2017 International Symposium on Cocoa Research (ISCR), Lima, Peru, 13-17 November 2017

The use of chloroplast markers for the traceability of certified sustainably produced cacao (*Theobroma cacao*) in the chocolate industry

P. Lafargue Molina[1], A. Wetten[1], J.M. Allainguillaume[2], A.J. Daymond[2] and J. Allainguillaume[1]

[1]Department of Applied Sciences, University of the West of England, Coldharbour Lane, Bristol, UK, BS16 1QY

[2]School of Agriculture, Policy and Development, the University of Reading, Whiteknights, Reading, UK, RG6 6AR

Abstract

Recently the chocolate industry has changed to a higher demand for sustainably certified cacao (Rainforest Alliance, UTZ, and FairTrade) and closer attention is being paid to how this sustainably produced cocoa can be traced. Companies like Mars, Hershey and Ferrero have stated that by 2020, all their cocoa will be certified sustainable. There is therefore a need for methodologies to be developed enabling the characterisation and geographical tracking of certified cocoa products. Research on chloroplast ultra-barcoding in cacao has revealed a level of DNA polymorphism sufficient to reliably identify lineages below the species level such as subspecies or varieties (Kane *et al.* 2012). This level of variation, in conjunction with the high copy number of the chloroplast genome, offers possibilities to develop reliable DNA assays which being, less susceptible to industrial DNA degradation than single locus nuclear markers, are suitable for the characterisation of sustainably produced chocolate products.

DNA was extracted from 159 representative trees of major cacao cultivars present in the International Cocoa Quarantine Centre in Reading (UK). All accessions were screened with four chloroplast simple sequence repeat (cpSSR) markers also known as microsatellites, to assess the chloroplast haplotype diversity of *Theobroma cacao*. These loci were designed from polymorphic sites identified by Kane *et al.* (2012) to allow for multiplex PCR amplification. Fluorescently labelled products were screened using capillary analysis and all markers scored using GeneMarker. Eleven cpSSR alleles were identified across the four loci revealing six unique cacao chloroplast haplotypes. All markers were screened on DNA extracted from a range of commercially available chocolate products. The capillary profiles generated were normalised to determine the proportion of each specific cpSSR alleles per locus identified in each chocolate sample. Principal Component Analysis (PCA) of all samples for the proportion of all alleles gave contrasting results with distinct clustering observed for chocolate produced from beans harvested by small cooperatives in Peru, Ecuador, Venezuela, Trinidad and Madagascar but no differentiation was observed for chocolate derived from West African plantations reflecting the lack of allelic diversity found in cultivars in West Africa. These results indicate that this sensitive and relatively low cost barcoding approach has potential to support cocoa certification programmes for the Fine cocoa/premium cocoa market but is not likely to be appropriate for the identification of bulk cocoa production.

Introduction

Tracking the origin and quality of cacao products

Theobroma cacao (cocoa) is one of the most studied commodities around the world and the source of one of the world’s most delicious and familiar products, chocolate. Native to Central America and Northern South America, it has been introduced as a crop to the equatorial regions of West Africa and South East Asia.

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. Cultivars for bulk cocoa production often lack the range of flavour observed in premium cocoa in favour of higher yield capability. For instance, CCN-51 a hybrid developed in Ecuador in the 1960’s, is a remarkable variety for production standards giving yields up to

70% higher than Fino and Aroma premium cacao accessions but its organoleptic characteristics are lower and lead to less specific flavour development following fermentation.

Only 2% of the cocoa produced in the world is of premium quality with the remaining 98% aimed at bulk usage. There is however increased interest in this premium market and a much higher demand for certified cacao. As a result, instances of fraud in the quality or product origin are likely to rise. For example in Ecuador, mixing bulk cacao (CCN-51) or Forastero with a high-quality variety is already a common practice to increase the price and quality for medium quality chocolates (Herrmann *et al.*, 2015). There is therefore a need for methodologies to be developed enabling the characterisation and geographical tracking of certified cocoa products and DNA markers can be utilised for such purpose.

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 as, for example, described by Singh *et al.* (2013) who compared these two types of marker to characterised Indian rice varieties. The same marker can be also found on the chloroplast genome and have been used for many investigations in plant tracking (Schroeder *et al.*, 2016).

