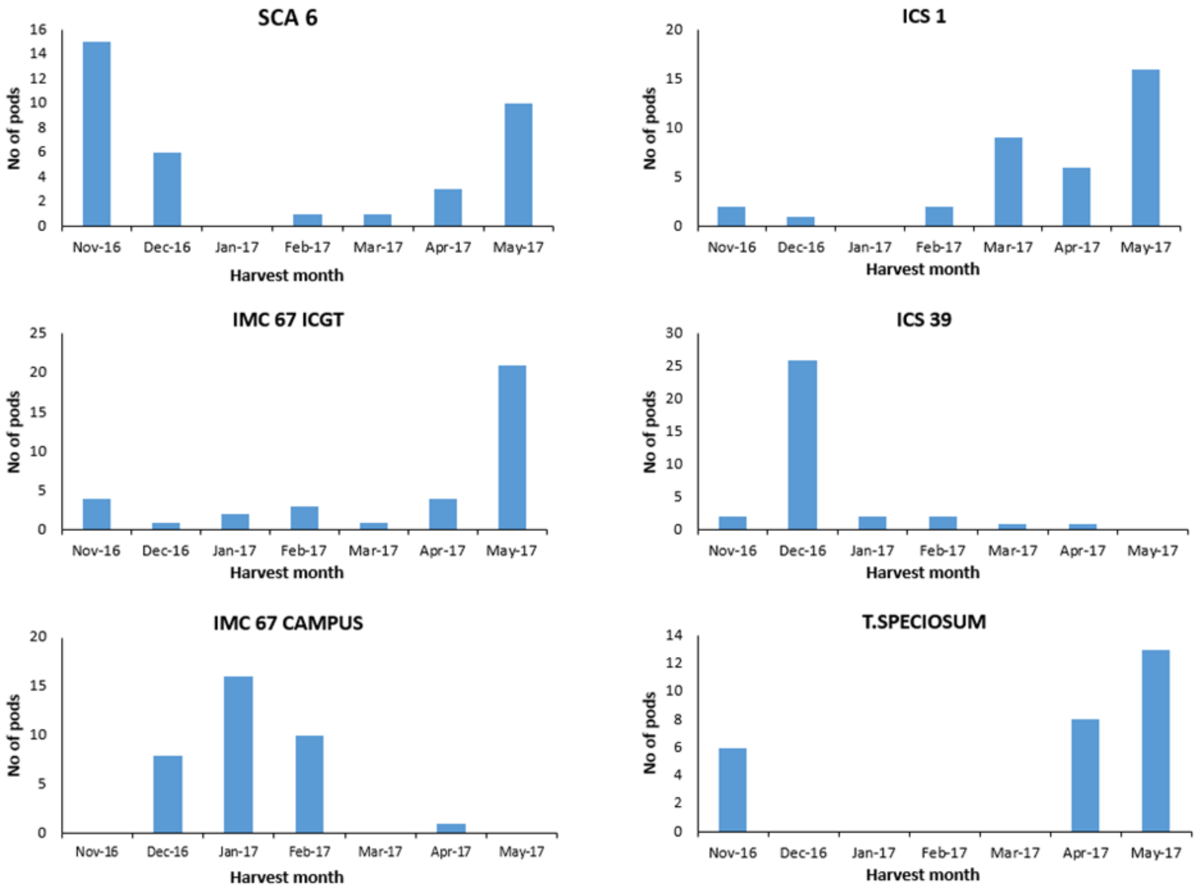


Figure 18. Graphical visualisation of harvest time for all biological replicates analysed to investigate the proteome changes dependent on genotypes

2.2.3 Method development for analysis of free peptides

Beans of the Amelonado variety were naturally fermented in cocoa farms in Ghana with the heaps method as described in section 1.2 and sun-dried after fermentation. The dried beans were stored at room temperature and shipped to Reading with no control over the temperature. Upon arrival, the beans were stored at -20° C prior to analysis.

2.3 Fat and polyphenol removal

2.3.1 Characterisation of cocoa proteome

The seeds were freeze-dried for 12-14 hours, snap-frozen using liquid nitrogen and subsequently ground using a mortar and pestle. These samples were used for all subsequent analyses. Fat from aliquots of approximately 500 mg were extracted with 10 ml of petroleum ether (boiling point 40-60°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 [12], this class of compounds was removed following a slight modification of a published method [23]. In brief, polyphenols were extracted from the defatted seeds with 10 ml of wash solution. The suspensions were vortexed for 1 minute and centrifuged at 3100 g for 10 minutes at 4° C. The supernatant was discarded and the extraction repeated twice. Residual water was removed by extraction with 10 ml of cold acetone. The sample was then dried under a stream of nitrogen, resulting in acetone-dried powder (ACDP). Taking into account that the fat and polyphenols content of dried cocoa beans is around 40% and 10% [11], respectively, it can be estimated that approximately 250 mg of ACDP was obtained from 500 mg starting material.

2.3.2 Method development and proteomic analysis of cocoa genotypes

Initial analyses were carried out on samples prepared as described in section 2.3.1. Defatting and polyphenols removal was also performed on aliquots of approximately 600 mg of freeze-dried cocoa powder, using the same procedure as described in section 2.3.1.

To optimise the extraction process and reduce the amount of sample required to carry out protein analysis, samples aliquots were reduced to approximately 160 mg of freeze-dried cocoa beans. Fat and polyphenols removal was performed as described in section 2.3.1, however, in this case the extraction volumes were scaled down from 10 to 3.5 ml.

2.3.3 Peptidomic analysis

The seeds were freeze-dried for 12-14 hours, snap-frozen using liquid nitrogen and subsequently ground using a grinder. These samples were used for all subsequent peptidomic analyses. Fat from aliquots of approximately 200 mg were extracted with 3.5 ml of petroleum ether (boiling point 40-60° 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.

2.4 Proteins extraction

2.4.1 Characterisation of cocoa proteome

Albumin and vicilin fractions were extracted following a slightly modified method reported in the literature [23]. To obtain the albumin fraction 13.5 ml of buffer 1 was added to the ACDP. The suspension was stirred at 250 rpm for 1 hour at 4°C and subsequently centrifuged at 3100 g for 20 minutes at 4°C. The extraction was repeated, and all the supernatants collected. The combined supernatant solutions were centrifuged at 3100 g for 20 minutes and their supernatant was transferred into a 50 ml centrifuge tube. To precipitate the albumin fraction, TCA was added to this solution to get a final concentration of 10% (w/v). The solution was stored at -20° C for 60 minutes and subsequently centrifuged at 3100 gf for 20 minutes. The supernatant was discarded and the pellet was washed three times with 15 ml of cold acetone. After each acetone wash, the suspension was centrifuged at 4000 rpm for 15 minutes at 4° C. Residual acetone was removed under a stream of nitrogen.

In order to obtain the vicilin (globulin) fraction, the precipitate obtained following the albumin extraction with buffer 1 was extracted with 13.5 ml of buffer 2. The extraction was performed by TCA precipitation as described for the albumin extraction above.

Proteins from the pellet obtained following the albumin and vicilin fractionation steps were extracted with 14.0 ml of solubilisation solution. The suspension was stirred at 250 rpm for 1 hour at room temperature and subsequently centrifuged at 4000 rpm for 15 minutes at 4 ° C. The supernatant was collected and stored at -20° C.

For the total protein extraction from the unfractionated ACDP, 14.0 ml of solubilisation solution was added to this sample, and the suspension was stirred at 250 rpm for 1 hour at room temperature and subsequently centrifuged at 4000 rpm for 15 minutes at 4° C.

2.4.2 Method development and proteomic analysis of cocoa genotypes

Initial experiments were carried out on unfractionated samples prepared as described in section 2.4.1. The ACDP obtained from the 600 mg aliquots were extracted following the same procedure described in section 2.4.1 for the unfractionated workflow. However, the volume of solubilisation solution was reduced to 12.0 ml in this case.

To the ACDP obtained from the 160 mg aliquots, 3.5 ml of solubilisation solution were added. The suspensions were vortexed 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.

2.5 Peptide extraction

For the initial experiments, approximately 25-30 mg of PVPP (Fisher, UK) were added to the defatted cocoa beans and subsequently 4 ml of a methanol:water 70:30 (v/v) solution were added. The samples were vortexed for 1 minute and then 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. An aliquot of 0.2 ml of the peptides solutions was transferred into a 0.5 ml Eppendorf^{TM} centrifuge tube, dried down in a Genevac MIVAC^{TM} duo concentrator and reconstituted in 0.2 ml of a 0.1% TFA solution in water. These solutions were transferred into HPLC vials for LC-MS/MS analysis.

Additional analyses were performed using 0.5% TFA in water and 0.5% TFA in water:methanol 80:20 (v/v) as extraction solutions, using the same volumes and protocol as described above. For every extraction, approximately 20-25 mg of PVPP were added to the defatted cocoa beans.

2.6 Desalting of peptides solutions

The peptide solutions extracted with 0.5% TFA in water 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 aqueous 0.5% (v/v) TFA. After loading an aliquot of 0.4 ml of the samples solutions, the cartridges were washed with 0.2 ml of 0.5% TFA, and then eluted with 2x 25 μl of 0.1% TFA in acetonitrile:water 75:25

(v/v) solution. The SPE eluates were dried down in a speed vacuum and reconstituted in 0.1 ml of aqueous 0.1% TFA prior to LC-MS/MS analysis.

The peptide solutions extracted with 0.5% TFA in water:methanol 80:20 (v/v) were also desalted with SOLAμ HRP 96 Well Plate 2 mg sorbent mass SPE cartridges following the same protocol described above. However, in this case, an aliquot of 0.4 ml of each peptide extract solution was diluted to a final volume of 2 ml with 0.5% aqueous TFA and the whole solution loaded on the SPE cartridges.

2.7 SDS-PAGE

2.7.1 Method development

Protein extracts obtained from 600 mg aliquots as described in section 2.4.2 were diluted with water to have a final concentration of approximately 1.9 or 3.0 mg/ml of protein. An aqueous 8.7-mg/ml BSA solution was diluted with water to have an approximate concentration of 1.8 or 0.4 mg/ml. Samples and BSA solutions were subsequently diluted with 1 volume of Laemmli Buffer containing mercaptoethanol and incubated for 10 minutes at 70° C. Precision Plus Protein™ (BIO-RAD) markers which cover a range of 10-250 kDa were used to assess the molecular weight of the electrophoretic bands. Samples and standards (10 μl) were loaded onto a 12% Mini-PROTEAN® TGX™ gel (BIO-RAD) and the electrophoresis analysis was carried out on a Mini PROTEAN Tetra Cell (BIO-RAD) at a constant voltage of 300 or 250 V. The running buffer was 25 mM Tris, 192 mM glycine, 0.1% SDS. The gels were subsequently stained with Bio-Safe Coomassie Stain (BIO-RAD). Images of the gels were acquired on a BIO-RAD Molecular Imager® Gel Doc™ XR System with Quantity One Software version 4.6.9.

2.7.2 Analysis of cocoa genotypes

Protein extracts obtained from 160 mg aliquots as described in section 2.4.2 were diluted with water to have a final concentration of approximately 3.0 mg/ml. An aqueous BSA solution (8.7 mg/ml) was diluted with water to have an approximate concentration of 0.4 mg/ml. Samples and BSA solutions were reacted with Laemmli Buffer and analysed on a Mini PROTEAN Tetra Cell with the same conditions as described in section 2.7.1, setting the voltage at 250 V for the whole duration of the run. Images of the gels were acquired as described in section 2.7.1.

2.8 Protein quantitation and trypsin digestion

2.8.1 Characterisation of cocoa proteome

A volume of 3.0 ml of solubilisation solution was added to the albumin and globulin TCA precipitation pellets and the pH was adjusted to 8.5-9.0 with the addition of 10 µl of a 1-M NaOH solution. The pellets were incubated for 30 minutes at 30° C and subsequently vortexed until complete dissolution. The solutions were then centrifuged at 4000 rpm for 15 minutes at 4° C. The supernatant was collected and stored at -20° C prior to further analysis. The protein concentration in each sample was assessed with the Bradford assay [72]. BSA was used as reference standard for quantitation purposes.

Aliquots of each sample solution containing a total amount of 10 µg of proteins based on the Bradford assay were transferred into 1.5-ml microcentrifuge tubes and spiked with 7 µl of an aqueous 10-mg/l BSA solution. A volume of 10 µl of an aqueous 200-mM DTT solution was added to each tube, and the final concentration of DTT was adjusted to 10 mM by adding 160-180 µl of 50 mM aqueous ammonium bicarbonate. The solutions were incubated for 30 minutes at 37° C. A volume of 23 µl of an aqueous 200-mM IAA solution was then added to each sample solution in order to obtain a final IAA concentration of 20 mM. After keeping these solutions in the dark at room temperature for 30 minutes, 40 µl of a 50-mM aqueous ammonium bicarbonate solution was also added to reduce the concentration of urea below 1 M. The pH of each solution was measured to ensure it was around 8. To each sample tube a volume of 1 µl of a 0.2-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 10 µl of a 10% (v/v) solution of aqueous TFA to each sample tube. The sample solutions were dried down under vacuum using a centrifugal evaporator and stored at -20°C. Prior to MS/MS analysis, the tryptic digests were defrosted and solubilised with 10 µl of 0.1% (v/v) TFA in water and desalted using ZipTips (Merck Millipore, Watford, UK) with 0.6 µl C18 resin according to the manufacturer's protocol. The ZipTip eluates were diluted to a final volume of 20 µl with the addition of 0.1 % (v/v) TFA in water.

2.8.2 Method development and proteomic analysis of cocoa genotypes

Aliquots of unfractionated protein extracts prepared as described in section 2.4.1 containing a total amount of 10 µg of protein based on the Bradford assay were digested following the protocol outlined in section 2.8.1. Some of these aliquots were desalted using ZipTip as described in section 2.8.1. A desalting step was also carried out on STRATA X 96 Well Plate 2 mg sorbent mass SPE cartridges (Phenomenex, Macclesfield, UK) with a sorbent mass of 2 mg. In this case the cartridges were initially conditioned with 0.2 ml of methanol and subsequently equilibrated with 0.2 ml of 0.4% v/v TFA in 50 mM aqueous ammonium bicarbonate. After loading the whole tryptic digest solutions, the cartridges were washed with 0.2 ml of a water:methanol 97:3 (v/v) solution, and then eluted with

3x 25 µl of 1% TFA in acetonitrile:water 75:25 (v/v) solution. The eluates were dried down under vacuum using a centrifugal evaporator and stored at -20° C. Prior to MS/MS analysis, the tryptic digests were diluted to a final volume of 50 µl with 0.1% TFA in water.

Aliquots of unfractionated protein extracts solution obtained from approximately 600 mg of cocoa beans (see section 2.4.2), containing a total amount of 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 13 µl of an aqueous 200-mM DTT solution was added to each tube, and the final concentration of DTT was adjusted to 10 mM by adding 170 µl of 86 mM aqueous ammonium bicarbonate. The solutions were incubated for 30 minutes at 37° C. A volume of 22 µl of an aqueous 200-mM IAA solution was then added to each sample solution to obtain a final IAA concentration of 20 mM. To each sample tube a volume of 20 µl of a 0.15-µg/µl trypsin 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 10 µl of a 10% (v/v) aqueous solution of TFA to each sample tube. A volume of 20 µl of the tryptic peptide solutions was transferred into 0.2 ml microcentrifuge tubes, dried down under vacuum using a centrifugal evaporator and stored at -20° C. Prior to MS/MS analysis, the tryptic digests were defrosted and solubilised with 10 µl of 0.1% (v/v) TFA in water and desalted using ZipTip with 0.6 µl C18 resin according to the manufacturer's protocol. The ZipTip eluates were diluted to a final volume of 20 µl with the addition of 0.1 % (v/v) TFA in water. The tryptic digest solutions which were left in the 0.5 ml tubes after removing the aliquots for desalting with ZipTip, 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 aqueous 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 LC-MS/MS analysis.

Aliquots of protein extracts obtained from 160 mg of cocoa beans (see section 2.4.2), 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 DTT solution was added to each tube, and the final concentration of DTT was adjusted to 10 mM by adding 290 µl of 77 mM aqueous ammonium bicarbonate. The solutions were incubated for 30 minutes at 37° C. A volume of 43 µl of an aqueous 200-mM IAA solution was then added to each sample solutions to obtain a final IAA concentration of 20 mM. Samples were diluted with a 7-M urea solution to give a final concentration of urea of 0.6-0.7 M. To each sample tube a volume of 20 µl of a 0.15-µg/µl trypsin 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 TFA to each sample tube. Prior to MS/MS analysis, the whole tryptic digest solutions were desalted employing the same protocol with Thermo Sola SPE cartridges as described above.

2.9 Nano-UHPLC-ESI MS/MS analysis

2.9.1 Characterisation of cocoa proteome

The desalted tryptic digests were analysed on a nano-UHPLC-ESI MS/MS system consisting of an Orbitrap Fusion (Thermo Scientific, Waltham, MA USA) mass spectrometer coupled to a Dionex Ultimate 3000 nano-RSLC (Thermo Scientific) nano-UHPLC system. The injection volume for each sample was 1 µl. The nano-UHPLC system was kept at 40° C and the column configuration included an Acclaim PepMap C18 100 µm × 2 cm 3 µm particle size trap column (Thermo Scientific) and an Acclaim PepMap C18 75 µm × 25 cm 3 µm particle size analytical column (Thermo Scientific). The chromatographic separation of the tryptic digests was carried out under a linear gradient elution using 0.1 % (v/v) formic acid in water as solution A and 0.1 % (v/v) formic acid in acetonitrile as mobile phase B with a flow rate of 300 nl/min. The gradient conditions were as follows: 4% B at 0-4 minutes, 50% B at 144 minutes, 90% B at 180-185 minutes, 4% B at 186-196 minutes. The nano-ESI source was operated in positive ion mode. MS analysis was carried out using the Orbitrap mass analyser, setting the resolution at 120,000 and the AGC target at 400,000 with a maximum injection time of 100 ms. The MS scan covered an *m/z* range between 300 and 1500. For MS/MS analysis a data dependent experiment was performed with the Quadrupole mass analyser as the initial filter, setting the isolation window width at *m/z* 1.6. For this experiment, the resolution of the Orbitrap was set to 30,000 with an AGC target of 5,000 and a maximum injection time of 35 ms. Fragmentation was performed by HCD with a normalized collision energy of 32% and an activation q value of 0.25. Dynamic exclusion was enabled in order to reduce the occurrence of redundant sequencing. MS peaks detected more than once over a 30 s window were not automatically fragmented for 40 s. The threshold for triggering a data-dependent scan was set to 5,000 and only ions with a charge state between 2 and 7 were selected.

2.9.2 Method development for quantitative proteomic analysis

The same instrumentation, trap column, mobile phases and MS parameters described in paragraph 2.9.1 were employed. Analyses were carried out an Acclaim PepMap C18 75 μm × 50 cm analytical column (Thermo Scientific). The temperature of the column oven was set at 50° C. The following gradient conditions were employed: 4% B at 0-4 minutes, 30% B at 150 minutes, 60% B at 160 minutes, 90% B at 180-185 minutes, 4% B at 186-196 minutes. The flow rate was set at 300 nl/min and the injection volume was 1 μl.

2.10 Microflow UHPLC-ESI MS/MS analysis

2.10.1 Method development

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 varied between 10 and 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 tryptic 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. Injection volumes of 10 and 15 μl were evaluated. The ESI source was operated in positive ion mode. MS analysis was carried out using the Orbitrap mass analyser, setting the resolution at 70,000 and the AGC target at 1,000,000 with a maximum injection time of 200 ms. The MS scan covered an *m/z* range between 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 at *m/z* 2.0. For this experiment, the resolution of the Orbitrap was set to 17,500 with an AGC target of 100,000. Maximum injection time of 200 and 300 ms were evaluated. Fragmentation was performed by HCD with a normalised collision energy of 28%. Dynamic exclusion was enabled setting the filter at 15 seconds. The threshold for triggering a data-dependent scan was set to 67,000 and only ions with a charge state between 2 and 5 were selected.

2.10.2 Proteomic analysis of cocoa genotypes

The same instrumentation, column, mobile phases and MS parameters described in section 2.10.1 were employed. The injection volume for all samples analysed was 15 µl, and the injection time for MS/MS analysis was 300 ms.

2.10.3 Peptides analysis

The same instrumentation, and MS parameters described in section 2.10.1 were employed. For the MS/MS analysis injection times of 200 and 300 ms were evaluated. The injection volume for all samples analysed was 10 µl.

2.11 Data analysis

2.11.1 Characterisation of cocoa proteome

All MS/MS spectra were processed using Mascot Distiller software (Matrix Science Ltd, London, UK; Version 2.5.1.0) to convert the raw LC-MS/MS data into peak lists suitable for database searching using the Mascot search routine (Matrix Science Ltd; Version 2.4.1). 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; accessed on 31st May 2015; 59,577 sequences; 23,720,084 residues), The cacao Criollo genome v2.0 *Theobroma cacao* database [73] (downloaded on 25th July 2018; 30,655 sequences; 14,782,063 residues) the NCBInr database (downloaded on 16th June 2015; 67,841,823 sequences; 24,324,060,020 residues), the Uniprot database (downloaded on 31st March 2014; 542,782 sequences; 193,019,802 residues), a custom NCBInr database with entries restricted to *Theobroma cacao* only (downloaded on 7th July 2015; 43,683 sequences; 19,146,837 residues) and a custom Uniprot database with entries restricted to *Theobroma cacao* only (downloaded on 1st July 2015; 40,941 sequences; 17,501,566 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. Taxonomy of *viridiplantae* was specified when searching against the NCBInr and Uniprot databases. The Mascot protein reports were exported as .csv files using Report Builder within Mascot with a filter of at least 2 'significant sequences'. The data were subsequently processed using Excel software for protein differential analysis based on the emPAI value of each protein normalized against BSA which had been added to the samples as an internal standard. The emPAI value is calculated using the following equation: emPAI = $10^{\frac{N\ observed}{N\ observable}} - 1$, where *N observed* is the number of experimentally observed peptides and *N observable* is the calculated number of observable peptides for each protein [44].

The amino acid sequence of BSA was added to the *Theobroma cacao* databases. Functional annotation was carried out by matching the protein accession codes from the Cacao Matina 1-6 Genome v1.1 *Theobroma cacao* to the GoMapMan database (http://protein.gomapman.org/).

2.11.2 Method development and proteomic analysis of cocoa genotypes

All MS/MS spectra were processed using Mascot Distiller software as described in section 2.11.1. Mascot searches were carried out against the Cacao Matina 1-6 Genome as described in section 2.11.1 and a custom-made database (70 sequences; 31,845 residues) containing the most common contaminants. For method development and 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). Label-free quantitation was carried out using a replicate protocol with Mascot Distiller software. The FDR was adjusted to 1% for all searches related to quantitative experiments. Normalisation of the proteins' intensities was carried out against BSA. Protein quantitation was performed using 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.

