Marker development for the traceability of certified sustainably produced cacao (*Theobroma cacao*) in the chocolate industry

Pedro Lafargue-Molina

"A thesis submitted in partial fulfilment of the requirements of the University of the West of England (UWE), Bristol for the degree of **Doctor of Philosophy (PhD)**"

Faculty of Health and Applied Sciences
Centre for Research in Biosciences

Director of studies
Dr Joel Allainguillaume Ph.D.

March 2021

Word Count: 47795

Marker development for the traceability of certified sustainably produced cacao (*Theobroma cacao*) in the chocolate industry

Pedro Lafargue-Molina

"A thesis submitted in partial fulfilment of the requirements of the University of the West of England (UWE), Bristol for the degree of **Doctor of Philosophy (PhD)**"

Faculty of Health and Applied Sciences
Centre for Research in Biosciences

Director of studies
Dr Joel Allainguillaume Ph.D.

March 2021

Word Count: 47795

PUBLICATION STATEMENT

The candidate confirms that the work submitted is his/her own, except where work which has formed part of jointly-authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

ABSTRACT

Theobroma cacao (cocoa) is one of the most studied commodities around the world and the source of one of the world's most consumed and familiar products, chocolate. The multibillion-pound industry has changed to a higher demand for sustainably certified cacao (Rainforest Alliance, UTZ, and Fair Trade) and closer attention is being paid to how this such cacao can be traced. The present work describes a new concept, "From Shelf to Farm & Cooperative", a study to identify the geographical origin of the fermented cacao beans used to manufacture premium and bulk chocolate products. The research sought to assess how DNA based approaches for traceability of food products can be utilised within the supply chain of cacao and chocolate. To identify the factors that influence cacao traceability and the importance of assessing it in different supply chain systems, multi-disciplinary stakeholders from policy makers, small-scale farmers in South and Central America, to the biggest cacao and chocolate manufacturers in Europe were interviewed. Two stages in chocolate production were identified as key to be screened for tracking implementation: The farm (**Stage 1**) to identify cacao trees genotype composition and the cooperative (**Stage 2**) where fermentation of cacao beans occur. A reliable modified cacao DNA extraction protocol was developed using the DNeasy mericon Food Kit which enable higher DNA yield from a range of chocolate products including, for the first time, 'cocoa butter'. DNA markers characterising the chloroplast genome of *T. cacao* were assessed to trace back the chocolate to **Stage 1** (farm). Reference genotypes from the International Cocoa Quarantine Centre at the University of Reading were screened with 25 chloroplast single sequence repeat (cpSSR) markers revealing a level of DNA polymorphism sufficient to reliably identify lineages below the species level to characterise farms. Allelic proportions for nine cpSSR were quantified and compared in DNA extracted from 116 chocolate samples revealing distinct clustering in single-origin chocolate produced from beans harvested in Peru, Ecuador, Venezuela, Trinidad and Madagascar. In contrast, no differentiation was observed for bulk chocolate samples (Mars, Nestle) and beans originating from Ivory Coast farms thus reflecting the lack of allelic diversity found in cultivars in West Africa. To identify unique biomarkers for **Stage 2** (cooperative), the fermentation microbiome was assessed by performing amplicon Illumina sequencing on 47 single origin chocolate using the universal 16S v3-v4 ribosomal region and three housekeeping genes from *Acetobacter pasteurianus*. Variation in microbiome diversity was characterised with unique Amplicon Sequence Variants (ASV) identified per continent, country and fermentation location for which signature bacterial profile was found to be conserved across years. Markers identified in **Stage 1** and **Stage 2** can be used for tracking cocoa beans origin. To make these biomarkers applicable in industrial scenarios, it will be essential to create a machine learning model that could recognize the specific markers from multiple regions.

ACKNOWLEDGEMENT

Thank you! Gracias! Merci! What a journey! Writing up during the first Pandemic of my life is clearly one thing that I'd have never imagined. It has been ten years since I started working in Cacao, which went from planting the trees as a small farmer, creating a chocolate company from scratch to finally finishing this PhD research thesis. Was it my passion, tenacity or something else that drove me? I certainly don't know. But what I know is that without the advice and support of my family, farmers, academics, business partners and researchers, this work would have not been accomplished. Firstly I would like to thank my director, Assoc. Prof Joel Allainguillaume for all the guidance, continued support, enthusiasm and advice throughout these three years. I could not have imagined having a better mentor to build this innovative project, who constantly challenged and guided me towards the right direction to connect the dots. I really appreciated how Joel has shared the same passion for the project and taught me all the scientific skills to support and confidently lead the research to a bigger scope that can be applied into the industry. Besides Joel, I would like to thank the rest of my research supervisory team at UWE: Prof Dawn Arnold, Assoc. Prof Emma Weitkamp and Dr Rey Loor from the National Cacao Research Institution in Ecuador, for their insightful comments and encouragement, but also for the hard questions which helped to shape my research from various perspectives and for future projects.

A huge thanks to Brigitte Laliberté, who coordinates Cocoa of Excellence Programme (CoEx), Bioversity International and the International Cocoa Awards (ICA) for supplying unique single-origin chocolates from across the globe. To Dr Andrew Daymond for all the DNA samples and metadata from the ICQC at the University of Reading, UK. To Dr Rey Loor, who personally drove me more than 400km to selected farms in the Amazon, rainforest and coastal Ecuador for collecting antique cacao genotypes, and who in addition granted me complete access to the INIAP national research laboratories. A special thanks to Tree of Wisdom Chocolates Ltd (TOW) crew in Ecuador for supplying single-origin chocolates. A further big thank you to Dr Vicente Norero from Latiali and Camino Verde for his initial mentoring, sourcing and fermenting TOW cacao beans under unique conditions. To the Shaman Gilberto from the community Cayapas in Esmeraldas for hosting me and supporting my fieldwork in the tropical rainforest. To Karen and Paul from the NGS sequencing facility at the University of Exeter, UK, who gave to me in-depth feedback on the experimental design of Illumina sequencing and for processing the DNA samples.

I would like to express my sincere appreciation to all my trainers in metagenomics and bioinformatics from the Harvard T.H Chan School of Public Health from University in Boston Massachusetts, USA Catchen Laboratory from the University of Illinois, USA to the QIIME developers: Dr Antonio Gonzales, Dr Yoshiki Vázquez from Knight labs from University of California San Diego, USA who originally thought I was a bit insane to try to write a program to study a 5 genes multiplexed experiment. A huge thank you to Prof Cherif Guermat from the UWE Faculty of Business and Law, UK who took the time to understand my vision and help me to build a prediction model based on financial algorithms (who nicely constantly complain about the genetic terminology, which lead me to translate my research into business vocabulary to be in the same page). I would like to thank all my interviewees for their time, but especially to Dr Martin Gilmour, Global head of Agronomy at Barry Callebaut who granted me the access to show my research to the RD board of directors in Belgium.

I must thank Dr Tim Moss at UWE for awarding me my scholarship. I am extremely grateful to the finance and graduate school departments (Caterina, Neil, Vicky, Annie, Joel and Lyn) who made my self-funding agreements possible and approved all my fieldworks. Without them, I wouldn't have been able to present the project to multiple audiences in Cambridge, Berlin, Netherland, Belgium and Indonesia. I would like to thank Dr Robin Thorne, Shonna Nelson, Andy Wetten, Gareth Robinson, David Veal and Man-Kim Cheung for making me part of the lecturing teams, which helped me to learn more about research in applied science and gave me feedback in order to accomplish my research goals. To all my friends at the HAS and FBL PGR centre, I salute you and thank you for making me feel at home.

Last but not the least, I would like to thanks my mother Sandra, my sister Michelle, my little nephew Elias, my big Molinas and Cucalon family (Diegos, Mireya, Lorena, Edgar, Lucho, Jose, Andres), my father Miguel and my grandmas Fanny and Juana for all the support and encouragement to follow my dreams. To my amazing best friends and second family (Polits, Calderon, Garces, Polanco, Moscoso, Urresta), and my Bristolian community (Tanmay, Leo, Isabel, Goncalo, Ewan, Berrbizne, Chris, Buffy, Cyril) for believing in me and helping me to cope with the stress and difficulties throughout these three years.

"THE PURSUIT OF KNOWLEDGE CAN LEAD TO CHAOS, BUT YOU MUST ACCEPT THE FACT THAT THE JOURNEY WILL NOT BE SMOOTH AND THAT YOU CAN SHAPE CHAOS TO BUILD A MAGNIFICENT PIECE OF ART".

PEDRO LAFARGUE

'''

CONTENTS

List of Figures

List of Tables

ABBREVIATIONS

AAB	Acid Acetic Bacteria
B.C	Before Christ
BC	Barry Callebaut
Cacao	Cocoa
CAGR	Compound annual growth rate
CATIE	The Tropical Agricultural Research and Higher Education Centre
CB	Cacao Butter
CCN-51	Coleccion Castro Naranjal
CFC	Common Fund for Commodities
Chocolate Makers	Manufacture who make chocolate from the beans or nibs.
cpSSR	Chloroplast Single Sequence Repeat
CRC	Cocoa Research Centre
CTAB	Cetyl trimethylammonium bromide
DNA	deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
EST-SNP	Expressed Sequence Tags – Single Nucleotide Polymorphisms
EU	European Union
FMCG	Fast-moving consumer goods
ICCO	The International Cocoa Organization
ICGD	International Cocoa Germplasm Database
ICQCR	International Cocoa Quarantine Centre at the University of Reading
INIAP	Instituto Nacional Autónomo de Investigaciones Agropecuarias
ISO	International Organization for Standardization
Ivoseed	cacao beans
LAB	Lactic Acid Bacteria
PCoA	Principal Coordinate Analysis
PCR	Polymerase chain reaction
PCR-DGGE	Polymerase chain reaction denaturing gradient gel electrophoresis
PERMANOVA	Permutational multivariate analysis of variance
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
SC	Supply Chain
SIMPER	Similarity Percentage analysis
SME	Small-Medium Enterprise
TAE	Tris base, acetic acid and EDTA
TE	Tris base and EDTA
TOW	Tree of Wisdom Chocolate Ltd
TSC	Theoretical supply chain (TSC
UBC	Ultra-barcoding
USP	Unique selling proposition

UNITS

°C	Celsius
µF	Micro farad
µg	Micro grams
µL	Micro liter
µM	Micro Molar
bp	Base Pairs
g	grams
ha	Hectare
hr(s)	hours
kg	Kilo grams
min	minutes
mL	millilitre
mM	millimolar
mm	millimetre
ng	nanograms
ppb	parts per billion
RFU	Relative Fluorescent Units

Chapter 1. Introduction

1.1 *Theobroma cacao*

Native to Northern South America and Central America, the cacao tree is the source of one of the world's most delicious and familiar products, chocolate. In 1680, Carl von Linne named the cacao tree as *Theobroma cacao*, from the Greek combination of two words, "Theos", meaning God and "Broma", meaning food (Humphries, 1944). The word cacao has been adapted from the language of the Aztecs, "Cacahuatl" and Mayas, "Kakaw". The term "cacao" was incorporated later in northern countries to describe products derived from processed cacao beans (Afoakwa, 2010).

T. cacao is a perennial tree that grows around the Equatorial line in tropical conditions with temperatures ranging from 20°C to 30°C and with water requirement reported as 3–6 mm d^{-1} during the rainy and < 2mm d^{-1} during the dry season (Carr and Lockwood, 2011). It belongs to the genus *Theobroma* and it is classified in the subfamily Sterculioidea of the family Malvaceae. The genus includes 22 species sub-divided into six sections based on their morphological characteristics (Motamayor *et al.*, 2002). The main *T. cacao* varieties have been further clustered in 10 principal cultivated groups according to flavour characteristics, geographical origin and genetic composition. Cocoa varieties have been previously described and categorized http://www.icgd.reading.ac.uk/icqc/ in the ICQC database.

Due to its economic importance, this tropical-fruit tree has been the subject of scientific research to assess its genetic diversity with its whole genome sequenced in 2010 and several germplasm collection centres established around the world (Argout *et al.*, 2011). The main commercial or bulk varieties identified as Forastero (has been superseded by reclassification based on DNA genotyping); fine & flavour: Criollo, Trinitario (a hybrid between Forastero and Criollo also known as cacao Guinensis) and Nacional (Risterucci *et al.*, 2000). Generally, the difference between fine or flavour and bulk cacao is based on the flavour profile. Fine and flavour defines a group of varieties that have unique flavour notes in contrast with bulk varieties. Ancient Criollo and Nacional (Fine and Flavour or Arriba) beans have been categorized by

sensorial panels as having the highest aromatic flavours and mostly considered as premium varieties (Sukha *et al.*, 2008).

Ancient Criollo and Nacional are still claimed to be the source of various single origin industrial chocolates, though in reality they have almost disappeared due to their high susceptibility to diseases and pests (Afoakwa, 2016a; Kongor *et al.*, 2016; Fowler, 2009). Using molecular genotyping Motamayor *et al.* (2008) characterised the above-mentioned varieties and seven additional varieties including Maranon, Curaray, Iquitos, Nanay, Cotamana, Purus and Guiana.

The varieties of *T. cacao*, which dominate current world cacao production, are known as Forastero, which means in Spanish "stranger" or "foreign". These types of varieties are also known as bulk cacao and have higher yields than fine cacao but also show lower quality, exhibiting basic or ordinary and weak flavour profiles (CFC/ICCO/Bioversity, 2009). For instance, CCN-51 a hybrid developed in Ecuador in the 1960s is a remarkable variety for production standards giving up to 70% higher yield compared to fine and flavour (profiles) premium cacao accessions. However, its organoleptic characteristics are lower and lead to poor flavour development following traditional fermentation. Due to this plain flavour, bulk cacao is used mainly for the production of milk chocolates, cacao powder and butter as ingredients for bakeries and pharmaceutical products. Interestingly, in Ecuador, Forastero also includes Nacional and premium quality varieties (Kongor *et al.*, 2016).

Premium quality is categorized by fine and flavours that are present in rare and native genetic varieties. Premium Ecuadorian cacao including Arriba, Criollo, Fino and Aroma, is commonly known as "fine and flavoured" or "Ancient Criollos" (in Spanish "native"). These varieties exhibit flavours such as fresh fruits, mature fruits, yellow fruits, floral, herbal, wood, wood notes, nut and caramel notes as well as rich and balanced chocolate bases developed during fermentation. These are used mostly in trade and for dark chocolates. This is quite a distinction from bulk cacao such as Forastero that do not have these flavour notes (Seguine and Meinhardt, 2014).

Only 2% of the cacao produced in the world is premium quality mainly from Ecuador, Colombia, Peru, Madagascar and others with the remaining 98% being bulk

cacao from Africa. Bulk cacao production originates mainly from West Africa with 70% of the production in Ivory Coast and Ghana (Davrieux *et al.*, 2007) The production of fine and flavour cacao is largely concentrated in Central and South America with Ecuador recognised as the biggest producer and exporter. While Ecuador produces only 18% of the cacao in the world, it is responsible for 70% of the total premium quality world production, which make quality tracking in the country crucial (Rey *et al.*, 2015).

1.2 Reproduction and breeding

T. cacao is a diploid tree fruit species (2n = 2x = 20) and can be propagated by both sexual and vegetative methods. *T. cacao* is entomophilous and is pollinated by midges from the genus *Forcipomyia* (Boza *et al.*, 2014). The species has hermaphrodite flowers and can be self-pollinated but self-incompatibility is often observed and has been reported in Criollo (Central America), Comun (Brazil) and Nacional (Ecuador) (Maximova *et al.*, 2006; Argout *et al.*, 2011) varieties. Bawa (1986) demonstrated that these upper-Amazon cacao genotypes are self-incompatible due to a gameto-sporophytic system, with a late-acting incompatibility system, resulting in non-fusion of the gametes or non-development of the zygote, after normal pollen tube growth to the ovules. *T. cacao* produces on average 125 000 flowers per tree each year with high variation observed between genotypes only 1 to 5% of all flowers get pollinated. Flowering occurs mainly between January and June in Ivory Coast and Ghana and in Brazil (Paulin, Decazy and Coulibaly, 1983). This is partly due to the typically lower-Amazon Amelonado genotypes grown in these regions, which cease flowering from July to November (Mossu, 1990). In contrast, upper-Amazon genotypes usually produce flowers throughout the year with two high seasons corresponding to the main crop (March to June) and the mid-crop (December to January). Bouharmont (1960) observed that pollen tubes reach the embryo sacs about 4 hours after pollination and double fertilization is completed within 24 hours following pollination. The first division of the zygote does not occur before 55 days following pollination, while embryo growth begins 90 days and is completed 110 days after pollination. Maturity is reached approximately 150 days after anthesis when the green or red pods turn yellow (Falque *et al.*, 1995).

1.2.1 *Theobroma cacao* dissemination and cultivation around the world

The history of the spread *T. cacao* around the world can be useful to understand the pattern of genetic variations observed for the crop across the equatorial line. It is indigenous to South and Central America and believed to have originated from the Amazon and Orinoco valley which corresponds to present day Ecuador, Colombia and Venezuela (Motamayor *et al.*, 2002). The domestication of *T. cacao* is thought to have occurred during the mid-Holocene 7000 to 5000 years ago. The analysis of archaeological samples indicated that the south of Ecuador, an area that used to belong to the Mayo-Chinchipe culture, is likely to be the oldest centre of cacao domestication. This study reported three independent lines of evidence including DNA residues in cacao starch and absorbed theobromine which all date from at least 5300 years ago (Zarrillo *et al.*, 2018).

Genetic studies support the hypothesis that the crop was spread by humans from the Amazon to Central America. Evidence of cacao consumption by the Mokaya villagers in Central America date from 1900 B.C, further archaeological evidence indicating that the crop was part of the Olmec (1800-400 B.C) and the Mayan (1000B.C) cultures (Powis *et al.*, 2007). The Mayas and Aztecs were the first communities to process the whole cacao beans as a traditional drink for ceremonies and rituals. At that time, cacao was kept for the elite class and warriors (Vail, 2008) and in the Amazon region, the pulp was commonly processed for making alcoholic beverages (Belsky *et al.*, 2014; Crown and Hurst, 2009). The first contact of the crop with Europe came through the Spanish conquistadors at the end of the 15th Century with the arrival in 1492 of Christopher Columbus in the Caribbean followed later by Hernan Cortez in 1519 who led the conquest of the Aztec empire in 1521. Both became interested with cacao having noticed that some communities used the beans as currency, while others were producing a traditional beverage that contained cacao beans, honey and vanilla (Afoakwa, 2016; Belsky *et al.,* 2014; Henderson *et al.,* 2007). As an expensive and exotic product, it was brought back to Spain in 1520s where the drink was introduced. Cortes started to import small batches of the beans through 1528 to 1580 leading to the popularity of the chocolate drink in Spain, which then spread to Italy (1606), France (1615), Germany (1641), and Great Britain (1657). These

countries are still today the major consumers and processors of *T. cacao* worldwide (Afoakwa, 2010).

In the sixteenth century, large-scale (commercial) cultivation started in Central America and spread to the British, French and Dutch West Indies (Jamaica, Martinique and Surinam). Its cultivation further spread in Brazil during the 17th and 18th centuries. From Brazil, it was taken across the Atlantic, to what is now known as Equatorial Guinea and from there to other regions of West Africa namely, Ghana, Nigeria and Ivory Coast. It is also now grown in several Pacific islands, including Papua New Guinea and the Solomon Islands and South East Asia with production present in Vietnam, Sri Lanka, Malaysia and Indonesia (Afoakwa *et al.*, 2013).

The industrialization of cacao properly began in 1828 following the development of the cacao press (Dhoedt, 2009). Joseph Fry, a Bristolian, developed in 1847 the first plain (bitter) eating chocolate bar, incorporating each derived products from the cacao e.g. powder, butter and mass (Beckett, 2008). The demand for raw cacao beans dramatically increased and put a further impetus in the spread of *T. cacao* cultivation (Afoakwa, Paterson and Fowler, 2008)

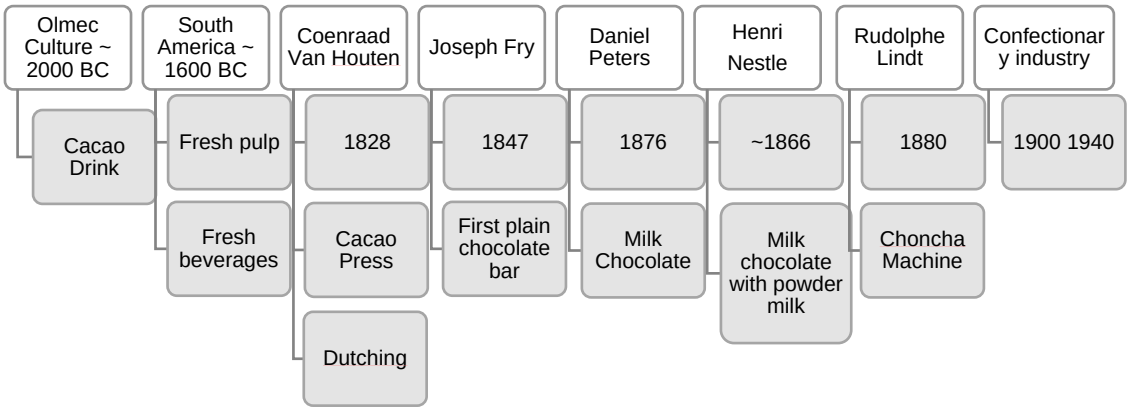


Figure 1.1 Evolution of the cacao industry, adapted from (Afoakwa, 2016)

In 1879, an African farmer, Tetteh Quarshie a native from Accra in Ghana, introduced the variety Amenolado (a high yielding cultivar), establishing a farm and nursery. Farmers started buying cacao pods from him to multiply the crop (Jonfia-Essien, 2006) and Amelonado became the prominent variety across West Africa and

still is today. As a result, the crop in West Africa exhibits a low level of genetic diversity when compared to other places in the world (Colombia, Ecuador, and Peru).

As of 2016, the cacao crop was grown on an estimated land size of 8 million hectares around the equatorial line countries. Fifty million people worldwide have a direct relation with the production of the fruit and eight million of them are smallholder farmers with a land size of fewer than two hectares (Afoakwa, 2016a; Ozturk and Young, 2017; International Cocoa Organization, 2012). Africa has 72.3% of the world production with 1.5 million smallholders, and Ivory Coast is recognised as the leading exporter and third worldwide processor. Since 2014, Ecuador has improved the crop and due to a higher global demand produces 13% of cacao from 37.43% of planted land. Ecuador exports more than 230 000 tonnes a year of cacao making it the fifth largest producer of cacao in the world (PROECUADOR, 2015).

1.3 Post-harvest of cacao beans

Cacao beans extracted from pods are subjected to postharvest procedures which include a fermentation stage followed by drying of the beans. These are key to the development of the correct flavours and aromas in both premium price and bulk cacao (Belsky *et al.,* 2014; Owusu, *et al.,* 2013). This process lasts 6 to 8 days in bulk cacaos but only 3 to 4 in fine and flavour varieties. The fermentation develops the essential colour and flavour so that the cacao can be processed for the chocolate industry (Ozturk and Young, 2017). It involves a unique spontaneous microbial development that is initiated with native yeasts that can have a relationship with the pulp of each variety of cacao, with Lactic Acid Bacteria (LAB) (*Lactobacillus fermentum*, *Lactobacillus plantarum,* 73%) at the initial stage of fermentation and Acetic Acid Bacteria (AAB) (*Acetobacter pasteurianus*, 92%) at the final stage completing the fermentation process. The main species, *Lactobacillus* and *Saccharomyces cerevisiae* are always active from the first 22 hours (Illeghems *et al.*, 2012).

There is a lack of information about the link between the microbiota complex of each geographical region and the development of the flavour, but it is assumed that in combination with weather conditions, these factors give the unique flavour to each

variety. The degradation of the pulp and the revolving (maceration) of the beans generates adequate conditions for providing precursors for the next microbial stages. As a dynamic process, when the yeast population starts to decrease after consuming the sugars of the mucilage, the population of LAB and AAB increase starting a new stage at around 56 hours of fermentation (Meersman *et al.*, 2013). Although the population densities of the secondary yeast species are relatively limited by 22 hours, their presence or absence may have characteristic specific effects on the fermentation efficiency and final product quality, either directly or indirectly. Techniques for affecting the fermentation such as removing the pulp or pre-drying the beans leads to a decrease in the sugar content which in turn results in low acidity and shorter fermentation times.

Many studies have focused on understanding the fermentation process for flavour and colour development (De Roos and De Vuyst, 2018; Kongor *et al.*, 2016). Cacao beans are microbiologically sterile when they have not been extracted from the pod. The extraction, the beans are exposed to several microbiota from farmers' hands, tools and environment. The initial fermentation over the pulp starts as an anaerobic hydrolytic phase and anaerobic yeasts generate CO2 and alcohol (Afoakwa, 2016a). Lopez and Dimick (1995) show that the cacao seed structure and anatomy promote the development of wild microbiota during the fermentation. The fresh average weight of the bean is one gram and the seed consists of two cotyledons, germ (embryo) a semipermeable testa (coat) and surrounding pulp (spongy parenchyma). On average, the pulp is composed of 80-90% water, 6-13% glucose, 6-7% fructose and 0.3% sucrose, 0.5–1% citric, aspartic, asparagine and glutamic acid with pH of 3 to 3.5 (Nielsen *et al.,* 2007; Crafack *et al.,* 2013; Mikkelsen, 2010). The pulp (mucilage) which surround the fresh bean is composed of 1-1.5% pectin, thus creating a good environment for microbial growth but also 10 – 15 % simple sugars, 2 – 3% pentosans and 1 – 3% citric acid with the concentration of glucose, fructose and sucrose varying with the age of the pod (Afoakwa *et al.,* 2013).

Biehl and Voigt (1999) demonstrated that the testa is permeable to water, acids, alcohols and volatile organic compounds (VOC), a study that was confirmed by Fowler (2009). This supports the hypothesis that the microbiota that helps to develop

the flavour would be in the nucleus or core of the cotyledons of the bean. Polyphenolic cells contain a single large vacuole filled with polyphenols and alkaloids such as theobromine, theophylline and caffeine. In unfermented beans when they have not been altered with acidification or high temperatures, the polyphenols give a purple pigmentation to the cotyledons (Osman *et al.*, 2004).

As a complex, the wild microorganisms that populate the mass starts with the activity of yeast liquefying and degrading the pulp by depectinisation, which reduces its consistency and produces ethanol under aerobic conditions. These conditions are responsible for the increase of pH and temperature, which will create ideal conditions for LAB and AAB growth. As the pulp is degraded, it increased air penetration that will lead to the cooking of the cotyledons stage when the embryo will die. Different fermentation techniques (wooden boxes, plastic tanks or hips) and drying approaches for the beans (sun, mechanical or gas) will have a major impact on the development of the flavour and aroma that gives chocolate it's quality and flavour profile (Guehi *et al.,* 2010; Bertoldi *et al.,* 2016).

After fermentation, cacao beans are either sundried or artificially dried, to reduce the moisture content from about 60% to about 7.5%. Drying is a key process that must be carried out under certain controls to ensure that off-flavours don't develop (Hamdouche *et al.*, 2015).

1.4 Chocolate making process from bean to bar

Dried fermented cacao beans are roasted and separated from their hull to produce cacao nibs. These, when processed, generate primary materials all used in chocolate manufacturing including liquor, cacao butter, cacao presscake and chocolate powder. These primary materials in addition to other ingredients (including milk, fruits, nuts and fats) are mixed in different proportions depending on the specification of the final producers who can be traders, confectionaries, ice cream, drinking chocolate manufacturers or grinders and pharmaceutical companies (Table 1.1).

Table 1.1 Production process from bean to cacao cake and butter

Initial product	Process	Final product	Industry
Dry and fermented Beans	Roasting, Crushing and winnowing (separation of the shell)	Roasted beans, Nibs, Shells	Roaster/Grinder/Bean to bar chocolate makers. Shells: Tea companies, animal feed, Agri-products
Nibs	Grinding	Liquor (liquid 100% cacao solids)	Grinders/Press/Dark Chocolate makers
Liquor	Press or refine	Cacao cake (Dry block), Cacao butter or Dark chocolate 100% cacao solids after conching.	Cacao butter; Chocolate makers, Confectionaries Pharmaceuticals, cosmetic, covertures for ice-cream. Cacao cake: Chocolate makers, beverages or powder products.

Source: Adapted from TOW Superfoods Chocolate making process 2017.

Classifications and cacao solids composition of cacao products ranging from milk chocolate to couverture chocolate has been developed by major multinationals since 1849's and are standardised in the Codex Alimentarius (Table 1.2).

Table 1.2 Chocolate types and composition

Chocolate types	Formula (composition)
White chocolate	Cacao butter, sugar and milk powder or others. (Bulk use vegetable fats)
Milk chocolate	Mix of cacao powder, cacao butter (34% at least), sugar, condensed, powder or whey milk. (Bulk use vegetable fats, glycerine, additives like emulsifier; soya lecithin)
Ruby (pink) Chocolate	A patent by 'BC', compound of sugar, cacao butter 29.5%, skimmed milk, whole milk powder, cacao mass 4.5%, emulsifier; soya lecithin, citric acid, natural vanilla flavouring.
Dark chocolate	Chocolate above 65% of cacao solids
Taza chocolate	Dark chocolate 8% at least, flour/starch from rice, maize or wheat
Coverture chocolate	No less than 35% of cacao solids of which no less than 31% shall be cacao butter and no less than 2.5 % of fat free cacao solids

Source: Author and adapted from Codex Alimentarious standard 2003

Since the technology to make chocolate bars were developed, the markets have diversified into a range of cacao bean-derived products including cacao butter, powder, alkaline powder, mass and liquors. Each derivative targets specific producer needs with the demand for cacao butter and powder still the main products in the market. Hence, they are used for most commercial candy bars, ice cream covertures and milk chocolates. It is estimated that 65% of the world grind is pressed into about 55% of cake (powder) and about 45% of butter. While 35% is processed into cacao mass and almost entirely used directly for the manufacture of chocolate. As an example, by 2010 as reported in the International cacao trade, the Netherland had the capacity to produce more than 120 types of cacao powder, which comprises different aromas, colours, and fat content (Dand, 2010). For 2019, the compound annual growth rate is estimated to be 2.3% from 2014 to 2019. The world cacao market is expected to be worth about USD 2.1 billion and USD 13.7 billion in 2019 (Beg *et al.*, 2017; ICCO, 2012). Chocolate production companies specialising in bulk chocolate are also investigating the possibility of supplementing cacao butter with *Theobroma Grandiflorum* butter because of its higher butter yield (Lim, 2012).

1.5 Chocolate market share insights

The supply chain and processing of the cacao beans is very complex (see Chapter 2); it depends on the quality of chocolate, colour patterns, market and product to be developed, These factors are influenced by the origins and genetics of the beans (Kadow *et al.*, 2013). The combination of these last two can differ between high-grade (fine and flavour) and common grade (bulk) cacao, which also leads to new industrial blends of powders and chocolate depending on demand.

In recent years, with the opening of Asian markets by introducing chocolate with frutal and natural flavours (Ruby and Whole Fruit chocolate developed by Barry Callebaut) and a new generation of young chocolate tasters and makers, the consumption of cacao products has increase by more than 13% worldwide. Dark chocolate now represents 10% of the global market. This new product development and opening of niche markets has influenced mainly the demand of cacao beans from South and Central America. South American countries like Ecuador, Peru and Venezuela hold a vast collection of cacao varieties, which represents a wide range of flavour profiles. Therefore, cacao and chocolate producers are looking to characterise these varieties to make more personalized products. Moreover, the expected compound annual growth rate (CAGR) of cacao market is 3.1%, whereas the chocolate market was estimated to grow at a CAGR of 2.3% from 2014 to 2019. The demand for cacao by 2020 is forecasted to rise by 30%, and the industry is struggling to obtain sufficient quantities from their current supplies (Beg *et al.*, 2017). Various strategies are being established to increase productivity but new requirements in sustainability and quality standards such as cadmium levels are limiting the trade of cacao between some countries (European Commission, 2014).

As these markets are shifting and growing, more studies describing flavour profiles and the benefits of consuming dark chocolate arise. Moreover, consumers are starting to learn that the flavour of the chocolate changes according to the origin of the beans and the percentage of cacao solids. This highlights new consumer expectations and demand. Dark chocolate production now bears a resemblance to the wine or coffee market with an attraction to more sophisticated and complex flavours than these offered by candy bars made from bulk chocolate. Also, dark

chocolate has less sugar and is richer in antioxidants and flavonoids. Research indicates that it could be considered as a more healthy product, which is also an incentive for consumers to acquire it (Petyaev and Bashmakov, 2017; Sorond *et al.*, 2008).

1.6 Food traceability and security

Malicious or unconscious fraud can happen for several reasons. When markets start to become interesting for new consumers or become a profitable commodity, there is a major opportunity for blending and replacing products. Most commodities from crops to bonbons have the risk to be fraudulent or mislabelled. This can affect the country of origin, the ingredients and welfare, eco-friendly, organic or fair trade claims (Bertoldi *et al.*, 2016). Rationalization or substitution fraud can happen by replacing higher-value food, like fine and flavour cacao from Amazonia or coastal Ecuador with lower qualities such as CCN-51 in a similar way to what was observed with the mixing of horse and beef meat (Doosti, Ghasemi Dehkordi and Rahimi, 2014), olive and palm oil (Ben-Ayed, Kamoun-Grati and Rebai, 2013) or frozen fishes (Warner *et al.*, 2013). Food fraud in 2014 was estimated to cost the global food industry US$30 to $40 billion every year and this number has increased by 12 billion in the last two years (Johnson, 2014).

With an increased interest in the premium market, food fraud relating to product origin is likely to increase. For instance, in various regions of Ecuador, bulk cacao (CCN-51 or Forastero) may be mixed with high-quality beans(Herrmann *et al.*, 2015) and this happened in the 1980s and in 2010. Ecuador was penalised for reducing the production of fine and flavour beans by the ICCO panel and for mixing the bulk variety “CCN-51” into batches of cacao sold as pure fine and flavour or Nacional Ecuadorian variety (2015). As a result, the Ecuadorian national research institute on cacao (INIAP) has been asked to developed nuclear markers to identify mixed cacao beans and mixed breeds in plant nurseries (Motamayor *et al.*, 2008; Loor Solorzano *et al.*, 2012). From 2019, the implementation of these technologies to identify mixed beans is being installed in border and customs controls.

One of the main strategies for cacao traceability comes through education and the development of protocols to characterised products. Brigitte Laliberte, responsible for CoEX and Bioversity International, highlighted the opportunities of developing such protocols following the World Cacao Foundation forum in Paris 2017. As director of Cacao of Excellence Programme, during the Managua, International consultation, she stated that to improve this mislabelling and confusion between genetic and flavour profiles, the cacao stakeholders from farmers to manufacturers need to be trained, educated and provided with tools (Cocoa of Excellence and Biodiversity, 2017). There is also a need to understand what buyers want and how producers can achieve it, and there should be a clear understanding of the connections between the ways that attributes in fermented/dried beans translate into roasted beans, liquor and chocolate (Cocoa of Excellence and Biodiversity, 2017).

Due to the adoption and the awareness of new consumers and changing markets, many multinationals and major cacao and chocolate manufacturers emphasize the need for traceability systems and certifications along the supply chain (Afoakwa, 2016a). One of the most used technologies for tracking the origin of food products is stable isotope ratio mass spectrometry. Isotope tracking allows the identification of adulteration or mislabelling of premium and protected foods ensures organic practices and confirms the addition of additives to premium products. Isoprime from Elementar UK company use it to trace the provenance of fruits, vegetables, meats, wines and more.

Since the flavour variation of dark chocolate is mainly due to the origin of the beans, there is a growing interest in tools that allow food control, quality and sustainability to be assessed. This can be addressed by improving supply chain transparency and traceability by developing novel tracking methodologies from Radio Frequency Identification (RFID) for the cacao bags to genetic barcoding of cacao beans (Germani *et al.*, 2015; Hawkins *et al.*, 2015; Galimberti *et al.*, 2013; Kane *et al.*, 2012). For the study of chocolate, some of these techniques can involve genetic analysis targeting the plant material but also the characterisation of the microbiological individual species/communities related to the manufacturing process.

1.7 Tracking the quality and origin of cacao products: current quality control approaches

More than 80% of cacao is consumed and processed in North America and Europe. Therefore, quality control for determining bean standards has been designed following the ISO 2292:2017 for cacao beans sampling. This standard focuses on medium-sized chocolate companies that purchase their raw products from medium to large traders. This is a morphological assessment done in triplicate batches of 100 fermented and dried beans involving a physical cut to expose the cotyledons. This allows the identification of slaty, violet and purple beans. Purple beans, for instance, are indicative of incomplete fermentation and will lead to undesired astringent and acidic flavours. This procedure also helps to determine if there has been mould or insect damage when the fermentation or storage has not been controlled (International Organization for Standards, 2017). One of the main issues for this test is that the margin of error can increase for batches exceeding one metric ton (MT) (Hii *et al.,* 2006; Ludlow *et al.,* 2016). At the 2015 annual symposium of the World Cacao Foundation, a group of advisors identified the need to develop standards for differentiating flavours by genetic profile. To fulfil this, it is first necessary to establish accepted, credible and verifiable protocols for assessing and communicating about cacao quality attributes, to facilitate comparison of samples and feedback, with the aim of improving fermentation and drying processes for each cacao variety.

Various organizations (e.g. ICCO, ISO, CAOBISCO and FCC) have developed quality control methodologies for production and trading. This means there are several analyses available to identify premium cacao varieties and good fermentation profiles. However, the main approach follows a sensory analysis using organoleptic panels and long chemical assays. This makes the companies dependent on human sensory skills to establish standards for flavour profiles (Branch *et al.,* 2015). As with any other fermented product, this method has been used for centuries to conserve and develop flavours and aromas in the final product. To date, quality control in the premium *T. cacao* market and chocolate industry has been mostly visual for fermentation, chemical for toxicity and physical for identification of varieties. Under EU law (European Commission, 2014), several standards for traceability tracking and

geographical certifications for controlling the supply chain are now required for food safety and sustainability.

Traceability is a core responsibility in the food industry and has several implications for cost effective processes, logistics and verification of quality. It is assessed to achieve consumer protection by targeting precisely the recall and elimination of the non- consumable food products and promoting the investigation of the causes of food safety issues. It is governed by acts, such as the Food Modernizations and Safety Act FMSA, 2011. H.R. 2751, (Badia-Melis, Mishra and Ruiz-García, 2015). On the other hand, genetic characterization and cacao traceability for quality control are required by nations and companies. For example, due to the increase in worldwide competition for dark chocolate production, Ecuadorian chocolate makers and farmers have been improving the fermentation process and characteristics of flavour notes (Otter, Prechtel and Theuvsen, 2014) . This has led to several chocolate awards with more than 400 recognitions from the International Chocolate Awards and Cacao of Excellence organizations (Cocoa of Excellence Programme, 2015; Cocoa of Excellence, 2017). Provenance mapping of the varieties of cacao from Ecuador has been characterised by private companies and the Institute of Agricultural Research (INIAP) in Ecuador, these data have been generated from tasting panels. The panels have been used to characterise regional chocolate origins (Manabí, Esmeraldas, Los Rios, Amazonia and Guayas) with the aim of controlling single origin claims.

1.8 Molecular DNA markers

DNA markers have become the key tools for tracking the provenance of food products (Krapp *et al.,* 2012). While food authentication can be achieved with protein, isotopes or metabolite markers, processing methods are less damaging to DNA, and more likely to provide accurate information for identification purposes. DNA markers can target the genome of species and cultivars utilised in the manufacturing of a product. Simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) are the two most robust markers used for identifying variations in plant DNA and usually target nuclear genomes as described by Singh, *et al.* (2013) who compared these two types of marker to characterise Indian rice varieties. An application in cacao

was to identify mislabelling in gene bank accessions with more than 15 – 44% of these accessions found to be wrongly labelled in 2003 (Motilal and Butler, 2003). The SSR and SNP markers can also be found on the chloroplast genome and have been used for investigations in plant tracking (Schroeder *et al.,* 2016). Several studies have also, looked at markers specific to the microbiota involved in the production of the food or present in the raw material used in the food chain. For instance 16s ribosomal amplicon screen using Next Generation Sequencing (NGS) were performed on cheese (Planý *et al.,* 2016; Escobar-Zepeda, Sanchez-Flores and Quirasco Baruch, 2016) wine (Bokulich, *et al.*, 2014) and commercial table salt (Gibtan *et al.,* 2017), revealing unique complex microbiota patterns indicative of origin or in some cases even the quality of the product.

1.8.1 Markers associated with *Theobroma cacao* genome

1.8.1.1 Single locus markers in the nuclear genome

Due to the economic importance of this tropical-fruit tree, *T. cacao* nuclear genome has been extensively studied, as the genetic improvement of the crop is essential to provide protection against major diseases and improve chocolate quality. Preliminary studies produced high density linkage mapping (Argout *et al.,* 2008) and were followed by NGS analysis of the whole genome of the crop (Allegre *et al.,* 2012). Further Expressed Sequence Tags – Single Nucleotide Polymorphisms (EST-SNP) and SSR polymorphisms were screened in a collection of 249 diverse genotypes representing the major part of the *T. cacao* diversity with 409 new SSR markers detected on the Criollo genome (Allegre *et al.,* 2012). The high-density map generated, and the set of new genetic markers identified are crucial in cacao genomics and for marker-assisted breeding, but they also offer a platform for chocolate tracking with the identification of variety specific markers. For instance, in 2015, Herrmann *et al.* conducted a comparative study to identify nuclear specific SSR alleles to the Ecuadorean variety CCN-51. Using the published 10 SSR markers generated by CIRAD (mTcCIR) they compared CCN-51 to the varieties Arriba, Criollo "Nacional", "Fino and Aroma", and demonstrated clear allele differentiation between these. They also observed a level of genetic variation within CCN-51 accession, which demonstrate that it is not a clone. While single-locus nuclear markers are highly informative, they have the disadvantage of being more susceptible to degradation due to low copy number.

Various research groups have used molecular markers to improve breeding programmes or study pest and virus controls. The international cacao germplasm database (ICGD) has incorporated molecular data in addition to other information and literature to support cacao research since 1991. This information has supported research in 40 centres including cacao genebanks and quarantine stations. The work of the ICGD has produced thousands of profiles for 15 SSR primer pairs on more than two thousand clones including the collections of CATIE, CRC, INIAP, Trinidad and Tobago.

1.9 Chloroplast genome and ribosomal regions

In contrast, markers associated with the chloroplast genome and nuclear ribosomal regions are less genetically variable but offer the advantage of being multicopy, which is important when studying DNA extracted from processed food. The chloroplasts are semi-autonomous organelles, which have their own genome and gene expression system. They are maternally inherited and therefore cannot be altered by cross-breeding. It has now become the marker of choice for plant DNA barcoding (de Vere *et al.,* 2012; Braukmann *et al.,* 2017) and used for example in the plant composition analysis of honey products (Hawkins *et al.*, 2015). In 2012, Kane *et al.* assessed the diversity of these genomic regions in *T. cacao*. The authors used high-throughput NGS to examine the whole plastid genomes as well as nearly 6000 bases of nuclear ribosomal DNA sequences for nine genotypes of *T. cacao* and an individual of the related species *T. grandiflorum*. This ultra-barcoding (UBC) approach demonstrated that all individuals examined were uniquely distinguishable. A later study (Hermann *et al.,* 2014) identified distinct chloroplast markers differentiating the two cacao types, Arriba (fine cacao type) and CCN-51 (bulk cacao) being cultivated in Ecuador. Most variation observed were SNPs but a different repeat of the sequence TAAAG in the inverted repeat region resulted in a different length of PCR amplicons for the two cacao types, which could be detected by agarose gel electrophoresis. These sequence variations were confirmed for a comprehensive cultivar collection of Arriba and CCN-51, for both bean and leaf samples.

1.10 Metagenomics of fermented cacao beans

Genetic markers representative of the microbiota involved in the fermentation of the beans are likely to be independent from the crop but could potentially be used for geographical tracking, as they are likely to be unique to the location where bean fermentation is conducted. Fungal and bacterial genotyping studies of microbial populations in food matrices have been extensive (Savazzini and Martinelli, 2006; Ludlow *et al.*, 2016; Camin *et al.*, 2017) and can be utilised as a tool for finding biogeographical locations according to genetic variation and distribution in microbe strains. Most in depth microbiological assessments for genotyping involves DNA-based protocols for single colonies which can be analysed by PCR amplification and DNA sequencing of highly conserved ribosomal regions including the 16S ribosomal RNA for the study of bacteria and archaea (Garcia-Armisen *et al.*, 2010; Moreira *et al.*, 2013) and the 18s ribosomal RNA for the study of fungus (Belsky *et al.*, 2014). Specific studies of genera or species can include the use of housekeeping genes from *Acetobacter pasteurianus* which is one of the most common and abundant bacteria in food fermentations (Li *et al.*, 2014; Ho, Zhao and Fleet, 2015; Meersman *et al.*, 2013). Admixture samples can also be analysed through electrophoretic profiling following amplified fragment length polymorphism "RFLP" procedures (Acierno *et al.*, 2016) or NGS (Araujo *et al.*, 2014). To perform these assessments with high efficiency and develop methodologies to fight against fraud or mislabelled products, DNA extraction generating reliable quality and standard yield is essential (Smulders *et al.*, 2009).

The quantification of markers specific to bacterial and fungal species in cacao products should also be indicative of the quality of the fermentation process, which is key to chocolate quality. Research and working groups from industry and academia have combine efforts to set up international standards (ISO 2451:2017, ICCO, CR, ENCI, INEN). Which indicate that exportable cacao must have more than 60% of fermented beans, not more than 5% of slaty beans, 5% of defective and 7% of humidity to be considered of a good quality (Loureiro *et al.*, 2017; International Organization for Standards, 2016, 2017). The majority of these quality control factors are dependent on the fermentation process.

Microbiological studies of cacao bean fermentation have been originally based on culture-dependent (Schwan and Wheals 2004; Meersman *et al.,* 2013) or culture-independent approaches with (Papalexandratou *et al.*, 2011b) using PCR-DGGE of 16S rRNA gene PCR amplicons of DNA directly extracted from fermentation samples to analyse the bacterial diversity. Metagenomics approaches, to generate a 16S rRNA clone library from total DNA extracted from heap and box fermentations has been set up in Ghana and Brazil. These samples were sequenced and revealed a low bacterial diversity in the fermentation samples, in accordance with the results obtained through culture-dependent and culture-independent analysis (PCR-DGGE). The common groups are lactic acid bacteria (LAB), acetic acid bacteria (AAB) and have important dynamics in relation to the metabolomics system.

Illeghems *et al. (2012)* performed the first (NGS) metagenomics analysis of spontaneous cacao bean fermentation. Using the 454 pyrosequencing platform, they were able to characterise not only bacteria but also yeast species including *Hanseniaspora uvarum, Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum,* and *Acetobacter pasteurianus* as common species in the fermentation process. They also identified a wider range of species in comparison to previous metagenomics studies including occasional temporal members of the cacao bean fermentation process such as *Erwinia tasmaniensis, Lactobacillus brevis, Lactobacillus casei, Lactobacillus rhamnosus, Lactococcus lactis, Leuconostoc mesenteroides*, and *Oenococcus oeni*. This demonstrated the efficacy of NGS for metagenomics analysis of cacao, identification of different processes of fermentation, and perhaps quality controls in the postharvest process (Camu *et al.*, 2007).

1.11 Aims

The overall aim of this research was to develop an advanced traceability system based on genomic methodologies that could improve the supply chain of cacao and chocolate by tracing the origin of sustainable fermented cacao beans. DNA biomarkers that target plants to obtain farm composition and microbiome from the post-harvest processes were assessed. To achieve this, the research sought to accomplish four specific goals:

1) Liaising with international stakeholders (traders, chocolate manufacturer, farmers), to assess how the development of DNA marker tracking system could be optimised for time and cost in quality control protocols in the supply chain of chocolate.
2) Assess DNA extraction methodologies on chocolate products.
3) Assess chloroplast markers usefulness in discriminating chocolate products for geographical origin and pattern identification in order to build a model, which could retrieve information about the haplotype composition of the farm in a commercial chocolate product.
4) Characterise and assess unique genetic biomarkers representative of the microbiota complex involved in cacao fermentation for the detection of geographical origin.

Chapter 2. Stakeholder outlooks on traceability within the cacao and chocolate supply chain

2.1 Introduction

Within the chocolate industry, the origin of cacao beans plays an important role in both quality assurance and brand storytelling (e.g. single-origin, fair trade) and the marketplace for premium cacao products has grown (Squicciarini and Swinnen, 2016). This rise has been accompanied by an increase in fraud (Mattevi and Jones, 2015), suggesting the need for tools to enhance quality assurance in the supply chain (SC). Alongside this, consumers have become more interested in the origins of the food they consume (Rousseau, 2015). Currently, certifications are used to demonstrate that companies have good control of the (SC) and pay fair prices to the farmers and stakeholders (Dragusanu and Nunn, 2014). However, these certifications have been shown to exert a negative impact on the SC, as they put pressure on non-certified actors to make unsupported claims about their SC, increasing the risk of fraud (Deppeler, Fromm and Aidoo, 2014). Such claims has led to the generation of new regulatory agencies, tougher international standards and governmental pressures (Recanati, Marveggio and Dotelli, 2018).

This chapter sets out to explore the challenges faced by the industry with regards to factors that influence the traceability of cacao at all stages of the SC, with a view to understanding the potential role of a traceability tool. The chapter starts by reviewing pertinent literature, before moving on to present research with stakeholders. Stakeholders were involved to develop a greater understanding of the SC and to gain insights into the need for a tool to identify the authenticity and origin of the fermented beans used in the chocolate manufacturing processes. Specifically, the chapter explores the potential for biomarkers (genetic fingerprinting) to be used within specific stages of cacao processing that require greater traceability control.

2.2 Literature review

Previous studies of the cacao supply chain (SC) have focused on commodity trading (Makhloufi *et al.*, 2018), manufacture optimization (Saltini, Akkerman and

Frosch, 2013), sustainability (Sonwa *et al.*, 2019b), fair trade and labour (Dragusanu and Nunn, 2014) and quality controls (Sukha *et al.*, 2014). This has highlighted the complexity of the SC between regions. However, even if organizations are mapping their SCs, issues around fraud (Teye *et al.*, 2020), child slavery (Berlan, 2013) and deforestation (Kroeger *et al.*, 2017) are still present in the production of premium and bulk chocolate production.

The need for controls as quality standards in chocolate manufacture dates back to 1850 when the chocolate market was emerging in Europe. At this time, tax on cacao bean imports to the UK rose from 17% to 43% of the price (Dand, 2010). This steep rise in taxation prompted confectionaries to find alternative ingredients to reduce production costs: for example, additives such as brick dust were used to maintain colour, while reducing the percentage of cacao in the product (Perlin, 2015). Similar practices were evident in France and The Netherland, which led to the development of food laws (Squicciarini and Swinnen, 2016), such as the '*Food and Adulteration act*' which was implemented as a result of discoveries by Mitchel (1848), Normandy (1850) and Hassall, (1855). As the market has changed and manufacturing technology has developed, new types of fraud have begun to emerge, including mixing of beans derived from diverse origins and genetic varieties (Teye *et al.*, 2020). This is an issue that affects sustainable-sourcing initiatives and continues to be pressing with the majority of companies and organisations setting a deadline of the end of 2025 to establish traceability and sustainability improvements (e.g. Cargill, Mars, Nestle, the International Organisation for Standardizations (ISO), and the Sustainable Trade Initiative (IDH)) (Nelson and Phillips, 2018).

At present, the main tools used for traceability are certifications such as 'Fair-trade©' and 'UTZ©'. The number of certifiers and self-certified companies is increasing, powered by a market of consumers that wants to know the origin of their food and to verify that the production processes and bean source are fair (Lalwani *et al.*, 2018). Regardless of the quality of the product (bulk or premium), organizations need to improve the SC visibility and transparency to adapt to the new customer and legal requirements (Recanati, Marveggio and Dotelli, 2018). Certification processes which depend on traceability have also contributed to improvements in the sensory

quality of chocolate products, merging both characteristics: sustainability and flavour profiles as a quality metric (Lalwani *et al.*, 2018). However, little is known about the specific needs of stakeholders within the cacao SC, a gap this research sets out to address.

2.3 Research questions

The global cacao SC has not been completely mapped. Various attempts to define it as a whole system have been performed by economic and logistic agencies such as The World Bank and Port of Amsterdam (Makhloufi *et al.*, 2018). This mapping shows a complex system, with farmers selling through a series of agents and traders. This is further complicated by the characteristics of products aimed at different markets. This research also suggests there are non-visible gaps and actors, which need to be mapped. This had led to research question 1:

***RQ1*:** Who is involved in the cacao supply chain and how?

In addition, little is known about the needs of these stakeholders concerning traceability and quality control. Therefore, this project explored the current actors and their interactions, as well as their perspectives and knowledge of the SC; and how this relates to their need for traceability. Such an understanding of stakeholders' needs is essential for the development of an effective molecular tracking technology for the chocolate industry and leads to research questions 2 and 3:

***RQ2*:** Which of these participants may be interested in genomic traceability technologies?

RQ3: What would stakeholders need from a genomic traceability technology?

2.4 Methodology

A grounded theory approach (which involves construct theories from systematically obtained data using inductive reasoning) was taken to understand and clarify uncertainties that are embedded in the global chocolate SC (Denzin and Strauss, 2006). Little is known about the interactions between stakeholders with respect to the global SC. The SC and interlinkages developed in this research were constructed from archival and respondent data.

An initial, theoretical supply chain (TSC) was constructed from a documentary analysis of public records and data in the form of recordings and notes gathered from conferences (forum) and fieldwork [including 1 Small-Medium Enterprise (SME) and 1 fast-moving consumer goods business (FMCG)]. These were also assessed to identify the types of stakeholders involved in the SC and their potential linkages. The data drew on previous case studies focused on particular countries (Ivory Coast, Ghana, Ecuador and Guatemala) (Cocoa of Excellence and Biodiversity, 2017) as well as more global data available through the international standards (ISO 34104-2019). In addition, the International Cocoa Organization (ICCO) annual report, for 2015, was analysed. These documents highlighted the process operations and quality controls in place, as well as the world cacao market, state of international trading and an initial list of stakeholders. This resulted in a theoretical SC running from farm to commercial product. This TSC was used to identify stakeholders representing a wide range of backgrounds and who could provide various perspectives on the trade (Orcher, 2007).

Stakeholder involvement

An email introducing the research aims was sent to 63 potential interviewees representing a range of interests in the SC, including Farmers, chocolate makers (*Maker: someone who transforms the beans or nib into chocolate),* manufacturers and grinders representatives, policymakers and governmental representatives. Contact was made with 46 and 18 responded agreeing to participate as advisors; of these, 13 were interviewed. A further group of international stakeholders was identified as attendees at the Chocoa professional forum and conference held in February 2018 (https://www.chocoa.nl/). Twenty-Five were contacted through the Chocoa networking contact list, with information on the research and an invitation to participate. 14 responded to the invitation agreeing to participate in an interview, 10 were interviewed. Details of stakeholders contacted and interviewed are provided in Figure 2.1 (see also Appendix II: Interviewees). Further opportunity to assess the SC was performed by attending cacao sourcing and chocolate makers forums, where public discussions were recorded. In addition, industrial fieldwork at a SMEs and FMCG grinder provided further information about the SC.

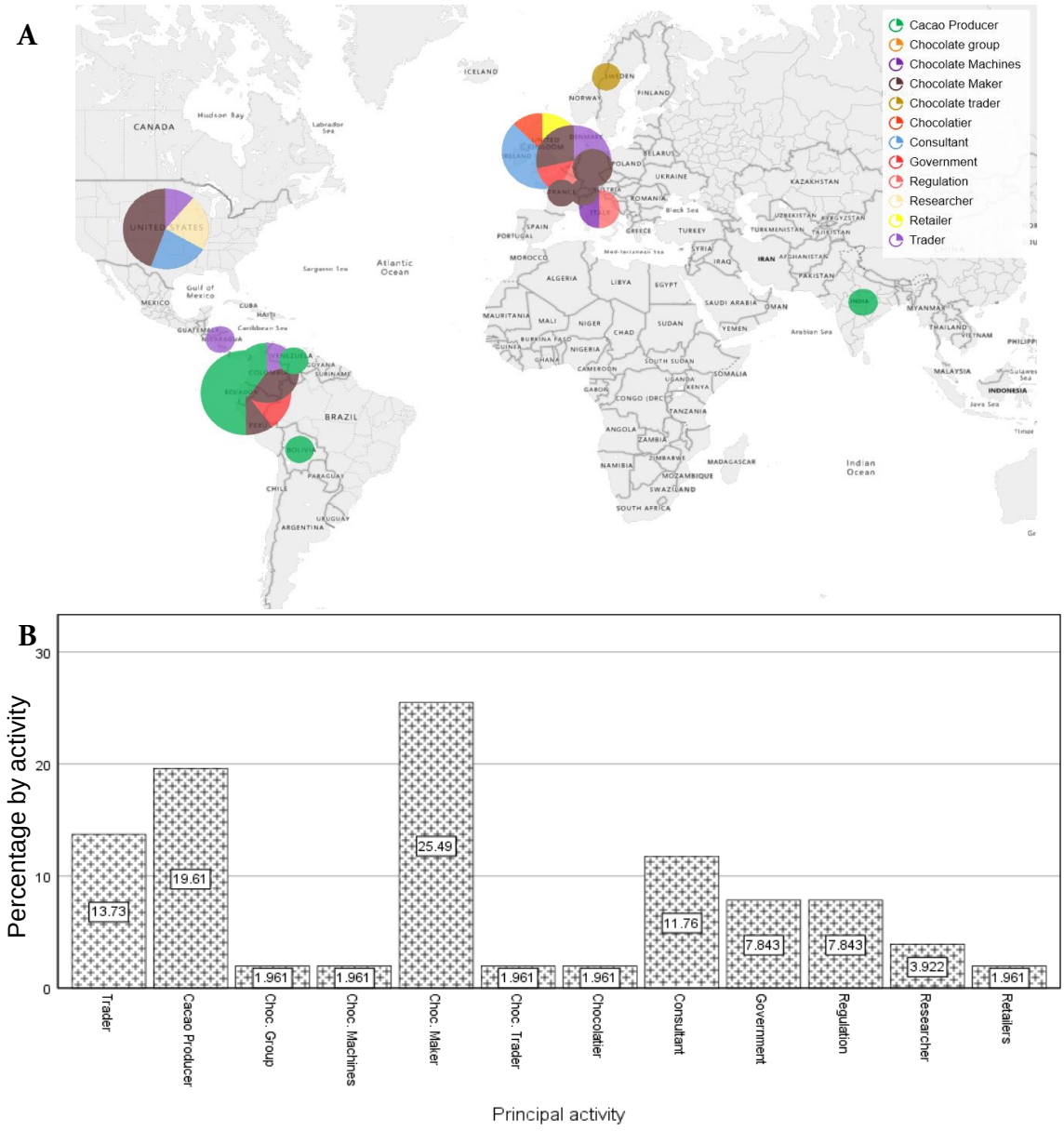


Figure 2.1 Stakeholders contacted classified by principal activities

A. The pie charts show the activities that the contacted stakeholders perform in each country while **B.** shows the percentage of interviewees by principal activity within the SC. Some interviewees have multiple roles, e.g. Choc. Makers can be traders and consultants, see (Appendix II: Interviewees).

2.4.1 Semi-structure interview

A semi-structured interview schedule was used to explore the potential for a tool to enhance traceability in the SC. This involved probing to understand the different elements of the cacao SC and how they are interlinked. The interviews considered existing quality control mechanisms and whether improvements are needed. It went on to explore how important it is to know the cacao genotype and if

there are any regions that are considered particularly important in premium chocolates.

A flexible set of questions was created and adapted to the specific cacao SC expertise of the interviewee (Silverman, 2004). Overall, interview questions were chosen to reflect the activities of interviewees, rather than formulaically asking each participant all questions (Foddy, 2009; Drever and Scottish Council for Research in Education, 1995). The interview schedule was designed to identify stakeholders' roles and their interactions in the SC. All interviews asked respondents to provide details about their (1) role in the organization, (2) business activities and (3) awareness of existing cacao and chocolate traceability mechanisms (Leal Filho and Kovaleva, 2015), which includes the personal or corporate quality control points and factors that influence their products (Gordon-Finlayson, 2010; Henderson, 2011; Somerfield and Clarke, 1997). Interviews' were conducted either face to face (4), through video or telephone call (4) and where this was inconvenient, the questions were sent by email and the respondent provided written comments (14).

All farmer and bean trader interviewees (N=7) were questioned about the genetics of their growing region and existing traceability controls. Manufacturers (N=5) were asked about the type of products they produce and if they lease other factories to manufacture their main chocolates, powder, liquor or butter; this question should reveal the types of control in their production process and the risk of cross-contamination and mixing of products. All 23 participants were asked about their participation in policy-making with certifications or governmental regulation and how they control the SC and sourcing of beans.

Stakeholders from production/manufacture backgrounds (N=11) were asked additional questions about quality protocols and tracking methodologies, while stakeholders involved in policy-making (N=11) were asked about current approaches for traceability control, the impact of policy on sourcing and products, current needs for traceability and their perspectives about a potential biomarker tool. All information gathered from; interviews were transcribed, and where appropriate translated from Spanish (by the author).

2.4.2 Thematic Analysis

The study adopted a qualitative thematic analysis approach to analyse the interviews and address the research questions (Clarke and Braun, 2017; Fletcher, 2015). Each code and theme was studied to understand how it related to current available documentation and information on the cacao SC constructed through documentary analysis. The inductive analytical approach allowed the identification of generalizable themes as they emerged from the data. Data were fractured into many patterns to find similarities and differences between stakeholders and also from the theoretical SC. Explanatory statements showed relationships between themes and the core categories (Quality, Traceability, Needs, Fraud) (Corbin and Strauss, 2012).

2.4.3 Ethics

This research was approved by the Faculty Ethics Research Committee. Consent forms, interview questions and questionnaires were produced in both English and Spanish.

