

# Optimization of a SNP assay for genotyping *Theobroma cacao* under field conditions

Donald S. Livingstone III · Barbie Freeman · Juan Carlos Motamayor ·  
Raymond J. Schnell · Stefan Royaert · Jemmy Takrama ·  
Alan W. Meerow · David N. Kuhn

Received: 25 August 2010 / Accepted: 2 June 2011  
© Springer Science+Business Media B.V. (outside the U.S.A) 2011

**Abstract** The tropical tree crop *Theobroma cacao* L. is grown commercially for its beans, which are used in the production of cocoa butter and chocolate. Although the upper Amazon region is the center of origin for cacao, 70% of the world's supply of cacao beans currently comes from small farms in West Africa. While cacao breeding programs in producer nations are the source of improved planting material, modern marker-based breeding is difficult to perform due to the lack of genotyping facilities in these countries. While DNA extraction can be routinely performed, the equipment needed to analyze simple sequence repeats (SSRs) is seldom available, forcing the outsourcing of genotyping to foreign laboratories and delaying the breeding process. We describe a 5' nuclease (TaqMan)-based single nucleotide

polymorphism (SNP) assay for genotyping cacao plants under conditions similar to those found in most cacao-producing areas. The assay was tested under field conditions by planting open pollinated seeds of seven pods from four different maternal plants. The resulting 171 seedlings were successfully genotyped with 18 SNP markers representing 12 loci. The ability to use temperature-stable reagents and rapid DNA extraction methods is also explored. Additionally, by examining the seedling genotypes for the SNP markers and 14 additional SSR markers, we investigated whether seeds in a pod are the result of single or multiple pollination events. This simple, effective method of genotyping cacao seedlings in the field should allow for more efficient resource management of seed gardens and is currently being implemented in Ghana.

---

**Electronic supplementary material** The online version of this article (doi:[10.1007/s11032-011-9596-4](https://doi.org/10.1007/s11032-011-9596-4)) contains supplementary material, which is available to authorized users.

---

D. S. Livingstone III · B. Freeman ·  
R. J. Schnell · S. Royaert · A. W. Meerow ·  
D. N. Kuhn (✉)  
Subtropical Horticulture Research Station, USDA-ARS,  
13601 Old Cutler Rd., Miami, FL 33158, USA  
e-mail: David.Kuhn@ars.usda.gov

J. C. Motamayor  
Mars, Inc., Hackettstown, NJ 07840, USA

J. Takrama  
Cocoa Research Institute of Ghana (CRIG), Tafo, Ghana

**Keywords** SNP · Cacao · Genotype · SSR ·  
Parentage

## Introduction

*Theobroma cacao* L. (cacao), a member of the Malvaceae family, is a primarily cross-pollinating perennial tree crop. With its center of origin in the South American Amazon region, cacao is now grown in the tropical regions from 10 to 20 degrees latitude north or south of the equator (Motamayor et al. 2002). Cultivation occurs primarily for the cacao

beans which are used for the production of chocolate as well as cocoa butter for the cosmetic industry. After devastating outbreaks of disease in the major growing areas of Central and South America, West Africa has emerged as the largest cacao-producing region (Borrone et al. 2004; Brown et al. 2005; Duguma et al. 2001). Worldwide production of cacao is estimated at over 3.5 million tons (ICCO 2010) with almost 70% coming from West Africa.

Cacao has become one of the most important cash crops in western Africa. Small farmers grow the majority of Africa's cacao crop with over 80% of the cacao produced in this region coming from farms smaller than 10 ha (Duguma et al. 2001; Rice and Greenberg 2000). Local cutbacks in government-subsidized pesticides and fungicides have forced many farmers to reduce expenses, especially for fertilizers and fungicides, and to rely on alternative cropping practices and natural resistance to control losses from pests and diseases (Duguma et al. 2001). Despite improvements provided by better agronomic practices, global yield losses from diseases are estimated at about 30% (Duguma et al. 2001). This underscores the importance of breeding programs for identification and incorporation of disease resistance traits.