2.11.3 Peptide analysis

All MS/MS spectra of the peptide extract raw files were processed using Proteome Discoverer software (Version 2.1) to obtain peaks lists from the LC-MS/MS raw files which were subsequently loaded onto the Mascot server (Version 2.4.1). Searches were carried out against a custom-made database which contained the 100 most abundant proteins listed in Appendix 1 (100 sequences; 43,226 residues), the Cacao Matina 1-6 Genome v1.1 *Theobroma cacao* database (http://www.cacaogenomedb.org/Tcacao_genome_v1.1

#tripal_analysis-downloads-box; accessed on 31st May 2015; 59,577 sequences; 23,720,084 residues), a custom-made database (897 sequences; 374,606 residues) containing the 897 proteins which had been identified during the characterisation of the cocoa proteome using a custom-made Uniprot\Tremble database with entries restricted to *Theobroma cacao* only, as described in section 3.1. Only peptides with an FDR and a q-value <= 0.01 were selected.

3 Results

3.1 In-depth characterisation of the cocoa bean proteome by LC-MS

Cocoa seeds were from ripe pods of the West African Amelonado variety harvested at the Cocoa Research Centre of the University of West Indies, St. Augustine, Trinidad. Approximately 240 beans from 6 different pods were combined. Cocoa beans have a fat content of around 30-40% w/w [11]. Therefore, it is advisable to extract the fat fraction and thus to remove highly hydrophobic compounds which can negatively affect the protein extraction yield and cause interferences in chromatographic separation. The most common procedure to extract fat involves the use of a Soxhlet apparatus and petroleum ether as extraction solvent. However, this procedure requires overnight extraction at a temperature ranging between 30-40° C which could lead to protein degradation. As a result, a quicker method for the removal of fat was devised for this study, involving three quick extractions with petroleum ether. The data from this quicker fat removal protocol showed that the amount of fat extracted was similar to the Soxhlet method, while substantially reducing the extraction time.

As mentioned in section 2.3, polyphenols can form insoluble complexes with proteins during extraction [12]. Therefore, this class of compounds was also removed prior to protein extractions, employing aqueous acetone as extraction solvent [23].

Two samples were then taken, one for a sample preparation workflow with fractionation and another for a workflow without fractionation. In the workflow with fractionation, albumin and vicilin fractions were extracted with a low and high ionic strength buffer, respectively. Membrane and hydrophobic proteins are not highly soluble in aqueous buffer, therefore a solubilisation solution containing chaotropic and denaturing agents was employed to extract these proteins from the cocoa powder remaining after the extractions with the low and high ionic strength buffers. This solution was also used for the protein extraction of the cocoa powder in the workflow without fractionation.

The majority of the peptides detected in the BPC chromatograms of the samples analysed were eluted in a range of 20-90 minutes, which corresponds to a percentage of organic solvent from approximately 10 to around 30%, see Figures 19-22. The rise in the baseline at the end of the gradient is due to strongly retained compounds present either in the sample solution or in the mobile phases. Some differences could be observed in the peptide elution patterns of these samples, specifically in the chromatogram of the unfractionated tryptic digest which showed a higher number of peaks compared to the chromatograms of the fractionated extracts, see Figures 19-22.

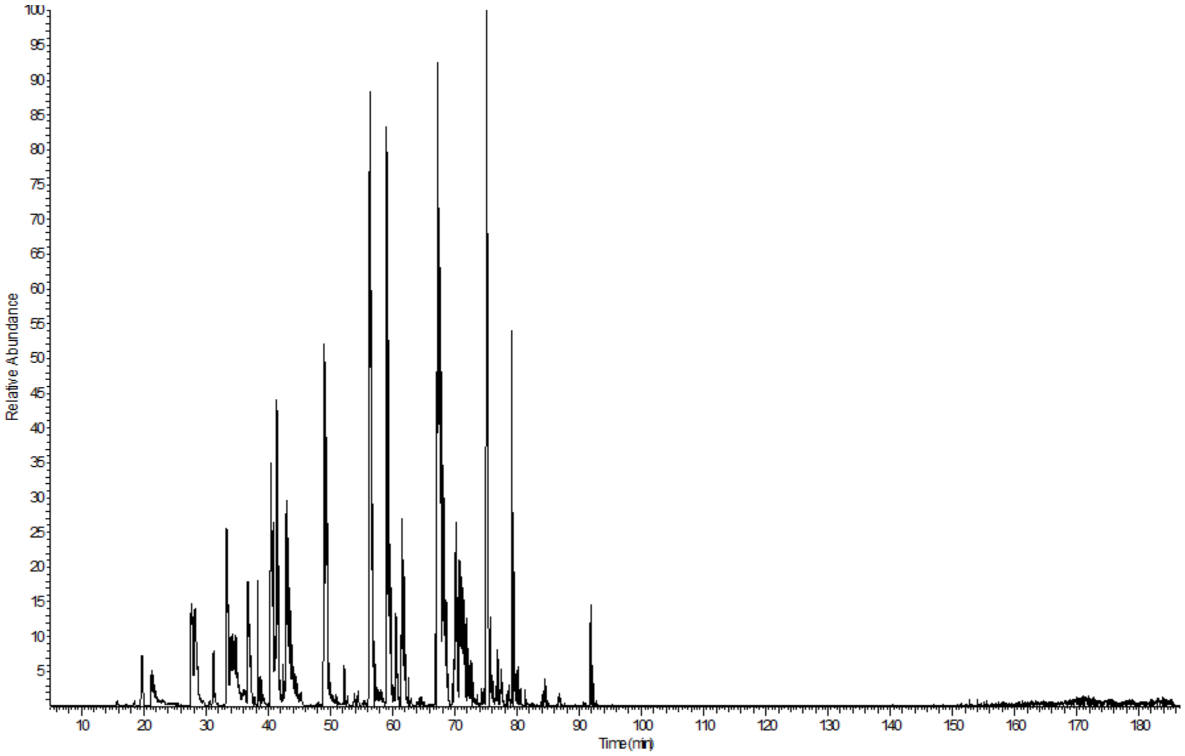


Figure 19. BPC chromatogram of full MS scan of salt-soluble fraction extracted from Amelonado cocoa beans with a high ionic strength buffer (0.5 M NaCl, 5 mM sodium ascorbate, 2 mM EDTA, 10 mM Tris), and analysed on the Orbitrap Fusion

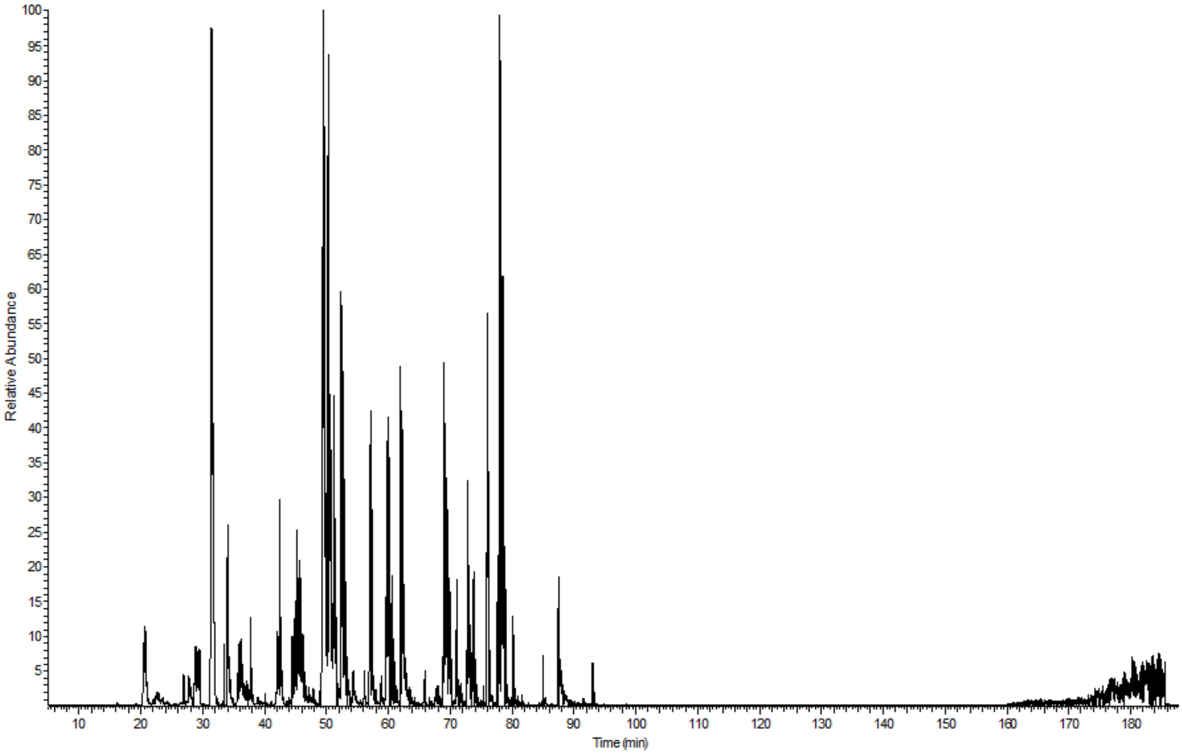


Figure 20. BPC chromatogram of full MS scan of water-soluble fraction extracted from Amelonado cocoa beans with a low ionic strength buffer (5 mM sodium ascorbate, 2 mM EDTA, 10 mM Tris), and analysed on the Orbitrap Fusion

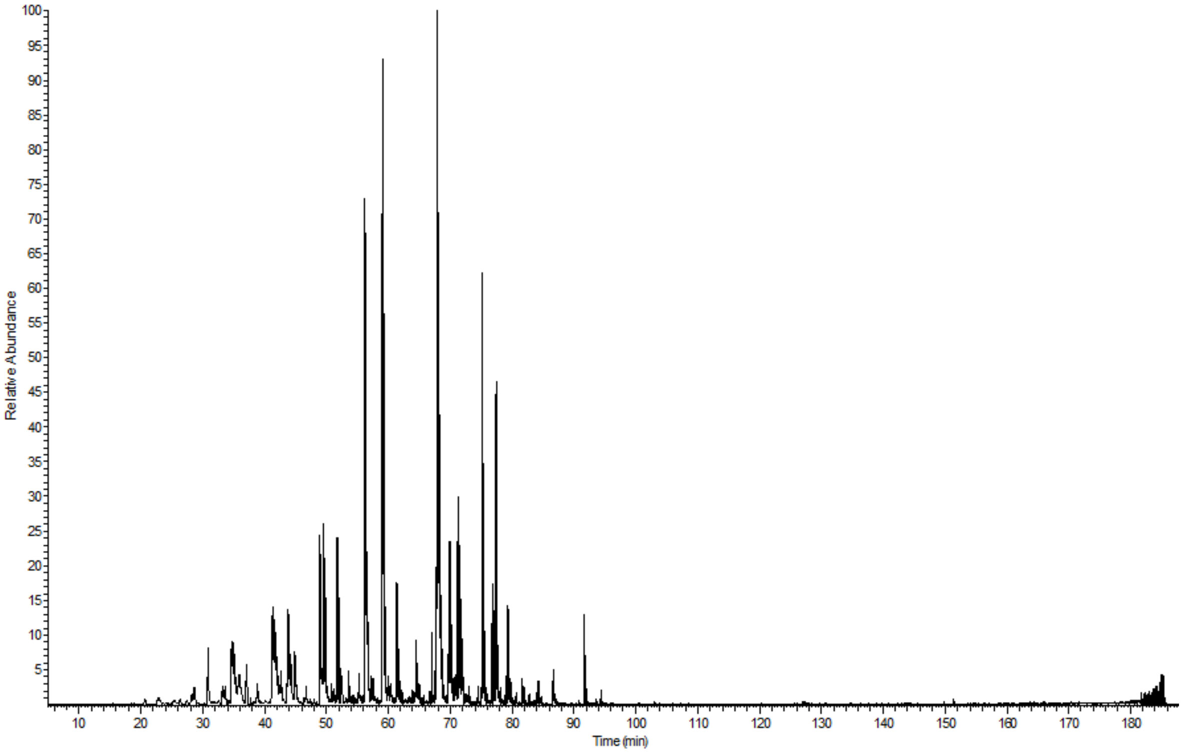


Figure 21. BPC chromatogram of full MS scan of urea-soluble fraction extracted from Amelonado cocoa beans with a solution consisting of 7 M urea, 2 M thiourea, 10 mM DTT and analysed on the Orbitrap Fusion

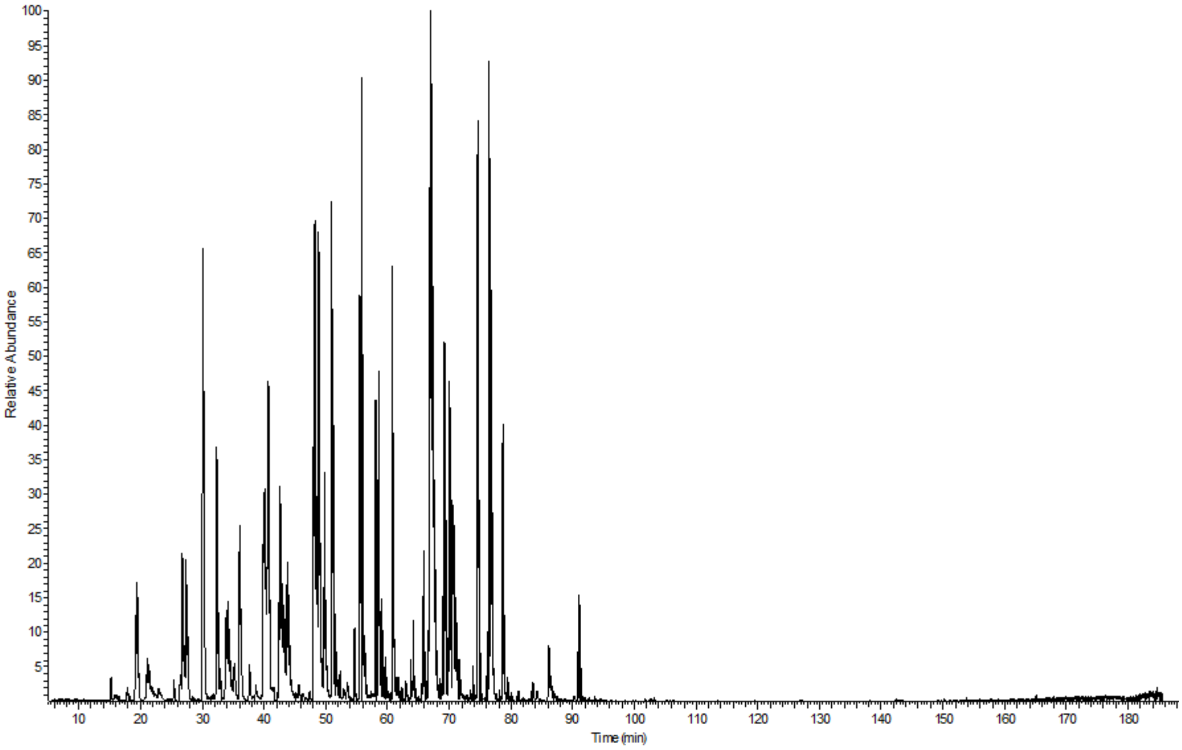


Figure 22. BPC chromatogram of full MS scan of the unfractionated sample of Amelonado cocoa beans extracted with a solution consisting of 7 M urea, 2 M thiourea, 10 mM DTT and analysed on the Orbitrap Fusion

Overall, the searches against the Cacao Matina 1-6 Genome database published by Motamajor *et al.* [74] of the MS/MS data obtained from the tryptic digests returned a total of 906 and 704 proteins hits for the fractionated and unfractionated sample, respectively, see Figure 23. About 86% of the proteins detected in the unfractionated sample (607 hits) were also detected in the fractionated sample, while 97 proteins were identified in the unfractionated sample only, see Figure 23. A total of 1003 proteins were identified when the entries from the fractionated and unfractionated samples were combined. A graphical representation of the proteins detected in both the fractionated and unfractionated sample is shown in Figure 23:

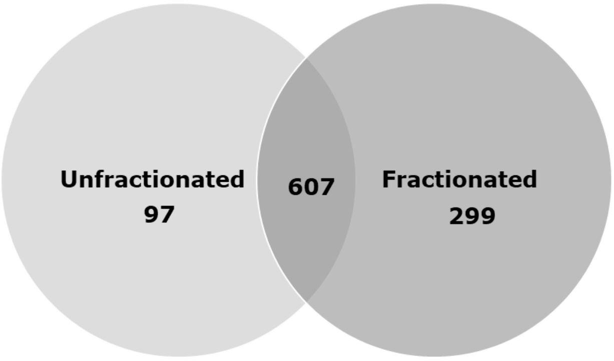


Figure 23. Venn diagram of proteins detected in the fractionated and unfractionated samples searched against the Cacao Matina 1-6 Genome database published by Motamajor *et al*. [74]

As for the fractionated sample, 590 protein identifications were recorded when the identifications of the water-soluble and salt-soluble fractions were combined, of which 376 were also present in the urea-soluble fraction, see Figure 22. A higher number of proteins were detected in the urea-soluble fraction compared with the water- and salt-soluble fractions. The distribution of the proteins detected in the fractionated sample is depicted in the Venn diagram shown in Figure 22.

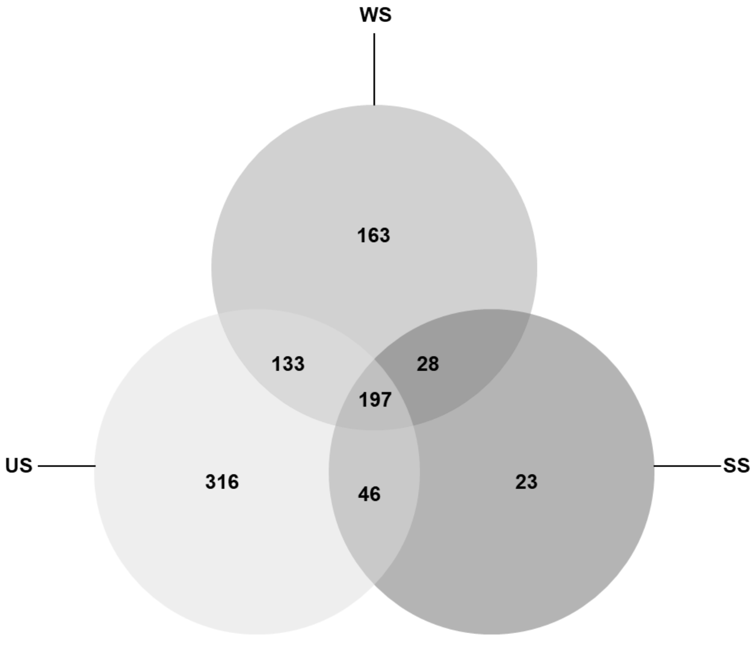


Figure 24. Venn diagram showing the numbers of proteins identified in the fractionated T. cacao sample searched against the Cacao Matina 1-6 Genome database published by Motamajor *et al*. [74]. The protein fractions are labelled as follows: WS, water-soluble fraction; SS, salt-soluble fraction; US, urea-soluble fraction.

In order to assess whether searching different databases would yield a higher number of protein hits, searches using NCBInr and Uniprot\Swissprot databases with taxonomy *Viridiplantae*, and custom databases containing only *Theobroma cacao* entries from Uniprot\Tremble and NCBInr were also carried out. The Uniprot\Tremble version was chosen for the custom database as the Uniprot\Swissprot version accessed on 7th July 2015 returned only 7 hits (Endochitinase 1, CHI1_THECC; Arginine decarboxylase, CHI1_THECC; Casparian strip mebrane protein, SPE2_THECC; Vicilin, VCL_THECC; Maturase K, MATK_THECC; 21 kDa seed protein, ASP_THECC; Coat protein, COAT_CAYMV) when filtered for the keyword *Theobroma cacao* . The results of these searches are shown in Table 6.

Table 6. Number of identified proteins by nano-UHPLC-ESI MS/MS at 1% FDR for data searches using the search engine Mascot and different protein sequence databases

Database	No. of Sequences in Database	No. of Identified Proteins
Cacao Matina 1-6 Genome (*T. cacao)*	59,577	906
Uniprot\Swissprot (*Viridiplantae*)	34,907	364
NCBInr (*Viridiplantae*)	3,047,619	759
Uniprot\Tremble (*T. cacao*)	40,941	897
NCBInr (*T. cacao*)	43,683	870

The highest number of identified proteins (906) was obtained when searching the Cacao Matina 1-6 Genome database. A slightly lower number of proteins were detected in the Uniprot/Tremble database (897) and the NCBInr database (870) when restricted to *Theobroma cacao* entries. Only 364 proteins could be identified from a search using the Uniprot\Swissprot database with the taxonomy set to

Viridiplantae, while 759 proteins were identified when searching the same data against the NCBInr atabase with taxonomy *Viridiplantae*. Searches against the recently sequenced cacao Criollo Genome database [73] were also carried out. However, in this case only 781 proteins were identified.