2.5 Results

2.5.1 *RQ1*: Who is involved in the cacao supply chain and how? Theoretical cacao/chocolate supply chain (TSC)

The theoretical cacao/chocolate supply chain (TSC) was constructed from a case study made in 2017 by the Port of Amsterdam focused on the Ivory Coast and the ISO standards (Figure 2.2), (Mujica Mota, El Makhloufi and Scala, 2019; Makhloufi *et al.*, 2018; International Organisation for Standardisation, 2017b). Each stakeholder and stage of this SC was studied to identify if there were similarities between other countries (Ecuador, Colombia and Nicaragua) and company operations. The professional open forums held at Chocoa 2018 were used to reflect on this initial TSC, looking for similarities (cacao production, trading and logistics) and differences (between farmers, internal trading, logistics of cacao beans and processed products). The forum reflected main differences in the primary buyer of the beans and operations at farm, cooperative and manufacturing segments. The TSC was amended as an overview including the newly discovered gaps (Figure 2.2) and the different interactions (Figure 2.3).

Figure 2.2 Overview of the chocolate supply chain

An overview of the theoretical supply chain of cacao/chocolate (TSC) included multiple stakeholders per stage. Stages 1-5 are performed in the growing regions while 6-8 are often operated in consuming countries. 1, 4 can include direct collaboration between NGOs, Governments, Manufacturers. 2) Is performed by family members or by outsourcing pickers. 3) Include multiple private buyers (traders, farmers, cooperatives and manufacturers). 4) Performed by some farmers, coops, fermentaries and traders. 5) Mostly by third parties from logistic firms. 6) Beans come from multiple stakeholders, the SC varies depending on the products (Powder, butter, chocolate) and subsequently the trade-in stage 7.

Fieldwork to manufacturers (SME) and grinders (FMCG) (stage 6-7, Figure 2.2), showed that the initial material for production could be supplied as different blends of products: Beans, Nibs, Liquor, Cacao Mass, Chocolate or Covertures in its liquid form or blocks and Powder from a wide range of geographical origins and even various factories (Table 1.2). This means that the product can move from stages 2 and 3 directly to 6-8 (Figure 2.2), depending on which products are sourced by the company. Sourcing material directly at stage 2, places more of the SC in the control of the manufacturer. The TSC was correlated with the fieldwork and revealed stages 2, 3, as the first gap in the SC and the lack of traceability during harvest and post-harvest. Stages 1-7 were categorised by stakeholders as high risk of mislabelling, smuggling, beans mixing and therefore the risk of traceability loss. (Figure 2.2).

2.5.2 Mapping the cacao supply chain as a global system: stakeholder input

Stakeholder interviews were used to understand stakeholders concerns around control of the SC and their strategies for mitigating these. First, several uncategorized stakeholders were identified through the interviews. These are multiple segments of farmers, internal traders, chocolate producers (including self-declare makers) which operate as chocolate traders and certifying bodies with multiple missions and business approaches. Through the interviews, it became clear that farmers are not a uniform category, but are better split into micro, small, medium-sized and commercial farms. These different sized producers interfaced with the SC in different ways, leading to particular benefits and vulnerabilities for manufacturers. During Chocoa forum, chocolate makers in Europe talked about farmers being the weakest link in the SC.

Micro and small farmers make up 90% of worldwide producers and are the main suppliers to corporations. They have few or no quality controls for managing the beans as they grow cacao as a cash crop. They generally sell to traders or bigger farmers. Interviews revealed that they inherit the *T.cacao* plots or buy land with established trees. As Interviewee 7 says, “My plants come from more than five generations since my great grandparents got antique cacao plots”. In contrast, chocolate enthusiasts who are not natives from the country or regions acquire their own plots for experimental and niche chocolate business purposes. For example, Interviewee 8, says “I've planted 4 micro-plots and manage one additional plot that was planted decades ago and then abandoned; these are planted in the experimental agroforestry”.

There are also bigger farms, comprising: Medium size and commercial plantations. These farmers Most of these producers are part of associations (certified) and managed as established companies; they are often able to undertake collective bargaining to increase prices for their crops. Associations may supply directly to international traders and comply with established quality controls. The commercial-scale model is widely found in South America where intensive farming is applied as explained by Interviewee 19 “Large scale intensive farms seem to work in Ecuador and Brazil especially, in Ecuador it went to 90000tons, it is the number 3 in production”.

When medium-sized and commercial-scale farmers have shortages of beans they become co-dependent on smaller producers. This increases the risk that traceability will be lost, which is the gap that Fairtrade©, UTZ© and ISO accounting protocols aim to control. Cacao commercialization is also shaped directly by manufacturer requirements and is carried out among farmers, associations, cooperatives, wholesalers-regional traders and brokers.

Traders can also be further broken down into those which handle Bulk, Mix, and Premium products. For example in Ecuador, regional traders (acopiadores) operate as the first bulk buyer and quality control point. They are the main point of contact for bulk cacao sourcing. Mix traders source blends of beans from different regions, cultivars and with different post-harvest qualities. Mixed beans have the lowest prices on the market. Interviewee 6 explains "we sell them everything, CCN51, Fine and Flavour and also mixed, they do not have problems with the mix of genetics or premium quality for their bars".

Premium cacao traders claim and highlight the importance of getting to know personally their suppliers to acquire beans with high flavour profiles and quality. They explained that it's important for the chocolate maker to know whom they are working with and encourage them to visit the farmer, the farm and learn about where the beans come from and it is quality. The premium traders in Europe consider that they care about traceability because high-quality chocolate starts on the farm and needs to be shown with marketing and storytelling which is what consumers currently ask for.

Fermentation and drying are the most important stages of the T. cacao production, thus individual smallholders rarely perform these post-harvest steps at the farm. These processes have been standardized at cooperative and commercial farms. The need for quality and standardization has created the role of 'fermentaries' which are included with premium traders**.** Fermentaries receive fresh beans and do all the processes up to the point of export. This allows them to establish quality and traceability controls following a vertically integrated business model and helps them eliminate the risk of mixed beans and fraudulent claims. Interviewee 15: "Our business

model is about a centralized post-harvest…we have long-term contracts, so they have guaranteed access to the market. We supply the trees".

Manufactures were separated into two main groups**:** Premium chocolate makers and FMCG manufacturers, The first, usually claim to control the whole SC from beans to chocolate 'bars' and this forms part of their unique selling proposition (USP). Interviewee 20 explains "Origin and flavour of beans are important to how we get the flavours we love. We feel that direct sourcing is the best way to know what is going on and adds value at origin". The stakeholder analysis and fieldwork showed that premium chocolate makers can be segmented in three groups, makers that travel to origin to source their beans (Bean to bar from the source of origin), those that rely on traders to obtain the beans and provenance documentation (bean to bar). These two premium providers make up the smallest segment in the cacao market. The third type of premium chocolate maker is those that purchase chocolate from business-to-business (B2B) suppliers, such as grinders (confectionaries). The B2B model provides to them with the certifications, claims (Free from child slavery, Organic, Sustainably sourced, Genotypes) and quality controls that they require giving them a role more like a chocolate trader, which eliminates the difficulties of dealing with farmers and cacao trader. B2B suppliers represent the second largest segment of the cacao market.

Most of the cacao SC revolves around FMCG. As Interviewee 18 states, "Only three or four companies control de 80% of the market. The market share of sustainable cocoa is 15% to 20% and is increasing". They influence the whole SC, setting up guidelines and conditions to buy the beans from the farmers. These organizations act as big traders managing most of the beans across the world. They are mostly vertically integrated industries, which control the plant supply to farmers as explained by Interviewee 19, "We are the biggest supplier of plants to farmers. In Indonesia, we supply 5 different clones, but the farmer doesn't know".

FMCG manufacturers are the main industries that source Fairtrade©, Rainforest-alliance/UTZ© certified cacao and they also trade their beans under their own certification labels. By sourcing cacao with certifications, the companies are able to market (Story-telling, claims) and trade (Premium prices) their products accordingly. All of the activities that these firms do are overseen by several trading

and quality control policymakers, who play an important role in regulating the SC. The full SC and linkages, as uncovered through documentary analysis and interviews is shown in (Figure 2.3).

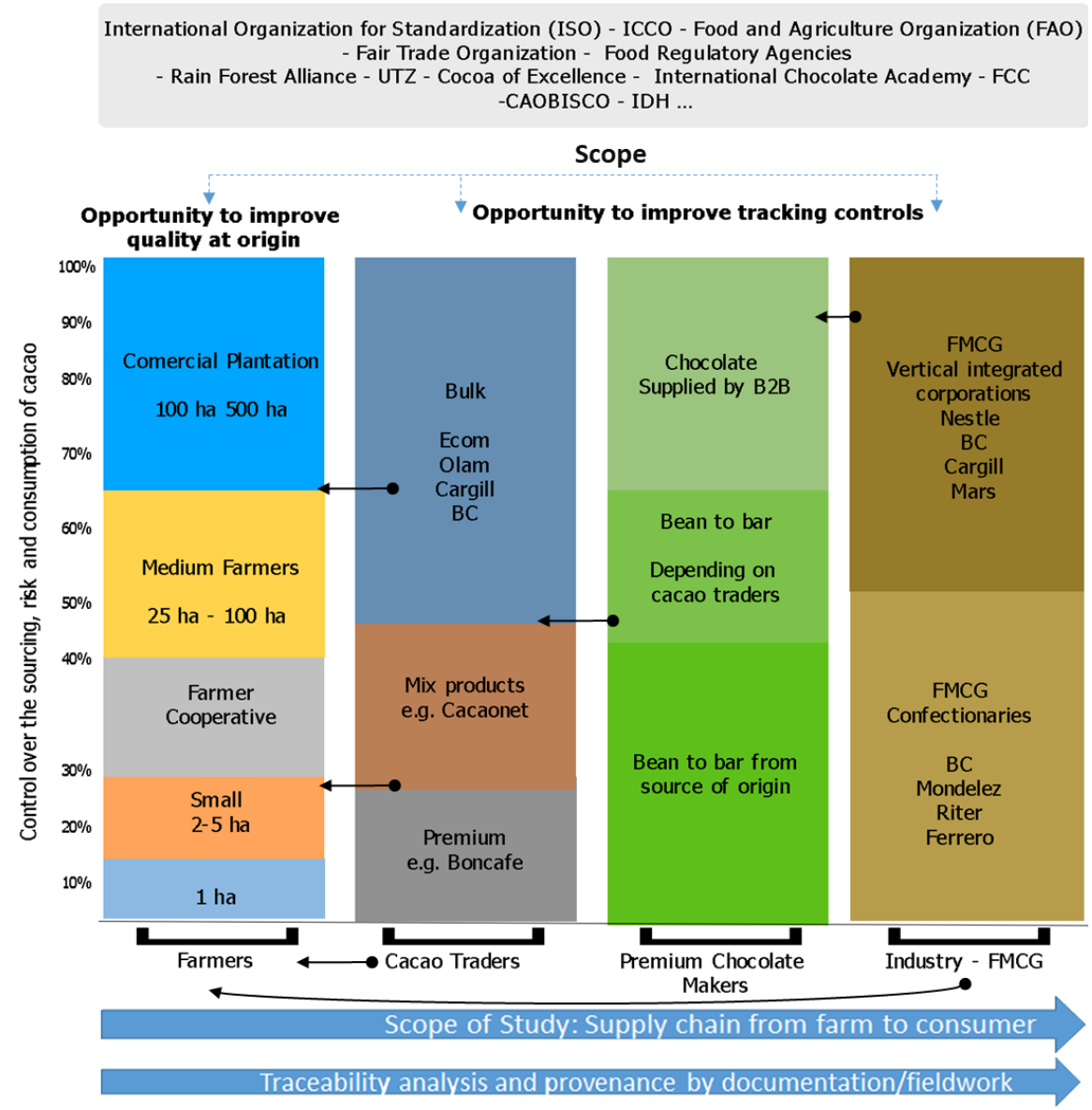


Figure 2.3 *T.Cacao* supply chain system, stakeholders and interactions in the chocolate production summarize the *RQ1*

Cacao beans flow from left to right. The 4 stacked bar charts increase in size to express the quantity of cacao/chocolate that these groups manage and therefore their influence over the market share and industry. The proportion of the stacked bars (y-axis) are based on the quantity of supplies that they move or require in order to fulfil the demand of their customer (shown with arrows). European chocolate demand accounts for approximately 60% of the market. The lead manufacturer Barry Callebaut© uses more than 25% of world cacao production. In contrast, the USA accounts for 27% of the market. Above the stacked bars, the regulatory bodies and policymakers who influence every cacao/chocolate stakeholder are listed, showing the scope of their influence.

2.5.3 *RQ2*: Which of these participants may be interested in genomic traceability technologies?

Two main factors drive the need to improve the traceably and visibility of the cacao SC: international trade regulations and marketing strategies. The analysis presented above shows that Policy Makers, Chocolate Makers and Traders agreed that cacao production is the main site of bean mixing, presenting a risk to quality control and traceability. As Interviewee 8 states, "This [small scale farmers] is the first point of contamination, and the hardest to control, because there are thousands of growers, and the mixing happens in the privacy of their own farm". Likewise, premium traders indicated that even if traders/makers visit the farmer, controls and consistency are difficult to achieve. A trader during the sourcing forum at Chocoa 2018 highlighted that, within cacao trading there are risks along the whole SC: Internal trading, storage bags and logistics.

Traders and chocolate makers expressed their discomfort about quality (fermented or not) and the mixing of beans (cultivars) from various origins. For example in Ecuador, Interviewee 7 explains, "Separating mixed cacao is impossible. We have as a policy here, different prices for selected and classified nationals and CCN51 beans". The beans that they gathered are priced on the weight of the bag and depending on the trader; prices might change if there was some post-harvest processing or for approved high-quality flavour genotypes. Traders selling both types (Mix and Premium) are potential sources of fraud, because only premium quality beans, beans with high-quality flavours and beans certified as from sustainable farms, traded under fair practices secure higher prices. This means that traders selling Mix or Premium beans may include a percentage of bulk beans to increase profits.

To solve the issues of consistency, price and sustainable practises, governments, policymakers and chocolate makers encourage the creation of associations that follow certification schemes. This facilitates international trading and tracking (Figure 2.3). However, ensuring certification is costly, as explained by Interviewee 19, "It is too expensive, [there are] difficulties to take people to the places. Inertia, poor training, getting farmers to do anything! If the recommendation comes up with 5 different things, make the farmer do that! hmm".

2.5.4 RQ3: What do stakeholders need in relation to traceability?

Interviews showed that as part of the branding and supply chain transparency (SCT), each FMCG and some SMEs have started creating their own certifications, traceability system or methodology to demonstrate their sourcing practices and their positive impacts towards sustainability. Interviewee 19 explains, "We have our own certifications process (Cocoa Horizons)". This increase in private certifications are not exclusive for chocolate makers, it is also a form of branding for premium traders. Interviewee 16 explains, "We have even made our own certification system called COCOA ID, so cocoa with an identity".

There are various reasons why traders (bulk-premium) work with specific farmers, associations and communities. Some are related to quality and others about social aspects. Interviewee 6 explains "Producers of these regions were cheated; losings more than a hundred thousand dollars and they needed someone to help to trade and connecting their community with the world". These difficulties in crop production and quality controls have been documented since the 1900s and are still a problem in *T. Cacao* producing countries. Moreover, it was found that a group of NGOs and researchers have developed a type of certification to award 'premium' genotypes so traders and chocolate makers can claim flavour profiles relate to cultivar. Interviewee 22: "I work with HCP [Heirloom Cacao Preservation Fund], it is meant to preserve cacao genetics and flavour thus their groups are already privilege cacao producers and they want to sell it at higher prices". Nevertheless, all the efforts for implementing traceability systems to prove the origin are more related to fulfilling policies and are not related to flavour profiles, as the main aim of most certifying bodies is to improve sourcing control to avoid smuggling, child labour and deforestation. Interviewee 19 explains, "The worst thing that can happen is getting cocoa from deforestation side; there is child labour and the whole poverty thing. Trying to get farmers out of poverty".

Interviews showed that the main actors involved in certifications are NGOs and governments that have established cacao recovery programmes such as in the cases of Ecuador, Guatemala, Nicaragua, Colombia (Cocoa of Excellence Programme, 2015). The majority of these programmes aim to connect farmers directly with agriculture

and chocolate professionals, to guide farmers in the production of high-quality varieties and to find approaches for creating farmer cooperatives linked to premium and sustainable markets. Interviewee 18 argues that "The farmers must be helped by organizations and work together with a cooperative or farmer organization. To document and be recognized from the beginning as sustainable". However, even if organizations approach farmers and offer assistance, most small and medium-scale farmers are not able to meet the basic requirements of certification, or are not willing to change their practices, as explained by Interviewee 19: "If the guy has only half a hectare and is not part of the cooperative, it is hard for us to do anything. It can be quite challenging but we are trying".

Certification schemes require a certain level of accounting and management knowledge, which makes the small farmer dependent on business brokers with experience, such as traders, buyers or international advisors. Cooperatives sometimes have a business and legal schemes, which enable farmers or internal traders to mix, top-up and smuggle uncertified cacao beans in the batches, even when the cooperative is certified. Interviewees also highlighted the risks and inconsistencies that appear when using only document-based traceability as it is imperative to visit the farm and map the boundaries of the land. For example, some farmers 'borrow' documentation and certificates from neighbours who sell the beans for them. Interviewee 6 explains 'Union Nacional Association (UNA)' legally has the certifications, but Association 'Pepa de Oro' sells their beans independently using these [UNA] certifications because they can't afford to pay for their own [certification] yet".

2.5.5 Current approaches to avoid mixed beans and to improve traceability control

Cacao consuming countries and organizations create policies and standards designed to ensure that the product coming to their market meets appropriate standards. These standards are the main approach to improve traceability and are part of governmental regulations, private policy projects or provided by independent bodies such as; ISO, FAO, WHO, Fairtrade©, Rainforest-alliance/UTZ© and Codex Alimentations. Interviewee 18 explains that these tools do not work in isolation "The ISO certification will not be beneficial alone but will help to improve all countries

infrastructure. The higher the quality of cacao and chocolate the better the traceability".

Fieldwork uncovered a particular problem for farmers: premium or certified chocolate industries are niche markets and do not purchase beans on regular schedules. This means that farmers are likely to have to sell some beans to bulk purchasers (at a lower price), even if they are certified. Since certification can be expensive, this presents a challenge for the farmer that discourage them in adapting to certifications. Moreover, the certification scheme and ISO standard allows the process of segregation "process that separates conforming from nonconforming cocoa, but allows mixing of conforming cocoa from different cocoa supply chain actors" (ISO 30102-3), which aims to support non-certified farmers. This, in turn, means that farmers and buyers following good practices need a simple system that can identify the physical origin of their cacao at affordable prices and where information can't be corrupted so they can sell their beans at a fair price. The premium cacao traders in Chocoa explained, that they fulfil a function in the SC that is more than just importing and exporting beans. In this sense it is more about long-term relationships that they create directly with the farmer, taking on the risk for the chocolate makers that don't know about the logistics and quality controls at the origin.

All the stakeholders highlighted the importance of controlling the source of each batch of cacao beans. While others visit the farmers at least once a year to forecast the harvest, which suggests that bean quality and traceability needs to be controlled at the farm. Fieldwork to farms showed multiple controls required by the industry, this includes strong controls such as not mixing different cultivars, or basic controls by documenting traceability by batches and by plots (Figure 2.4). For the chocolate makers, there are substantial advantages of working directly with the farmer and vice versa.

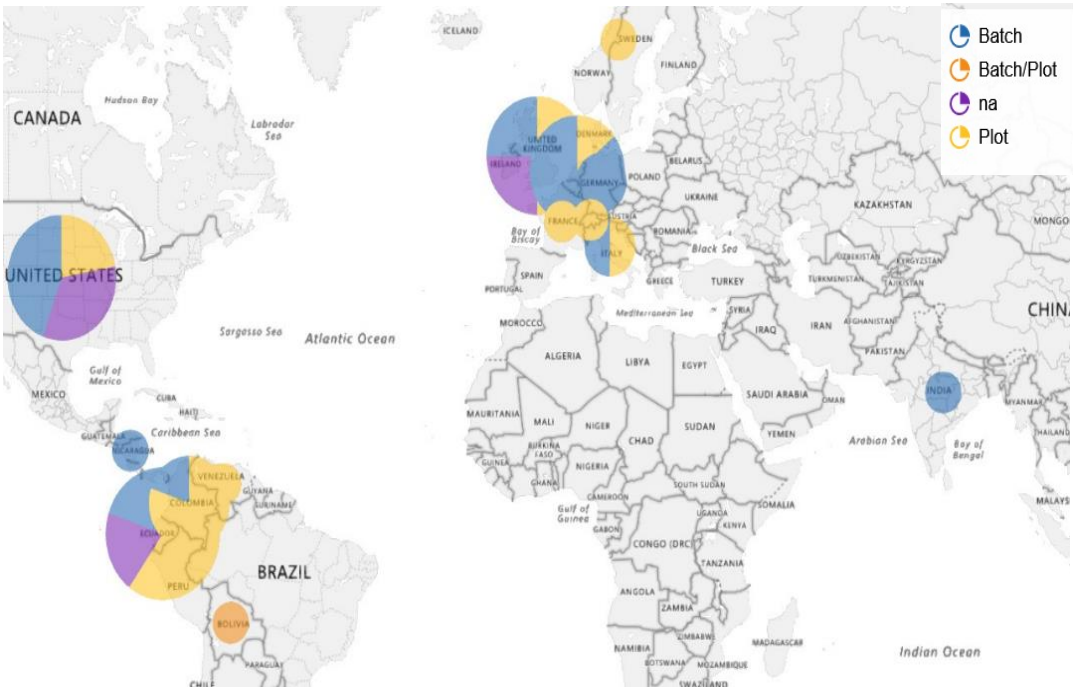


Figure 2.4 Stakeholders that have to do farm traceability inspection per year or plot (farm)

The pie chart shows the distribution of stakeholders that visit each farm (Plots) which are mainly regional traders or certified companies and companies who source the beans from traders control each production batch by collecting documentation about the farm and quality standards. Buyers who buy bulk cacao don't control any traceability. Batch: Each Lot of beans, Batch/Plot: Each lot of beans per plot of harvested land, na: No inspection of the farm, Plot: Inspection of the land.

Fieldwork to farms, helped to understand the history of their land and economic situation, it assisted to identify their structure, why and how organizations set up direct agreements with them. Interviewee 19, explains "Cocoa farming is especially from small size, is a couple of hundred of farms on a daily basis, so we had to do a lot of emphasis on the origin". Farm and batch inspection is performed by the premium traders, manufacturers and policymakers and is the most controlled way of verifying the origin, process of postharvest and acquiring unique flavour profiles from different regions. However, fieldwork and visits are the most expensive approach to take. Some companies are willing to implement technologies (Satellite mapping, Blockchains, and DNA biomarkers) to optimize this process. For example, the collaboration between Barry Callebaut and SAP [Systems, Applications, and Products] company, combined satellite mapping and metadata from the farms and processing facilities to verify the volumes that cooperatives are supplying. Interviewee 19: "We have a huge and growing database called KATCHILE©, where every farmer we work with is either

in there or going to be. It has got the amount of cocoa he produced last year, family [Farmers] for transparency and impact".

2.5.6 Stakeholder needs for a biomarker-based tool

As most of the quality claims come from descriptions about origin and cultivar, the stakeholder categorisation explored the assessments stakeholders use to identify the farm and cultivar (genotype) of cacao beans. The majority uses the morphology of the pod and trees as a guideline, while others profile the flavour of the bean, liquor or chocolate and a few have been able to confirm the genotype from the farm by Single Nuclear Polymorphism (SNP) DNA tests or fingerprinting assessments (Figure 2.5). This is a one-time genetic profile test of the plot and can be used to identify varieties that have a unique flavour profile that the producers, buyers or claimant want to use as a marketing tool. SNP may also be used to reproduce cultivars. Interviewee 22 explains that by using SNP tests "they hope … [to find that] they have something exclusive. To create a new market around the genetics of this cacao [plant]". The unique characteristics of the trees on the To'ak company farms have helped it become a luxury brand and a conservation project at the same time. Interviewee 8 explains "There are tons and tons of Fair-Trade certified cacao on the market--that's not such a special thing. How much cacao out there is certified 100% pure Nacional [special Ecuadorian variety]? Very little".

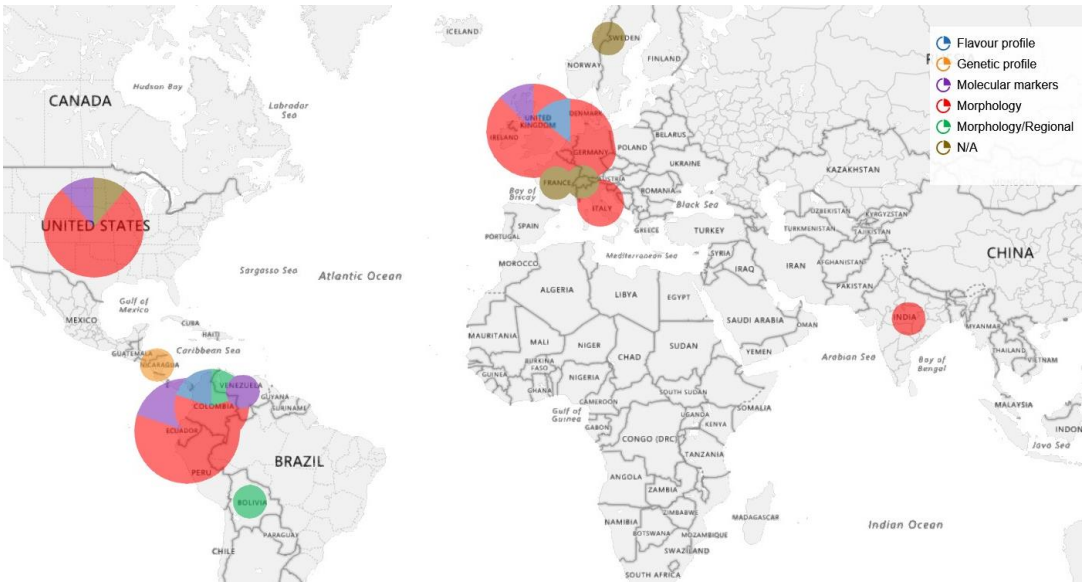


Figure 2.5 Current technical assessments to determine the *T. Cacao* genotype

The figure illustrates the types of cultivar verification that the stakeholders do before making an agreement with a farmer or as an internal control.

The main reason why stakeholders require biomarker-based technology is to avoid sourcing beans from deforested, unsustainable areas and to support marketing strategies that increase the price of their products (Table 2.1). However, as Interviewee 16, notes, it is important that such a technology is woven into the certification process, "Even if you control the whole supply, chain and you can prove the tracing, it is important to certify. Because consumers understand certifications, they don't understand assurances".

Table 2.1 RQ3 summary: What do stakeholders need from a genomic traceability technology?

The needs of various organizations and the role of the interviewees and fieldwork notes.

Organization	Role	Reasons a biomarker tool could be helpful
UTZ/Rainforest Alliance	Policy	Identify beans smuggled from deforested areas.
Cocoa of Excellence/Bioversity International		Demonstrate that genotypes are regional and have unique flavour profiles.
ISO Cocoa sustainability		Proving transparency of the supply chain and that cocoa has been obtained from sustainable and fair sources.
Cocoa Research Association (CRA)	Research	Scientifically support claims as a marketing tool
Cacao Cayapas	Trader	Demonstrate that genotypes are regional and have unique flavour profiles instead of certifications.
Cacao Desidente		Marketing tool to support post-harvest claims by showing research about the microbiota and origin
Cacao Caribe	Farmer	Demonstrate that genotypes are regional and have unique flavour profiles.
Chocolate Norway	Chocolate trader	Demonstrate that genotypes are regional and have unique flavour profiles instead of certifications.
Fine Cacao and Chocolate (FCCI)	Academic	Marketing tool for luxury brands and to determine beans from unsustainable and forced labour regions.

2.6 Discussion: Conceptualizing 'Biomarkers as a system to improve sustainability and supply chain'

RQ1: Who is involved in the cacao supply chain and how?

The cacao SC is complex. As a commodity product, multiple industries require different components and sub-products made from the beans and multiple contradictory claims (Deppeler, Fromm and Aidoo, 2014) about the origin are made. In turn, the TSC has categorized cacao production as one group of farmers, one for cooperatives, one for traders and all the manufacturers as chocolate makers

(Makhloufi *et al.*, 2018; Dand, 2010; International Cocoa Organization, 2019). This research showed that the TSC has various uncategorized stakeholders and non-visible gaps through the SC. These can be segmented depending on their operations (Figure 2.3) e.g. small-farmers rely on internal traders, brokers and NGOs to sell the beans under fair prices, some chocolate producers claim to make chocolate but are supplied with the final product by B2B firms, giving them a role as branders (Folds, 2002). In addition, it was identified that fermenters or post-harvest stakeholders were missing from the TSC. These stakeholders may have a large input on farmers operations as they implement controls between the farm and certification processes. It is important to notice that most of the interactions are influenced by the requirements of FMCG manufacturers, which consume 80% of the world cacao production and require multiple types of certification.

***RQ2*: Which of these participants may be interested in genomic traceability technologies?**

From the interviews and fieldwork, it was identified that not all the stakeholders are interested in acquiring or improving traceability. Interviewee 10 and a firm from Switzerland which participated in the Chocoa forum stated that they don't require genomic tools as they have selected farms that they work with. These are either long-term contracts or indeed farms they own. Moreover, they argued that the chocolate profile is based on their recipe and therefore the genotype of cacao is irrelevant in terms of quality. In contrast, governmental and NGOs representatives highlighted that farmers and premium chocolate makers need tools to showcase the high-quality of specific cacao genotypes and regions. This mutual collaboration helps to brand countries and improve their rank in the ICCO. This also enables market segmentation within vertically integrated companies, allowing them to identify specific genotypes for their premium or bulk markets, thus also for their breeding programs. Apart from the marketing aspects related to flavour profiles, NGOs, SMEs and FMCG have an urgent need to improve the traceability and ways of identifying farms for purposes of sustainability. A number of organisations (e.g. Cargill©, Mars©, Nestle©, The International Organization for Standardization (ISO), The Sustainable Trade Initiative (IDH), Centre for the Promotion of Imports from developing countries

(CBI)), have set deadlines for sustainability and traceability improvements by the end of 2025. These address issues such as deforestation, forced labour and the common aim of achieving the United Nations Sustainable Development Goals (Johnston, 2016).

***RQ3:* What would stakeholders need from a genomic traceability technology?**

Analysis of manufacturers by size (Artisan, SMEs, and FMCG) showed that businesses claiming to sell 'Premium' products, typically use the origin (Farm, Cooperative, Country) or assumed genotype of the beans as a quality marker. From available products and the opinion of these stakeholders when referring to quality, it was noticed that claims are being correlated with different native growing regions that are known for producing specific "high flavour profiles", pushing forward the production of single-origin chocolates (Cidell and Alberts, 2006). Participants in the SC (Commercial farmers, traders and manufacturers) build a story around the attributes of beans from particular locations and market it to their clients as a flavour profile and story of origin. This research suggests that there is potential to add value in the area of traceability of single-origin products regardless of the size of the company; traceability could be useful to support sustainability and transparency claims **Error! Reference source not found.**.

Stakeholders identified two main products that could benefit from a biomarker tool: 1) Certified (bulk) products and 2) Single-origin chocolate (Premium). Bulk products need to show transparent practices as part of the certification process. From the discussions with FMCG representatives, while this tool could improve control of bulk produced chocolates. At bulk scale, cacao genotyping is currently used to identify breeding and yield characteristics rather than flavour patterns as the amount of cacao used in the bulk chocolate bars is relatively low. However, bulk chocolate products use country flavour profiles which are mixed in the factory with chocolate from other origins and which have a legal requirement to improve their sustainability. Thus, there is a need to prove that companies are sourcing chocolate from sustainable, non-slave labour, fair-trade, non-deforested areas (UTZ, 2016; International Organisation for Standardisation, 2017a). A biomarker-based system could provide an efficient and reliable tool to barcode farms or cooperatives, giving this assurance of sustainability to the purchaser.

Single-origin chocolate manufacturers also typically make claims related to sustainability, such as the social impact on the region on farmers and poor communities (Acierno *et al.*, 2018; Petyaev and Bashmakov, 2017). For single-origin chocolates, stakeholder analysis revealed that there is a direct link between the regional origin and cultivar (genotypes) and that various players (firms and governments) have been genotyping and cloning varieties with unique characteristics to use as a marketing tool. Various stakeholders from several countries linked the potential of a biomarker tool to the ability to detect genetic varieties like Nacional (Ecuador), Matina 1-6 (Costa Rica), Sca6 (Costa Rica) and ancient Criollo (Mexico), Chuncho (Peru). These varieties not only have historical value to the country but are also used as breeding stock for future premium chocolate plantations. Initially, genetics and flavour profiles were studied privately by chocolate industries such as Nestle©, Mars©, Lindt© and by government institutions in cacao producing countries and other big importers. One such project resulted in a publicly available whole-genome sequence of ten cacao accessions (Zhang *et al.*, 2009).

Historically, chocolate and cacao stakeholders have categorized cacao origins into different quality ranks depending on the market that they want to reach, flavour profiles, yield and resistance (Motamayor *et al.*, 2002; Cornejo *et al.*, 2018; Kongor *et al.*, 2016). This research indicates that there is no clear standardized concept about how the quality of the final product relates to the origin of the beans and that flavour profile does not mean quality chocolate. While there are various guidelines (International Cocoa Organization (ICCO), 2009) this remains a subjective characteristic. Most of the participants' consulted provided different assessments that they used to measure the quality of the cacao beans and chocolate: Variety, Growing regions, Post-harvest process, and Exportation standards. This showed that any bulk or premium variety can be of high post-harvest quality and chocolate confection but it does not mean it has a high-quality flavour profile related to a particular genotype (Kongor *et al.*, 2016; Silva *et al.*, 2014; Dand, 2010).

Moreover, it was identified that there are pressures, which drive governments to develop and acquire technologies to control fine and flavour cacao qualities, protected areas and fraud. Some of the stakeholders in the cacao industry explained

the need for these controls. For example, Ecuador has always been known for its premium quality and production of fine and flavour beans. However, in the past, the country was penalized for exporting bulk cacao mixes (CCN-51), when claiming to be exporting fine and flavour beans. This affected Ecuador's rank by the ICCO and influenced pricing. This led the government to generate national strategies along with the SC of the crop, which would maximise value.

2.7 Conclusion

The use of biomarkers can help to support supply chain transparency

Stakeholder interviews underscore the need for standardization of quality terms, sustainable practices and linkage between growers, policymakers and manufacturers. The research fills a gap in our knowledge of the specific needs of stakeholders within the cacao SC, in relation to traceability of bean origin and type, highlighting the need for new tools. Such tools could help mitigate the risk they face of contamination from mixed beans at key points in the SC. In addition, certified (UTZ, Fair-trade) stakeholders and some other stakeholders are interested in a tool that would demonstrate to customers their sustainability and flavour credentials. However, the potential economic savings were the main reason most stakeholders were interested in such a tool (i.e. if it helps to cut the cost of certification).

Cacao genotypes have been grouped as fine and flavour, and bulk varieties, which comprise of hundreds of cultivars. This means there is a large number of genetic varieties that are presented as unique national varieties (Motamayor *et al.*, 2010). This leads to three main reasons why it is important to stakeholders to identify the characteristics of provenance in each country: 1) Governmental pressures related to national flavour profiles that can improve country ranks within the ICCO and other global flavour markets. 2) Supply chain transparency of products with impact on various sustainable development goals (e.g. No poverty, climate action, responsible consumption) and 3) Scientific support for marketing and brand-story telling of the beans and chocolates. As there is a wide variety of players in the cacao SC, there is a need for a simple system that could offer the presence/absence of a proportion of a sample coming from a specific region or supplier.

This proposed development focuses on a biomarker to be used in the first stage of production. This is likely to be the best option to optimize the traceability control, as the genotype of the farm will not change nor the microbiome composition of the specific cooperative Figure 2.6. The DNA pattern of the crop or microbial composition of the place should be replicable between seasons and give the opportunity of tracing beans back to their source. This sets the path to develop two types of marker system, one that could target the farm genotype composition and a biomarker targeting the cooperative or location where the beans are fermented post-harvest. Moreover, it presents the opportunity for generating new tools to identify unique qualities per region or process, which can lead to new policies and standards based on the genetics linked to the farm, cooperative or country, which will need to be related to traceability and provenance.

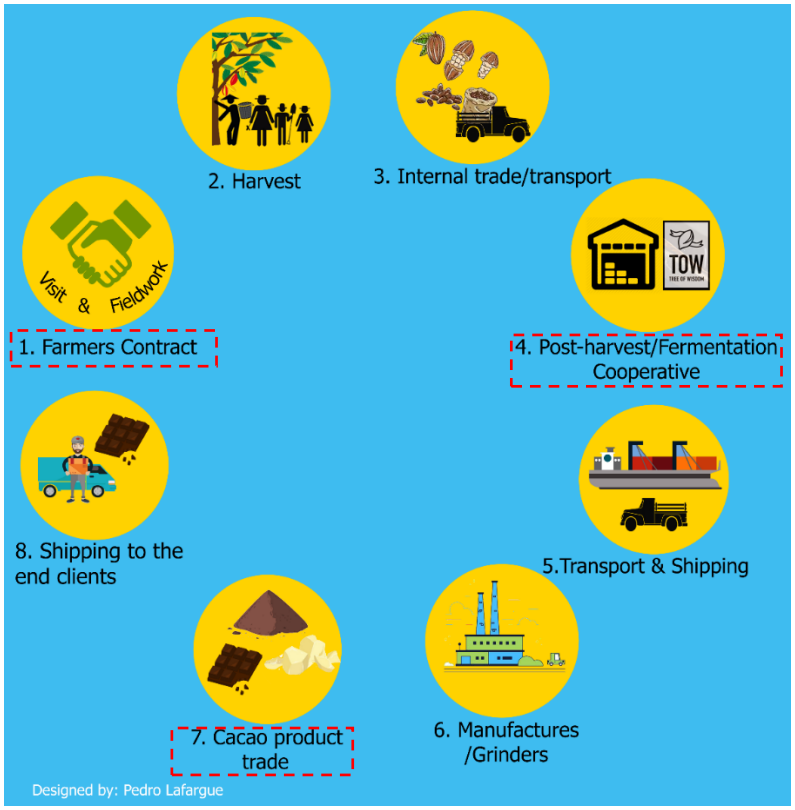


Figure 2.6 Identification of stages that can be improved by implementing biomarker controls in the current cacao/chocolate supply chain system

Feedback from the stakeholders reinforces the hypothesis that the development of biomarkers that could target DNA composition at the farm while doing the Technical Visit/ Fieldwork (1) and a biomarker that could target Post-harvest/Fermentation Cooperatives (4) would be useful controls. This would allow verification at stage 7 by obtaining DNA from the final product (Beans, Nibs, Chocolate, Powder or Butter).

Chapter 3. Chocolate and Beans total DNA extraction

3.1 Introduction

DNA extraction for food authentication has been performed in a range of raw and minimally processed products including wine, honey, olive oil, meat and dairy products (Drummond *et al.*, 2013; Gryson, Messens and Dewettinck, 2004). This, if successful, enables the amplification via Polymerase Chain reaction (PCR) of mitochondrial, nuclear or chloroplast markers that can be used to identify the presence of species of interest in a raw product.

When studying plant-derived products, a standard protocol is generally created which includes a control set of DNA extractions from reference fresh tissue plant material (leaf, seeds) in addition to the raw material that should appear in the mixed and processed products. This helps to identify the presence of DNA from the plant through physical transformations from one stage (raw) to the next one (final product) (Metzger, 2003; Bieber *et al.*, 2016). DNA extraction protocols are well established for use on reference plant tissues such as leaves but can be more challenging for processed material as DNA quality and quantity can be more affected (Özgen Arun, Yilmaz and Muratoĝlu, 2013).

Processed food has been categorized according to the extent and purpose of food processing, rather than in terms of nutrients using (NOVA) classification (Monteiro *et al.*, 2010). Processed food that goes through multiple procedures of high temperatures, physical-chemical transformations and incorporation of various ingredients such as casein, flavours and processing aids such as de-foaming, salts, firming, bulking, anti-caking, glazing agents, emulsifiers, sequestrants and humectants are categorized as ultra-processed food products with chocolate being one of them. These processes lead to DNA degradation. In addition, incorporated ingredients during the chocolate manufacturing process may include inhibitors and secondary metabolites that will be co-extracted with DNA and can bind directly either

with single or double-stranded DNA or sequestrate the magnesium co-factor ion (Schiefenhövel and Rehbein, 2013) leading to failure in PCR amplification (Lo and Shaw, 2018). Indeed, previous studies looking at DNA extraction from chocolate have confirmed that the key challenge has been to extract good yield of high quality DNA (Viet Ha *et al.*, 2015) with the presence of high levels of the oxidized form of polyphenols revealed following extraction. These can bind covalently to proteins and nucleic acids and prevent the DNA to be used for molecular testing (Ha *et al.*, 2015a) with (Moreira and Oliveira, 2011; Rosman *et al.*, 2016).

Comparative studies on DNA extraction methods from cacao sub-products were conducted by Viet Ha *et al.* (2015) using well established protocols. The main approach was performed with the Cetyl trimethylammonium bromide (CTAB) method involving a long extraction protocol with chloroform which is more efficient in generating a high yield of DNA but does add more inhibitors to the end product (Rosman *et al.*, 2016; He *et al.*, 2007). Depending on the cacao sub-product and component used for extraction, amplification of plant-specific DNA from cacao butter (CB) and dark chocolate can be difficult or sometimes impossible (Ha *et al.*, 2015b). In most cases, studies in chocolate and cacao derivatives have focused on identifying animal fat or vegetable fats from Genetically Modified crops in mixed products but no genetic studies have been conducted that have identified CB presence and content in food fraud or provenance adulteration (Che Man *et al.*, 2005; Rosman *et al.*, 2016).

In premium chocolates and confectionary, CB is considered one of the most essential components of cacao that determines the chocolate texture and melting behaviour (Afoakwa, Paterson and Fowler, 2008; Beg *et al.*, 2017). CB can be produced by physical pressing of liquor (100% CB pure prime pressed -premium quality), expeller pressing (poorer quality) and by solvent extraction of residues from pressing which is the lowest quality. Due to the importance of this product, authenticity issues emerge in the market of premium speciality fats where different mixes of vegetables fat, non-cacao fats and lower grades of CB from low-quality cacao nibs/liquor/residues are mixed (Jahurul *et al.*, 2013). For instance, due to the high price of CB, the chocolate industry has utilised a range of oleic substitutes originating

from soy, palm, related species to *T. cacao* (*Theobroma grandiflorum*) and even animal fat (lard) or sugars like glycerine to mimic the texture produced from cacao butter (de Oliveira and Genovese, 2013; Azir *et al.*, 2017).

In chocolate, the final product is a mix of four main cacao derivatives, cacao mass, cacao powder, cacao butter and cacao liquor. As Rosman *et al.* (2016) indicated, it is important to identify the DNA purity and yield from each product to interpret appropriately the molecular analysis of the final product. For instance, CB production goes through various chemical and physical transformations that involve high pH which may cause DNA hydrolysis and lead to a limited amount of DNA template of sufficient length for PCR application. The remaining three components are likely to generate a higher yield of DNA but might contain also higher levels of inhibitors. If the PCR amplification fails, the ingredient with the higher levels of inhibitor should be the one which is inhibiting the PCR amplification. Conversely, if the PCR is successful, results might represent mainly the products producing the highest yield of DNA. Viet Ha (2015) compared different DNA extraction approaches with the CTAB-SDS (sodium dodecyl sulphate) shown to be the most efficient on cacao mass but the plant DNA extraction kits exhibiting the lowest yield and no PCR amplification. Furthermore, none of these methods generated reliable DNA yield from cacao butter.

Food extraction protocols have now moved toward silica-based membrane for the isolation and purifications of degraded DNA. These still use chloroform extraction present in CTAB-SDS approaches but employ also a membrane to trap the DNA instead of using ethanol/isopropanol precipitation. Smulders *et al.* (2010), showed that the DNA extraction from chocolate bars using the QIAamp DNA Stool kit needed further improvement. Smulders reached the same conclusion when using the DneasyTM Mericon food DNA extraction kit as it also generated low yield and low quality DNA.

This chapter has three aims:

Aim 1: To assess if silica-based column DNA extraction protocol could be improved to increase DNA yields and quality of all cacao-sub products and dark chocolate samples.

Aim 2: To assess if DNA yields are adequate as a template to perform accurate PCR amplification. DNA should be specifically suitable as a template for the screening of *T. cacao* chloroplast markers and all bacteria commonly involved in the fermentation process of cacao beans. These markers will be utilised in the later part of the research to assess geographical traceability of the processed material and the traceability of cacao in its supply chain.

Aim 3: To assess the differences obtained in term of extraction both qualitatively and quantitatively when comparing cacao butter and cacao solid. This will inform the "noise level" for DNA markers expected to be observed from the cacao butter used in a sample.

3.2 Materials and Methods

3.2.1 Cacao and derived products

A total of 54 single origin chocolate samples and 3 samples combined from 3 single origins (TOW 8, 9, 14) were utilised in this section to assess DNA extraction from cacao products (Appendix III). These produced in 2013, 2015, 2017 and 2018 included samples of dark chocolate ranging in content from 70% to 100% cacao-solids but also more heavily processed samples with lower cacao-solids ranging from 0% to 30%. Ten non-commercial samples were supplied by Tree of Wisdom Chocolate Limited (TOW; Ecuador-UK) from three regions of Ecuador including a farm in Manabí, a cooperative in Guayas and a cooperative in Esmeraldas. Samples originating from the Guayas cooperative included chips (TOW4-100% cacao-solids), nibs (TOW7-100% cacao-solids) and 100% cacao butter pure prime pressed (TOW13-0% cacao-solids). Forty non-commercial single-origin award-winning chocolate bars from Cacao of Excellence originating from thirty-two fermentation location of fourteen countries were supplied by Bioversity International. Two samples mixed with dry fruit/nuts and milk: INSP Coverture Fraise 1" and "INSP Coverture Amande 1, from Valrhona and five commercial samples from Colombia, Peru and Ecuador from Cacao Hunters, QRN

and CX respectively were analysed. In addition to these, the study included 10 DNA reference samples previously extracted in 2013 by Dr Joel Allainguillaume, from commercial chocolate samples using the standard extraction protocol from QIAGEN Mericon kit were included for yield comparison (Appendix IV). Samples were stored in sealed packaging at room temperature until DNA extraction was performed.

3.2.2 DNA extraction methods

Dneasy™ Mericon Food Kit (QIAGEN, UK) was used as per manufacturer's instructions following the standard protocol to establish standard DNA extraction yield. The small fragment protocol suited to more degraded DNA was then applied and modified to assess extraction steps that might lead to higher DNA yield. For all extractions, two grams of cacao sample were homogenised. Cacao nibs were warmed at 65°C for 20 minutes, then disrupted and grounded using a pestle and mortar to produce a soft mass. Chocolate and cacao butter samples were incubated at 65°C until melted, twenty minutes before starting the DNA extraction procedure. Homogenised samples were mixed in 50 mL falcon tubes with 10 mL of food lysis buffer and 25 µL of Proteinase K. To improve homogenization and allowing for processing of multiple samples, the incubation was carried out in a shaking incubator (Innova 4230, New Brunswick Scientific, UK) at 65°C for 30 minutes at a speed of 400 rpm. Samples were then cooled down on ice to room temperature (15 – 25°C) and centrifuged for 5 minutes at 2500-x g (MST Mistral 2000).

Three main factors within the small fragments protocol were modified to assess improvement in DNA yield return. The experiments corresponding to the protocols including the combination of these modifications are described in (Table 3.1).

1. Food Lysis Buffer incubation time was increased from 30 min to 12 hours.
2. PB binding buffer volume was increased from 350 µL to 500 µL.
3. Elution Buffer (EB) incubation time on the silica membrane was increased from 5 minutes to 6 hours at 65°C in a shaking incubator at a speed of 100 rpm.

Table 3.1 Protocol amendment descriptions

Protocol code	Amendment
1	Standard 1 as per manufacturer instructions
2	Protocol 1 with homogenized sample incubating for 12 hours in lysis buffer. (Discarded)
3	Small DNA fragment protocol warm elution buffer added to the silica membrane
4	Protocol 3 + 150 µL PB added and warm elution buffer added for 5 minutes of incubation before centrifugation.
4A	Replicate of protocol 4
5	Protocol 3 + 150 µL PB added and warm elution buffer added for 6 hours of incubation before centrifugation.
5A	Replicate of protocol 5
6	Small DNA fragment protocol with 12 hours lysis buffer added for 12 hours and warm elution buffer added for 5 minutes incubation.
7	Protocol 3, added with 12 hours lysis buffer for 12 hours and warm elution buffer added for 6 hours incubation.

3.2.3 DNA quantification

DNA quantification of all DNA stocks was determined initially with a spectrophotometer at the absorbance (A260 nm). The purity of DNA extraction was assessed by using the ratio of the absorbance at 260 nm, 230 nm and 280 nm measured in a NanoDrop 100 (Thermo Scientific, UK). DNA samples were also quantified and compared to the NanoDrop 100 measurement by using fluorimetric (Qubit™) procedures. The kit Qubit™™ 1X dsDNA high sensitivity (HS) chemistry for small fragments reactions was used to perform quantification assay as per manufacturer instruction. Negative controls for both approaches were done with the DNA elution buffer (EB) through the whole process.

3.2.3.1 NanoDrop 1000 quantification

The measurement pedestal was cleaned twice with 2 µL of sterile distilled water and lint-free tissue. Before measuring any DNA samples, the equipment was blanked with 2 µL of Qiagen elution buffer used to elute the DNA from the column. The

measurement pedestal was then cleaned and 2 µL of DNA solution was added and measured at 260nm for double-stranded DNA. Measurements were done twice per samples with the measurement pedestal cleaned between each measurement.

3.2.3.2 Qubit™ Fluorometer

Concentrated assay reagent, dilution buffer, and pre-diluted DNA standards kept at 4°C were held at room temperature (22–28°C) for 30 min before being used to measure the DNA samples. Prior DNA measurement of the samples, a fresh standard was prepared with 2 µL of provided standard and 198 µL of buffer thoroughly pipetted and vortexed. Since the standards provided for the measurement last for a maximum of 3 hours before losing their fluorescence, all DNA measurements where undertaken in batches of no more than 40 per session. The measurement of the samples followed the manufacturer protocol with 9 µL of reagent combined with 1 µL of the extracted DNA and this performed in duplicates.

3.2.4 Statistical analysis of DNA yields

A student T-test (Mann-Whitney test) was performed on the raw yield data to compare each extraction assay value with the mean of all values for that group. A correlation analysis of the A260/280, A260/230 ratios of contaminations versus the raw yield data was performed to assess the significance between the extraction protocol yield and NanoDrop measurement. The correlation between cacao solids percentage present in the sample and DNA yield was evaluated with a two way ANOVA followed by Dunnett's multiple comparisons test. This was performed on Qubit™ and NanoDrop data to determine the significance between yield concentration measurements. GraphPad Prism version 8.0.0 for Windows statistical package was used for the analysis (GraphPad Software, San Diego, California USA, www.graphpad.com).

3.2.5 PCR assays DNA quality assessment

DNA quality was assessed using the selected PCR assays used throughout the project including chloroplast microsatellites (cpSSR) and 16s ribosomal bacterial marker. Following PCR amplification, quality assessment was conducted from the final output generated for each specific marker. Fluorescent capillary sequencing

analysis was performed on chloroplast microsatellite (cpSSR) and Illumina amplicon sequencing on 16s ribosomal microbial markers.

3.2.5.1 Chloroplast microsatellites

PCR assays targeting three chloroplast cpSSR loci from *T. cacao* were screened on duplicate PCR reactions from 57 chocolate DNA extractions including a cacao butter sample TOW 13 (Table 0.7). These were selected to generate a range of product size that can be used to assess the effect of amplicon size on successful PCR amplification from degraded DNA. cpSSR3 (209-215 bp product size), cpSSR4 (146-151 bp product size) and cpSSR14 (170-173 bp product size) were all amplified and labelled with the fluorochrome HEX and submitted for analysis at Aberystwyth University (see chapter 4 for details of loci, PCR protocol). Samples were analysed using GeneMarker with the total raw fluorescent values obtained from the sum of fluorescence at all allelic peak for each specific microsatellite locus. The average values across all PCR generated from chocolate templates was compared per locus to the cacao butter sample PCR output.

3.2.5.2 16s v3-v4 ribosomal regions

The universal primer pair v3-v4 was used for amplicon Illumina screening on samples originating from the Guayas cooperative included chips (TOW4-100% cacao-solids), nibs (TOW7-100% cacao-solids) and 100% CB pure prime pressed (TOW13-0% cacao-solids) following the procedure described in section 5.2.4 from Chapter 5. Illumina amplicon sequence analysis was performed on TOW 4, 7 and 13 to further compare the quality and quantity of DNA isolated from cacao butter. Sequences generated from these samples were analysed following the procedure described in Chapter 5, section 5.3.2 with comparison made between the three samples generated in terms of quality and quantity of sequences.

3.3 Result

3.3.1 Assessing DNA quantity using NanoDrop/Qubit™

3.3.1.1 Protocol reproducibility in the extraction of DNA from the same chocolate samples

The accuracy of the extractions measured by NanoDrop™ spectrophotometer were assessed by comparing a range of duplicated DNA samples extracted on two separate dates. DNA samples were extracted twice from 2g of the nine TOW chocolate samples using the standard protocol. DNA concentrations were measured in triplicates and an average value recorded. The DNA concentration analysis following a student T, Mann-Whitney test confirmed that overall no significant differences were observed between replicate measurements, but a high variance was seen between the two extraction replicates across all samples. For example, TOW1 (chocolate 70%) first extraction was 3.4 ng while the second one yielded 23.4 ng, the same effect was seen in sample TOW 7 (cacao nibs) which yielded from 0.9 to 43 ng. This was not observed for TOW 13 Cacao butter which yielded 0.9 ng and 0.8 ng. Since chocolate samples are fairly homogeneous in their content, separate extraction from the same sample should generate very similar DNA yield but that was not observed. The variable values measured by Nanodrop, exemplified by some high values from the outliers TOW 1 (23 ng/ µL), TOW 7 (43 ng/ µL) represented with stars in the error bars (Figure 3.1) gave some indication that the amount of inhibitors in the samples might cause apparent variation in DNA yield.

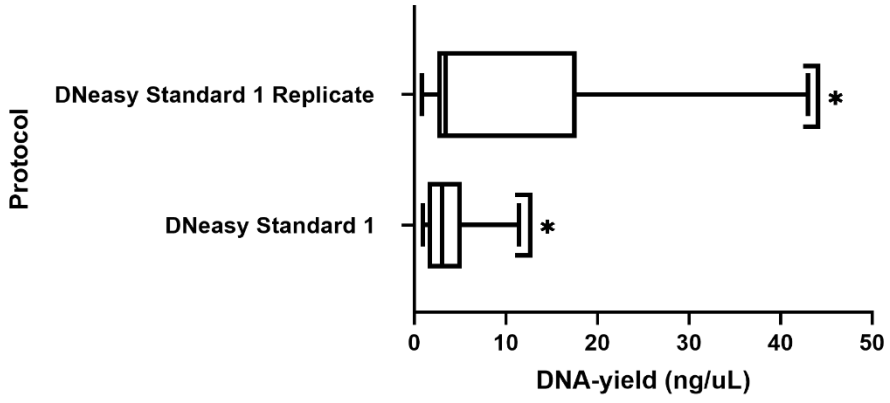


Figure 3.1 Comparing DNA concentration of seven TOW chocolate samples measured by NanoDrop from two extractions performed in a different day

Using the DneasyTM standard (Recommended for process food as per manufacturer instructions) protocol (y-axis) and showing the DNA yield concentration (x-axis) in ng/uL which indicate two outliers, one in each extraction day. In DNeasy standard 1 the outlier TOW 9 is a mix of chocolate with added short branch amino acids and in the case of the replicate

the outlier is TOW 1 chocolate 70%. Mann-Whitney test showed the Median of protocol 1= 3, Median of protocol 1 replicate = 3.4, U test= 29; the replicate has 0.4 higher yields but is not significantly different. *P* = 0.3281.

3.3.2 NanoDrop assessment of DNA impurity levels

Analysis by NanoDrop provided two purity ratios for all samples extracted with the same standard protocol (Table 3.1). These measurements indicated overall a high level of contamination with all A260/280 measurements ranging from 1.19 to 1.63 and lower than 1.8 (clean DNA value) and all A260/230 measurements were four times lower than 2 (clean DNA value) ranging from 0.39 to 0.63. These might be indicative of leftover protein and phenol contaminants remaining from the extraction method but also are probably inherent to the template extracted. When comparing all extractions there was not a significant correlation between the DNA yields and the contamination ratios A260/280 ($\boldsymbol{R}^2$ = 0.1691, *P* < 0.11) and A260/230 *(*$\boldsymbol{R}^2$ = 0.01943, *P* = 0.60) (Figure 3.2).

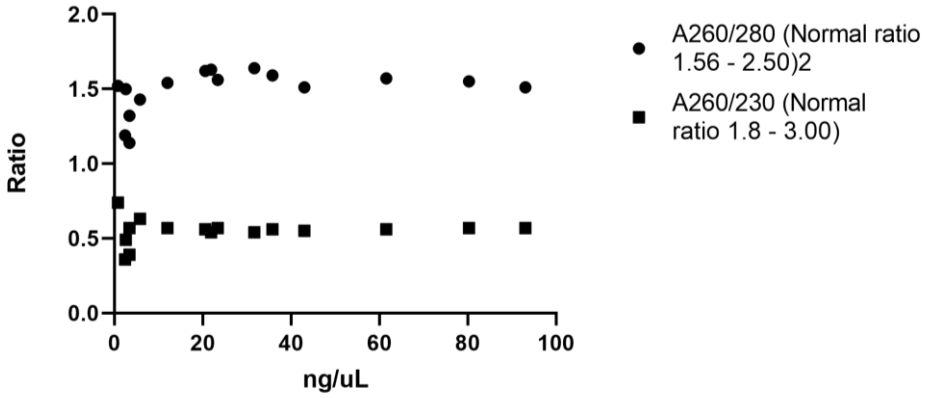


Figure 3.2 Comparison of DNA extraction yield in ng/µL vs Nanodrop impurity ratios

Purity rations were measured with absorbance at A260/280 ($\boldsymbol{R}^2$ = 0.1691, *P*< 0.11) represented with the circle symbol and A260/230 ($\boldsymbol{R}^2$ = 0.01943, *P*< 0.60) with a square symbol showing that none of these ratios is within normal values. It Similar levels of DNA contamination from the inhibitors can be observed.

3.3.3 Comparison of NanoDrop and Qubit™ DNA measurement

The small fragment protocol was used to extract 60 chocolate samples (50 from the current research and 10 as controls from previously extracted DNA). These were diverse in their composition including 20% to 100% of cacao solids and for some samples included milk and other additives. Initial measurements of total DNA extractions (average of triplicate measurement) using the NanoDrop showed unequal yields that were above 30 ng/µL up to 40 ng/µL (Figure 3.3). "INSP Coverture Fraise 1" and "INSP Coverture Amande 1" composition was different from all other samples with the inclusion lyophilized pulverised fruit, milk and nuts as ingredients. These ingredients would have endured a much lower level of processing than the cacao beans and are likely to have contributed to the majority of the DNA yield.

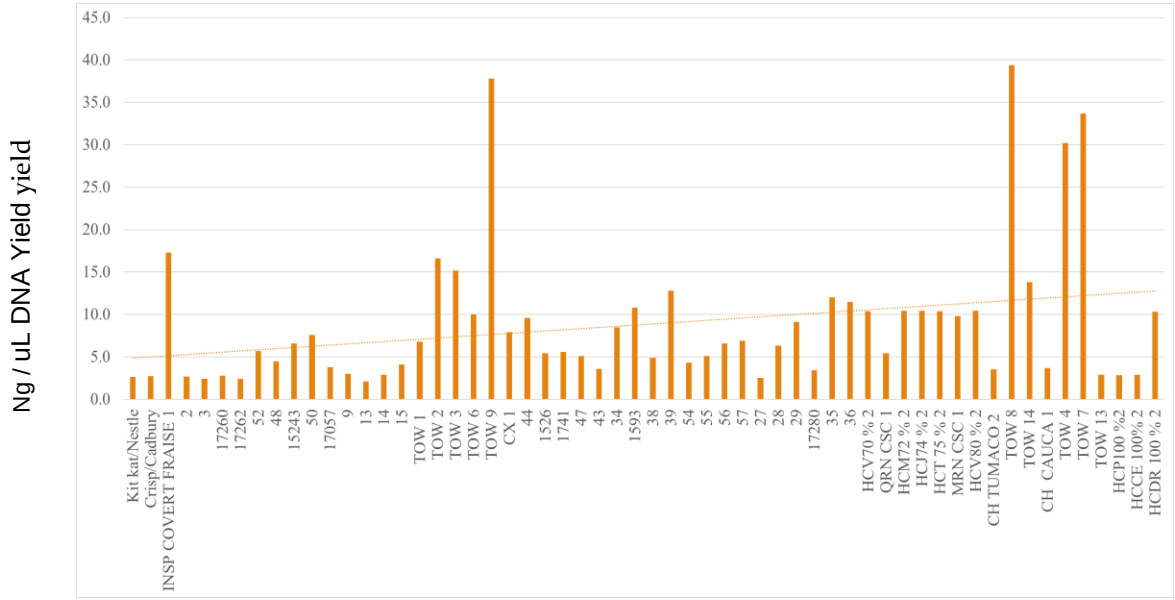


Figure 3.3 Total DNA extraction measurements by NanoDrop following protocol 3

Comparison of 57 cacao-derived/chocolate samples listed on the x-axis (excluding cacao butter) and DNA yield (y-axis) in ng/µL with chocolate samples ordered according to cacao solid content increasing from left to right. All samples were analysed by one-sample t-test (R^2 = 53% P< 0.0001) showing significant variation between them and the trend line shows yield increment while the cacao percentage increase in the sample. Sample TOW 8 shows a yield of 39.4 ng/µL being the highest concentration with protocol 3.