Cacao breeding programs focus on establishing improved cacao varieties through the selective crossing of trees with desired horticultural and resistance traits. The results of these crosses are evaluated and eventually distributed to farmers as seedlings or, more often in West Africa, as seed produced in clonal seed gardens. However, as a perennial tree crop, *T. cacao* presents many challenges not present in annual commodity crop breeding programs such as increased land usage and long juvenile periods (Schnell et al. 2007). For example, a major drain on breeding program resources are the costs of maintaining trees that do not contain the desired traits to reproductive age (3–4 years). The use of molecular markers in a marker-assisted selection (MAS) program promises to identify desired traits earlier in the lifecycle of a tree, thereby eliminating undesirable trees while still in the nursery and focusing phenotypic data gathering on more promising candidates.

MAS uses molecular markers to identify and follow the transfer of traits at the genetic level. This allows for seedlings in a cross to be screened for the desired traits within a few weeks after germination,

significantly earlier than many traits manifest phenotypically. Substantial success has been achieved in cacao marker development, with the genetic mapping of over 400 markers and the identification of 83 horticulture traits and 22 disease resistance quantitative trait loci (QTL) (Brown et al. 2005; Faleiro et al. 2006; Lanaud et al. 2009; Pugh et al. 2004; Schnell et al. in press). Unfortunately, MAS programs have drawbacks that need to be addressed, especially as they apply to cacao breeding efforts.

Perhaps one of the greatest barriers for a MAS breeding program is the development of useful molecular markers, which requires a large investment of time and resources. The commercial importance of cacao has helped to direct research interests in marker development and has resulted in the availability of hundreds of simple sequence repeat (SSR) markers and dozens of single nucleotide polymorphism (SNP) markers for cacao (Borrone et al. 2004, 2007; Kuhn et al. 2003; Lanaud et al. 1999; Lima et al. 2009; Livingstone et al. 2011; Zhang et al. 2009). The large number of identified markers and their highly polymorphic nature has resulted in the use of SSR markers as the de facto marker for cacao breeding (Irish et al. 2010). Unfortunately, while cacao breeders often are able to perform the initial sample preparation steps for marker analysis, such as DNA extraction and PCR, equipment needed to perform high-throughput SSR analysis is often too costly. Additionally, the analysis of SSR data requires highly trained personnel who are commonly not available at small breeding centers. As a result, many cacao breeders outsource their marker analysis to laboratories in other countries.

SNP markers may be better suited for MAS at cacao breeding stations around the world. SNP analysis using TaqMan-based probes requires significantly less upfront equipment costs, approximately 10 times less (USD 35,000 vs. USD 350,000) than equipment needed to perform high-throughput SSR analysis, and can be performed by minimally trained personnel without sacrificing high-throughput capabilities (Livingstone et al. 2011). Unfortunately, when compared to SSR markers, relatively few SNP markers are currently available for cacao, and the biallelic nature of SNPs means that more markers need to be analyzed to provide as much information as the more polymorphic SSR markers (Kuhn et al. 2008; Lima et al. 2009; Livingstone et al. 2011;

Rafalski 2002). However, the cacao genome sequencing project ([www.cacaogenomedb.org](http://www.cacaogenomedb.org)) and other ongoing SNP discovery projects utilizing new sequencing data and existing expressed sequence tag resources will provide an ample supply of SNP markers to suit the needs of breeders and researchers alike.

In this manuscript, we describe a 5' nuclease SNP assay (TaqMan) that is suitable for use under field conditions similar to those found at breeding sites in cacao producer regions. We explore modifications to previously published 5' nuclease SNP assays (Livingstone et al. 2011) such as the use of alternative equipment, reagents and DNA extraction techniques to better suit the needs of cacao growing regions. Parental determination analysis is an important control of the output quality of seed gardens. Therefore, an examination of parental determination using SNPs and SSRs is also detailed, with attention given to single or multiple pollination events occurring within collected cacao pods.