Proteins were also classified according to their main biological function using the results from the search of the Cacao Matina 1-6 Genome database of the fractionated sample (906 protein identifications). The biological process for each protein was obtained by loading the protein accessions on the GoMapMan database. If the information provided by this database was ambiguous, the proteins sequences were searched against Uniprot for homology using BLAST to gain information on the function of the specific protein. A graphical overview of the proteins classification is shown in Figure 25 with two representations, providing the abundance-weighted and unweighted percentage for each protein class (function group). The percentage of each protein function for the abundance-weighted classification was calculated by summing the normalised emPAI responses of each protein within the same function detected in all three fractions. The following function definitions were employed to classify proteins based on their function:

Table 7. Function definitions for the classification of proteins based on their biological processes, based on the results obtained from the characterisation of the proteome of *Theobroma cacao* beans

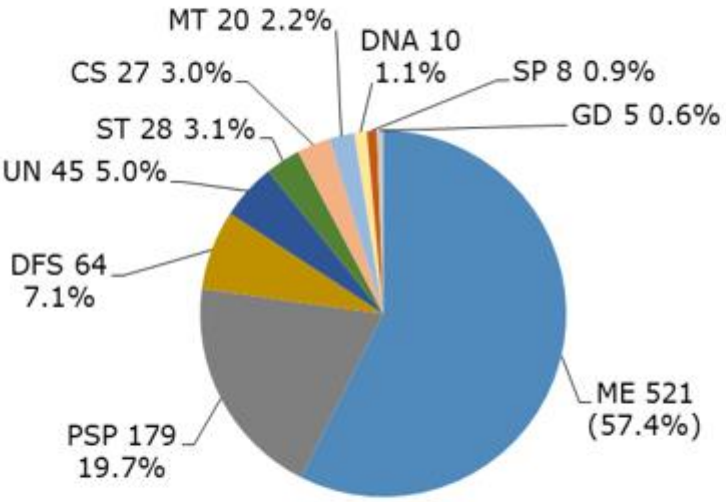
Biological process	Function
Cell organisation	CS
Cell structure	
Cell vescicle transport	
Cell wall modification/pectinesterase activity	
Cell wall proteins/glucosyl transferase activity	
Cell wall structure	
Constituent of cytoskeleton	
Endoplasmic reticulum	
Outer membrane constituent	
Structural molecular activity	
Chitin hydrolase/response to stress	DFS

Defence response	
Response to stress	
Senescence	
Tumor reversion/response to stress	
DNA binding	DNA
DNA repair	
DNA repair/DNA recombination	
DNA synthesis	
Embryo development	GD
Acid phosphatase activity	ME
Amidase activity	
Amino acid metabolism	
Amino acid metabolism/oxidation reduction process	
ATP binding/phosporylation	
ATP hydrolysis/proton transport	
ATP synthesis/proton transport	
Biodegradation of xenobiotics	
Biosynthetic process/strictosidine synthase activity	
Calcium-dependent phospholipid binding	
Carbohydrate metabolic process	
Carbonate dehydratase activity	
Catalytic activity	
Chlorophyll catabolic process	
Copper ion binding/electron carrier activity	
Cyanate metabolic process	
Decarboxilase activity/phospholipid biosynthetic process	
Desphosporylation	
Electron transport	
Electron transport/ATP synthesis	
Endopeptidase inhibitor activity	
Gluconeogenesis	
Glucose catabolic process	
Glycerol metabolic process	
Glycolytic process	
Hormone metabolism	
Hydration of carbon dioxide	
Hydrolase activity	
Iron binding	
Iron binding/iron sulphur assembly	
Isomerase activity	
Kinase activity/phosporylation/ATP binding	
LIPIDS metabolism	
LIPIDS metabolism/oxidation reduction process	
Lypase actvity/cellular modified amino acid biosynthetic process	
Metabolic process	
Metabolic process/catalytic activity	
Metabolic process/oxidation reduction process	
Metabolic process/transferase activity	
Metabolism formate-tetrahydrofolate ligase	
Metal handling chelation and store	
Methyltransferase activity	
Monolayer-surrounded lipid storage body	
Nitrogen compound process/hydrolase activity	
Nucelotide metabolism	

Oxidation-reduction process	
Pectinesterase inhibitor activity	
Protease inhibitor	
Protein methabolic process/oxidation reduction	
Proteolysis	
Proton transport/hydrolase activity/ATP binding	
Regulation of protein catabolic process	
Ribulose-bisphosphate carboxylase activity/carbon fixation /plastid	
Selenium binding	
Strictosidine synthesis	
TCA cycle	
Transaminase activity	
Transferase activity	
Vitamin metabolism/catalytic activity	
Anion transport/transmembrane transport	MT
Membrane transport	
Metabolite transport	
Potassium transport	
Protein transport	
Protein transport through membrane	
Sterols carrier	
GTPase activity/protein synthesis/elongation factor	PSP
mRNA binding	
Negative regulation of translation	
Proteasome assembly	
Protein binding	
Protein folding	
Protein glycosylation	
Protein phosphorylation	
Protein polymerization	
Protein postranslational modification	
Protein synthesis	
Protein synthesis/translation	
Protein phosphorylation	
Protein synthesis and translation	
Protein folding	
RNA binding	
RNA processing	
RNA regulation	
RNA transcription regulation	
Transcription regulation	
Transcription regulation/protein metabolic process	
Translation/RNA binding	
T-RNA ligase activity	
T-RNA ligase activity/TRNA binding	
T-RNA synthesis	
Endopeptidase inhibitor/storage protein	SP
Nutrient reservoir activity	
Protease inhibitor/seed storage	
Signalling	ST
Unspecified biological process	UN

The functions are labelled as follows: ME, metabolism and energy; PSP, protein synthesis and processing; SP, storage proteins; MT, membrane transport; ST, signal transduction; UN, unclassified; CS, cell structure; DNA, DNA synthesis and processing; GD, growth and development; DFS, defence and stress.

A list of all proteins identified in the fractionated sample searched against the Cacao Matina1-6 genome database is provided in Appendix 1.



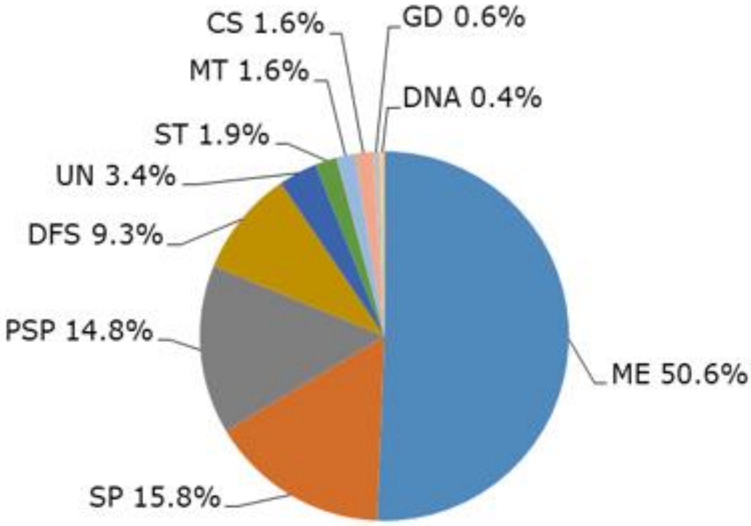


Figure 25 Classification of cocoa bean proteins based on their function. The percentages in the upper pie chart represent the number of proteins in each function group relative to the total number of proteins. Each function group is also labelled with the number of proteins. The lower pie chart provides the sums of the BSA-normalized emPAI values of the proteins in each function group relative to the total sum of the BSA-normalized emPAI values of all proteins.

3.2 Method development for quantitative proteomic analysis by LC-MS

3.2.1 Improvement of the desalting step and chromatographic separation for the Fusion Orbitrap analysis

The linear gradient employed for the chromatographic separation of the tryptic digests of the samples analysed for the characterisation of the cocoa bean proteome, generated a chromatographic pattern where the majority of the peptides were eluted within 90 minutes, see Figures 19**-**22**,** which would translate to a final percentage of acetonitrile of approximately 30%. To improve the chromatographic separation of the tryptic peptides, a shallower linear gradient was employed, reducing the gradient steepness from 0.33%/minute to 0.18%/minute increase in the percentage of the organic modifier, and doubling the column length from 25 to 50 cm. This gradient was labelled as 'long gradient' and is described in detail in section 2.9.2. The BPC chromatogram of an unfractionated sample analysed on the Fusion Orbitrap with the optimised long gradient is shown in Figure 26. In this case most of the peptides peaks were eluted between 20 and 170 minutes, see Figure 26.

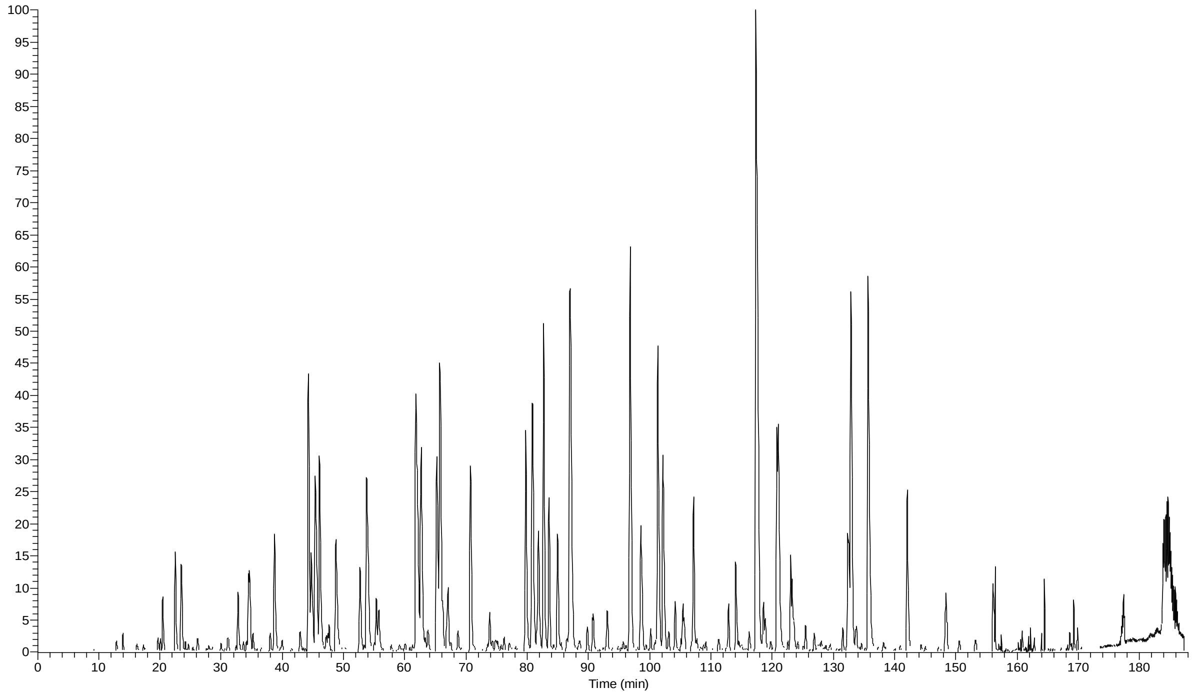


Figure 26. BPC chromatogram of an unfractionated sample of Amelonado cocoa beans extracted with urea 7 M, thiourea 2 M and 10 mM DTT and analysed with the 'long gradient' on the Dionex Ultima 3000 nano-RSLC coupled to the Fusion Orbitrap

An unfractionated sample was injected 10 times on the Dionex Ultima 3000 nano-RSLC coupled to the Fusion Orbitrap with the long gradient method to assess the reproducibility of the instrumentation in terms of the number of identified proteins. The results showed that an average of 880 proteins were identified in the 10 replicate injections searching the Cocoa Matina 1-6 database, which is considerably higher than the number of proteins (704) identified in the same sample analysed using a steeper gradient, see section 3.1. The number of identified proteins in each replicate sample are shown in Table 8.

Table 8. Number of proteins identified in 10 replicate injections of the same sample

Replicates	No of proteins	Average	RSD %
1	836	880	2.0
2	903		
3	897		
4	885		
5	886		
6	864		
7	867		
8	896		
9	889		
10	881		

To assess whether the identified proteins could be reproducibly quantified, a label-free quantitative proteomic analysis on replicate samples number 2 to 5 from Table 8 was carried out. In this case Mascot Distiller software with replicate protocol was employed. A total of 880 proteins were quantified in all the four replicate samples, of which 837 (95%) showed an RSD between the replicates <20%. The RSD is expressed as the relative standard deviation of the intensity of each protein in the four replicate samples analysed. Protein abundance was normalised against BSA which had been added to the samples as internal standard prior to trypsin digestion.

To evaluate the reproducibility of the trypsin digestion and desalting step, four aliquots of the unfractionated sample extract solution containing 10 µg of proteins were digested separately, and the resulting tryptic peptides were desalted using C18 ZipTips. The desalted samples were analysed using the Dionex Ultima 3000 nano-RSLC coupled to the Fusion Orbitrap LC-MS and label-free quantitative proteomic analysis with replicate protocol was carried out. A total of 714 proteins could be quantified in all four aliquots. However, only 90 of these proteins had an RSD <20% between the four aliquots analysed, indicating that these results were not reproducible.

An additional experiment was carried out using a different desalting protocol. In this case four aliquots from an unfractionated sample extract containing each 10 µg of proteins were digested separately. The resulting tryptic digests were desalted using a Phenomenex Strata SPE 96 well plate. The desalted peptides were analysed using the Dionex Ultima 3000 nano-RSLC coupled to the Fusion Orbitrap LC-MS and a label-free quantitative proteomics analysis was performed with replicate protocol as described above. The results of this analysis showed that 435 proteins were quantified in all the four aliquots, of which 210 (48%) had an RSD <20% between the four aliquots.

To understand whether the poor reproducibility of the results was mainly due to the trypsin digestion step, an aliquot of the unfractionated sample containing 60 µg of protein was digested. After digestion, six separate aliquots containing each 10 µg of peptides were taken from the tryptic digest solution and desalted separately on a Phenomenex Strata SPE 96 well plate. The LC-MS analysis of the desalted peptides and data processing for protein quantitation was carried out as previously described. The result of this experiment showed that a total of 505 proteins could be quantified in all the six aliquots, and 293 proteins (58%) had an RSD<20% between the six different aliquots.

Desalting tryptic peptides with the Phenomenex Strata SPE protocol resulted in a higher reproducibility for proteins quantification when compared to the results obtained with C18 ZipTips. However, a significant drop in sensitivity using the STRATA SPE protocol was observed, as the number of identified and quantified proteins was lower compared to the initial experiments carried out on tryptic digests desalted with C18 ZipTips. Therefore, to understand whether the drop in sensitivity was due to the desalting step or could be ascribed to issues with the instrumentation, eight aliquots of an unfractionated sample containing each 10 µg of proteins were digested separately, and then four of the digested aliquots were desalted with C18 ZipTips and the remaining four aliquots were desalted with Phenomenex Strata SPE 96 well plate. The desalted aliquots were analysed using the Dionex Ultima 3000 nano-RSLC coupled to the Fusion Orbitrap within the same sequence batch. A total of 681 proteins were quantified in the four aliquots desalted with Strata SPE, of which 557 (82%) showed an RSD<20% between the four aliquots. A lower number of proteins (571) were quantified in the aliquots desalted with C18 ZipTips. The reproducibility for these samples was considerably lower compared to the aliquots desalted with Strata SPE, as only 95 proteins (17%) had an RSD <20% between the four aliquots analysed. These results indicated that the reduced sensitivity in terms of number of identified proteins was not related to the use of Strata SPE for the desalting step.

To assess whether the low reproducibility observed for samples desalted with both ZipTips and Strata SPE was due to the low amount of proteins used for digestion, a further experiment increasing the amount of digested proteins from 10 to 160 µg was carried out. In this case proteins were extracted from approximately 600 mg of freeze-dried samples as described in section 2.4.2. The average protein contents of four replicates extracted from the same cocoa powder based on the Bradford Assay was 11.55 % (w/w) with an RSD (expressed as the variation in the protein amount between the four replicates) of 4.0 %. After the trypsin digestion was completed, aliquots equivalent to approximately 10 µg of proteins were taken from each of the digested sample solutions and desalted with C18 ZipTips. The remaining solutions were desalted with a Thermo SOLA reverse phase SPE cartridges mounted on a 96 well plate. Unlike the SPE strata 96 well plate which had cartridges fixed to the plate, the cartridges on the SOLA plate could be freely removed and located in any of the 96 positions of the plate. For this experiment the SPE cartridges used to desalt the tryptic digests were placed on the same positions of the well plate, and the solutions were desalted one at a time. Placing the SPE cartridges on the same position of the well plate would minimise variation in the flow rate of the solutions pushed through the cartridges, which may have occurred if the cartridges had been mounted on different positions of the well plate. A total of 511 and 484 proteins were quantified in the samples desalted with SPE SOLA and C18 ZipTips, respectively. The reproducibility of the quantified proteins was considerably improved in both experiments, as 96% of the quantified proteins in the sample desalted with SPE SOLA showed an RSD<20% in the four samples analysed, and 93% of the quantified proteins in the samples desalted with C18 ZipTips showed an RSD<20%. Replicate protocol with Mascot Distiller software was used in this case for the quantitative proteomics analysis.

To assess whether scaling down the amount of sample extracted and the volume of extraction solution for the workflow with no fractionation would affect the reproducibility and yield of the extraction, six portions of approximately 160 mg of cocoa powder were extracted with 3.5 ml of solubilisation solution. An average protein amount of 9.6 % w/w based on the Bradford assay was obtained from the six replicate extractions with a RSD between the replicates of 4.8 %.

3.2.2 Improvement of the LC-MS/MS method for the Q Exactive Orbitrap analysis

For the improvement of the LC-MS/MS method using the Q Exactive Orbitrap, tryptic digests obtained from 160 mg of freeze-dried cocoa powder, with a final concentration of peptides of 0.5 mg/ml were analysed. These solutions were desalted on SOLA SPE prior to LC-MS/MS analysis. In order to assess the reproducibility of the method and the proteins extraction procedure, 6 replicates of the same cocoa powder sample were extracted with no fractionation and digested as described in the materials and method section. In this case the injection volume was 10 µl and the injection time for the ions accumulated in the C-Trap prior to entering the Orbitrap for MS and MS/MS acquisition was 200 ms. The BPC chromatogram of an unfractionated sample aliquot analysed on the Q Exactive is shown in Figure 27.

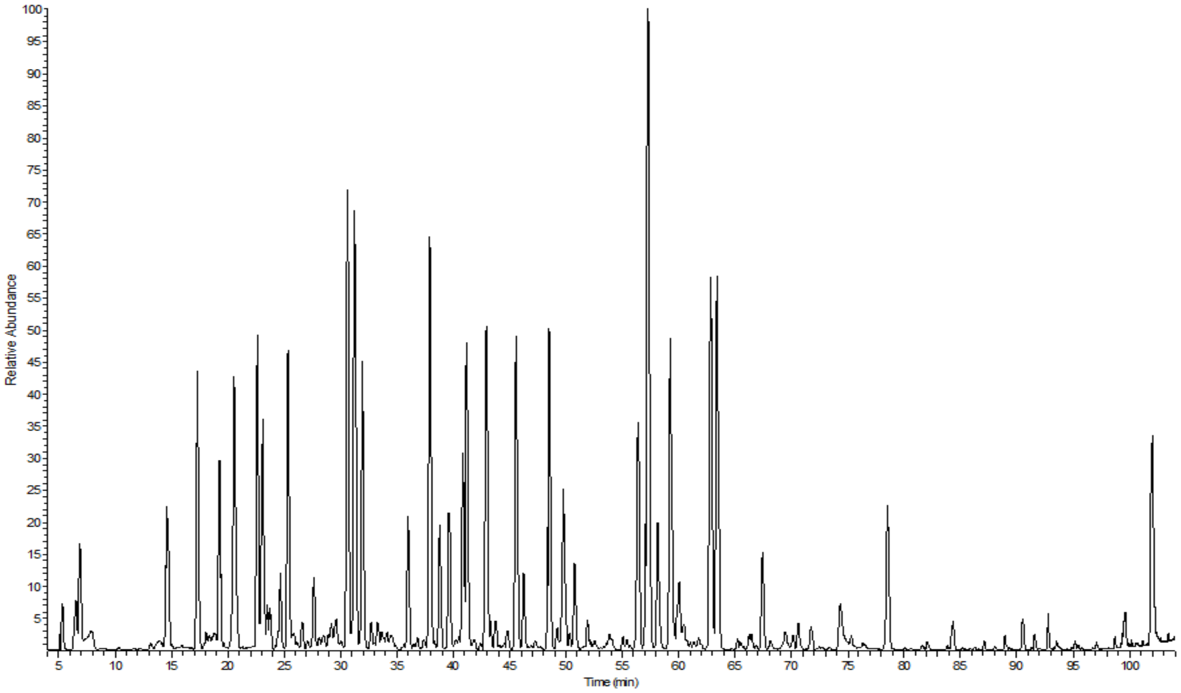


Figure 27. BPC chromatogram of an unfractionated tryptic digest of Amelonado cocoa beans extracted with urea 7 M, thiourea 2 M and 10 mM DTT sample analysed on the microflow Dionex Ultima 3000 UHPLC coupled to the Qexactive with an IT time of 200 ms and injection volume of 10 µl

The most abundant peptides were eluted within 15 and 65 minutes, with minor peptides detected up to 100 minutes. The chromatographic separation of the tryptic digests using the Dionex Ultima 3000 UHPLC looked acceptable, therefore the gradient was not further optimised. The raw MS/MS data from the six replicates were processed with Mascot Distiller and searched against the Cacao Matina 1-6 database. Label-free quantitation was carried out employing replicate protocol in Mascot Distiller. The results showed that a total of 419 proteins could be quantified in all the six replicates, and 406 (97%) of the quantified proteins had an RSD of <20% among the six replicates, confirming that both the extraction and the acquisition method were reproducible. In this section the RSD is expressed as the relative standard deviation of the intensity of each protein in the replicate samples analysed, unless stated otherwise.

To assess whether the length of the injection time for MS/MS analysis would have an impact on the number of identified proteins, a tryptic digest sample obtained from 160 mg of freeze-dried cocoa powder as described above was injected in triplicate with an injection time of both 200 and 300 ms. A total of 286 proteins were identified and quantified in the replicates analysed with an injection time of 200 ms, while increasing the injection time to 300 ms allowed identification and quantitation of 348 proteins in three replicates of the cocoa tryptic digest sample. In both cases, 98% of the quantified proteins showed an RSD of <20% between the three replicates. Once it was determined that an injection time of 300 ms for the MS/MS scan would provide a higher number of identified proteins, an additional experiment was performed to evaluate whether increasing the injection volume from 10 to 15 µl would result in a higher number of identified and quantified proteins. In this case the same tryptic digest sample as described above was analysed in triplicate, with an injection volume of 15 µl and an injection time for MS/MS of 300 ms. A total of 503 proteins were identified and quantified in three replicates of this sample, of which 493 (98%) showed an RSD of <20% between the three replicates. These results confirmed that using an injection volume of 15 µl and injection time of 300 ms for the MS/MS scan allowed a higher sensitivity in terms of identified and quantified proteins.

3.3 Quantitative proteomics analysis of cocoa genotypes by LC-MS

Although the label-free quantitation method using the Obitrap Fusion mass analyser had been developed and the initial results were highly reproducible, several hardware related issues were experienced after the method was developed, which affected significantly the reproducibility of the quantitative results. Since these issues could not be resolved within the timeframe of this PhD, all the quantitative analyses to assess the effects of shipment, different trees, harvest time and the investigation of proteome changes dependent on the genotype were performed on the Orbitrap Q Exactive mass analyser using the label-free replicate protocol with Mascot Distiller software as described in sections 2.10.2 and 3.2.2.

3.3.1 Effect of sample shipment

In order to understand whether the shipment would affect the proteomic profile of the cocoa bean samples, a portion of ground and freeze-dried cocoa beans from an Amelonado variety as described in section 2.2.1 was shipped back and forth from the University of Reading to Trinidad, while the remainder of the sample was stored at the University of Reading. A label-free quantitative proteomic analysis was carried out on the shipped and not-shipped sample to assess whether differences in the proteomic profile of the two samples could be found. Three portions of approximately 160 mg were taken from each sample and each of these portions were extracted and prepared separately as described in the material and method section 2.4.2, so as to have three analytical replicates for each sample. Each analytical replicate was injected only once when analysed by LC-MS/MS. A total of 450 proteins were identified and quantified in the two samples analysed. An S-adenosyl-L-methionine-dependent methyltransferases superfamily protein was detected at a higher level in the shipped sample (1.6-fold difference), while no other protein showed a fold difference >1.5 between the two samples.

3.3.2 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 harvested on the same day, one from four different trees grown in the ICGT field were selected. 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. The reproducibility of UHPLC-MS/MS analysis 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.

Only proteins which were identified and quantified in at least three preparative replicates of a biological replicate were selected for comparative label-free quantitative proteomic analysis. A total of 511 proteins were detected in the four biological replicates with these filters, and only six proteins showed a fold increase/decrease from the overall mean >2 in at least one biological replicate, while none showed any increase/decrease from the overall mean of >2.7, see Table 9.