Markers associated with Theobroma cacao nuclear genome

The *Theobroma cacao* nuclear genome has been extensively studied as 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 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 of the Ecuadorean variety CCN-51. Using 10 published SSR markers generated by CIRAD (mTcCIR) they compared CCN51 to the varieties Arriba, Criollo, “Nacional”, “Fino and Aroma”, and demonstrated clear allelic differentiation between these. While single locus nuclear markers are highly informative they have the disadvantage of being more susceptible to degradation due to low copy number.

Chloroplast Genome and ribosomal regions

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.* in 2012. They 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 *T. grandiflorum*. This ultra-barcoding (UBC) approach demonstrated that all individuals examined were uniquely distinguishable. A later study (Hermann *et al.*, 2014) focusing purely on the chloroplast genome, identified distinct chloroplast markers differentiating the two cocoa types, Arriba (fine cocoa type) and CCN-51 (bulk cocoa) being cultivated in Ecuador. 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. In turn, 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.

In the present study using chloroplast microsatellite specific to *T. cacao*, we assessed if this allelic frequency can be used as an indicator of origin by examining a range of chocolate product of known origin.

Methods

Using the 12 chloroplast genome sequences generated by Kane *et al.* (2012) for *T. cacao*, we selected three polymorphic mononucleotide and one polymorphic hexanucleotide cpSSR. Allelic diversity for each locus was assessed by amplifying the targeted region via polymerase chain reaction (PCR). Amplified products were then separated using fluorescent capillary electrophoresis. Primer pairs for the four selected loci were designed using Primer3 software (Rozen and Skaletsky, 2000) so that the four expected product size ranges did not overlap and would be suitable for multiplex analysis (Table 1).

Total genomic DNA was extracted from fresh leaf samples collected from 159 accessions 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 to assess the level allelic diversity that could be expected at each cpSSR locus. DNA was also extracted using a DNeasy mericon Food Kit (Qiagen) from duplicate chocolate samples purchased from food shops in the UK (Hotel Chocolat© single origin: Peru Pichanaki, Coastal Ecuador Hacienda Iara, Venezuela Chuao, Trinidad Cocoa association and Madagascar Somia Plantation; Mars© Mars bar, Nestlé© Kit Kat and Cadbury© Crisp bar) and pools of dry beans obtained from six farm location in Côte d' Ivoire.

PCR amplifications were performed as a multiplex with all cpSSR loci 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 specific primers for the four loci and the universal fluorescent primer (0.05 µ M forward primer, 0.5 µ M reverse primer, and 0.5 µ M Hex-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. Each chocolate sample was replicated four times and all beans were sampled twice.

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. Allelic diversity and chloroplast haplotypes were identified from the reference samples using GenAlEx 6.1 software (Peakall and Smouse, 2006). 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). All positions were recorded even if these might correspond to single base stuttering artefact inherent to cpSSR mononucleotides PCR amplification. 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. The allelic proportions generated for each sample were then analysed by Principal Component Analysis (PCA) using Minitab17 to identify clusters sharing common allelic frequencies.

Results and Discussion

Plants and haplotype and allelic diversity

The four cpSSRs were assessed on 159 diverse accessions from the International Cocoa Quarantine Centre at the University of Reading (ICQC, R). The aim here was to assess the likely allelic diversity observed for each locus in plantations of cacao around the world. All four loci were polymorphic, ranging from two to four alleles with an average of 2.5 alleles per locus (Table 2). CpSSR 3 and 4 exhibited the same level of allelic diversity as observed by Yang *et al.* (2011). The unbiased haploid diversity per locus calculated using GenAlEx 6.501 software (Peakall and Smouse, 2006) was from 0.172 to 0.551. We detected six haplotypes out of the 159 samples using all four cpSSR (Table 2). The six haplotypes varied in frequency ranging from 1.8% (3 out of 159) to 63.5% (101 out of 159). This preliminary analysis established the allelic range to be screened for allelic proportion within chocolate samples.