## Materials and methods

### Plant material

Mature open pollinated cacao pods were collected from four different maternal cacao trees from the germplasm collection at the USDA-ARS Subtropical Horticulture Research Station (SHRS) in Miami, Florida. The maternal trees consisted of two genetic clones of Gainesville II 164 (Parent ID M1 and M2, Table 1), and two clones of EET59 (Parent ID M3 and M4, Table 1) with two pods collected from M1, three pods from M2, and one pod each from M3 and M4. Seeds from each pod were germinated in Jiffy horticultural peat pellets (Hummert, Earth City, MO, USA), later transplanted into larger containers (4.16 L) and grown in a greenhouse. Each planted seedling was given a unique identifying number of the style GCSXXX. In all, 171 seedlings were planted. Potential fathers that radiated outward from the maternal trees were chosen for sampling in our parental pool (Table 1).

Leaf samples from all the seedlings and the four maternal trees, as well as trees in our paternal pool present in the field collection at the SHRS, were harvested, labeled and used for DNA extractions.

DNA was extracted from 300 trees representing the four maternal plants, 171 progeny, and 125 possible paternal trees. Additionally, DNA from cultivar Matina 1-6 was used in the production of Genome-Walker libraries (Clontech, Mountain View, CA, USA), and DNA from the cultivars in the diversity panel (Table 2) was extracted for SNP discovery.

### DNA extraction, quantification, and normalization

DNA was isolated from leaf discs following a modified protocol of the Fast DNA Kit (Bio101, Carlsbad, CA, USA) as described in Kuhn et al. (2003) and hereafter referred to as FastPrep. All isolated DNA was first quantified using a SYBR Green I based assay (Livingstone et al. 2009). In brief, 24 µl of water, 1 µl of DNA sample, and 25 µl of 60× SYBR Green I (diluted in 10 mM Tris pH 7.5, 1 mM EDTA buffer) were combined in a 96-well plate. The reaction plate was then incubated at room temperature in the dark for 5 min before fluorescence was measured in a plate reader. Fluorescence was converted to concentration by comparison to a standard curve. After quantifying, all samples were normalized to 4 ng/µl.

### QuickExtract DNA isolation

Rapid DNA extractions suitable for PCR amplification were performed using the QuickExtract™ Plant DNA Extraction Solution (Epicentre Biotechnologies, Madison, WI, USA). The standard protocol was modified slightly from the protocol provided. A 5-mm diameter leaf punch was collected from each sampled plant and placed into a 96-well plate. Following the addition of 100 µl of QuickExtract Plant DNA Extraction Solution to each leaf disk, the samples were placed in a thermocycler for 16 min at 65°C, then 4 min at 98°C, and thereafter immediately placed on ice.

### SNP marker discovery

For this study, 18 SNP markers representing 12 different loci were used: w17s189, w3s41, w3s463, w3s558, e0050s274, w7s104, e4785s737, w8s119, w8s131, w8s288, c4s123, c4s536, c5s639, Cir37s112, Cir160s384, Cir211s1036, Cir222s296, and Cir222s384. SNP markers were created by one of three different