Table 9. Proteins with fold increase/decrease from the mean >2 in at least one biological replicate of the genotype IMC 67 analysed to evaluate the effect of different trees. Fold increase/decrease from the overall mean is expressed as the ratio between the sample mean and the overall mean

Accession	Description	Biological process	Max fold increase/ decrease
Thecc1EG042578t1	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein	Hormone metabolism	2.69
Thecc1EG025391t1	Beta-amylase 6	Carbohydrate metabolic process	2.67
Thecc1EG000326t1	Salicylate O-methyltransferase	Hormone metabolism	2.64
Thecc1EG026589t1	Eukaryotic aspartyl protease family protein, putative	Protein degradation	2.47
Thecc1EG027146t1	HSP20-like chaperones superfamily protein	Stress	2.44
Thecc1EG041163t1	Glycosyl hydrolase family protein	Cell wall	2.09

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 number of proteins with a fold increase/decrease from the mean >2 in each biological replicate is shown in Table 10

Table 10. Number of proteins with a fold increase/decrease from the mean >2 detected in each biological replicate (effect of different trees)

Sample	No of proteins with fold increase/decrease >2	Accession	Description
T13-6	1	Thecc1EG027146t1	HSP20-like chaperones superfamily protein
T2-2	3	Thecc1EG025391t1	Beta-amylase 6
		Thecc1EG000326t1	Salicylate O-methyltransferase
		Thecc1EG042578t1	S-adenosyl-L-methionine-dependent methyltransferases
T5-3	1	Thecc1EG026589t1	Eukaryotic aspartyl protease family protein
T6-4	2	Thecc1EG025391t1	Beta-amylase 6
		Thecc1EG041163t1	Glycosyl hydrolase family protein

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 (20th Dec 2016, 21st Feb 2017, 23rd March 2017, 17th 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. Only proteins which were quantified in at least three preparative replicates of a biological replicate were evaluated

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 Table 11. 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.

Table 11. Proteins with fold increase/decrease from the mean of >2 in at least one biological replicate (effect of different harvest times) Fold increase/decrease from the overall mean is expressed as the ratio between the sample mean and the overall mean

Accession	**Description**	**Biological process**	**Max fold increase/ decrease**
Thecc1EG042149t1	Serine carboxypeptidase-like 48	Protein degradation	11.0
Thecc1EG047098t1	Uncharacterized	Unspecified process	5.0
Thecc1EG025043t1	Adenine nucleotide alpha hydrolases	Stress	2.9
Thecc1EG036608t1	Laccase 14	Secondary metabolism	2.7
Thecc1EG026543t1	Lipoxygenase 1	Hormone metabolism	2.5
Thecc1EG027146t1	HSP20-like chaperones superfamily protein	Stress	2.2
Thecc1EG006498t1	Basic chitinase	Stress	2.1
Thecc1EG005507t1	N-terminal nucleophile aminohydrolases	Protein degradation	2.0
Thecc1EG016209t1	Osmotin 34	Stress	2.0

The number of proteins with a fold increase/decrease from the overall mean >2 in each biological replicate was between three and five, see Table 12

Table 12. Number of proteins with a fold increase/decrease from the mean >2 detected in each biological replicate (effect of different trees)

Sample	No of proteins with >2-fold diff. from mean
T10-1	3
T10-3	4
T10-5	4
T10-6	5

Biological processes to the proteins listed in Table 9 and Table 11 were assigned based on the output of the GoMapMan software.

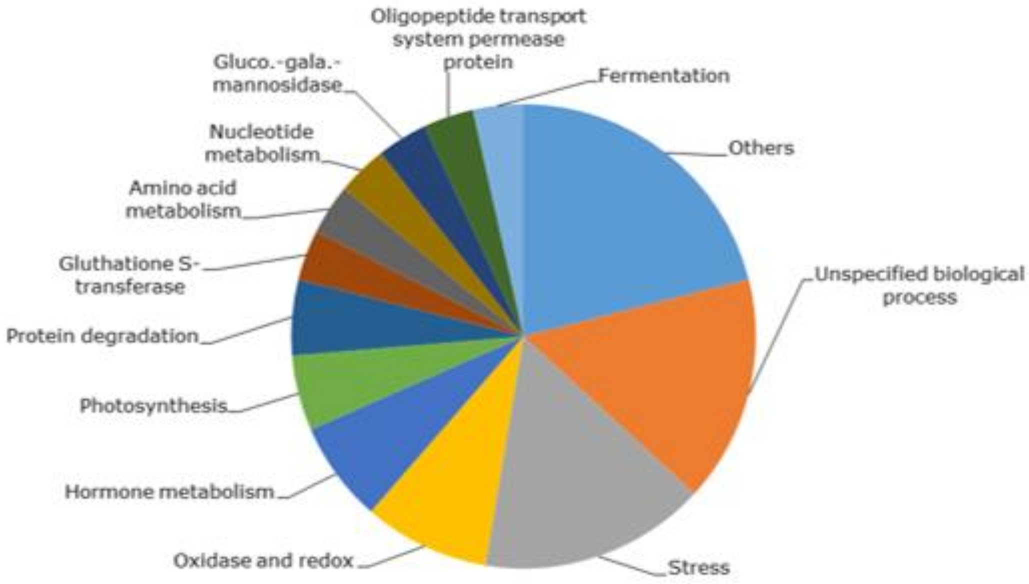
3.3.3 Investigation of proteome changes dependent on the genotype

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. Aliquots of approximately 160 mg of 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 2.4.2.

The methodology used to extract proteins from cocoa pods was not suitable for beans of *T. speciosum*, as these samples formed a gel-like solution when extracted with the urea solution. As a result, protein extracts of *T. speciosum* could not be processed further and therefore a quantitative proteomic analysis on these samples was not performed. Details of these samples are provided in Table 3.

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 once. 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 Appendix 2). 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 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 using the non-parametric Wilcoxon test) among the four cocoa genotypes. Among these proteins, those which showed a sum of the sample-to-reference ion signal ratios 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 investigated. A list of the differentially abundant proteins, including their biological process and function, is provided in Table 13. A pairwise comparison between the genotypes for each protein listed in Table 13 was also carried out, using the non-parametric Whitney Mann test to assess the significance of the differential expression (p <0.05). A graphical representation of the proteins' classification based on their biological processes and functions is provided in Figure 28. Biological processes, for which only one protein was identified and quantified, are labelled as 'Others' in Figure 28.



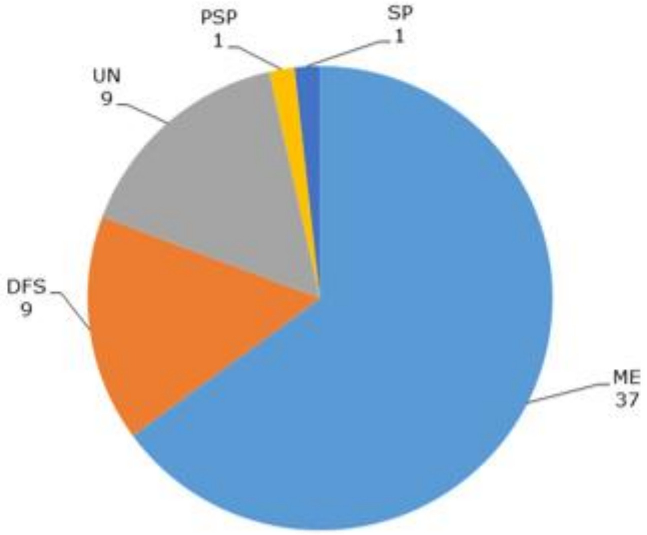


Figure 28. Classification of the differentially abundant proteins listed in Table 13 based on their biological process (upper pie chart) and their function (lower pie chart). 'Others' in the upper pie chart refers to all biological process, for which only one protein was found. The function group labels are as follows: DFS, defence and stress; ME, metabolism and energy; PSP, protein synthesis and processing; SP, storage proteins; UN, unclassified.

Table 13 Pair comparison of differentially abundant proteins among all genotypes obtained from the four cocoa genotypes analysed by label-free LC-MS/MS. The fold change is reported for significantly different proteins only. Proteins which were not significantly different for each pair have been labelled as "/".

ID	Accession	Description	Biological process	Function	ICS 1 vs IMC 67	ICS 1 vs SCA 6	ICS 39 vs IMC 67	ICS 39 vs SCA 6	ICS 1 vs ICS 39	SCA 6 vs IMC 67
1	Thecc1EG029400t1	N-terminal nucleophile aminohydrolases	Protein degradation	ME	6.82	6.95	5.70	5.81	/	/
2	Thecc1EG029392t1	Glutathione S-transferase family protein	Gluthatione S-transferase	ME	5.28	/	5.11	/	/	2.88
3	Thecc1EG025391t1	Beta-amylase 6	Carbohydrate metabolism	ME	2.33	4.90	/	/	2.66	2.10
4	Thecc1EG017184t1	Sulfite oxidase	S-assimilation	ME	2.21	/	/	3.87	4.26	2.01
5	Thecc1EG038258t1	Molybdenum cofactor sulfurase	Co-factor and vitamin metabolism	ME	/	2.20	/	4.12	/	2.59
6	Thecc1EG030320t1	Ethylene-forming enzyme	Hormone metabolism	ME	/	/	3.61	/	4.03	2.21
7	Thecc1EG021639t1	PEBP	Unspecified biological process	UN	/	2.60	/	3.36	/	/
8	Thecc1EG020604t1	Primary amine oxidase	Oxidase	ME	/	/	/	3.25	2.41	/
9	Thecc1EG047098t1	Uncharacterized protein	Unspecified biological process	UN	/	2.09	/	2.19	/	3.18
10	Thecc1EG036433t1	HSP20-like chaperones protein	Stress	DFS	/	/	2.95	2.66	3.15	/
11	Thecc1EG026543t1	Lipoxygenase 1	Hormone metabolism	ME	2.64	2.92	2.72	3.01	/	/
12	Thecc1EG026589t1	Eukaryotic aspartyl protease	Protein degradation	ME	2.91	/	2.88	/	/	/
13	Thecc1EG042578t1	S-adenosyl-L-methionine-dependent methyltransferases protein	Hormone metabolism	ME	/	2.30	2.85	/	2.30	2.85
14	Thecc1EG019372t1	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin protein	Protease inhibitor/seed protein/lipid transfer	SP	/	2.28	/	/	/	2.78
15	Thecc1EG027146t1	HSP20-like chaperones protein	Stress	DFS	/	2.03	/	/	2.77	/
16	Thecc1EG026193t1	Threonine aldolase 1	Amino acid metabolism	ME	2.66	/	/	/	/	/
17	Thecc1EG037345t1	17.6 kDa class II heat shock protein	Stress	DFS	/	2.24	2.21	/	2.65	/
18	Thecc1EG012673t1	21 kDa seed protein*	Stress	DFS	/	/	2.65	/	/	/
19	Thecc1EG025860t2	Uncharacterized protein	Unspecified biological process	UN	/	/	2.26	/	/	2.64
20	Thecc1EG012662t1	21 kDa seed protein*	Stress	DFS	/	/	2.63	/	/	/
21	Thecc1EG038931t1	Xyloglucan endotransglycosylase 6	Cell wall degradation	ME	/	2.57	/	/	/	2.49
22	Thecc1EG006471t1	Flavin-dependent monooxygenase 1	Oxidase	ME	/	/	2.55	/	/	/
23	Thecc1EG030938t1	Cc-nbs-lrr resistance protein	Unspecified biological process	UN	2.40	/	/	2.41	2.53	2.29
24	Thecc1EG036938t1	Aldolase-type TIM barrel	Nucleotide metabolism	ME	2.12	2.47	/	/	2.21	/
25	Thecc1EG006154t1	Glycinamide ribonucleotide synthetase	Nucleotide metabolism	ME	2.45	/	/	/	/	/
26	Thecc1EG041496t1	Stress responsive A/B Barrel Domain	Unspecified biological process	UN	/	/	2.29	2.44	/	/
27	Thecc1EG030354t1	Fumarylacetoacetase	Amino acid metabolism	ME	/	/	/	2.06	/	2.40
28	Thecc1EG019909t2	Carrot EP3-3 chitinase	Stress	DFS	/	/	/	2.39	/	/
29	Thecc1EG040975t1	Alpha/beta-Hydrolases protein•	Gluco.-gala.-mannosidase	ME	/	2.37	/	/	/	/
30	Thecc1EG020603t2	Primary amine oxidase	Oxidase	ME	/	/	/	2.36	/	/

31	Thecc1EG016747t1	Acyl-CoA-binding protein 6	Lipid metabolism	ME	/	/	/	/	/	2.33
32	Thecc1EG022426t1	Thioredoxin protein	Redox	ME	2.33	/	/	/	/	2.21
					2.20	/	2.21	/	/	2.30
33	Thecc1EG008318t1	Aldolase-type TIM barrel	Oligopeptide transport system permease protein	ME	/	/	/	/	/	2.28
34	Thecc1EG022506t1	Monodehydroascorbate reductase seedling isozyme	Redox	ME	/	2.27	/	/	/	/
35	Thecc1EG047057t1	Cystathionine beta-synthase	Unspecified biological process	UN	/	/	/	2.21	/	/
36	Thecc1EG029923t1	Larreatricin hydroxylase	Unspecified biological process	UN	/	/	2.20	/	2.05	/
37	Thecc1EG000245t1	Serine carboxypeptidase S28	Protein degradation	ME	/	2.20	/	/	/	/
38	Thecc1EG021820t1	Tau class glutathione transferase GSTU45	Gluthatione S-transferase	ME	/	/	2.20	/	/	/
39	Thecc1EG025715t1	Uncharacterized protein	Unspecified biological process	UN	/	/	/	/	/	2.19
40	Thecc1EG043707t1	Anti-oxidant 1	Metal handling	ME	2.18	/	/	/	/	/
41	Thecc1EG016386t1	6-Phosphogluconate dehydrogenase	Oligopeptide transport system permease protein	ME	/	/	/	/	2.17	/
42	Thecc1EG014591t1	Malate synthase glyoxysomal	Gluconeogenesis	ME	/	/	2.14	/	/	2.16
43	Thecc1EG042584t1	S-adenosyl-L-methionine-dependent methyltransferases protein	Hormone metabolism	ME	/	2.16	/	/	/	/
44	Thecc1EG034339t1	Dehydrin 2	Stress	DFS	/	/	/	/	/	2.16
45	Thecc1EG035433t1	Alcohol dehydrogenase 1[+]	Fermentation	ME	/	2.13	/	/	/	/
46	Thecc1EG010364t2	Carbonic anhydrase 2 CA2	TCA/organic transformation	ME	/	/	/	/	/	2.11
47	Thecc1EG006694t2	Triosephosphate isomerase	Photosynthesis	ME	/	2.10	/	/	/	/
48	Thecc1EG006498t1	Basic chitinase	Stress	DFS	/	/	2.10	/	/	/
49	Thecc1EG026326t2	Pathogenesis-related protein P2	Stress	DFS	2.07	/	/	/	/	/
50	Thecc1EG029913t1	Alpha/beta-Hydrolases protein[•]	Gluco.-gala.-mannosidase	ME	/	/	/	/	2.05	/
51	Thecc1EG036604t1	Secretory laccase	Secondary metabolism	ME	/	/	/	/	/	2.05
52	Thecc1EG015253t1	RNA binding Plectin/S10 domain-containing protein	Protein synthesis	PSP	/	/	/	/	/	2.05
53	Thecc1EG005533t1	Transketolase	Photosynthesis	ME	2.03	/	/	/	/	/
54	Thecc1EG000770t1	Acetamidase/Formamidase	Photosynthesis	ME	/	2.01	/	/	/	/
55	Thecc1EG014683t1	Hydroxysteroid dehydrogenase 1	Dehydrogenase	ME	/	/	/	/	/	2.01
56	Thecc1EG001447t1	Alcohol dehydrogenase 1[+]	Fermentation	ME	/	2.00	/	/	/	/
57	Thecc1EG001141t1	Lipase/lipooxygenase PLAT/LH2	Unspecified biological process	UN	6.82	6.95	5.70	5.81	/	/

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 protein entries have a 99.5% homology and can therefore be considered to be proteoforms of the same gene. [+]These protein entries have an 87% homology. [•]These protein entries have a 35% homology.

The number of proteins listed in Table 13 whose intensity was highest and lowest in each genotype compared to the others are graphically represented in a histogram in Figure 29.

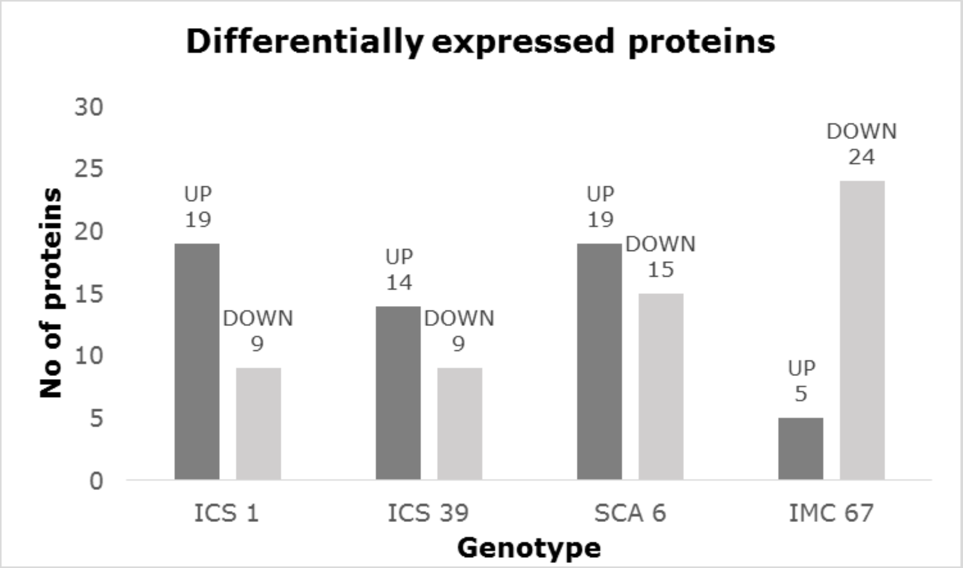


Figure 29. Number of differentially abundant proteins (fold difference of >2) detected at the higher (UP) and lower (DOWN) level in each genotype analysed to investigate the proteome changes dependent on the genotype.

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 abundant proteins listed in Table 13 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 which explained 61% of the total variance, could clearly separate the four cocoa genotypes (see Figure 30). 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 abundant proteins listed in Table 13 is also shown in Figure 30 (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.

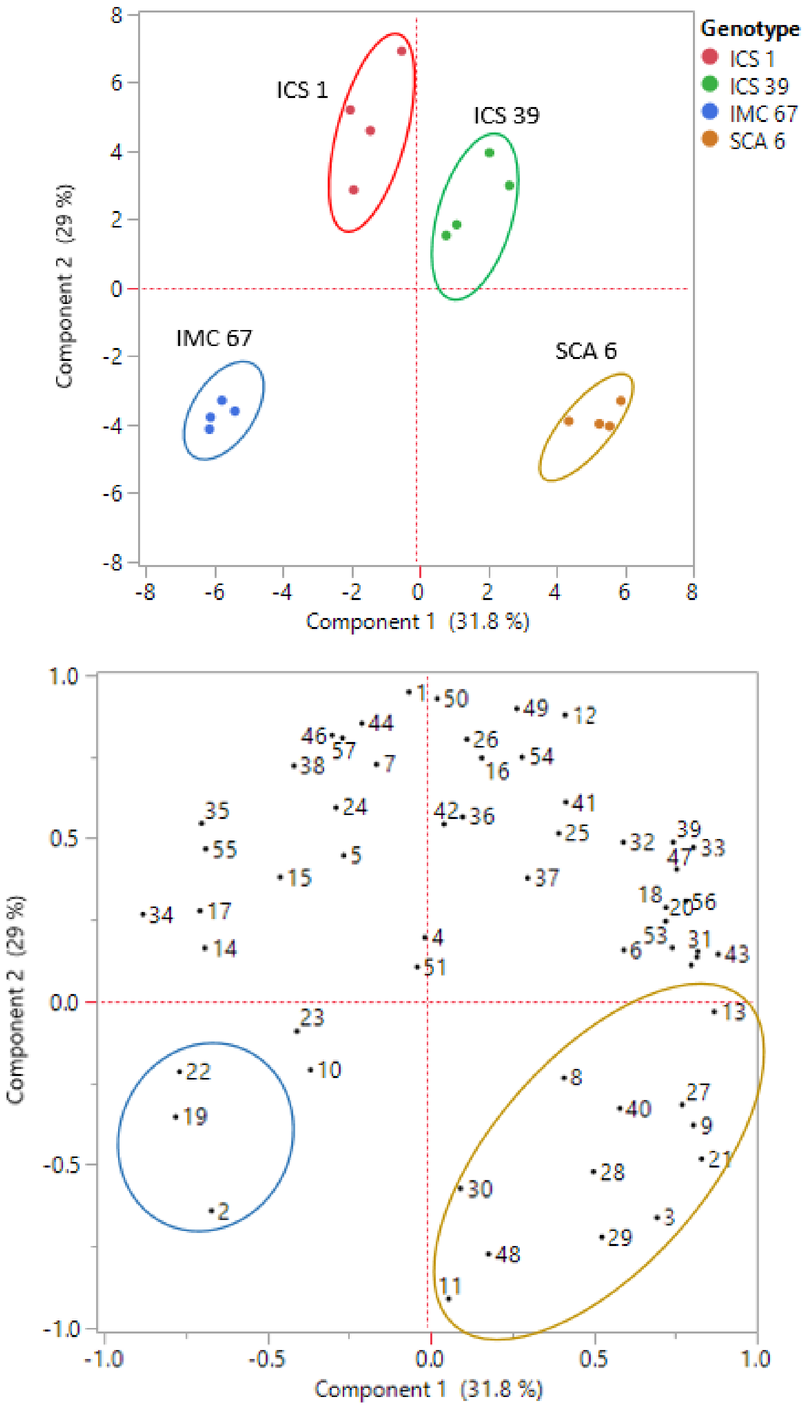


Figure 30. PCA score plot (upper plot) of the 57 differentially abundant proteins listed in Table 1. Preparative replicates of the same genotype are displayed with the same colour. The lower plot shows the PCA loading plot. Each variable is labelled with the corresponding ID number as listed in Table 13. 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 Table 14). The latter two proteins were detected at low levels while the others had a fold change of <3.4.

Table 14. Differentially expressed cocoa bean proteins from IMC 67 trees grown in two different fields: Campus and ICGT.

Accession	Description	Biological process	Function	Fold change	Field*
Thecc1EG029400t1	N-terminal nucleophile aminohydrolases	Protein degradation	ME	3.39	Campus
Thecc1EG030320t1	Ethylene-forming enzyme	Hormone metabolism	ME	2.40	Campus
Thecc1EG012246t1	Oleosin family protein	Lipids metabolism	ME	2.23	ICGT
Thecc1EG016994t2	Esterase	GDSL lipase	ME	2.12	Campus
Thecc1EG036604t1	Secretory laccase	Secondary metabolism	ME	Only in ICGT	ICGT
Thecc1EG005525t1	Ribosomal protein S13A	Protein synthesis	PSP	Only in ICGT	ICGT

* This column indicates in which field the protein is more abundant.
ME, metabolism and energy; PSP, protein synthesis and processing; ICGT, International Cocoa Genebank Trinidad.

3.4 Proteomic analysis by SDS-PAGE

3.4.1 Method development

Method development for the SDS-PAGE analysis was carried out using cocoa beans proteins extract from an Amelonado variety grown in Trinidad. Protein standards in the range 10-250 kDa were also employed for each analysis to assess the general purity and molecular weight distribution of the proteins extracted from cocoa bean. BSA solutions were also loaded on the gel as quality controls to evaluate whether the migration distance on the gel was consistent with the molecular weight of the proteins detected.