TOW 9 (Chocolate 70%), TOW 4 (Liquor), TOW 7 (Nibs) and TOW 8 (Mix of nibs and chocolate 70%) all showed higher yields 37.8, 30.2 33.7 and 39.4 ng/µL respectively. The remaining samples which include 62% to 100% of cacao solids all

exhibited lower yield when measured with the NanoDrop ranging from 2.1 to 18.9 ng/µL. To assess the accuracy of the Nanodrop measurement, all 60 samples were analysed by the high sensitivity Qubit™ Fluorometer. The DNA yield measured were much lower than those observed with the Nanodrop with all DNA concentration measured below 6 ng/µL. TOW13 (cacao butter 100%) which contains no cacao solid showed the lowest yield while the "INSP coverture fraise 1" and "INSP couverture Amande 1" (CB + Milk + lyophilized pulverised fruit/nuts) again generated the highest levels of total DNA. However, all other chocolate samples measurement were fairly homogeneous irrespectively of their cacao content ranging from 0.021 ng/µL to 6.75 ng/µL (Figure 3.4).

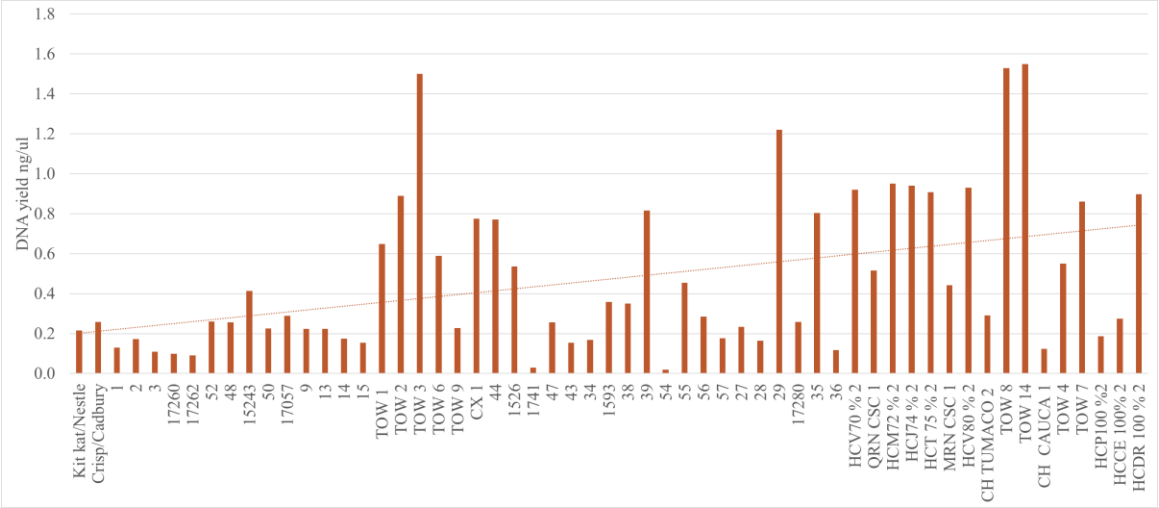


Figure 3.4 Total DNA extraction measurements by Qubit™ following protocol 3

Comparison of 57 cacao-derived/chocolate samples listed on the x-axis (excluding cacao butter) and DNA yield (y-axis) in ng/µL with chocolate samples ordered according to cacao solid content increasing from left to right. All samples were analysed by one-sample t-test (R^2 = 33% P< 0.0001) showing significant variation between them and the trend line shows yield increment when cacao percentage increase in the sample. Sample TOW 14 shows a yield of 1.54 ng/µL being the highest concentration with protocol 3.

The comparison of DNA measurements for the same samples using NanoDrop and Qubit™ revealed that the outliers observed with the NanoDrop did not necessarily correspond to a true larger DNA yield and these large variations were not observed when the measurement was done with Qubit™. TOW 9 measured by NanoDrop showed values from 11.4 to 37.83 ng/µL, in contrast, Qubit™ range from 0.22 to 0.27 ng/µL. Similarly, DNA from a chocolate 70% (TOW 6) measured by

Nanodrop showed values from 3.1 to 64 .4 ng/µL with values by Qubit™ ranging from 0.472 to 4.99 ng/µL (Table 3.2).

Table 3.2 Comparison of NanoDrop and Qubit™ measurements from protocol 3 extractions in relation to the cacao solid percentage and type of product

DNA concentration measured by Nanodrop and Qubit™ on a selected sample representative of a range of product from the first stage of production including cacao nibs, cacao butter, chocolate and covertures which are a mix of cacao butter and other ingredients. Each sample was classified according to its cacao solid percentage (n=8 biological samples, measured in triplicates).

Sample Code	**Type of product**	**Cacao solids**	**NanoDrop** ng/µL	**Qubit™** ng/µL
TOW 3	Chocolate	70%	15.2	0.97
TOW 4	Chocolate	100%	30.2	0.682
TOW 6	Chocolate	70%	10	0.472
TOW 13	Cacao Butter	0%	2.9	0.0216
TOW 9	Chocolate	70%	37.8	0.274
TOW 7	Nibs	100%	43	1.03
Insp covert Fraise 1	Covertures	37.9%	17.3	3.34
Insp covert Amande 1	Coverture	30.6%	18.9	5.54

The analysis of student T, Mann-Whitney test indicated a significant variance between the yields measured by NanoDrop and Qubit™. The variation and high yield observed from the Nanodrop might be caused by contamination which generated an overestimate of DNA concentration in the samples (Figure 3.5).

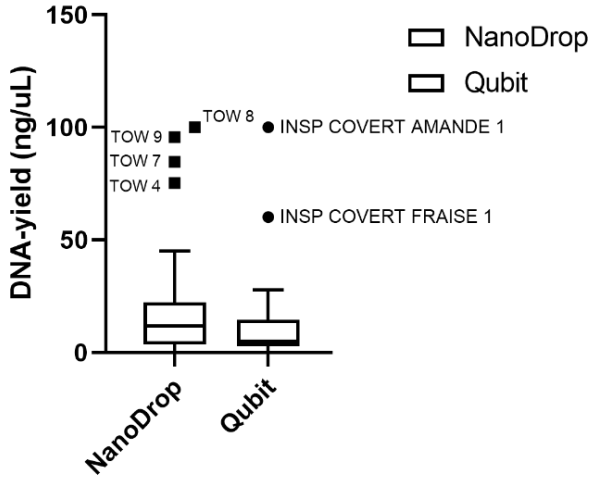


Figure 3.5 Comparison of DNA extraction yield in ng/µL performed with protocol 3 between Nanodrop vs Qubit™ fluorometer

The comparison of both measurements is showing the outliers from the 60 samples (Appendix IV: Chocolate samples DNA quantification
Table 0.7). Normalised data including outliers (TOW 4,7,8,9 represented with a square symbol and INSP Coverture Amande 1/Fraise1 represented with a circle symbol) assessed with Mann-Whitney test, showed NanoDrop median standard Mdn = 18.19 and Qubit™ Mdn = 10.49, and significant difference when comparing the two methods of quantification $P<$ 0.0043. R^2=0.03993, *F, DFn, Dfd* = 2.129, 59, 59.

While there is a direct correlation observed in the average measurement of the 60 DNA samples (Appendix IV) between Nanodrop and Qubit™ ($P<$ 0.0008), many samples also exhibited strong contrasting differences in DNA measurements. To validate this, an unpaired t-test for Nanodrop and Qubit™ was performed and no significant variance was observed related to sample effect. The Nanodrop median was 6.62 ng/µL, with a min = 2.1 ng/µL and a max = 39 ng/µL. The former high values were considered to be erroneous observations as they varied by replicate and were too far apart from the median. For Qubit™ measurement, the median was 1.48 ng/µL, with a min = 0.128 ng/µL and a max = 3.71 ng/µL. There is a significant difference between Nanodrop and Qubit™ measurements $P<$0.0001, where Qubit™ shows more consistency and less variability between replicate samples as per variation between samples. This methodology was therefore used to assess the DNA extraction protocol modification aiming at improving yield. Qubit™ was also used for accurate assessment of the DNA content of all samples preliminary to all PCR generated for capillary analysis and Illumina sequencing in the present study.

3.3.4 Extraction protocol comparison Qubit™

Four chocolate samples derived from the same batch of beans were used to compare the modifications implemented on the small fragment DNA extraction protocol. These included samples with 70% cacao solids (TOW 6), 100% cacao solids (TOW 4) Chocolate/Liquor, 100% cacao solids as nibs (TOW 7) and 100% cacao butter pure prime pressed (TOW 13). To assess the efficiency of the protocols, a two way ANOVA was performed for the samples and sample composition vs protocols. The protocol comparisons were extremely significantly different ($P< 0.007$) and account for 52.64% of the variations between the yields while there was also a significant effect between chocolate samples ($P<0.0045$) with a total variation of 24.01%.

All samples registered an increment in yield while using the protocol amendment Extra PB/EB 6 hours (5 and 5A). This protocol had two modifications which included the PB binding buffer volume increased from 350 µL to 500 µL and the Elution Buffer (EB) incubation time on the silica membrane increased from 5 minutes to 6 hours at 65°C in a shaking incubator at a speed of 100 rpm. This protocol increased the DNA yield in most of the samples, such as in (TOW 4) from 0.55 to 6.75 ng/µL, (TOW 6) from 0, 59 to 4.99 ng/µL, (TOW 7) from 0.862 to 4.19 ng/µL and cacao butter (TOW 13) from 0.0106 obtained with protocol 3 to 0.27 ng/µL obtained with protocol 5A (Figure 3.6). Importantly, with the standard protocol failing to generate any DNA yield when applied to cacao butter, the amended protocol did work with 0.27 ng/µL DNA yield which could be potentially used successfully as the template in PCR. Qubit™ quantification indicated that the small fragment protocol adaptations and the use of extra binding buffer improved consistency in extractions.

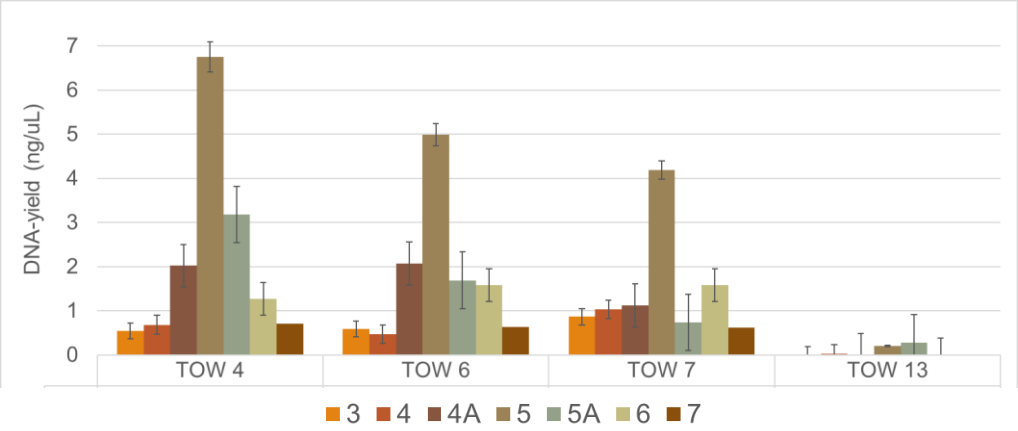


Figure 3.6 DNA yield comparison from four cocoa samples extracted with 5 protocols

DNA extracted from TOW 4, 6, 7, 13 using 5 protocol extraction 3, 4 (replicate 4A), 5 (replicate 5A), 6 and 7 were measured by Qubit™. The value corresponds to the average between the triplicates measurements of ng/uL with error bars represented as standard error (SE). Highest yield observed in all samples using protocol 5.

All extraction protocols were compared using a two ways ANOVA test. The use of each protocol depending on the type of samples is very significant ($P< 0.0045$, $F = 6.168$, $DFn = 3$, $DFd = 18$). This is striking when looking at the increase in DNA yield obtained from cacao butter template when using protocol 5.

3.3.5 Effect of cacao composition on DNA yield

Chocolate products vary in their content and this will likely influence the quality and yield of the DNA obtained. In the present work centrifugation of different types of chocolate and cacao derived products following lysis showed three distinct phases of particle separation corresponding to a supernatant containing lipids and low-density particles, an aqueous intermediate solution containing the DNA and a pellet formed of debris, contaminants, polyphenols and leftover cacao solids. Visual differences were observed between the sizes of these layers depending on the cacao percentage of the sample (Figure 3.7). In the case of chocolate 100%, the formulation has 48 % of cacao butter and 52 % of cacao solids, while chocolate 70% has 70% of cacao liquor (composed of 33% cacao butter and 37% cacao solids), 28% of refined sugar and 2% of sunflower lecithin. While these two types of chocolate showed three different phases and amounts of debris, supernatant and low-density particles, the 100% cacao butter sample separated in two phases showing a supernatant of lipids rich in cacao butterfat and a bottom aqueous clear phase containing the DNA (Figure 3.7).

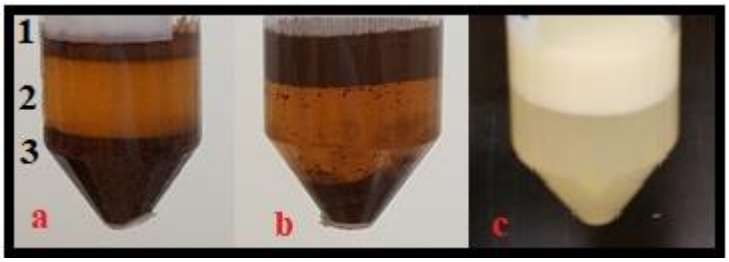


Figure 3.7 Chocolate 100% lysate in 50 ml falcon tube following centrifugation for 5 minutes at 2500-x g

The three different sections generated following centrifugation correspond to 1) supernatant containing fat, lipid and low-density particles; 2) intermediate aqueous phase with DNA; 3) Precipitate of debris, contaminants, polyphenols and leftover cacao solids. a) Chocolate 100%; b) Chocolate 70%; c) cacao butter.

The analysis of TOW13 (100% cacao butter) showed the lowest yield while "INSP coverture fraise 1" and "INSP coverture Amande 1" samples which contain lyophilized pulverised fruit and nuts exhibited the highest levels of total DNA. These samples clearly differ from all other samples analysed in the study with an unknown proportion of DNA likely to be derived from added fruits and nuts (INSP samples) or as a residual of no cacao solid being present (cacao butter). To assess the correlation between cacao solid percentage and DNA yield, they were therefore excluded from the analysis. The majority of the samples analysed in this study are from chocolate containing 70% cacao solid, therefore a comparison between chocolates below and above this value was performed.

A correlation test and t-test were performed to compare the relationship and significance between cacao solids and the optimised protocol. The analysis indicated that this protocol has a positive correlation *(P*< 0.0156) and significance between cacao solid percentages and yield. It is significantly different (*P*< 0.0008) from the small DNA fragment, protocol 3, which again confirms that these amendments improve the DNA extraction and yield.

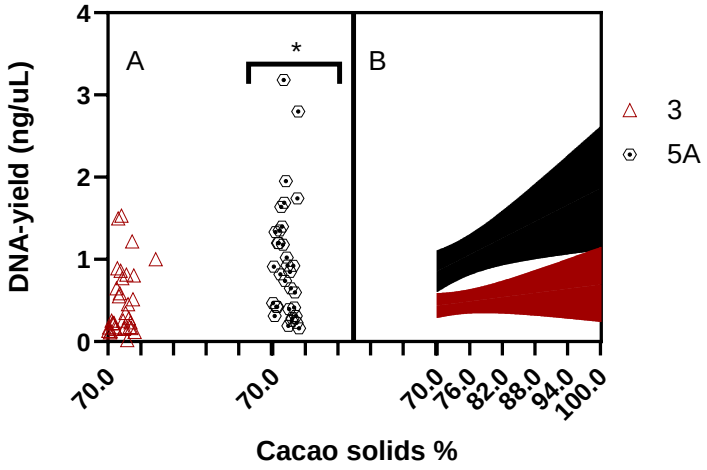


Figure 3.8 Correlation test for 34 chocolate samples, between cacao solids percentage and DNA yield measured by Qubit™; Comparison of protocol 3 and 5A DNA extraction

A: Red triangle and surface represent protocol 3 for small DNA fragment use in the research as control showing lower DNA yield(y-axis) ranging from 0.0193 to 1.53 ng/uL. Black hexagon and surface represent protocol 5A amendment, showing to be the most efficient which is significant correlated $P<0.0156$* with the cacao solids (x-axis) percentages per chocolate sample and also showing higher yields ranging from 0.16 to 3.18 ng/uL. B: The diagram shows that there is an increment in DNA-yield when the cacao solid percentages are higher.

3.3.6 Assessing DNA quality for PCR analysis

3.3.6.1 Amplification of DNA using chloroplast microsatellites

Markers amplified via PCR for capillary analysis are not always detectable via agarose gel. A more accurate approach is to assess the amplification of the product using a fluorescent capillary analysis (Figure 3.9).

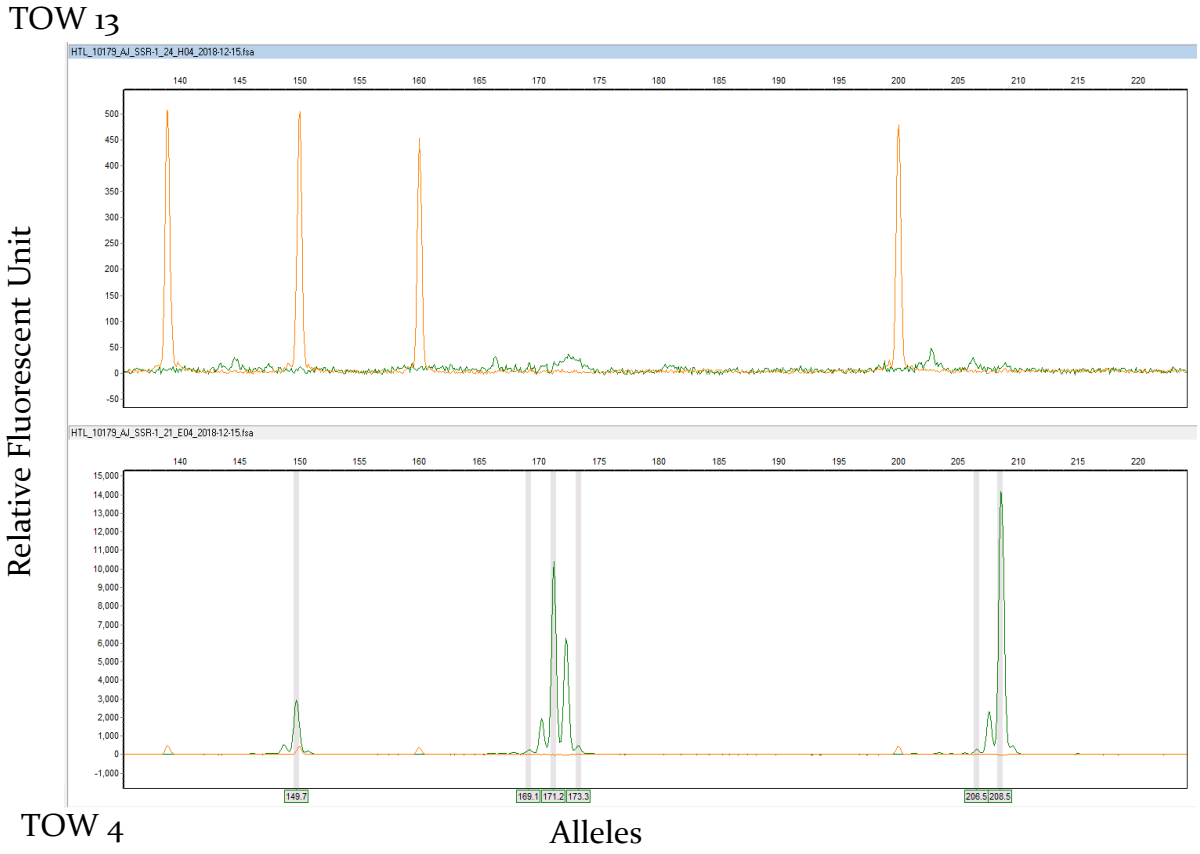


Figure 3.9 Comparison of capillary analysis showing the difference between amplifying cacao butter TOW 13 and chocolate TOW 4

The capillary profile generated from loci cpSSR 3, 4 and 14 on sample TOW 13 and TOW 4 and peak fluorescence intensity visualised and measured on GeneMarkers. The green line corresponds to the amplification peak of the markers and the orange line shows the DNA standard used to assess alleles size. (). TOW13: Cacao butter showing very low fluorescence value lower than 50TOW4: All markers amplified with values of fluorescence ranging from 2000 to 14000.

The total peak height observed for each of the three chloroplast microsatellites cpSSR 3, cpSSR 4 and cpSSR 14 were combined from all fluorescence raw data for the position of all potential alleles as described in Chapter 4 section 4.2.7.2. Fluorescence at each locus was then averaged across all PCR from samples of chocolate (118 reactions) and compared to the duplicate PCR reaction produced on cacao butter DNA. The average fluorescence value for the cacao butter samples across cpSSR4, cpSSR14 and cpSSR3 were 26, 46 and 115 times less strong respectively than all the average value observed for chocolate extract (Figure 3.10). For instance, the average fluorescence value across all alleles within locus cpSSR3 was 8161 in chocolate samples (TOW 4) with a value of 71 observed for cacao butter (TOW 13).

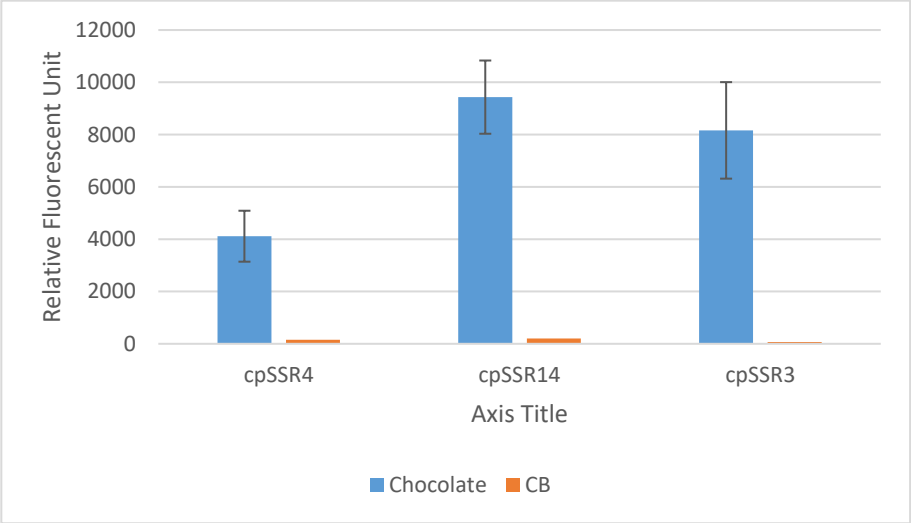


Figure 3.10 Comparison of chocolate and CB DNA as a template for chloroplast microsatellite analysis

The total raw fluorescence values observed in cpSSR 3, cpSSR 4 and cpSSR 14. Average raw fluorescence is indicated on the y-axis with each specific cpSSR listed for both chocolate and CB on the x-axis. Average values were obtained across 118 PCR reaction from 118 chocolate samples and CB values calculated as the average between duplicates. Error bar: 2SE

3.3.6.2 Amplification of DNA using 16s rRNA amplicon Illumina sequencing

Comparison and quality control metrics of 16s amplicon Illumina generated from chocolate and CB

Chocolate samples TOW 4 and 7 were compared to CB TOW 13to assess sample type would affect the quality and sequences count. The three samples shared the same cocoa origin. The preliminary FastQC quality report indicated differences in quality between the three samples.

The average median quality score of Illumina amplicon reads which typically start out lower over the first 5-7 bases, will then rise and steadily drop over the length of the read. With paired-end reads, the average quality scores for the first read will almost always be higher than for replicate reads. This plot provides the distribution of quality scores at each position in the read-across all reads. The quality reports are highlighted with three colours representative of quality with Green for high quality, orange representing the threshold and red which indicate poor quality sequence and failure to amplify the whole sequence. Good quality scores were observed for TOW 7,

with lower quality for TOW 4 and poor quality for TOW13 indicative of the failure to obtain complete sequences of the amplicon (Figure 3.11).

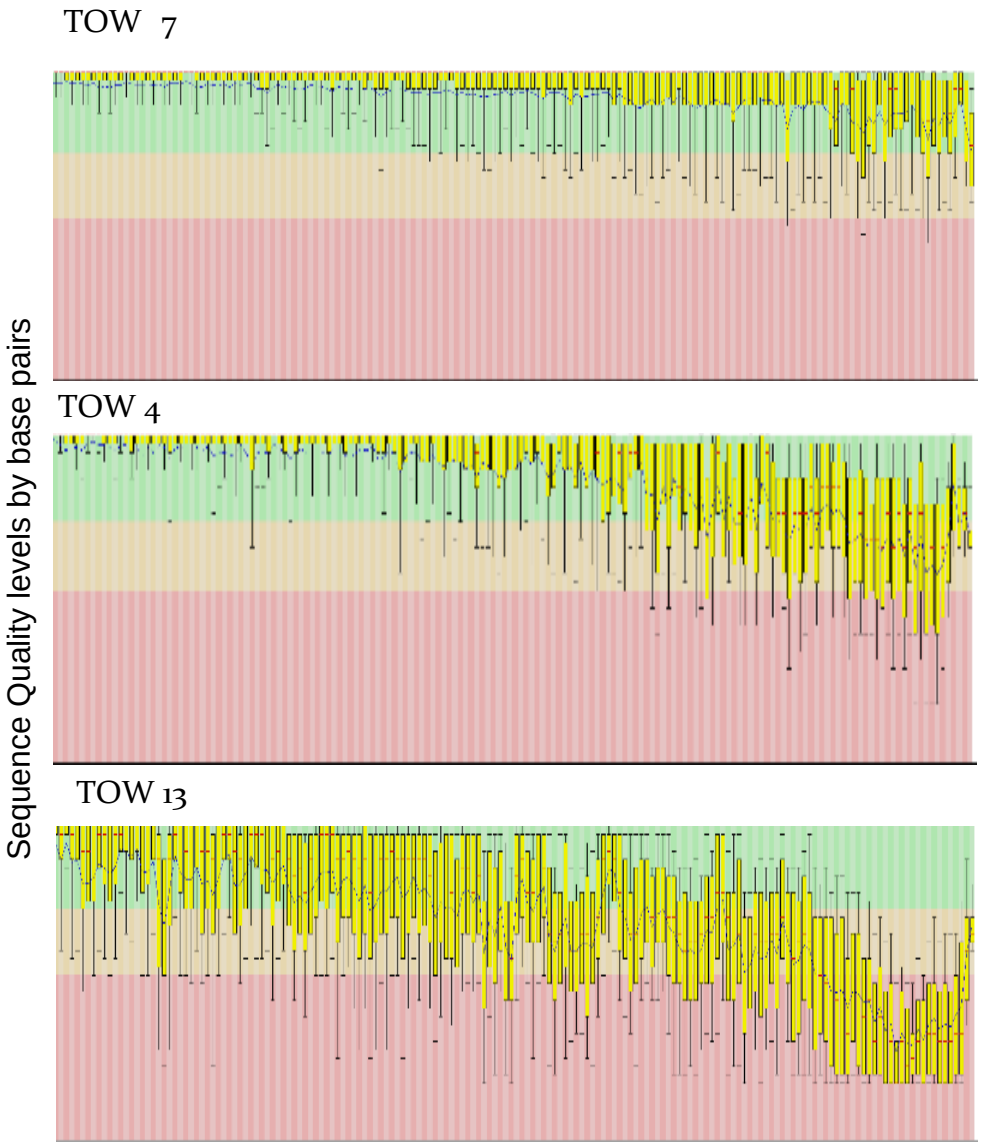


Figure 3.11 FastQC Per base sequence quality report of the Illumina raw sequences for TOW 7, TOW 4 and TOW 13

A box-and-whisker plot showing aggregated quality score statistics at each position along all sequence reads. The y-axis gives the quality scores, while the x-axis represents the position in the read. The colour-coding of the plot denotes what is considered high (Green), medium (Orange) and low (Red) quality scores. Chocolate TOW 7 shows a good quality read for most of the 300bp sequenced. Chocolate TOW 4 presented reduced sequence quality with low quality scored observed from position 250bp. TOW 13 showed the lowest sequence quality indicative of amplification failure with low-quality scores observed at earlier position (150bp) in the reads.

Quality filtering was performed with a combination of DADA2 and QIIME2 scripts to determine the number of sequences that could be truly used for sample

analysis. The number of sequences identified following filtering reflects the FastQC per base sequence quality report (Table 3.3). TOW 7 generated the highest number of usable sequences with 127727 non-chimeric sequences with TOW 4 yielding 78744 sequences. In contrast, TOW 13 only generated 694 sequences.

Table 3.3 Quality sequence filtering of 16S v3-v4 amplicons for TOW 4, TOW 7 and TOW 13 CB

Input samples (raw data) have gone through quality filtering and controls performed with DADA and QIIME2 scripts combinations which denoised and eliminates low quality and short reads.

Sample-id	Input	Filtered	Denoised	Merged	Non-chimeric
TOW 4	134248	79833	79488	78785	78744
TOW 7	204318	128104	127927	127727	127727
TOW 13	1223	738	717	694	694

The identity of the predominant sequences obtained for all samples were compared. The two most abundant sequences found in TOW 4 and TOW 7 were identical and representative of the two main bacteria expected during the fermentation process Seq1 *Acetobacteraceae* and Seq2 *Lactobacillaceae*. In contrast, only the Seq1 *Acetobacteraceae* appear in the top 5 sequences observed in Tow 13 and only in 5th position. While this is indicative that CB is produced from fermented cacao beans, the four top sequences for Tow 13 are rare or absent from TOW 4 and TOW 7. This might indicate that the amplicons have been randomly amplified due to the high degradation of the DNA yielded from CB or corresponded to bacteria present during the CB manufacturing process (Table 3.4).

Table 3.4 Predominant sequences observed for 16S v3-v4 ribosomal regions Illumina sequencing of TOW 4, TOW7 and TOW 13 cacao samples with their taxonomic assignment

The predominant 5 species identified for each sample TOW 4, TOW 7 and TOW 13 are listed in a separate section for each sample. Shared sequences are highlighted in red and bold.

Taxonomy	Seq	TOW 4	TOW 7	TOW 13
TOW 4				
f__Acetobacteraceae	1	2014	218	26
f__Lactobacillaceae	2	1974	130	0
c__Bacilli; o__Bacillales	3	1133	10	0
f__Lactobacillaceae; g__Lactobacillus; s__paraplantarum	4	682	19	0
f__Leuconostocaceae; g__Leuconostoc; s__fallax	5	445	0	0
TOW 7				
f__Acetobacteraceae	1	2014	218	26
f__Lactobacillaceae	2	1974	130	0
f__Rickettsiaceae; g__Rickettsia; s__Rickettsia endosymbiont of Deronectes platynotus	6	0	42	0
f__Oxalobacteraceae	7	101	40	235
f__Pseudomonadaceae; g__Pseudomonas	8	0	31	0
TOW 13				
f__Oxalobacteraceae	9	101	40	235
f__Pseudomonadaceae; g__Pseudomonas	10	4	8	166
c__Actinobacteria; o__Actinomycetales	11	0	0	44
f__Propionibacteriaceae; g__Propionibacterium	12	0	0	36
f__Acetobacteraceae	1	2014	218	26

3.4 Discussion

Attempts to extract DNA efficiently from cacao beans sub-products have proved to be difficult and sometimes unmanageable (Rosman *et al.*, 2016; Ha *et al.*, 2015a). The present study demonstrated that silica-based column DNA extraction protocol could be improved to increase DNA yields and quality of all cacao-sub products and dark chocolate samples.

DNA extractions from clean nibs (no-husk contamination), were grounded to a cacao mass, chocolate and butter were melted resulting in the optimal particle size and homogeneity which showed to be key to obtain a good DNA yield (Moreano, Busch and Engel, 2005). This pre-processing of the samples is important when assessing commercial cacao products as the food ingredients and matrix varies and needs to be homogenised so that DNA is not obtained from one ingredient more than another. Different combinations of ingredients that form complex matrices with various levels of contaminants and different particle sizes may generate obstacles for extracting high-quality DNA. DNA may be trapped in fat globules from the cacao butter or

lecithin emulsifier substitutes (Becket, 2008; Glicerina *et al.,* 2014) or within the silica gel from the DNA extraction column. Therefore, the possibility of improving DNA extraction from cacao beans without husk, nibs, pure and mixed fruits/nuts chocolates and cacao butter was investigated. By using samples of clean nibs derived solely from cotyledons, DNA yields ranging from 0.16 ng/µL – 6.75 ng/µL with an average of 1.83 ng/µL were obtained.

Following the small fragment protocols from Dneasy™ Mericon Food Kit (QIAGEN, UK) designed to be more efficient with processed samples, five amendments of the protocol were made. The extraction includes stages of lysis, protein precipitation, filtering and purification of the sample. The food lysis buffer in this kit uses CTAB extraction method developed by (Dung, 2011) which effectively eliminates polysaccharides and polyphenols by employing the cationic detergent CTAB (hexadecyltrimethylammonium bromide or cetyltrimethylammonium bromide), and the polyphenol binding agent, Polyvinylpyrrolidone by forming an ionic complex (Dneasy™ Mericon Food Kit (QIAGEN, UK). After the centrifugation of the samples, the quantity of aqueous phase and amount of debris (precipitants) was dependent on the starting material. The amount of debris varied depending on the percentage of cacao solids present in the sample and chocolate type, with a clear difference in precipitate observed between 70% and 100% cacao solid chocolate extract and 100% cacao butter (Figure 3.7).

Evaluations of DNA purity and protein content by NanoDrop™ 1000 (Thermo-Scientific) A260/A280 ratio are not a sufficient criterion for complex food samples like the chocolate matrix. The natural cacao inhibitors e.g. polyphenols, polysaccharides, salts and proteins do affect the reads potentially giving erroneous yields. Nonetheless, as previous investigations have shown, the quantification of chocolate DNA by using Nanodrop, an initial standardisation for DNA quality and yield was assessed following that benchmark. Abnormalities in DNA quantification were observed with large apparent differences observed in DNA yield between replicate extractions. DNA yield assessment was therefore measured by Qubit™ which proved to be more reliable and reproducible. By increasing the binding buffer volume and the incubation time of elution buffer over the silica membrane for 6 hours at

65°C, the DNA yield improved in all cases across 60 chocolate samples and up to 345% for some samples.

One of the reasons for this improvement is that these adaptations do change the physicochemical structure of the silica membrane. In most extraction columns, the silica membrane is negatively charged (weakly at pH 5, more strongly at pH 8). Therefore, interactions between the amount of buffer, sample composition and temperatures over the silica membrane can be modulated by electrostatic charge and depend on the solution pH, which dictates the charge on the sample and surface. As the temperature increases in the 65°C incubation stage, the pH of the surface will increase to 8.5 pH values, which then can influence the positive ions of the surface to interact with other components like, amino acids that will bind to the surface and to build new salt bridges with other components. This change in pH and, negatively charged groups and the increment in negative charge of the surface will then result in a greater electrostatic repulsion between DNA and silica surface (Meng, Stievano and Lambert, 2004), with lower ionic strength conditions increase repulsive electrostatic DNA –surface and DNA-DNA interactions (Vandeventer *et al.*, 2013).

The majority of commercial chocolates are a mix of cacao mass, powder and butter, where the cacao powder has been treated with alkali salts to improve colour and texture (Kawash, 2010; Afoakwa, 2016b). This alkali process called Dutching can also increase the difficulty of extracting DNA, induce higher contamination from inhibitors and in turn cause PCR amplification errors or failure. Indeed, the oxidized form of polyphenol can bind covalently to proteins and nucleic acids, which may cause failure in DNA detection using PCR (Perry *et al.*, 1998; Smith *et al.*, 2007; Drummond *et al.*, 2013). Even if traces of the cacao matrix in the sample are inevitable, following the optimized amendment protocol for extraction, it was found that there is no correlation between the high level of contamination detected in the samples (260/280 and 260/230 ratios) and the PCR amplification with most samples being positives. An experiment increasing lysis time was performed to determine if this would help to separate DNA from debris and fat globules, different lysis times and temperatures were also assessed to improve and increase the yield and purity of DNA.

No improvement was observed in DNA yield, but an increase in protein and salt contamination was detected (260/280 and 260/230 ratios).

Samples from dark chocolates and derivatives from TOW have been roasted and processed under lower temperatures (120°C) compared to industrial processing temperatures (173°C). This in principle could have an impact on DNA degradation and yield return and could explain some of the outliers observed when measuring samples by NanoDrop. However, these high values did not correspond specifically to the samples containing high measurements of DNA as seen by Qubit™ and could be interpreted as variability due to contamination. Therefore, there is no evidence of a significant effect of an increase in roasting temperature on DNA return.

The sixty samples were compared and showed that cacao butter premium 100% (TOW 13) yield was optimized to an average yield within the group. This is important as, cacao butter is a key component of all chocolate products, but due to the percentage of fat globules in the structure, DNA can be trapped preventing the lysis process to be effective and making it difficult for the DNA to be extracted similarly to other oil extractions (Busconi *et al.*, 2003). Moreover, using the optimum protocol, as low as 0.002 ng/µL and up to 0.027ng/µL of cacao butter. Two other cacao butter base products "INSP Coverture Fraise 1" and "INSP Coverture Amande 1" (CB + Milk + lyophilized pulverised fruit/nuts) were also analysed and showed the highest level of total DNA across all samples. These samples only contain 30-37% of cacao butter and the origin of the DNA extracted is therefore likely to be derived mainly from the other ingredients.

The correlation analysis and comparison between cacao solids and DNA extraction protocols demonstrated a significant difference ($P< 0.501$) in yield between the products with higher percentages of cacao solids and derivatives containing low cacao solid content (Gryson, Dewettinck and Messens, 2007; Ha *et al.*, 2015a). These former samples from commercial chocolates show more homogeneous yields, probably due to the standard industrial processing steps which eliminate nuclear and chloroplast DNA, when comparing to high variations from single-origin samples that are processed in different conditions and in a more artisanal way.

Having demonstrated that the modified Dneasy™ Mericon Food Kit protocol could generate adequate yield of DNA from a diverse range of cacao products, these were successfully tested as template in PCR. There are different reports comparing DNA extraction kits and assays with nuclear SSR markers on cacao leaves but with poor PCR amplification of cacao products (Ha *et al.*, 2015a; Gryson, Dewettinck and Messens, 2007). The present project focused on multicopy DNA markers. In the study, it was identified by PCR that chloroplast, DNA and bacterial DNA can be detected in cacao-derived products such as chocolate, nibs, cacao and liquor. The results of this study demonstrate that the concentrations of DNA obtained using the modified protocol, enable amplification of chloroplast markers of different size in length and targeted gene amplification by Illumina sequencing.

The analysis of cacao butter provided contrasting results. Sufficient DNA was extracted to enable PCR amplification using v3-v4 16S ribosomal specific primers, but no sufficient amplification was observed for cpSSR analysis. When the amplification products were analysed qualitatively via Illumina sequencing, it revealed a lower number of sequences in the cacao butter when compared to nibs and dark chocolate samples from the same origin. The sequences identified in the cacao butter were also different to some extent to the dark chocolate extract. This is important in terms of DNA tracking and appears to suggest that the addition of cacao butter to a sample, which is routine in chocolate production is not likely to interfere with the DNA analysis as it will provide only low background DNA which is likely to be outcompeted during PCR amplification by the DNA template from cacao solid. Conversely, it is likely to complicate the identification of the origin of cacao butter added to a chocolate sample, using a DNA approach.

3.5 Conclusion

The extraction of DNA from cacao products was improved by amending the standard small fragment protocol from Dneasy™ Mericon Food Kit, this has enabled

to amplify DNA biomarkers from all chocolate samples including in some cases cacao butter. This is the first study which emphasized the importance of standardising the extraction process and has succeeded in generating DNA from CB. While the CB DNA might be more difficult to be used for traceability studies, all other chocolate samples have yielded sufficient DNA that can be utilised for methodologies to develop biomarkers into traceability of fermented cacao beans.

Chapter 4. The use of chloroplast markers for the traceability of chocolate products

4.1 Introduction

Tracking the origin and quality of cacao products

There are several varieties of cacao cultivated around the world and these fall in two categories described as premium cacaos or bulk cacaos. Premium cacao including Arriba, Criollo, Fino and Aroma, are commonly known as fine and flavoured or Criollo and exhibit flavours such as fresh fruits, mature fruits, yellow fruits, floral, herbal, wood notes, nut and caramel notes as well as rich and balanced chocolate bases developed in the fermentation. These are typically produced from beans harvested within a specific geographical location. This is quite a distinction from bulk cacao which can originate from multiple geographical areas around the world including a large contribution from cocoa producers from West Africa where 70% of the world production is generated (Makhloufi *et al.*, 2018). Cultivars for bulk cocoa production often lack the range of flavour observed in premium cocoa in favour of higher yield capability.

Only 2% of the cocoa produced in the world is of premium quality with the remaining 98% aimed at bulk usage and there is an increased interest in this premium market and a much higher demand for certified cacao. There is, therefore, a need for methodologies to be developed enabling the characterisation and geographical tracking of certified cocoa products. As described in Chapter 2, for the development of an efficient sustainable tracking process, one important screening stage has been

identified as the farming production level, which is the main stage where traceability is lost.

The use of genomic markers in chocolate tracking

DNA markers have become the key tools for tracking the provenance of food products. While food authentication can be achieved with protein or metabolite markers, processing methods are less damaging to DNA and more likely to provide the right information for identification purposes. DNA markers can target the genome of species and cultivars utilised in the manufacturing of a product. Simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) are the two most robust markers used for identifying variations in plant DNA and usually target nuclear genomes. The same markers can be also found on the chloroplast genome and have been used for many investigations in plant tracking (Schroeder *et al.*, 2016).

The nuclear genome of *T. cacao* has been extensively studied as the genetic improvement of the crop is essential to provide protection against major diseases and enhance chocolate quality. Preliminary studies produced high-density linkage mapping (Argout *et al.*, 2008) and were followed by next-generation sequencing analysis of the whole genome of the crop (Allegre *et al.*, 2012). Further Expressed Sequenced Tagged-SNP and SSR polymorphisms were screened in a collection of 249 diverse genotypes representing the major part of the *T. cacao* diversity with 409 new SSR markers detected on the Criollo genome (Allegre *et al.*, 2012). The high-density map that was generated and the set of new genetic markers identified are crucial in cacao genomics and for marker-assisted breeding, but they also offer a platform for chocolate tracking with the identification of variety-specific markers.

In contrast, markers associated with the chloroplast genome but also the nuclear ribosomal regions are less genetically variable but offer the advantage of being multicopy which is important when studying potentially degraded DNA extracted from processed food. The diversity of these two genomic regions was assessed by Kane *et al.* (2012) who used high-throughput next-generation sequencing to examine the whole plastid genomes as well as nearly 6000 bases of nuclear ribosomal DNA sequences for nine genotypes of *T. cacao* and an individual of the related species

Theobroma grandiflorum. This ultra-barcoding approach demonstrated that all individuals examined were uniquely distinguishable. Chloroplast markers are maternally inherited and only exhibit one allele per locus per plant. Unlike ribosomal markers, this means that any bean produced by a single tree will have an identical chloroplast genome to the maternal tree irrespective of its paternal progenitor.

This chapter as two main aims and focuses on the development of biomarkers to characterise farms. The uses of allelic frequencies in chloroplast microsatellite specific to *T. cacao* is assessed as an indicator of farm or production origin. Since cacao plantations last for over 25 years, at each harvest beans collected from trees planted in a single plot should exhibit an allelic frequency per chloroplast locus which should mirror the allelic frequency found in the tree population. The DNA pattern of the crop from a specific place should be replicable between seasons and give the opportunity of tracing beans back to their source.

This chapter has two aims:

Aim 1: The genetic diversity level of the chloroplast genome in the crop will be characterised using chloroplast microsatellites across reference cultivars from the International Cocoa Quarantine Centre at the University of Reading (ICQC, R).

Aim 2: Chloroplast allelic diversity and proportion will be then examined and compared in a range of chocolate products and cacao beans.

4.2 Materials and methods

4.2.1 ICQC reference DNA samples

Total genomic DNA was extracted from leaf samples collected from 159 reference genotypes from the International Cocoa Quarantine Centre at the University of Reading (ICQC, R) using DNeasy Plant Mini Kit (Qiagen). These accessions were selected to represent the diversity of crop grown worldwide which would enable the assessment of the level allelic diversity that could be expected at each cpSSR locus. DNA purity and concentration of all DNA leaf extracts were measured by Nanodrop 1000.

4.2.2 Chocolate and Beans total DNA extraction

DNA extracted from chocolate samples using a DNeasy mericon Food Kit (Qiagen) as per manufacturers' instructions was obtained from duplicate chocolate samples purchased from food shops in Bristol and Reading in the UK (Hotel Chocolat© single origin: Peru Pichanaki, Coastal Ecuador Hacienda Lara, Venezuela Chuao, Trinidad Cocoa Association, Saint Lucia (Rabot Estate 70% and 100% Saint Lucia) and Madagascar Somia Plantation; Mars© Mars bar and Nestlé© Kit Kat). DNA was also extracted from pools of five dry beans and five dry beans roasted for 30min at 107 ^{o}C obtained from five farm location in Côte d' Ivoire (Martin Gilmour, Mars Wrigley).

4.2.3 Chloroplast locus primer design.

The complete chloroplast genomic sequences from 12 *T. cocoa* genotypes representing a range of accession from South and Central America and one related species *Theobroma grandifolium,* were aligned to identify loci exhibiting polymorphism suitable for fragment analysis screening via capillary electrophoresis (Kane *et al.*, 2012; Swensson *et al.*, 2011; Jansen *et al.*, 2011).

Table 4.1 Chloroplast whole genomic sequences utilised to generate alignment for the identification of polymorphic loci

The accessions include a range of varieties representative of the main cultivars of *T. cacao* and one related species *T. grandiflorum*. Both available sequences of Scavina-6 accession, independently generated were included in the analysis to account for technical variation resulted from Illumina sequencing that Kane et al, 2012, reported.

Sequence ID	Name of accessions	Country of origin	Type of accessions	Notes	Author
JQ228389	Catongo *(T. cacao)*	Unknown	Traditional variety	Amelonado	Kane *et al.*, 2012
JQ228384	EET-64 (*T. cacao*)	Ecuador	Breeding line	Hybrid between Upper Amazon Forastero (Nacional from Ecuador) and Trinitario (from Venezuela).	Kane *et al.*, 2012
JQ228379	Criollo-22 (*T. cacao*)	Costa Rica	Traditional variety	Ancient Criollo variety	Kane *et al.*, 2012
JQ228385	Stahel (*T. cacao*)	Costa Rica	Traditional variety	Trinitario with similarities to lower Amazon Forastero	Kane *et al.*, 2012
JQ228386	Pentagonum (*T. cacao*)	Costa Rica	Traditional variety	Trinitario (Criollo-type)	Kane *et al.*, 2012
JQ228380	Amelonado (*T. cacao*)	Venezuela	Traditional variety	Lower Upper Amazon Forastero	Kane *et al.*, 2012
JQ228387	ICS 39 (*T. cacao*)	Trinidad	Farmer selection	Trinitario	Kane *et al.*, 2012
JQ228383	ICS 6 (*T. cacao*)	Trinidad	Farmer selection	Trinitario	Kane *et al.*, 2012
JQ228381	ICS 1 (*T. cacao*)	Trinidad	Farmer selection	Trinitario	Kane *et al.*, 2012
JQ228388	*T. grandiflorum* (Cupuaçu)	Costa Rica	Related species	Species related to *T. cacao*. Wild and cultivated in Amazon Basin	Kane *et al.*, 2012
JQ228382	Scavina-6 (*T. cacao*)	Peru	Uncultivated tree	Upper Amazon Forastero collected in Ucayali River, Peru	Kane *et al.*, 2012
HQ244500	Scavina-6 (*T. cacao*)	Peru	Uncultivated tree	Upper Amazon Forastero collected in Ucayali River, Peru	Swensson *et al.*, 2011
HQ336404	*T. cacao*	Unknown	Unknown	Clone obtained from United States Department of Agriculture-Agriculture Research Service-Subtropical Horticulture Research Station, Miami, FL	Jansen *et al.*, 2011

4.2.4 Polymerase Chain Reaction (PCR) of chloroplast specific loci

The allelic diversity for each identified polymorphic locus was assessed by amplifying the targeted region via polymerase chain reaction (PCR). For the assessment of each locus, PCR was performed on DNA accession IMC 71 (RUQ 734). PCR volume consisted of 2 µL of DNA (10 ng), 1 µL of forward primer (0.2 µm) and 1 µL of reverse primer (0.2 µm), 12.5 µL of PCR mastermix (DreamTaq PCR mastermix, Thermo Fisher) and 8.5 µL of nanopure water. PCR was performed on a Flexigene thermal cycler (TECHNE) with the following steps of 94°C for 10 mins, 35 cycles of 94°C for 30 secs, 72°C for 1 min and a final extension at 72°C for 10 mins with the product then hold at 4 °C.

To generate products for capillary analysis, PCR amplifications were performed in a final volume of 10 µL, containing 2 µL of DNA (1-10 ng), 5 µL of 2 × Type-it Microsatellite PCR mix (Qiagen), 1 µL of primer mix containing all forward and reverse primers for each specific multiplex mix and the appropriate universal fluorescent primer (0.5 µ M forward primer, 0.125 µ M reverse primer, and 0.5 µ M Hex/Fam-labelled universal primer). PCR reactions were performed on an Applied Biosystems thermocycler with the following programme: 95 ° C for 5 min; followed by 30 cycles of 95 ° C for 30 s, 56 ° C for 90 s and 72 ° C for 30; followed by a final extension at 60 ° C for 30 min. PCR was performed once on all reference samples to obtain sufficient replicates for each specific chloroplast haplotype. Six PCR were performed for each chocolate sample and all bean pools were analysed in triplicates.

4.2.5 Agarose gel electrophoresis for chloroplast PCR products

All primer pairs were tested individually using DNA accession IMC 71 (RUQ 734) and PCR amplification was confirmed for all by screening through agarose gel electrophoresis. Agarose gel of 0.75% were prepared by dissolving agarose (Bioline, UK) in 30 mL of 1x TAE Buffer (Tris-acetate-EDTA) and microwaving for 3 mins at 450w. Agarose gels were added with 3 µl of 10% v/v Sybr Safe (Invitrogen, USA) with a concentration of 10000x (as manufacturer's units). 8 µL of PCR products were combined with 2 µL of loading dye. The gels were loaded with 5 µL of 1kb size ladder (Bioline, UK) and 10 µL of PCR sample per well. The gels were then placed in tanks containing 1x TAE and electrophoresis was performed for 80 mins at 80V. The gels were visualized and photographed using UV gel doc system (Syngene, UK). Because of the close similarity in size of all the loci screened, multiplex PCR samples were directly assessed via capillary electrophoresis.

4.2.6 Fluorescent capillary analysis of multiplex Chloroplast PCR loci

Multiplexed PCR products labelled with FAM and Hex were combined as described in (section 4.2.3) following the identification of suitable polymorphic loci

(Mix1-Mix2 and Mix3-Mix4) (Table 4.2). Samples were sent for capillary analysis to the University of Aberystwyth (UK) and results analysed using the software GeneMarker version 3.2 (Softgenetics) to call the allele sizes. Alleles generated with Hex labelled loci appear as green peaks and Fam labelled loci as blue peaks.

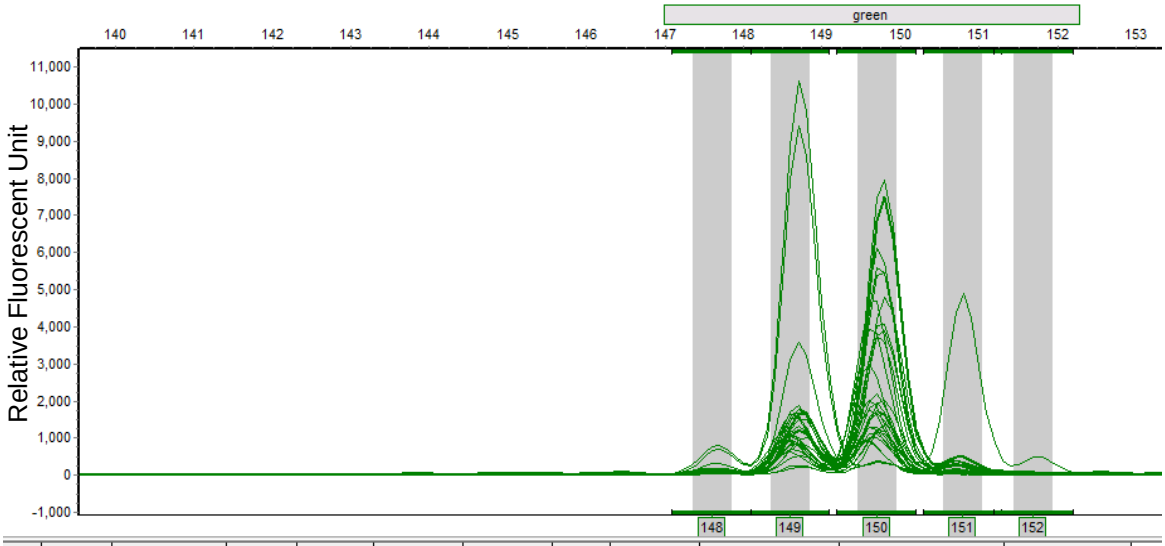
4.2.7 Data Analysis

4.2.7.1 Chloroplast haplotype and allelic diversity

Allelic diversity and chloroplast haplotypes were identified from the reference samples using GenAlEx 6.1 software (Peakall and Smouse, 2006). To visualise the resolution and grouping of chloroplast haplotypes, the results generated from capillary profiles were converted to Fasta DNA sequences summarising all loci screened and to reflect the number of mutation steps between alleles, Appendix VII Figure 0.2 Appendix VII: Haplotype proportion in chocolateFigure 0.2. These summary sequences were used to generate the haplotype network using Network 10 (fluxus-engineering.com).

4.2.7.2 Allelic quantification using fluorescence peak height

For the calculation of the proportion of specific chloroplast alleles per locus, peak height fluorescence was recorded at the position of all know alleles observed in the control panel and potential neighbouring alleles (+/-1 base for mononucleotide cpSSR) using GeneMarker version 2.4 (Softgenetics). The region to be assessed for peak height fluorescence per locus was determined by a preliminary assessment of peaks generated from all samples screened (Figure 4.1). All positions detected were then recorded even if these might correspond to single base stuttering artefact inherent to cpSSR mononucleotides PCR amplification. In the combined data set of reference plants, beans and chocolate, the proportion of alleles in each sample per locus was then calculated by dividing peak height at all positions by the sum of all peak heights. To ensure that allelic proportions could be accurately assessed, samples with an overall peak fluorescence below 200 observed at any specific locus were discarded.



No.	Dye	Marker	Size	Left Range	Right Range	Allele Name	Control	Distance/kb	Chromosome Position	Recombination Frequency	Comments
1	Green	green	147.6	0.5	0.5	148	0	0.00		-1.00	
2	Green	green	148.6	0.5	0.5	149	0	0.00		-1.00	
3	Green	green	149.7	0.5	0.5	150	0	0.00		-1.00	
4	Green	green	150.8	0.5	0.5	151	0	0.00		-1.00	
5	Green	green	151.7	0.5	0.5	152	0	0.00		-1.00	

Figure 4.1 Allele amplification recording with capillary analysis for cpSSR4

The signal per allele are recorded in Relative Fluorescent Units (RFUs) which are shown along the y-axis. The x-axis indicate the base pair sizes of the fragments identified using GeneMarkers. Table below the figure list the size values for each peaks and the left and right flanking region (+/- 0.5) delimiting the bins for each allele. For cpSSR4, five peaks were observed and expected for the allele 148, 149, 150, 151 and 152. Axis: X=RFU, Y=Allele position.

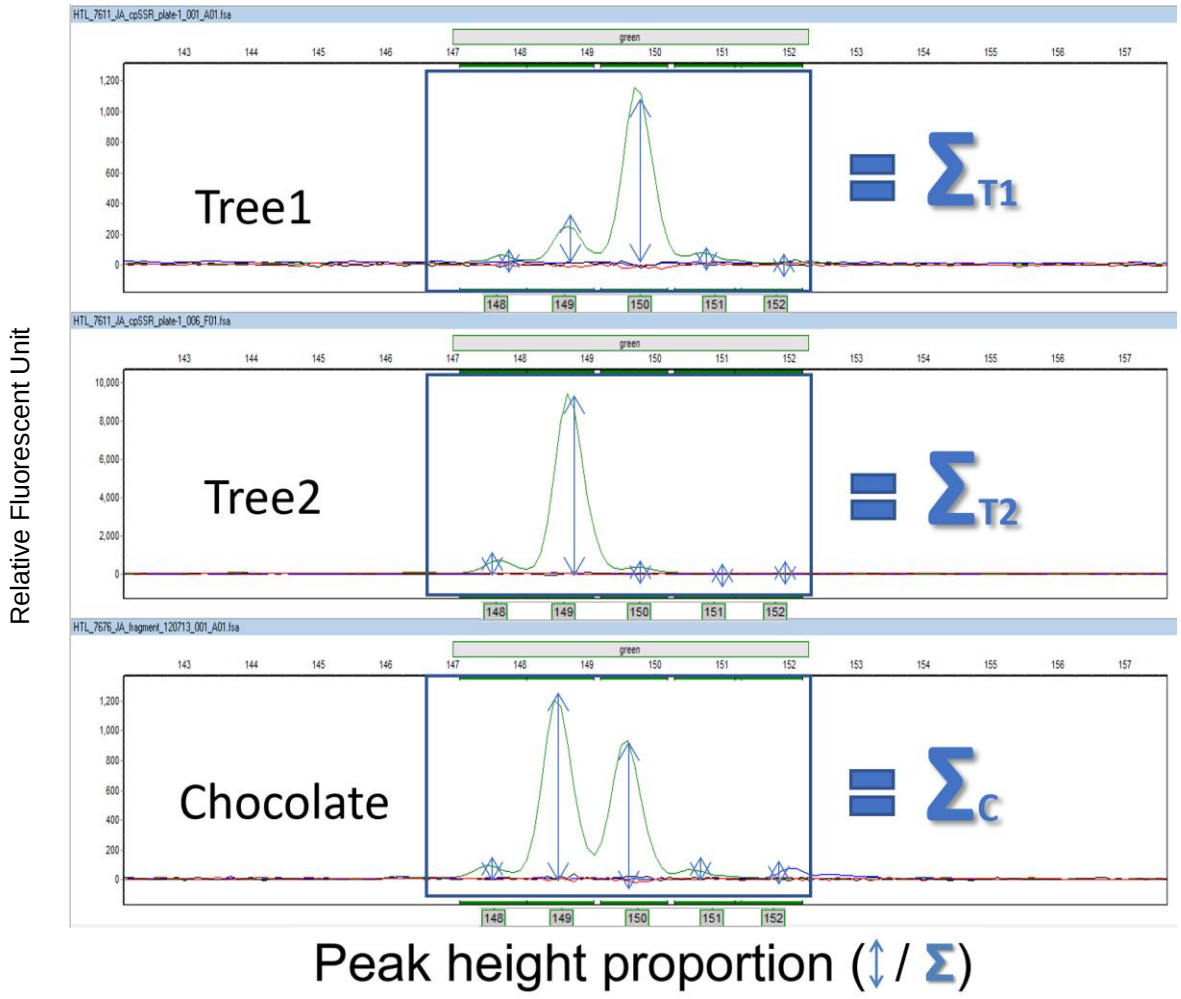


Figure 4.2 Example of cpSSR4 allelic profile for Tree1, Tree2 and a chocolate sample

The contribution of each of the 5 expected peaks for cpSSR4 to the profile is determined as a proportion from the total fluorescence for this marker in each sample. Axis: X=RFU, Y=Allele position.

4.2.7.3 Quantitative analysis of cpSSR allelic proportion in samples

The allelic proportions generated for each sample were then transformed by square root normalization. Ordinations of the data were performed using Principal Coordinates Analysis (PCoA) (Primer-7 v 7.0.13 from Primer-e) to identify clusters from common allelic frequencies. A SIMPER analysis with Bray-Curtiss dissimilarity matrix was performed to calculate the contribution of each allelic percentage to the sim/dissimilarity between each cluster and paired group. The individuals with higher percentage determined the group dissimilarity. The analysis maximises the Spearman rank correlation (Rho) between plant, seed and chocolate RFU reads sample dis/similarity matrices, by checking all combinations of variables.

4.2.8 Quantitative analysis of cpSSR haplotype proportion in samples

Nine polymorphic chloroplast molecular markers identified in this study were analysed as linked markers or haplotypes. These markers were assigned as locus (Region of the chloroplast sequence where one or more alleles exhibits a peak signal), with each locus exhibiting from 2 to 9 alleles for a total of 35 alleles. The data was obtained from normalized readings of the relative fluorescent unit (RFU) signals from capillary electrophoresis sequencing analysis, with a level of detection (0 to *n*) per allele per locus. The proportion of allele RFU in each sample per locus was determined as described in (section 4.2.7.2). Two matrices were structured. One matrix contained 15 Haplotypes formed from previously undetected patterns with no pre-existing labels identified by performing unsupervised machine learning techniques on reference plant material. The second matrix comprehended readings from 116 (Chocolate-DNA) to be analysed against the 15 haplotypes defined in the first matrix.

4.2.8.1 Prediction model for cpSSR haplotype proportion in samples with a supervised machine learning approach

A program was written in the econometrics and time-series analysis software, Regression Analysis of Time Series (RATS), version 7. The model included a combination of a general linear model (GLM), which needs to estimate the beta coefficient in two steps. The chocolate samples were analysed to identify the unique contributors (Haplotypes) as the predictors. The adjusted $\boldsymbol{R^2}$ was used to determine the best model to predict the contribution of haplotypes.