**Table 1** List of trees used in parental analysis

Parental pool					
Variety name	Parent ID	Variety name	Parent ID	Variety name	Parent ID
106R	F1	EQXZ	F31	LCTEEN84_S10	F61
75R	F2	GAINESVILLEII032	F32	LCTEEN86	F62
AMAZ1515	F3	GAINESVILLEII058	F33	LCTEEN86_02	F63
AMELONADO	F4	GAINESVILLEII164	M1,M2	LCTEEN86_13	F64
Bz000205_SCA6xICS6	F5	GAINESVILLEII164_F	M1,M2	MO20	F65
BZ000208_SCA6xICS6	F6	GAINESVILLEII330	F36	P30B	F66
Bz000214_SCA6xICS6	F7	GAINESVILLEII360	F37	PA120	F67
Bz000219_SCA6xICS6	F8	GS46	F38	PA120_a	F68
Bz000219_SCA6xICS6_04	F9	GU221H	F39	PA120_b	F69
Bz000220_SCA6xICS6	F10	GU296H	F40	PA150	F70
Bz000221_SCA6xICS6	F11	GU296H_06	F41	PA300	F71
Bz000226_SCA6xICS6	F12	ICS1	F42	PA7	F72
Bz000229_SCA6xICS6	F13	ICS43	F43	PA70_a	F73
Bz000234_SCA6xICS6	F14	IMC105	F44	POUND12	F74
Bz000239_SCA6xICS6	F15	IMC11	F45	POUND7	F75
Bz000241_SCA6xICS6	F16	IMC45	F46	SANMIGUEL3161	F76
Bz000244_SCA6xICS6	F17	IMC47	F47	SCA6	F77
Bz000250_SCA6xICS6	F18	IMC51	F48	SCA6_01	F78
Bz000251_SCA6xICS6	F19	LCTEEN127	F49	SPA16_11	F79
Bz000257_SCA6xICS6	F20	LCTEEN162_S1010	F50	SPEC138_15	F80
Bz000265_SCA6xICS6	F21	LCTEEN300_S201	F51	U12	F81
Bz000268_SCA6xICS6	F22	LCTEEN302	F52	U12_16	F82
Bz000283_SCA6xICS6	F23	LCTEEN31	F53	U15	F83
Bz000284_SCA6xICS6	F24	LCTEEN338_S201	F54	U26	F84
COCA3308_1_15	F25	LCTEEN37_02	F55	U48_01	F85
COCA3308_1_16	F26	LCTEEN37_03	F56	U48_a	F86
EET400	F27	LCTEEN46	F57	UF676	F87
EET400_05	F28	LCTEEN73_01	F58	UNKNOWNCACAO	F88
EET59	M3,M4	LCTEEN73_03	F59		
EET59_F	M3,M4	LCTEEN80	F60		

discovery schemes. All three schemes utilized a diversity panel consisting of DNA samples from a total of 23 different cacao cultivars, representing most of the diversity found within cacao including all ten STRUCTURE groups (population clusters) as identified by Motamayor et al. (2008) as well as other cultivars representing populations of local interest. SNPs were identified by comparing the sequences of the amplicons from each locus for all members of the diversity panel and a minor allele frequency was estimated by determining the percentage of the least common allele with respect to the total number of

alleles observed (Table 2). The consensus sequence for each locus was used to design TaqMan-MGB probes and primers using Primer Express 3.0 software (Applied Biosystems, Carlsbad, CA, USA).

The first scheme identified SNPs by using previously developed single strand conformation polymorphism (SSCP) markers and converting them into SNP markers by sequencing these regions across the diversity panel as detailed in Livingstone et al. (2011). This method was successfully used to create w17s189, w3s41, w3s463, w3s558, w7s104, w8s119, w8s131, and w8s288. The second scheme for