For the initial analysis four aliquots from the same protein extract solution with a final protein concentration of 1.9 mg/ml, and two aliquots of a 1.8 mg/ml BSA solution were analysed, setting the voltage of the Mini PROTEAN Tetra Cell to 300 V. The four aliquots showed a similar electrophoretic pattern, with three main electrophoretic bands migrating at approximately 50, 30 and 20 kDa, see Figure 31. Minor bands between 15 and 10 kDa were also observed, however, these bands were not well resolved. BSA was visualised between the 75 and 50-kDa markers, therefore, its migration distance was consistent with its actual molecular weight of 66 kDa. BSA overloading was also observed. The protein markers could clearly be resolved over the range of 10-250 kDa.

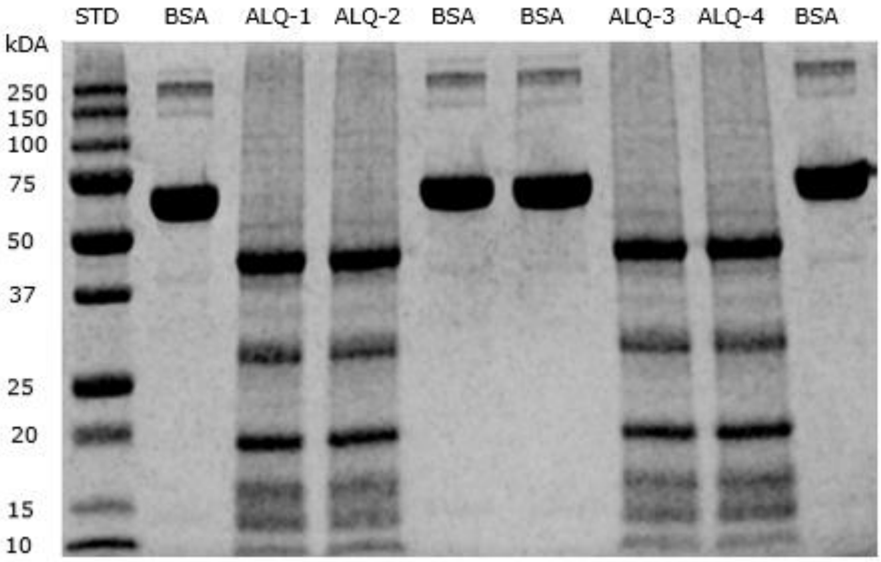


Figure 31. SDS-PAGE gel of protein markers (STD), cocoa protein extracts 1.9 mg/ml (ALQ 1-4) and BSA 1.8 mg/ml. For each sample and standards 10 µl of solution were loaded on a Mini Protean TGX 12% polyacrylamide gel. The gels were stained with Bio-Safe Coomassie Stain

An additional analysis was carried out reducing the concentration of the BSA solution to 0.4 mg/ml, and increasing the concentration of the cocoa protein extracts to approximately 3 mg/ml. In this case 5 aliquots of the same cocoa proteins extract as described above were analysed. The voltage of the Mini PROTEAN Tetra Cell was set to 250 V for the whole duration of the run. A gel acquired with these conditions is shown in Figure 32.

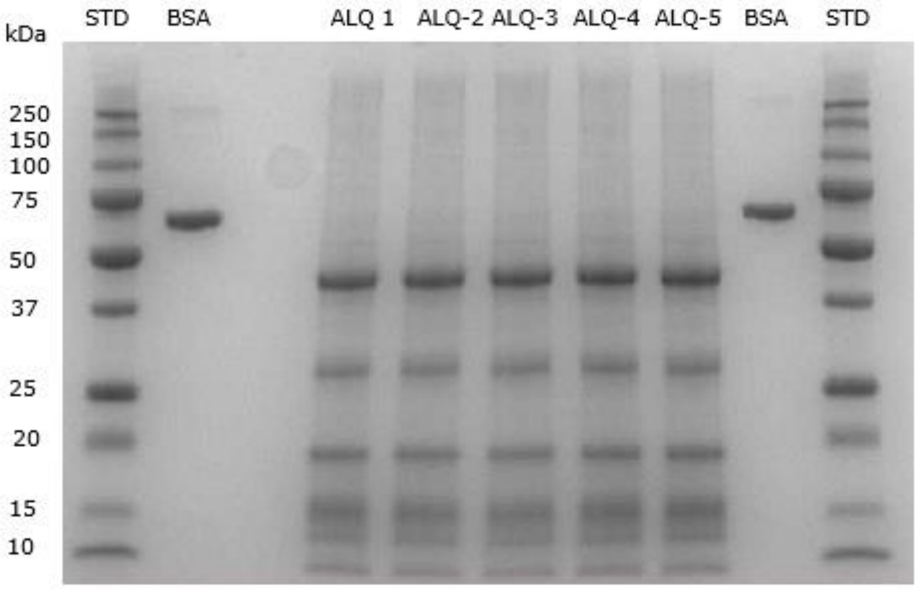


Figure 32 SDS-PAGE gel of protein markers (STD), cocoa protein extracts 3 mg/ml (ALQ 1-5) and BSA 0.4 mg/ml. For each sample and standards 10 µl of solution were loaded on a Mini Protean TGX 12% polyacrylamide gel. The gels were stained with Bio-Safe Coomassie Stain

Decreasing the BSA concentration from 1.8 to 0.4 mg/ml improved significantly the shape of the electrophoretic band of this protein, as in this case no overloading was observed, see Figure 32. The electrophoretic pattern of the cocoa protein extract was similar in the five aliquots analysed. The three main bands migrating at approximately 50, 30 and 20 kDa, were manually integrated and their intensities reported in Table 15. The RSD is expressed as the relative standard deviation of the areas of the bands in each aliquot.

Table 15. Area of the major proteins bands detected in 5 aliquots of the cocoa proteins extracts described in Figure 32

Band kDa	**Areas**					**RSD %**
	ALQ-1	**ALQ-2**	**ALQ-3**	**ALQ-4**	**ALQ-5**	
~ 50	23765	23700	28375	23937	23687	8.3
~ 30	16251	16054	16948	16274	17293	3.2
~ 20	18757	18557	20190	18789	19475	3.5

3.4.2 Effect of harvest time and different trees

To evaluate the effect of different trees on the electrophoretic protein profile of cocoa beans, four biological replicates from the cocoa variety IMC 67 grown at the ICGT on different trees and harvested on the same day were analysed by SDS-PAGE. The electrophoretic gel of these samples is shown in Figure 33.

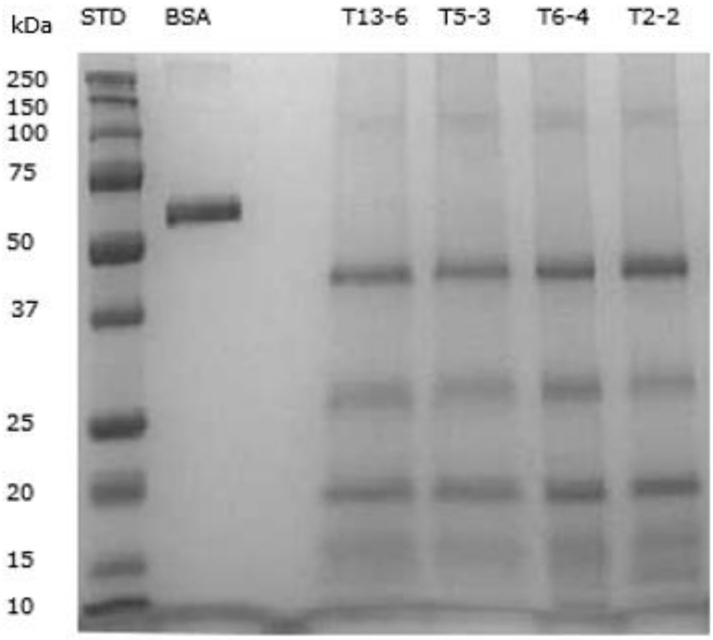


Figure 33. SDS-PAGE of proteins marker (STD), BSA and biological replicates (T13-6, T5-3, T6-4 and T2-2) to assess the effect of different trees and a BSA standard solution. A total of 4 µg and 30 µg of BSA and cocoa samples, respectively were loaded on a Mini Protean TGX 12% polyacrylamide gel. The gels were stained with Bio-Safe Coomassie Stain

Three main electrophoretic bands migrating at approximately 50, 30 and 20 kDa, were observed for all samples analysed, see Figure 33. These bands were integrated, and their intensities reported in Table 16

Table 16 Intensities of the main electrophoretic bands for the experiment to assess the effect of different tree. Fold change is the ratio between the highest and lowest intensity for each band

Band kDa	Intensities of the bands				Fold change
	T2-2	T6-4	T5-3	T13-6	
~ 50	11596	9614	10204	12082	1.3
~ 30	12577	7925	8591	7848	1.6
~ 20	8100	7649	9304	12382	1.6

Minor bands with an apparent molecular weight between 15 and 10 kDa were also detected. Since these bands were not properly resolved, their intensities could not be measured.

Furthermore, four biological replicates from the cocoa variety IMC 67 grown at the ICGT filed on the same tree but harvested on different days, were analysed by SDS-PAGE with the aim to evaluate the effect of harvest time on the electrophoretic proteins profile of cocoa beans. The electrophoretic gel of these samples (see Figure 34) showed a pattern similar to the one observed for the biological replicates analysed to assess the effect of different trees (see Figure 33), with three main well resolved electrophoretic bands showing an apparent molecular weight of approximately 50, 30 and 20 kDa and a series of unresolved bands with a migration distance between 15 and 10 kDa, see Figure 34.

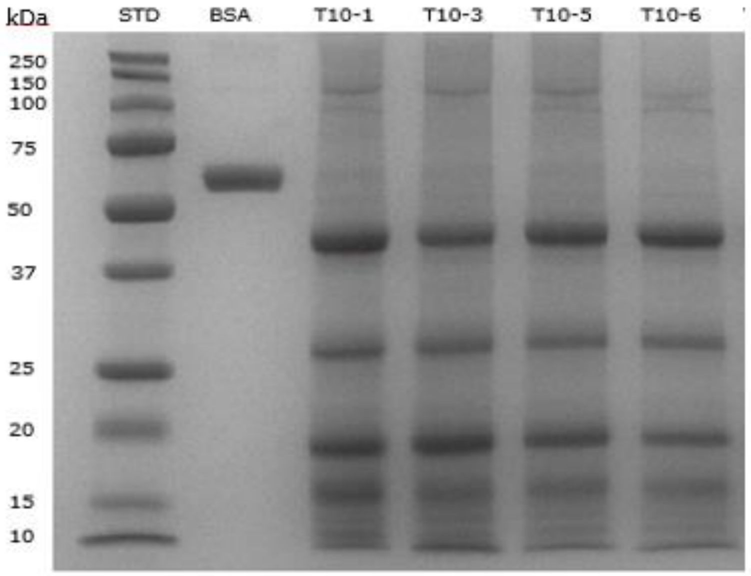


Figure 34. SDS-PAGE gel of protein markers (STD), cocoa protein extracts to assess the effect of harvest time (T10-1, T10-3, T10-5, T10-6) and a BSA standard solution. A total of 4 µg and 30 µg of BSA and cocoa samples, respectively were loaded on a Mini Protean TGX 12% polyacrylamide gel. The gels were stained with Bio-Safe Coomassie Stain

The intensities of the electrophoretic bands between 20 and 50 kDa detected in the biological replicates analysed to assess the effect of different harvest time are listed in Table 17. Fold change is the ratio between the highest and lowest intensities for each band.

Table 17. Intensities of the main electrophoretic bands for the experiment to assess the effect of different harvest time. Fold change is the ratio between the highest and lowest intensity for each band

Band kDa	Intensities of the bands				Fold change
	T10-1	T10-3	T10-5	T10-6	
~ 50	14726	13144	15741	15771	1.2
~ 30	8853	7785	7195	7523	1.2
~ 20	6891	10960	9247	9430	1.6

3.4.3 Investigation of proteome changes dependent on the genotype by SDS-PAGE

The same cocoa varieties described in section 3.3.4 were also analysed by SDS-PAGE in order to assess whether differences in the proteomic profiles of these varieties would be detected using this technique. The electrophoretic gel profiles of the four different cocoa genotypes grown at the ICGT field and the genotype IMC 67 grown at the Campus field is shown in Figure 35. Three main, well resolved, electrophoretic bands with apparent molecular weights of approximately 50, 30 and 20 kDa, and unresolved bands with a migrating distance between the 15 and 10-kDa protein markers were observed in all samples analysed, see Figure 35.

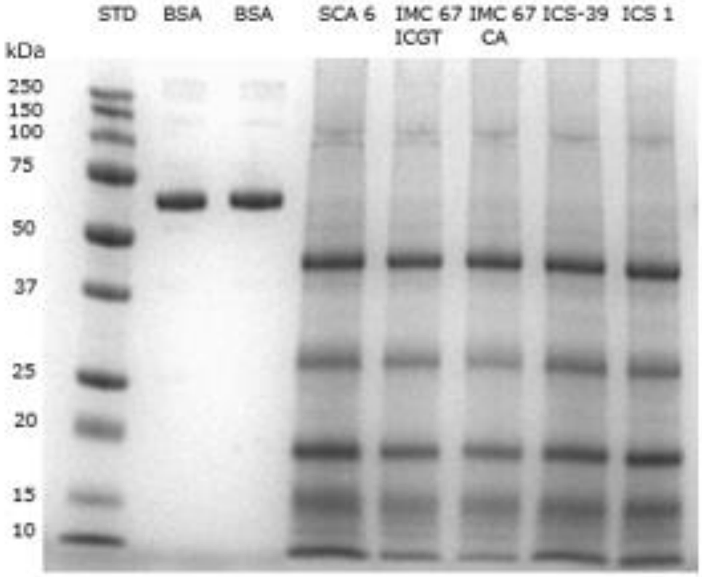


Figure 35. SDS-PAGE gel of protein markers (STD), protein extracts from different cocoa varieties (IMC 67 ICGT, IMC 67 CA, ICS 39 AND ICS 1) and BSA standard solutions. A total of 4 µg and 30 µg of BSA and cocoa samples, respectively were loaded on a Mini Protean TGX 12% polyacrylamide gel. The gels were stained with Bio-Safe Coomassie Stain

The electrophoretic bands with apparent molecular weight of approximately 50, 30 and 20 kDa were integrated and the intensities reported in Table 18. Fold change is the ratio between the highest and lowest intensities for each band.

Table 18. Intensities of the main electrophoretic bands for the experiment to assess changes in proteome dependent on genotype by SDS-PAGE. Fold change is the ratio between the highest and lowest intensity for each band

Band kDa	Intensities of the bands					Fold change
	SCA 6	IMC 67 ICGT	IMC 67 CA	ICS 39	ICS 1	
~ 50	38876	29527	30190	37779	33641	1.3
~ 30	29195	19404	19007	29693	24110	1.5
~ 20	43784	30263	27257	36607	32974	1.6

3.5 Method development for analysis of free peptides by LC-MS

Beans of Ghanaian origin fermented with the heaps method were used for developing a methodology to characterise and quantify free peptides in cocoa beans by LC-MS/MS. For the initial analysis described in section 2.5, approximately 200 mg of fermented beans were defatted and subsequently extracted with a methanol:water 70:30 (v/v) solution containing PVPP, as this polymer binds polyphenols which could form complexes with peptides [75]. Injection of a solution with a high percentage of methanol on a reverse phase column could cause loss of very polar peptides which are poorly retained on this column, and peak shape distortion as well. As a result, an aliquot of 200 µl of the methanolic peptide extracts solution was dried down under a stream of nitrogen and reconstituted in aqueous 0.1% (v/v) TFA. For this analysis three portions of the same sample were extracted separately to give three preparative replicates. The peptide extracts were analysed using the microflow Dionex Ultima 3000 UHPLC coupled to the Q Exactive Orbitrap with a DDA experiment. Each replicate was injected only once. The raw files were initially searched against the Cacao Matina 1-6 genome database published by Motamajor *et al.* [74]. Since the aim of this experiment was to identify and quantify free peptides, the samples were not digested and 'no enzyme' was selected for protein digestion in the search parameters. The results of this analysis show that a total of 14 peptides were detected in all three replicates analysed, of which 12 could be identified as proteolysis products of a 21 kDa albumin, and the remaining two peptides shared amino acids sequences with cocoa vicilin. The same raw files were also searched against a custom-made protein database containing the 100 most abundant cocoa proteins as listed in Appendix 1, see section 3.1 and Table 6 for more details. A total of 82 peptides were detected in all three replicates searched against this database. As shown in Table 19, the vast majority of these peptides were assigned to a 21 kDa albumin and a cocoa vicilin.

Table 19. Free peptides detected in fermented beans extracted with a methanol:water 70:30 (v/v) solution and searched against a database containing the 100 most abundant cocoa beans proteins as listed in Appendix 1

Accession	Protein description	No. of peptides	Empai value
Thecc1EG012658t1	21 kDa seed protein	32	39.73
Thecc1EG020665t1	Vicilin-A	28	19.66
Thecc1EG030267t1	Peroxygenase 2	7	2.55
Thecc1EG029926t1	Saposin-like aspartyl protease family protein	3	1.13
Thecc1EG041714t1	Larreatricin hydroxylase	2	1.50
Thecc1EG015612t1	HSP20-like chaperones superfamily protein	2	1.53
Thecc1EG042785t1	UDP-Glycosyltransferase superfamily protein	2	0.51
Thecc1EG041085t1	Pyrophosphate-fructose 6-phosphate 1-phosphotransferase	2	2.10
Thecc1EG026589t1	RmlC-like cupins superfamily protein	1	6.31
Thecc1EG036433t1	Eukaryotic aspartyl protease family protein	1	1.76
Thecc1EG020975t1	HSP20-like chaperones superfamily protein	1	3.46
Thecc1EG000075t1	Catalase 2	1	0.43
	Total	82	

The drying step to remove methanol and the reconstitution in an aqueous buffer prior to LC-MS/MS analysis can result in loss or degradation of peptides during this process. Therefore, an additional experiment was performed using aqueous 0.5 % (v/v) TFA as extraction solution. The amount of fermented and defatted beans and the volume of the extraction solution were not changed, and the same amount of PVPP was also added to the extraction solution. In this case 400 µl of the peptide extracts were loaded on SPE cartridges. The eluate was dried down and re-dissolved in 100 µl of aqueous 0.1% (v/v) TFA. Although a drying step was also required with this approach, in this case the volume to be dried down was reduced from 200 to 50 µl, shortening the drying time from approximately 4 hours to around 2 hours, minimising as a result loss or degradation of peptides during this process. Using SPE cartridges allowed the removal of salts and other high polar interferences from the peptides extracts, and the concentration of the peptides solutions by a factor of 4. Three preparative replicates were prepared for this experiment and analysed by LC-MS/MS. Each replicate was injected once. The results of this analysis showed that a total of 123 peptides originating mainly from a 21 kDa albumin and vicilin were detected in all three replicates analysed, as shown in Table 20:

Table 20. Free peptides detected in fermented beans extracted with a 0.5% TFA solution by searching the LC-MS/MS data against a home-made database of the 100 most abundant proteins in cocoa beans

Accession	Protein description	No. of peptides	emPAI value
Thecc1EG020665t1	Vicilin-A	60	19.66
Thecc1EG012658t1	21 kDa seed protein	45	39.73
Thecc1EG020975t1	Peroxygenase 2	4	3.46
Thecc1EG041714t1	HSP20-like chaperones superfamily protein	3	1.50
Thecc1EG022427t1	Enolase	2	2.71
Thecc1EG030267t1	Saposin-like aspartyl protease family protein	2	2.55
Thecc1EG026326t2	Pathogenesis-related protein P2	2	1.32
Thecc1EG015612t1	UDP-Glycosyltransferase superfamily protein	1	1.53
Thecc1EG017080t2	Glyceraldehyde-3-phosphate dehydrogenase C2	1	6.49
Thecc1EG026589t1	Eukaryotic aspartyl protease family protein, putative	1	6.31
Thecc1EG034805t1	Pyrophosphate--fructose 6-phosphate 1-phosphotransferase subunit alpha	1	3.43
Thecc1EG037346t1	Heat-shock protein	1	5.03
	Total	123	

The same data were also searched against a larger database consisting of 897 proteins which had been identified during the characterisation of the cocoa proteome using a custom-made Uniprot\Tremble database with entries restricted to *Theobroma cacao* only, see Table 6. In this case a total of 130 peptides were detected in all three replicates analysed. A list of the proteins associated with these peptides is provided in Table 21.

Table 21. Free peptides detected in fermented beans extracted with a 0.5% TFA solution by searching the LC-MS/MS data against a database consisting of 897 proteins from a custom-made Uniprot\Tremble database (see section 3.1 and Table 6 for database details)

Accession	Protein description	No. of peptides	Empai value
A0A061EM85	Vicilin-A, putative	57	19.66
A0A061G2K6	21 kDa seed protein	46	39.73
A0A061EMW3	Peroxygenase 2 isoform 1	4	3.46
A0A061F2P1	Lipoxygenase	4	0.23
A0A061GFX2	Saposin-like aspartyl protease family protein	3	2.55
A0A061F2B4	Pathogenesis-related protein P2 isoform 2 (Fragment)	2	1.32
A0A061GZI5	RmlC-like cupins superfamily protein	2	2.10
A0A061FVK5	21 kDa seed protein	2	4.41
A0A061FEW0	Pyrophosphate--fructose 6-phosphate 1-phosphotransferase	2	3.43
A0A061GKC6	Heat-shock protein, putative	2	5.03
A0A061GFY7	Larreatricin hydroxylase isoform 1	1	1.13
A0A061F3Y1	Eukaryotic aspartyl protease family protein	1	6.31
A0A061DK50	Glutathione S-transferase PHI 9 isoform 1	1	0.49
A0A061F0S7	Enolase	1	0.48
A0A061ECJ8	Glyceraldehyde-3-phosphate dehydrogenase	1	6.49
A0A061FKT5	HSP20-like chaperones superfamily protein	1	1.76
	Total	130	

To evaluate whether increasing the length of the injection time for the LC-MS/MS analysis would result in a higher number of identified peptides, analysed with injection times of 200 and 300 ms were performed out using the peptide extract obtained with aqueous 0.5% TFA. A total of 139 peptides were detected when increasing the injection time for the LC-MS/MS analysis to 300 ms, while only 94 peptides were detected when the same samples were analysed with an injection time of 200 ms.

Peptides with hydrophobic amino acids may not be highly soluble in aqueous TFA, therefore the addition of an organic modifier may be required to increase the solubility of these peptides. To assess whether adding an organic solvent to the extraction solution would yield a higher number of extracted peptides, three portions of approximately 200 mg of fermented beans were extracted with an aqueous 0.5% TFA:methanol (80:20; v:v) solution containing PVPP. Prior to desalting with SPE, 400 µl of the aqueous 0.5% TFA:methanol extracts were diluted to a final volume of 2 ml with aqueous 0.5% TFA. The number of peptides detected in the experiments to assess the injection time and addition of organic modifier are shown in Table 22.