Frequency of alleles in chocolate samples

All plant accessions were included in the assessment for allelic frequency since, as reference samples with unique haplotypes, they would be expected to exhibit unique allelic profiles. As predicted, PCA revealed distinct and unique clusters corresponding to each of the six haplotypes identified and these have been highlighted in various shades of grey on Figure1. PCA gave contrasting results when assessing the allelic composition of chocolate samples and beans from Côte d' Ivoire. The analysis of DNA extracted from beans from Côte d' Ivoire and chocolate samples from bars made by Mars©, Nestlé© and Cadbury© clustered together and closely to the haplotype 6, exhibiting no evidence of allelic variation at the cpSSR analysed. This makes sense in terms of genetic diversity since the majority of the crop grown in Ghana and Côte d' Ivoire originates from a limited genepool mainly derived from the variety Amelonado and a number of upper Amazon types, and the accessions from this variety screened in the reference panel all exhibited Haplotype 6. With 70% of the current production of bulk cocoa originating from West Africa, it is also logical that DNA extracted from bars made from bulk chocolate would also exhibit a profile close to haplotype 6 and therefore group in that cluster. In contrast to this, all replicate samples from Hotel Chocolat© grouped according to their geographical origin in position intermediate to all reference single haplotype clusters. These samples all originated from specific geographical locations with beans gathered by small cooperatives. The differential clustering observed here reflect the chloroplast genetic diversity present in the chocolate samples which is absent from West Africa. 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. Replicate batches of chocolate for each origin did cluster together which supports the idea that cpSSR genetic diversity in chocolate samples mirrors the genetic diversity of the beans used to make this chocolate. This might in turn reflect the genetic diversity of the production area the beans have been collected from. These results indicate that this sensitive and relatively low cost barcoding approach has potential to support cocoa certification programmes for the Fine cocoa/premium cocoa market but is not likely to be appropriate for the characterisation of bulk cocoa production.

Acknowledgements

This project is funded by UWE QR. We would like to thank Dr Martin Gilmour from Mars UK for providing us with cacao beans from Côte d' Ivoire farms and Dr Kirsten Skøt from Aberystwyth University for the analysis of the chloroplast microsatellites by fluorescent capillary screening.

References

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., Umaharan, P., Brunel, D. and Lanaud, C. (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.

Argout, X., Fouet, O., Wincker, P., Gramacho, K.P., Legavre, T., Sabau, X., Risterucci, A.M., Da Silva, C., Cascardo, J., Allegre, M., Kuhn, D., Verica, J., Courtois, B., Loor, R., Babin, R., Sounigo, O., Ducamp, M., Guiltinan, M.J., Ruiz, M., Alemanno, L., Machado, R., Phillips, W., Schnell, R., Gilmour, M., Rosenquist, E., Butler, D., Maximova, S. and Lanaud, C. (2008) Towards the understanding of the cocoa transcriptome: Production and analysis of an exhaustive dataset of ESTs of *Theobroma cacao L.* generated from various tissues and under various conditions. *BMC Genomics* 9 (1), pp. 1–19.

Herrmann, L., Haase, I., Blauhut, M., Barz, N. and Fischer M. (2014). DNA-Based Differentiation of the Ecuadorian Cocoa Types CCN-51 and Arriba Based on Sequence Differences in the Chloroplast Genome. *Journal of Agricultural and Food Chemistry,* 2014, 62, 12118−12127.

Herrmann, L., Felbinger, C., Haase, I., Rudolph, B., Biermann, B. and Fischer, M. (2015) Food fingerprinting: Characterization of the Ecuadorian type CCN-51 of *Theobroma cacao L.* Using microsatellite markers. *Journal of Agricultural and Food Chemistry,* 63 (18), pp. 4539–4544.

Kane, N., Sveinsson, S., Dempewolf, H., Yang, J., Zhang, D., Engels, J., and Cronk, Q. (2012). Ultra-barcoding in cacao (*Theobroma* spp.; Malvaceae) using whole chloroplast genomes and nuclear ribosomal DNA. *American Journal of Botany,* 99 (2):320–329.

Peakall, R., and Smouse, P.E. (2006). GenAlEx 6: Genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 6, 288–295.

Rozen, S., and Skaletsky, H. (20000. Primer3 on the WWW for general users and for biologist programmers. In S. Krawetz and S. Misener [eds.], Bioinformatics methods and protocols: Methods in molecular biology, 365–386. Humana Press, Totowa, New Jersey, USA.

Singh, N., Choudhury, D.R., Singh, A.K., Kumar, S., Srinivasan, K., Tyagi, R.K., Singh, N.K. and Singh R. (2013) Comparison of SSR and SNP markers in estimation of genetic diversity and population structure of Indian rice varieties. *PLOS ONE*, 8(12): e84136.

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):e0158221.

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):e372–e374.