**Table 2** Diversity panel genotypes at SNP loci and minor allele frequencies

Genotype	STRUCTURE group <sup>a</sup>		Wrky3		Wrky7		Wrky8		Wrky17		Est4785		c5		CIR160		CIR37		CIR211		CIR222	
	41	463	558	104	119	131	288	189	274	737	123	536	639	384	112	1036	296	316				
TSH516	G/T	C/G	A/G	A/T	C	C/T	A/G	C/T	C/T	C/T	A/G	A/G	A/G	A	G	A/T	T	G/T				
PA 41	G	G	G	A	C	C	G	T	C	C/T	A	G	A	A	C	T	T	G/T				
IMC 47	G/T	C	A/G	A	C	C	G	C/T	C	C	A	G	A	A	G	T	T	T				
NA 194	G	G	G	A	C	C	G	T	C	C	A	G	A	NA	NA	NA	NA	NA				
COC 3335	G/T	C	A	A/T	C	C	G	C	C	C	A	A/G	A	A	G	T	T	T				
GU 124A	G	G	G	A	C	C	G	C	C	C	A	G	A	G	G	T	T	G				
Las Brisas 17 17	G	C	A	A	C	C	G	C	C	C	A	A	A	A	G	T	T	T				
RB 40 PL 1	T	C	A	A	C	C	C	C	C	C	A	G	A	NA	NA	NA	NA	NA				
BE 4 PL 3	T	G	G	A	T	C	G	C/T	C	T	A	G	A	A	G	A	T	G				
CAB 0331 PL 4	G	C	A	A	T	C	G	C/T	C	C	A	G	A	A	G	T	T	T				
CAB 0339 PL 1	G	C	A	A	C	C	G	C	C	C	A	G	NA	NA	NA	NA	NA	NA				
SCA 6	T	C	A	A	C	T	A	C	C	C/T	A	G	G	A	G	T	T	T				
Criollo 13	T	C	A	T	C	C	G	C	T	T	G	A	A	A	G	T	T	T				
UF273 type2	G	C/G	A/G					C/T	C	C/T	A	G		G	A/T	T	G/T					
UF273 type1	G	C/G	A/G					C	C	C		A/G	A	G	A/T	T	G/T					
Pound# 7	G	C/G	A/G	A				C/T	C	C/T	A	G	A	A	C/G	A/T	T	T				
KA2-101		C/G	A/G								A	G	NA	NA	NA	NA	NA	NA				
K82		C/G									A	G	A	NA	NA	NA	NA	NA				
SCA 12		C	A					C	C	C/T	A	G	A/G	A	G	T	T	T				
SHA 1 (PERU)	G/T	C	A					C	C	C/T	A	G	A	G	T	T	T	T				
ICS1	G/T	C/G	A/G					T	C/T	T	G	A	A	A	G	A	T	G/T				
Pa7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G	T	C/T	G/T					
P30	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	A	G	A	T	G				
Minor allele frequency (%)	36	36	38	15	15	12	13	32	11	34	13	17	14	11	8	28	3	33				
Minor allele	T	G	G	T	T	T	A	T	T	T	G	A	G	G	C	A	C	G				

NA individual was not examined at this locus

<sup>a</sup> STRUCTURE group determined by Motamayor et al. (2008)

identifying SNPs utilized the sequence of genes of interest to design primers that amplify across introns within those genes. Introns were amplified as they generally contain more SNPs than exons (Kuhn et al. 2010). These primers were then used to sequence across the members of the diversity panel and SNPs identified by sequence comparison using the phred, phrap, and polyphred software pipeline (Ewing and Green 1998; Ewing et al. 1998; Gordon et al. 1998). Using this method SNP markers e4785s737, e0050s274, c4s123, c4s536, and c5s639 were generated. The third SNP discovery scheme was to convert previously established SSR markers into SNP markers. Often, the sequence flanking the repeat was too small and the repeat itself too repetitive to find useful SNPs. Instead of directly sequencing the SSR region across the diversity panel, GenomeWalker Universal Kit (Clontech) libraries were created using DNA from cacao variety Matina 1-6. The Matina 1-6 DNA was used in four separate restriction digests (*DraI*, *EcoRV*, *PvuII*, *StuI*), and adaptors were added to the digested ends creating GenomeWalker libraries. The SSR flanking sequence was used to design primers that amplify away from the repeat element and towards the GenomeWalker adaptor (Siebert et al. 1995). PCR reactions using a nested set of such primers were performed and the product sequenced. This provided an additional flanking sequence from Matina 1-6. This sequence was used to increase the length of known repeat flanking sequences until primers could be designed to amplify and sequence across the diversity panel for SNP identification. The SNP markers Cir37s112, Cir160s384, Cir211s1036, Cir222s296, and Cir222s384 were created through this method. Some of these SNP markers have been described previously (Livingstone et al. 2011); the remaining markers are described in Table 3. When possible, SNP markers were mapped as previously described (Livingstone et al. 2011) or, in cases of converted SSR markers, the linkage group locations are based upon the original SSR marker location.