Table 22. Number of peptides detected in files searched against Uniprot\Tremble database containing 897 entries

Extraction solution	IT time for MS/MS injection	No. of peptides detected in all replicates
Aqueous 0.5 % TFA	300 ms	139
Aqueous 0.5 % TFA:methanol (80:20; v:v)	300 ms	155
Aqueous 0.5 % TFA	200 ms	94

The results listed in Table 22 indicate that using aqueous 0.5 % TFA:methanol (80:20; v:v) as extraction solution, and increasing the injection time for the MS/MS analysis from 200 to 300 ms resulted in the highest number of identified peptides. A higher proportion of peptides originating from vicilin was detected using these conditions, as shown in in Table 23. The complete list of these peptides including their amino acid sequences and intensities is provided in Appendix 3.

Table 23. Free peptides detected in fermented beans extracted with an aqueous 0.5 % TFA:methanol (80:20; v:v) solution by searching the LC-MS/MS data against a database including 897 proteins from a custom-made Uniprot\Tremble database (see section 3.1 for database details)

Accession	Protein description	No. of peptides	emPAI value
A0A061EM85	Vicilin-A	80	19.66
A0A061G2K6	21 kDa seed protein	55	39.73
A0A061EMW3	Peroxygenase 2 isoform 1	5	3.46
A0A061GFX2	Saponin-like aspartyl protease family protein	4	2.55
A0A061F2P1	Lipoxygenase	3	0.23
A0A061F2B4	Pathogenesis-related protein P2 isoform 2 (Fragment)	2	1.32
A0A061GZI5	RmlC-like cupins superfamily protein	2	2.10
A0A061GKC6	Heat-shock protein, putative	2	5.03
A0A061DK50	Glutathione S-transferase PHI 9 isoform 1	1	0.49
A0A061F3Y1	Eukaryotic aspartyl protease family protein	1	6.31
	Total	155	

The vast majority of the free peptides detected in all experiments carried out were associated with sequences of vicilin and a 21 kDa seed albumin. As the samples extracted with the aqueous 0.5 % TFA:methanol solution returned the highest number of detected peptides when searched against a custom-made Uniprot\Tremble database containing 897 proteins, only the peptides detected in these samples were further evaluated with respect to the sequence coverage and cleavage sites on vicilin and the 21 kDa albumin. A sequence coverage of 69% was observed for the 21 kDa albumin, and a total of 55 cleavage sites were localised in the sequence of this protein, see Figure 36.

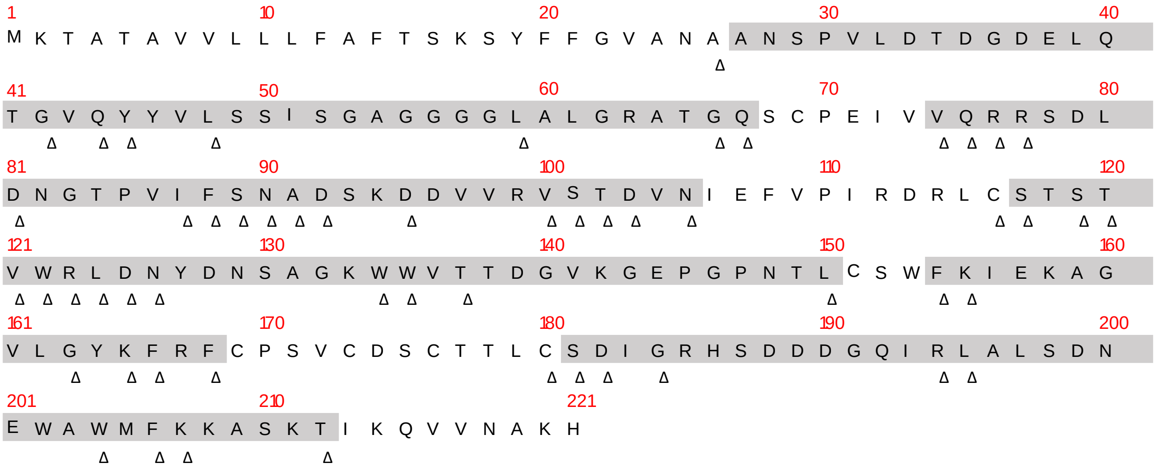
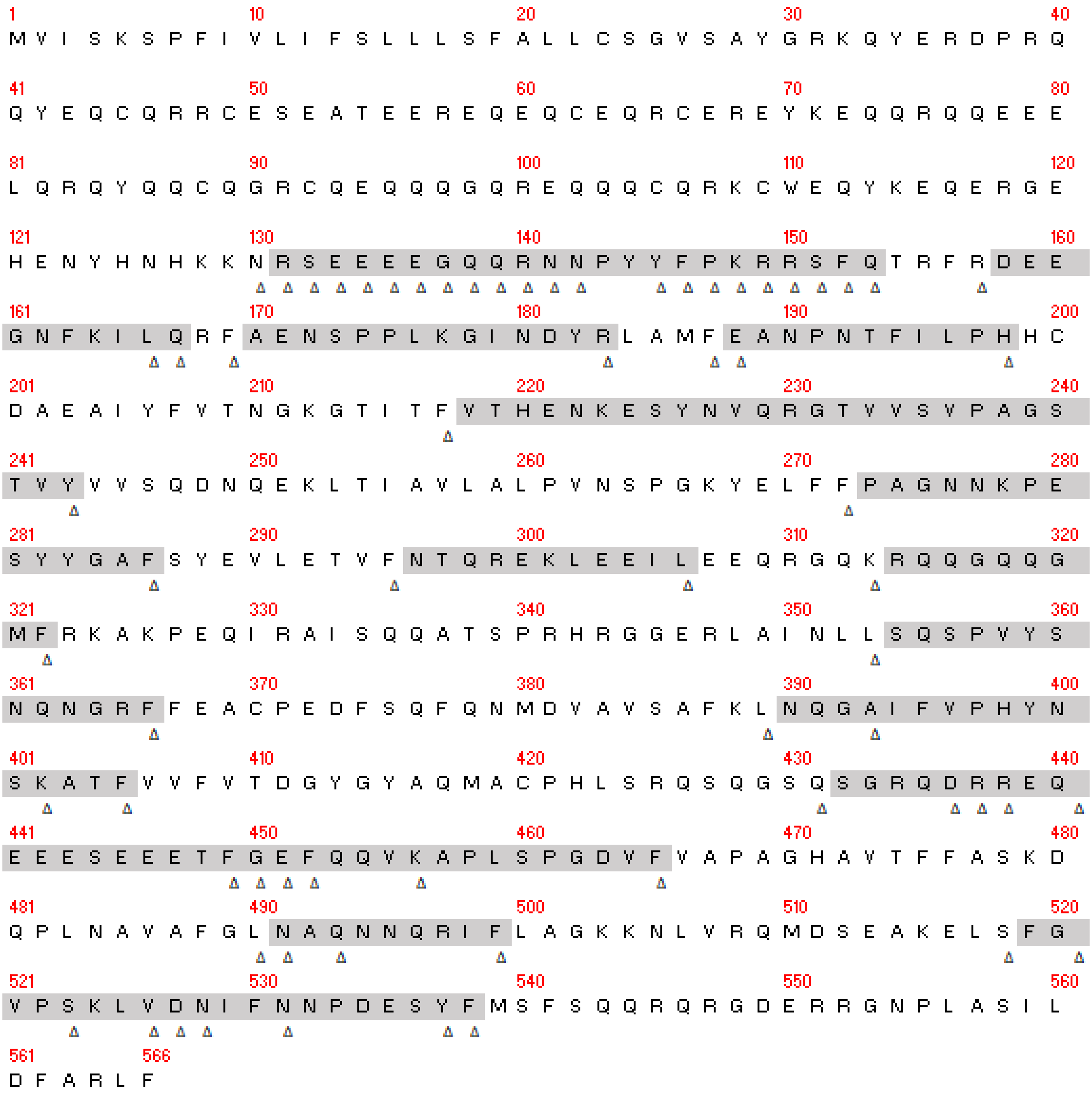

1 10 20 30 40
M K T A T A V V L L L F A F T S K S Y F F G V A N A A N S P V L D T D G D E L Q
41 50 60 70 80
T G V Q Y Y V L S S I S G A G G G G L A L G R A T G Q S C P E I V V Q R R S D L
81 90 100 110 120
D N G T P V I F S N A D S K D D V V R V S T D V N I E F V P I R D R L C S T S T
121 130 140 150 160
V W R L D N Y D N S A G K W W V T T D G V K G E P G P N T L C S W F K I E K A G
161 170 180 190 200
V L G Y K F R F C P S V C D S C T T L C S D I G R H S D D D G Q I R L A L S D N
201 210 221
E W A W M F K K A S K T I K Q V V N A K H

Figure 36 Sequence of 21 kDa cocoa albumin. Identified peptides sequences are highlighted. Cleavage sites are indicated by a small triangle.

The amino sequences of the combined free peptides originating from vicilin covered only 39% of the sequence of this protein, and a total of 68 cleavage sites were found in the sequence of this protein, see Figure 37.



```
1                 10                  20                  30                  40
M V I S K S P F I V L I F S L L L S F A L L C S G V S A Y G R K Q Y E R D P R Q

41                50                  60                  70                  80
Q Y E Q C Q R R C E S E A T E E R E Q E Q C E Q R C E R E Y K E Q Q R Q Q E E E

81                90                  100                 110                 120
L Q R Q Y Q Q C Q G R C Q E Q Q Q G Q R E Q Q Q C Q R K C W E Q Y K E Q E R G E

121               130                 140                 150                 160
H E N Y H N H K K N R S E E E E G Q Q R N N P Y Y F P K R R S F Q T R F R D E E

161               170                 180                 190                 200
G N F K I L Q R F A E N S P P L K G I N D Y R L A M F E A N P N T F I L P H H C

201               210                 220                 230                 240
D A E A I Y F V T N G K G T I T F V T H E N K E S Y N V Q R G T V V S V P A G S

241               250                 260                 270                 280
T V Y V V S Q D N Q E K L T I A V L A L P V N S P G K Y E L F F P A G N N K P E

281               290                 300                 310                 320
S Y Y G A F S Y E V L E T V F N T Q R E K L E E I L E E Q R G Q K R Q Q G Q Q G

321               330                 340                 350                 360
M F R K A K P E Q I R A I S Q Q A T S P R H R G G E R L A I N L L S Q S P V Y S

361               370                 380                 390                 400
N Q N G R F F E A C P E D F S Q F Q N M D V A V S A F K L N Q G A I F V P H Y N

401               410                 420                 430                 440
S K A T F V V F V T D G Y G Y A Q M A C P H L S R Q S Q G S Q S G R Q D R R E Q

441               450                 460                 470                 480
E E E S E E E T F G E F Q Q V K A P L S P G D V F V A P A G H A V T F F A S K D

481               490                 500                 510                 520
Q P L N A V A F G L N A Q N N Q R I F L A G K K N L V R Q M D S E A K E L S F G

521               530                 540                 550                 560
V P S K L V D N I F N N P D E S Y F M S F S Q Q R Q R G D E R R G N P L A S I L

561      566
D F A R L F
```

Figure 37 Sequence of cocoa vicilin. Identified peptides sequences are highlighted. Cleavage sites are indicated by a small triangle.

4 Discussions

4.1 In-depth characterisation of the cocoa bean proteome by LC-MS (results section 3.1)

The chromatographic profiles of the fractionated and unfractionated tryptic digests showed a series of peaks eluting between 20 and 90 minutes, corresponding to a percentage of organic modifier ranging from 10 to 30 %. These results indicate that most of the detected peptides are moderately hydrophilic, since a low percentage of organic modifier was required to elute these compounds from a reverse phase C18 chromatographic column.

This work is the first attempt to characterise the whole cocoa bean proteome by nano-UHPLC-ESI MS/MS analysis using tryptic digests of cocoa bean protein extracts. The results obtained in this study show that the genome of *Theobroma cacao* published by Motamayor *et al.* [74] provides a relatively comprehensive protein database for proteomic analyses, as more than 1000 proteins were identified when the entries from the fractionated and unfractionated samples were combined. A similar approach was employed by Wang *et al.* [54] to quantify and identify proteins in cocoa beans, as described in section 1.7.3. However, although the authors provided a list of all identified proteins in cocoa beans, only 887 proteins were identified, which is considerably lower when compared to the over 1000 proteins found in this PhD project. In addition to that, classification of proteins based on their functions and abundances was not reported. More importantly, results related to this PhD project were initially presented at a BSPR conference in July 2015, which took place several months before the work by Wang *et al.* [54] was published. Therefore, to the best of my knowledge the statement about the originality of the presented approach to characterise the cocoa bean proteome is still valid. Approximately 30 % more proteins were identified in the fractionated sample compared to the unfractionated one. Due to the stochastic nature of DDA acquisition, the precursor ions of low abundant peptides selected for MS/MS fragmentation will vary each time the same sample is re-injected. Therefore, a higher number of replicates for the same sample will result in a higher number of identified proteins. Considering that the entries for the fractionated sample resulted from merging the results of three separate raw files, while only one raw file was used for the unfractionated sample, the increased number of identified proteins observed for the fractionated sample could also be ascribed to the higher number of raw files processed for this sample compared to the unfractionated sample.

Most of the proteins detected in the unfractionated sample were also present in the fractionated one, indicating a high degree of overlap between the two samples sets. Some degree of overlap was also observed between the fractions of the fractionated sample. The highest number of proteins were detected in the urea-soluble fraction, as this fraction was extracted with a solution which can cover a higher range of protein solubility. The salt-soluble fraction represented the fraction with the lowest number of identified proteins.

Searches carried out using different databases showed that the three *Theobroma cacao*-specific databases led to the search results with the highest numbers of proteins being identified, confirming previous plant proteomic results [76] and showing that the use of a more species-specific database can increase the number of identified proteins.

The majority of the identified proteins were linked to the function group *metabolism and energy* with 521 entries which accounted for 57.4% of the total number of identified proteins. The most abundant proteins within this function group belonged to the oleosins family. There is a direct correlation between the content of oil in seeds and the level of oleosin [77]. However, the exact role of oleosins in oil accumulation has not yet been elucidated [77]. Several phosphoglycerate kinases and glyceraldehyde-3-phosphate dehydrogenases, two enzyme families which play an important role in glycolysis and the Calvin cycle, were among the most abundant proteins linked to *metabolism and energy*. Proteins belonging to the aldolases family were also present at significant levels among the proteins involved with *metabolism and energy*.

Proteins involved with *protein synthesis and processing* accounted for 19.7% of the total number of identified proteins (179 entries). A cyclophilin peptidyl-prolyl cis-trans isomerase showed the highest abundance within this group. Cyclophilins are present in all organisms which have been studied so far and have a common domain of 109 amino acids [78]. This class of proteins reduce the energy state of the peptidyl-prolyl bond in the sterically unfavoured cis form, increasing the conversion rate of the cis-trans isomerization of this peptide bond [78]. A high number of molecular chaperones were found within the *protein synthesis and processing* group. This class of proteins interacts with the non-native state of other protein molecules and plays a crucial role in the folding of newly synthesized polypeptides and assembly of structures with multiple units, keeping proteins in unfolded states suitable for translocation across membranes, and ensuring that proteins remain unfolded during cellular stress [79]. Several ribosomal proteins were also detected within this category. These proteins are responsible for translating RNA code into amino acid sequences during protein synthesis [80]. A significant number of identified proteins (64 entries and 7.1% of the total protein number) could be linked to *defence and stress*. Several heat-shock proteins and chitinases were also found within this function group. Heat-shock proteins are synthesised when the plant is exposed to adverse environmental factors and therefore determine the ability of plants to survive under such unfavourable conditions [81]. These proteins keep other proteins in an unfolded state when the plant is subjected to elevated temperatures, ensuring that irreversible aggregation and denaturation does not occur in these conditions [81]. Chitinases play a crucial role in plant defence against pathogens as they hydrolyse chitin, which is a structural component of the cell wall of many phytopathogenic fungi [82].

Vicilin and albumin were classified as *storage proteins* and showed the highest BSA-normalised emPAI value among all detected proteins, although only eight entries could be linked to this function group. However, if the relative percentages of the classified proteins are based on the sum of the BSA-normalised emPAI values of all proteins within the same function group, *storage proteins* are the second most abundant protein group after *metabolism and energy* (Figure 25, lower pie chart). The ratio albumin/vicilin in the water-soluble and salt-soluble fractions were 3.5 and 0.9, respectively, indicating that fractionation resulted in an enrichment of vicilin in the salt-soluble fraction. The ratio albumin/vicilin in the urea-soluble fraction was 4.3. The relative amounts of vicilin and albumin compared to the total protein amount in the whole fractionated sample were 3.9 and 11.5 %, respectively. These levels are lower than the reported values of between 43 and 23 % for vicilin [12, 19], and 52 and 14 % for albumin [12, 19] reported in the literature. However, these values have been calculated based on selective solubilisation of proteins [12, 19], or by quantifying the bands of these proteins in SDS-PAGE gels [12, 19]. Separation of specific protein classes into different fractions using selective extraction solutions may be inefficient, as complete isolation of the proteins of interest into distinct fractions may not be achieved, which can lead to an overestimation of the protein amount. Protein bands in a SDS-PAGE gel usually contain several different proteins, therefore, using this technique to quantify specific proteins can result in an overestimation in the quantitation of specific proteins.

A total of four entries were identified within the albumin protein class, while only one vicilin was detected. Among the albumin class, three entries were identified as 21 kDa seed proteins whose gene function is associated with an endopeptidase inhibitor activity, and the remaining entry was described as a 2S storage albumin with a nutrient reservoir activity. Two of the 21 kDa seed albumins contained 219 residues while the other entry was composed of 221 residues. The 21 kDa seed albumins with 219 residues showed a homology of 99.5 %, as they differed only by the presence of an aspartic acid residue at the N-terminal position 157 and an asparagine residue at the same position. As a result, these two proteins are likely to be expressed from the same gene. However, a homology of only 80% was observed between the two 21 kDa albumins with 219 residues and the one with 221 residues, indicating that there were some differences between the sequences of these albumins. The 21 kDa seed albumin with 221 residues was detected at a much higher level compared to the other two entries. Two proteins belonging to the cupins family proteins with nutrient reservoir activity were also detected within the *storage proteins* function.

4.1.1 Conclusions

More than 1000 proteins could be identified in tryptic digests of cocoa bean protein extracts analysed by nano-UHPLC-ESI MS/MS. The highest number of proteins were identified with searches carried out using the three *Theobroma cacao*-specific databases showing that the use of a more species-specific database can increase the number of identified proteins.

The majority of the identified proteins were linked to the function group *metabolism and energy* and *protein synthesis and processing*. The storage proteins vicilin and albumin showed the highest abundance among the identified proteins.

The presented methodology may benefit from further optimisation in terms of protein extraction and chromatographic separation. However, its current performance and the dataset obtained already provide a good platform for studies aimed at gaining a better understanding of the proteomic profile of cocoa beans and present the largest proteome dataset for cocoa beans to date.

4.2 Method development for quantitative proteomic analysis by LC-MS (results section 3.2)

4.2.1 Improvement of the desalting step and chromatographic separation for the Fusion Orbitrap analysis (results section 3.2.1)

The initial gradient employed for the characterisation of the cocoa bean proteome produced a chromatographic pattern for the tryptic peptides which was compressed between 20 and 90 minutes, see Figures 19**-**22**.** Reducing the steepness of the gradient and doubling the column length at the same time resulted in a wider peptide elution pattern ranging from 20 to 170 minutes, which also improved considerably the chromatographic separation of the tryptic peptides, see Figure 26, when compared to the chromatographic profile of the peptides analysed for the initial characterisation of the cocoa beans proteome, see Figures 19-22. The improved chromatographic conditions allowed a higher sensitivity of the method in terms of identified proteins, as in this case an average of 880 proteins were identified in an unfractionated sample, while 704 proteins had been identified in the same sample analysed with the initial method. As the chromatographic separation of the method is improved, the number of co-eluted peptides is reduced, which in turn decreases ion suppression in the ESI source due to signal saturation, as a lower number of peptides are simultaneously ionised, reducing therefore charge competition in the ESI process. Data dependent acquisition is also more efficient, as a higher number of MS/MS spectra can be acquired for low abundant peptides, if these are chromatographically separated from the high abundant ones. A high level of reproducibility was achieved in terms of protein identification, as an average of 880 proteins was identified in 10 replicate injections of the same sample with an RSD of 2%, expressed as the variation in the number of identified proteins between the replicates, see Table 8. The method was also highly reproducible for label-free quantitation using replicate protocol with Mascot Distiller, as most of the proteins quantified in 4 replicate injections of the same sample (837 out of 880), showed an RSD expressed as protein abundances between the replicates within 20%.

Although the method delivered reproducible quantitative results for replicate injections of the same sample, the reproducibility of four aliquots of the same sample extract containing 10 µg of proteins digested and desalted separately using C18 ZipTips® was poor, suggesting that either the digestion or desalting step were contributing to the variability of the results. Replacing C18 ZipTips with Strata SPE reverse phase cartridge resulted in an increased reproducibility of the label-free quantitation method, although the results were still not optimal. Additional experiments confirmed that the digestion process was not affecting the reproducibility of the method, as the results from the quantitative analysis of six aliquots taken from the same tryptic digest solution and desalted separately were similar to those obtained from aliquots digested and desalted separately. A drop in sensitivity in terms of identified proteins was observed in the instrument during method development. This issue could not be ascribed to the replacement of C18 ZipTips with Strata SPE cartridges, as when aliquots desalted with both protocols were analysed within the same sequence, a slightly high number of proteins were identified in the aliquots desalted with Strata SPE compared to C18 ZipTips.

A higher reproducibility was obtained increasing the amount of digested proteins from 10 to 160 µg, keeping the ratio proteins:trypsin 50:1. For the desalting step carried out with C18 ZipTips, the amount of peptides loaded was kept to 10 µg as this is the maximum capacity of these tips. As the digested solutions had a concentration of peptides of approximately 0.5 µg/µl, only 20 µl of this solution were aliquoted, dried down and reconstituted into 10 µl of 0.1% TFA prior to loading on the ZipTip. This solution had a much lower concentration of salts and chaotropic agents due to the reduced volume which had to be dried down, resulting in a lower viscosity. This made it easier to pipette the solutions in and out of the ZipTip which may have accounted for the improved reproducibility of the results. The higher concentration of digested proteins also minimised the loss of peptides due to absorption to the tube walls which could have negatively affected the reproducibility of the results. The combining effect of replacing Strata SPE cartridges with SOLA reverse phase cartridges and increasing the amount of digested proteins from 10 to 160 µg, considerably improved the reproducibility of the results. The fact that SOLA cartridges were placed on the same position of the well for each desalted solution, may have increased the reproducibility of the results, as in this case the flow rates for the various steps involved in the desalting protocol could be better controlled. A similar protocol was employed for the three clean-up cartridges.

4.2.2 Improvement of the LC-MS/MS method for the Qexactive Orbitrap analysis (results chapter 3.2.2)

The chromatographic separation of the tryptic digests with an injection volume of 10 µl was considered satisfactory for the analytical column used for this test, see Figure 27. It was not possible to evaluate the performance of the column with a greater length, as 15 cm is the maximum available length for 0.1 mm ID columns. The tryptic digests were desalted using SOLA reverse phase cartridges, as this desalting protocol proved to be the optimal one when method development with the Fusion Orbitrap was performed. The method improvement then focused on the injection time for MS/MS acquisition and the injection volume of tryptic digests on column. Increasing the injection time for MS/MS acquisition allows accumulation of a larger number of ions on the C-Trap before the AGC target is reached, which in turn may favour the acquisition of low abundant peptides by increasing their signal to noise ratio above the threshold set in the method. This may explain why more proteins were identified when the injection time for MS/MS acquisition was increased from 200 to 300 ms. By increasing the injection volume, a higher amount of peptides is loaded on the column. However, if this amount is too high and exceeds the column capacity, overloading will occur, which would result in a distorted peak shape and a loss of resolution. Increasing the injection volume from 10 to 15 µl allowed the identification of a high number of proteins, and it neither affected the peak shape of peptides nor caused a loss of resolution. Based on these results the parameters for the LC-MS/MS method which allowed the identification of the highest number of proteins were an injection time for the MS/MS acquisition of 300 ms, and an injection volume of 15 µl.