Therefore, for a chocolate sample with N=35 alleles:

$$C = [a_1^c, \dots, a_N^c]'$$

It would be associated with a combination of haplotypes s=1,..,k (k=15):

$$H_s = [a_1^s, \dots, a_k^s]'$$

The first step was to run a regression of the chocolate sample on the 15 haplotypes:

$$a_i^c = \left(1 - \Sigma_1^k \beta_s\right) E(a_i^c) + \beta_1 a_i^1 + \cdots + \beta_k a_i^k$$

$i = 1, \dots, N$ $(N = 35)$, and $E(a_i^c)$ is the mean of the RFU from the chocolate sample.

If, $\left(1 - \Sigma_1^k \beta_s\right)$ is equal to zero, it should demonstrate that the combination of haplotypes is a perfect match and therefore it explains the source of the beans to make that chocolate.

In the second regression, only the haplotypes with significant coefficients were selected. For example, if only haplotype 2 and 15 are significant, the contribution of other haplotypes is deemed insignificant, and the regression is:

$$a_i^c = (1 - \beta_2 - \beta_{15}) E(a_i^c) + \beta_2 a_i^2 + \beta_{15} a_i^{15}$$

The intercept, $(1 - \beta_2 - \beta_{15}) E(a_i^c)$ is the unexplained part of the sample, which could come from molecular contamination, data handling or statistical error of the system.

Furthermore, the haplotype proportions generated for each chocolate sample were then analysed by Principal Coordinate Analysis (PCoA) using Primer-e to confirm clear discrimination between products.

4.3 Results

4.3.1 Chloroplast marker design

4.3.1.1 Identification of *T. cacao* chloroplast genome polymorphic sites

Thirteen whole chloroplast genome sequences from plants of the genus *Theobroma* were obtained from NCBI. These include a sequence for *Theobroma grandiflorum* (JQ228388), twelve sequences from *T.cacao* accessions including, Scavina 6 (HQ244500), Scavina 6 (NC014676), EET-64 (JQ2283840) Criollo-22 (JQ228379), Stahel (JQ228385), Pentagonum (JQ228386), Scavina-6 (JQ228389), Amelonado (JQ228380), ICS39 (JQ228387), ICS06 (JQ228383), ICS01 (JQ228381) and Matina1/6 (HQ336404). No variations were observed between the two sequences from the same varieties Scavina 6 which support the robustness of the sequence data available from NCBI. The screen of the alignment for the presence of cpSSR, Indel polymorphic variants revealed 25 sites exhibiting polymorphism due to variable number of nucleotide repeats (Figure 4.3).

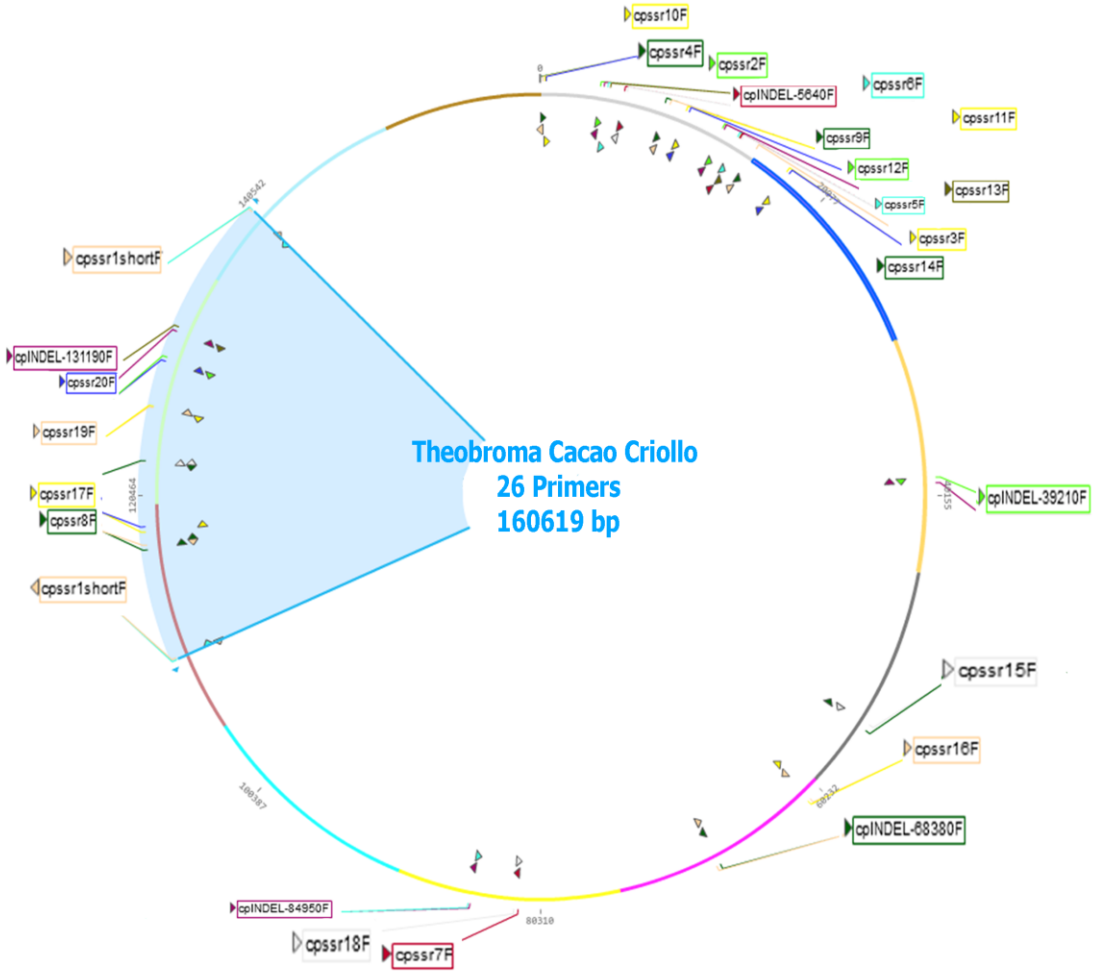


Figure 4.3 Whole Chloroplast genome of *Theobroma cacao* circular representations including the position of forward and reverse primers for all loci

Position of the twenty five polymorphic loci identified following *T. cacao* chloroplast genome sequence analysis. Twenty four markers amplified at unique position (single locus) with cpSSR1 amplifying in the two inverted repeated regions (multi-locus). Made in Genome Compiler 2015.

These include mainly mononucleotide repeats **a/t** and **c/g** but also larger repeat unit ranging from 4 to 80 nucleotides (Figure 4.4; A). For example, cpSSR1 correspond to a highly repeated pentanucleotide motif showing differences in only one of the sequence namely Matina1/6 (HQ336404) (Figure 4.4; B). This repeat is unique as it amplifies in both inverted regions of the chloroplast genome. Indel5 exhibit a larger repeat motif of 80bp also observed in Matina1/6 (HQ336404) which appear, when comparing all whole chloroplast genome of *T. cacao* to be the variety the most distinct from all accessions sequenced (Figure 4.4; C).

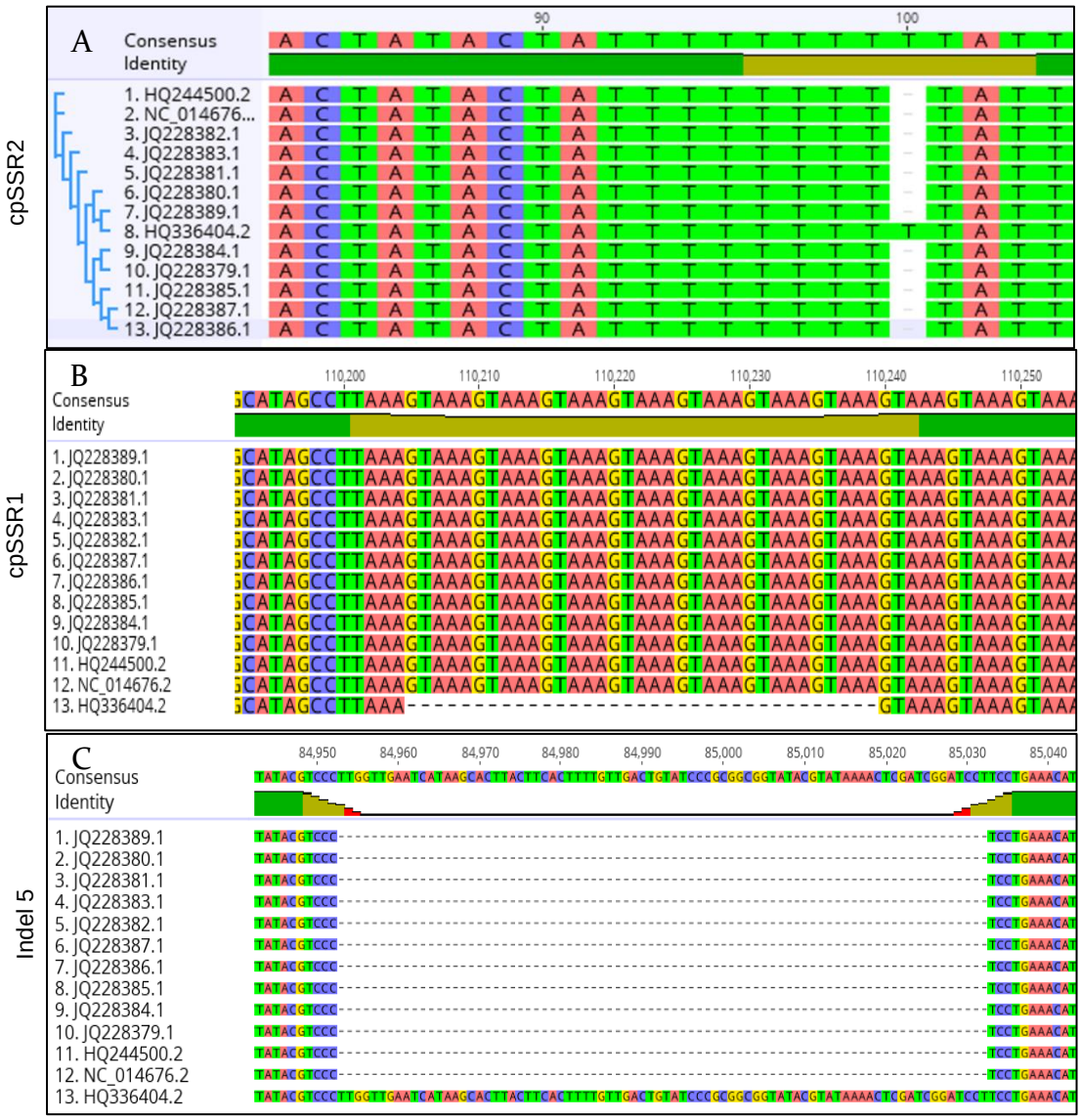


Figure 4.4 Examples of polymorphic loci detected following the alignment of the reference published chloroplast genome sequences

Multi-locus marker identified from the alignment of 13 *Theobroma* chloroplast published sequences. A: Microsatellite cpSSR2 mononucleotide T.B: Multi-locus marker cpSSR1 with repeat unit identified as GTAAA C: Indel 5 locus detected as an insertion of 80bp observed in sequence HQ336404. Alignment generated with Geneious® 10.2.2 (www.geneious.com, Kearse *et al.*, 2012).

4.3.2 Qualitative assessment of loci primer pairs

Primers have been designed for ten *T. cacao* loci in a previous study by Yang *et al.* (2011). For the present work, new primer pairs for all selected loci (including those identified by Yang *et al.*, 2011) were redesigned using Primer3 software (Rozen and Skaletsky, 2000) so that all expected product size ranges did not overlap and would be suitable for multiplex analysis (Table 4.2). All new primers generated single PCR products when applied to DNA accession IMC 71 (RUQ 734) which was confirmed by

screening the products on agarose gel electrophoresis. However, products were always checked by capillary analysis which was the approach used to generate the data for the study. PCR products appearing weak on agarose might not correspond to failed PCR and are likely to be detectable via capillary analysis.

To facilitate multiplexing analysis using capillary electrophoresis, loci primer pairs were combined according to product size to be amplified and labelled with one of two fluorescent tags, Hex and Fam (Figure 4.5). This was achieved by using universal labelled primers complementary to a sequence tag added at the 5' end of all reverse primers. Four multiplexes were produced with the final capillary screening of the PCR products generated performed as Mix1+Mix2 and Mix3+Mix4. The loci used in each mix are indicated in (Table 4.2).

Figure 4.5 Highlight of a multiplex capillary profile of 11 cpSSR labelled FAM and Hex applied on four *T. cacao* accessions from the ICQC, R

The visible 11 microsatellite SSR on this profile amplify fragments between 90 and 250bp and are part of Mix1 FAM labelled (blue peaks) and Mix 2 Hex labelled (green peaks) with expected fragment size from published sequence (**Table 4.2**). Highlighted in red is an example of product generated from cpSSR4, a poly T repeat showing an allelic difference of 1 base to produce two fragments of 149bp and 150bp.

Table 4.2 cpSSR and Indel marker specifications for T. cacao and chocolate tracking

Description of 25 polymorphic chloroplast loci identified from *Theobroma cacao*. Information for these loci includes their name, their previous use by Yang *et al.* (2011), the expected product size in base pairs (bp), the position of the loci on *Theobroma cacao* accession JQ228389 and the repeat motif.

Locus id	Primer name	Primer sequence	Expected PCR amplicon size in bp (JQ228389)	Multiplex Mix - PCR product labeling FAM/HEX	Nucleotide repeat motif	Position of unit (JQ228389)	Yand et al. 2011
cpssr1	cpssr1F	CCTTTCCTCGTTTGAACCTC	114	Mix4 (HEX)	CTTTA	139915	CaCrSSR1 (JF979116)
	cpssr1R	acagctatgaccatgGCACCTTAGGATGGCATAGC					
cpssr2	cpssr2F	CAACCCAATCGCTCTTTTGA	235	Mix1 (FAM)	A	3967	CaCrSSR2 (JF979117)
	cpssr2R	acagctatgaccatgTTTGAATGATTACCCGATCT					
cpssr3	cpssr3F	AGAACGAATCCGCTCCTCTT	215	Mix2 (HEX)	TAAAAG	17261	CaCrSSR3 (JF979118)
	cpssr3R	acagctatgaccatgGGTCACGGCAACATAACAAC					
cpssr4	cpssr4F	GCATGGTGGATTCACAATCC	148	Mix2 (HEX)	T	92	CaCrSSR4 (JF979119)
	cpssr4R	acagctatgaccatgATGATGAATCGTAGAAATGG					
cpssr5	cpssr5F	TCACTTTCACTCCTTTTCCA	195	Mix1 (FAM)	T	13443	CaCrSSR5 (JF979120)
	cpssr5R	acagctatgaccatgTGACTCCGTTTAGACATAGG					
cpssr6	cpssr6F	AATCCCTTTCTTCATACAAA	173	Mix1 (FAM)	C	4490	CaCrSSR6 (JF979121)
	cpssr6R	acagctatgaccatgTTCATGTTTTGATTGCATCG					
cpssr7	cpssr7F	AGGGCTCCGTAAGATCCAGT	265	Mix2 (HEX)	T	81846	CaCrSSR7 (JF979122)
	cpssr7R	acagctatgaccatgGTCTTAGGCCTTGCGATTCA					
cpssr8	cpssr8F	TTTCTGATTCACCGGCTCTT	304	Mix1 (FAM)	T	117430	CaCrSSR8 (JF979123)
	cpssr8R	acagctatgaccatgTGGTGGAATTCTTTGCATTG					
cpssr9	cpssr9F	TCCACTCAGCCATCTCTCCT	339	Mix1 (FAM)	TACTTTAT	8332	CaCrSSR9 (JF979124)
	cpssr9R	acagctatgaccatgGTCCCTTTTGAGCGAAATCA					
cpssr10	cpssr10F	TTAGTCCATAACGGACGTAT	118	Mix1 (FAM)	T	372	
	cpssr10R	acagctatgaccatgCTATTTTATTTTACCATAAG					
cpssr11	cpssr11F	ATGTTATGGGCCGAATTTGT	214	Mix1 (FAM)	T	9747	
	cpssr11R	acagctatgaccatgCGCATCGTTAGCTTGGAAGG					
cpssr12	cpssr12F	TGCGAATAGTATTCAAGATC	134	Mix2 (HEX)	T	12354	
	cpssr12R	acagctatgaccatgTCGGGAAGGGATCATGGAAT					
cpssr13	cpssr13F	CCTATGTCTAAACGGAGTCA	166	Mix1 (FAM)	T	13580	
	cpssr13R	acagctatgaccatgGGAGTGTTGCAACAAATGAG					
cpssr14	cpssr14F	ACAAGAGTTGGCTTATAGCC	173	Mix2 (HEX)	C	14674	
	cpssr14R	acagctatgaccatgCGTCCATTTAGATTGTATCC					
cpssr15	cpssr15F	TGTGAGCGAGCTTATGGGAA	435	Mix1 (FAM)	TAGTTTGCTT	55316	
	cpssr15R	acagctatgaccatgGAATCAGAGCACATGGAACC					
cpssr16	cpssr16F	GAGCCAAGTATCACAACTTC	194	Mix1 (FAM)	A	61261	
	cpssr16R	acagctatgaccatgATTTCCCCATGTTGTGTAAA					
cpssr17	cpssr17F	ATATTGGGCGGCGTTGAAAG	328	Mix2 (HEX)	T	118599	
	cpssr17R	acagctatgaccatgACATAACAAATGGGCGATGG					
cpssr18	cpssr18F	AGATACGCTTGGACCAGAAA	134	Mix1 (FAM)	A	122632	
	cpssr18R	acagctatgaccatgCCCCACTTGAATCCATTTTG					
cpssr19	cpssr19F	CCAATGGAATTCTGTCTGCT	99	Mix2 (HEX)	A	126104	
	cpssr19R	acagctatgaccatgACGGGCGAGATCAATTCAGA					
cpssr20	cpssr20F	TCTTATTGGTGAAACTGACT	277	Mix1 (FAM)	T	129240	
	cpssr20R	acagctatgaccatgGAATCAAATTGAAAACCGGA					
Indel1	Indel1F	CGATTCGATAAACGGCTCAT	184	Mix4 (HEX)	GTATGGAA	5636	
	Indel1R	acagctatgaccatgTGATGGGTTTATGGATCTTTTTG					
Indel3	Indel3F	CTATCCCGCGATGAAAAGAA	348	Mix3 (FAM)	GGGATAT	39207	
	Indel3R	acagctatgaccatgCGCGAAATAGTGCACCTACA					
Indel4	Indel4F	CCAGATTTTTCAGAACCCTTTC	273	Mix4 (HEX)	TTTG	68376	
	Indel4R	acagctatgaccatgCGACCAGATATCAAGCAAACC					
Indel5	Indel5F	AAGCGTTCGAATCCTTGTTG	135	Mix3 (FAM)	TGGTTGAATCATAAGCACTT ACTTCACTTTTGTTGACTGTA TCCCGCGGCGGTATACGTAT AAAACTCGATCGGATCCTT	84954	
	Indel5R	acagctatgaccatgGAATCGCTTCCCAATGGTAT					
Indel6	Indel6F	TTTACTTGGAATGGTGATCCA	307	Mix3 (FAM)	TTTCGT	131195	
	Indel6R	acagctatgaccatgACCAGGACAATGGCGAAATA					

Anchor *acagctatgaccatg*
AnchorFam [FAM]ggaaacagctatgaccatg
AnchorHex [HEX]ggaaacagctatgaccatg

4.3.3 Chloroplast haplotypes and allelic diversity

The 25 chloroplast markers were applied to the reference panel from the International Cocoa Quarantine Centre at the University of Reading (ICQC, R) with 147/159 accessions producing a complete marker profile. The aim here was to assess the likely allelic diversity observed for each locus in plantations of cacao around the world. All 25 loci were polymorphic, ranging from two to 10 alleles with an average of 3.24 alleles per locus (Table 4.3). The results corroborate previous studies with for instance cpSSR 3 and 4 exhibiting the same level of allelic diversity as observed by Yang *et al.* (2011). It was noted that although cpSSR1 target two regions of the chloroplast genome, only one allele per plant was observed within the reference panel. This was also true for all the chloroplast loci screened with a single allele observed per plant per locus. The unbiased haploid diversity per locus calculated using GenAlEx 6.501 software (Peakall and Smouse, 2012) varied from 0.078 to 0.802. Twenty-six haplotypes were detected out of 147 samples using all markers (Table 4.3) and varied in frequency from 0.68% (1 out of 147) to 25.2% (37 out of 147). This preliminary analysis established the allelic range to be screened for the analysis of allelic proportion within chocolate samples. Out of all haplotypes, *Theobroma speciosum* represented by the haplotype H15, exhibited the highest level of allelic differences when compared to the other samples, with unique alleles generated at 8/25 loci. Similarly, the two accessions from *Theobroma grandiflorum* were also clearly different from all remaining *T. cacao* accessions with 7/25 loci generating unique alleles for these individuals. The two haplotypes H24 and H25 that are assigned to these species only differed by one allele.

Table 4.3 Initial screening of the 25cpSSR

Initial screening of the 25 cpSSRs on 147 accessions from The International Cocoa Quarantine Centre at the University of Reading (ICQC, R). Allelic diversity (Na) and unbiased haploid diversity index (h) is provided for each locus. Identified chloroplast haplotypes (Ha), haplotype frequencies in the reference panel, and the alleles pertaining to each haplotype are indicated in base pairs (bp). cpSSR used for the analysis with 4 markers (cp4) are highlighted in red; cpSSR used for the analysis with 9 markers are highlighted in red and brown (cp9).

Species	*T. cacao* Group	Haplotypes 25 loci	9 loci	4 loci	Ha%	Accessions (147)	h	Na	CP locus	cpSSR3	cpSSR4	cpSSR14	cpSSR20	cpSSR1	cpIndel-1	cpIndel-6	cpIndel-5	cpIndel-3	cpSSR19	cpSSR10	cpSSR12	cpSSR18	cpSSR13	cpSSR6	cpSSR16	cpSSR5	cpSSR11	cpSSR2	cpSSR7	cpSSR8	cpSSR17	cpSSR9	cpSSR15	cpIndel-4
T. cacao	1	H11	cp9-2	cp4-3	1.36	2	0.161	2		208	149	171	276	102	177	302	215	349	98	116	132	133	165	176	195	193	213	236	285	303	325	340	443	280
T. cacao	1	H12	cp9-5	cp4-4	4.76	7	0.261	4		215	149	172	275	87	177	302	215	349	98	116	132	133	165	176	195	193	212	236	285	303	325	340	443	280
T. cacao	1	H13	cp9-4	cp4-5	1.36	2	0.423	4		215	149	172	276	82	185	302	133	349	98	116	132	133	166	176	194	194	212	236	285	303	325	340	443	280
T. cacao	1	H14	cp9-7	cp4-5	2.72	4	0.557	5		215	149	172	276	126	185	302	133	349	98	116	132	133	166	176	194	194	212	237	285	303	325	340	443	280
T. cacao	1	H26	cp9-3	cp4-3	0.68	1	0.802	10		208	149	171	276	92	177	302	215	349	98	116	132	133	165	176	195	193	213	236	285	303	325	340	443	280
T. cacao	2	H1	cp9-13	cp4-1	5.44	8	0.078	2		208	150	171	277	92	177	296	215	343	97	115	133	134	165	175	195	193	213	235	286	302	324	340	433	284
T. cacao	2	H3	cp9-13	cp4-1	2.72	4	0.194	2		208	150	171	277	92	177	296	215	343	97	116	133	134	165	175	195	193	213	236	286	302	324	340	433	284
T. cacao	2	H2	cp9-14	cp4-1	2.72	4	0.078	2		208	150	171	277	97	177	296	215	343	97	115	133	134	165	175	195	193	213	235	286	302	324	340	433	284
T. cacao	2	H16	cp9-13	cp4-1	0.68	1	0.377	2		208	150	171	277	92	177	296	215	343	97	115	133	134	165	175	195	193	213	234	286	302	324	340	433	284
T. cacao	3	H9	cp9-1	cp4-2	4.76	7	0.380	3		208	150	171	275	97	177	296	215	349	98	115	133	134	165	175	195	198	213	234	286	303	324	340	443	280
T. cacao	3	H10	cp9-1	cp4-2	6.80	10	0.351	3		208	150	171	275	97	177	296	215	349	98	115	133	134	165	175	195	198	213	235	286	303	324	340	443	280
T. cacao	3	H23	cp9-1	cp4-2	0.68	1	0.218	3		208	150	171	275	97	177	296	215	349	98	115	133	134	165	175	195	198	213	235	286	303	324	340	443	284
T. cacao	4	H4	cp9-10	cp4-7	25.17	37	0.194	2		208	150	172	278	102	177	296	215	343	97	116	133	134	165	175	195	203	213	236	286	302	324	324	433	284
T. cacao	4	H5	cp9-10	cp4-7	4.76	7	0.078	2		208	150	172	278	102	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H6	cp9-11	cp4-7	10.20	15	0.194	2		208	150	172	278	112	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H7	cp9-6	cp4-7	17.69	26	0.078	2		208	150	172	278	136	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H8	cp9-9	cp4-9	1.36	2	0.194	2		208	151	172	278	97	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H17	cp9-8	cp4-6	0.68	1	0.116	3		208	150	173	278	102	177	296	215	343	97	115	133	134	165	175	195	203	213	236	286	302	324	348	433	284
T. cacao	4	H18	cp9-12	cp4-7	0.68	1	0.547	4		208	150	172	278	107	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H19	cp9-16	cp4-7	0.68	1	0.198	4		208	150	172	278	112	177	296	215	343	97	116	133	134	165	175	195	203	213	237	286	302	324	348	433	284
T. cacao	4	H20	cp9-8	cp4-6	0.68	1	0.677	5		208	150	173	278	102	177	296	215	343	97	116	133	134	165	175	195	203	213	236	286	302	324	348	433	284
T. cacao	4	H21	cp9-15	cp4-8	0.68	1	0.228	3		208	150	173	279	102	177	296	215	343	97	116	133	134	165	175	195	203	213	236	286	302	324	348	433	284
T. cacao	4	H22	cp9-9	cp4-9	0.68	1	0.370	2		208	151	172	278	97	177	296	215	343	97	116	133	134	165	175	195	203	213	238	286	302	324	348	433	284
T. grandiflorum		H24	cp9-17	cp4-10	0.68	1	0.230	4		208	148	172	277	100	177	296	215	349	98	116	131	134	165	175	194	194	216	235	289	303	323	346	443	280
T. grandiflorum		H25	cp9-18	cp4-11	0.68	1	0.474	3		208	149	172	277	100	177	296	215	349	98	116	131	134	165	175	194	194	216	235	289	303	323	346	443	280
T. speciosum		H15	cp9-19	cp4-12	0.68	1	0.377	2		208	149	166	275	97	177	296	215	349	95	113	133	134	165	175	193	194	214	234	285	302	316	340	443	275
							0.380	3		208	149	171	276	102	177	302	215	349	98	116	132	133	165	176	195	193	213	236	285	303	325	340	443	280

Preliminary observations of the haplotype distribution in *T. cacao* did not reveal any specific structuring according to the regional origin. For example, haplotype H4 which account for 25.17% of all accessions screened (37/ 147) can be found in Peru, Ecuador French Guiana and Surinam. The same pattern of diverse geographical distribution can be seen in haplotype H12 (4.17% of all accessions, 7/147) found in Peru, Trinidad & Tobago, Colombia, Malaysia and Costa Rica (Table 4.3). This suggests that different chloroplast haplotypes could be found within specific farm location and that differences in chloroplast allelic frequencies are likely to be detected between sites. Haplotype network generated using Network 10 revealed four clear separate groups of haplotypes within *T. cacao* (Figure 4.6). These include Group1 (H11, H12, H13, H14, H26), Group2 (H1, H2, H3, H16), Group 3 (H9, H10, H23), and Group 4 (H4, H5, H6, H7, H8, H17, H18, H19, H20, H21, H22).

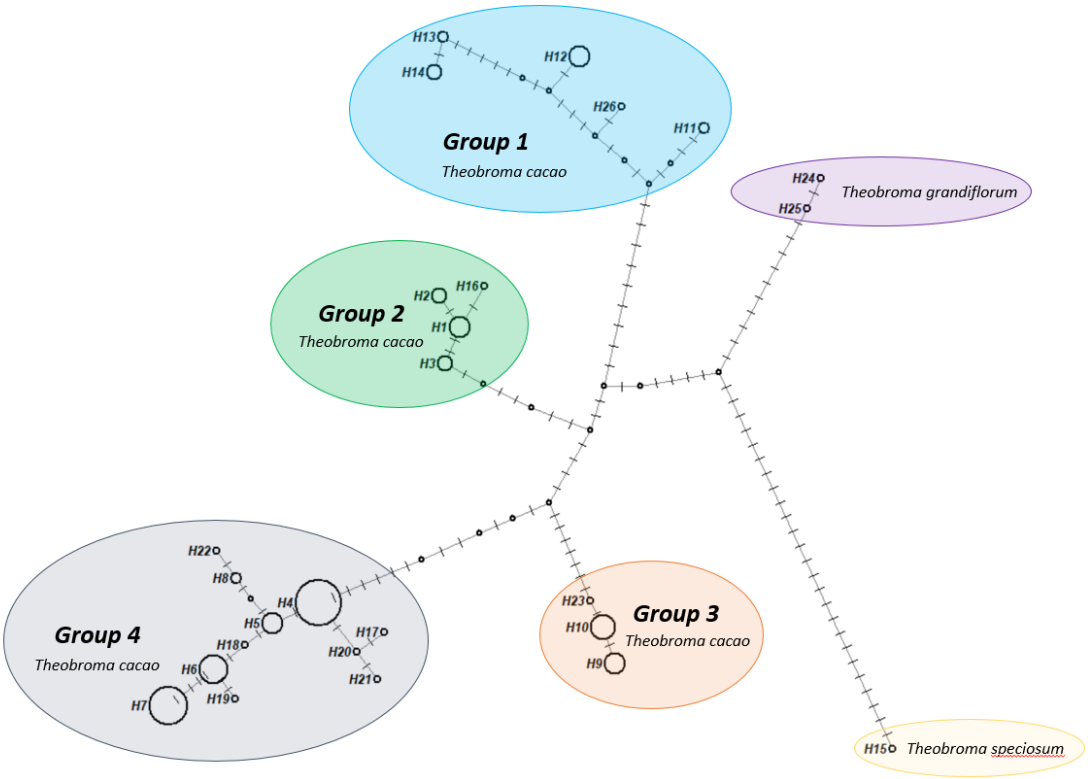


Figure 4.6 Haplotype chloroplast network in *Theobroma cacao* and related species

Haplotype network based on combined chloroplast haplotypes generated from 25 cpSSR loci. H1–25 represent haplotypes detected in *Theobroma*. The size of each circle is approximately proportional to the size of samples (n) harbouring a certain haplotype, with the smallest circles representing n = 1 and the largest representing n = 37. Each solid line represents one mutational step between two haplotypes corresponding to a change in motif repeat. Four clear separate groups of haplotypes within T. cacao include Group1 (H11, H12, H13, H14, H26), Group2 (H1, H2, H3, H16), Group 3 (H9, H10, H23), and Group 4 (H4, H5, H6, H7, H8, H17,

H18, H19, H20, H21, H22). The two other *T. speciosum* and *T. grandiflorum* showed higher distances to the *T. cacao* groups.

4.3.4 How does the number of markers influence the haplotype grouping?

The maintenance of the 4 main haplotype groups was assessed by reducing the number of marker from 25 to 9 (cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Indel1, Indel3, Indel5, Indel6) and 4 (cpSSR3, cpSSR4, cpSSR14, cpSSR20). This reduction in number of markers lower the total number of haplotypes observed from 26 for 25 markers, to 19 for 9 markers and 12 for 4 markers (Table 4.2). In the representative accession selected from the ICQC, R the majority of the *T. cacao* samples (64.6%) were part of group 4, with similar frequencies observed between Group 1, 2 and 3 (11.1, 11.8 and 12.5%). Importantly, the four groups of *T. cacao* haplotypes were still detected when using 9 markers and 4 markers (Table 4.4).

Table 4.4 Comparative distribution and frequencies of haplotypes according to the number of markers used

Columns indicate the number of markers assessed (25, 9 and 4 loci) with the frequency per haplotype. All samples have been grouped by colour as indicated in the haplotype network (Figure 4.6).

***T. cacao* haplotypes frequencies**							
25 loci	**%**	**9 loci**	**%**	**4 loci**	**%**	**Group**	**%**
H11	1.4	cp9-2	1.39	cp4-3	2.08	1	11.1
H12	4.9	cp9-5	4.86	cp4-4	4.86	2	11.8
H13	1.4	cp9-4	1.39	cp4-5	4.17	3	12.5
H14	2.8	cp9-7	2.78	cp4-1	11.8	4	64.6
H26	0.7	cp9-3	0.69	cp4-2	12.5		
H1	5.6	cp9-13	9.03	cp4-7	60.4		
H3	2.8	cp9-14	2.78	cp4-9	2.08		
H2	2.8	cp9-1	12.5	cp4-6	1.39		
H16	0.7	cp9-10	30.6	cp4-8	0.69		
H9	4.9	cp9-11	10.4				
H10	6.9	cp9-6	18.1				
H23	0.7	cp9-9	2.08				
H4	26	cp9-8	1.39				
H5	4.9	cp9-12	0.69				
H6	10	cp9-16	0.69				
H7	18	cp9-15	0.69				
H8	1.4						
H17	0.7						
H18	0.7						
H19	0.7						
H20	0.7						
H21	0.7						
H22	0.7						

4.3.5 Assessing cpSSR markers for plant accessions clustering according to haplotypes using relative fluorescent units (RFU)

4.3.5.1 Assessing the reproducibility of RFU measurement across chloroplast loci

The analysis of chloroplast markers in complex samples such as chocolates required the quantification of all alleles generated by each locus, determined by assessing the relative fluorescence of each allele generated from capillary analysis within a sample. Since a reduced number of markers can still generate sufficient differences to maintain the 4 distinct haplotype groups observed in the ICQC, R accessions, quantitative analysis was performed using the combination of either 9 or 4 markers described in section 4.2.3. The robustness of using allelic proportions to characterise samples was assessed on the reference plant material corresponding to known haplotypes (and therefore expected allelic frequencies). Only *T. cacao* accessions were included in the analysis and PCR products generating peaks lower than 200 RFU for any loci screened were excluded. A total of 35 peaks measurement were included in the analysis corresponding to all expected alleles generated from each marker within the reference panel and two additional peak for locus cpSSR4 (allelic position 148 and 152), one additional peak for cpSSR14 (allelic position 170) and two additional peaks for locus cpSSR20 (allelic position of 273 and 274). These three loci are mononucleotide repeats and generated a higher level of stutter than all other markers and these additional positions were assessed to account for the true variability observed at these loci. In total, 103 accessions were included corresponding to 15 out of the 16 *T. cacao* haplotypes group generated from 9 markers (cp9-5, cp9-4, cp9-7, cp9-3, cp9-2, cp9-13, cp9-14, cp9-1, cp9-8, cp9-12, cp9-15, cp9-9, cp9-10, cp9-11, cp9-6) and 9 haplotypes groups obtained using 4 markers (cp4-4, cp4-5, cp4-3, cp4-1, cp4-2, cp4-6, cp4-7, cp4-8, cp4-9). These were still representative of all main 4 groups of *T. cacao* haplotypes.

The reproducibility of the fluorescent profile for each allele generated was assessed across all accessions generating the same allele. For the 6 loci containing larger motif repeats (cpSSR3, cpSSR1 and all cpIndels), the presence of an allele generated a clear peak contributing to 94.80% to 99.98% of total fluorescence with 2SE ranging from 2.07 to 0.02. For the 3 loci containing mononucleotide repeats, the

fluorescence was distributed across the stutter peaks. For locus cpSSR4 and cpSSR14, 90% of the fluorescence was observed across the main peak of fluorescence and the immediate lower peak (-1bp). cpSSR20 presented a higher level of stuttering with 90% of the fluorescence spread across the main peak of fluorescence and the next two lower peaks (-1bp, -2bp). For these three loci, the main peaks of fluorescence and all associated peaks decreased in size when the allele size increased. For instance, when comparing alleles 171, 172 and 173 in cpSSR14,the main peaks observed were 79.29% (+/- 0.78), 63.30% (+/- 0.41) and 59.41% (+/- 0.27) respectively. At the same time, the main stutter below the peak increased with values of 16.13% (+/- 0.68), 26.27% (+/- 0.18) and 30.62% (+/- 0.42) respectively. This is likely to be a reflection on PCR competition for primers between fragments with lower size fragment preferentially amplified. Overall the peak proportion for all specific alleles was highly reproducible (Table 4.5; Figure 4.7).

Table 4.5 Relative Fluorescent Peak (RFU) measured from 9 cpSSR alleles assessed on 103 plant accessions from the ICQC, R

Table separated in 9 sub-tables for each cpSSR analysed. Means peak contribution per cpSSR for each allele is indicated with corresponding 2 standard error (2SE). n correspond to the number of plants exhibiting the specific allele measured.

		peak 148		peak 149		peak 150		peak 151		peak 152	
	n	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE
cpSSR4- allele 149	12	7.09	0.57	89.32	0.46	2.99	0.17	0.45	0.11	0.16	0.03
cpSSR4- allele 150	89	1.80	0.04	15.66	0.28	77.30	0.26	4.82	0.07	0.42	0.03
cpSSR4- allele 151	2	0.54	0.14	3.07	0.12	19.09	1.03	71.27	0.81	6.03	0.20

		peak 170		peak 171		peak 172		peak 173	
	n	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE
cpSSR14 -allele 171	29	16.13	0.68	79.29	0.78	3.98	0.18	0.60	0.19
cpSSR14 -allele 172	72	4.19	0.14	26.27	0.18	65.30	0.41	4.24	0.33
cpSSR14 -allele 173	2	1.64	0.19	8.33	0.34	30.62	0.42	59.41	0.27

		peak 208		peak 215	
	n	Mean	2SE	Mean	2SE
cpSSR3-allele 208	94	99.84	0.02	0.16	0.02
cpSSR3- allele 215	9	0.38	0.13	99.62	0.13

		peak 273		peak 274		peak 275		peak 276		peak 277		peak 278		peak 279	
	n	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE
cpSSR20- allele 275	19	6.06	0.61	34.00	2.23	56.32	2.44	2.12	0.25	0.57	0.13	0.49	0.18	0.44	0.18
cpSSR20- allele 276	6	1.47	0.16	8.85	1.24	36.56	4.51	49.82	5.05	2.44	0.48	0.52	0.15	0.33	0.14
cpSSR20- allele 277	13	0.75	0.20	2.15	0.21	10.35	1.16	37.52	2.39	45.79	3.17	2.82	0.40	0.62	0.24
cpSSR20- allele 278	64	0.48	0.10	0.83	0.08	2.83	0.11	11.92	0.46	37.67	0.71	43.17	0.95	3.10	0.13
cpSSR20- allele 279	1	0.70	NA	0.56	NA	1.53	NA	4.18	NA	16.02	NA	40.25	NA	36.77	NA

		peak 297		peak 303	
	n	Mean	2SE	Mean	2SE
Indel6- allel 297	91	99.51	0.18	0.49	0.18
Indel6- allele 303	12	0.70	0.59	99.30	0.59

		peak 133		peak 215	
	n	Mean	2SE	Mean	2SE
Indel5- allele	3	99.09	1.46	0.91	1.46
Indel5- allele	100	0.23	0.09	99.77	0.09

		peak 343		peak 349	
	n	Mean	2SE	Mean	2SE
Indel3- allele 343	78	98.60	0.57	1.40	0.57
Indel3- allele 349	25	1.17	0.62	98.83	0.62

		peak 177		peak 185	
	n	Mean	2SE	Mean	2SE
Indel1- allele	100	99.89	0.04	0.11	0.04
Indel1- allele	3	1.46	2.22	98.54	2.22

		peak 82		peak 87		peak 92		peak 97		peak 102		peak 107		peak 112		peak 126		peak 136	
	n	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE	Mean	2SE
cpSSR1-allele 82	19	98.98	0.02	0.14	0.11	0.24	0.08	0.16	0.23	0.16	0.27	0.13	0.21	0.07	0.08	0.03	0.03	0.09	0.12
cpSSR1- allele 87	6	0.27	0.08	97.86	1.16	0.24	0.07	0.14	0.13	0.18	0.11	0.43	0.31	0.13	0.07	0.06	0.06	0.68	1.14
cpSSR1- allele92	13	0.25	0.07	0.38	0.05	98.48	0.29	0.11	0.04	0.15	0.12	0.32	0.13	0.12	0.04	0.03	0.02	0.15	0.18
cpSSR1- allele 97	64	0.14	0.04	0.10	0.03	0.84	0.13	96.65	2.07	0.79	0.84	0.41	0.13	0.16	0.08	0.10	0.04	0.82	1.04
cpSSR1- allele 102	1	0.21	0.06	0.08	0.04	0.22	0.09	1.16	0.19	97.36	0.61	0.59	0.21	0.14	0.05	0.07	0.04	0.17	0.13
cpSSR1- allele 107	6	0.12	NA	0.06	NA	0.14	NA	0.15	NA	1.62	NA	97.60	NA	0.24	NA	0.04	NA	0.01	NA
cpSSR1- allele 112	13	0.13	0.05	0.06	0.03	0.13	0.03	0.08	0.02	0.31	0.13	2.20	0.18	96.86	0.45	0.05	0.04	0.17	0.24
cpSSR1- allele 126	64	0.17	NA	1.02	NA	0.67	NA	1.08	NA	0.52	NA	0.73	NA	0.22	NA	94.80	NA	0.79	NA
cpSSR1- allele 136	1	0.20	0.05	0.08	0.03	0.25	0.11	0.60	0.65	0.38	0.30	0.49	0.12	0.21	0.07	0.40	0.05	97.39	1.05

4.3.5.2 Assessing RFU measurement across cpSSR loci for clustering analysis of *T. cacao* haplotype groups

Principal Coordinate Analysis (PCoA) of the RFU data generated by the 9 markers across 103 ICQC, R accessions revealed 15 clusters corresponding to the *T. cacao* haplotypes identified (Table 4.6). A SIMPER analysis with Bray-Curtiss dissimilarity matrix was performed to calculate the sim/dissimilarity between each cluster (Figure 4.7). High similarity was observed for each cluster (92.32% to 100%) with a low level of dissimilarity observed between clusters from the same haplotype groupings. For instance, Bray-Curtis dissimilarity of 4.69% was observed when comparing cluster cp9-8 and cp9-15. These two haplotypes are part of group 4 and only differ in allelic composition at locus cpSSR20 (alleles 278 and 279). Similarly in the same group, a low dissimilarity of 5.98% was observed between cp9-8 and cp9-10 corresponding to a single allelic difference from locus cpSSR14 (alleles 172 and 173). The haplotypes from group 2, c9-13 and cp9-14 also exhibited a low dissimilarity index of 9.86% with only a single allelic difference from cpSSR1 (Alleles 92 and 97). This low level of dissimilarity for haplotype grouping was reflected in the PCoA diagram with the expected 4 haplotypes group clustering independently (Figure 4.7).

Table 4.6 Bray-Curtis Dissimilarity matrix between the 15 haplotype groups generated using 9 markers (cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Indel1, Indel3, Indel5, and Indel6)

All haplotypes listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 16 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Haplotype	BC-similarity	cp9-9	cp9-8	cp9-6	cp9-11	cp9-10	cp9-12	cp9-14	cp9-13	cp9-15	cp9-1	cp9-7	cp9-4	cp9-5	cp9-2	cp9-3
		Bray-Curtis Dissimilarity 9 mkrs														
cp9-9	92.32		18.15	15.62	17.58	15.07	17.1	19.41	25.37	20.95	30.65	52.98	60.63	49.52	43.36	43.91
cp9-8	100			13.25	13.3	5.98	11.86	20.75	21.07	**4.69**	34.49	55.62	62.12	51.18	37.62	44.49
cp9-6	96.42				10.44	10.41	10.25	19.46	19.15	16.43	32.43	51.74	58.76	47.19	42.95	41.97
cp9-11	97.32					10.39	8.95	19.79	19.2	16.79	32.92	52.84	59.13	47.66	43.12	42.33
cp9-10	96.57						9.19	18.62	19.05	9.39	31.59	51.61	58.52	47.26	35.05	41.75
cp9-12	96.72							18.55	18.37	15.62	32.59	51.17	57.25	46.28	40.51	40.86
cp9-14	97.22								9.86	24.36	18.13	51.8	58.17	48.94	31.91	31.56
cp9-13	96.9									24.94	25.5	53.08	59.11	49.4	33.97	25.91
cp9-15	100										35.16	57.6	64.31	51.89	39.9	46.87
cp9-1	96.23											48.44	53.3	35.03	26.78	26.63
cp9-7	100												14.53	28.12	34.64	34.45
cp9-4	95.96													29.81	36.93	36.02
cp9-5	95.77														27.32	27.02
cp9-2	97.3															10.79
cp9-3	100															

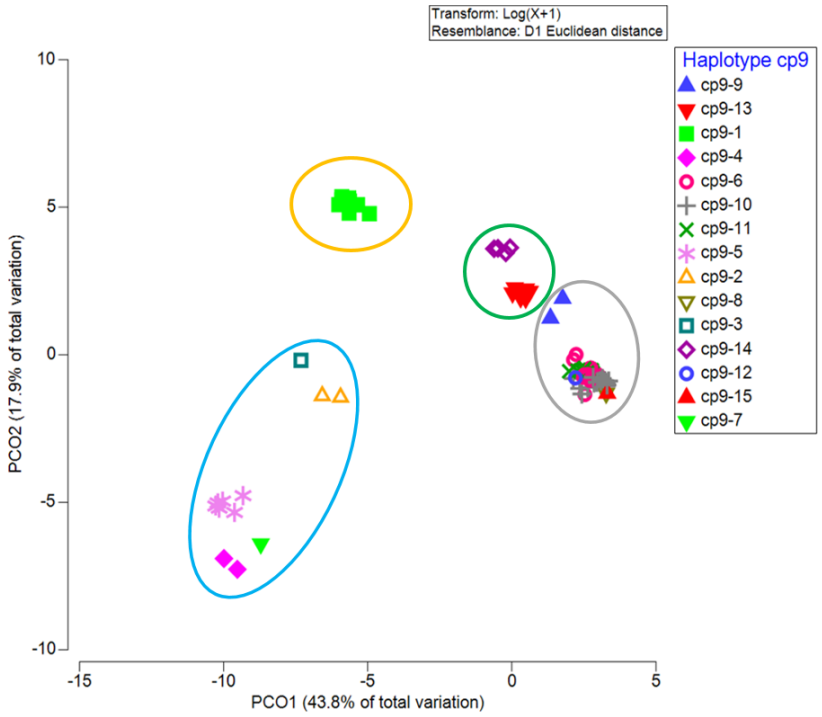


Figure 4.7 Principal Co-ordinates Analysis (PCoA) of the allelic distribution of cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Indel1, Indel3, Indel5 and Indel6 measured as RFU per cpSSR locus in all ICQ, R

Individual plants sharing the same haplotype cluster in units groups. Green, orange, blue and grey circles indicate *T. cacao* haplotype indicates haplotypes member of the four main *T. cacao* network grouping as seen in Figure 4.6. Graph generated with Primer-e.

When the analysis was restricted to 4 markers, PCoA of the RFU, generated distinct clusters corresponding to each 9 *T. cacao* haplotypes (Figure 4.8). A SIMPER analysis with Bray-Curtiss similarity/dissimilarity matrix confirmed the results observed when using 9 markers with high similarity observed for each cluster (96.09% to 100%). Low dissimilarities were observed between the 4 haplotypes from group 4. For instance, these included values of 6.68% when comparing cp4-6 and cp4-8 and 7.65% when comparing cp4-6 to cp4-7 which corresponded to single allelic changes between haplotypes (278 to 279 and 173 to 172 respectively). Similarly, haplotypes cp-4-4 and cp4-5 from group 1 presented a low dissimilarity value of 11.68% with only one allelic difference observed (275 to 276). In contrast, haplotype cp4-3, also part of group 1 exhibited a higher dissimilarity index when compared to cp4-4 and cp4-5 (30.94% and 23.06% respectively) which could be explained by two additional alleles differences (171 to 172 and 208 to 215). Cp4-2 and cp4-3 which also different by two alleles (149 to 150 and 276 to 275) appear to cluster more closely on

the PCoA graph but have a high dissimilarity value at 19.81%. This could be related to the nature of the marker providing the allelic differences with cp4-2 allelic differences to cp4-3 observed at two loci which generated a higher level of stutter (cpSSR4 and cpSSR20) with this effect possibly buffering the main difference between haplotypes. Conversely, these two haplotypes shared common alleles at sites showing less to no stuttering (cpSSR3 and cpSSR14) resulting in proportion observed for these alleles ranging from 79% to 99% (Table 4.5).

Table 4.7 Bray-Curtis Dissimilarity matrix between the 9 haplotype groups using four markers (cpSSR3, cpSSR4, cpSSR14, cpSSR20)

All haplotypes listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 11 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Haplotype	BC-similarity	cp4-9	cp4-6	cp4-8	cp4-7	cp4-1	cp4-2	cp4-5	cp4-4	cp4-3
		Bray-Curtis Dissimilarity 4 mkrs								
cp4-9	97.87		16.55	21.06	11.46	26.49	38.73	44.25	49.17	41.44
cp4-6	100			6.68	7.65	21.01	34.36	43.99	49.93	39.66
cp4-8	100				12.93	27	35.62	47.45	50.75	43.21
cp4-7	97.59					17.02	30	38.57	43.87	35.49
cp4-1	97.16						18.14	38.71	47.5	20.01
cp4-2	96.92							39.78	32.59	19.81
cp4-5	96.09								11.68	23.06
cp4-4	96.5									30.94
cp4-3	97.64									

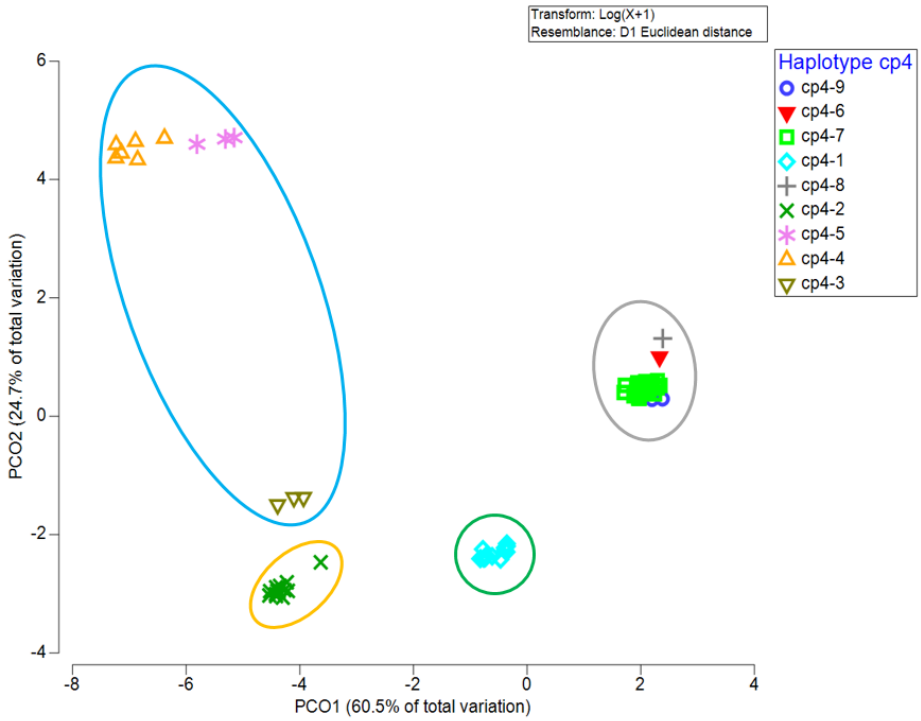


Figure 4.8 Principal Co-ordinates Analysis (PCoA) of the allelic distribution of cpSSR3, cpSSR4, cpSSR14 and cpSSR20measured as RFU per cpSSR locus in all ICQ, R

Individual plants sharing the same haplotype cluster in units groups. Green, orange, blue and grey circles indicate *T. cacao* haplotype indicates haplotypes member of the four main T. cacao network grouping as seen in Figure 4.6. Graph generated with Primer-e.

4.3.6 Analysis of cpSSR allelic frequencies in Chocolate samples using relative fluorescent units (RFU)

4.3.6.1 Comparison of chocolate samples from Hotel Chocolat, Mars Wrigley and Nestle using a combination of 9 and 4 cpSSR markers

The 9 markers preliminary screened on the reference panel were selected to screen a panel of chocolate samples bought commercially. The profile generated via fluorescent capillary analysis revealed good reproducibility within replicates from chocolate samples as observed in the replicates for haplotypes of *T. cacao* (Figure 4.9).

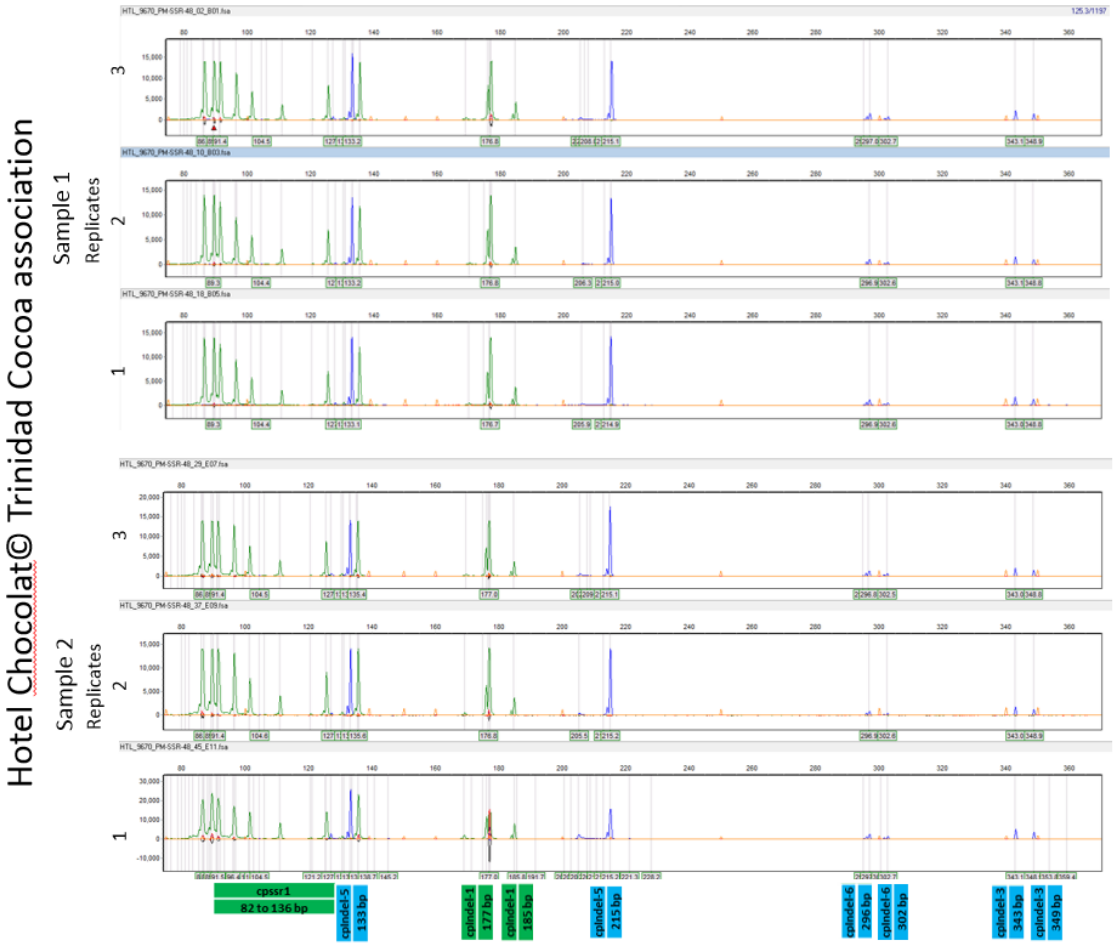


Figure 4.9 Capillary fragment analysis of alleles generated from loci cpSSR1, Indel1, Indel3, Indel5 and Indel 6 labelled with Hex and FAM

Similar profiles observed on triplicate PCR performed from separate DNA extractions of two chocolate samples manufactured by Hotel Chocolate Hotel (75% Trinidad Cocoa Association).

The profiles are triplicates of two chocolate extractions from which exhibit similar signal amplification across technical and chocolate replicates which shows the reproducibility of the system (Table 4.8).

Table 4.8 Assessment of 9 microsatellites markers in 12 chocolates

The fluorescent relative unites (RFU) peaks were scored for each real allele present in the chocolate sample, n = the number of peaks identified in each sample and its replicate.