Locus name	Yang *et al.* (2011)	Amplicon size (bp)	Position on JQ228389	cpSSR motif
cpSSR1	Yes	209	17260	taaaag
cpSSR2	Yes	149	92	t
cpSSR3	No	170	14676	c
cpSSR4	No	275	129240	t

Table 1. Description of four polymorphic chloroplast microsatellite loci identified from *Theobroma cacao*. Information for these loci include their name, their previous use by Yang *et al.* (2011), the expected product size in base pairs (bp), the position of the loci on the full *Theobroma cacao* chloroplast genome GenBank accession JQ228389 and the repeat motif.

cpSSR locus		cpSSR1	cpSSR2	cpSSR3	cpSSR4
Na		2	2	2	4
h		0.172	0.202	0.397	0.551
Haplotypes	**Ha %**	**Amplicon size (bp)**			
H1	1.9	209	149	171	276
H2	4.4	215	149	172	275
H3	5.1	215	149	172	276
H4	11.9	209	150	171	275
H5	13.2	209	150	171	277
H6	63.5	209	150	172	278

Table 2. Initial screening of the four cpSSRs on 159 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).

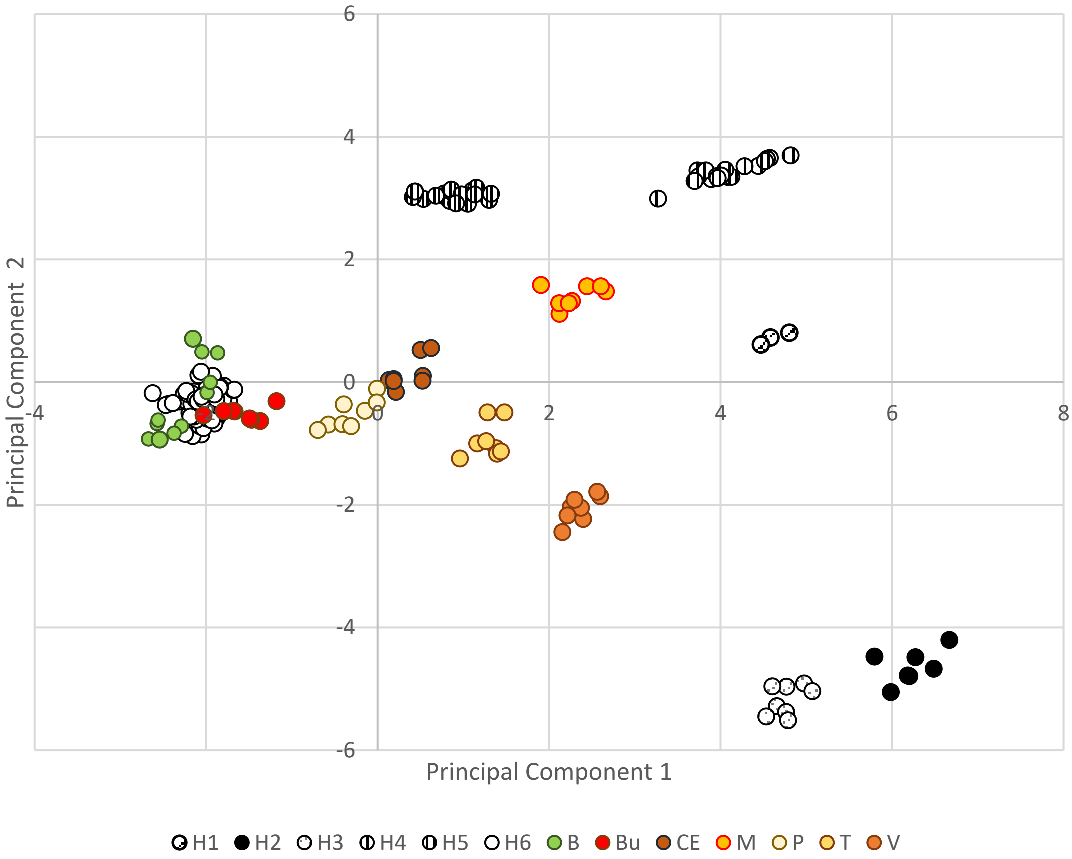

Figure 1. Principal Component Analysis (PCA) of the allelic distribution of four cpSSR markers measured as peak fluorescence proportion per cpSSR locus. PCA1 is compared to PCA2. Reference plant individuals cluster according to haplotypes (H1 to H6) and are represented by circles in shade of grey. All other samples are highlighted as follow: Beans from Côte d' Ivoire (B: green circles); Mars© Mars bar, Nestle© Kit Kat and Cadbury© Crisp bar (BU: red circles) Hotel Chocolat© samples: Peru Pichanaki (P: cream circle), Coastal Ecuador Hacienda Iara (CE: brown circles), Venezuela Chuao (V: Dark orange circles), Trinidad Cocoa association (T: yellow circles) and Madagascar Somia Plantation (M: orange circles).