4.2.3 Conclusions

Reducing the steepness of the gradient and doubling the column length increased considerably the number of identified proteins in unfractionated cocoa beans extracts analysed on the Orbitrap Fusion.

The most reproducible quantitative results were obtained when digesting a total of 160 µg of proteins keeping the ratio protein:trypsin at 50:1, and carrying out the desalting step on Sola Thermo reverse phase SPE cartridges.

The highest number of identified proteins on the Qexactive Orbitrap were achieved by increasing the injection time for the MS/MS analysis to 300 ms and the injection volume for the tryptic digest to 15 µl. The comparison of the proteomic profile of cocoa beans from different genotypes was carried out on the Qexactive as this instrument showed a higher reproducibility compared to the Orbitrap Fusion for quantitative proteomic analysis, although the sensitivity of the Qexactive in terms of the number of identified proteins was lower when compared to the Fusion.

4.3 Quantitative proteomics analysis of cocoa genotypes by LC-MS (results section 3.3)

4.3.1 Effect of shipment, different trees and harvest time (results section 3.3.1-3.3.2)

Cocoa beans for all the analyses carried out in this part of the project were provided by the University of West Indies in Trinidad and air-freighted to Reading. It was not known whether the shipment would cause changes to the abundance of proteins in cocoa beans. Therefore, an experiment to evaluate the effect of shipment on the proteomic profile of cocoa beans was performed. As described in section 3.3.1 a portion of ground and freeze-dried cocoa beans from an Amelonado variety was shipped back and forth from the University of Reading to Trinidad, while the remainder of the sample was stored at the University of Reading. When this sample arrived in Trinidad, it was held in customs for several days before being sent to the University of West Indies. It took around 10 days for the sample to be shipped back and forth from Reading to Trinidad. Although the abundance of a S-adenosyl-L-methionine-dependent methyltransferases superfamily protein was 1.6-fold higher in the shipped sample, no other protein was detected with a fold difference higher than 1.5-fold between the shipped and not-shipped sample. These results indicate that the proteins' distribution and levels in the cocoa bean samples were not affected by the shipment to and from Trinidad, even though the temperature of the shipped sample was not controlled during the shipment. The high stability of the cocoa proteins was probably due to the fact that the samples had been freeze-dried prior to shipping, which minimised protein degradation.

As mentioned in the material and methods section 2.2.2, biological replicates from the four selected cocoa varieties were grown on different trees and harvested at different times. As there was no prior knowledge about differences in the proteomic profiles of biological replicates from the same genotype, experiments to evaluate the effect of different trees and harvest time on the proteomic profile of biological replicates from the cocoa genotype IMC 67 grown in the ICGT field were carried out. It was decided to select this genotype for the assessment of the effects of harvest time and different trees, as it provided a selection of biological replicates grown on the same tree but with different harvest time, and biological replicates harvested on the same day and grown on different trees. The results of these experiments are reported by descriptive statistics only.

The vast majority of the proteins detected in the assessment of the effect of different trees did not show considerable variation between the four biological replicates evaluated for this experiment, as only 6 proteins out of a total of 511 were detected with a fold increase/decrease from the mean >2, see Table 9. A beta-amylase, salicylate O-methyltransferase and a S-adenosyl-L-methionine-dependent methyltransferases were detected in the biological replicate T2-2 with a fold increase/decrease >2 from the overall mean, while the biological replicate T6-4 showed a beta-amylase and a glycosyl hydrolase with a fold increase/decrease from the mean >2, see Table 10. A HSP20-like chaperone and an eukaryotic aspartyl protease were detected with a fold increase/decrease from mean >2 in the biological replicates T13-6 and T5-3, respectively.

A total of 502 proteins were quantified in the four biological replicates analysed to evaluate the effect of harvest time, of which 9 were detected with a fold increase/decrease from the mean >2, see Table 11, confirming that also for this experiment most of the detected proteins were consistent between the biological replicates analysed. A serine carboxypeptidase and an uncharacterised protein showed the highest variation with fold increase/decrease of 11 and 5, respectively.

All the other proteins listed in Table 11 had fold increase/decrease values <3. The biological replicate T10-6 showed the highest number of proteins with a fold increase/decrease from the mean >2 with 5 entries, see Table 12. A total of 4 proteins with a with a fold increase/decrease from the mean >2 were detected in the biological replicates T10-3 and T10-5, while 3 proteins with a value >2 were found in the biological replicate T10-1, see Table 12. These results showed that the distribution of proteins with a fold increase/decrease from the mean >2 was similar in the biological replicates analysed to assess the effect of harvest time.

4.3.2 Investigation of proteome changes dependent on the genotype (results section 3.3.1-3.3.2)

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 (ID 1, see Table 13). 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 is involved in the hydrolysis of the amide bonds in either proteins or small molecules [83]. 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 [83].

The next highest fold change was recorded for a glutathione S-transferase (GST) family protein (ID 2) 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 [84].

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 [85].

With respect to the 57 proteins in Table 13, 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 29). 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 abundant 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 30, higher plot). Both ICS 1 and ICS 39 belong to the same genetic group Trinitario, which originates from hybridisations 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 [3]. 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 abundant proteins can be used as a tool to differentiate cocoa genotypes.

Loading the differentially abundant 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 30, lower plot), reflecting the separation of these genotypes observable in the PCA score plot. The genotypes ICS 1 and ICS 39 showed 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 the differentially abundant proteins could 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 [86], and includes two primary amine oxidases (ID 30 and ID 8 in Table 13) and two alcohol dehydrogenases identifications (ID 45 and ID 56 in Table 13), 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 [87], while alcohol dehydrogenases catalyse the oxidation of primary and secondary alcohols to the corresponding aldehydes and ketones [88]. 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 [89]. These reactions are endothermic as they require high temperatures to be activated and are not catalysed by enzymes. Aldehydes and ketones can also be produced from oxidation of amines and alcohols during fermentation catalysed by amine oxidases and alcohol dehydrogenases, as both these enzymes have been linked to the production of volatile compounds responsible for the aroma of other plants [90, 91]. However, it is not known whether these enzymes are activated during fermentation, 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.

The flavour profile of the genotype SCA 6 includes a floral flavour note which is not present in the other genotypes selected for this project, see Table 3. This flavour attribute has been associated to both aldehydes and ketones [89], therefore its presence in the genotype SCA 6 could be linked to a higher amount of primary amine oxidase and alcohol dehydrogenase found in this genotype, as these enzymes could release aldehydes and ketones which could induce floral flavour notes.

A total of 9 proteins involved in stress response were differentially abundant. Four of these proteins (ID 10, 15, and 17 in Table 13) were heat shock proteins by name which are linked to the response of the plant to stress conditions [81]. 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. It has been reported that the amount of proteins involved in defence and stress increases as the cocoa bean ripens [54].

A eukaryotic aspartyl protease (ID 12 in Table 13) 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 [7]. 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 13) was detected at a significantly 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 [24]. These compounds are important flavour precursors which react with sugars during roasting to form volatile 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. However, it is not known what flavour precursor can be linked to specific flavour notes for the cocoa genotypes selected for this project.

Sensory analysis of cocoa beans to evaluate the aroma and flavour are performed by a series of trained panellists who can assess flavour profile based on their experience in recognizing characteristic notes and attributes. Even though the panel members are properly trained, and the results of their assessment are subjected to statistical test to validate the data, human bias cannot always be eliminated in these tests. Another tool often employed to provide a more robust assessment of the flavour profile is the PAC (potent aroma compounds) analysis. In this case the cocoa beans are analysed by GC-MS in order to quantify volatile compounds which can be linked to flavour. Since this analysis is based on the quantification of aroma compounds against reference standards, the human bias is removed, and the results are more accurate. However, specific flavour characteristic cannot always be linked to the amount of aroma compound quantified with this technique, therefore, sensory analysis is also required when assessing flavour profiles of cocoa beans.

A β-amylase was detected at a significantly higher level in the genotype SCA 6 compared to ICS 1 and IMC 67 (ID 3 in Table 13). β-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 [92]. This disaccharide can react with nitrogen containing compounds such as amino acids and peptides during roasting through the Maillard reaction which results in the generation of volatile compounds [93]. Therefore, the release of maltose can be affected by the levels of β-amylase present in cocoa beans, which in turn could influence the flavour profile of roasted cocoa beans. However, the abundance of this specific β-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 albumins were present at a significant higher level in the genotype ICS 39 compared to IMC 67 (ID 18 and 20 in Table 13). 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. Due to this high homology, these two proteins would generate the same tryptic peptides, and therefore can be regarded as one single entry. 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 showed no abundance difference between the cocoa genotypes analysed (see Appendix 2). LC-MS/MS identifications of free peptides released from this protein during fermentation have been reported by several authors [62, 63, 66]. 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.

In another study, two proteins identified as a degraded albumin subunit at 17 kDa and an internal vicilin subunit at 15 kDa were detected at significant different levels in cocoa beans of different origins and varieties by 2D gel electrophoresis and subsequent MALDI-TOF-MS analysis of the digested spots [59]. Even though the proteomic analysis performed for this PhD project should provide a deeper dataset with a higher amount of identified proteins, these two entries are subunits of larger proteins, and therefore cannot detected with a bottom-up approach employed for this analysis, as the peptides generated from the subunits would be considered as part of the larger proteins. Moreover, the cocoa varieties analysed by the authors of this work differed from those evaluated for this PhD project.

A 2S albumin was significantly more abundant in the genotypes ICS 1 and IMC 67 compared to SCA 6 (see ID 14 in Table 13). 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 [94]. Degradation of this protein during fermentation has not been reported in the literature.

4.3.3 Conclusions

Shipment did not have a significant impact on the proteomic profile of cocoa beans, even though there was no control over the temperature of the sample shipped back and forth from Reading to Trinidad. The high stability of the cocoa proteins was probably due to the fact that the samples had been freeze-dried prior to shipping, which minimised protein degradation.

The experiment to assess the effect of different trees on the proteomic profile of cocoa beans showed that only 6 proteins were detected with a fold increase/decrease from the mean >2 in at least one biological replicate. These results indicate that no significant differences were found in the proteomic profile of biological replicates harvested on the same date but grown on different trees.

Similar results were obtained for the experiment to evaluate the impact of harvest time on the proteomic profile of cocoa beans from different biological replicates. In this case, 9 proteins were present with a fold increase/decrease of >2 from the mean in at least one biological replicate, confirming that the harvest time did not have a significant impact on the proteomic profile of the biological replicates analysed either.

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 also possible to graphically visualise proteins positively correlated with each genotype. 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 with differential abundance 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 primary amino oxidases and alcohol dehydrogenases, enzymes that could potentially lead to flavour-inducing compounds, have been detected. These enzymes could be linked to the characteristic floral flavour note reported for the genotype SCA 6 only. Thus, further experiments should be performed to assess whether the different amounts of these enzymes, present during fermentation, affect the final flavour profiles obtained.

4.4 Proteomic analysis by SDS-PAGE (results section 3.4)

4.4.1 Method development (results section 3.4.1)

The electrophoretic bands with a migration point close to the 20-kDa marker can be attributed to cocoa albumins, see Figure 32, as a polypeptide with a similar molecular weight has been previously reported to be a main component of this protein class [12, 19]. Polypeptides detected in gel bands at 47 and 31 kDa were identified with vicilin fractions of cocoa seeds [12, 19], suggesting that the electrophoretic bands with apparent molecular weights of approximately 50 and 30 kDa detected here can be assigned to vicilin polypeptides, see Figure 32. The unresolved electrophoretic bands migrating between 10 and 15 kDa can also be assigned to vicilin, as polypeptides with similar molecular weights have been previously found in vicilin fractions [12, 19].

4.4.2 Effect of harvest time and different trees (results section 3.4.2)

A similar electrophoretic pattern was observed between the samples analysed to assess the effect of different trees on the proteome of the biological replicates for the variety IMC 67 grown in the ICGT field, see Figure 33, even though the bands for T13-6 and T5-3 were slightly wider than the other two biological replicates analysed for this experiment. There was minimal variation in the intensities of the bands at an apparent molecular weight of approximately 50 kDa between the biological replicates analysed for this experiment. A higher variability was observed for the intensities of the 30-kDa bands, as the fold change between the biological replicates with the highest (T2-2) and lowest (T13-6) intensities for this band was 1.6, see Table 16. The same fold change was observed for the 20-kDa bands, however in this case the biological replicates which showed the highest and lowest intensities were T13-6 and T6-4, respectively, see Table 16.

A similar electrophoretic pattern was observed among the four biological replicates analysed to assess the effect of harvest time, see Figure 34. The electrophoretic bands at 50 and 30 kDa showed similar intensities between the biological replicates, see Table 17. The intensity of the band at 20 kDa was lower in the biological replicate T10-1 compared to the other samples analysed, however the fold change for this band was only 1.6 when compared to the highest values observed for the biological replicate T10-3, see Table 17. These results indicate that the effect of different trees and harvest time did not have a significant impact on the proteomic profile of the cocoa beans analysed by SDS-PAGE.

4.4.3 Investigation of proteome changes dependent on the genotype by SDS-PAGE (results section 3.4.3)

No visual differences were observed between the electrophoretic pattern of the cocoa varieties analysed to assess the proteome changes dependent on the genotype by SDS-PAGE, see Figure 35. The electrophoretic bands at apparent molecular weights of 50 and 30 kDa could be attributed to vicilin polypeptides, while the band at 20 kDa could be assigned to albumin. These intensities of these bands were lower in the cocoa genotype IMC 67 grown at both the ICGT and Campus fields compared to the other genotypes, see Table 18. The cocoa genotypes SCA 6 and ICS 39 showed the highest intensities for both the vicilin bands at 50 and 30 kDa, and the albumin band at 20 kDa, see Table 18. The intensities of the vicilin and albumin bands in the genotype IMC 67 grown in two different fields were very similar, see Table 18. The albumin band at 20 kDa showed the highest fold change (1.6) between the different cocoa genotypes, followed by the globulin bands at 30 and 50 kDa, with values of 1.5 and 1.3, respectively, see Table 18. Based on these results there were no significant changes in the proteomic profile of the cocoa varieties analysed by SDS-PAGE. Besides, growing the same cocoa variety in two fields did not have an impact on the proteomic profile either. No significant differences have ever been reported in the proteomic profiles of cocoa beans of different varieties and origin analysed by SDS-PAGE [47, 48]. With this technique only the most abundant proteins such as albumin and vicilin can be detected. These class of proteins did not show significant differences in the samples analysed by LC-MS/MS, confirming the results obtained by SDS-PAGE.

4.4.4 Conclusions

Three main electrophoretic bands at approximately 50, 30 and 20 kDa were observed. The band at 20 kDa was attributed to albumin, while the remaining bands at 50 and 30 kDa were assigned to vicilin polypeptides.

The analysis of biological replicates harvested at different times and grown on different trees showed that these two variables did not have a significant impact on the electrophoretic pattern of the biological replicates evaluated for these experiments.

The bands' distributions and intensities of the protein extracts of cocoa beans from different genotypes were similar, indicating that there were no detectable differences in the proteomic profile of these sample analysed by SDS-PAGE. Similarly, growing the same cocoa genotype in two different fields did not cause changes in the electrophoretic pattern.

4.5 Method development for the analysis of free peptides by LC-MS (results section 3.5)

Searches of the MS/MS data from the initial analysis of free peptides extracted with a methanol:water 70:30 (v/v) solution against the Cacao Matina 1-6 genome database [74] returned a low number of identified peptides. However, when the same MS/MS data were searched against a database including the 100 most abundant proteins identified in non-fermented cocoa beans, see Appendix 1, a significantly higher number of peptides were detected, see Table 19. When carrying out searches on protein databases, a peptide score threshold for peptide identifications is defined as the probability that an identified peptide is not a random event. This threshold increases proportionally with the size of a database, therefore weaker matches with lower scores are lost and not detected when using larger databases. Since only the most abundant proteins from cocoa beans are likely to be fermented, it was reasonable to reduce the size of the database to the proteins identified in cocoa beans, in particular the most abundant ones. Using aqueous TFA as an extraction solution and concentrating the peptides extracts by SPE resulted in an increased number of identified peptides, see Table 20, as this step allows loading a higher amount of peptides on the column and at the same time removes salts and strongly polar compounds from the peptide extracts. To cover a wider range of proteins, the MS/MS data of the peptide solutions extracted with aqueous TFA and concentrated by SPE were also searched against a database containing the 897 proteins detected in unfermented cocoa beans protein extracts searched against Uniprot\Tremble database restricted to *Theobroma cacao* entries only, see Table 21. A similar number of peptides originating from vicilin and albumin were detected when using the databases with 100 and 897 proteins, see Tables 20-21**,** however, with the larger database additional peptides originating from a higher number of proteins were identified. Adding methanol to the peptide extraction solution and increasing the injection time for the MS/MS acquisition from 200 to 300 ms achieved the highest number of identified peptides, see Table 22. The addition of an organic modifier may have favoured the extraction of hydrophobic peptides which are not highly soluble in a 100% aqueous buffer. By increasing the injection time for the MS/MS acquisition, a higher number of ions are stored in the C-Trap before being transferred into the mass analyser, improving the signal to noise and enabling the acquisition of higher quality spectra particularly for low abundant peptides.

With these improved conditions a total of 155 peptides were identified and quantified in all three technical replicates of fermented beans of Ghanaian origin. Over 800 peptides were identified by D'Souza *et al.* [66] by LC-MS/MS analysis of cocoa beans at different stages of fermentation. However, in this case, the list of identified peptides resulted from combining entries in samples from 7 fermentation stages and included peptides with two and three residues as well. The results reported in this PhD project were obtained from the analysis of cocoa beans at the final fermentation stage only and were limited to peptides with a minimum of 4 residues, as the software cannot identify peptides shorter than 4 residues. Therefore, all peptides generated at the initial fermentation stages and the di and tri-peptides could not be identified, which could explain why the number of peptides reported in this project is considerably lower compared to the results published by D'Souza *et al.* [66].

The vast majority of the detected peptides originated from vicilin and a 21 kDa albumin with 221 amino acids, although a few peptides released from other proteins such as aspartyl proteases, peroxygenases and lipoxygenases were also detected, see Table 23. As vicilin and albumin are the most abundant proteins in cocoa seeds, see Appendix 1, it is expected that cocoa endogenous proteases degrade more of these proteins compared to other proteins present in cocoa beans. A considerably higher number of peptides originating from vicilin compared to peptides from the 21 kDa albumin were detected, see Table 23, as the sequence of vicilin is 566 residues, while the 21 kDa albumin is only 221 residues. It has been reported that vicilin is the protein which release the highest number of free peptides in fermented beans [66], followed by albumin. Therefore the results of this work are in agreement with previously published papers [66].

It has been reported in the literature that the 21 kDa cocoa albumin undergoes little or no degradation during fermentation of cocoa beans [12, 19, 20], due to its inhibitor properties towards endoproteases. However, LC-MS/MS analysis of free peptides from fermented beans have shown that proteolysis product generated from this protein are released during fermentation [61-63, 66]. Besides, it has been demonstrated that although the 21 kDa albumin can inactivate trypsin and chymotrypsin, other proteases such as serine proteases, aspartic proteases, pepsin and cocoa endoproteases are not inhibited by this albumin [56]. The free peptides originating from the 21 kDa albumin covered 69% of the sequence of this protein and were localised in specific zones spread throughout the sequence except around the initial 26 residues at the N-terminal, see Figure 36. These results were in agreement with findings from previous studies [26, 62], suggesting that the N-terminal region of the 21 kDa albumin is not degraded during fermentation. No peptides were detected in regions of the sequence localised at the amino acid residues 68-73, 106-116, 151-153, 169-180.

Peptides from the C-terminal region at 213-221 were not present either, although a cleavage site at amino acid residue 212 was identified. It is possible that the peptides at the C-terminal with sequence IKQVVNAKH may have been further degraded to smaller peptides which cannot be detected with the methodology used for this experiment. Free peptide rich zones were localised in the region of the sequence at amino acid residues 27-67, 74-105, 117-150 154-168 and 180-212. A total of 7 peptides were detected in the region localised at position 27-67, see Appendix 3. A peptide with sequence ANSPVLDTDGDELQTGVQYYVL at position 27-48 was further degraded into three smaller peptides sharing the same N-terminus, suggesting carboxypeptidase enzyme activity, see Appendix 3. Another peptide with sequence SSISGAGGGGLALGRATGQ located at position 49-67 was cleaved at the C-terminus into a smaller peptide, providing additional evidence of the action of a carboxypeptidase, see Appendix 3.

The region of the 21 kDa albumin covering the amino acid residues at positions 74-105 yielded the highest number of peptides for this protein, as a total of 23 peptides with amino acid sequences from these positions were detected, see Appendix 3. Carboxylase activity was also found in peptides generated from this region, for example the peptide at positions 89-103 and amino acid sequence SNADSKDDVVRVSTD which was further cleaved into four smaller peptides sharing the same N-terminus, see Appendix 3. Peptides sharing the same C-terminus at position 105 were also detected in the region at positions 89-105, indicating that the action of an aminopeptidase may also be present, see Appendix 3.

A peptide located at positions 117-134 with the amino acid sequence STSTVWRLDNYDNSAGKW was further cleaved at the N-terminus into 9 smaller peptides sharing the same C-terminus, indicating that these cleavages were also the results of the action of an aminopeptidase, see Appendix 3. It was interesting to note that no carboxypeptidase activity was observed in this region of the 21 kDa albumin. However, carboxypeptidase activity was observed in the region of the 21 kDa albumin covering the residues 181-212, which included a total of 10 detected peptides, see Appendix 3. The cleavage sites resulting from the actions of aspartic proteases, aminopeptidases and carboxypeptidases for the peptides originating from the 21 kDa albumin were mapped on the sequence of this protein, and a total of 55 sites were localised, see Figure 36. The majority of the cleavage sites resulting from the action of cocoa aspartic proteases were in agreement with the work carried out by Janek *et al.* [26] who evaluated the cleavage specificity of the cocoa aspartyl proteases by *in vitro* degradation of cocoa storage proteins.