#### Field-suitable SNP assay

A comparison of PCR reagents was performed and the ability of these reagents to produce viable SNP marker data with a microplate reader was evaluated. SNP assays for six markers were performed using six greenhouse-grown seedlings and reactions were

performed with one of three different methods. The TMM (TaqMan Master Mix) method is identical to that used in Livingstone et al. (2011). A total reaction volume of 25  $\mu$ l included 2.25  $\mu$ l of each primer (10 mM), 0.5  $\mu$ l of FAM-labeled probe (10 mM), 0.5  $\mu$ l of VIC-labeled probe (10 mM), 6  $\mu$ l water, 12.5  $\mu$ l TaqMan Genotyping Master Mix with ROX (2 $\times$ ) (Applied Biosystems), and 1  $\mu$ l of template DNA (4 ng/ $\mu$ l).

With the method from GE Healthcare (the GE method; Niskayuna, NY, USA), SNP reactions were set up using illustra<sup>TM</sup> PuReTaq Ready-To-Go PCR Beads which contain all PCR reagents and Taq polymerase as a dried pellet in a 96-well plate. For SNP assays using this method, a 25- $\mu$ l reaction contained 2.25  $\mu$ l of each primer (10 mM), 0.5  $\mu$ l of FAM-labeled probe (10 mM), 0.5  $\mu$ l of VIC-labeled probe (10 mM), 18.5  $\mu$ l water and 1  $\mu$ l of template DNA (4 ng/ $\mu$ l) with Ready-To-Go PCR Beads.

The final SNP assay method, Taq, used a standard Taq DNA polymerase (New England Biolabs, Beverly, MA, USA) in the reaction. A 25- $\mu$ l reaction contained 2.25  $\mu$ l of each primer (10 mM), 0.5  $\mu$ l of FAM-labeled probe (10 mM), 0.5  $\mu$ l of VIC-labeled probe (10 mM), 15.3  $\mu$ l water and 1  $\mu$ l of template DNA (4 ng/ $\mu$ l) with 0.2  $\mu$ l Taq polymerase (New England Biolabs), 0.5  $\mu$ l dNTPs (10 mM each), and 2.5  $\mu$ l 10 $\times$  buffer.

Allelic determination was performed by first recording the background fluorescence of a standard 96-well PCR plate containing assay reagents using an FLx 800 TBP Fluorescence Microplate Reader (Bio-Tek Instruments, Winooski, VT, USA). PCR amplification was then carried out on a standard thermocycler using the following cycling parameters: one cycle at 95°C for 10 min, 40 cycles at 95°C for 15 s, then 60°C for 1 min. After amplification, an end-point fluorescence measurement was taken using the FLx 800 TBP Fluorescence Microplate Reader and the Gen 5 control software (BioTek). The fluorescence values (Relative Fluorescence Units, RFU) for each plate were exported from Gen 5 to an Excel spreadsheet which subtracts background fluorescence and plots FAM versus VIC fluorescence for each sample. Samples homozygous for the FAM-labeled allele cluster along the *x*-axis while those homozygous for the VIC-labeled allele cluster along the *y*-axis. Heterozygous samples cluster between the two. Allelic data for each SNP marker were collected