		Hotel Chocolat samples											
		GroupCE100%	GroupCE100%	GroupM72%	GroupM72%	GroupP100%	GroupP100%	GroupSL70%	GroupSL100%	GroupT75%	GroupT75%	GroupV70%	GroupV70%
Peaks	**n**	6	6	6	6	6	6	6	4	6	6	6	5
cp4-148	Mean	2.79	3.10	3.32	3.11	3.25	3.36	3.69	3.50	4.08	4.07	5.19	5.00
	2SE	*0.43*	*0.48*	*0.40*	*0.23*	*0.43*	*0.45*	*0.45*	*0.08*	*0.53*	*0.52*	*0.36*	*0.43*
cp4-149	Mean	22.85	26.86	25.30	25.86	28.56	28.19	25.44	31.27	38.75	39.77	50.88	51.58
	2SE	*0.70*	*1.22*	*0.19*	*0.24*	*0.49*	*0.95*	*0.32*	*1.91*	*1.85*	*0.52*	*0.49*	*0.15*
cp4-150	Mean	69.16	65.51	66.81	66.37	63.61	64.20	63.44	60.95	53.37	52.68	41.06	40.57
	2SE	*1.82*	*2.89*	*1.57*	*1.36*	*1.67*	*1.99*	*1.42*	*0.89*	*2.40*	*1.79*	*0.99*	*1.04*
cp4-151	Mean	4.82	4.18	4.27	4.30	4.27	3.92	6.59	3.90	3.52	3.23	2.56	2.44
	2SE	*0.90*	*1.17*	*1.10*	*0.89*	*0.83*	*1.19*	*0.88*	*1.01*	*0.83*	*0.83*	*0.55*	*0.89*
cp4-152	Mean	0.37	0.35	0.30	0.37	0.31	0.34	0.84	0.38	0.29	0.26	0.32	0.25
	2SE	*0.12*	*0.09*	*0.10*	*0.03*	*0.07*	*0.06*	*0.07*	*0.16*	*0.07*	*0.08*	*0.05*	*0.10*
cp14-170	Mean	6.33	6.45	10.28	9.65	5.45	6.27	7.91	6.22	6.37	5.83	5.45	4.53
	2SE	*1.29*	*0.58*	*1.68*	*1.39*	*0.89*	*0.95*	*1.27*	*2.07*	*1.18*	*1.06*	*1.28*	*1.15*
cp14-171	Mean	42.27	45.86	60.61	59.55	36.27	38.45	48.96	40.94	40.75	40.80	32.46	33.03
	2SE	*0.61*	*2.11*	*0.96*	*0.98*	*0.37*	*0.65*	*1.35*	*0.78*	*1.15*	*0.86*	*0.51*	*0.81*
cp14-172	Mean	47.83	44.65	27.05	28.66	54.60	51.49	39.64	49.40	49.58	50.34	57.75	59.05
	2SE	*1.43*	*1.51*	*1.07*	*0.83*	*0.89*	*1.13*	*0.47*	*1.73*	*1.23*	*0.78*	*1.87*	*1.59*
cp14-173	Mean	3.57	3.04	2.06	2.14	3.68	3.79	3.50	3.44	3.30	3.03	4.34	3.38
	2SE	*0.27*	*0.37*	*0.26*	*0.13*	*0.32*	*0.48*	*0.58*	*0.41*	*0.60*	*0.26*	*0.62*	*0.49*
cp3-208	Mean	93.55	91.97	91.35	91.66	90.80	90.75	93.47	87.02	67.43	70.27	54.91	53.18
	2SE	*3.53*	*3.16*	*4.08*	*3.16*	*3.91*	*5.95*	*6.06*	*2.34*	*1.38*	*3.28*	*1.25*	*0.67*
cp3-215	Mean	6.45	8.03	8.65	8.34	9.20	9.25	6.53	12.98	32.57	29.73	45.09	46.82
	2SE	*3.53*	*3.16*	*4.08*	*3.16*	*3.91*	*5.95*	*6.06*	*2.34*	*1.38*	*3.28*	*1.25*	*0.67*
cp20-273	Mean	1.32	1.35	2.30	2.59	0.76	0.90	2.86	1.67	1.49	1.38	2.32	1.84
	2SE	*0.25*	*0.23*	*0.67*	*0.58*	*0.22*	*0.09*	*0.35*	*0.85*	*0.17*	*0.25*	*0.31*	*0.86*
cp20-274	Mean	6.98	6.61	16.42	13.52	3.70	3.94	12.17	8.26	7.85	7.25	10.19	9.42
	2SE	*1.11*	*0.74*	*2.75*	*4.98*	*1.14*	*0.21*	*1.99*	*4.38*	*0.46*	*0.53*	*1.21*	*1.88*
cp20-275	Mean	27.49	26.91	43.20	41.38	13.98	16.54	42.94	28.34	26.87	26.26	32.95	34.27
	2SE	*0.99*	*0.63*	*2.27*	*3.99*	*0.81*	*0.40*	*1.80*	*3.68*	*0.84*	*0.51*	*1.56*	*0.97*
cp20-276	Mean	11.78	13.52	7.91	9.53	15.05	14.51	8.16	8.66	17.93	18.20	20.32	20.31
	2SE	*0.60*	*0.67*	*0.45*	*1.29*	*0.83*	*0.26*	*0.89*	*1.35*	*0.49*	*0.19*	*0.92*	*0.74*
cp20-277	Mean	16.84	17.26	11.33	11.65	21.94	21.67	12.03	19.08	20.28	21.10	12.53	12.36
	2SE	*0.55*	*0.58*	*0.63*	*1.40*	*0.93*	*0.14*	*0.64*	*0.64*	*0.33*	*0.29*	*0.29*	*0.57*
cp20-278	Mean	32.64	31.60	17.24	19.49	41.16	38.94	19.62	31.38	23.56	23.88	19.80	20.04
	2SE	*0.83*	*1.69*	*1.15*	*1.37*	*1.21*	*0.82*	*1.32*	*1.95*	*0.46*	*0.43*	*0.36*	*1.37*
cp20-279	Mean	2.95	2.76	1.60	1.84	3.41	3.50	2.22	2.62	2.02	1.94	1.90	1.76
	2SE	*0.34*	*0.44*	*0.35*	*0.37*	*0.65*	*0.53*	*0.35*	*0.29*	*0.33*	*0.21*	*0.37*	*0.45*
Indel5-133	Mean	35.07	39.26	18.42	20.99	38.09	36.96	13.03	15.90	50.31	49.11	58.05	60.33
	2SE	*1.39*	*2.20*	*1.58*	*2.82*	*1.40*	*2.04*	*2.33*	*1.85*	*2.25*	*5.03*	*1.87*	*3.99*
Indel5-215	Mean	64.93	60.74	81.58	79.01	61.91	63.04	86.97	84.10	49.69	50.89	41.95	39.67
	2SE	*1.39*	*2.20*	*1.58*	*2.82*	*1.40*	*2.04*	*2.33*	*1.85*	*2.25*	*5.03*	*1.87*	*3.99*
Indel6-297	Mean	89.26	85.05	86.35	84.25	84.55	85.29	85.26	83.98	68.42	68.15	51.47	51.75
	2SE	*1.18*	*1.24*	*1.07*	*1.51*	*1.37*	*0.55*	*1.58*	*4.67*	*0.52*	*0.91*	*1.45*	*2.41*
Indel6-303	Mean	10.74	14.95	13.67	15.75	15.45	14.71	14.74	16.02	31.58	31.85	48.53	48.25
	2SE	*1.18*	*1.24*	*1.07*	*1.51*	*1.37*	*0.55*	*1.58*	*4.67*	*0.52*	*0.91*	*1.45*	*2.41*
Indel3-343	Mean	58.79	59.49	36.24	38.48	77.67	73.98	41.53	58.15	58.76	59.72	41.61	40.60
	2SE	*1.43*	*0.77*	*0.68*	*0.79*	*0.51*	*0.82*	*0.65*	*2.86*	*0.51*	*0.75*	*0.51*	*0.72*
Indel3-349	Mean	41.21	40.51	63.76	61.52	22.33	26.02	58.47	41.85	41.24	40.28	58.39	59.40
	2SE	*1.43*	*0.77*	*0.68*	*0.79*	*0.51*	*0.82*	*0.65*	*2.86*	*0.51*	*0.75*	*0.51*	*0.72*
Indel1-177	Mean	92.04	91.37	96.98	95.43	88.50	92.14	97.48	96.80	79.63	81.70	74.85	75.10
	2SE	*1.51*	*0.99*	*0.83*	*0.83*	*1.87*	*2.92*	*0.52*	*0.53*	*3.19*	*7.38*	*4.90*	*2.67*
Indel1-185	Mean	7.96	8.63	3.02	4.57	11.50	7.86	2.52	3.20	20.37	18.30	25.15	24.90
	2SE	*1.51*	*0.99*	*0.83*	*0.83*	*1.87*	*2.92*	*0.52*	*0.53*	*3.19*	*7.38*	*4.90*	*2.67*
cp1-82	Mean	0.53	0.42	0.49	0.38	0.39	0.40	2.36	0.38	0.37	0.34	0.61	0.41
	2SE	*0.23*	*0.35*	*0.19*	*0.18*	*0.05*	*0.14*	*0.21*	*1.39*	*0.20*	*0.15*	*0.14*	*0.29*
cp1-87	Mean	2.18	8.08	9.52	12.68	8.02	6.85	18.96	26.42	19.17	20.52	35.42	34.62
	2SE	*1.49*	*0.81*	*4.18*	*1.17*	*0.89*	*2.33*	*1.62*	*3.21*	*1.03*	*1.08*	*4.61*	*6.62*
cp1-92	Mean	1.08	3.16	2.57	3.11	1.66	1.43	4.46	4.78	18.41	19.34	3.29	2.86
	2SE	*1.02*	*0.27*	*0.67*	*0.26*	*0.07*	*0.41*	*0.60*	*0.38*	*0.55*	*0.44*	*0.41*	*0.83*
cp1-97	Mean	29.90	51.05	70.05	63.23	30.02	33.90	45.18	31.49	15.58	14.70	19.49	18.09
	2SE	*0.70*	*1.39*	*1.59*	*0.56*	*0.39*	*1.46*	*1.02*	*2.00*	*0.28*	*0.95*	*0.20*	*0.59*
cp1-102	Mean	33.89	5.03	5.75	6.56	5.94	6.45	12.97	14.91	9.77	9.04	1.84	1.32
	2SE	*4.33*	*2.41*	*8.75*	*2.56*	*1.56*	*2.90*	*1.85*	*2.11*	*0.49*	*1.00*	*2.56*	*2.24*
cp1-107	Mean	0.69	0.33	0.29	0.29	0.63	0.53	1.01	0.43	0.29	0.26	0.21	0.18
	2SE	*0.06*	*0.16*	*0.07*	*0.11*	*0.05*	*0.13*	*0.07*	*0.96*	*0.05*	*0.05*	*0.07*	*0.09*
cp1-112	Mean	13.11	2.89	3.35	4.07	16.83	13.94	8.53	11.63	4.62	4.03	1.02	0.97
	2SE	*0.39*	*2.31*	*0.79*	*0.38*	*0.56*	*0.75*	*0.81*	*2.07*	*0.45*	*1.13*	*0.18*	*0.33*
cp1-126	Mean	2.41	6.22	1.90	2.21	8.24	7.36	1.34	1.81	11.86	11.65	18.32	19.85
	2SE	*0.56*	*1.35*	*0.40*	*0.18*	*0.57*	*1.76*	*0.13*	*0.32*	*0.48*	*0.21*	*1.39*	*2.03*
cp1-136	Mean	16.21	22.83	6.07	7.46	28.28	29.13	5.19	8.16	19.93	20.11	19.81	21.70
	2SE	*1.17*	*2.25*	*1.32*	*0.34*	*1.36*	*3.91*	*0.56*	*0.25*	*1.14*	*0.95*	*0.86*	*1.57*

The data generated from the proportion of specific alleles per locus was analysed separate PCO analysis for 9 and 4 markers respectively to assess the clustering of chocolate samples according to provenance. A SIMPER analysis with Bray-Curtiss dissimilarity matrix was performed to calculate the sim/dissimilarity between each cluster.

Table 4.9 Bray-Curtis Dissimilarity matrix between 8 chocolate samples using cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Inde1, Indel 4, Indel 5, Indel

All chocolate samples including V70% (Hotel Chocolat Venezuela 70%), T75% (Hotel Chocolat Trinidad 75%), CE100% (Hotel Chocolat coastal Ecuador 100 %), M72% (Hotel Chocolat Madagascar 72 %), Mars©(Africa), Nestle© (Africa), P100% (Hotel Chocolat Peru 100 %),SL100% and SL70% (Islands Saint Lucia) are listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 10 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Chocolate	BC-similarity	CE100%	Kitkat	M72%	Mars	P100%	SL	T75%	V70%
		Bray-Curtis Dissimilarity 9 mkrs							
CE100%	95.42		11.8	9.72	12.92	6.54	8.4	10.15	15.47
Kitkat	92.53			18.07	6.62	10.55	14.85	17.25	22.56
M72%	96.88				18.43	12.7	7.52	14.51	16.89
Mars	95.08					11.56	14.85	19.33	25.01
P100%	97.2						10.31	10.4	15.69
SL	95.29							11.64	16.14
T75%	98								8.47
V70%	97.82								

When 9 markers were analysed, distinct clusters were observed corresponding to the various chocolate samples analysed. Clustering was very distinct for samples from Venezuela, Trinidad and Tobago, Peru and Madagascar with Bray Curtiss similarities values ranging from 96.09% to 100%. Groups corresponding to Mars and Nestle clustered closely with a dissimilarity value of 6.62% (Figure 4.11). This could reflect the West African origin of the beans used for the manufacturing of these products. Chocolate samples from Saint Lucia while producing a value of 96.92% via Bray Curtiss analysis exhibited two clusters corresponding to the 70% and 100% chocolate samples. These two samples while from the same geographical location originate from separate farms and the distinct clustering might reflect some differences in the beans origin between the two products. Chocolate samples from the same origin in Ecuador also separated in distinct grouping though also exhibited

a high similarity value of 97.87. These clusters correspond to each of the chocolate samples used in the analysis.

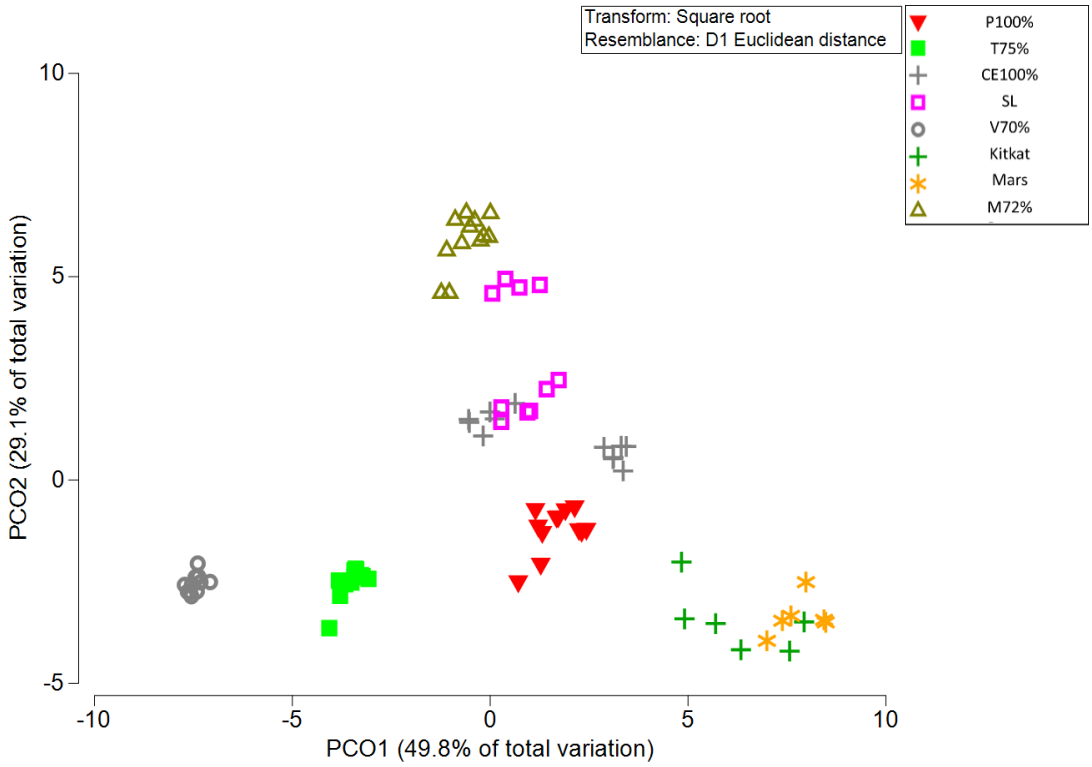


Figure 4.10 Principal Co-ordinates Analysis (PCoA) of cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Indel1, Indel3, Indel5 and Indel6 applied to eight chocolate samples

Principal Co-ordinates Analysis (PCoA) of the allelic distribution of 9 cpSSR markers measured as RFU per cpSSR locus in all 8 chocolate samples including V70% (Hotel Chocolat Venezuela 70%), T75% (Hotel Chocolat Trinidad 75%), CE100% (Hotel Chocolat coastal Ecuador 100 %), M72% (Hotel Chocolat Madagascar 72 %), Mars©(Africa), Nestle© (Africa), P100% (Hotel Chocolat Peru 100 %),SL100% (Above) and SL70% (Below), Islands Saint Lucia. PCoA and graph generated with Primer-e.

Table 4.10 Bray-Curtis Dissimilarity matrix between 8 chocolate groups using cpSSR3, cpSSR4, cpSSR14, cpSSR20

All chocolate samples including V70% (Hotel Chocolat Venezuela 70%), T75% (Hotel Chocolat Trinidad 75%), CE100% (Hotel Chocolat coastal Ecuador 100 %), M72% (Hotel Chocolat Madagascar 72 %), Mars©(Africa), Nestle© (Africa), P100% (Hotel Chocolat Peru 100 %),SL100% and SL70% (Islands Saint Lucia) are listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 10 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Chocolate	BC- similarity	CE100%	Kitkat	M72%	Mars	P100%	SL	T75%	V70%
		Bray-Curtis Dissimilarity 4 mkrs							
CE100%	96.65		9.63	8.24	10.35	5.61	5.31	7.6	12.78
Kitkat	96.08			16.19	**3.75**	7.15	11.99	14.44	18.44
M72%	96.57				16.96	11.98	7.24	11.75	14.37
Mars	96.51					8.06	12.82	15.58	14.37
P100%	96.98						7.66	8.31	13.51
SL	94.72							7.84	11.85
T75%	98.06								6.53
V70%	98.11								

The profile generated by cpSSR4, cpSSR14, cpSSR3, cpSSR20 were analysed all together and provided very similar clustering to the one generated with 9 markers Table 4.10. The separation observed between the accessions from Saint Lucia was maintained but the Ecuadorian replicate samples now formed a unique grouping. A clear separation was observed again for milk chocolate from the Mars and Kit Kat group when compared to all the samples with a variance of Principal components 58%. While geographically different, the allelic frequencies observed might reflect the use of specific trinitario or forastero variety combination in the plantations from which beans were collected to produce these chocolate samples. Venezuela Trinidad and Vietnam samples were widely separated from each other with a variance of PC1 58% and PC2 32%.(Figure 4.11).

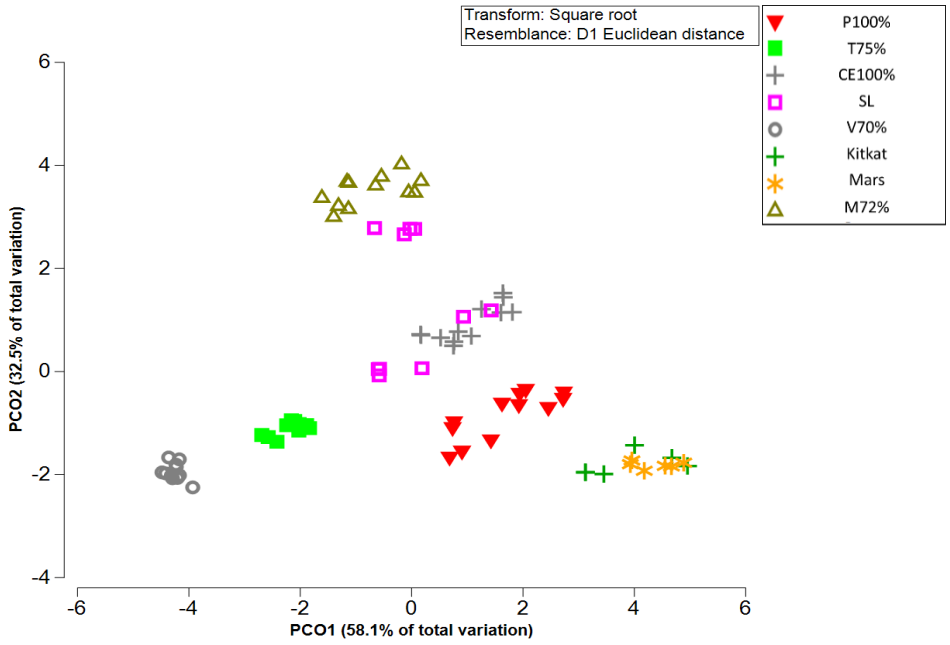


Figure 4.11 Principal Co-ordinates Analysis (PCoA) of cpSSR3, cpSSR4, cpSSR14 and cpSSR20 applied to eight chocolate samples

Principal Co-ordinates Analysis (PCoA) of the allelic distribution of four cpSSR markers measured as RFU per cpSSR locus in all 8 chocolate samples including V70% (Hotel Chocolat Venezuela 70%), T75% (Hotel Chocolat Trinidad 75%), CE100% (Hotel Chocolat coastal Ecuador 100 %), M72% (Hotel Chocolat Madagascar 72 %), Mars©(Africa), Nestle© (Africa), P100% (Hotel Chocolat Peru 100 %),SL100% and SL70% (Islands Saint Lucia). PCoA and graph generated with Primer-e.

4.3.6.2 Comparison of bulk chocolate samples, cacao beans from Ivory Coast and haplotype commonly found in West Africa using a combination of 9 and 4 cpSSR markers

Bulk samples from Mars and Nestle are produced from cultivars grown in West Africa, where more than 70% of the bulk cocoa originate from. Using the 9 selected markers, chocolate samples from these companies were compared to DNA extracts from pools of 5 roasted beans Cote d'ivoire (Ivoseeds) originating from farms in the Ivory Coast and reference plant bearing the haplotype of cultivars most commonly grown in West Africa, cp9-1 and cp9-10. The seeds clustered into two separate groups intermediate between cp9-1 and cp9-10 thus reflecting that pools of seeds extracted might contain each of the haplotypes in various proportion (Figure 4.12). Cp9-10 is a much more common haplotype than cp9-1 according to the reference panel (30.6% compared to 12.5%, (Figure 4.4) which might explain why the cluster leans towards this haplotype with only 3 samples truly intermediate. Chocolate samples from Mars

and Nestle follow an identical pattern with an intermediate position between these two haplotypes though leaning more again towards cp9-10. The clear overlap between Group Mars and Kit Kat chocolates with Ivory Coast seeds reflects shared genotypes and a common region of origin from West Africa.

Table 4.11 Bray-Curtis Dissimilarity matrix between Mars and Nestle chocolate samples, beans from Ivory Coast and haplotypes cp9-1 and cp9-10 using cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Inde1, Indel 4, Indel 5 and Indel6

All samples compared are listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 7 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Haplotype	BC- similarity	cp9-1	cp9-10	Kitkat	Mars	IvoSeed
		Bray-Curtis Dissimilarity 9 mkrs				
cp9-1	96.99		33.12	35.85	35.14	35.2
cp9-10	97.18			16.67	13.28	10.39
Kitkat	92.53				6.62	16.02
Mars	95.08					13.35
IvoSeed	89.48					

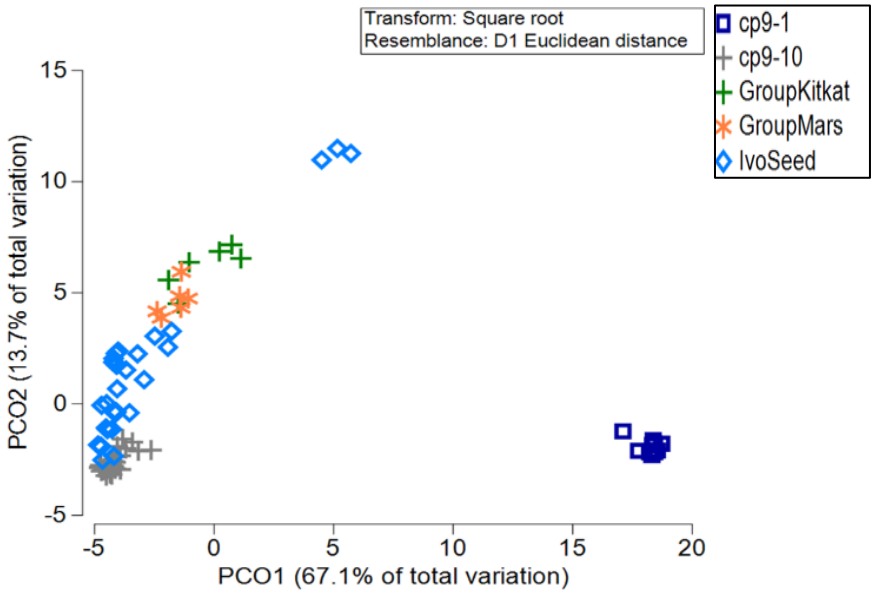


Figure 4.12 Principal Coordinate Analysis (PCoA) of the allelic distribution of cpSSR1, cpSSR3, cpSSR4, cpSSR14, cpSSR20, Indel1, Indel3,

Indel5 and Indel6applied to Mars, Kit Kat, Ivory Coast beans and haplotypes cp9-1 and cp9-10

PCoA and graph generated with Primer-e.

Samples from Mars and Nestle were assessed with the 4 markers and compared to DNA extracted from pools of 5 roasted beans (Ivoseeds) originating from farms in the Cote d'ivoire (Ivory Coast) and reference plant bearing the haplotype of cultivars likely to be commonly grown in West Africa, cp4-2 and cp4-7. cp4-7 is the most common haplotype detected in the ICQC, R panel with a frequency of 60%. The pattern observed with 9 markers was similar with four markers exhibiting strong clustering between both chocolate sample, most of the beans and haplotype cp4-7 with an increase in explained total variation from 67% to 81%. Three bulk seeds samples DNA extracts were intermediate between both haplotypes suggesting a mix of these two haplotypes in these samples. This confirm that the low genetic diversity found in the crop in West Africa is also observed when analysing the chloroplast genome. The allelic proportion typical of cp4-7 might therefore be used to assess bulk products origin. By using just cp4-7 the pattern is much higher compared to the 9 marker analysis (Table 4.12).

Table 4.12 Bray-Curtis Dissimilarity matrix between Mars and Nestle chocolate samples, beans from Ivory Coast and haplotypes cp4-2 and cp4-7 using (cpSSR3, cpSSR4, cpSSR14, cpSSR20)

All samples compared are listed in column 1 with Bray Curtis similarity index for each cluster indicated in column 2. Bray Curtiss dissimilarity index when comparing haplotype groups indicated from column 3 to 7 with higher similarity represented by low values (green) and the most dissimilar haplotypes represented by high values (red).

Haplotype	BC- similarity	cp4-2	cp4-7	Kitkat	Mars	IvoSeed
		Bray-Curtis Dissimilarity 4 mkrs				
cp4-2	97.13		33.12	35.85	35.14	35.2
cp4-7	97.65			16.67	13.28	10.39
Kitkat	96.08				6.62	16.02
Mars	96.51					13.35
IvoSeed	98.55					

Following the dissimilarity analysis, the PCoA showed the same clustering patters.

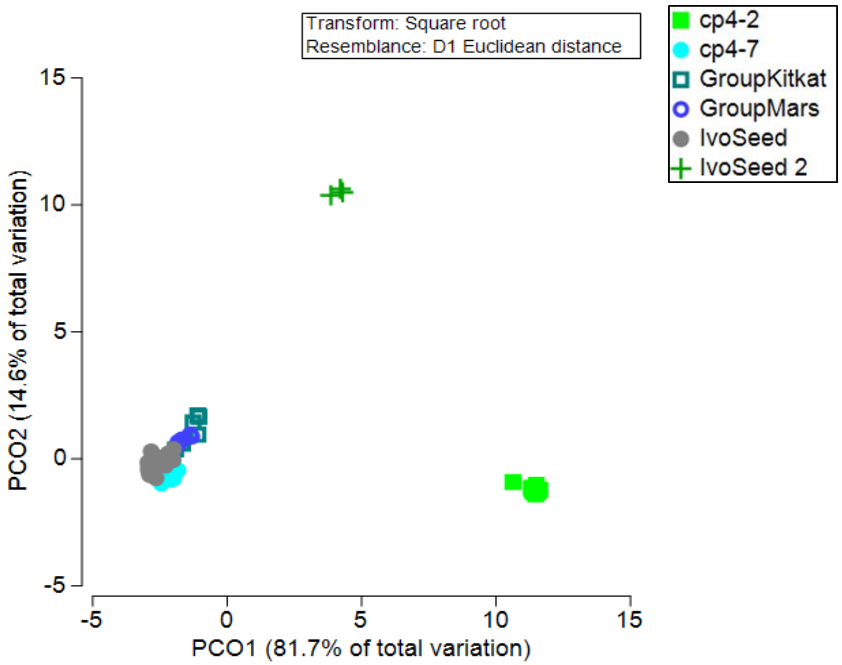


Figure 4.13 Principal Coordinate Analysis (PCoA) of the allelic distribution of cpSSR3, cpSSR4, cpSSR14 and cpSSR20 applied to Mars, Kit Kat Ivory Coast beans and haplotypes cp9-1 and cp9-10

PCoA and graph generated with Primer-e.

4.3.7 Analysis of Haplotype frequencies in Chocolate samples

To determine the proportion of haplotypes in chocolate sample, linear regressions were performed using the regression implemented in RATS software enabling the calculation of haplotype proportions within samples. The 116 chocolate samples were analysed to identify the unique contributors in the sample using haplotype RFU as the predictors. The adjusted $\boldsymbol{R}^2$ per linear regression determined the best model to predict the haplotype contribution and 1-SUM corresponded to the error of the system. Profile generated from 103 plants from the ICQC, R reference panel were assessed with the model to confirm its ability to detect known haplotypes. All plants were allocated to the correct haplotype with value all higher than 0.99 and $\boldsymbol{R}^2$ value superior to 0.99 (Appendix VII). Data generated from technical replicate PCR

of specific samples and replicate sample extraction identified similar haplotype proportions for specific haplotypes which appear to be predominant in their profile (Appendix VII). Strong similarities in profile was for instance observed in samples from Venezuela and Trinidad and Tobago. In some cases, haplotypes were only observed in one of the two chocolate replicates analysed with for instance cp9-10 only observed in sample B of HC Ecuador and cp9-1 present in sample A from Peru. This might be reflecting a slight difference in the composition of the beans used to produce these chocolate replicates. However, some haplotypes were also observed across all chocolate samples in single reaction. The detection of unique haplotype per reaction, sometimes at high level might be caused by error generated by the molecular marker used for this approach or data generated by specific locus. For instance, the analysis of DNA extracted from the pools of 5 seeds from Ivory Coast, exhibit such a pattern with the four haplotypes cp9-6, cp9-10, cp9-11 and cp9-12, consistently identified in various proportions. These haplotypes are very similar to each other with only differences due to one allele at locus cpSSR1 and they compared they did cluster very closely as seen in the analysis of ICQC, R reference samples (Figure 4.8). The haplotype proportions for each replicate sample were averaged between replicates of chocolate accession with country of origin indicated when known.

Table 4.13 Averaged haplotype contribution classified by country/origin in chocolate samples identified by the model for proportion

Information regarding origin, replicate, range haplotype prediction contribution R2, haplotype proportion contribution and unknown/error contribution 1-SUM listed for all chocolate samples analysed.

Chocolate Claim replicate	CHOC	R2	cp9-1	cp9-2	cp9-3	cp9-4	cp9-5	cp9-6	cp9-7	cp9-8	cp9-9	cp9-10	cp9-11	cp9-12	cp9-13	cp9-14	cp9-15	SUM
Ecuador	A	0.871-0.89	0.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.01	0.00	0.00	0.05	0.01	0.05	0.00	0.22
Ecuador	B	0.89-0.933	0.30	0.00	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.36	0.05	0.00	0.00	0.00	0.00	0.19
Madagascar	A	0.939-0.972	0.55	0.00	0.00	0.00	0.05	0.11	0.00	0.00	0.11	0.00	0.00	0.02	0.00	0.03	0.00	0.14
Madagascar	B	0.9-0.984	0.62	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.03	0.01	0.00	0.04	0.01	0.06	0.00	0.16
Peru	A	0.888-0.939	0.18	0.00	0.00	0.00	0.00	0.36	0.05	0.00	0.01	0.00	0.20	0.01	0.00	0.04	0.00	0.16
Peru	B	0.852-0.939	0.00	0.00	0.00	0.00	0.00	0.40	0.08	0.00	0.00	0.00	0.28	0.02	0.00	0.04	0.00	0.18
Saint Lucia 70%	A (70%)	0.937-0.983	0.27	0.00	0.00	0.00	0.15	0.18	0.00	0.00	0.00	0.11	0.22	0.00	0.01	0.00	0.00	0.07
Saint Lucia 100%	B (100%)	0.943-0.978	0.48	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.06	0.26	0.00	0.02	0.00	0.00	0.08
Trinidad and Tobago	A	0.836-0.912	0.08	0.00	0.00	0.00	0.01	0.31	0.26	0.00	0.01	0.00	0.00	0.03	0.20	0.00	0.00	0.09
Trinidad and Tobago	B	0.813-0.897	0.10	0.00	0.00	0.00	0.00	0.30	0.28	0.00	0.02	0.00	0.00	0.01	0.18	0.02	0.00	0.10
Venezuela	A	0.794-0.907	0.02	0.00	0.00	0.00	0.19	0.32	0.32	0.00	0.04	0.00	0.00	0.03	0.00	0.00	0.00	0.09
Venezuela	B	0.761-0.906	0.06	0.00	0.00	0.00	0.19	0.32	0.30	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.11
Mars		0.915-0.985	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.60	0.22	0.00	0.00	0.00	0.00	0.14
Nestle		0.801-0.93	0.00	0.00	0.00	0.02	0.00	0.07	0.00	0.00	0.00	0.46	0.22	0.04	0.00	0.00	0.00	0.19
Ivory Coast Seeds		0.7-0.995	0.00	0.02	0.00	0.00	0.06	0.15	0.00	0.00	0.02	0.54	0.19	0.01	0.00	0.00	0.00	0.01

All chocolate samples exhibited a unique composition and proportion of specific haplotypes with no samples solely represented by a single haplotype Table 4.13. This does reflect the chloroplast haplotypic diversity found in *T. cacao* cultivation. The

proportion of non-assignment to specific haplotypes varied across the samples with the lowest (0.01) observed in the beans from Ivory Coast and the Hotel Chocolat samples from St Lucia, Trinidad and Tobago and Venezuela (0.07 to 0.11). Higher levels were observed in the remaining samples with values up to 0.22 in one sample from Ecuador.

There were high similarity in the proportion of haplotypes per sample cluster for the products from Group Mars and Group Nestle when compared to the beans from Ivory Coast. As mentioned above, the main haplotypes detected for these groups were haplotypes cp9-10 and cp9-11 with a combined frequency for Mars, Nestle and Ivory Coast beans of 0.82, 0.68 and 0.73 respectively. These two haplotypes only differ by one allele within cpSSR1 (102bp and 112bp) and clustered very closely in the plant haplotype analysis. The analysis confirmed cp9-10 (see section 4.3.5.2) as the main haplotype present in West Africa and representative of Amelonado genotypes. This genotype is known to be the main contributor to bulk chocolate produced from this region. It is likely that cp9-11 is also a haplotype found in Amelonado but not represented in the reference panel used in the present study. Conversely, cp9-1 found in the reference panel in Amelonado accessions was not detected in the analysis of beans and samples from Mars and Nestle. This might suggest that this haplotype is rare in Amelonado or other cultivars grown in West Africa.

4.4 Discussion

4.4.1 Marker design and chocolate testing

DNA-based identification systems are being developed for defining premium cacao varieties as “Fino and Aroma”, “Arriba”, “Criollo” and locating the region of where blends of *T. cacao* originated. In Ecuador, Germany, The Netherlands and the UK, most of the investigations to date have been using nuclear sequencing methods for tracking the origins of each variety and to classify the flavour as a geographical mapping. These approaches focus mainly on characterising individual cultivars via nuclear genetic markers. In this chapter, the use of chloroplast DNA markers to characterise the profile of chocolate samples enabled the characterisation and grouping of distinct chocolate samples according to their origin. To complete this,

chloroplast SSR loci were screened via fluorescent capillary analysis and analysed quantitatively both independently from each other and as part of haplotype grouping.

Markers found on the chloroplast genome have been used for many investigations in plant tracking, due to the high copy number and conserved structure of their genome (Schroeder *et al.*, 2016). They can be even used on complexed processed samples as demonstrated by the characterisation of plants species contributing to the making of honey (Hawkins *et al.*, 2015). All chloroplast markers screened in the present study were confirmed to be polymorphic and exhibited unique allelic profile per locus. Previous research in chloroplast structure has shown that the primers developed for cacao ancestry and breeding programs have been mainly designed from the inverted repeat (IR) structure to enable the genetic mapping of the chloroplast. These studies indicated that it is a well-conserved region that separates large and small single-copy (LSC) and (SSC) regions (Daniell *et al.*, 2016) but also provide a sufficient level of variability that can be used for genetic differentiation. Like genes, introns in land-plants chloroplast genomes are conserved but it has been also reported that several plant species have loss intron within protein coding regions (Yagi and Shiina, 2014). Chloroplast evolved from endosymbiosis of a cyanobacterium, with some of the genes transferred from the chloroplast to the nucleus. Some of the genes that have been lost from evolution and transfer to the chloroplast include for instance *infA*, *rpl22*, and *ndh*. Even though studying these genes can be a reliable way of determining the ancestry of plants, their independent study via PCR analysis might generate nuclear data and produce erroneous phylogeny in comparison to the use of whole chloroplast sequences (Daniell *et al.*, 2016).

This was confirmed by Singh *et al.* (2013) who compared these two types of approaches to characterise Indian rice varieties. While all cpSSR markers used in this study were located in introns or intergenic regions of the chloroplast genome, most (shown in Figure 4.1) where near coding regions and genes which might have been transferred into the nuclear genome. For instance Indel 1 was found next to the gene *rps16* (Keller *et al.*, 2017). However, none of the chloroplast markers screened on the reference panel generated more than 1 allele per locus, which would suggest that that none of these regions have been transferred to the nuclear genome. Indeed, the gene

rp16 sequence from accession JQ228389 was submitted to a Blast search on NCBI but did not yield any significant matching results from the *T. cacao* nuclear genome. If sequences have been transferred to the nuclear genome at early stages of *T. cacao* evolutionary history, they might have been subject to DNA mutation and would now be too different to be recognised by the primers designed here to be amplified by PCR. Finally, during the PCR process, the chloroplast locus simply might out-number the nuclear copy which would therefore never be amplified and detected.

A total of 25 cpSSR and Indel loci were screened on a reference panel of cultivars representative of the genetic diversity of the crop. The study reveals the highest number of haplotypes detected in *T. cacao* so far when compared to previous work (Yang *et al.*, 2013; Kane *et al.*, 2010, 2012; Yang *et al.*, 2011; Song *et al.*, 2014) with 22 *T. cacao* haplotypes, two haplotypes specific to *T. grandiflorum* and one to *T. speciosum*. *T. cacao* haplotypes separated into 4 distinct groups with Group 4 accounting for 64.6% of all haplotypes observed and included cultivars grown in West Africa. Previous studies on *T. cacao* chloroplast diversity have focused on assessing the diversity of *T. cacao* in specific countries. For instance, (Yang *et al.*, 2013) assessed the diversity of Trinitario cultivars in Trinidad and Tobago, revealing eight haplotypes which clustered into three highly distinctive groups present in Trinitario cultivars, each corresponding to genotypes for the Criollo (CRI), Upper Amazon Forastero (UAF) and Lower Amazon Forastero (LAF) varietal groups. The authors concluded that these three groups were likely to represent the founding lineages of the cacao crop in Trinidad and Tobago. Similarly, (Gutiérrez-López *et al.*, 2016) used trnH-psbA chloroplast DNAs sequencing and identified 12 chloroplast haplotypes within Criollo accessions from different farms in the Soconusco region of southern Mexico reflecting again on the history of farming and breeding of *T. cacao*.

While the study did not aim at addressing the specificity to haplotypes to *T. cacao* genotype groups, the four main haplotype groups contained accessions originating for diverse origin. It is possible that some cultivars might bear more commonly certain type of haplotypes, but the diversity of origin observed in the ICQC, R panel probably reflect the breeding history of the crop. Indeed, the population genomic analysis of over 100 clones from collections around the world has

revealed that many accessions identified originally as specific *T. cacao* types were in fact admixtures from different genotypic groups and potentially could harbour a range of chloroplast haplotypes (Cornejo *et al.*, 2018). While it would be preferable to be cautious regarding assigning a specific haplotype to a group of genotypes or to a single geographical location, the presence of different haplotypes within cultivars indicates that different level of chloroplast allelic frequencies are likely to be observed within single field thus allowing the proposed tracking approach presented in this chapter to be developed. The assessment of the overall data generated revealed that to maintain the four distinct *T. cacao* haplotypes group and therefore enough level of allelic diversity, a reduced number of loci could be used. This allowed simplification of the analysis and would also reduce the cost of chocolate screening.

All chloroplast markers applied to the chocolate sample produced amplicon that could be scored thus confirming the robustness of such markers on degraded DNA from processed chocolate. Assessing the chloroplast allelic proportion in chocolate samples required the conversion and normalisation of the RFU data generated through capillary analysis. The capillary profiles generated were normalised to determine the proportion of each specific cpSSR alleles per locus identified in each sample studied. Since the data was generated via PCR, it was necessary to first assess the effect of the PCR on all loci screened. Indeed microsatellite units are sensitive to PCR often generating stutter bands that could have an impact on the screening of chocolate (Hosseinzadeh-Colagar *et al.*, 2016).

All markers exhibited clear allelic patterns with all Indel with 4 or more bases motifs not showing any stuttering and all mononucleotide repeats loci exhibiting various levels of stuttering. To account for this variable, fluorescence was recorded at additional peaks which do not correspond to any alleles found in *T. cacao*. This was done to ensure that any variation pattern would be recorded and specified in the Haplotype models designed for the haplotype quantitative analysis. A high level of reproducibility was observed across all plants with identical allelic profiles both in term of the main allele but also for PCR generated stutters. When analysed with 9 markers or 4 markers, all plants clustered according to their haplotype profile. It was noticed that some alleles had a greater influence on the clustering of the haplotypes

compared to others. This could be dependent on whether alleles present stutters and are therefore likely to show lesser differentiation when compared to Indel markers. Choice of the type of markers to be used for the analysis might depend on the number of alleles produced and the level of stuttering observed.

Principal Coordinate Analysis (PCoA) of all samples for the proportion of all alleles using 9 or 4 markers gave contrasting results. Chocolate samples from Mars and Nestle are thought to be mainly derived from chocolate produced in West Africa. The allelic variation observed for Mars © and Nestle© samples were low and very similar to those observed in beans from Ivory Coast. The closest haplotype observed for this group was cp9-10 which was the most common haplotype in the largest *T. cacao* chloroplast Group 4. This is consistent with the current geographical origin of 70% of the bulk chocolate which is West Africa. It makes sense in terms of genetic diversity since most of the crop grown in Ghana and Côte d' Ivoire originates from a limited gene pool mainly derived from the variety Amelonado a type of Forastero and a number of upper Amazon types.

Distinct clustering was observed for Hotel Chocolat samples which are believed to be produced from beans harvested by small cooperatives in Peru, Ecuador, Venezuela, Trinidad, Madagascar and Saint Lucia. These samples originated from specific geographical locations with beans gathered by small cooperatives. The differential clustering observed here reflected the chloroplast genetic diversity present in the chocolate samples which is absent from West Africa. Differences were observed between the two accessions from Saint Lucia which are produced from separate farms. The replicate accession from Ecuador appeared to be clustering in different groups when using 9 markers but this difference disappeared when reducing the panel of markers to four. Since the total number of haplotypes detected with 9 markers was 15 compared to 9 for four cpSSRs markers, it is possible that the differences observed is due to the presence of a specific haplotype. Conversely, some of the additional markers used might have generated noise that promoted this difference. Chocolate extracts from Peru and Ecuador appear to be more closely related in their chloroplast genetic composition and this could be explained by the geographical proximity of the location and a more common occurrence of specific

chloroplast haplotypes in these regions. The remaining samples from Venezuela, Madagascar and Trinidad clearly differentiated in unique clusters for most of the marker combinations.

Identifying clusters specific to chocolate according to chloroplast allelic diversity is useful. This analysis suggests that any characters that could be measured as a proportion within these samples could generate similar results for geographical tracking. But to enable a more complex and informative description, it would be better to be able to assess the proportion of linked markers characteristic of a batch of beans (i.e. field origin) used to produce one sample. These markers could be of diverse nature, including DNA markers from either the chloroplast or nuclear genomes, DNA markers associated to the metagenome from organisms involved in the fermentation of the beans or other metabolites characteristic of specific cultivars. To enable this type of analysis, it is necessary to develop a model to assess the presence of linked markers.

The design of this model was adapted from similar approaches in a different setup. For instance, in human forensics, nuclear SSRs are still the most common type of markers used for genotyping and are screened using capillary analysis. The Combined Probability of Inclusion or Exclusion (CPI/CPE) is regularly assessed to find admix DNA and paired alleles in samples from crime scenes (Perlin, 2015; Bieber *et al.*, 2016). The genetic profile and genotype can be assessed by likelihood ratios, Analysis of Molecular Variance (MANOVA), hierarchical Bayesian probability and use an assignment test to infer population membership and admixed background with linkage disequilibrium to infer ancestry or background of admixed samples and models in DNA mixture datasets (Bieber *et al.*, 2016). Models typically use frequency analysis to determine the proportion of a "pattern" genotype in a mix gene pool, or percentage of a sample in a population that have a specific genotype "pattern" (National Research Council (US) Committee on DNA Technology, 1992).

This approach proposes to find the probability that a randomly chosen unrelated sample could be included as a possible contributor to a mix DNA profile. Prior to comparison with known profiles, peak heights are used to determine whether contributors can be distinguished (Primrose, Woolfe and Rollinson, 2010). When a

known DNA sample can reasonably be expected to be present, the known contribution can be subtracted. The peak high ratio of the RFU are approximately proportional to the amount of DNA from each contributor. And depends on the quality and quantity of the input DNA (Bieber *et al.*, 2016).

Perlin, Szabady, (2001) described that matrix algebra linearly combines genotype allelic pairs to form a markers peak height pattern vector as the mean of a multivariate distribution. The same principle can be adapted for the comparison of unknown cacao samples by utilizing a control panel where more than two combination of alleles generating all known haplotypes have been recorded in reference samples. These results would potentially lead to identify the mixed of beans or plants from different countries and from different batches of production (Belsky *et al.*, 2014).

Chloroplast markers were an ideal model to develop such approach and a quantitative model was developed using 15 common haplotypes observed across reference panel for the crop. Using this approach, distinct chocolate samples could be separated according to haplotype proportions. Clear pattern of similarity was observed between beans from Ivory Coast and bulk chocolate samples from Mars and Nestle, and Hotel Chocolat accession exhibited very different pattern with clearly distinct haplotype contribution to their profile. The results were not as significant as those observed with the direct measurement of allelic proportion.

The error level observed in the analysis could be caused by the type of marker used and indicate that additional work needs to be conducted to refine the model, assess the effect of the combination of markers or the effect of specific cpSSR locus. To further test the resolution level of the system, it would be good to include samples containing a known proportion of haplotypes. This could be achieved by assessing the mixture of DNA from known haplotypes but ideally should be performed on a known mix of DNA from chocolate.

The effective use of data and information is one of the core requirements for new discoveries. In this research, it has been shown that simple molecular biology laboratory techniques and capillary separation methods can be an easy and reliable

option to gather data and re-share data sets to implement new metrics that can be analysed by combining multiple methodological approaches from unsupervised and supervised machine learning. The system could be tested with a different type of markers for which proportion could be more accurately recorded such as the number of sequences with SNP. Several projects have been conducted in recent year to generate genomic information from many *T. cacao* cultivars (Cornejo *et al.*, 2018). The information generated here could be used and incorporated in a similar model as the one described in this chapter by replacing cpSSR markers with a range of SNP characterised on these newly sequenced chloroplast genomes. This system is likely to be more reliable in terms of accuracy but would require the inclusion and development of cheaper Next-generation sequence analysis approaches to make it affordable for large scale screening. In an era of big data and machine learning implementation, genetics and genomics cannot be left behind (Libbrecht and Noble, 2015; Swan *et al.*, 2013). New approaches being developed in forensic analysis rely on building of databases to be used as a reference control (Vlam *et al.*, 2018; Woolfe and Primrose, 2004). For cacao tracking, there is now a clear need for adapting various quantitative methodologies to understand and mine insights from biological trials while correlating them with informative metadata-sets (Country, Source, Weather conditions, Quality profiles).

4.5 Conclusion

The present chapter confirmed the suitability of chloroplast biomarker to be used in characterising the first stage of chocolate production. DNA extracted from a range of commercially available chocolate products were screened using chloroplast microsatellites markers to assess allelic distribution within these samples. These clustered in very distinct groups and support the idea that chocolate produced from beans from specific farms will have an allelic pattern signature for that farm reflecting the diversity of chloroplast haplotypes present.

Chapter 5. Chocolate microbiome as a tool to identify the origin of fermentation/post-harvest location

5.1 Introduction

Food fraud and misleading claims have encouraged researchers, industry and policymakers to improve technologies to detect and avoid illegal entitlements from premium food products. There is a growing interest in barcoding and mapping the vast diversity of microorganism to identify microbiome signatures diagnostic of environmental conditions. These signatures can be used to resolve forensics cases but also have other applications in agriculture and microbiome biotechnology. Much of these studies have been driven by marker-gene surveys (for example, bacterial 16S rRNA genes, fungal internal regions and eukaryotic 18S rRNA genes), which profile microbiota with varying degrees of taxonomic specificity and photeylogenetic information (Attwood *et al.*, 2019; Bolyen *et al.*, 2018). Metagenomics studies of the microbiome of fermented food products (Wine, Cheese, Chocolate, Beer) has been traditionally performed to identify variations in microbial communities during the fermentation process (Cotter and Beresford, 2017; Ludlow *et al.*, 2016; Pinto *et al.*, 2015). In chocolate production, these studies have been used widely to identify harvest seasons and fermentation variables related to chocolate quality (Illeghems *et al.*, 2012; Papalexandratou *et al.*, 2011a).

For instance, a combination of culture-dependent techniques and metagenomics have been assessed in cacao with the aim of isolating communities that can be grown in culture media. These strain can then be used as a starter mix on freshly harvested beans in an attempt to replicate the flavour profiles or characteristics of a chocolate product from a specific area, regions or country (Packard *et al.*, 2019; Papalexandratou and Nielsen, 2016). Studies in other fermented food product like wine have indicated that the microbiota can be associated to the region where it has been produced and the whole production process (Bokulich *et al.*, 2014; Lima *et al.*, 2012). A similar approach involving the analysis of the fermentation microbiome of cacao beans from various geographical locations might therefore reveal information/patterns regarding all manufacturing stages leading to chocolate production. If successful, they could offer a new approach for chocolate quality control and regional signature or terroir.

The cacao fermentation is still one of the last wild "native", traditional fermentation practices and it is believed that these processes in combination with varieties of cacao grown in specific regions enhance regional "typical" characteristics. Indeed, the process of cacao beans fermentation, also known as curing, is critical in the development of key flavour and colour precursors in well-fermented usable cacao beans (Agyirifo *et al.*, 2019). Numerous factors including microclimatic conditions, cacao varieties genotypes and geophysical characteristics can explain the variation observed among the quality of postharvest samples. These factors will determine the composition of the microbiota on the beans following fermentation. Following the harvest of the pods, these are opened with a machete or other sharp artefacts to extract the beans. Beans are covered in a sterile mucilaginous fruit pulp which get rapidly colonised with a variety of microorganisms contributing to the fermentation process. This microbiota is highly variable, mostly due to a range of environmental factors and from the harvest and postharvest practices. These microorganisms mainly present on the pods surface can originate from a range of donors including the hands of farmers/workers, the machetes, vehicles of transportation, animals present on the farm and the mode of fermentation. This fermentation process is conducted differently around the world and include heaps of beans on the ground, heaps covered with banana leaves, the use of wooden or plastic boxes or no fermentation (Afoakwa, 2016a).With so many factors involved in the fermentation, characterising the microbial community of the terroir (origin) where cacao pods are produced, open and where beans are fermented can be a challenge. In addition to this, the microbiota diversity present during the fermentation of the beans varies as fermentation progress. Schwan and Wheals, 2004, showed that culture-dependent, aerobic spore-forming bacteria can be isolated (10^4 CFU/g) during the initial three days of fermentation, thus the population remains unchanged. After the turning of the beans, which increase in oxygen tension, pH and temperature (40°C to 50°C) measurements (5.5×10^7 CFU/g) shows that the microbial population starts to dominate with and extend over 80% of the microbiota of the mass.

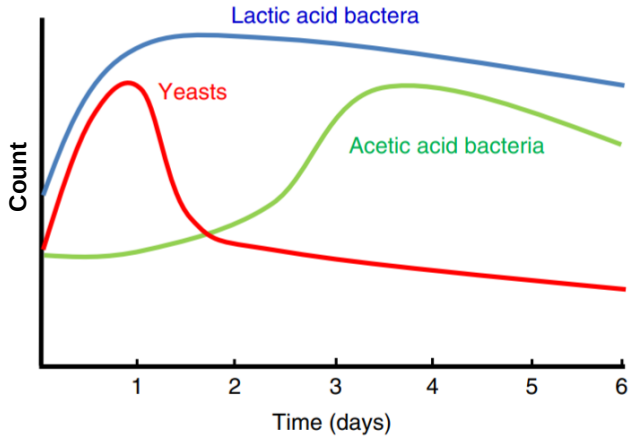


Figure 5.1 Fermentation process; count of microbiological development and time in days of fermentation (De Vuyst and Weckx, 2016).

After more than 130 hours of fermentation, the process of drying starts and eventually only microorganisms that are able to form spores, bacilli, and filamentous fungi can survive *(Lima et al*>, 2012).

Studies of culture-independent communities by amplicon sequencing of 16S rRNA v3-v4 regions and Denaturing Gradient Gel Electrophoresis (DGGE) revealed that species of *Tatumella* and *Pantoea* observed in the initial fermentation microbiome were subsequently and progressively replaced with the dominant species *L. plantarum* and *L. fermentum* (LAB) throughout the fermentation processes, with a peak of *Acetobacter* species observed at the end of the fermentation stage (Lefeber *et al.*, 2011a). Not all the microbiota that is potentially present in the fermentation can be identified by culture-dependent methods. Taxonomic profiling amplicon sequences generated via Next Generation Sequencing (NGS) is now accepted as the standard for inferring the composition of complex microbial ecosystems (Kalyuzhnaya *et al.*, 2008). These approaches generate typically high number of sequences from conserved genes region present in all your targeted species and determine species diversity and abundance within your study sample. For instance, the identification of fungal species in metagenomic samples, the internal transcribed spacer 1 (ITS1) region of the rRNA cistron is a commonly used DNA marker (Schoch *et al.*, 2012). Bacterial diversity screening typically involved the sequencing of the 16S rRNA gene with the hypervariable region v3-v4 most commonly used (Bukin *et al.*, 2019). Other universal bacterial genes with a finer taxonomic resolution can also be

utilised (Ogier *et al.*, 2019), and allow for a more specific and targeted screening of specific species within samples.

Aims

Metagenomics studies of various fermentation stages of cacao have already revealed that there is not an extensive diversity when comparing microbiomes in fermented cacao beans with some common species of bacteria and fungus always detected (Hamdouche *et al.*, 2019). This gives the opportunity to identify from a universal gene region specific Amplicon Sequence Variants (ASV) within all species present but also to focus the analysis on some of the dominant species by analysing species specific genes.

Aim 1: The first aim was to identify unique Amplicon Sequence Variant (ASV) in chocolate samples of microbial species from the cacao bean fermentation microbiota profile.

Aim 2: When identified, the second aim was to verify if these unique markers could be characteristic to specific geographical origin (post-harvest area, country and continent).

The approach focused on the characterisation of genomic DNA from bacterial species within chocolate samples and used an Illumina amplicon screening multiplex approach which included four genes. The reference ribosomal 16s v3-v4 was included to assess and compare the diversity of bacteria present in each sample. A more focused analysis on *Acetobacter pasteurianus*, one of dominant species in the fermentation process, was performed using three housekeeping genes (*dnaK*, *groEL* and *rpoB*) specific to that species. Following bioinformatics analysis, the presence and distribution of AVS was assessed across a worldwide range of single origin chocolate.

5.2 Materials and methods

5.2.1 Chocolate samples and DNA Extraction

Forty-five samples of single-origin dark chocolates from 70% to 100% cacao solids, cacao butter 100% and nibs made of fermented beans from known geographic

origins, were supplied by Tow SuperFood Chocolate, Cocoa Hunters, Bioversity International/Cocoa of Excellence and QRN (Peru). The samples have been classified from fermentation/post-harvest location, country and the continent of provenance (2 Australia, 2 Bolivia, 4 Colombia, 2 Ivory Coast, 15 Ecuador, 2 Ghana, 2 Guatemala, 2 Haiti, 2 Honduras, 2 Indonesia, 2 Malaysia, 4 Peru, 2 Trinidad and Tobago and 2 Vietnam). All samples were assigned with the predominant genetic variety as described by the supplier of each region. Total genomic DNA extractions from cacao samples were performed with Dneasy™ Mericon Food Kit (QIAGEN, UK), as per manufacturer instructions following the standard protocol. Yield optimisation and quality assessment were performed as described in Chapter 2 (DNA Chocolate extraction). DNA concentration per sample was obtained using Qbit and expressed in ng/µl, (Appendix VIII: Chocolate samples.)

5.2.2 Multiplex primer design for Illumina amplicon sequencing

To analyse bacterial diversity, the v3-v4 region of the 16S rRNA gene was amplified by PCR using the universal primers (341F) and (785R). *Acetobacter pasteurianus* specific primer pairs for single-copy housekeeping genes were designed using Primer3 (http://primer3.uee/) from alignments of similar sequences of type and cultured strains found in the Genebank database (NCBI) (http://blasncbi.nlm.nih.gov/) (Table 5.1). To enable co-amplification for all housekeeping genes, the product sizes were all similar ranging from 318 bp to 444 bp. All primers were also modified at the 5-prime end with Nextera adapters target sequences and stabilised with the inclusion at the 3-prime end of a Phosphothioate Oligonucleotide (PTO). The addition of Nextera adaptors enable the construction of individual Illumina library for each 96 samples via a second run of amplification using 96 pairs of individual indexes (N5, N7). This step was performed by Exeter University Sequencing service.

Table 5.1 Multiplexed locus information for Illumina amplicon screening of chocolate samples

Information for all four loci amplified for Illumina analysis are listed including locus ID, forward and reverse primer sequences, targeted microbial species and product expected size. * indicate a Phosphothioate Oligonucleotide (PTO) modification at the second last nucleotide at the 3-primer end to strengthen the bond between the last two nucleotides. W: Weak (A or T), N: Any nucleotide, r: Purine (A or G), r1: Forward read, r2: Reverse read.

Locus id	Primer sequence	Target	Expected PCR amplicon size in bp
16S _V3V4_r1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCa*G	Microbiome	444
16S _V3V4_r2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATc*C		
2RPOBAP_r1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGACCGTAAGCGTAAGCTGc*C	*rpoB* *A. pasteurianus*	355
2RPOBAP_r2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTTCACCCACCAGCACTt*C		
3DNAKAP_r1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCTTGGGCATACGGGTCa*T	*dnaK* *A. pasteurianus*	361
3DNAKAP_r2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGTGAAGACTTTGATAACc*G		
1GROELAP_r1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACGGCTACGGTTCTGGCTCAg*G	*groEL* *A. pasteurianus*	318
1GROELAP_r2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTAGCCrCGGTCrAACTGc*A		

5.2.3 Multiplex Polymerase Chain Reaction (MPCR)

All genes screened were amplified in two separate reactions. All housekeeping genes primers were combined to produce a 2 µm multiplex working solution with a 2 µm working solution produced separately for 16S v3-v4. All PCR amplifications were carried out in a final reaction of 50 µl composed of 25 µl of NEB Next high-fidelity PCR master mix (New England Biolabs), 5 µl of 2 µm primer), 10 µl of DNA template and 10 µl of HyClone™ sterile DNA free water. PCR was performed on a Flexigene thermal cycler (TECHNE) with the following programme: Initial denaturation; 98 ° C for 30 sec for 1 cycle followed by 25 cycles of 98 °C for 10 s, annealing and extension at 65°C for 75s and a final extension of 1 cycle at 65°C for 5 min. The negative PCR controls included 2 µl of HyClone™ DNA free water as a template. 96 PCR were

generated for each primer mix from 45 chocolate DNA templates with x samples triplicated and all remaining samples duplicated. All PCR products were measured using Qbit to assess positive amplification. Both plates were then sent to Exeter University Sequencing Services for completion of the Illumina sequencing procedure.

5.2.4 Illumina sequencing library construction and multiplex

All final steps for the generation of Illumina raw data were performed by Exeter Sequencing Services which provided us for all protocols. A second amplification was performed for each 96 samples with individual dual Indexing (N7, N5) primers complementary to the Nextera adapters attached to the primers of the four genes. PCR was performed as described in section 5.2.3 for both plates separately. This step provided a unique barcode for all sequences generated for each 96 samples. PCR amplicons were purified using MgNa beads on the Bravo LHR and eluted in 30 µl elution buffer. PCR products from both plates were multiplexed as a single pool using equivalent molecular weights (20 ng). This pool was purified using the solid-phase reversible immobilization (SPRI) method (Agencourt AMPure XP beads (Beckman Coulter # A63882, 450ml) to remove free primers and primer-dimer. The library was produced using the TruSeq DNA sample preparation kit (REF 15026486, Illumina Inc, San Diego, CA). The library concentration was verified by qPCR using the NEBNext® Library Quant Kit for Illumina® and library expected size assessed on a Bioanalyzer. Sequence data were generated across 2 runs with a third run performed to repeat replicate samples showing weak signal due to technical Illumina issue.

5.3 Bioinformatics Analysis of Illumina data

5.3.1 Illumina Sequence quality control

Data were generated across two Illumina pair ends runs. For each sequence analysed, a Forward (r1) and Reverse (r2) reads were obtained. These reads when aligned corresponded to a single amplicon sequence from the multiplex pool. Illumina amplicon screening required an extensive sequence quality assessment to remove any erroneous sequences generated from sequencing error. The 16S rRNA and HKG gene sequences generated were processed through the open source software pipeline Quantitative Insights Into Microbial Ecology 2 (QIIME 2) version 2017.2 (http://qiime2.org) to obtain the clean amplicons for each gene of interest (Bolyen *et*

al., 2018). A preliminary screen was performed to remove low-quality sequences determined by (FastQC v0.11.4) using a threshold for high-quality bases equal to Q30 (probability of an incorrect base call is 1 in 1000 and the inferred base call accuracy is 99.9%). Sequenced reads were then trimmed; sequencing adapters removed from the 5' end of all sequences using cutadapt version 1.13 (Martin, 2011).

The reads were then separated in 4 groups corresponding to each gene using cutadapt with the gene primers used as reference. Each amplicon construction was performed from demultiplexed reads (r1) and (r2) Quality control, Illumina-sequenced amplicon errors filtering, and feature table construction followed by denoising was performed with DADA2 incorporated in QIIME2. The DADA2 program filtered out PhiX Control v3 Library reads, removed chimeric sequences, and assigned reads into Amplicon Sequence Variants (ASVs) (Callahan *et al.*, 2016; Janssen *et al.*, 2018). PCR duplicates, ASVs, singletons and features generated from sequencing errors were discarded using feature contingency filtering set to a minimum of at least one feature present in two samples. After obtaining all the ASVs per gene and per run both feature-table from each run were merged together using the sequence overlap and sum of sample frequency.

5.3.2 Sequence identification 16s and HKG

All 16s rRNA ASVs were aligned with SATé-enabled phylogenetic placement (SEPP) technique. To identify and compare the microbial communities in each chocolate sample the metagenomic reads were associated to existing species via q2 alignment (Mirarab, Nguyen and Warnow, 2012) and used to construct a phylogeny insertion tree with fasttree2 (Price, Dehal and Arkin, 2010) using the plugin q2-phylogeny. This step was not required for housekeeping genes since the species was already known as *Acetobater pasteurianus.* ASVs for all genes were grouped into clusters at 97% sequence similarity, taxonomy was assigned using the DADA2 pipeline, which implemented the Ribosomal Database Project (RDP) naïve Bayesian classifier training set (Callahan *et al.*, 2016). ASVs for 16S amplicon were classified taxonomically against a representative subset of the Greengenes 16S rRNA database (McDonald *et al.*, 2012b) with 99% sequence similarity using q2-feature-classifier, a QIIME 2 (https://qiime2.org) plugin for taxonomy classification of marker-gene

sequences (Bokulich *et al.*, 2018; Pedregosa *et al.*, 2011) and assignment bar charts were generated using Phyloseq in from QIIME2. This last procedure was repeated after screening for contaminants allowing removing mitochondrial and chloroplast sequences. The same approach was performed with each *Acetobacter pasteurianus* HKG utilising Greengenes databases.

5.3.3 Statistical Analysis

16s Analysis of bacterial communities

To identify ASVs of interest and their dispersion between the samples, a heat map was constructed from the raw ASV features table by using Euclidean distances and centroid method. To identify core ASVs per origin, all samples were classified using the metadata according to Sample name, Fermentation Location, Continent, Country and year of production. All statistical analyses were completed using QIIME2 core-metrics, R studio and PRIMER-e software version 7. Species evenness estimate was calculated using qiime-diversity, to determine statistical significance in alpha diversity (within-samples) a non-parametric Kruskal–Wallis tests using Pielou's evenness values was used to complete pairwise comparisons between fermentation region, country, continent and year of production. Beta diversity (between-samples) metrics clustering was performed with Jaccard and Bray-Curtis distance matrix for an initial screening and a phylogenetic approach (unweighted UniFrac) to identify dissimilarity using the unique locus of the ASVs of the communities per sample. Principle Coordinate Analysis (PCoA) was performed using qiime-diversity after samples were subsampled "rarefied-method" to 1000 sequences per sample.