Although a higher number of peptides originating from the vicilin were identified compared to the peptides originating from the 21 kDa albumin, see Table 23, the combined sequences of all peptides originating from vicilin provided a coverage of only 39% for this protein, which is considerably lower than the sequence coverage observed for the 21 kDa albumin, see Figure 37**.**

No peptide sequences localised in the N-terminal region between 1-130 of the cocoa vicilin were identified, indicating that this N-terminus is not degraded during fermentation, as reported in previous studies [20, 62, 64]. Since the oligopeptide at positions 1-130 of the cocoa vicilin has never been identified, it has been suggested that the annotation of vicilin at the N-terminus may not be correct [20]. Almost half of the identified peptides originating from vicilin were localised in the region between amino acid residues 131-153, see Appendix 3. A peptide with sequence RSEEEEGQQRNNPYYFPKRRSFQ located at positions 131-153 was cleaved at the C-terminus into 6 peptides with lower molecular weights, suggesting activity of a carboxypeptidase, see Appendix 3. This peptide was most likely cleaved by the action of aspartic protease into smaller peptides which underwent further degradation at both the C-terminus and N-terminus, indicating activity of both carboxypeptidases and aminopeptidases, see Appendix 3.

Only single vicilin peptides were identified in the regions of the sequence at positions 170-183, 218-243, 273-286, 296-306, 314-322, 354-366, see Appendix 3. A peptide at positions 158-167 with the sequence DEEGNFKILQ identified in this study, was also found in beans fermented in a lab-scale fermentation and lost after roasting [67]. This peptide released the C-terminal amino acid to form the peptide DEEGNFKIL. A peptide at positions 188-198 with the sequence EANPNTFILPH was cleaved at the N-terminal amino acid residue, see Appendix 3. A total of three peptides were identified in the regions of the vicilin sequence between positions 390-405. In this case, only activity of aspartyl protease was observed. A peptide located at positions 491-499 with sequence NAQNNQRIF was cleaved at the first Q residue from the N-terminus by the action of the cocoa aspartyl protease to form a smaller peptide with sequence NNQRIF, see Appendix 3. The peptide with sequence NAQNNQRIF was also cleaved at the N-terminal residue indicating activity of an aminopeptidase. The region of the amino acid positions 432-465 was particularly rich in peptides, as a total of 15 peptides resulting from the putative activities of carboxypeptidases, aminopeptidases and aspartyl proteases were identified within this region, see Appendix 3. A peptide APLSPGDVF localised at the amino acid sequence 457-465 which had been previously reported as a potential flavour precursor [60] was also detected. A total of 11 peptides were identified in the cocoa vicilin region within the amino acid positions 519-538. Peptides cleaved at both the C-terminus and N-terminus were detected in this region, indicating carboxypeptidase and aminopeptidase activities. A total of 69 cleavage sites potentially resulting from the actions of aspartic proteases, aminopeptidases and carboxypeptidases were localised on the sequence of cocoa vicilin, see Figure 37. Most of the cleavage sites of the aspartic proteases matched the results published by Janek *et al.* [26].

4.5.1 Conclusions

A method to identify and quantify free peptides from fermented cocoa beans by LC-MS/MS analysis has been developed. The extraction conditions which provided the highest number of identified peptides were achieved using a water:methanol 80:20 (v/v) solution acidified with TFA as extraction solvent, and subsequently desalting the peptides extracts on Thermo Sola reverse phase SPE cartridges. Setting the injection time for MS/MS analysis to 300 ms resulted in an increased number of identified peptides.

The highest number of identified peptides was obtained when the data files were searched against a custom database contained 897 proteins previously identified using a Uniprot\Tremble database specific to *Theobroma cacao* only. The majority of these peptides were related to vicilin and a 21 kDa albumin, which are the most abundant proteins in cocoa beans. These results confirmed that the 21 kDa albumin is extensively degraded during fermentation although this protein has endopeptidase inhibitory activity.

An evaluation of the sequences of the identified free peptides showed activity of carboxypeptidases, aspartyl proteases and aminopeptidases on both vicilin and 21 kDa albumin. The combined sequences of the peptides cleaved from the 21 kDa albumin covered 69% of the sequence of this protein, while a protein sequence coverage of only 39% was obtained from the combined peptides sequences cleaved from vicilin. The cleavage sites for the activity of the aspartic proteases in both these proteins were mainly in agreement with results previously reported in the literature [26].

This methodology can be applied to assess the free peptides profiles of fermented cocoa beans from varieties showing different flavour characteristic in order to understand whether similar peptides are released in these cocoa beans during fermentation.

5 Future work

5.1 Proteomic and peptidomic analyses of cocoa beans at different stages of fermentation

Cocoa beans from selected cocoa genotypes could be fermented under controlled conditions and sampled at different stages of fermentation, to get a better understanding of the impact of this process on the proteomic profile of cocoa beans. Protein and free peptide analyses by LC-MS/MS could be carried out on these samples to evaluate what proteins are degraded during fermentation and what proteolysis product are generated during this process. These experiments would also be useful to understand whether proteins breakdown and release of peptides during fermentation is similar in cocoa beans from genotypes showing different flavour characteristics.

5.2 Roasting and sensory evaluation of fermented beans

Fermented beans from cocoa genotypes evaluated in this PhD project could be roasted and cocoa liquor produced. Sensory analysis could be performed on these liquors to assess the cocoa flavour profiles of these cocoa genotypes and evaluate whether the changes in the levels of proteases and enzymes producing flavour-inducing compounds can be linked to the flavour notes of the roasted beans. A potent aroma compound analysis on the roasted beans could also be carried out to assess whether volatile compounds responsible for the cocoa flavour are detected at different levels among the selected genotypes.

Analysis of free peptides of fermented cocoa beans prior and after roasting, would allow to get a better understanding of what peptides react with sugars through Maillard reaction during roasting leading to the formation of volatiles compounds responsible for the cocoa flavour.

5.3 Proteogenomics on the MS/MS data

The characterisation of the cocoa bean proteome carried out in this project has allowed identification of over 1000 proteins, which provides a sufficient proteome coverage to perform proteomics aiding genome annotation using these data [95].

In this case, the MS/MS data generated for the characterisation of the cocoa beans proteome could be searched against genome sequence databases specific for *Theobroma cacao*. The results of this search could then be used to map the identified peptides to the existing gene annotation model, providing valuable evidence of protein translation, and thus improving the accuracy of the algorithm employed to predict gene prediction for the *Theobroma cacao* genome.

6 Reference

1. Afoakwa, E., O., *Chocolate Science and Technology*, in *Chocolate Science and Technology*, Wiley-Blackwell, Editor. 2010.

2. Lima, L.J.R., et al., *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, 2011. **51**(8): p. 731-761.

3. Motamayor, J.C., et al., *Geographic and Genetic Population Differentiation of the Amazonian Chocolate Tree (Theobroma cacao L).* Plos One, 2008. **3**(10).

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

5. Schwan, R.F. and A.E. Wheals, *The microbiology of cocoa fermentation and its role in chocolate quality.* Critical Reviews in Food Science and Nutrition, 2004. **44**(4): p. 205-221.

6. Camu, N., et al., *Fermentation of cocoa beans: influence of microbial activities and polyphenol concentrations on the flavour of chocolate.* Journal of the Science of Food and Agriculture, 2008. **88**(13): p. 2288-2297.

7. Voigt, J., et al., *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, 1994. **49**(2): p. 173-180.

8. Afoakwa, E.O., et al., *Flavor formation and character in cocoa and chocolate: A critical review.* Critical Reviews in Food Science and Nutrition, 2008. **48**(9): p. 840-857.

9. Pettipher, G.L., *ANALYSIS OF COCOA PULP AND THE FORMULATION OF A STANDARDIZED ARTIFICIAL COCOA PULP MEDIUM.* Journal of the Science of Food and Agriculture, 1986. **37**(3): p. 297-309.

10. Schwan, R.F., *Cocoa fermentations conducted with a defined microbial cocktail inoculum.* Applied and Environmental Microbiology, 1998. **64**(4): p. 1477-1483.

11. Gu, F., et al., *Comparison of Cocoa Beans from China, Indonesia and Papua New Guinea.* Foods, 2013. **2**(2): p. 183-197.

12. Voigt, J., B. Biehl, and S.K.S. Wazir, *The Major Seed Proteins of Theobroma cacao L.* Food Chemistry, 1993. **47**(2): p. 145-151.

13. Rodriguez-Campos, J., et al., *Dynamics of volatile and non-volatile compounds in cocoa (Theobroma cacao L.) during fermentation and drying processes using principal components analysis.* Food Research International, 2011. **44**(1): p. 250-258.

14. Hansen, C.E., M. del Olmo, and C. Burri, *Enzyme activities in cocoa beans during fermentation.* Journal of the Science of Food and Agriculture, 1998. **77**(2): p. 273-281.

15. de Brito, E.S., et al., *Structural and chemical changes in cocoa (Theobroma cacao L) during fermentation, drying and roasting.* Journal of the Science of Food and Agriculture, 2001. **81**(2): p. 281-288.

16. Biehl, B. and D. Passern, *PROTEOLYSIS DURING FERMENTATION-LIKE INCUBATION OF COCOA SEEDS.* Journal of the Science of Food and Agriculture, 1982. **33**(12): p. 1280-1290.

17. Hashim, P., et al., *Changes in free amino acid, peptide-N, sugar and pyrazine concentration during cocoa fermentation.* Journal of the Science of Food and Agriculture, 1998. **78**(4): p. 535-542.

18. Kirchhoff, P.M., B. Biehl, and G. Crone, *PECULIARITY OF THE ACCUMULATION OF FREE AMINO-ACIDS DURING COCOA FERMENTATION.* Food Chemistry, 1989. **31**(4): p. 295-311.

19. Lerceteau, E., et al., *Evolution of cacao bean proteins during fermentation: a study by two-dimensional electrophoresis.* Journal of the Science of Food and Agriculture, 1999. **79**(4): p. 619-625.

20. Kumari, N., et al., *Biochemical fate of vicilin storage protein during fermentation and drying of cocoa beans.* Food Research International, 2016. **90**: p. 53-65.

21. Amin, I., S. Jinap, and B. Jamilah, *Vicilin-class globulins and their degradation during cocoa fermentation.* Food Chemistry, 1997. **59**(1): p. 1-5.

22. Jinap, S., et al., *Aroma precursors and methylpyrazines in underfermented cocoa beans induced by endogenous carboxypeptidase.* Journal of Food Science, 2008. **73**(7): p. H141-H147.

23. Voigt, J., et al., *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, 1994. **51**(2): p. 197-205.

24. Bytof, G., et al., *SPECIFICITY AND STABILITY OF THE CARBOXYPEPTIDASE ACTIVITY IN RIPE, UNGERMINATED SEEDS OF THEOBROMA-CACAO L.* Food Chemistry, 1995. **54**(1): p. 15-21.

25. Voigt, J., et al., *IN-VITRO STUDIES ON THE PROTEOLYTIC FORMATION OF THE CHARACTERISTIC AROMA PRECURSORS OF FERMENTED COCOA SEEDS - THE SIGNIFICANCE OF ENDOPROTEASE SPECIFICITY.* Food Chemistry, 1994. **51**(1): p. 7-14.

26. Janek, K., et al., *The cleavage specificity of the aspartic protease of cocoa beans involved in the generation of the cocoa-specific aroma precursors.* Food Chemistry, 2016. **211**: p. 320-328.

27. Guilloteau, M., et al., *Identification and characterisation of the major aspartic proteinase activity in Theobroma cacao seeds.* Journal of the Science of Food and Agriculture, 2005. **85**(4): p. 549-562.

28. Laloi, M., et al., *Molecular and biochemical characterisation of two aspartic proteinases TcAP1 and TcAP2 from Theobroma cacao seeds.* Planta, 2002. **215**(5): p. 754-762.

29. Amin, I., et al., *Oligopeptide patterns produced from Theobroma cacao L of various genetic origins.* Journal of the Science of Food and Agriculture, 2002. **82**(7): p. 733-737.

30. Lovric, J., *Introducing proteomics: from concepts to sample preparation, mass spectrometry and data analysis*. 2011, Wiley-Blackwell.

31. BIO-RAD. *Guide to Polyacrylamide Gel Electrophoresis and Detection*. Accessed on 31st October 2015; Available from: http://www.bio-rad.com/webroot/web/pdf/lsr/literature/Bulletin_6040.pdf.

32. Skoog, D.A. and L. J.J., *Principal of Instrumental Analysis*. 1992: Saunders College Publishing.

33. Gross, J.H., *Mass Spectrometry a Textbook*. 2004, Springer.

34. CHROMACADEMY. *Electrospray ionisation tutorial*. Accessed on 20th October 2015; Available from: http://www.chromacademy.com/Electrospray-Ionization-ESI-for-LC-MS.html?tpm=1_1.

35. M., D., et al., *Molecular beams of macroions.* Journal of Chemical Physics, 1968. **49**: p. 2240-2249.

36. Iribarne, J.V. and B.A. Thomson, *On the evaporation of small ions from charged droplets.* Journal of Chemical Physics, 1976. **64**(6): p. 2287-2294.

37. Singh, C., et al., *Higher Energy Collision Dissociation (HCD) Product Ion-Triggered Electron Transfer Dissociation (ETD) Mass Spectrometry for the Analysis of N-Linked Glycoproteins.* Journal of Proteome Research, 2012. **11**(9): p. 4517-4525.

38. Cottrell, J.S., *Protein identification using MS/MS data.* Journal of Proteomics, 2011. **74**(10): p. 1842-1851.

39. Hu, Q.Z., et al., *The Orbitrap: a new mass spectrometer.* Journal of Mass Spectrometry, 2005. **40**(4): p. 430-443.

40. *Exactive Plus Operating Manual*. 2012, Revision A - 1323060, Thermo Fisher Scientific.

41. CHROMACADEMY. *Fundamental LC-MS Mass Analysers*. Accessed on 12th October 2019; Available from: https://www.chromacademy.com/lms/sco36/Fundamental_LC-MS_Mass_Analysers.pdf.

42. Chen, C.H., *Review of a current role of mass spectrometry for proteome research.* Analytica Chimica Acta, 2008. **624**(1): p. 16-36.

43. Zhu, W.H., J.W. Smith, and C.M. Huang, *Mass Spectrometry-Based Label-Free Quantitative Proteomics.* Journal of Biomedicine and Biotechnology, 2010: p. 6.

44. Ishihama, Y., et al., *Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.* Molecular & Cellular Proteomics, 2005. **4**(9): p. 1265-1272.

45. Spencer, M.E. and R. Hodge, *CLONING AND SEQUENCING OF A CDNA-ENCODING THE MAJOR STORAGE PROTEINS OF THEOBROMA-CACAO - IDENTIFICATION OF THE PROTEINS AS MEMBERS OF THE VICILIN CLASS OF STORAGE PROTEINS.* Planta, 1992. **186**(4): p. 567-576.

46. Reisdorff, C., et al., *Comparative study on the proteolytic activities and storage globulins in seeds of Theobroma grandiflorum (Willd ex Spreng) Schum and Theobroma bicolor Humb Bonpl, in relation to their potential to generate chocolate-like aroma.* Journal of the Science of Food and Agriculture, 2004. **84**(7): p. 693-700.

47. Amin, I., et al., *Analysis of vicilin (7S)-class globulin in cocoa cotyledons from various genetic origins.* Journal of the Science of Food and Agriculture, 2002. **82**(7): p. 728-732.

48. Hue, C., et al., *Impact of fermentation on nitrogenous compounds of cocoa beans (Theobroma cacao L.) from various origins.* Food Chemistry, 2016. **192**: p. 958-964.

49. Spencer, M.E. and R. Hodge, *Cloning and sequencing of the cDNA-encoding the major albumin of Theobroma cacao-Identification of the protein as a member of the Kunitz protease inhibitor family.* Planta, 1991. **183**(4): p. 528-535.

50. Hansen, C.E., et al., *Comparison of enzyme activities involved in flavour precursor formation in unfermented beans of different cocoa genotypes.* Journal of the Science of Food and Agriculture, 2000. **80**(8): p. 1193-1198.

51. Awang, A., R. Karim, and T. Mitsui, *Proteomic analysis of Theobroma cacao pod husk.* Journal of Applied Glycoscience, 2010. **57**(4): p. 245-264.

52. Noah, A.M., et al., *Comparative proteomic analysis of early somatic and zygotic embryogenesis in Theobroma cacao L.* Journal of Proteomics, 2013. **78**: p. 123-133.

53. Quinga, L.A.P., et al., *Insights into the conversion potential of Theobroma cacao L. somatic embryos using quantitative proteomic analysis.* Scientia Horticulturae, 2018. **229**: p. 65-76.

54. Wang, L., et al., *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, 2016. **87**(3): p. 318-332.

55. Kratzer, U., et al., *Subunit structure of the vicilin-like globular storage protein of cocoa seeds and the origin of cocoa- and chocolate-specific aroma precursors*, in *Food Chemistry*. 2009. p. 903-913.

56. Kochhar, S., K. Gartenmann, and M.A. Juillerat, *Primary structure of the abundant seed albumin of Theobroma cacao by mass spectrometry.* Journal of Agricultural and Food Chemistry, 2000. **48**(11): p. 5593-5599.

57. Kochhar, S., et al., *Isolation and characterization of 2S cocoa seed albumin storage polypeptide and the corresponding cDNA.* Journal of Agricultural and Food Chemistry, 2001. **49**(9): p. 4470-4477.

58. Bertazzo, A., et al., *The protein profile of Theobroma cacao L. seeds as obtained by matrix-assisted laser desorption/ionization mass spectrometry.* Rapid Communications in Mass Spectrometry, 2011. **25**(14): p. 2035-2042.

59. Kumari, N., et al., *Origin and varietal based proteomic and peptidomic fingerprinting of Theobroma cacao in non-fermented and fermented cocoa beans.* Food research international (Ottawa, Ont.), 2018. **111**: p. 137-147.

60. Buyukpamukcu, E., et al., *Characterization of peptides formed during fermentation of cocoa bean.* Journal of Agricultural and Food Chemistry, 2001. **49**(12): p. 5822-5827.

61. Kochhar, S., et al., *Cocoa flavour precursors peptides.* Patent number 2004/0202761, 2004.

62. Marseglia, A., et al., *Extraction, identification and semi-quantification of oligopeptides in cocoa beans.* Food Research International, 2014. **63**: p. 382-389.

63. Caligiani, A., et al., *Influence of fermentation level and geographical origin on cocoa bean oligopeptide pattern.* Food Chemistry, 2016. **211**: p. 431-439.

64. Voigt, J., et al., *Partial purification and characterisation of the peptide precursors of the cocoa-specific aroma components.* Food Chemistry, 2016. **192**: p. 706-713.

65. Voigt, J., K. Textoris-Taube, and J. Woestemeyer, *pH-Dependency of the proteolytic formation of cocoa- and nutty-specific aroma precursors.* Food Chemistry, 2018. **255**: p. 209-215.

66. D'Souza, R.N., et al., *Degradation of cocoa proteins into oligopeptides during spontaneous fermentation of cocoa beans.* Food Research International, 2018. **109**: p. 516-516.

67. Warren A., J., et al., *Forcing fermentation: Profiling proteins, peptides and polyphenols in lab-scale cocoa bean fermentation.* Food Chemistry, 2019. **278**: p. 786-794.

68. Sukha, D.A. and D.R. Butler, *Trends in flavour profiles of the common clones for the CFC/ICCO/INIAP Flavour Project.* Cocoa Research Unit The University of the West Indies Annual Report, 2005: p. 55-61.

69. Sukha, D.A., et al., *An Assessment of the Quality Attributes of the Imperial College Selections (ICS) Cacao (Theobroma cacao L.) Clones.* III International Conference on Postharvest and Quality Management of Horticultural Products of Interest for Tropical Regions, 2014. **1047**: p. 237-243.

70. De Witt, K.W., *The flavour assessment of cacao, a report on cacao research.* The Imperial College of Tropical Agriculture, *1954*: p. **77-81**.

71. Turnbull, C.J. and P. Hadley. *International Cocoa Germplasm Database (ICGD) [Online Database]*. CRA Ltd./ICE Futures Europe/University of Reading, Accessed on 7th July 2018; Available from: http://www.icgd.reading.ac.uk.

72. Bradford, M.M., *A Rapid and Sensitive Method for the Quantitation of Microgram Quantities of Protein Utilizing the Principle of Protein-Dye Binding.* Analytical Biochemistry, 1976. **72**(1-2): p. 248-254.

73. Argout, X., et al., *The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies.* Bmc Genomics, 2017. **18**: p. 9.

74. Motamayor, J.C., et al., *The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color.* Genome Biology, 2013. **14**(6): p. 1-48.

75. Hoving, H.D., H.R. Kattenberg, and D.A.J. Starmans, *Process for extracting polyphenolic antioxidants from purine-containing plants.* European Patent, 2007.

76. Bryant, L., et al., *Proteomic analysis of Artemisia annua - towards elucidating the biosynthetic pathways of the antimalarial pro-drug artemisinin.* Bmc Plant Biology, 2015. **15**.

77. Parthibane, V., et al., *Oleosin Is Bifunctional Enzyme That Has Both Monoacylglycerol Acyltransferase and Phospholipase Activities.* Journal of Biological Chemistry, 2012. **287**(3): p. 1946-1954.

78. Wang, P. and J. Heitman, *The cyclophilins.* Genome Biology, 2005. **6**(7).

79. Burston, S.G. and A.R. Clarke, *Molecular chaperones: Physical and mechanistic properties.* Essays in Biochemistry, 1995. **29**: p. 125-136.

80. de la Cruz, J., K. Karbstein, and J.L. Woolford, *Functions of Ribosomal Proteins in Assembly of Eukaryotic Ribosomes In Vivo*, in *Annual Review of Biochemistry, Vol 84*, R.D. Kornberg, Editor. 2015, Annual Reviews: Palo Alto. p. 93-129.

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

82. Punja, Z.K. and Y.Y. Zhang, *PLANT CHITINASES AND THEIR ROLES IN RESISTANCE TO FUNGAL DISEASES.* Journal of Nematology, 1993. **25**(4): p. 526-540.

83. Marchler-Bauer, A., et al., *CDD/SPARCLE: functional classification of proteins via subfamily domain architectures.* Nucleic Acids Research, 2017. **45**(D1): p. D200-D203.

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

85. Remacha, M., et al., *Proteins P1, P2, and P0, components of the eukaryotic ribosome stalk. New structural and functional aspects.* Biochemistry and Cell Biology, 1995. **73**(11-12): p. 959-968.

86. Scollo, E., et al., *Characterization of the Proteome of Theobroma cacao Beans by Nano-UHPLC-ESI MS/MS.* Proteomics, 2018. **18**(3-4).

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

88. Svensson, S., et al., *Crystal structures of mouse class II alcohol dehydrogenase reveal determinants of substrate specificity and catalytic efficiency.* Journal of Molecular Biology, 2000. **302**(2): p. 441-453.

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

90. Jin, Y.Z., et al., *The Alcohol Dehydrogenase Gene Family in Melon (Cucumis melo L.): Bioinformatic Analysis and Expression Patterns.* Frontiers in Plant Science, 2016. **7**: p. 18.

91. Tavladoraki, P., A. Cona, and R. Angelini, *Copper-Containing Amine Oxidases and FAD-Dependent Polyamine Oxidases Are Key Players in Plant Tissue Differentiation and Organ Development.* Frontiers in Plant Science, 2016. **7**: p. 11.

92. Rejzek, M., et al., *Chemical genetics and cereal starch metabolism: structural basis of the non-covalent and covalent inhibition of barley beta-amylase.* Molecular Biosystems, 2011. **7**(3): p. 718-730.

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

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

95. Ruggles, K.V., et al., *Methods, Tools and Current Perspectives in Proteogenomics.* Molecular & Cellular Proteomics, 2017. **16**(6): p. 959-981.