Assessment of diversity differences between regions and identification of markers specific to region/fermentation, country and continent

Differential abundance, analysis of composition of microbiomes (ANCOM) test were run to determine if there were significant differences in the relative abundance of any individual driving taxa discrimination between regions (Ghannam *et al.*, 2020; Paun *et al.*, 2019; Mandal *et al.*, 2015) and tested for difference from zero using a one-sample t-test with Benjamini-Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 1995). After identifying all the dominant species with 16S v3-v4 and benchmarking with previous taxa diversity results, all ASVs with their

barcodes from 16S and the three *Acetobacter pasteurianus* HKG were recorded and stored in a database per gene. All feature tables and frequencies per group were stored as TSV and BIOM files (McDonald *et al.*, 2012a). To determine whether sample classifications (region, year of the collection) contained differences in phylogenetic or taxa diversity, analysis of similarities (ANOSIM) and permutational multivariate analysis (MANOVA) with 999 permutations were used to test significant differences between sample groups based on unweighted UniFrac and Bray–Curtis distance matrices. In addition, (PCoA) was performed in Primer-e and used to analyse beta diversity in different chocolate, which was conducted on the basis of the calculated unweighted UniFrac, Jaccard and Bray Curtis distances (Lozupone and Knight, 2005). To identify if the markers were unique from the locations the data obtained of 100% presence of the ASV in the sample from the ANCOM analysis was selected and PERMANOVA was assessed to determine whether the groups of samples are significantly different from one another (Anderson, 2001). ASV were then assessed across all samples for their specificity to samples, country and continent.

5.4 Results

5.4.1 16S rRNA v3-v4 and Housekeeping genes sequence output overview

The demultiplexed MiSeq sequencing analysis from chocolate samples yielded an overall sequence count of 17million raw reads with a quality score above 33 (Phred) before trimming and an average length of 417 base pairs per sequence. Amplicons were created with the merged runs and 16S v3-v4 amplicon yielded 6,743,192 raw sequences with a frequency per sample of 46,340.063 reads. Following quality filtering, and removal of samples with reads lower than, a total of 594 ASVs were generated for downstream analysis. All HKG specific to *Acetobacter pasteurianus* loci amplified efficiently. This was predicted for a bacteria present in abundance in the last stages of the fermentation process. Overall for all 96 samples, ***rpoB*** yielded 7,952,320 raw 313 bp sequences, a frequency per sample of 44706.06 reads count and 276 ASVs; ***groEL*** *yielded* 1,150060 raw 313 bp sequences, a frequency per sample of 14662.54 and 246 ASVs; ***dnaK*** yielded 1,030173.5 raw 330bp sequences, a frequency per sample of 11276.87 reads count and 256 ASVs.

5.4.2 Microbial community analysis with 16S rRNA v3-v4

5.4.2.1 Distinguishing patterns of microbial composition profiles

Levels of discrimination of the retained 94 samples were visualized using a QIIME2 heatmap (Janssen *et al.*, 2018; McKinney, 2010; Hunter, 2007) via hierarchical clustering method (UPGMA) from the centroid under Euclidean distances, showing the relationship of the country of production. There is a consistent pattern of high frequency features clustering together, high frequency in single samples, low frequency heats across all samples and also isolated heats (Figure 5.2). This qualitative approach using the frequency of the relative abundance of the features (Figure 5.3), shows a pattern which separate the different regions of the world with unique sequences AVS identified per regions. A second heatmap using relative abundance with taxonomic assignments, disclosed that these communities clustering are mainly from family of bacteria associated to the fermentation: *Acetobacteraceae, Lactobacillaceae, Corynebacteriaceae, Bacillaceae, Micrococaceae* (Figure 5.3), which confirms that it is possible to identify microbiome from the fermentation stage in chocolate DNA. In addition some sequences are only found in some regions and other features are apparently shared between regions, which suggests that there might be a pattern into the distribution of specific bacteria in the process or origin. High frequency features (avg 53000) found in most samples were related to *Acetobacteraceae* and *Lactobacillaceae*, further analysis showed that features with high frequency only found in some places where identified as specific species e.g. *Lactobacillaceae Pediococcus.* Isolated heats in low frequency (avg < 5000) across all samples showed specific bacterial ASVs for *Bacillaceae* thus, other low frequency features found for example in Bolivia and Peru where found as *Lactobacillus Hamsteri* which can point to identifying unique sequences for those regions or process.

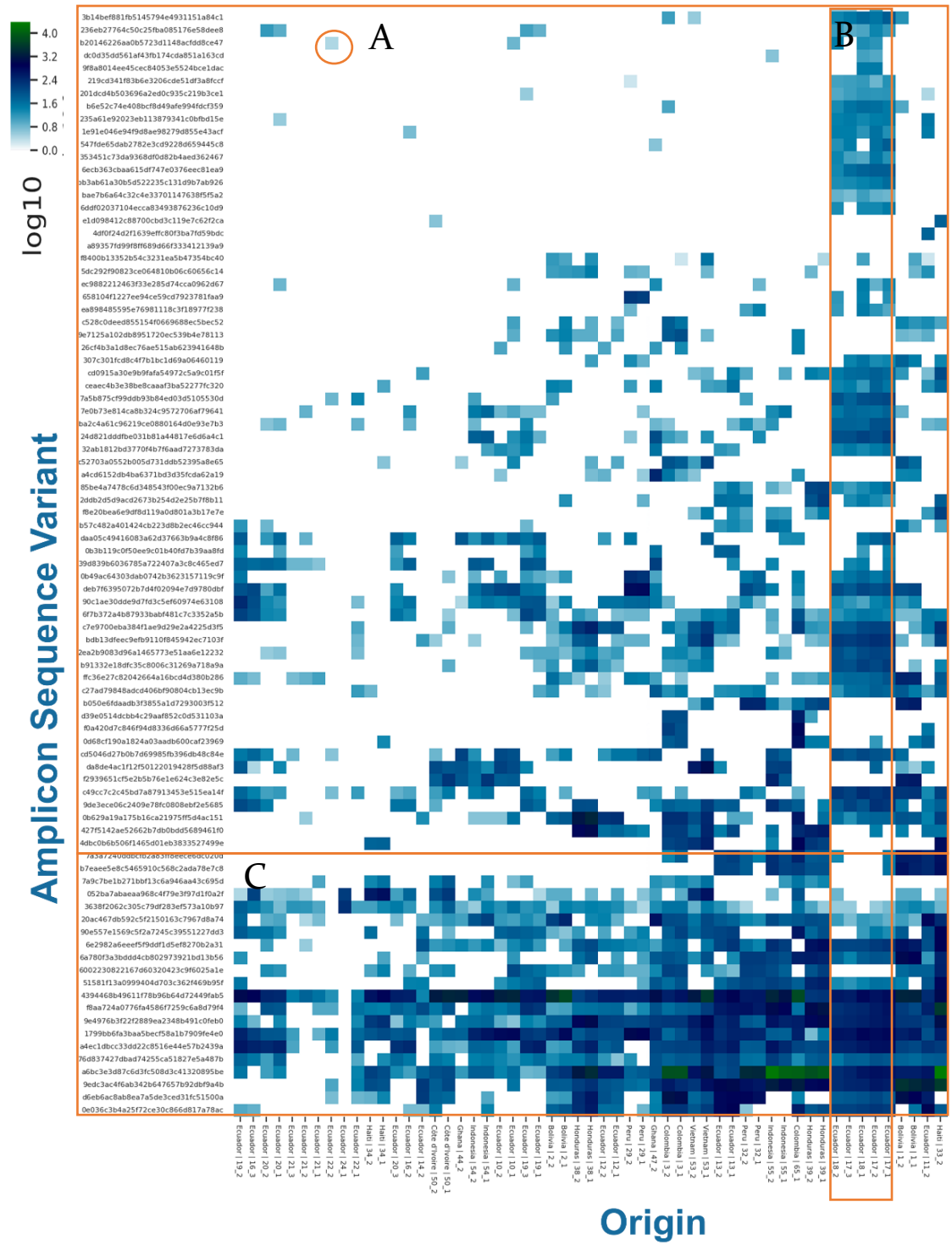


Figure 5.2 Levels of discrimination of the retained 94 samples showing the frequency of 80 ASVs across countries (Available in html file due to size).

Three main patterns were identified which drove further analysis (Dominance, rare species and communality), each cell refers to the -log 10 (feature counts). (A) Isolated heats with light colour intensity showed specific bacterial ASVs (y-axis) assigned to various genus e.g., *Streptomyces* with low-frequency reads (5000) which can point to identifying unique sequences for that region (x-axis) or specific for that sample. (B) High colour intensity determines higher frequency reads (53000) of the feature, in this example, Ecuadorian samples from Esmeraldas region showed a high number of ASVs that are not present in other

e.g. Manabí, Haiti, and Indonesia. (C) Bottom part of the heatmap indicate a pattern according to common bacteria present in fermentation, clustering mainly *Acetobacteraceae* and *Lactobacillaceae* ASVs with high intensity shared across most of the samples as expected.

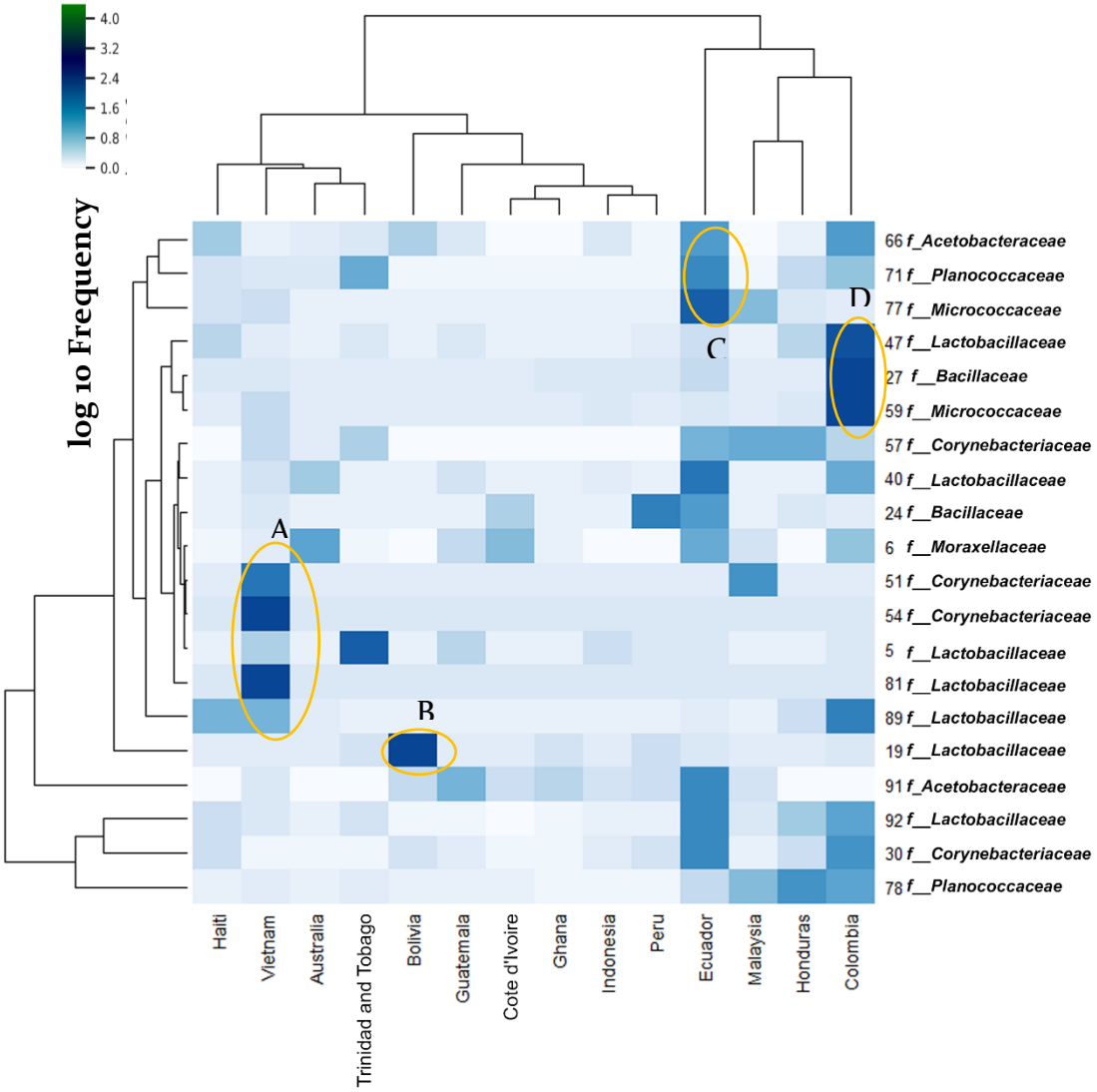


Figure 5.3 Randomized subsample of 20 unique ASVs segmented by country of origin

Heatmap of high frequency of a subsample of 589 filtered ASVs represented with high-intensity colours, each cell refers to the log 10 (Relative abundance). The taxonomic analysis of these ASVs was performed at family level which showed common bacteria from the fermentation process. Some ASV are unique to countries or showed high frequencies. (A) Vietnam, (B) Bolivia, (C) Ecuador, (D) Colombia. These where 35% of sequences comes from Lactobacillaceae, 20% of *Corynebacteriaceae*, 10% for *Acetobacteraceae*, *Micrococcaceae* and *Bacillaceae* and 5% for *Planococcaceae*.

5.4.3 Core microbiome diversity using 16S rRNA v3-v4

Microbiome diversity has been widely described for fermented cacao beans. Here it was important to calculate the alpha (within the sample) and beta (between the samples) diversity metrics to determine if the ASVs identified per fermentation

locations, country and continent were even or dominant in their habitat and if there were similarities between samples according to origin. Eighty-nine samples were retained and assessed for core-metric analysis for alpha (Observed ASVs and Evenness) and beta diversity (Dissimilarities). Samples 14-2, 24-1, 22-1, 21-2, 21-1, 21-3 and 22-2 were excluded as generated less than 1000 reads.

5.4.3.1 Alpha diversity metrics (within groups diversity)

The total number of ASVs per sample was assessed for the presence of contaminants with 5 sequences detected as mitochondrial and chloroplast genomic sequences. A total of 589 unique ASVs were selected for further studies with an average count of 45 per region (P= 0.0014 significant difference, no significance in q value > 0.1) between regions. This accounts for both abundance and evenness of the taxa present in each sample. As a bacterial population/community structure there is a big difference in number of observed ASV at a local scale. For example, Ouanaminthe, Haiti has the lowest number with only 2 ASV while Esmeraldas Ecuador showed the highest count with 87. The frequency of these ASVs was used for further analysis of evenness. Alpha diversity was assessed through observed ASVs and Evenness indexes (Figure 5.4). The presence/absence of taxa (ASV, absolute frequency) were considered for qualitative (observed ASV) analysis and the frequencies accounting each taxon were also indicated in the species evenness plots. There was a higher number of specific ASV observed when looking at the cooperative level. Which appear to average when looking at country and continent as the main bacteria across all samples have higher frequency counts.

The samples were first labeled by their post-harvest location and grouped by origin (fermentation location, country and continent) to perform the analysis. Samples analysed by fermentation location had more significant differences (Kruskal–Wallis) between evenness indexes as it is the analysis of each chocolate extraction individually P= 0.000. Each sample was correlated to the known cooperative (or producer) indicating that there could be higher (same-bacteria dominating the sample) and lower (multiple-bacteria in the sample) indexes according to the fermentation location (Figure 5.4 C). The samples from the same cooperatives showed similar Pielou's index which should indicate common dominant

bacteria during the fermentation. This is important for further analysis as it can be a signature of the location. This had the highest significant differences (*P*= 0.0007) and species diversity evenness across all post-harvest facilities, it can be seen that samples with high indexes have species that are more dominant which can depend on the post-harvest process. As shown in the example from Ecuador (Figure 5.4 A, C), the coloured boxes represent specific cooperatives and the same area of fermentation with similar dominant species evenness (Pink "Camino Verde", Yellow "Esmeraldas", Red "Fortaleza del Valle").

When all the samples from different fermentation locations where grouped together by country and then continent, the index progressively became more similar (*Pielou's* Index 0.5/ 1). There were significant differences in species evenness across all countries *P*>0.005. Indonesia (0.35) had the lowest index which means that various organisms will be dominating the sample and in contrast Honduras (0.68), Australia (0.65), Malaysia (0.60) and Ecuador (0.59) showed to be more even with only some species dominant. As the frequency of the shared sequences started to add up, no significant difference was observed (*P*> 0.098) when samples were grouped and analysed by continent. This was confirmed with the q-values (corrected p-value) all higher than 0.05 (Figure 5.4, A).

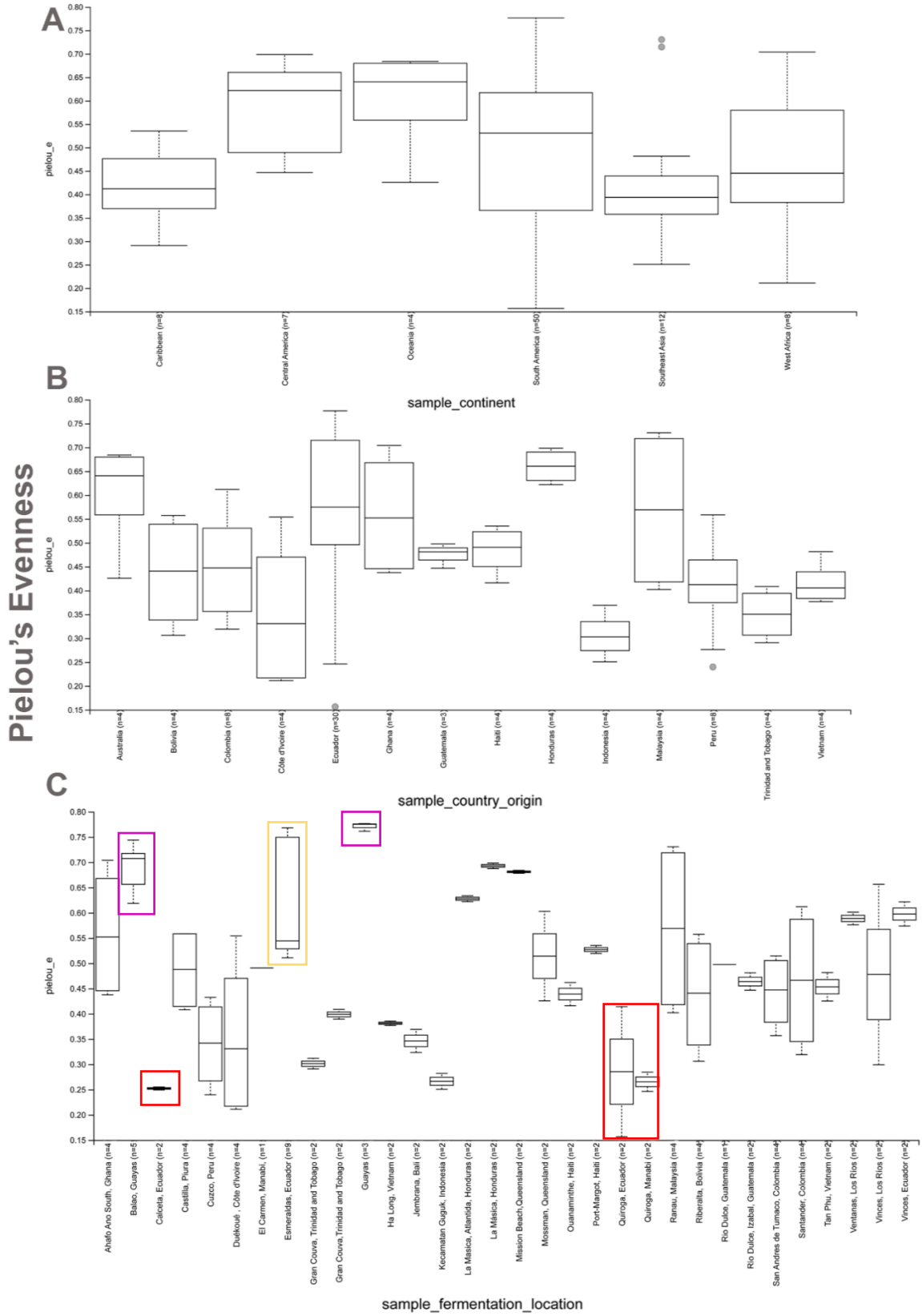


Figure 5.4 *Pielou's* evenness by continent, country and fermentation location

Pielou's Evenness (y-axis) comparison within the sample by location in alphabetical ascending order (x-axis) showed differences in composition and dominance. (A) Comparison by Continent had the lowest species evenness index *P*=0.09, H= 9.2. (B) There were significant differences in species evenness across all countries (*P*=0.00, H = 29.5), (C) The origins

(Fermentation Location) had the highest significant differences in species evenness across all post-harvest facilities, *P*= 0.00, H 62.1 (Kruskal–Wallis). The coloured boxes represent cooperatives and the same area of fermentation with similar dominant species evenness; Pink "Camino Verde", Yellow "Esmeraldas", Red "Fortaleza del Valle".

5.4.3.2 Beta diversity and community comparison between groups

Characterisation of country of origin for chocolate according to microbial community signature

The 89 retained chocolate samples were first analysed to determine if there were significant differences in the microbial communities according to country of origin. A principal component analysis (PCoA) was performed to visualize the dispersion with Bray-Curtis distance matrixes between the ASVs frequency from each sample and significant differences per communities were compared using Unifrac phylogenetic analysis. The PCoA performed after grouping samples by country didn't showed clear separations or clusters, with a total variation of 31.7% explained by the first 2 principal components (Figure 5.5). For instance, samples from Malaysia, Vietnam, Ecuador, Indonesia and Trinidad and Tobago overlapped and in addition some replicates didn't clustered together (e.g. Haiti) but exhibited a pattern of similar distribution. This was indicative of possible structuring at the cooperative level and as a result, a second analysis by individual PCoAs of fermentation locations in specific countries was performed.

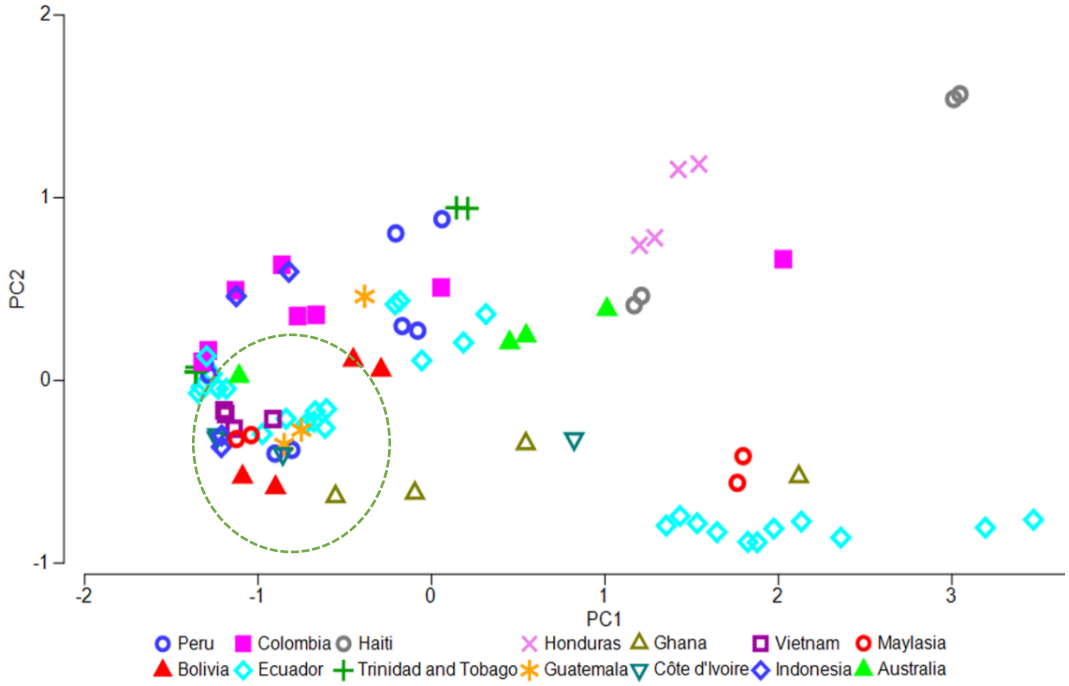


Figure 5.5 Principal Component Analysis of all samples classified by country

Beta diversity measurements for community comparisons using features frequency of each sample grouped by country. The PCoA shows discrimination of 31.7% of communities variation and samples clustering together (dashed circle) and overlapping correspond to samples from Malaysia, Vietnam, Ecuador, Indonesia and Trinidad and Tobago.

5.4.3.2.1 Characterisation of chocolate origin according to microbial community signature at Cooperative level

Specific analysis was performed looking at cooperative in Ecuador, Peru and Ivory Coast.

Ecuador

Bray-Curtis dissimilarity analysis indicated that it is possible to discriminate samples by fermentation location, with beans fermented in the same cooperative sharing the same microbiome (Figure 5.6). Samples 15 (Green upside down Triangle), 9 (Red Rhombus), 13 (Blue ring) from Cocoa of Excellence coming from Quiroga and Calceta, in the province of Manabí, Ecuador which are 13.2 km apart were found to cluster together. After carefully reviewing the traceability of the samples through the documentation of Cocoa of Excellence and International Chocolate Awards 2015 and 2017 report, both samples were found to come from the same cooperative

"Corporation Fortaleza del Valle" (Coop C). This is an important result as it indicates that samples fermented in the same place and in different years could still bear the same microbiome inherited from the fermentation area (Boxes) or surface. All replicate samples grouped together exhibiting similar profiles with the exception of one sample (11) from Vinces Los Rios (Coop E) which produced a very distinct profile more similar to these observed from samples in Coop C (Figure 5.5). Interestingly and assuming that a technical error caused this differences in pattern, the remaining replicate from Vinces Los Rios cluster with another samples from the sample processed in the same Coop E but in a different year confirming what has been observed in Coop C. Samples 20, 16 (Pink Rhombus) coming from Balao and 19 Guayas (dark green upside down triangle) manufactured by TOW cluster together. Both have been fermented with a controlled method by "Camino Verde" in Guayas and in the same facility (Coop A). This indicated that variation in processing following fermentation is not likely to affect the metagenomic profile of the sample.

A third cluster of samples TOW 1 (17) and TOW 2 (18) (Yellow triangles) was identified which comes from a single cooperative in Esmeraldas, Ecuador (Coop D). These samples come from the same fermented beans batch and the same chocolate factory. Their production flow followed the same process until the stage of moulding (bars and drops) when the chocolate was directed to different pipelines. These results indicated that different type of chocolates made from the same beans cluster together and that there are no significant differences when chocolate flow is stopped or redirected to a different pipeline in the factory. Other Ecuadorian samples without a cluster comes from independent cooperatives. The analysis of similarities (ANOSIM) showed that samples from Balao (S) are 82.88% similar between them, Esmeraldas (N), 96.88%, Quiroga (W) 97.74% and Calceta (W) 98.96%, respectively; moreover, products from Calceta vs Quiroga have 94.36% similarity (Figure 5.6).

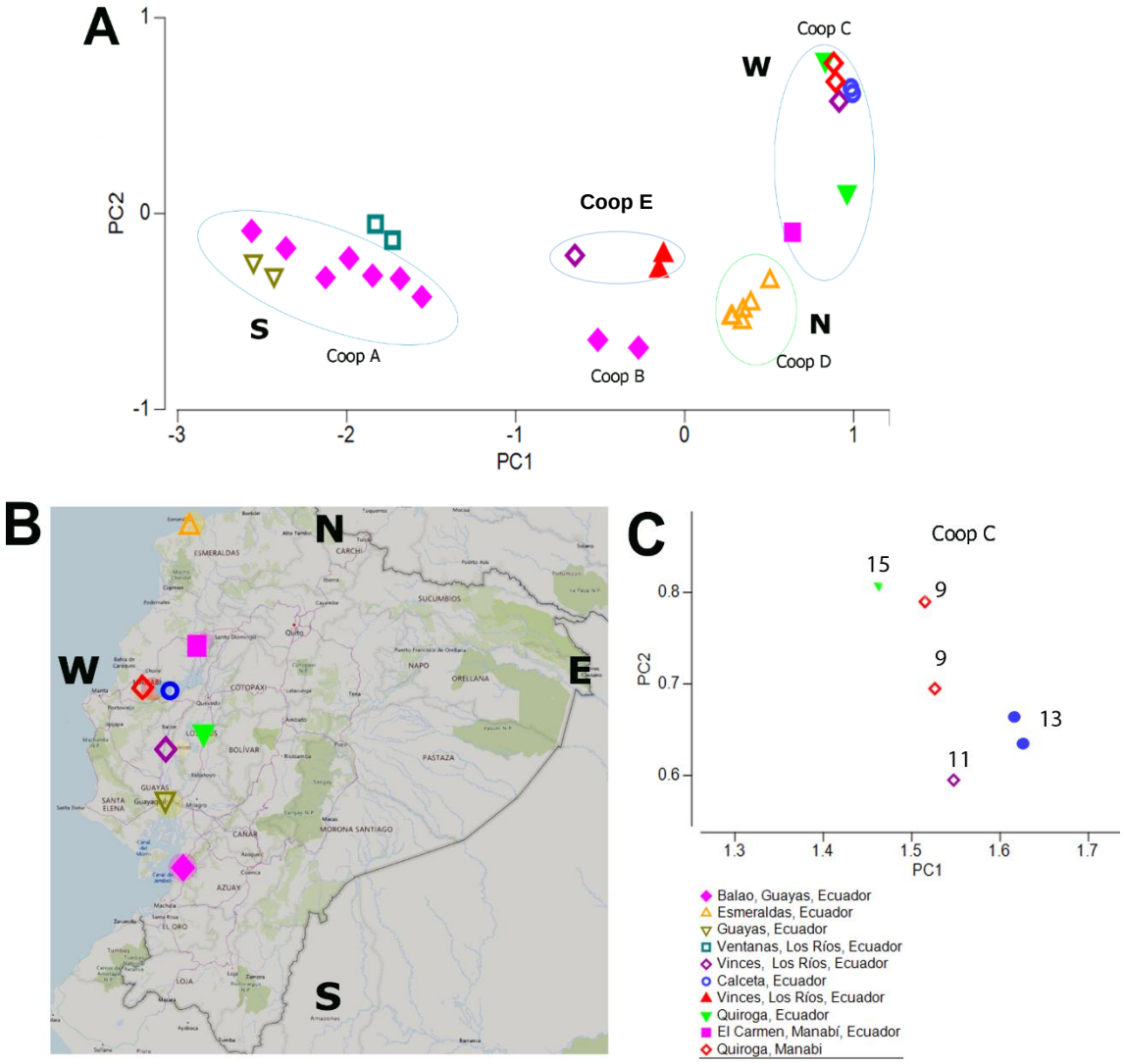


Figure 5.6 Characterisation of chocolate origin according to microbial community signature in Ecuador at Cooperative level Identification of different regions in Ecuador at Cooperative level

Beta diversity measurements for community comparisons in Ecuador using features frequency of each sample grouped by chocolate identity. PCoA shows discrimination (31.8% variation between microbiota components) at the cooperative level in the North (N), south (S) and West (W) of Ecuador, Tests for differences between communities at fermentation location using randomised Pairwise test, R^2_{ANOSIM} of 0.688. Circle indicate cluster for samples from cooperative A, B C and D. (B) Geographical origin of beans for each chocolate sample in Ecuador with colour key symbols for each of the 10 locations. (C) PCoA of samples in the west (W) from 2015 and 2017 clustering together belonging to the same cooperative.

Peru

The same pattern of variation and discrimination between different locations was observed in Peru for samples from premium claimed origins., Three samples from Cocoa of Excellence and a commercial sample were compared (Figure 5.7, A). Samples 27 and 28 were post-harvested at the cooperative "Agraria Nor Andino" (Coop A) and

clustered together with 80% of similarity between their microbiome. The two other samples were from the region of Cuzco with sample 29 from the cooperative "Asocasel" and sample 32 (QRN CSC) from the cooperative QARANA-Echeria. These are 22KM apart and showed 81.6% of similarity but did not cluster together. When comparing the two region in Peru, there was an overall dissimilarity of 40.73% SIMPER, between Piura and Cuzco.

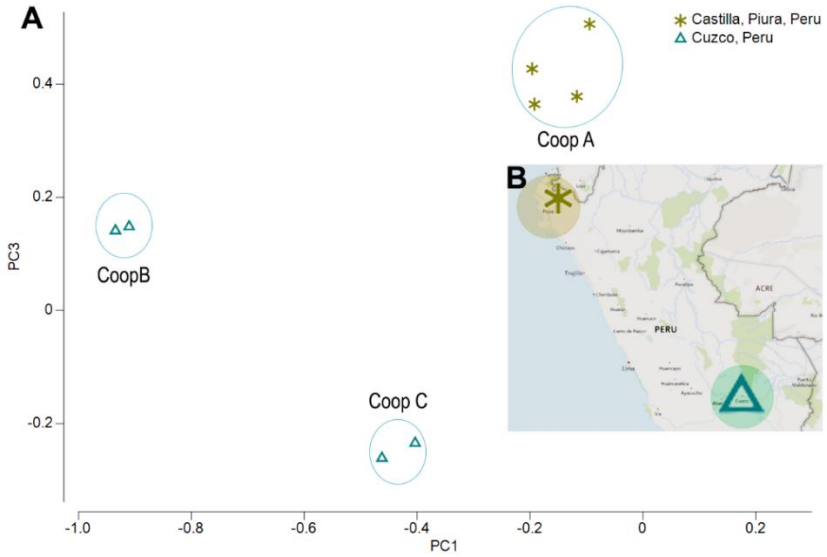


Figure 5.7 Characterisation of chocolate origin according to microbial community signature in Peru at Cooperative level identification of different regions in Peru at Cooperative level

(A) Beta diversity measurements for community comparisons in Peru using features frequency of each sample grouped by chocolate identity. PCoA showing discrimination (57.3% variation between microbiome communities using Bray-Curtis Dissimilarity) between 3 cooperatives A, B and C, in two regions from Peru (North and South), R^2ANOSIM of 0.667. (B) Map and geographical location of cooperative in the Piura, North (Asterix) and Cuzco, South (Triangle).

Ivory Coast

When assessing samples from Ivory Coast, a similar pattern of clustering was observed according to fermentation location. While no comparison was possible between different samples from the same cooperative these samples originated from specific single origin producers in Duekoue and not from bulk mixes gathered across the whole country. This is an important discovery since cooperatives which are in protected areas can potentially have their own signature bacterial profile. Cooperative A, (DISSA NAFON KARIM / Cooperative CAVAZA) has post-harvested sample 50.

Samples 48 (PIHI Lambert) showed 51.44% of dissimilarity between them. There is an overall dissimilarity of 48.66% SIMPER, between A and B (Figure 5.8).

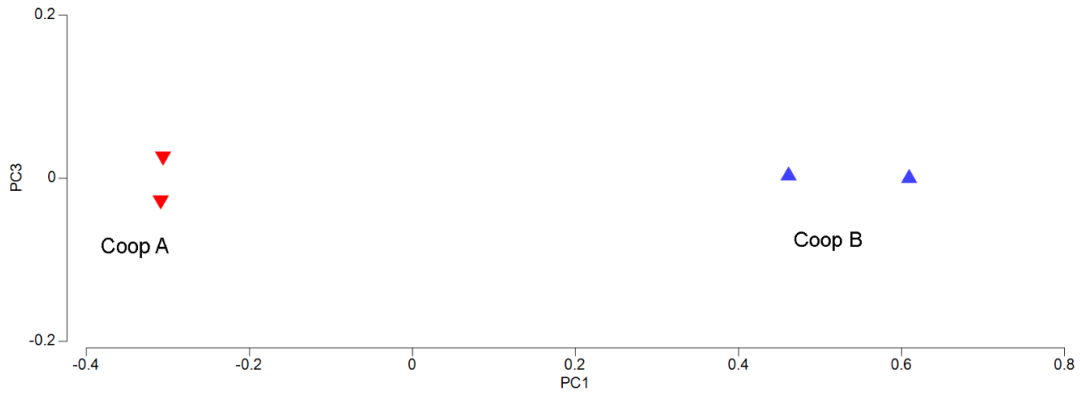


Figure 5.8 Identification of different regions in Duekoue Ivory Coast at Cooperative level

Samples from Ivory Coast: PCoA showing discrimination (86% variation between microbiome communities) between 2 cooperatives (A) and (B) in the same region of Duekoue, Guemon using Bray-Curtis dissimilarity.

5.4.3.3 *A. pasteurianus* diversity analysis using housekeeping genes (HKG)

All samples were analysed by fermentation locations with duplicate or triplicates for each accessions. Sequences generated from all chocolate samples only included *Acetobacter pasteurianus* sequences confirming the high specificity of the primers designed for the study. The three housekeeping genes were specifically screened in samples from Ecuador where more samples were available for comparisons. Clear clusters were identified reflecting a similar patterns previously identified by screening 16v3-v4 but also with additional subgrouping possibly indicative of field origin. This highlights the importance of targeting the main bacteria of fermentation for diversity assessment to determine chocolate origin. Duplicate samples exhibited consistently similar patterns and clustered together. This was for instance even observed for samples from one sample from Vinces, Los Rios which when screened with 16s presented very different pattern between replicates. This suggest that one of the sample from Vinces, Los Rios screened with 16s might actually not been from this region as it produced a very distinct profile more similar to these observed from samples in cooperative C.

5.4.3.3.1 Characterisation of chocolate origin according to *A. pasteurianus* population signature in Ecuador at Cooperative level

***rpoB* encoding the b-subunit of bacterial RNA polymerase**

Bray-Curtis dissimilarity analysis of *rpoB* indicated that it is possible to discriminate samples by fermentation location, with beans fermented in the same cooperative sharing the same microbiome. The results observed mirrored the grouping founds when using 16S. For instance, samples from Quiroga from 2015 and 2017, which were processed in the same cooperative "Corporation Fortaleza del Valle" (Coop C) clustered together. This confirmed that samples have been fermented in the same place and in different years with the same strain of *Acetobacter pasteurianus* inherited from the fermentation area (boxes) or surface. Samples TOW 1 (17) and TOW 2 (18) (Yellow triangles) which originate from the same batch of chocolate and a single cooperative in Esmeraldas (Coop D) but followed a different process of manufacturing still clustered together.

Samples from Balao and Guayas made by TOW cluster together showing the same origin to cooperative A. For this analysis with *rpoB*, a sufficient number of sequences were obtained for the cocoa butter (CB) sample Tow 13 (24) which indicated that Bacteria involved in the fermentation process can be detected in CB (Figure 5.8 A). The duplicate reactions for CB group separately at the top of the graph (brown dash circle). While this sample comes from the same region in Guayas and cooperative A as the cocoa nibs Tow 4 (19) and chocolate samples Tow 7 (21), they do not share the same profile. It is possible that the beans used to produce the CB were not fermented in the same way or not in the same batch. But this is more likely to reflect the slightly differential amplification of *A. pasteurianus* strains present in the samples as previously noted in the raw data generated from 16s and described in Chapter 3. Both cocoa nibs and chocolate (Tow samples) clustered together, showing that even when the beans are roasted the same microbiome is present.

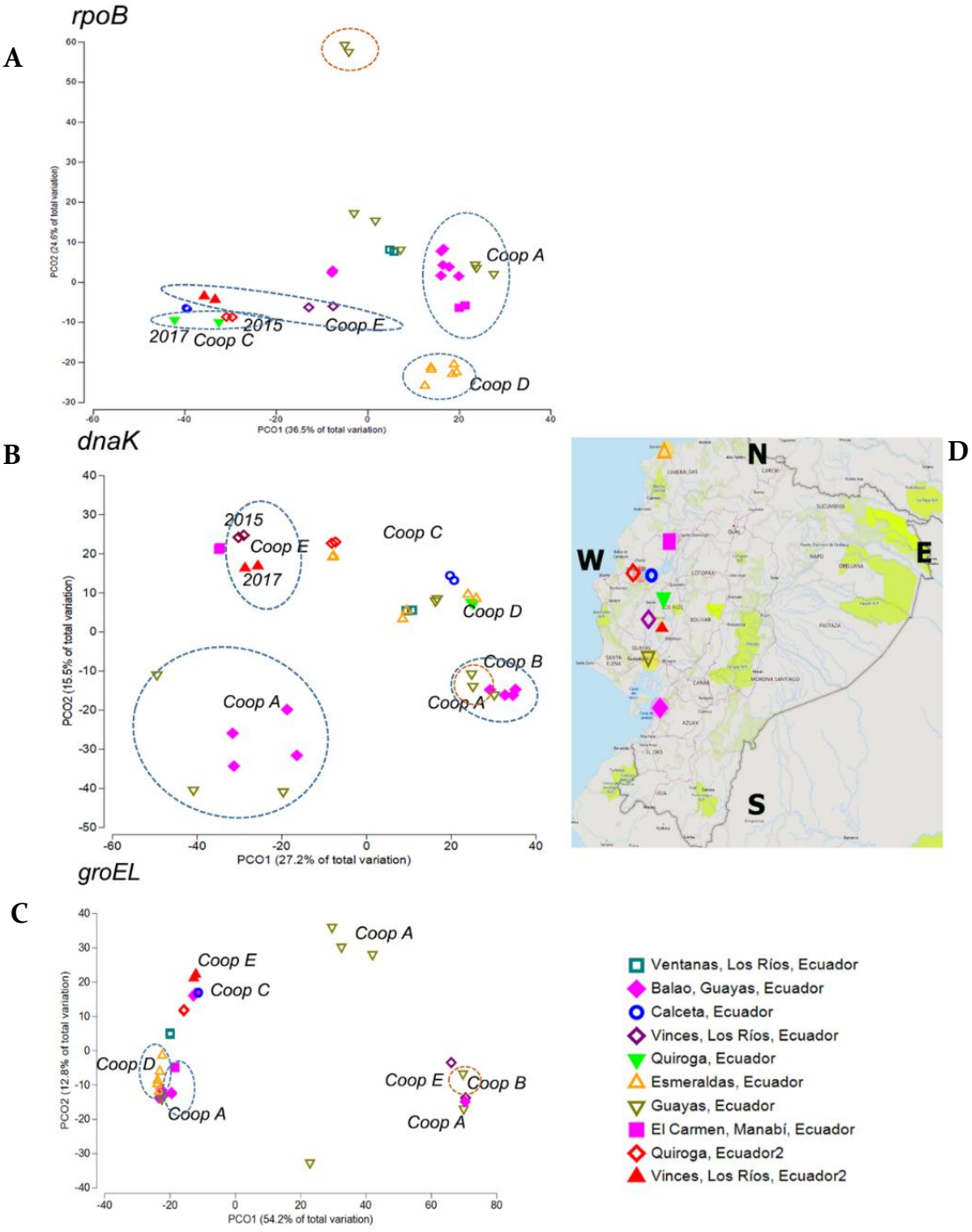


Figure 5.9 Characterisation of chocolate samples according to *A. pasteurianus rpoB*, *dnaK* and *groEL* ASV driving discrimination of different regions in Ecuador at Cooperative level

Principal Coordinate Analysis of ASVs diversity in chocolate samples from Ecuador using Bray-Curtis distance matrix in *A. pasteurianus* HKG. (A) PCoA for rpoB. (B) PCoA for dnaK (C) PCoA for groEL (D) Map for the sample origin in Ecuador. Clustering of samples according to Coop is indicated with a dash circle. Position of coca butter samples indicated with brown dashed circle.

dnaK encoding a heat-shock protein

The analysis of chocolate accessions using *dnaK* did not exactly match the pattern of clustering as observed in *rpoB* and 16s but enable the characterisation of different grouping that were not observed so strongly previously. For instance samples from "Vinces Los Rios" both produced from cooperative E did group together. One of the replicates correspond to a sample which is likely to have been wrongly labelled in the analysis of 16s. These samples are interesting as they were produced in two different years. Indeed one of the replicate does clustered in 16s with the other two samples from the same cooperative. This confirm that the pattern produced in cooperatives can be maintained across years not only when looking at broad bacterial population, but also specific species involved in the fermentation process (Figure 5.9 B). Samples processed in Cooperative A did separate in two distinct clusters but did not separate according to chocolate origin which are either from TOW or CoEX. It is possible that *dnaK* highlights the presence of unique bacteria from specific locations in the region of Guayas. Interestingly, and possibly supporting this idea, the cocoa butter (dashed brown circle) clustered with some of samples from cooperative A. Samples from cooperative D also group together though clustering with additional samples from other origins. This pattern might also be seen in samples from Cooperative C with still evidence of a large cluster but split in three distinct groups (Figure 5.9, B).

groEL encoding a chaperonin protein

The analysis of chocolate accessions using groEL provided a similar results cluster to these observed with all other genes when looking at Samples TOW 1 (17) and TOW 2 (18) (Yellow triangles) which originate from the same batch of chocolate and a single cooperative in Esmeraldas (Coop D) (Figure 5.9, C). Samples processed in Cooperative A did again very distinct groups with the cocoa butter again clustering together with samples from Balao, Guayas Ecuador. Whilst the cocoa butter was directly provided by TOW (Fermented by Camino Verde Cooperative) the sample clustering with it (Pink rhombus) comes from the same cooperative but was provided by CoEX. Also, the cocoa butter sample was produced in 2017 whilst the chocolate sample was process in 2015. This is another indication that DNA obtained from a

range of chocolate products (nibs, butter or chocolate) from different years can generate similar profiles and cluster together by cooperative. Conversely, samples from "Vinces Los Rios" both produced from cooperative E but in different year did not clustered when using *groEL* as seen in the three previous genes. This might suggest that in this instance, *groEL* actually highlight differences between years of fermentation. Finally, for *groEL*, the sample from Quiroga, Ecuador (green triangle) did not amplified in both samples and was therefore eliminated from the analysis (Figure 5.9, C).

5.4.4 Identifying unique Biomarkers to predict the origin of chocolate samples

The Illumina dataset was assessed to identify potential unique ASV characteristics of specific locations such as field (single source chocolate accession), country and continent. These markers of origin were selected after screening for unique sequences (presence/absence) present in all samples belonging to the same geographical group. These unique markers per sites could potentially indicate a signature bacterial profile that could be utilise individually or in combination to identify a specific location. The analysis was done independently for 16s and the three Housekeeping genes.

Table of ASV distribution across all genes showing a large number of unique markers observed with different level of frequency according to the geographical screening performed. At the sample level, 212 unique ASV were identified across all 4 genes, this increasing to 380 for country specificity country level and up to 493 at the continent level. Differences were observed when comparing 16s to *HKG*. At the sample level, 191 16s ASV (44.7% of all ASV) were unique and specific to samples. The value observed was much lower in the housekeeping genes with no specific ASV detected in *groEL*, 8 in *rpoB* (4% of all ASVV) and 13 (9.4% of all ASV) in *dnaK*. These value change dramatically when looking at ASV specific to countries and continent. While the total number of specific ASV only increase marginally when looking at 16s from 191 to 215 per countries and 246 per continent (44.75, 50.4% and 57.6% respectively), the increase was more significant with the house keeping genes with for example *groEL* increase from 0 ASV to 41 and 65 (0%, 32% and 50.8% increase respectively). These differences reflect the screening performed with 16s identifying

different bacterial species and therefore more likely to characterise unique profile from sites. Conversely, since the housekeeping gene are specific to one of the key bacteria involved in the fermentation process, similar strain of *A. pasteurianu*s are more likely to be present in locations geographically closer to each other which could explain the characterisation of specific ASV per country and continents.

It is worth noting the large differences in number of ASV observed across samples specifically when looking at ASV generated from 16s. A larger number of ASV were observed in samples from Ecuador and Peru (18, 27) and also samples from Malaysia (13 and 25). This might reflect the method of fermentation for these samples with a higher microbial diversity presents at sites in South America but also possibly the length of fermentation in Malaysia. Importantly the large number of ASV found for Ecuador (139 across the four genes) reflect the larger number of samples screened for this country compared to all other countries with for instance only 4 and 3 ASV characterised for Ghana and Ivory Coast. This indicates that the screening of a larger number of locations per country is crucial for a wider characterisation of specific locations.

Table 5.2 Number of unique biomarkers per single origin using 16S v3-v4 gene and the three *A. pasteurianus* HKG *rpoB, dnaK and groEL*

ASV specific to 44 samples, 14 countries and 5 continents are indicated according to 16s and three HKG from *A. pasteurianus*. Total number of ASVs are indicated by gene and sample, country and continent with the proportion of specific ASV indicated against the total number of ASV generated per gene. Total number of sequences generated per gene, total ASV number and average sequence number per ASV indicated as reference.

		Gene			
		Universal	*A. pasteurianus* HKG		
Geographical location	ID	16s	*rpoB*	*groEL*	*dnaK*
Ventanas, Los Ríos, Ecuador	10	1	0	0	0
Vinces, Los Ríos, Ecuador	14	1	0	0	0
Vinces, Los Ríos, Ecuador	11	0	0	0	0
Balao, Guayas, Ecuador	12	0	0	0	0
Balao, Guayas, Ecuador	16	0	1	0	0
Balao, Guayas, Ecuador	20	1	3	0	6
Guayas, Ecuador	21	3	0	0	1
Guayas, Ecuador	19	4	0	0	4
El Carmen, Manabí, Ecuador	22	0	0	0	0
Quiroga, Ecuador	15	0	0	0	0
Calceta, Ecuador	13	3	0	0	0
Esmeraldas, Ecuador	17	5	2	0	2
Esmeraldas, Ecuador	18	15	1	0	0
Riberalta, Bolivia	1	4	0	0	0
Riberalta, Bolivia	2	2	0	0	0
Cuzco, Peru	32	9	0	0	0
Cuzco, Peru	29	8	0	0	0
Castilla, Piura, Peru	27	16	0	0	0
Castilla, Piura, Peru	28	9	0	0	0
Santander, Colombia	65	0	0	0	0
Santander, Colombia	67	2	0	0	0
San Andres de Tumaco, Colombi	8	10	0	0	0
San Andres de Tumaco, Colombi	3	2	0	0	0
San Andres de Tumaco, Colombi	9	3	0	0	0
Port-Margot, Haiti	33	7	0	0	0
Ouanaminthe, Haiti	34	5	0	0	0
Gran Couva,Trinidad and Tobagc	35	1	0	0	0
Gran Couva,Trinidad and Tobagc	36	3	0	0	0
La Másica, Honduras	38	2	0	0	0
La Masica, Atlantida, Honduras	39	3	0	0	0
Río Dulce, Izabal, Guatemala	42	na	0	0	0
Río Dulce, Izabal, Guatemala	43	4	1	0	0
Ahafo Ano South, Ghana	44	2	0	0	0
Ahafo Ano South, Ghana	47	2	0	0	0
Duékoué , Côte d'Ivoire	48	0	0	0	0
Duékoué , Côte d'Ivoire	50	3	0	0	0
Ha Long, Vietnam	52	4	0	0	0
Tan Phu, Vietnam	53	3	0	0	0
Kecamatan Guguk, Indonesia	54	5	0	0	0
Jembrana, Bali	55	6	0	0	0
Ranau, Malaysia	56	13	0	0	0
Ranau, Malaysia	57	25	0	0	0
Mossman, Queensland	58	0	0	0	0
Mission Beach,Queensland	60	5	0	0	0
Total		*191*	*8*	*0*	*13*
% unique ASV		*44.7*	*4.0*	*0.0*	*9.4*

	Gene			
	Universal	*A. pasteurianus* HKG		
Country	16s	*rpoB*	*groEL*	*dnaK*
Ecuador	51	36	17	35
Bolivia	6	2	1	2
Peru	46	8	5	10
Colombia	19	1	1	0
Haiti	12	0	0	0
Trinidad and Tobago	4	0	0	0
Honduras	5	0	0	0
Guatemala	4	7	5	2
Ghana	4	0	0	0
Côte d'Ivoire	3	0	2	0
Vietnam	7	8	4	4
Indonesia	11	2	1	1
Malaysia	38	2	4	2
Australia	5	2	1	0
Total	*215*	*68*	*41*	*56*
% unique ASV	*50.4*	*34.3*	*32.0*	*40.3*

	Gene			
	Universal	*A. pasteurianus* HKG		
Continent	16s	*rpoB*	*groEL*	*dnaK*
South America	150	73	39	59
Central America	26	8	8	3
West Africa	7	0	2	0
South East Asia	58	21	15	16
Oceania	5	2	1	0
Total	*246*	*104*	*65*	*78*
% unique ASV	*57.6*	*52.5*	*50.8*	*56.1*

	Gene			
	Universal	*A. pasteurianus* HKG		
	16s	*rpoB*	*groEL*	*dnaK*
Total sequences	2404840	2301195	2095636	1076592
Total ASV	427	198	128	139
Average seq per ASV	5631.944	11622.2	16372.16	7745.266

While 16s generated a larger number of sample specific ASV, the majority of these (60.2%) were represented by less than 100 sequences from a total of 2,404,840 sequences with only 3.3% found with more than 1000 sequences. In contrast *groEL* presented 24.6% of ASV with less than 100 sequences but 29.2% with more than 1000 sequences from a total of 2,095,636. A similar pattern was observed for the remaining two HKG. This difference is partly due to the higher number of ASV detected in 16s (427) compared to *groEL* (128) for an overall similar total number of sequences

generated. Nonetheless, it is important to consider the depth of sequences available for ASV to assess the reliability of the markers to characterise location. This could be done by resequencing or quantifying de novo specific sequences of interest. Importantly, from the present pilot study, 50 specific ASV were detected across all samples which generated more than 1000 sequences which suggest that a more in depth study should generate additional markers.

Table 5.3 Number of sequences per specific ASV generated from 16S v3-v4 gene and the three *A. pasteurianus* HKG *rpoB, dnaK and groEL*

The number of sequences per ASV indicated in distinct group ranging from >1000, 99-500, 499-250, 250-100 and <100. ASV fitting in each category indicated per gene with the proportion for each category specified. Total unique ASV, Total ASV, Total sequence per gene and Average sequence number per gene indicated.

	Gene Universal		*A. pasteurianus* HKG					
Seq per unique ASV	16s	%	*rpoB*	%	*groEL*	%	*dnaK*	%
>1000	8	3.3	14.0	13.5	19	29.2	9	11.5
99-500	16	6.5	12.0	11.5	8	12.3	13	16.7
499-250	19	7.7	9.0	8.7	13	20.0	6	7.7
250-100	55	22.4	24.0	23.1	9	13.8	20	25.6
<100	148	60.2	45.0	43.3	16	24.6	30	38.5
Total Unique ASV	246		104		65		78	
Total ASV	427		198		128		139	
Total sequences per gene	2404840		2301195		2095636		1076592	
Average seq per ASV	5631.944		11622.2		16372.16		7745.266	

5.5 Discussion

In this chapter, Amplicon Sequence Variant (ASV) specific to the cacao bean fermentation microbiota were identified in a range of cacao products. While metagenomics analysis using next generation sequencing had been performed previously from fermented beans (Ardhana and Fleet, 2003; Camu, Bernaert and Lohmueller, 2010) this is the first time it has been achieved in a range of cocoa products including chocolate, nibs and cocoa butter. Confirmation that metagenomic signature from cacao fermentation can be detected in chocolate products was performed using not only the universal 16s ribosomal assay but also three separate genes specific to one of the key bacteria involved in the fermentation of coca beans, *A. pasteurianus*. The analysis of the profile generated from all genes screened has revealed that it is possible to group chocolate samples according to the cooperative where the raw material has been processed with strong evidence that it

is reproducible even across years of production. Overall, ASV specific to samples, countries and continents have been identified demonstrating that metagenomic screening has a potential for generating biomarkers that could be used for chocolate tracking.

While DNA has been extracted previously from chocolate, most markers used have been related to the cacao plant genome (Petiard, no date; Gryson, Messens and Dewettinck, 2004) and it was not clear if any DNA might remain following the processing steps to identify the microbiota. Taxonomic assignments analysis using the universal 16s gene region v3-v4 revealed that high frequency ASV detected belonged mainly from family of bacteria associated to the fermentation including *Acetobacteraceae, Lactobacillaceae, Corynebacteriaceae, Bacillaceae, Micrococaceae*, which confirms that it is possible to identify microbiome from the fermentation stage in chocolate DNA. The microbiota is thought to be mainly associated with the shells of the cacao beans, but the present work demonstrate that its DNA can be detected in single origin chocolate product but also in highly process derivative cocoa product such as cacao butter.

Alpha diversity analysis revealed large differences in the number of ASV (3-87) detected across all chocolate samples. This measure can be useful for preliminary identification of target regions that might contain diagnostics ASV. But it is important to assess the results according to the depth of sequencing for each sample analysed, which in this study was low. In microbiome analysis, biological replicates yield a different number of reads, different community compositions and different levels of diversity, which gives a level of uncertainty about its real composition. A study by (Rivera-Pinto *et al.*, 2018; Willis, 2017), showed that the statistics of microbiome are relative, estimates of many alpha diversity indices (Richness - Shannon diversity) are negatively biased for the environment alpha diversity parameters, that is, they underestimate the true alpha diversity when the environment is not sampled exhaustively (Willis, 2017). Phylogenetic Diversity index increases with sampling size and sample completeness, environments can be identical with respect to one alpha diversity metric, but the different abundance structures will induce different biases when samples or sample size have been adjusted using

methods as rarefaction. The unique property of microbiome experiments and alpha diversity analysis is that each sample does not faithfully represent the entire microbial community under study, thus it helps to identify and corroborate the data and downstream analysis from beta diversity insights, (Willis, 2019; Crits-Christoph *et al.*, 2013).

Beta diversity was assessed using ASV frequencies by calculating (Bray-Curtis and Unifraq metrics) the number of species that are dissimilar in two different environments. Bray-Curtis distance metrics measures the overall change in the whole community, considering the distribution of species as a whole, without using phylogenetic patterns. With this assessment it can be seen that there is high diversity between fermentation locations and there is high bacterial discrimination between each sample. Even with the depth of sequencing performed in this study, the results from evenness and beta diversity confirmed possible countries with a high diversity of species that are known to have standardized fermentation techniques such as Ecuador, Colombia, and Honduras while countries in Asia or Africa with lower diversity are known for not doing box fermentation or in some cases none supervised fermentation. This could mean that the post-harvest process can help to develop different microbial dynamics with a broader variety of species than non-fermented beans.

Some samples from very different regions were also overlapping which could be explained by the detection of the same approximate reads counts for a particular ASV. The calculation of Bray-Curtis dissimilarities for geographical groups (Fermentation Location, Country and Continent) showed some biased results as the samples does not have the same amount of observation or size. The analysis between two sites (e.g. Ecuador n=36 vs Australia n =4) assume that both sites are the same size, either in area, number of samples per origin or as in relevance to species counts. This bias happened in microbiome analysis as the equation of Bray Curtis doesn't include any notion of space and only included feature counts. The set was then assessed according to regions within a single country focusing on the location were fermentation would have occurred which would be often in a cooperative. Fermentation protocols are different between each producer were the length of the process (days), the timing

(hours), season (dry/wet), method for the bean mixture and environmental conditions give variations in chocolate flavour and colour of the beans, this is also reflected in the microbiome composition as every sample has a fairly different microbial ecology and proportion. Cacao bean producers and companies have invested various efforts to obtain the core microbiome (bacteria and fungi) to homogenise the fermentation process and also to reproduce dominant and rare species as starter cultures to make this process replicable (Lefeber *et al.*, 2011b, 2012). This process is highly relevant for the industry in different aspects: first regulating the fermentation can develop multiple combination of flavours and colour profiles. The right use of these starter cultures can also optimize the time that takes to ferment properly the beans and develop further attributes.

For instance, a method registered in 2010 by Barry Callebaut AG involved the addition to the beans of a microbial composition including *Lactobacillus plantarum, Lactobacillus fermentum, Acetobacter pasteurianus, Lactobacillus parafarraginis,* and at least one strain of a yeast species of a genus selected from the group consisting of *Saccharomyces* and *Candida"* (Camu, Bernaert and Lohmueller, 2010). The approach developed here could use markers from the microbiome spontaneous fermentation markers of specific localities combined with the starter cultures from these companies as a new unique target for improving traceability systems. A case study was specifically performed on Ecuador with 15 samples originating from 7 fermentation locations and better discriminations was observed between these samples. Pielou's index was similar between samples from the same fermentation area suggesting common dominant bacteria for these sites and this similarity was confirmed through Beta diversity analysis. Cooperative A has a standardized fermentation method with starter-cultures (unknown mix) which can be the main reason for the clustering of samples from Balao and Guayas, and that could potentially help to determine origin authentication (Böhme *et al.*, 2019). Importantly, samples in Ecuador produced in the same cooperative but in different years exhibited similar profiles and this was confirmed in both Coop C and Coop E. This indicates that fermentation signatures can be maintained across year though it would need to be confirmed and assessed to check if patterns are similar even at different time of harvest within the same year. The results obtained from chocolate samples in West

Africa is also interesting since it indicates that while a low genetic variation observed in the crop (Belsky *et al.*, 2014) might be problematic to implement DNA tracking of products, there is evidence of sufficient variability in the fermentation metagenome to be able to identify single origin cocoa from Ghana or Ivory Coast.

By screening the 16S gene, the biomarkers detected corresponded to a range of microbes related to the fermentation process itself. Common to all fermentation location, continent and country, acetic acid bacteria (AAB) belonging to the family *Acetobacteraceae* were highly abundant. It can be assumed that every batch of cacao beans reached the alcoholic stage and therefore were fermented under local protocols. The amount of (AAB) could explain that the fermentation was accomplished as these species belongs to the genera *Acetobacter, Gluconacetobacter, Gluconobacter* and have high capacity to oxidise ethanol to ethanoic and acetic acid with high resistance to acetic acid released into the fermentative medium (Reis *et al.*, 2012). Sometimes more specific bacteria were detected. For instance, ASV from *Streptomyces* genus were found to be present only in samples from Ecuador made by Tree of Wisdom Chocolate Ltd (TOW). *Streptomyces* is predominantly found in soil and are important for initial decomposition of organic material, they produce spores which release an "earthy-aroma" that is the results from the production of a volatile metabolite called geosmin. These species can move only when grown in the presence of fungi, especially yeast which could explain the type of fermentation and storage that the beans had previously to manufacture (Jones *et al.*, 2017; Chater *et al.*, 2010).

Conversely, all samples showed the presence of the species *Pediococcus parvulus* from the family *Lactobacillaceae* with high proportions of 45.29%. *P. parvulus* has been isolated from various fermented foods (cider, wine and kimchi). *Leuconostoc fallax* strains were found in samples from Malaysia, Colombia and Ecuador, this species is usually found in hetero-fermentative stages from Sauerkraut but some of the LAB strains are also responsible for the initial acid fermentation which might be similar in various spontaneous fermentation (Barrangou *et al.*, 2002). Finally, the analysis of taxonomic assignments and profiling shows not only the microbiome of the region but also everything involved in the manipulation of the

product from bean to chocolate and many rare ASVs that can give false positive discrimination of the samples. These might appear to be unique markers for a cooperative or regions of post-harvest but habitually exhibited low sequence number and might not be reproducible. Often, they can be related to pathogenic bacteria including *Bordetella ansorpii*, *Corynebacterium pilosum* and *Rickettsia endosymbiont of Deronectes platynotus.* These bacteria has been reported as a cause of human and animal disease (Küchler, Kehl and Dettner, 2009) were *C. pilosum* has been shown that can be frequently found on the skin and in the upper respiratory and gastrointestinal tracts, a fact that can be linked to the manipulation of the cacao product in farm, factory of laboratory (Yanagawa and Honda, 1978).

Pattern generated by the infrequent presence of such microbes is likely to be not desirable in the identification of stable profiles for sites. For this reason, housekeeping genes (HKG) from one of the key bacteria of the fermentation process, *A.pasteurianus* were assessed to evaluate their ability to generate diagnostic biomarkers. The assays were confirmed to be specific generating only amplified sequences from *A. pasteurianus*. The analysis of these genes focused specifically on the Ecuadorian samples and compared the clustering of the samples according to cooperatives. Identical grouping to these observed ASV with the gene 16s were found when looking at the gene *rpoB*. This indicate that a sufficient level of sequence variation can be found within the HKG to produce specific pattern for fermentation sites. When using dnaK and groEL, most of the main groupings like Esmeraldas (Coop D) were again observed but some additional clustering were generated within cooperatives. These might well be corresponding to the variability of microbes observed between field sites and present on the surfaces of the pods. Interestingly, because the PCR assay target specifically *A. pasteurianus*, it was possible to generate sufficient sequence numbers to analyse the cocoa butter (CB) while was not possible with 16s. While analysis with rpoB provided a unique separate cluster for CB, analysis with the remaining two genes did produce a cluster with samples processed in the same cooperative A. While it is always going to be difficult to detect a marker specific to CB within a mix samples, this result demonstrates that ASV from HKG can be found for CB.

Overall the ASV number generated by HKG was lower than 16s for a similar number of total sequences produced. Since HKG only target one species of bacteria, and these genes are conserved it is logical to obtain a reduced number of ASV per sites. When fermentation is established, only a certain number of *Acetobacter* strains are likely to be present and the PCR process will detect the most common one only. While diverse microbial population has been detected using 16s when looking at different sites, the results for the house keeping genes differ. Hardly any ASV were found specific to sites but a much larger number were observed diagnostic for countries. Similar *A. pasteurianus* strains are likely to be found in sites with close proximity, for instance within countries, but more differences should appear as the geographical distance increase. This is what is observed here with a similar proportion of ASV specific to countries and continent found when using 16s or HKG. This is important as it demonstrates that a range of markers, targeting specific organism always found in the fermentation of cacao beans, can be utilised to generate diagnostics markers. While 16s might produce ASV characteristic of unique bacterial species within sites, HKG can produce more robust markers even if is only specific to larger regions of production. Similar markers could be generated from other bacteria present during including *Lactobacillus fermentum*. Instead of analysing bacteria, yeast population could be assessed either by using universal 18s ribosomal fungal markers (Klindworth *et al.*, 2013) or specific genes targeting cacao fermentation yeast. To generate similar pattern of local specificity obtained with 16s but without using markers related to bacterial not involved in the fermentation process, strategies of marker screening could involve targeting different genome regions more variable than HKG. For instance, a more refine analysis of *A. pasteurianus*, highly variable genome regions of the genome could be performed with for example the screening of transposon insertion sites. Metagenomic studies of coca fermentation profile has revealed the presence of bacteriophage specific to *A. pasteurianus* and these could also be used to potentially generate unique site specific markers (Illeghems *et al.*, 2012).

5.6 Conclusion

The approach presented here demonstrates that it is possible to build an array of diagnostic markers specific to fermentation sites. This is only a preliminary study and it is evident that care must be taken in assessing the validity of all marker identified with many 16s specific ASV only characterised by a few sequences. This would go through a preselection approach to identify potential useful markers which could be targeted for different scale of study (cooperatives, regions, country and continent) and a secondary higher depth screening to verify the uniqueness of each ASV. Confirmed markers could be then included in database to be used for further characterisation and tracking of chocolate products. These unique ASVs per origin could be defined by (brand, post-harvest facility, country and continent) and be used to determine inclusion or exclusion of particular groups in a controlled sample set. In order to be able to compare a random sample in a trained data set, there is a need of implementing in-depth supervised machine learning, which can handle a large number of labelled independent variables and this system should be able to identify the best predictive model to trace back randomized data to the original clusters.

Chapter 6. From shelf to farm: Final conclusions and major findings to support fundamental and industrial research

This chapter presents the major findings and conclusions from the work undertaken during this study which bridges fundamental and applied research into industrial processes. The chapter combines two areas in food research explored in this study: supply chain operations and the use of genomic biomarkers technology to support the food industry to demonstrate how they can be used as a control tool in sustainability and food safety.

The way we see food is changing and claims about provenance, origin and "farm to fork" trends are becoming more important (Kamath, 2018). The present research introduced a new concept "From Shelf to Farm" as a way to verify where chocolate products come from. The chocolate supply chain has multiple quality and sustainability issues, some of which have persisted since the 19th century (Hasian Jr., 2008). The analysis described in chapter 2 has identified big gaps in knowledge of traceability, origin and cultivars (genotypes) across the producers, manufacturers and by extension consumers. This has led to a mix of concepts and perspectives about what quality or premium really means. As a subjective characteristic, these labels are often associated with quality and variations in price per region or market. For example, for claims of chocolate made from native genotypes, some companies sell 50g of chocolate for less than £2 while others charge more than £300. This creates an issue in understanding whether quality claims are related to post-harvest processes, single-origin beans, genotypes or certifications.

The main factor leading to quality claims was identified as unique genotypes, which were designated by country. Apart from West Africa, almost every country and stakeholder claimed to have premium or unique varieties of cacao beans, with fine and flavour profiles. Different countries claim to produce unique native cacao with a signature flavour, for example in Bolivia the variety is called 'Amazonian', while in Ecuador it is called 'Nacional' or 'Arriba' and in Mexico it is referred to as 'Criollo'. The research also showed that these names are products of each countries linguistical approach, rather than proven genetic differences. This suggests a problem with the terminology used to describe cacao; this could be addressed by using scientifically-

based descriptors, which could be incorporated in the quality standards. Moreover, there were contradictory views amongst manufacturers; some stated that each genotype has the potential to become premium chocolate independently of the genetic characteristics, while others claimed that quality could only be achieved through understanding the genetics. This led some stakeholders to clone their private breeds. This research shows that there is a huge interest in learning more about cacao genotypes and that the majority of stakeholders would like to map the source of high-quality cacao crops. Furthermore, as there is no international standard lexicon for cacao, the stages of cacao production and cooperative trading were categorised by stakeholders as at high risk of mislabelling, smuggling and bean mixing. This translates into an increased risk of loss of traceability (Figure 2.2). Thus the present research indicates that the main issues is not related to the quality of the flavour alone, but also about sustainability and fair trade constraints.

The research also revealed that cacao producers and policymakers' interactions are shaped by FMCG chocolate manufactures who control more than 80% of the worldwide cacao trade and whose supply chain structures are, as yet, not sustainable. Whilst all the supply chain actors are affected by multiple types of fraud (mixing of beans of un-proven quality or beans without sustainability certification), the main issues that beset the system include aspects of farmers' livelihoods, child slavery and deforestation (Sonwa *et al.*, 2019a). Issues related to deforestation and child slavery were highlighted by the interviewees as the core problems within the FMCG supply chains and self-declared sustainable companies. Representatives of global corporations highlighted the urgent need to implement new technologies that would combat these issues, as justified by interviewee 19 *"As the world's leading manufacturer of high-quality chocolate and cocoa products, we have a moral obligation as well as a business interest to tackle the structural issues in the chocolate supply chain"* interviews carried out here clearly demonstrate that policymakers and self-declared premium or sustainable cacao firms emphasize that ISO, Cocoa of Excellence, UTZ and other certifications are not sufficient, but used in combination, might help improve infrastructure in all countries.

Independently of the goal of the stakeholder (fairness, fine-flavours and other claims), outcomes from the interviews clearly highlighted the needs for new, accurate technologies based on inalterable data sources, such as DNA-Biomarkers, that could be applied to ensure sustainable cacao production. Such a tool would complement document-based and fieldwork inspections of farms and cooperatives. Moreover, stakeholder interviews indicate that policymakers and chocolate manufacturers are interested in applying multiple technologies to audit and improve supply chain transparency, border controls and provide stories for marketing purposes. These tools could include: The Internet of Things (IoT), Polygon Mapping, Data Analytics, Artificial Intelligence and importantly for this research, the use of biomarkers, which could be combined with blockchain technologies. Two stages in chocolate production were identified as key to be screened for tracking implementation: The farm to identify cacao beans and the cooperative where fermentation of cacao samples occur.

The present work explored the possibility of using DNA markers to characterise these two stages in the chocolate manufacturing. Before marker testing could be performed, the first step was to assess that DNA could be isolated reliably from a range of cacao product and this was demonstrated in chapter 3. A key development worth noting has been the detection of low DNA concentration from cacao butter (CB) for the first time with previous studies failing to detect any DNA at all (Gryson, Dewettinck and Messens, 2007; Ha *et al.*, 2015a). Using DNA extracted from a range of cacao samples, markers targeting the plant chloroplast genome and markers targeting the microbiota characterising cacao beans fermentation stages were successfully amplified and analysed.

In chapter 4, the use of chloroplast markers was shown to be a reliable and cost-effective approach to identify the farm haplotype composition of the sample. This simple model showed that assessing a group of unknown plants in a plot/farm by unsupervised machine learning approaches can give clear discrimination between single-origin products. It also generates data that can be used for quantitative models to identify proportions of the haplotype groups in chocolate products. This pilot system could be adapted to other biomarkers including DNA markers characteristics of the nuclear genome of *T. cacao* or volatile components associated with specific

cultivars. These could be more accurately identified and might characterise the presence or absence of specific genotypes of interest in specific regions.

One of the key stages of chocolate production for tracking was identified as the cooperative were beans are processed and markers related to the fermentation process were investigated. The targeted microbiome analysis of chocolate samples conducted in Chapter 5 revealed that it is possible to identify good quality biomarkers related to organism link to the fermentation. After performing unsupervised machine learning screening for pattern discovery in the microbiome genomics of chocolate samples, signature biomarkers were identified for cooperative locations preserved across years. This data generated through Illumina sequencing enables various hypotheses to be drawn about the quality, stages of post-harvest, traceability of cacao beans, manufacturing process and chocolate products from single origins.

These unique ASVs per origin could be defined by brand, post-harvest facility, country and continent. This in turn, can be used to determine inclusion or exclusion of particular groups in a controlled sample set. To make these biomarkers applicable in industrial scenarios, it is vital to create a machine learning model that could recognize the ASVs from multiple regions. To do this a controlled data set with all the identified biomarkers needs to be generated and trained. In order to be able to compare a random sample in a trained data set, in-depth supervised machine learning needs to be implemented. Such a system should be able to handle a large number of labelled (known origin) samples as independent variables. The system should be able to identify the best predictive model to trace back randomized data to the original clusters.

As a proof of concept and using data generated in this research, a pilot model was built using a closed reference method which assigned the closest match of the ASV to a taxonomic profile from the green genes database. Following the process used for personalized medicine biomarker benchmarks (Lloyd-Price *et al.*, 2017; Vázquez-Baeza *et al.*, 2018), the samples were analysed to predict and forecast microbial abundances based on taxa from different geographical regions. An Ordinary Least Square (OLS) model was designed, using fixed factors (origin): fermentation location and country of origin were run as independent variables. To evaluate the explanatory

model of a single covariate (origin), a leave-one-variable-out approach (Shi, Zhang and Li, 2016) was used on each combination of microbiomes (y0 and y1). This suggested that the current unique ASVs are associated with the area of fermentation (Richards *et al.*, 2019) (Figure 6.1) and can be use as markers.

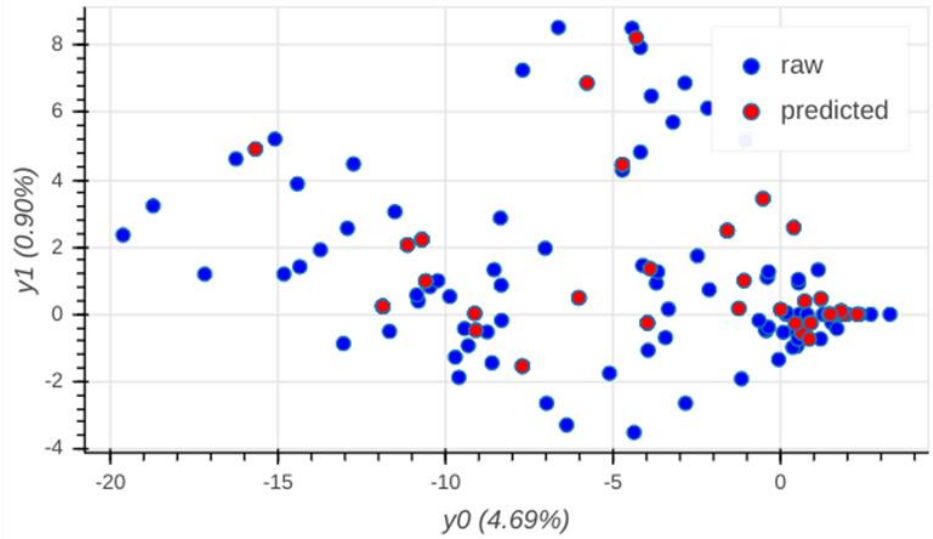


Figure 6.1 Predicted points lie within the same region as the original microbiome and location

The figure shows the assignment (4.69%) of communities which can be predicted after running the OLS model. Raw log ratios of the microbial communities balances (y1 in y-axis and y0 in x-axis) are represented with blue dots and the predicted red points (R^2 62%) which represent: Vietnam: Tan Phu, Ha Long; Ivory Coast: Duékoué, Indonesia: Kecamatan Guguk; Bali, Jembrana; Malaysia: Ranau; Ghana: Ahafo Ano South; Bolivia: Riberalta; Peru: Cuzco, Castilla; Ecuador: Balao, Guayas, Esmeraldas, Vinces; Colombia: San Andres de Tumaco; Guatemala: Río Dulce, Izabal. These lie within the same labelled regions as the original communities and that further analysis could be performed on the resulting regions to identify its unique sequences which that can be used for prediction of each region.

The ideal scenario to predict origins with reliable accuracy would require recording the frequency of a larger number of ASV, from multiple controls for each locations of likely origin for a product. To develop this further, a new classifier (reference data base) for taxonomic assignments or geographical locations would need to be built. These proposed classifiers need to be developed from the actual library of new ASVs which will become the training (control data base) data set for chocolate tracking (Knights *et al.*, 2011)Knights et al., 2011).

Chapter 7. Future work

This thesis shows how DNA biomarkers could be used in industrial scenarios to improve supply chain visibility (SCV), sustainable operations and the societal value of a product. The use of DNA biomarkers can be an unalterable target in the supply chain of a product that can give insights about the process or location. This is an advantage when compared to other current approaches, including RFID tags attached to the cocoa bags which can be transferred to other bags, visual inspections made by the staff that could be mistaken, or isotope tracking that can be degraded through the process. As stated in this research, certifications such as ISO or UTZ help to standardize processes related to organizational structures and accounting within the farmers and trader; however, they need the support of technology to help decrease fraud or bias that is embedded in the national and international trading system. The findings of this research showed that even if farmers/traders/chocolate companies are certified “fair, organic, sustainable, traceable”, farmers which have jointed plots or associated companies still share the certificates between them to sell their beans.

Companies have been assessing the use of QR codes printed on the packaging of products to give information about the location of origin or the storytelling of the chocolate bars. This underlined that a QR code could be used to inform to the client any other evidence related to the origin, processes or certifications that the company have. Some start-ups such as Farmers Connect in Europe and USA have implemented new metrics such as the price of the cocoa or coffee beans paid to the farmer as a starting point to show to the final client how does their process add value to the product through the supply chain. This model aims to improve the visibility of the trade and SC of companies. However, as previously shown this information can be altered through the process. Which is why “Blockchain” technology was studied within this research, that by design, is resistant to modification of the data once recorded. Blockchain is "an open, distributed ledger that can record transactions between two parties efficiently and in a verifiable and permanent way" (Marco Iansiti and Karim R. Lakhani, 2017). Therefore, the data (DNA sequence, Price or Barcodes) registered in any given block cannot be altered retroactively without the alteration of all subsequent blocks.

These insights could lead to future work in the area of Innovation & Technology, Metagenomics and Supply Chain Management within a framework "From Farm to Shelf: a science-based bio-visibility audit Framework" (*Lafargue et al 2020 – Peer review*), Figure 7.1. The forensic DNA-biomarkers assessment performed in this research generated sufficient reliable and robust data that is stored in a database with information about the crop, farm, and geo-location of the product and DNA sequences for each origin. This creates an important data pool that can be used to compare new samples that need to be traced to an origin. The application of this framework could be used for further research in academia and as an auditing process for certifiers and industry.

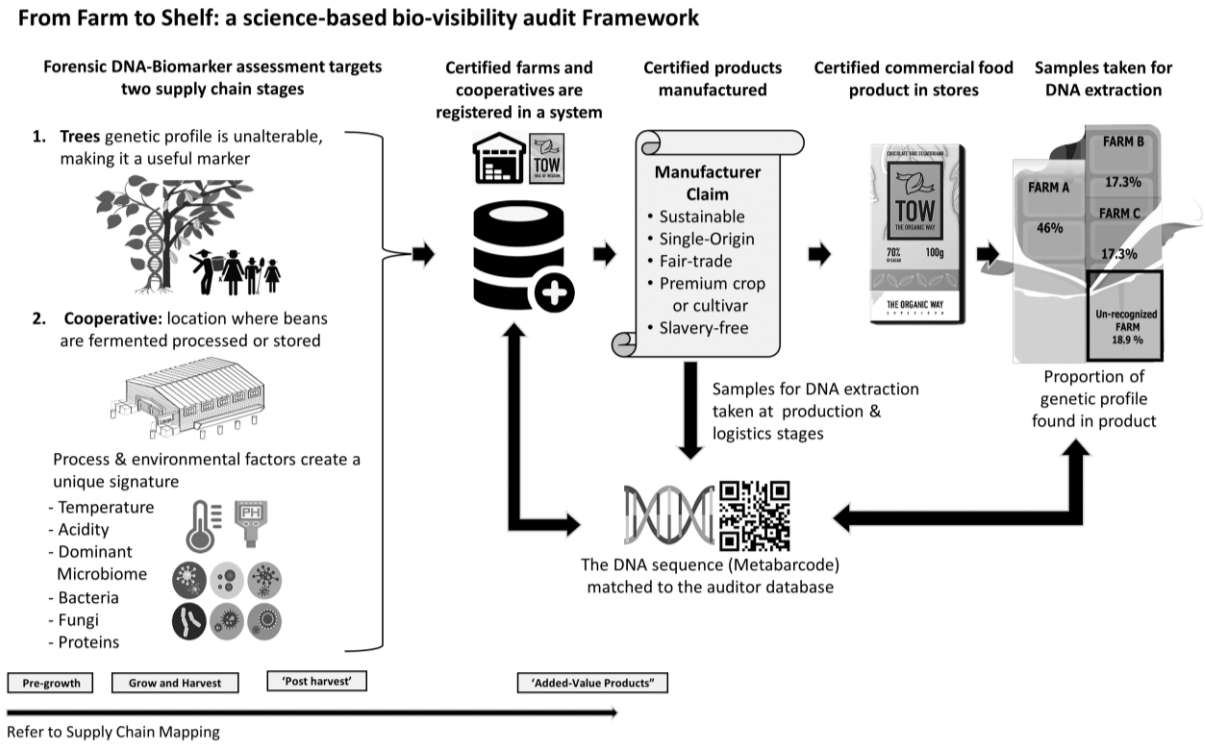


Figure 7.1 From Farm to Shelf: A science-based bio-visibility audit Framework (Lafargue. P, Rogerson. M, Allainguillaume. J, Parry. G, 2020) International Supply Chain Management Journal (Peer Review).

7.1 Academic Research

7.1.1 DNA Markers development and Metagenomics analysis

To strengthen the efficiency of this tool there are three main areas of research that are required:

1) Increasing the chocolate sample size from known origins and cultivars, thereby building a controlled data set of unique beans for national or international mapping, e.g. South vs North of Ivory Coast or the borders between Colombia and Ecuador.

2) Improving the accuracy of the chloroplast DNA markers to identify variations in genotypes by using whole-genome or shotgun sequencing. This will include the development of new SNP markers and DNA Microarray technology (A multiplex lab-on-a-chip with multiple DNA probes) instead of SSRs which can improve the chloroplast ultra-barcoding accuracy and allow further comparison of SNP data from other research institutes (e.g Reading, CATIE). Additionally, the combination of all the unique SSR markers could show the allele frequency variations and therefore the difference, presence or absence of cultivars such as CCN51 or Nacional genotypes.

3) Incorporating DNA markers from fungi or viruses within metagenomics studies, which could reveal important information on each fermentation process and location, and could be correlated with the previous bacterial determination from the coops.

7.1.2 Innovation and Technology into Supply Chain Management

New developments in the field of innovation and technology are changing how data is collected and analysed. For decades supply chain management has focused on studying how tangible assets, operations and hardware add value to companies or processes. More recently, a growing interest has been shown by world-leading countries to become more financially sustainable and environmental friendly. However, to accomplish these goals a strong analysis of environmental data (e.g. Carbon footprint, Human and environmental resources) needs to be performed and to date, it hasn't been exploited.

7.2 Industrial Research

The outputs of this thesis were presented to multinational manufacturing companies, traders and certifiers. The result of these discussions highlighted that companies could use a DNA barcode tool to improve their current procurement and

storytelling processes. To do so the companies could supply samples of single-origin beans and chocolate from each location of interest to the internal/external auditor or laboratory of preference. This will follow with the proper analysis so they can claim that there is no mixing of unwanted/unsustainable countries in their product. Additionally, it was identified that cooperation such as Knowledge Transfer Partnerships between higher education, manufacturers and policymakers could help to implement innovative technology on real-time and support governmental development. By working collaboratively faster technologies such as Nanopore sequencing could be used to sample each batch of beans at port or when products are received by the client.

References

Acierno, V., Alewijn, M., Zomer, P. and van Ruth, S.M. (2018) Making cocoa origin traceable: Fingerprints of chocolates using Flow Infusion - Electro Spray Ionization - Mass Spectrometry. *Food Control*. doi:10.1016/j.foodcont.2017.10.002.

Acierno, V., Yener, S., Alewijn, M., Biasioli, F. and Van Ruth, S. (2016) Factors contributing to the variation in the volatile composition of chocolate: Botanical and geographical origins of the cocoa beans, and brand-related formulation and processing. *Food Research International* [online]. 84 pp. 86–95. Available from: http://dx.doi.org/10.1016/j.foodres.2016.03.022doi:10.1016/j.foodres.2016.03.022.

Afoakwa, E.O. (2016a) *Chocolate Science and Technology: Second Edition* Wiley - Blackwell (ed.). [online]. Second Edi. Accra, Ghana: Jhon Wiley & Sons, Ltd.

Afoakwa, E.O. (2010) *Chocolate Science and Technology*.

Afoakwa, E.O. (2016b) Food safety management systems in chocolate processing. In: *Chocolate Science and Technology*. (no place) John Wiley & Sons, Ltd. pp. 399–415. doi:10.1002/9781118913758.ch18.

Afoakwa, E.O., Kongor, J.E., Takrama, J.F. and Budu, A.S. (2013) Changes in acidification, sugars and mineral composition of cocoa pulp during fermentation of pulp pre-conditioned cocoa (Theobroma cacao) beans. *International Food Research Journal*. 20 (3), pp. 1215–1222.

Afoakwa, E.O., Paterson, A. and Fowler, M. (2008) Effects of particle size distribution and composition on rheological properties of dark chocolate. *European Food Research and Technology*. doi:10.1007/s00217-007-0652-6.

Agyirifo, D.S., Wamalwa, M., Otwe, E.P., Galyuon, I., Runo, S., Takrama, J. and Ngeranwa, J. (2019) Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities. *Heliyon* [online]. 5 (7), pp. e02170. Available from: https://doi.org/10.1016/j.heliyon.2019.e02170doi:10.1016/j.heliyon.2019.e02170.

Allegre, M., Argout, X., Boccara, M., Fouet, O., Roguet, Y., Bérard, A., Thévenin, J.M., Chauveau, A., Rivallan, R., Clement, D., Courtois, B., Gramacho, K., Boland-Augé, A., Tahi, M., et al. (2012) Discovery and mapping of a new expressed sequence tag-single nucleotide polymorphism and simple sequence repeat panel for large-scale genetic studies and breeding of Theobroma cacao L. *DNA Research*. 19 (1), pp. 23–35. doi:10.1093/dnares/dsr039.

Anderson, M.J. (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology*. 26 (1), pp. 32–46. doi:10.1046/j.1442-9993.2001.01070.x.

Araujo, Q.R., Fernandes, C.A.F., Ribeiro, D.O., Efraim, P., Steinmacher, D., Lieberei, R., Bastide, P. and Araujo, T.G. (2014) Cocoa Quality Index e A proposal. *Food Control* [online]. 46 pp. 49–54. Available from: http://dx.doi.org/10.1016/j.foodcont.2014.05.003doi:10.1016/j.foodcont.2014.05.003.

Ardhana, M.M. and Fleet, G.H. (2003) The microbial ecology of cocoa bean fermentations in Indonesia. *International Journal of Food Microbiology*. 86 (1–2), pp. 87–99. doi:10.1016/S0168-1605(03)00081-3.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., et al. (2011) The genome of Theobroma cacao. *Nature Genetics* [online]. 43 (2), pp. 101–108. Available from: http://www.nature.com/doifinder/10.1038/ng.736doi:10.1038/ng.736.

Attwood, G.T., Wakelin, S.A., Leahy, S.C., Rowe, S., Clarke, S., Chapman, D.F., Muirhead, R. and Jacobs, J.M.E. (2019) Applications of the soil, plant and rumen microbiomes in pastoral agriculture *Frontiers in Nutrition*. doi:10.3389/fnut.2019.00107.

Azir, M., Abbasiliasi, S., Tengku Ibrahim, T., Manaf, Y., Sazili, A. and Mustafa, S. (2017) Detection of Lard in Cocoa Butter—Its Fatty Acid Composition, Triacylglycerol Profiles,

and Thermal Characteristics. *Foods*. doi:10.3390/foods6110098.

Badia-Melis, R., Mishra, P. and Ruiz-García, L. (2015) Food traceability: New trends and recent advances. A review. *Food Control* [online]. 57 pp. 393–401. Available from: http://dx.doi.org/10.1016/j.foodcont.2015.05.005doi:10.1016/j.foodcont.2015.05.005.

Barrangou, R., Yoon, S.S., Breidt, F., Fleming, H.P. and Klaenhammer, T.R. (2002) Identification and characterization of Leuconostoc fallax strains isolated from an industrial sauerkraut fermentation. *Applied and Environmental Microbiology*. 68 (6), pp. 2877–2884. doi:10.1128/AEM.68.6.2877-2884.2002.

Beckett, S. (2008) *"The Science of chocolate"*.

Beg, M.S., Ahmad, S., Jan, K. and Bashir, K. (2017) Status, supply chain and processing of cocoa - A review. *Trends in Food Science and Technology*. 66 pp. 108–116. doi:10.1016/j.tifs.2017.06.007.

Belsky, J.M., Siebert, S.F., Argout, X., Salse, J., Aury, J.-M., Guiltinan, M.J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S.N., Abrouk, M., Murat, F., et al. (2014) Report on Cacao Products. *Cafe Cacao The* [online]. 16 (2), pp. v. Available from: http://sk8es4mc2l.search.serialssolutions.com/?sid=sersol%7B&%7DSS%7B_%7Djc=TC%7B_%7D00885199 7%7B&%7Dtitle=Impact%5CnOf%5CnStructural%5CnAdjustment%5Cnand%5CnAdoption%5CnTechnology%5Cnon%5CnCompetitiveness%5CnMajor%5CnCocoa%5CnProducing%5CnCountries$%5C$ndoi:BOOK_DOI 10.1201/b16546.

Ben-Ayed, R., Kamoun-Grati, N. and Rebai, A. (2013) An overview of the authentication of olive tree and oil. *Comprehensive Reviews in Food Science and Food Safety*. 12 (2), pp. 218–227. doi:10.1111/1541-4337.12003.

Benjamini, Y. and Hochberg, Y. (1995) Benjamini-1995.pdf *Journal of the Royal Statistical Society B* [online] 57 (1) p.pp. 289–300. Available from: http://www.jstor.org/stable/2346101doi:10.2307/2346101.

Berlan, A. (2013) Social Sustainability in Agriculture: An Anthropological Perspective on Child Labour in Cocoa Production in Ghana. *Journal of Development Studies*. 49 (8), pp. 1088–1100. doi:10.1080/00220388.2013.780041.

Bertoldi, D., Barbero, A., Camin, F., Caligiani, A. and Larcher, R. (2016) Multielemental fingerprinting and geographic traceability of Theobroma cacao beans and cocoa products. *Food Control*. 65 pp. 46–53. doi:10.1016/j.foodcont.2016.01.013.

Bieber, F.R., Buckleton, J.S., Budowle, B., Butler, J.M. and Coble, M.D. (2016) Evaluation of forensic DNA mixture evidence: Protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion. *BMC Genetics* [online]. 17 (1), pp. 1–15. Available from: http://dx.doi.org/10.1186/s12863-016-0429-7doi:10.1186/s12863-016-0429-7.

Böhme, K., Calo-Mata, P., Barros-Velázquez, J. and Ortea, I. (2019) Recent applications of omics-based technologies to main topics in food authentication *TrAC - Trends in Analytical Chemistry* 110 p.pp. 221–232. doi:10.1016/j.trac.2018.11.005.

Bokulich, N.A., Kaehler, B.D., Rideout, J.R., Dillon, M., Bolyen, E., Knight, R., Huttley, G.A., Gregory Caporaso, J. and Caporaso, J.G. (2018) Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* [online]. 6 (1), pp. 90. Available from: https://doi.org/10.1186/s40168-018-0470-zdoi:10.1186/s40168-018-0470-z.

Bokulich, N.A., Thorngate, J.H., Richardson, P.M. and Mills, D.A. (2014) Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proceedings of the National Academy of Sciences* [online]. 111 (1), pp. E139–E148. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1317377110doi:10.1073/pnas.1317377110.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., et al. (2018) QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints* [online]. 6 pp. e27295v1. Available from: https://doi.org/10.7287/peerj.preprints.27295v1doi:10.7287/peerj.preprints.27295v1.

Boza, E., Motamayor, J.C., Amores, F., Cedeño-Amador, S., Tondo, C., Livingstone, D., Schnell, R. and Gutierrez, O. (2014) Genetic Characterization of the cacao (Theobroma cacao L.) clone 'CCN 51' and its impact and significance on global cacao improvement and production. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*. 139 (2), pp. 219.

Branch, A., Byrne, P., Costa, A., Entzminger, C., Fredericq, A., Gilmour, M., Laird, G., Matissek, R., Quintana, S., Ruiz, S. and Sigley, P. (2015) *Cocoa Beans: Chocolate & Cocoa Industry Quality Requirements* [online]. (no place) ECA-Caobisco-FCC Cocoa.

Braukmann, T.W.A.A., Kuzmina, M.L., Sills, J., Zakharov, E. V. and Hebert, P.D.N.N. (2017) Testing the efficacy of DNA barcodes for identifying the vascular plants of Canada. *PLoS*

ONE. 12 (1), pp. 1–19. doi:10.1371/journal.pone.0169515.

Bukin, Y.S., Galachyants, Y.P., Morozov, I. V., Bukin, S. V., Zakharenko, A.S. and Zemskaya, T.I. (2019) The effect of 16s rRNA region choice on bacterial community metabarcoding results. *Scientific Data*. 6 . doi:10.1038/sdata.2019.7.

Busconi, M., Foroni, C., Corradi, M., Bongiorni, C., Cattapan, F. and Fogher, C. (2003) DNA extraction from olive oil and its use in the identification of the production cultivar. *Food Chemistry*. doi:10.1016/S0308-8146(03)00218-8.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*. 13 (7), pp. 581–583. doi:10.1038/nmeth.3869.

Camin, F., Boner, M., Bontempo, L., Fauhl-Hassek, C., Kelly, S.D., Riedl, J. and Rossmann, A. (2017) Stable isotope techniques for verifying the declared geographical origin of food in legal cases *Trends in Food Science and Technology*. doi:10.1016/j.tifs.2016.12.007.

Camu, N., Bernaert, H. and Lohmueller, T. (2010) *Microbial composition for the fermentation of cocoa material.* Available from: https://patents.justia.com/patent/9701986 [Accessed 23 December 2019].

Camu, N., De Winter, T., Verbrugghe, K., Cleenwerck, I., Vandamme, P., Takrama, J.S., Vancanneyt, M. and De Vuyst, L. (2007) Dynamics and biodiversity of populations of lactic acid bacteria and acetic acid bacteria involved in spontaneous heap fermentation of cocoa beans in Ghana. *Applied and Environmental Microbiology*. 73 (6), pp. 1809–1824. doi:10.1128/AEM.02189-06.

Carr, M.K.V. V and Lockwood, G. (2011) The water relations and irrigation requirements of cocoa (Theobroma cacao L.): A review. *Experimental Agriculture*. 47 (4), pp. 653–676. doi:10.1017/S0014479711000421.

Chater, K.F., Biró, S., Lee, K.J., Palmer, T. and Schrempf, H. (2010) The complex extracellular biology of Streptomyces: REVIEW ARTICLE. *FEMS Microbiology Reviews*. 34 (2), pp. 171–198. doi:10.1111/j.1574-6976.2009.00206.x.

Che Man, Y.B., Syahariza, Z.A., Mirghani, M.E.S., Jinap, S. and Bakar, J. (2005) Analysis of potential lard adulteration in chocolate and chocolate products using Fourier transform infrared spectroscopy. *Food Chemistry*. doi:10.1016/j.foodchem.2004.05.029.

Cidell, J.L. and Alberts, H.C. (2006) Constructing quality: The multinational histories of

chocolate. *Geoforum*. 37 (6), pp. 999–1007. doi:10.1016/j.geoforum.2006.02.006.

Clarke, V. and Braun, V. (2017) Thematic analysis *Journal of Positive Psychology*. doi:10.1177/1541344618777367.

Cocoa of Excellence (2017) *Cocoa of Excellence Programme 2017 - sample quota per country & Region Annex A . Allocation of sample quotas Annex B . Quotas per region & country*.

Cocoa of Excellence and Biodiversity (2017) International stakeholders ' consultations on the development and validation of proposed international standards on cocoa quality and flavour assessment Reports compiled by Brigi e Laliberte , Cocoa of Excellence Programme Bioversity Interna onal , Rome , *Reports Cocoa of Excellence Programme* (October).

Cocoa of Excellence Programme (2015) *CoEX High-quality Cocoa Origins Results of the 2015 Edition* (October).

Corbin, J. and Strauss, A. (2012) Practical Considerations. In: *Basics of Qualitative Research (3rd ed.): Techniques and Procedures for Developing Grounded Theory*. doi:10.4135/9781452230153.n2.

Cornejo, O.E., Yee, M.-C.C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., Livingstone, D., Stack, C., Romero, A., Umaharan, P., Royaert, S., Tawari, N.R., Ng, P., Gutierrez, O., et al. (2018) Population genomic analyses of the chocolate tree, Theobroma cacao L., provide insights into its domestication process. *Communications Biology* [online]. 1 (1), pp. 167. Available from: http://www.nature.com/articles/s42003-018-0168-6doi:10.1038/s42003-018-0168-6.

Cotter, P.D. and Beresford, T.P. (2017) Microbiome Changes During Ripening. In: *Cheese: Chemistry, Physics and Microbiology: Fourth Edition*. doi:10.1016/B978-0-12-417012-4.00015-6.

Crafack, M., Mikkelsen, M.B., Saerens, S., Knudsen, M., Blennow, A., Lowor, S., Takrama, J., Swiegers, J.H., Petersen, G.B., Heimdal, H. and Nielsen, D.S. (2013) Influencing cocoa flavour using Pichia kluyveri and Kluyveromyces marxianus in a defined mixed starter culture for cocoa fermentation. *International Journal of Food Microbiology*. 167 (1), pp. 103–116. doi:10.1016/j.ijfoodmicro.2013.06.024.

Crits-Christoph, A., Robinson, C.K., Barnum, T., Fricke, W.F., Davila, A.F., Jedynak, B., McKay, C.P. and DiRuggiero, J. (2013) Colonization patterns of soil microbial

communities in the Atacama Desert. *Microbiome*. 1 (1), . doi:10.1186/2049-2618-1-28.

Crown, P.L. and Hurst, W.J. (2009) Evidence of cacao use in the Prehispanic American Southwest. *Proceedings of the National Academy of Sciences* [online]. 106 (7), pp. 2110–2113. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0812817106doi:10.1073/pnas.0812817106.

Dand, R. (2010) *The International Cocoa Trade: Third Edition*. (no place) Elsevier Ltd.

Daniell, H., Lin, C.-S., Yu, M. and Chang, W.-J. (2016) Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biology* [online]. 17 (1), pp. 134. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1004-2doi:10.1186/s13059-016-1004-2.

Davrieux, F., Assemat, S., Sukha, D., Portillo, E., Boulanger, R., Bastianelli, D. and Cros, E. (2007) Genotype characterization of cocoa into genetic groups through caffeine and theobromine content predicted by near infra red spectroscopy. *Applications*. pp. 382–386.

Denzin, N.K. and Strauss, A.L. (2006) Qualitative Analysis for Social Scientists. *Contemporary Sociology*. 17 (3), pp. 430. doi:10.2307/2069712.

Deppeler, A., Fromm, I. and Aidoo, A. (2014) The Unmaking of the Cocoa Farmer: Analysis of Benefits and Challenges of third-party audited Certification Schemes for Cocoa Producers and Laborers in Ghana. *IFAMA and CCA Agribusiness & Food World Forum*.

Dhoedt, A.N.N. (2009) 'Food of the Gods' the rich history of chocolate. *Agro Food Industry Hi-Tech*. 20 (6 SUPPL. 1), pp. 2–5.

Doosti, A., Ghasemi Dehkordi, P. and Rahimi, E. (2014) Molecular assay to fraud identification of meat products. *Journal of Food Science and Technology*. 51 (1), pp. 148–152. doi:10.1007/s13197-011-0456-3.

Dragusanu, R. and Nunn, N. (2014) The Impacts of Fair Trade Certification: Evidence From Coffee Producers in Costa Rica. *Harvard Business School working paper (2013)*. (February), .

Drever, E. and Scottish Council for Research in Education, E. (1995) *Using Semi-Structured Interviews in Small-Scale Research. A Teacher's Guide.* [online].

Drummond, M.G., Brasil, B.S.A.F., Dalsecco, L.S., Brasil, R.S.A.F., Teixeira, L. V. and Oliveira,

D.A.A. (2013) A versatile real-time PCR method to quantify bovine contamination in buffalo products. *Food Control* [online]. 29 (1), pp. 131–137. Available from: http://dx.doi.org/10.1016/j.foodcont.2012.05.051doi:10.1016/j.foodcont.2012.05.051.

Escobar-Zepeda, A., Sanchez-Flores, A. and Quirasco Baruch, M. (2016) Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota. *Food Microbiology* [online]. 57 pp. 116–127. Available from: http://dx.doi.org/10.1016/j.fm.2016.02.004doi:10.1016/j.fm.2016.02.004.

European Commission (2014) COMMISSION REGULATION (EU) No 488/2014 of 12 May 2014 amending Regulation (EC) No 1881/2006 as regards maximum levels of cadmium in foodstuffs. *Official Journal of the European Commission*. doi:10.2903/j.efsa.2011.1975.

Falque, M., Vincent, A., Vaissiere, B.E. and Eskes, A.B. (1995) Effect of pollination intensity on fruit and seed set in cacao (Theobroma cacao L.). *Sexual Plant Reproduction* [online]. 8 (6), pp. 354–360. Available from: http://dx.doi.org/10.1007/BF02180403%5Cnhttp://springerlink.metapress.com/openurl.asp?genre=article&id=doi:10.1007/BF00243203doi:10.1007/BF00243203.

Fletcher, E. (2015) Interpreting qualitative data. *International Journal of Research & Method in Education*. doi:10.1080/1743727x.2015.1066173.

Foddy, W. (2009) *Constructing Questions for Interviews and Questionnaires*.

Folds, N. (2002) Lead Firms and Competition in "Bi-polar" Commodity Chains: Grinders and Branders in the Global Cocoa-Chocolate Industry. *Journal of Agrarian Change*. 2 (2), pp. 228.

Fowler, M.S. (2009) Cocoa Beans: From Tree to Factory. In: *Industrial Chocolate Manufacture and Use: Fourth Edition*. (no place) Wiley-Blackwell. pp. 10–47. doi:10.1002/9781444301588.ch2.

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S. and Labra, M. (2013) DNA barcoding as a new tool for food traceability *Food Research International* 50 (1) p.pp. 55–63. doi:10.1016/j.foodres.2012.09.036.

Garcia-Armisen, T., Papalexandratou, Z., Hendryckx, H., Camu, N., Vrancken, G., De Vuyst, L. and Cornelis, P. (2010) Diversity of the total bacterial community associated with Ghanaian and Brazilian cocoa bean fermentation samples as revealed by a 16 S rRNA gene clone library. *Applied Microbiology and Biotechnology*. 87 (6), pp. 2281–2292.

doi:10.1007/s00253-010-2698-9.

Germani, M., Mandolini, M., Marconi, M., Marilungo, E. and Papetti, A. (2015) A system to increase the sustainability and traceability of supply chains. *Procedia CIRP*. 29 pp. 227–232. doi:10.1016/j.procir.2015.02.199.

Ghannam, R.B., Schaerer, L.G., Butler, T.M. and Techtmann, S.M. (2020) Biogeographic Patterns in Members of Globally Distributed and Dominant Taxa Found in Port Microbial Communities. *mSphere*. doi:10.1128/msphere.00481-19.

Gibtan, A., Park, K., Woo, M., Shin, J.K., Lee, D.W., Sohn, J.H., Song, M., Roh, S.W., Lee, S.J. and Lee, H.S. (2017) Diversity of extremely halophilic archaeal and bacterial communities from commercial salts. *Frontiers in Microbiology* [online]. 8 (MAY), pp. 1–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5423978/pdf/fmicb-08-00799.pdfdoi:10.3389/fmicb.2017.00799.

Gordon-Finlayson, A. (2010) QM2: Grounded Theory. In: *Doing qualitative research in psychology : a practical guide*.

Gryson, N., Dewettinck, K. and Messens, K. (2007) Influence of cocoa components on the PCR detection of soy lecithin DNA. *European Food Research and Technology*. 226 (1–2), pp. 247–254. doi:10.1007/s00217-006-0533-4.

Gryson, N., Messens, K. and Dewettinck, K. (2004) Evaluation and optimisation of five different extraction methods for soy DNA in chocolate and biscuits. Extraction of DNA as a first step in GMO analysis. *Journal of the Science of Food and Agriculture*. 84 (11), pp. 1357–1363. doi:10.1002/jsfa.1767.

Guehi, T.S., Dadie, A.T., Koffi, K.P.B., Dabonne, S., Ban-Koffi, L., Kedjebo, K.D. and Nemlin, G.J. (2010) Performance of different fermentation methods and the effect of their duration on the quality of raw cocoa beans. *International Journal of Food Science and Technology*. 45 (12), pp. 2508–2514. doi:10.1111/j.1365-2621.2010.02424.x.

Gutiérrez-López, N., Ovando-Medina, I., Salvador-Figueroa, M., Molina-Freaner, F., Avendaño-Arrazate, C.H. and Vázquez-Ovando, A. (2016) Unique haplotypes of cacao trees as revealed by *trnH-psbA* chloroplast DNA. *PeerJ* [online]. 4 pp. e1855. Available from: https://peerj.com/articles/1855doi:10.7717/peerj.1855.

Ha, L.T.V., Vanlerberghe, L., Toan, H.T., Dewettinck, K. and Messens, K. (2015a) Comparative

Evaluation of Six Extraction Methods for DNA Quantification and PCR Detection in Cocoa and Cocoa-Derived Products. *Food Biotechnology*. 29 (1), pp. 1–19. doi:10.1080/08905436.2014.996761.

Ha, L.T.V., Vanlerberghe, L., Toan, H.T., Dewettinck, K., Messens, K., Viet Ha, L.T., Vanlerberghe, L., Toan, H.T., Dewettinck, K. and Messens, K. (2015b) Comparative Evaluation of Six Extraction Methods for DNA Quantification and PCR Detection in Cocoa and Cocoa-Derived Products. *Food Biotechnology*. 29 (1), pp. 1–19. doi:10.1080/08905436.2014.996761.

Hamdouche, Y., Guehi, T., Durand, N., Kedjebo, K.B.D., Montet, D. and Meile, J.C. (2015) Dynamics of microbial ecology during cocoa fermentation and drying: Towards the identification of molecular markers. *Food Control* [online]. 48 pp. 117–122. Available from: http://dx.doi.org/10.1016/j.foodcont.2014.05.031doi:10.1016/j.foodcont.2014.05.031.

Hamdouche, Y., Meile, J.C., Lebrun, M., Guehi, T., Boulanger, R., Teyssier, C. and Montet, D. (2019) Impact of turning, pod storage and fermentation time on microbial ecology and volatile composition of cocoa beans. *Food Research International*. 119 pp. 477–491. doi:10.1016/j.foodres.2019.01.001.

Hasian Jr., M. (2008) Critical Memories of Crafted Virtues: The Cadbury Chocolate Scandals, Mediated Reputations, and Modern Globalized Slavery. *Journal of Communication Inquiry*. 32 (2), pp. 249–270. doi:10.1177/0196859908316331.

Hawkins, J., De Vere, N., Griffith, A., Ford, C.R., Allainguillaume, J., Hegarty, M.J., Baillie, L. and Adams-Groom, B. (2015) Using DNA metabarcoding to identify the floral composition of honey: A new tool for investigating honey bee foraging preferences. *PLoS ONE*. 10 (8), pp. 1–20. doi:10.1371/journal.pone.0134735.

He, X., Carter, J.M., Brandon, D.L., Cheng, L.W. and McKeon, T.A. (2007) Application of a real time polymerase chain reaction method to detect castor toxin contamination in fluid milk and eggs. *Journal of Agricultural and Food Chemistry*. doi:10.1021/jf0707738.

Henderson, R. (2011) Doing qualitative research: a practical handbook. *Studies in Continuing Education*. doi:10.1080/0158037X.2011.609670.

Herrmann, L., Felbinger, C., Haase, I., Rudolph, B., Biermann, B. and Fischer, M. (2015) Food fingerprinting: Characterization of the ecuadorean type CCN-51 of theobroma cacao L. Using microsatellite markers. *Journal of Agricultural and Food Chemistry*. 63 (18), pp. 4539–4544. doi:10.1021/acs.jafc.5b01462.

Hii, C.L., Rahman, R.A., Jinap, S. and Man, Y.B.C. (2006) Quality of cocoa beans dried using a direct solar dryer at different loadings. *Journal of the Science of Food and Agriculture* [online]. 86 (8), pp. 1237–1243. Available from: http://doi.wiley.com/10.1002/jsfa.2475doi:10.1002/jsfa.2475 [Accessed 12 May 2017].

Ho, V.T.T., Zhao, J. and Fleet, G. (2015) The effect of lactic acid bacteria on cocoa bean fermentation. *International Journal of Food Microbiology* [online]. 205 pp. 54–67. Available from: http://dx.doi.org/10.1016/j.ijfoodmicro.2015.03.031doi:10.1016/j.ijfoodmicro.2015.03.031.

Hosseinzadeh-Colagar, A., Haghighatnia, M.J., Amiri, Z., Mohadjerani, M. and Tafrihi, M. (2016) Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. *Molecular biology research communications*. doi:10.22099/mbrc.2016.3789.

Hunter, J.D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*. doi:10.1109/MCSE.2007.55.

ICCO (2012) *International Cocoa Organization - The world cocoa economy: past and present*. Available from: https://www.icco.org/about-us/international-cocoa-agreements/cat_view/30-related-documents/45-statistics-other-statistics.html [Accessed 20 October 2016].

Illeghems, K., de Vuyst, L., Papalexandratou, Z. and Weckx, S. (2012) Phylogenetic analysis of a spontaneous cocoa bean fermentation metagenome reveals new insights into its bacterial and fungal community diversity. *PLoS ONE*. 7 (5), . doi:10.1371/journal.pone.0038040.

International Cocoa Organization (2019) *The Chocolate Industry*.

International Cocoa Organization (2012) The World Cocoa Economy: Past and Present. In: *One hundred and forty-sixth meeting*. 2012 pp. 1–43.

International Cocoa Organization (ICCO) (2009) Guidelines on Best Known Practices in the Cocoa Value Chain. *Nineteenth meeting*. (June), pp. 1–10.

International Organisation for Standardisation (2017a) ISO 34101-1 - Sustainable and traceable cocoa -- Part 1: Requirements for cocoa sustainability management systems *International Organization for Standards* [online] 44 (0). Available from: https://www.iso.org/standard/64765.html.

International Organisation for Standardisation (2017b) ISO 34101-2 - Sustainable and traceable cocoa -- Part 2: Part 2: Requirements for performance (related to economic, social, and environmental aspects). *International Organization for Standards* [online]. 44 (0), . Available from: https://www.iso.org/standard/64765.html.

International Organization for Standards (2016) BSI Standards Publication Foodstuffs — Detection of food allergens by molecular biological methods sequence in chocolate by real-time PCR. In: *BSI Standards Publication*. (no place) BSI Standards Limited 2016.

International Organization for Standards (2017) *ISO 2451:2017, Cocoa beans — Specification and quality requirements*. Available from: https://www.iso.org/obp/ui/#iso:std:iso:2451:ed-3:v1:en [Accessed 15 January 2018].

Jahurul, M.H.A., Zaidul, I.S.M., Norulaini, N.A.N., Sahena, F., Jinap, S., Azmir, J., Sharif, K.M. and Mohd Omar, A.K. (2013) Cocoa butter fats and possibilities of substitution in food products concerning cocoa varieties, alternative sources, extraction methods, composition, and characteristics *Journal of Food Engineering* 117 (4) p.pp. 467–476. doi:10.1016/j.jfoodeng.2012.09.024.

Janssen, S., Mcdonald, D., Gonzalez, A., Navas-molina, J.A., Jiang, L. and Xu, Z. (2018) Phylogenetic Placement of Exact Amplicon Sequences. *mSystems*. 3 (3), pp. e00021-18. doi:10.1128/mSystems.00021-18.

Johnson, R. (2014) *Food fraud and 'Economically motivated adulteration' of food and food ingredients*.

Johnston, R.B. (2016) Arsenic and the 2030 Agenda for sustainable development. *Arsenic Research and Global Sustainability - Proceedings of the 6th International Congress on Arsenic in the Environment, AS 2016*. pp. 12–14. doi:10.1201/b20466-7.

Jones, S.E., Ho, L., Rees, C.A., Hill, J.E., Nodwell, J.R. and Elliot, M.A. (2017) Streptomyces exploration is triggered by fungal interactions and volatile signals. *eLife*. 6 . doi:10.7554/eLife.21738.

Jonfia-Essien, W.A. (2006) Screening of new cocoa types for insect infestation and biochemical analysis of the stored beans. *Pakistan Journal of Biological Sciences*. 9 (14), pp. 2564–2571. doi:10.3923/pjbs.2006.2564.2571.

Kadow, D., Bohlmann, J., Phillips, W. and Lieberei, R. (2013) Identification of main fine or flavour components in two genotypes of the cocoa tree (Theobroma cacao L.). *Journal*

of Applied Botany and Food Quality. 86 pp. 90–98. doi:10.5073/JABFQ.2013.086.013.

Kalyuzhnaya, M.G., Lapidus, A., Ivanova, N., Copeland, A.C., McHardy, A.C., Szeto, E., Salamov, A., Grigoriev, I. V., Suciu, D., Levine, S.R., Markowitz, V.M., Rigoutsos, I., Tringe, S.G., Bruce, D.C., et al. (2008) High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*. doi:10.1038/nbt.1488.

Kamath, R. (2018) Food Traceability on Blockchain: Walmart's Pork and Mango Pilots with IBM. *The Journal of the British Blockchain Association*. 1 (1), pp. 1–12. doi:10.31585/jbba-1-1-(10)2018.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J.Y., Zhang, D., Engels, J.M.M.M. and Cronk, Q. (2012) Ultra-barcoding in cacao (theobroma spp.; malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany* [online]. 99 (2), pp. 320–329. Available from: http://www.amjbot.org/cgi/doi/10.3732/ajb.1100570doi:10.3732/ajb.1100570.

Kane, N.C., King, M.G., Barker, M.S., Raduski, A., Yatabe, Y., Knapp, S.J., Rieseberg, L.H. and True, J. (2010) *Comparative Genomic and Population Genetic Analyses Indicate Highly Porous Genomes and High*. 63 (8), pp. 2061–2075. doi:10.1111/j.1558-5646.2009.00703.x.COMPARATIVE.

Kawash, S. (2010) The candy prophylactic: danger, disease, and children's candy around 1916. *Journal of American culture*. doi:10.1111/j.1542-734x.2010.00742.x.

Keller, J., Rousseau-Gueutin, M., Martin, G.E., Morice, J., Boutte, J., Coissac, E., Ourari, M., Aïnouche, M., Salmon, A., Cabello-Hurtado, F. and Aïnouche, A. (2017) The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus. *DNA Research*. 24 (4), pp. 343–358. doi:10.1093/dnares/dsx006.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. and Glöckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research* [online]. 41 (1), pp. e1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22933715doi:10.1093/nar/gks808.

Knights, D., Parfrey, L.W., Zaneveld, J., Lozupone, C. and Knight, R. (2011) Human-associated microbial signatures: Examining their predictive value. *Cell Host and Microbe* [online].

10 (4), pp. 292–296. Available from: http://dx.doi.org/10.1016/j.chom.2011.09.003doi:10.1016/j.chom.2011.09.003.

Kongor, J.E., Hinneh, M., de Walle, D. Van, Afoakwa, E.O., Boeckx, P. and Dewettinck, K. (2016) Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile - A review *Food Research International* [online] 82 p.pp. 44–52. Available from: http://dx.doi.org/10.1016/j.foodres.2016.01.012doi:10.1016/j.foodres.2016.01.012.

Krapp, F., Wöhrmann, T., Pinangé, D.S.D.B., Benko-Iseppon, A.M., Huettel, B. and Weising, K. (2012) A set of plastid microsatellite loci for the genus Dyckia (Bromeliaceae) derived from 454 pyrosequencing. *American Journal of Botany*. 99 (12), pp. 2010–2013. doi:10.3732/ajb.1200153.

Kroeger, A., Koening, S., Thomson, A. and Streck, C. (2017) *Forest and climate smart cocoa in Côte d'Ivoire and Ghana: Aligning stakeholders to support smallholders in deforestation-free cocoa*. Available from: http://documents.worldbank.org/curated/en/317701513577699790/Forest-and-climate-smart-cocoa-in-Côte-D-Ivoire-and-Ghana-aligning-stakeholders-to-support-smallholders-in-deforestation-free-cocoa.

Küchler, S.M., Kehl, S. and Dettner, K. (2009) Characterization and localization of Rickettsia sp. in water beetles of genus Deronectes (Coleoptera: Dytiscidae). *FEMS Microbiology Ecology*. 68 (2), pp. 201–211. doi:10.1111/j.1574-6941.2009.00665.x.

Lalwani, S.K., Nunes, B., Chicksand, D. and Boojihawon, D.K. (2018) Benchmarking self-declared social sustainability initiatives in cocoa sourcing. *Benchmarking: An International Journal*. (just-accepted), pp. 0. doi:10.1108/BIJ-07-2017-0186.

Leal Filho, W. and Kovaleva, M. (2015) Research methods. In: *Environmental Science and Engineering (Subseries: Environmental Science)*. doi:10.1007/978-3-319-10906-0_5.

Lefeber, T., Gobert, W., Vrancken, G., Camu, N. and De Vuyst, L. (2011a) Dynamics and species diversity of communities of lactic acid bacteria and acetic acid bacteria during spontaneous cocoa bean fermentation in vessels. *Food Microbiology*. 28 (3), pp. 457–464. doi:10.1016/j.fm.2010.10.010.

Lefeber, T., Janssens, M., Moens, F., Gobert, W. and De Vuyst, L. (2011b) Interesting starter culture strains for controlled cocoa bean fermentation revealed by simulated cocoa pulp fermentations of cocoa-specific lactic acid bacteria. *Applied and Environmental Microbiology*. 77 (18), pp. 6694–6698. doi:10.1128/AEM.00594-11.

Lefeber, T., Papalexandratou, Z., Gobert, W., Camu, N. and De Vuyst, L. (2012) On-farm implementation of a starter culture for improved cocoa bean fermentation and its influence on the flavour of chocolates produced thereof. *Food Microbiology* [online]. 30 (2), pp. 379–392. Available from: http://dx.doi.org/10.1016/j.fm.2011.12.021doi:10.1016/j.fm.2011.12.021.

Li, L., Wieme, A., Spitaels, F., Balzarini, T., Nunes, O., Manaia, C., Van Landschoot, A., De Vuyst, L., Cleenwerck, I. and Vandamme, P. (2014) Acetobacter sicerae sp. nov., isolated from cider and kefir, and identification of species of the genus Acetobacter by dnaK, groEL and rpoB sequence analysis. *International Journal of Systematic and Evolutionary Microbiology*. doi:10.1099/ijs.0.058354-0.

Libbrecht, M.W. and Noble, W.S. (2015) Machine learning applications in genetics and genomics *Nature Reviews Genetics*. doi:10.1038/nrg3920.

Lim, T.K. (2012) Theobroma grandiflorum. *Edible Medicinal and Non-Medicinal Plants* [online]. 3 pp. 252–258. Available from: http://www.springerlink.com/index/10.1007/978-94-007-1764-0doi:10.1007/978-94-007-1764-0.

Lima, L.J.R., van der Velpen, V., Wolkers-Rooijackers, J., Kamphuis, H.J., Zwietering, M.H. and Rob Nout, M.J. (2012) Microbiota dynamics and diversity at different stages of industrial processing of cocoa beans into cocoa powder. *Applied and Environmental Microbiology*. 78 (8), pp. 2904–2913. doi:10.1128/AEM.07691-11.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., McDonald, D., Franzosa, E.A., Knight, R., White, O., et al. (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*. doi:10.1038/nature23889.

Lo, Y.T. and Shaw, P.C. (2018) DNA-based techniques for authentication of processed food and food supplements *Food Chemistry* 240 p.pp. 767–774. doi:10.1016/j.foodchem.2017.08.022.

Loor Solorzano, R.G., Fouet, O., Lemainque, A., Pavek, S., Boccara, M., Argout, X., Amores, F., Courtois, B., Risterucci, A.M. and Lanaud, C. (2012) Insight into the Wild Origin, Migration and Domestication History of the Fine Flavour Nacional Theobroma cacao L. Variety from Ecuador. *PLoS ONE*. 7 (11), . doi:10.1371/journal.pone.0048438.

Loureiro, G.A.H.A., Araujo, Q.R., Sodré, G.A., Valle, R.R., Souza, J.O., Ramos, E.M.L.S.,

Comerford, N.B. and Grierson, P.F. (2017) Cacao quality: Highlighting selected attributes *Food Reviews International* 33 (4) p.pp. 382–405. doi:10.1080/87559129.2016.1175011.

Ludlow, C.L.L., Cromie, G.A.A., Garmendia-Torres, C., Sirr, A., Hays, M., Field, C., Jeffery, E.W.W., Fay, J.C.C. and Dudley, A.M.A.M.A.M.A.M.A.M.A.M. (2016) Independent Origins of Yeast Associated with Coffee and Cacao Fermentation. *Current Biology* [online]. 26 (7), pp. 965–971. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0960982216300641doi:10.1016/j.cub.2016.02.012 [Accessed 17 September 2017].

Makhloufi, A. El, Mujica Mota, M., Damme, D. Van and Langenberg, V. (2018) *Towards a sustainable Agro-Logistics in developing countries: The case of cocoa's supply chain San Pedro Region/Côte d'Ivoire*.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R. and Peddada, S.D. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health & Disease*. 26 (1), pp. 27663. doi:10.3402/mehd.v26.27663.

Marco Iansiti and Karim R. Lakhani (2017) The Truth About Blockchain. *Harvard Business Review*.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*. doi:10.14806/ej.17.1.200.

Mattevi, M. and Jones, J. a. (2015) Traceability in the Food Supply Chain: Awareness and Attitudes of UK Small and Medium-sized Enterprises. *Food Control* [online]. 64 pp. 120–127. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0956713515303303doi:10.1016/j.foodcont.2015.12.014.

Maximova, S.N., Marelli, J.P., Young, A., Pishak, S., Verica, J.A. and Guiltinan, M.J. (2006) Over-expression of a cacao class I chitinase gene in Theobroma cacao L. enhances resistance against the pathogen, Colletotrichum gloeosporioides. *Planta*. 224 (4), pp. 740–749. doi:10.1007/s00425-005-0188-6.

McDonald, D., Clemente, J.C., Kuczynski, J., Rideout, J.R., Stombaugh, J., Wendel, D., Wilke, A., Huse, S., Hufnagle, J., Meyer, F., Knight, R. and Caporaso, J.G. (2012a) The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the omeome. *GigaScience*. 1 (1), pp. 7. doi:10.1186/2047-217X-1-7.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., Desantis, T.Z., Probst, A., Andersen, G.L., Knight, R. and Hugenholtz, P. (2012b) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME Journal*. doi:10.1038/ismej.2011.139.

McKinney, W. (2010) Data Structures for Statistical Computing in Python Stéfan van der Walt and Jarrod Millman (eds.). *Proceedings of the 9th Python in Science Conference* [online]. 1697900 (Scipy), pp. 51–56. Available from: http://conference.scipy.org/proceedings/scipy2010/mckinney.html.

Meersman, E., Steensels, J., Mathawan, M., Wittocx, P.J., Saels, V., Struyf, N., Bernaert, H., Vrancken, G. and Verstrepen, K.J. (2013) Detailed analysis of the microbial population in Malaysian spontaneous cocoa pulp fermentations reveals a core and variable microbiota. *PLoS ONE*. 8 (12), . doi:10.1371/journal.pone.0081559.

Meng, M., Stievano, L. and Lambert, J.F. (2004) Adsorption and thermal condensation mechanisms of amino acids on oxide supports. 1. Glycine on silica. *Langmuir*. doi:10.1021/la035336b.

Metzger, J.O. (2003) Book Review: Oils and Fats Authentication Edited by Michael Jee. *Angewandte Chemie International Edition*. 42 (15), pp. 1683–1684. doi:10.1002/anie.200390382.

Mirarab, S., Nguyen, N. and Warnow, T. (2012) SEPP: SATé-enabled phylogenetic placement. In: *Pacific Symposium on Biocomputing*. 2012 pp. 247–258.

Monteiro, C.A., Levy, R.B., Claro, R.M., Castro, I.R.R. de and Cannon, G. (2010) A new classification of foods based on the extent and purpose of their processing. *Cadernos de Saúde Pública*. doi:10.1590/s0102-311x2010001100005.

Moreano, F., Busch, U. and Engel, K.H. (2005) Distortion of genetically modified organism quantification in processed foods: Influence of particle size compositions and heat-induced DNA degradation. *Journal of Agricultural and Food Chemistry*. doi:10.1021/jf051894f.

Moreira, I.M. da V., Miguel, M.G. da C.P., Duarte, W.F., Dias, D.R. and Schwan, R.F. (2013) Microbial succession and the dynamics of metabolites and sugars during the fermentation of three different cocoa (Theobroma cacao L.) hybrids. *Food Research International* [online]. 54 (1), pp. 9–17. Available from: http://dx.doi.org/10.1016/j.foodres.2013.06.001doi:10.1016/j.foodres.2013.06.001.

Moreira, P.A. and Oliveira, D.A. (2011) Leaf age affects the quality of DNA extracted from Dimorphandra mollis (Fabaceae), a tropical tree species from the Cerrado region of Brazil. *Genetics and molecular research : GMR*. doi:10.4238/vol10-1gmr1030.

Mossu, G. (1990) Le cacaoyer *L'amélioration des plantes tropicales*.

Motamayor, J.-C., Lachenaud, P., Wallace, J., Loor-Solórzano, R.G., Martinez, W.J., Graham, J., Kuhn, D.N., Brown, S. and Schnell, R.J. (2010) No Mas ' Forastero ': a New Protocol for Meaningful Cacao Germplasm Classification. *Proceedings of the 16th International Cacao Research Conference, Denpassar, Bali 16-21st November 2009*.

Motamayor, J.C., Lachenaud, P., da Silva e Mota, J.W., Loor, R., Kuhn, D.N., Brown, J.S. and Schnell, R.J. (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (Theobroma cacao L). *PLoS ONE*. 3 (10), . doi:10.1371/journal.pone.0003311.

Motamayor, J.C., Risterucci, A.M., Lopez, P.A., Ortiz, C.F., Moreno, A. and Lanaud, C. (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* [online]. 89 (5), pp. 380–386. Available from: http://www.nature.com/articles/6800156doi:10.1038/sj.hdy.6800156.

Motilal, L. and Butler, D. (2003) Verification of identities in global cacao germplasm collections. *Genetic Resources and Crop Evolution*. 50 (8), pp. 799–807. doi:10.1023/A:1025950902827.

Mujica Mota, M., El Makhloufi, A. and Scala, P. (2019) On the logistics of cocoa supply chain in Côte d'Ivoire: Simulation-based analysis. *Computers and Industrial Engineering*. doi:10.1016/j.cie.2019.106034.

National Research Council (US) Committee on DNA Technology (1992) *DNA Technology in Forensic Science* [online].

Nelson, V. and Phillips, D. (2018) Sector, Landscape or Rural Transformations? Exploring the Limits and Potential of Agricultural Sustainability Initiatives through a Cocoa Case Study. *Business Strategy and the Environment*. doi:10.1002/bse.2014.

Nielsen, D.S., Teniola, O.D., Ban-Koffi, L., Owusu, M., Andersson, T.S. and Holzapfel, W.H. (2007) The microbiology of Ghanaian cocoa fermentations analysed using culture-dependent and culture-independent methods. *International Journal of Food Microbiology*. 114 (2), pp. 168–186. doi:10.1016/j.ijfoodmicro.2006.09.010.

Ogier, J.C., Pagès, S., Galan, M., Barret, M. and Gaudriault, S. (2019) RpoB, a promising marker for analyzing the diversity of bacterial communities by amplicon sequencing. *BMC Microbiology*. doi:10.1186/s12866-019-1546-z.

de Oliveira, T.B. and Genovese, M.I. (2013) Chemical composition of cupuassu (Theobroma grandiflorum) and cocoa (Theobroma cacao) liquors and their effects on streptozotocin-induced diabetic rats. *Food Research International*. doi:10.1016/j.foodres.2013.02.019.

Orcher, L.T. (2007) Conducting Research: Social and Behavioral Science Methods. In: *Routledge*.

Otter, V., Prechtel, B. and Theuvsen, L. (2014) The country-of-origin effect for chocolate made from ecuadorian cocoa: An empirical analysis of consumer perceptions. *Economia Agro-Alimentare*. doi:10.3280/ECAG2014-003005.

Owusu, M., Petersen, M.A. and Heimdal, H. (2013) Relationship of sensory and instrumental aroma measurements of dark chocolate as influenced by fermentation method, roasting and conching conditions. *Journal of Food Science and Technology*. 50 (5), pp. 909–917. doi:10.1007/s13197-011-0420-2.

Özgen Arun, Ö., Yilmaz, F. and Muratoĝlu, K. (2013) PCR detection of genetically modified maize and soy in mildly and highly processed foods. *Food Control*. doi:10.1016/j.foodcont.2013.01.023.

Ozturk, G. and Young, G.M. (2017) Food Evolution: The Impact of Society and Science on the Fermentation of Cocoa Beans *Comprehensive Reviews in Food Science and Food Safety* 16 (3) p.pp. 431–455. doi:10.1111/1541-4337.12264.

Packard, H., Taylor, Z.W., Williams, S.L., Guimarães, P.I., Toth, J., Jensen, R. V., Senger, R.S., Kuhn, D.D. and Stevens, A.M. (2019) Identification of soil bacteria capable of utilizing a corn ethanol fermentation byproduct. *PLoS ONE*. doi:10.1371/journal.pone.0212685.

Papalexandratou, Z., Camu, N., Falony, G. and De Vuyst, L. (2011a) Comparison of the bacterial species diversity of spontaneous cocoa bean fermentations carried out at selected farms in Ivory Coast and Brazil. *Food Microbiology*. 28 (5), pp. 964–973. doi:10.1016/j.fm.2011.01.010.

Papalexandratou, Z. and Nielsen, D.S. (2016) It's Gettin' Hot in Here: Breeding Robust Yeast Starter Cultures for Cocoa Fermentation *Trends in Microbiology* 24 (3) p.pp. 168–170. doi:10.1016/j.tim.2016.01.003.

Papalexandratou, Z., Vrancken, G., de Bruyne, K., Vandamme, P. and de Vuyst, L. (2011b) Spontaneous organic cocoa bean box fermentations in Brazil are characterized by a restricted species diversity of lactic acid bacteria and acetic acid bacteria. *Food Microbiology* [online]. 28 (7), pp. 1326–1338. Available from: http://dx.doi.org/10.1016/j.fm.2011.06.003doi:10.1016/j.fm.2011.06.003.

Paulin, D., Decazy, B. and Coulibaly, N. (1983) Etude des variations saisonnieres des conditions de pollinisation et de fructification dans une cacaoyer. *Cafe Cacao The*. 27 pp. 165–176.

Paun, V.I., Icaza, G., Lavin, P., Marin, C., Tudorache, A., Persoiu, A., Dorador, C. and Purcarea, C. (2019) Total and potentially active bacterial communities entrapped in a late glacial through holocene ice core from scarisoara ice Cave, Romania. *Frontiers in Microbiology*. doi:10.3389/fmicb.2019.01193.

Peakall, R. and Smouse, P. (2012) *6.5 ©2006*. (2006), pp. 2537–2539.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., et al. (2011) Scikit-learn: Machine learning in Python. *Journal of machine learning research*. 12 (Oct), pp. 2825–2830.

Perlin, M. (2015) Inclusion probability for DNA mixtures is a subjective one-sided match statistic unrelated to identification information. *Journal of Pathology Informatics*. 6 (1), pp. 59. doi:10.4103/2153-3539.168525.

Petiard, V. (no date) *Use of DNA identification techniques for the determination of genetic material of cocoa in fermented or roasted beans and chocolate*.

Petyaev, I.M. and Bashmakov, Y.K. (2017) Dark Chocolate: Opportunity for an Alliance between Medical Science and the Food Industry? *Frontiers in Nutrition*. doi:10.3389/fnut.2017.00043.

Pinto, C., Pinho, D., Cardoso, R., Custódio, V., Fernandes, J., Sousa, S., Pinheiro, M., Egas, C. and Gomes, A.C. (2015) Wine fermentation microbiome: A landscape from different Portuguese wine appellations. *Frontiers in Microbiology*. doi:10.3389/fmicb.2015.00905.

Planý, M., Kuchta, T., Šoltýs, K., Szemes, T., Pangallo, D. and Siekel, P. (2016) Metagenomic analysis of slovak bryndza cheese using next-generation 16S rDNA amplicon sequencing. *Nova Biotechnologica et Chimica*. 15 (1), pp. 23–34. doi:10.1515/nbec-2016-0003.

Powis, T.G., Hurst, W.J., del Carmen Rodriguez, M., Ortiz Ceballos, P., Blake, M., Cheetham, D., Coe, M.D. and Hodgson, J.G. (2007) Oldest chocolate in the New World. *Antiquity* [online]. 81 (314), pp. http://www.antiquity.ac.uk/ProjGall/powis/index.ht. Available from: http://www.antiquity.ac.uk/ProjGall/powis/index.html.

Price, M.N., Dehal, P.S. and Arkin, A.P. (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE*. doi:10.1371/journal.pone.0009490.

Primrose, S., Woolfe, M. and Rollinson, S. (2010) Food forensics: Methods for determining the authenticity of foodstuffs *Trends in Food Science and Technology*. doi:10.1016/j.tifs.2010.09.006.

PROECUADOR (2015) Perfil Sectorial de Cacao Y Ecuador un mundo de oportunidades ¿ Por qué invertir en Ecuador ? *Perfil Sectorial De Cacao* [online]. Available from: http://www.proecuador.gob.ec/wp-content/uploads/2014/07/PERFIL-DE-CACAO-Y-ELABORADOS.pdf.

Recanati, F., Marveggio, D. and Dotelli, G. (2018) From beans to bar: A life cycle assessment towards sustainable chocolate supply chain. *Science of the Total Environment* [online]. 613–614 pp. 1013–1023. Available from: https://doi.org/10.1016/j.scitotenv.2017.09.187doi:10.1016/j.scitotenv.2017.09.187.

Reis, J.A., Paula, A.T., Casarotti, S.N. and Penna, A.L.B. (2012) Lactic Acid Bacteria Antimicrobial Compounds: Characteristics and Applications *Food Engineering Reviews*. doi:10.1007/s12393-012-9051-2.

Rey, G., Lachenaud, P., Fouet, O., Argout, X., Peña, G., Macias, J.C., Amores, F.M., Lanaud, C., Valdez, F. and Hurtado, J. (2015) Rescue of Cacao Genetic Resources Related to the Nactional Variety: Surveys in the Ecuadorian Amazon. *Espamciencia*. 6 (E), pp. 7–15.

Richards, P., Fothergill, J., Bernardeau, M. and Wigley, P. (2019) Development of the caecal microbiota in three broiler breeds. *Frontiers in Veterinary Science*. doi:10.3389/fvets.2019.00201.

Risterucci, a.-M., Grivet, L., N'Goran, J. a K., Pieretti, I., Flament, M.H. and Lanaud, C. (2000) A high density linkage map of Theobroma cacao L. *Theoretical and Applied Genetics*. doi:10.1007/s001220051566.

Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M. and Calle, M.L. (2018) Balances: a New Perspective for Microbiome Analysis. *mSystems*. 3 (4), pp.

1–12. doi:10.1128/msystems.00053-18.

De Roos, J. and De Vuyst, L. (2018) Acetic acid bacteria in fermented foods and beverages. *Current Opinion in Biotechnology*. 49 pp. 115–119. doi:10.1016/j.copbio.2017.08.007.

Rosman, N.N., Mokhtar, N.F.K., Ali, M.E. and Mustafa, S. (2016) Inhibitory Effect of Chocolate Components Toward Lard Detection in Chocolate Using Real Time PCR. *International Journal of Food Properties*. doi:10.1080/10942912.2015.1137936.

Rousseau, S. (2015) The role of organic and fair trade labels when choosing chocolate. *Food Quality and Preference*. 44 pp. 92–100. doi:10.1016/j.foodqual.2015.04.002.

Saltini, R., Akkerman, R. and Frosch, S. (2013) Optimizing chocolate production through traceability: A review of the influence of farming practices on cocoa bean quality. *Food Control* [online]. 29 (1), pp. 167–187. Available from: http://dx.doi.org/10.1016/j.foodcont.2012.05.054doi:10.1016/j.foodcont.2012.05.054.

Savazzini, F. and Martinelli, L. (2006) DNA analysis in wines: Development of methods for enhanced extraction and real-time polymerase chain reaction quantification. *Analytica Chimica Acta*. doi:10.1016/j.aca.2005.10.078.

Schiefenhövel, K. and Rehbein, H. (2013) Differentiation of Sparidae species by DNA sequence analysis, PCR-SSCP and IEF of sarcoplasmic proteins. *Food Chemistry*. doi:10.1016/j.foodchem.2012.10.057.

Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A., Chen, W., Bolchacova, E., Voigt, K., Crous, P.W., Miller, A.N., Wingfield, M.J., Aime, M.C., An, K.D., et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.1117018109.

Schroeder, H., Cronn, R., Yanbaev, Y., Jennings, T., Mader, M., Degen, B. and Kersten, B. (2016) Development of molecular markers for determining continental origin of wood from White Oaks (Quercus L. sect. Quercus). *PLoS ONE*. 11 (6), pp. 1–15. doi:10.1371/journal.pone.0158221.

Seguine, E. and Meinhardt, L. (2014) Cacao Flavor Through Genetics — Anatomy of Fine Flavor *The Manufacturing Confectioner* [online] (November) p.pp. 25–30. Available from: http://www.finechocolateindustry.org/Resources/Documents/mc_HCP_2014_12_1.pdf.

Shi, P., Zhang, A. and Li, H. (2016) Regression analysis for microbiome compositional data. *Annals of Applied Statistics*. 10 (2), pp. 1019–1040. doi:10.1214/16-AOAS928.

Silva, A.R. de A., Bioto, A.S., Efraim, P. and Queiroz, G. de C. (2014) Impact of sustainability labeling in purchase intention and quality perception of dark chocolate. *Proceedings of the 9th International Conference on Life Cycle Assessment in the Agri-Food Sector (LCA Food 2014), San Francisco, California, USA, 8-10 October, 2014*.

Smulders, M.J.M., Esselink, D., Amores, F., Ramos, G., Sukha, D. a, Butler, D.R., Vosman, B. and Van Loo, E.N. (2009) Identification of Cocoa (Theobroma cacao L.) Varieties with Different Quality Attributes and Parentage Analysis of Their Beans. *INGENIC Newsletter*. (12), pp. 1–13.

Somerfield, P.J. and Clarke, K.R. (1997) A comparison of some methods commonly used for the collection of sublittoral sediments and their associated fauna. *Marine Environmental Research*. 43 (3), pp. 145–156. doi:10.1016/0141-1136(96)00083-9.

Song, S.L., Lim, P.E., Phang, S.M., Lee, W.W., Hong, D.D. and Prathep, A. (2014) Development of chloroplast simple sequence repeats (cpSSRs) for the intraspecific study of Gracilaria tenuistipitata (Gracilariales, Rhodophyta) from different populations. *BMC Research Notes* [online]. 7 (1), pp. 1–9. Available from: BMC Research Notesdoi:10.1186/1756-0500-7-77.

Sonwa, D.J., Weise, S.F., Schroth, G., Janssens, M.J.J. and Shapiro, H.-Y. (2019a) Structure of cocoa farming systems in West and Central Africa: a review. *Agroforestry Systems*. 93 pp. 2009–2025. doi:10.1007/s10457-018-0306-7.

Sonwa, D.J., Weise, S.F., Schroth, G., Janssens, M.J.J. and Shapiro, H.-Y.Y. (2019b) Structure of cocoa farming systems in West and Central Africa: a review. *Agroforestry Systems*. 93 (5), pp. 2009–2025. doi:10.1007/s10457-018-0306-7.

Sorond, F.A., Lipsitz, L.A., Hollenberg, N.K. and Fisher, N.D.L. (2008) Cerebral blood flow response to flavanol-rich cocoa in healthy elderly humans. *Neuropsychiatric Disease and Treatment*. doi:10.2147/ndt.s2310.

Squicciarini, M.P. and Swinnen, J. (2016) The Economics of Chocolate. In: *The Economics of Chocolate*. (no place) Oxford University Press. pp. 1–8. doi:10.1093/acprof:oso/9780198726449.003.0001.

Sukha, D.A., Butler, D.R., Comissiong, E.A. and Umaharan, P. (2014) The impact of processing

location and growing environment on flavor in cocoa (Theobroma cacao L.) - Implications for 'terroir' and certification - Processing location study. In: *Acta Horticulturae*. 2014 doi:10.17660/ActaHortic.2014.1047.31.

Sukha, D.A., Butler, D.R., Umaharan, P. and Boult, E. (2008) The use of an optimised organoleptic assessment protocol to describe and quantify different flavour attributes of cocoa liquors made from Ghana and Trinitario beans. *European Food Research and Technology*. doi:10.1007/s00217-006-0551-2.

Swan, A.L., Mobasheri, A., Allaway, D., Liddell, S. and Bacardit, J. (2013) Application of machine learning to proteomics data: Classification and biomarker identification in postgenomics biology *OMICS A Journal of Integrative Biology*. doi:10.1089/omi.2013.0017.

Teye, E., Anyidoho, E., Agbemafle, R., Sam-Amoah, L.K. and Elliott, C. (2020) Cocoa bean and cocoa bean products quality evaluation by NIR spectroscopy and chemometrics: A review *Infrared Physics and Technology*. doi:10.1016/j.infrared.2019.103127.

UTZ (2016) *UTZ*. Available from: https://www.utz.org/.

Vail, G. (2008) Cacao use in Yucatán among the Pre-Hispanic Maya. In: *Chocolate: History, Culture, and Heritage*. pp. 3–15. doi:10.1002/9780470411315.ch1.

Vandeventer, P.E., Mejia, J., Nadim, A., Johal, M.S. and Niemz, A. (2013) DNA adsorption to and elution from silica surfaces: Influence of amino acid buffers. *Journal of Physical Chemistry B*. doi:10.1021/jp405753m.

Vázquez-Baeza, Y., Gonzalez, A., Xu, Z.Z., Washburne, A., Herfarth, H.H., Sartor, R.B. and Knight, R. (2018) Guiding longitudinal sampling in IBD cohorts *Gut*. doi:10.1136/gutjnl-2017-315352.

de Vere, N., Rich, T.C.G.G., Ford, C.R., Trinder, S.A., Long, C., Moore, C.W., Satterthwaite, D., Davies, H., Allainguillaume, J., Ronca, S., Tatarinova, T., Garbett, H., Walker, K. and Wilkinson, M.J. (2012) DNA barcoding the native flowering plants and conifers of wales. *PLoS ONE*. 7 (6), pp. 1–12. doi:10.1371/journal.pone.0037945.

Vlam, M., de Groot, G.A., Boom, A., Copini, P., Laros, I., Veldhuijzen, K., Zakamdi, D. and Zuidema, P.A. (2018) Developing forensic tools for an African timber: Regional origin is revealed by genetic characteristics, but not by isotopic signature. *Biological Conservation* [online]. 220 (January), pp. 262–271. Available from: https://doi.org/10.1016/j.biocon.2018.01.031doi:10.1016/j.biocon.2018.01.031.

De Vuyst, L. and Weckx, S. (2016) The cocoa bean fermentation process: from ecosystem analysis to starter culture development *Journal of Applied Microbiology* 121 (1) p.pp. 5–17. doi:10.1111/jam.13045.

Warner, K., Timme, W., Lowell, B. and Hirshfield, M. (2013) Oceana study reveals seafood fraud nationwide. *Oceana*. (February), pp. 1–69.

Willis, A.D. (2017) Rarefaction, alpha diversity, and statistics. *bioRxiv* [online]. (1968), pp. 231878. Available from: https://www.biorxiv.org/content/10.1101/231878v1.fulldoi:10.1101/231878 [Accessed 24 December 2019].

Willis, A.D. (2019) Rarefaction, alpha diversity, and statistics. *Frontiers in Microbiology*. doi:10.3389/fmicb.2019.02407.

Woolfe, M. and Primrose, S. (2004) Food forensics: Using DNA technology to combat misdescription and fraud *Trends in Biotechnology* 22 (5) p.pp. 222–226. doi:10.1016/j.tibtech.2004.03.010.

Yagi, Y. and Shiina, T. (2014) Recent advances in the study of chloroplast gene expression and its evolution. *Frontiers in Plant Science* [online]. 5 (February), pp. 1–7. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2014.00061/abstractdoi:10.3389/fpls.2014.00061.

Yanagawa, R. and Honda, E. (1978) Corynebacterium pilosum and Corynebacterium cystitidis, two new species from cows. *International Journal of Systematic Bacteriology*. doi:10.1099/00207713-28-2-209.

Yang, J.Y., Motilal, L.A., Dempewolf, H., Maharaj, K. and Cronk, Q.C.B. (2011) Chloroplast Microsatellite Primers for Cacao (Theobroma cacao) and other Malvaceae. *American Journal of Botany*. 98 (12), pp. 372–374. doi:10.3732/ajb.1100306.

Yang, J.Y., Scascitelli, M., Motilal, L.A., Sveinsson, S., Engels, J.M.M., Kane, N.C., Dempewolf, H., Zhang, D., Maharaj, K. and Cronk, Q.C.B. (2013) Complex origin of Trinitario-type Theobroma cacao (Malvaceae) from Trinidad and Tobago revealed using plastid genomics. *Tree Genetics and Genomes*. 9 (3), pp. 829–840. doi:10.1007/s11295-013-0601-4.

Zarrillo, S., Gaikwad, N., Lanaud, C., Powis, T., Viot, C., Lesur, I., Fouet, O., Argout, X., Guichoux, E., Salin, F., Solorzano, R.L., Bouchez, O., Vignes, H., Severts, P., et al. (2018)

The use and domestication of Theobroma cacao during the mid-Holocene in the upper Amazon. *Nature Ecology and Evolution* doi:10.1038/s41559-018-0697-x.

Zhang, D., Boccara, M., Motilal, L., Mischke, S., Johnson, E.S., Butler, D.R., Bailey, B. and Meinhardt, L. (2009) Molecular characterization of an earliest cacao (Theobroma cacao L.) collection from Upper Amazon using microsatellite DNA markers. *Tree Genetics and Genomes*. 5 (4), pp. 595–607. doi:10.1007/s11295-009-0212-2.

Appendices

Appendix I: Interview schedule

1) What is your name and company that you represent?

2) How long have you been working in your current position? How long have you been working in the cacao research or industry?
3) Are you or your institution part of any regulatory committees, consultant, cacao certifier or governmental legislative body? If yes, please can you explain the role you or the organization has?

4) Why is traceability a critical control, is traceability a concern for you? Why? (Traceability: "The ability to follow the movement of a food through specified stage(s) of production, processing and distribution" FAO/WHO).

5) How does your institution or enterprise study the supply chain and verify the critical controls of the cacao beans and chocolate?

6) Which are the most important parameters to analyse for tracking back cacao beans? (Tools and systems)
7) How do you verify this information and how often do you think it should be verify?

8) What are the challenges you currently face regarding the traceability of cacao?

9) What kind of data do you receive from any tracking analysis requirement related to beans or chocolate? How well do these data meet your needs?

10) Do you believe that the actual certifications can verify 100% the origin of the product? If not, what would be your approach?

11) Do you consider that having a traceability control have an impact on the value chain of the cacao products and price?

12) Which industry is more affected by having wrong geographical regions or traceable cacao?
    a) Small Chocolate makers
    b) Industrial Chocolate makers
    c) Confectionaries
    d) Cacao Traders
    e) Grinders & other processors
    f) Others: Explain

13) Could you please identify for which regions a traceability control would be more beneficial or should it be compulsory? Please do not repeat the same value. (Scale importance 1 min – 8 max)

South America	Central America	Caribbean	Pacific Islands
Asia	Australia	Africa	Indian Oceans

14) We are keen to involve stakeholders throughout the process of developing (Genetic markers methodology, to identify the origin of the cacao beans and chocolate as a quality control technology for the industry). Would you like to be involved further with the project? If so, how would you like to contribute? What is the best way to reach you?

Appendix II: Interviewees

Table 0.1 Appendix II: Stakeholders (23) by principal activity which answer interviews, questionnaires or provide feedback

Interviewees	Organization type	Respondent role	Country
Interviewee 1	Cacao Farmer	Cacao Producer	Bolivia
Interviewee 2	Trading and quality control	Trader	Colombia
Interviewee 3	Governmental	Nacional research cacao coordinator	Ecuador
Interviewee 4	Governmental	Embassy Counsellor	Ecuador
Interviewee 5	Governmental	Trade Commissioner	Ecuador
Interviewee 6	Cooperative trader	Trader	Ecuador
Interviewee 7	Cacao Farmer	Cacao Producer and cooperative representative	Ecuador
Interviewee 8	Cacao Farmer and chocolate trading	Cacao Producer	Ecuador
Interviewee 9	Cacao Farmer	Cacao Producer	Ecuador
Interviewee 10	Chocolate maker	Cacao and sustainability researcher & development	France

Interviewee 11	Chocolate maker	Technical supervision and quality assurance cocoa and chocolate	Germany
Interviewee 12	Chocolate machinery and products	Director and sales manager	Italy
Interviewee 13	Research and Development - NGO	Coordination of the Global Network for Cacao Genetic Resources and the Cocoa of Excellence Programme	Italy
Interviewee 14	Policy making	Chair-person from ISO Cocoa	Netherlands
Interviewee 15	Post-harvest and trading	General Manager	Nicaragua
Interviewee 16	Chocolate trader	Chocolate trader - chocolatier	Norway
Interviewee 17	Chocolate trader	Partner and Co-Founder	UK
Interviewee 18	Consultancy and research	Chairman - Director	UK
Interviewee 19	Chocolate makers	Global Head Agronomy	UK
Interviewee 20	Chocolate maker	Director - Chocolate Maker	USA
Interviewee 21	Consultancy	Consultant – Chief chocolate correspondent	USA
Interviewee 22	Research and academy	Researcher - Director	USA
Interviewee 23	Chocolate maker	Founder – Chocolate maker	USA

Appendix III: Chocolate samples for extractions and Metagenomcis

Table 0.2 Appendix III: 1- Chocolate samples selected for DNA extraction: 81 cacao products were extracted for protocol optimization and further investigations in metagenomics analysis.

Code	Country	Origin	Cacao Solids	Description	Year
1	Bolivia	Riberalta	70%	Riberalta, Amazonia North	2015
2	Bolivia	Riberalta	70%	Riberalta, Amazonia North	2015
3	Colombia	San Andres de Tumaco	70%	Nariño	2015
4	Colombia	Arauca	70%	Bajo Cusay, Tame, Arauca, Piedemonte Llanero, Colombia	2017

5	Colombia	Arauca	70%	La Pica, Arauquita, Arauca, Piedemonte Llanero, Colombia	2017
6	Colombia	Arauca	70%	Bajo Cusay, Tame, Arauca, Piedemonte Llanero, Colombia	2017
CH TUMACO 2	Colombia	San Andres de Tumaco	82%		2018
CH CAUCA 1	Colombia	San Andres de Tumaco	100%		2018
48	Ivory Coast	Duékoué	70%	Guemon	2015
49	Ivory Coast	Divo	70%	Lôh Djiboua	2015
50	Ivory Coast	Duékoué	70%	Guemon	2015
51	Ivory Coast	Divo	70%	Bp 36 Divo, Village: Konankro (Cnra), Lôh Djiboua, Ivory Coast	2017

Table 0.3 Continuation 2 - Chocolate samples selected for DNA extraction: 81 cacao products were extracted for protocol optimization and further investigations in metagenomics analysis

Code	Country	Origin	Cacao Solids	Description	Year
10	Ecuador	Ventanas, Los Ríos	70%	Coastal	2015
11	Ecuador	Vinces, Los Ríos	70%	Coastal	2015
12	Ecuador	Balao, Guayas	70%	Coastal	2015
13	Ecuador	Calceta, Ecuador	70%	Km 1,5 Via Calceta-Canuto, Calceta, Manabi, Costa, Ecuador	2017

14	Ecuador	Vinces, Ecuador	70%	Km 1,5 Via Vinces - Guayaquil, Vinces (Antonio Sotomayor), Los Rios, Costa, Ecuador	2017
15	Ecuador	Quiroga, Ecuador	70%	Km 5 Via A Quiroga, Quiroga, Manabi, Costa, Ecuador	2017
TOW 1	Ecuador	Esmeraldas, Ecuador	70%	Coin Rojo	2017
TOW 2	Ecuador	Esmeraldas, Ecuador	70%	Bloque	2017
TOW 3	Ecuador	Esmeraldas, Ecuador	70%	Bars	2017
TOW 4	Ecuador	Guayas	100%	Chips café	2017
TOW 6	Ecuador	Balao, Guayas	70%	minibarras	2017
TOW 7	Ecuador	Guayas	100%	nibs	2017
TOW 8	Ecuador	El Carmen, Manabí,	86%	Manabi, Esmeraldas, Gye	2017
TOW 9	Ecuador	El Carmen, Manabí,	70%	Sport Bars BCA Manabi, Esmeraldas, Balao, Ecuador	2017
TOW 13	Ecuador	Guayas	100%	cacao butter	2017
TOW 14	Ecuador	El Carmen, Manabí,	86%	Manabi, Esmeraldas, Gye	2017
CX 1	Ecuador	Los Rios, Ecuador	70%	EDC	2018

Table 0.4 Continuation 3 - Chocolate samples selected for DNA extraction: 81 cacao products were extracted for protocol optimization and further investigations in metagenomics analysis

Code	Country	Origin	Cacao Solids	Year
9	Ecuador	Quiroga, Manabi	70%	2015
44	Ghana	Ahafo Ano South, Ghana	70%	2015
45	Ghana	Old Tafo, Ghana	70%	2015
46	Ghana	Kukurantumi, Ghana	70%	2017
47	Ghana	Ahafo Ano South, Ghana	70%	2017

42	Guatemala	Rio Dulce, Guatemala	70%	2015
43	Guatemala	Río Dulce, Izabal, Guatemala	70%	2017
33	Haiti	Port-Margot, Haiti	70%	2015
34	Haiti	Ouanaminthe, Haiti	70%	2017
37	Honduras	Yoro ,Honduras	70%	2015
38	Honduras	La Másica, Honduras	70%	2015
39	Honduras	La Masica, Atlantida, Honduras	70%	2017
INSP COVERT FRAISE 1	Honduras	Guanaja, Honduras	38%	2018
INSP COVERT AMANDE 1	Honduras	Guanaja, Honduras	31%	2018
54	Indonesia	Kecamatan Guguk, Indonesia	70%	2015

Table 0.5 Continuation 4 - Chocolate samples selected for DNA extraction: 81 cacao products were extracted for protocol optimization and further investigations in metagenomics analysis

Code	Country	Origin	Cacao Solids	Year
55	Indonesia	Jembrana, Bali	70%	2017
56	Malaysia	Ranau, Malaysia	70%	2015
57	Malaysia	Ranau, Malaysia	70%	2017
27	Peru	Castilla, Piura	70%	2015
28	Peru	Castilla, Piura	70%	2015
29	Peru	Cuzco, Peru	70%	2017
30	Peru	Morropón, Piura, Peru	70%	2017
MRN CSC 1	Peru	Cuzco, Peru	80%	2018
QRN CSC 1	Peru	Cuzco, Peru	0.72	2018
35	Trinidad and Tobago	Gran Couva,Trinidad and Tobago	70%	2015
36	Trinidad and Tobago	Gran Couva, Trinidad and Tobago	70%	2017
52	Vietnam	Ha Long, Vietnam	70%	2015
53	Vietnam	Tan Phu, Vietnam	70%	2017
HCV 70%	Venezuela		70%	2013
HCT 75 %	Trinidad and Tobago		75%	2013
HCVI 80 %	Vietnam		80%	2013
HCCE 100 %	Ecuador		100%	2013

Table 0.6 Continuation 5 - Chocolate samples selected for DNA extraction: 81 cacao products were extracted for protocol optimization and further investigations in metagenomics analysis

Code	Country	Origin	Cacao Solids	Description	Year
HCSL 100 %	Saint Lucia		100%	JA	2013
HCM 72 %	Madagascar		72%	JA	2013
HCSL 70 %	Saint Lucia		70%	JA	2013
HCP 75 %	Peru		75%	JA	2013
M 73 %	Mexico			JA	2013
IWC 2	Indonesia		69%	JA	2013
VWC 2	Venezuela		72%	JA	2013
VWC 3	Venezuela		70%	JA	2013
IWC 3	Indonesia		0.7	JA	2013
Mars	Mars		20%	JA	2013
Kit kat / nestle	Ghana		0.2	JA	2013
crisp / cadbury	Ghana		20%	JA	2013
HCP100 %2	Peru		100%	JA	2013
HCCE 100% 2	Ecuador		100%	JA	2013
HCDR 100 % 2	Dominican Republic		100%	JA	2013
HCT 75 % 2	Trinidad and Tobago		75%	JA	2013
HCV70 % 2	Venezuela		70%	JA	2013
HCV80 % 2	Vietnam		80%	JA	2013
HCJ74 % 2	Java		74%	JA	2013
HCM72 % 2	Madagascar		0.72	JA	2013

Appendix IV: Chocolate samples DNA quantification

Table 0.7 7.3 Appendix IV: 1- Sixty samples used for comparative analysis between Nanodrop and Qubit™

Code	NanoDrop	Qubit™	Cocoa Solids
Kit kat/Nestle	2.6	0.2	20%
Crisp/Cadbury	2.7	0.3	20%
INSP COVERT FRAISE 1	17.3	3.34	34%
INSP COVERT AMANDE 1	18.9	5.54	34%
1	2.2	0.13	70%
2	2.7	0.173	70%
3	**2.4**	**0.11**	**0.7**
4	2.8	0.099	70%
5	2.4	0.09	70%
6	5.7	0.26	70%
48	4.5	0.256	70%
49	6.6	0.414	70%
50	7.6	0.226	70%
51	**3.8**	**0.288**	**0.7**
9	3	0.224	70%
13	2.1	0.224	70%
14	2.9	0.175	70%
15	4.1	0.155	70%
TOW 1	6.8	0.648	70%
TOW 2	16.6	0.89	70%
TOW 3	**15.2**	**1.5**	**0.7**
TOW 6	10	0.59	70%
TOW 9	37.8	0.227	70%
CX 1	7.9	0.776	70%
44	9.6	0.772	70%
45	5.4	0.536	70%
46	5.6	0.0304	70%
47	**5.1**	**0.256**	**0.7**
42	3.6	0.154	70%
43	8.5	0.168	70%
33	10.8	0.358	70%
34	4.9	0.35	70%
39	12.8	0.816	70%
54	4.3	0.0193	70%
55	**5.1**	**0.454**	**0.7**
56	6.6	0.284	70%
57	6.9	0.177	70%
27	2.5	0.234	70%

Table 0.8 Continuation 2 - Sixty samples used for comparative analysis between Nanodrop and Qubit™

Code	NanoDrop	Qubit™	Cocoa Solids
28	6.3	0.164	70%
29	9.1	1.22	70%
30	3.4	0.258	70%
35	12	0.804	70%
36	11.5	0.118	70%
HCV70 % 2	10.4	0.9	70%
QRN CSC 1	**5.4**	**0.516**	**0.72**
HCM72 % 2	10.5	1	72%
HCJ74 % 2	10.4	0.9	74%
HCT 75 % 2	10.4	0.9	75%
MRN CSC 1	**9.8**	**0.442**	**0.8**
HCV80 % 2	10.4	0.9	80%
CH TUMACO 2	3.5	0.29	82%
TOW 8	39.4	1.53	86%
TOW 14	13.8	1.55	86%
CH CAUCA 1	3.7	0.124	100%
TOW 4	30.2	0.55	100%
TOW 7	**33.7**	**0.862**	**1**
TOW 13	2.9	0.0216	100%
HCP100 %2	2.8	0.2	100%
HCCE 100% 2	2.9	0.3	100%
HCDR 100 % 2	**10.3**	**0.9**	**1**

Appendix V: RFU Chocolate vs CB

Table 0.9 7.4 Appendix V: Microsatellites (cpSSR4, cpSSR 14, cpSSR3) relative fluorescence units (RFU) comparison between chocolate (TOW 4) and cacao butter (TOW 13)

Each microsatellite was assessed by capillary analysis in 118 chocolate samples. It showed to amplify (4114.479118; 9433.387; 8161.067) respectively while two samples from TOW 13 (cacao butter) didn't amplify as the chocolate. The samples below 200 RFUs were rejected and cocoa butter showed to be always much lower than the threshold. *2SE: Two Standard Error.

Samples	**Microsatellites (Relative Fluorescence)**		
	cpSSR4	**cpSSR14**	**cpSSR3**
TOW 4	4114.479118	9433.387	8161.067
2SE*	974.5285	1400.355	1845.709
TOW 13	156	205	71
2SE*	NA	NA	NA

Appendix VI: Haplotype allele conversion to FASTA format

CP loc	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR	cpSSR1short	cpIndel-1	cpIndel-6	cpIndel-5	cpInde	cpIndel-3
H1	aaa-	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	t--------	taaaag	tt---	aa---	ttt--	tt---	t-	tttttttt-	TACT	TAGT	ctttactttacttta--	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H2	aaa-	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	t--------	taaaag	tt---	aa---	ttt--	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H3	aaa-	tttt	ttt	aa	ttt-	t-	cccccc-	c-	aaa	t--------	taaaag	tt---	aaa--	ttt--	tt---	t-	tttttttt-	TACT	TAGT	ctttactttacttta--	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H4	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaa--	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H5	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H6	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttactttactttacttta----------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H7	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttactttactttactttactttactttactttacttta	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H8	aaa-	tttt	ttt	aa	tttt	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H9	aaaa	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	tttttt----	taaaag	tt---	a----	t----	tt---	tt	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H10	aaaa	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	tttttt----	taaaag	tt---	aa---	t----	tt---	tt	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H11	aaaa	tttt	tt-	a-	tt--	t-	cccccc-	cc	aaa	t--------	taaaag	tt---	aaa--	tt---	t----	tt	ttttttttt	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGTTTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H12	aaaa	tttt	tt-	a-	tt--	t-	ccccccc	cc	aaa	t--------	taaaagt	t----	aaa--	t----	t----	tt	ttttttttt	TACT	TAGT	ctttacttta---	gtatggaa	TTTCGTTTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H13	aaaa	tttt	tt-	a-	tt--	tt	ccccccc	cc	aa-	tt-------	taaaagt	t----	aaa--	tt---	t----	tt	ttttttttt	TACT	TAGT	cttta--	gtatggaagtatggaa	TTTCGTTTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H14	aaaa	tttt	tt-	a-	tt--	tt	ccccccc	cc	aa-	tt-------	taaaagt	t----	aaaa-	tt---	t----	tt	ttttttttt	TACT	TAGT	ctttactttactttactttactttactttactttactttacttta----------	gtatggaagtatggaa	TTTCGTTTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H15	a---	t---	ttt	aa	tt--	t-	c-------	c-	a--	tt-------	taaaag	ttt--	a----	t----	t----	t-	t--------	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	--------	ggatatggatat
H16	aaa-	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	t--------	taaaag	tt---	a----	ttt--	tt---	t-	tttttttt-	TACT	TAGT	ctttactttacttta--	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H17	aaa-	ttt-	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaa--	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H18	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttactttacttta--------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H19	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaa-	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttactttactttacttta----------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttg----	ggatat------
H20	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaa--	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H21	aaa-	tttt	ttt	aa	ttt-	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaa--	ttttt	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttactttacttta------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H22	aaa-	tttt	ttt	aa	tttt	t-	ccccccc	c-	aaa	ttttttttt	taaaag	tt---	aaaaa	tttt-	tt---	t-	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatat------
H23	aaaa	ttt-	ttt	aa	ttt-	t-	cccccc-	c-	aaa	tttttt----	taaaag	tt---	aa---	t----	tt---	tt	tttttttt-	TACT	TAGT	ctttactttactttacttta-----------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttgtttg	ggatatggatat
H24	aaaa	tttt	t--	aa	t---	t-	ccccccc	c-	aa-	tt-------	taaaag	ttttt	aa---	ttt--	ttttt	tt	ttttttt--	TACT	TAGT	ctttactttactttactt-------------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H25	aaaa	tttt	t--	aa	tt--	t-	ccccccc	c-	aa-	tt-------	taaaag	ttttt	aa---	ttt--	ttttt	tt	ttttttt--	TACT	TAGT	ctttactttactttactt-------------------------------------	gtatggaa	TTTCGT	TGGTTGAATC	tttg----	ggatatggatat
H26	aaaa	tttt	tt-	a-	tt--	t-	cccccc-	cc	aaa	t--------	taaaag	tt---	aaa--	tt---	t----	tt	ttttttttt	TACT	TAGT	ctttactttacttta--	gtatggaa	TTTCGTTTTCGT	TGGTTGAATC	tttg----	ggatatggatat

Figure 0.1 Appendix VI: Haplotype allele conversion to FASTA format

Appendix VII: Haplotype proportion in chocolate

Chocolate Claim replicate	CHOC	R2	cp9-1	cp9-2	cp9-3	cp9-4	cp9-5	cp9-6	cp9-7	cp9-8	cp9-9	cp9-10	cp9-11	cp9-12	cp9-13	cp9-14	cp9-15	∑-SUM
Ecuador1961	1	0.889	0.323	0	0	0	0	0.316	0	0	0.055	0	0	0.064	0.034	0	0	0.209
Ecuador1962	2	0.89	0.286	0	0	0	0	0.3	0	0	0	0	0	0.044	0	0.135	0	0.234
Ecuador1963	3	0.871	0.281	0	0	0	0	0.301	0	0	0	0	0	0	0	0.171	0	0.246
Ecuador1967	7	0.884	0.351	0	0	0	0	0.337	0	0	0	0	0	0.101	0	0	0	0.21
Ecuador1969	9	0.869	0.368	0	0	0	0	0.322	0	0	0	0	0	0.088	0	0	0	0.221
Ecuador19611	11	0.872	0.387	0	0	0	0	0.41	0	0	0	0	0	0	0	0	0	0.204
Ecuador2104	4	0.932	0.298	0	0	0	0	0.116	0	0	0	0.316	0.096	0	0	0	0	0.174
Ecuador2105	5	0.933	0.28	0	0	0	0	0.152	0	0	0	0.268	0.117	0	0	0	0	0.184
Ecuador2106	6	0.922	0.267	0	0	0	0	0.147	0	0	0	0.284	0.116	0	0	0	0	0.186
Ecuador2108	8	0.89	0.308	0	0	0	0	0	0	0	0	0.474	0	0	0	0	0	0.218
Ecuador21010	10	0.914	0.304	0	0	0	0	0.157	0	0	0	0.358	0	0	0	0	0	0.181
Ecuador21012	12	0.904	0.32	0	0	0	0	0	0	0	0	0.48	0	0	0	0	0	0.199
Madagascar19843	43	0.946	0.523	0	0	0	0.092	0.251	0	0	0	0	0	0	0	0	0	0.134
Madagascar19844	44	0.952	0.519	0	0	0	0.092	0.25	0	0	0	0	0	0	0	0	0	0.139
Madagascar19845	45	0.972	0.477	0	0	0	0.101	0.131	0	0	0.031	0	0	0	0	0.158	0	0.101
Madagascar19849	49	0.939	0.588	0	0	0	0	0	0	0	0.253	0	0	0	0	0	0	0.159
Madagascar19851	51	0.939	0.592	0	0	0	0	0	0	0	0.244	0	0	0	0	0	0	0.164
Madagascar19853	53	0.949	0.581	0	0	0	0	0	0	0	0.157	0	0	0.116	0	0	0	0.147
Madagascar21646	46	0.943	0.56	0	0	0	0	0	0	0	0.113	0	0	0	0.089	0.055	0	0.182
Madagascar21647	47	0.982	0.511	0	0	0	0.102	0.113	0	0	0.042	0	0	0	0	0.156	0	0.077
Madagascar21648	48	0.984	0.525	0	0	0	0.096	0.084	0	0	0	0.078	0	0.021	0	0.132	0	0.065
Madagascar21650	50	0.939	0.645	0	0	0	0	0	0	0	0	0	0	0.192	0	0	0	0.163
Madagascar21652	52	0.9	0.756	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.244
Madagascar21654	54	0.907	0.748	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.252
Peru20067	67	0.904	0.212	0	0	0	0	0.376	0	0	0	0	0.226	0	0	0	0	0.187
Peru20068	68	0.9	0.148	0	0	0	0	0.306	0	0	0.061	0	0.167	0	0	0.095	0	0.223
Peru20069	69	0.939	0.105	0	0	0	0	0.332	0.148	0	0	0	0.189	0.03	0	0.154	0	0.043
Peru20073	73	0.898	0.197	0	0	0	0	0.376	0	0	0	0	0.209	0	0	0	0	0.219
Peru20075	75	0.932	0.2	0	0	0	0	0.402	0.124	0	0	0	0.23	0	0	0	0	0.044
Peru20077	77	0.888	0.199	0	0	0	0	0.359	0	0	0	0	0.183	0.04	0	0	0	0.219
Peru20970	70	0.89	0	0	0	0	0	0.389	0.155	0	0	0	0.343	0	0	0	0	0.112
Peru20971	71	0.939	0	0	0	0	0	0.326	0.156	0	0	0	0.205	0.026	0	0.236	0	0.051
Peru20972	72	0.893	0	0	0	0	0	0.336	0.175	0	0	0	0.281	0.1	0	0	0	0.108
Peru20974	74	0.855	0	0	0	0	0	0.442	0	0	0	0	0.274	0	0	0	0	0.284
Peru20976	76	0.86	0	0	0	0	0	0.447	0	0	0	0	0.282	0	0	0	0	0.272
Peru20978	78	0.852	0	0	0	0	0	0.441	0	0	0	0	0.277	0	0	0	0	0.282
SaintLucia19779	79	0.983	0.245	0	0	0	0.162	0.14	0	0	0	0.222	0.181	0	0	0	0	0.049
SaintLucia19780	80	0.963	0.241	0	0	0	0.165	0.133	0	0	0	0.211	0.171	0	0	0	0	0.079
SaintLucia19781	81	0.965	0.255	0	0	0	0.163	0.137	0	0	0	0.206	0.169	0	0	0	0	0.069
SaintLucia19783	83	0.942	0.278	0	0	0	0.143	0.228	0	0	0	0	0.265	0	0	0	0	0.086
SaintLucia19785	85	0.939	0.295	0	0	0	0.122	0.211	0	0	0	0	0.252	0	0.05	0	0	0.07
SaintLucia19787	87	0.937	0.317	0	0	0	0.115	0.234	0	0	0	0	0.268	0	0	0	0	0.065
SaintLucia19982	82	0.978	0.441	0	0	0	0.126	0	0	0	0	0.228	0.131	0	0	0	0	0.075
SaintLucia19984	84	0.949	0.5	0	0	0	0.126	0	0	0	0	0	0.296	0	0	0	0	0.077
SaintLucia19986	86	0.958	0.465	0	0	0	0.098	0	0	0	0	0	0.291	0	0.08	0	0	0.066
SaintLucia19988	88	0.943	0.499	0	0	0	0.093	0	0	0	0	0	0.314	0	0	0	0	0.094
Trinidad19489	89	0.836	0	0	0	0	0	0.331	0.293	0	0	0	0	0	0.24	0	0	0.136
Trinidad19490	90	0.873	0.115	0	0	0	0	0.29	0.274	0	0.058	0	0	0	0.189	0	0	0.073
Trinidad19491	91	0.879	0.126	0	0	0	0	0.274	0.273	0	0	0	0	0.074	0.171	0	0	0.083
Trinidad19495	95	0.912	0.113	0	0	0	0.087	0.327	0.216	0	0	0	0	0	0.189	0	0	0.069
Trinidad19497	97	0.875	0.142	0	0	0	0	0.316	0.263	0	0	0	0	0	0.191	0	0	0.088
Trinidad19499	99	0.842	0	0	0	0	0	0.317	0.269	0	0	0	0	0.091	0.221	0	0	0.102
Trinidad21292	92	0.897	0.151	0	0	0	0	0.316	0.263	0	0	0	0	0	0.18	0	0	0.091
Trinidad21293	93	0.847	0	0	0	0	0	0.265	0.298	0	0.115	0	0	0	0.219	0	0	0.103
Trinidad21294	94	0.813	0	0	0	0	0	0.277	0.328	0	0	0	0	0	0.147	0.11	0	0.138
Trinidad21296	96	0.874	0.143	0	0	0	0	0.335	0.257	0	0	0	0	0	0.176	0	0	0.087
Trinidad21298	98	0.876	0.152	0	0	0	0	0.282	0.264	0	0	0	0	0.07	0.158	0	0	0.073
Trinidad212100	100	0.879	0.149	0	0	0	0	0.325	0.258	0	0	0	0	0	0.182	0	0	0.085
Venezuela193101	101	0.794	0	0	0	0	0.212	0.376	0.286	0	0	0	0	0	0	0	0	0.127
Venezuela193102	102	0.848	0	0	0	0	0.189	0.297	0.324	0	0.121	0	0	0	0	0	0	0.069
Venezuela193103	103	0.907	0.148	0	0	0	0.119	0.271	0.356	0	0	0	0	0.038	0	0	0	0.068
Venezuela193107	107	0.808	0	0	0	0	0.209	0.368	0.298	0	0	0	0	0	0	0	0	0.125
Venezuela193109	109	0.849	0	0	0	0	0.214	0.27	0.311	0	0.101	0	0	0.046	0	0	0	0.058
Venezuela193111	111	0.816	0	0	0	0	0.198	0.311	0.37	0	0	0	0	0.082	0	0	0	0.101
Venezuela213104	104	0.761	0	0	0	0	0.199	0.364	0.29	0	0	0	0	0	0	0	0	0.147
Venezuela213105	105	0.906	0.155	0	0	0	0.109	0.278	0.344	0	0	0	0	0.046	0	0	0	0.066
Venezuela213106	106	0.899	0.152	0	0	0	0.178	0.297	0.291	0	0	0	0	0	0	0	0	0.082
Venezuela213108	108	0.838	0	0	0	0	0.222	0.298	0.282	0	0	0	0	0.089	0	0	0	0.108
Venezuela213110	110	0.81	0	0	0	0	0.22	0.378	0.275	0	0	0	0	0	0	0	0	0.126
Mars20655	55	0.985	0	0	0	0	0	0.214	0	0	0	0.347	0.346	0	0	0	0	0.093
Mars20656	56	0.977	0	0	0	0	0	0.098	0	0	0	0.607	0.196	0	0	0	0	0.099
Mars20657	57	0.946	0	0	0	0	0	0	0	0	0	0.66	0.189	0	0	0	0	0.151
Mars20658	58	0.955	0	0	0	0	0	0	0	0	0	0.65	0.193	0	0	0	0	0.156
Mars20659	59	0.96	0	0	0	0	0	0	0	0	0	0.699	0.179	0	0	0	0	0.122
Mars20660	60	0.915	0	0	0	0	0	0	0	0	0	0.608	0.197	0	0	0	0	0.195
Nestle20761	61	0.801	0	0	0	0	0	0	0	0	0	0.675	0	0	0	0	0	0.325
Nestle20762	62	0.912	0	0	0	0	0	0	0	0	0	0.527	0.266	0	0	0	0	0.207
Nestle20763	63	0.93	0	0	0	0	0	0.151	0	0	0	0.422	0.278	0	0	0	0	0.148
Nestle20764	64	0.905	0	0	0	0.142	0	0.283	0	0	0	0	0.328	0.247	0	0	0	0
Nestle20765	65	0.922	0	0	0	0	0	0	0	0	0	0.589	0.241	0	0	0	0	0.17
Nestle20766	66	0.835	0	0	0	0	0	0	0	0	0	0.52	0.213	0	0	0	0	0.267
IvoSeed25013	13	0.995	0	0	0	0	0	0	0	0	0	0.99	0	0	0	0	0	0.012
IvoSeed25114	14	0.993	0	0	0	0	0	0.253	0	0	0	0.379	0.368	0	0	0	0	0
IvoSeed25215	15	0.842	0	0.263	0	0	0.434	0	0	0	0	0	0.277	0	0	0	0	0.025
IvoSeed25316	16	0.936	0	0	0	0	0.087	0	0	0	0	0.597	0.34	0	0	0	0	-0.025
IvoSeed25417	17	0.989	0	0	0	0	0	0.162	0	0	0.095	0.188	0.459	0.08	0	0	0	0.016
IvoSeed26118	18	0.994	0	0	0	0	0	0.278	0	0	0	0.722	0	0	0	0	0	0
IvoSeed26219	19	0.993	0	0	0	0	0.001	0	0	0	0	1.001	0	0	0	0	0	-0.002
IvoSeed26320	20	0.993	0	0	0	0	0	0.565	0	0	0	0.44	0	0	0	0	0	-0.003
IvoSeed26621	21	0.995	0	0	0	0	0	0	0	0	0.073	0.826	0.096	0	0	0	0	0.004
IvoSeed26722	22	0.955	0	0	0	0	0.664	0	0	0	0	0.314	0	0	0	0	0	0.021
IvoSeed27023	23	0.993	0	0	0	0.001	0	0	0	0	0	1.004	0	0	0	0	0	-0.005
IvoSeed27124	24	0.993	0	0	0	0.002	0.004	0.246	0	0	0	0.387	0.37	0	0	0	0	-0.006
IvoSeed27425	25	0.993	0	0	0	0	0	0.173	0	0	0	0.741	0.096	0	0	0	0	-0.01
IvoSeed27526	26	0.992	0	0	0	0	0	0	0	0	0.075	0.336	0.594	0	0	0	0	-0.005
IvoSeed27627	27	0.989	0	0	0	0	0	0.266	0	0	0	0.718	0	0	0	0	0	0.016
IvoSeedR25528	28	0.992	0	0	0	0	0	0.176	0	0	0	0.699	0.128	0	0	0	0	-0.003
IvoSeedR25629	29	0.958	0	0	0	0	0	0.22	0	0	0.182	0	0.385	0.206	0	0	0	0.006
IvoSeedR25730	30	0.993	0	0	0	0	0.003	0.294	0	0	0	0.711	0	0	0	0	0	-0.006
IvoSeedR25831	31	0.994	0	0	0	0	0	0	0	0	0	0.997	0	0	0	0	0	0.007
IvoSeedR25932	32	0.994	0	0	0	0	0.001	0.56	0	0	0	0.449	0	0	0	0	0	-0.01
IvoSeedR26033	33	0.993	0	0	0	0.002	0.002	0.151	0	0	0	0.43	0.429	0	0	0	0	-0.014
IvoSeedR26434	34	0.994	0	0	0	0	0.002	0.154	0	0	0	0.422	0.427	0	0	0	0	0.002
IvoSeedR26535	35	0.992	0	0	0	0	0	0.247	0	0	0	0.384	0.372	0	0	0	0	-0.003
IvoSeedR26836	36	0.993	0	0	0	0	0	0.165	0	0	0	0.692	0.143	0	0	0	0	0
IvoSeedR26937	37	0.99	0	0	0	0	0	0	0	0	0.071	0.222	0.613	0.096	0	0	0	-0.002
IvoSeedR27238	38	0.7	0	0.331	0	0	0.398	0	0	0	0	0	0	0	0	0	0	0.27
IvoSeedR27339	39	0.932	0	0	0	0	0.099	0	0	0	0	0.574	0.298	0	0.073	0	0	-0.044
IvoSeedR27740	40	0.992	0	0	0	0	0	0	0	0	0	0.995	0	0	0	0	0	0.005
IvoSeedR27841	41	0.993	0	0	0	0	0	0.535	0	0	0	0.466	0	0	0	0	0	-0.001
IvoSeedR27942	42	0.993	0	0	0	0	0	0.139	0	0	0	0.438	0.426	0	0	0	0	-0.003

Figure 0.2 Appendix VII: Haplotype proportion in chocolate

Appendix VIII: Chocolate samples

Table 0.10 Appendix VIII: 1 Chocolate samples used in metagenomics studies with its cacao solids and yield

Samples from Cocoa of Excellence program and the year of production has been anonymised. TOW; Tow Super Food Chocolate, Ecuador. CH; Cocoa Hunters, Colombia, QRN: Chocolates Peru. All samples were biological duplicates apart from the * marked which were biological triplicates and each sample had PCR duplicate.

Code	Country	Fermentation location	Genetic Background	Cacao solid	DNA ng/ul
1	Bolivia	Riberalta	Criollo	70	0.468
2	Bolivia	Riberalta	Criollo	70	0.91
3	Colombia	San Andrés de Tumaco	Híbridos, Tumaco and Clones	70	0.312
CH CAUCA 1	Colombia	San Andrés de Tumaco	Porcelana	100	1.33
9	Ecuador	Quiroga, Manabí	Criollo, Trinitario, Tipo Nacional	70	1.2
10	Ecuador	Ventanas, Los Ríos	Trinitario	70	2.38
11	Ecuador	Vinces, Los Ríos	Nacional	70	0.392
12	Ecuador	Balao, Guayas	Nacional x Trinitario	70	1.83
13	Ecuador	Calceta, Manabí	Nacional	70	1.2
14	Ecuador	Vinces, Los Ríos	Nacional Arriba	70	1.35
15	Ecuador	Quiroga, Manabí	Nacional	70	0.82
TOW 1 2017*	Ecuador	Balao	Arriba Nacional	70	1.64
TOW 2 2017*	Ecuador	Esmeraldas	Arriba Nacional	70	1.4
TOW 3 2017*	Ecuador	Esmeraldas	Arriba Nacional	70	1.18
TOW 4 2017*	Ecuador	Guayas	Arriba Nacional	100	3.18
TOW 6 2017*	Ecuador	Balao	Nacional hybrids	70	1.69
TOW 7 2017*	Ecuador	Guayas	Arriba Nacional	100	0.738
TOW 8 2017	Ecuador	El Carmen, Manabí	Nacional, Amazon, Criollo, Trinitario	86	1.95
TOWCB 13 2017	Ecuador	Guayas	Arriba Nacional	100	0.0216
27	Perú	Castilla, Piura	Criollo	70	0.6

Table 0.11 Appendix X: Continuation 2- Chocolate samples used in metagenomics studies with its cacao solids and yield

Code	Country	Fermentation location	Genetic Background	Cacao solid	DNA ng/ul
28	Perú	Castilla, Piura	Criollo	70	0.31
29	Perú	Cuzco	Chuncho	70	0.228
QRN CSC 1 2018	Perú	Cuzco	Chuncho	72	1.74
33	Haití	Port-Margot	Forastero	70	0.784
34	Haití	Ouanaminthe,	Criollo	70	0.396
35	Trinidad and Tobago	Gran Couva	Trinitario	70	2.8
36	Trinidad and Tobago	Gran Couva,	Trinitario	70	0.162
38	Honduras	La Másica	Híbridos Trinitarios	70	0.848
39	Honduras	La Masica, Atlantida,	Trinitario	70	0.648
42	Guatemala	Rio Dulce	Trinitario	70	0.53
43	Guatemala	Río Dulce, Izabal	Trinitario	70	0.192
44	Ghana	Ahafo Ano South	Forastero	70	1.02
47	Ghana	Ahafo Ano South	Forastero Hybrids	70	0.924
48	Ivory Coast	Duékoué	Forastero	70	0.424
50	Ivory Coast	Duékoué	Forastero	70	0.422
52	Vietnam	Ha Long	Trinitario	70	0.422
53	Vietnam	Tan Phu	Trinitario	70	0.448
54	Indonesia	Kecamatan Guguk	Trinitario	70	0.252
55	Indonesia	Jembrana, Bali	Trinitario	70	0.284
56	Malaysia	Ranau	Trinitario	70	0.922
57	Malaysia	Ranau	Trinitario	70	0.416
58	Australia	Mossman, Queensland	Trinitario	70	0.166
60	Australia	Mission Beach,Queensland	Trinitario and Forastero	70	0.226

Table 0.12 Appendix X: Continuation 3- Chocolate samples used in metagenomics studies with its cacao solids and yield

Code	Country	Fermentation location	Genetic Background	Cacao solid	DNA ng/ul
65	Colombia	Santander	Trinitario	70	0.434
67	Colombia	Santander	Trinitario and Regionales Nacional from Carmen de Chucuri	70	0.306