**Table 3** Primers and probes used for SNP assay and their corresponding linkage group

Locus <sup>a</sup>	Linkage group	Gene name (accession no.)	5' Primer	Probes <sup>b</sup>	3' Primer
c4s123	10	c4 (CU530428)	GGTTTTTTTTGGTTATGTTCTGAGTCT	FAM- TGCTAATATATGT[A]JTGACTGC VIC- TGCTAATATATGT[G]JTGACTG	TGCATTTGCATACTCATCAAT
c4s536	10	c4 (CU530428)	TGATGCATGTGAGCTCTCTAAA	FAM- TTTCCCTC[A]GTTCTGC VIC- TTTCCCTC[G]GTTCTGCT	CAATTTGATGCCGAATGAATCAT
c5s639	8	c5 (CU603656)	TGATGAGGCAGATGCTATGACAA	FAM- CTTTATCTC[A]JTGACAAAGTT VIC- TTTATCTC[G]JTGACAAAGTT	TC TTTGTGTATTTTCTCGATAAAGTGC
Cir37s112	10 <sup>c</sup>	mTeCIR37 (AJ271942)	AAAGTGCGTGTGAAGAGTTCCTATC	FAM- ATAATGGAAAGA[C]JAAC TTG VIC- ATAATGGAAAGA[G]JAAC TTG	GCATGGAAACGATCCAAGTTAGTC
Cir160s384	9 <sup>c</sup>	mTeCIR160 (AJ566490)	ATGATGGTGACAAACAGCAAGAAA	FAM- CAAGGATC[A]TTTTTTGCT VIC- AAGGATC[G]TTTTTTGCT	ATGCCTATTAATCACCTAGGTGAGACT
Cir211s1036	8 <sup>c</sup>	mTeCIR211 (AJ566534)	ACCTTAATTTTATGGGAAACGAGGT	FAM- CAATC[A]GTGCTGACTG VIC- AATC[T]GTGCTGACTGAT	CCAAAACAAAATCTTAATGCACCTGTG
Cir222s296	4 <sup>c</sup>	mTeCIR222 (AJ566543)	AGCAGTGCCTTCAACATACTCTGT	FAM- TTCCGGG[C]TTAAAGCT VIC- ATTCCGGG[T]TTAAAGCT	CCAGTTGGCTCAAAAAGTTTGG
Cir222s384	4 <sup>c</sup>	mTeCIR222 (AJ566543)	AGCAGTGCCTTCAACATACTCTGT	FAM- CACTTT[G]CCAAGAGA VIC- CCACTTT[T]CCAAGAGA	CCAGTTGGCTCAAAAAGTTTGG

<sup>a</sup> Additional SNP markers used can be found in Livingstone et al. (2011)

<sup>b</sup> SNP variants are presented inside *square brackets* in probe sequence

<sup>c</sup> Gene was mapped previously using SSR markers (Brown et al. 2008)

**Table 4** Primer sequences used for nested SNP assay

Locus	5' Nested primer	3' Nested primer
w3s41	NA	NA
w3s463	AAAGGCAATCCTTACCCAAGGT	ATCAGGAATGCTCCAAAATAATCAA
w3s558	NA	NA
w7s104	TCCAAATCCAAGGTTTTTGCC	CTCCTGCATTAAAGCAGTAGC
w8s119	GGATGCCCTGTCAAAAAGAAGGTAAGT	GGAACCCCTTTTACTGTTGCT
w8s131	GGATGCCCTGTCAAAAAGAAGGTAAGT	GGAACCCCTTTTACTGTTGCT
w8s288	AGCAACAGTAAAAGGGTTCC	CGCTGAACCTAGAGCCAGATGA
w17s189	GAGATGATTACACTGTTACACC	GCTGAGAGTCAACTTAAATCC
e0050s274	GAGAAAAAAGAGGAAGGG	TCTTGACACGTTCCATGA
e4785s737	CAGCCTGCTATTTATGTC	TGCATTGCTTCTCCTCTT
c4s123	CACGCTTCGCCAAGAAATTG	AAAATCGTGTGCCCTGCAAC
c4s536	CACGCTTCGCCAAGAAATTG	AAAATCGTGTGCCCTGCAAC
c5s639	TGGCACGGAAGCTTTATGGA	TGCCGGAATGATCTTGTGA
Cir160s384	GTTGCTGTTGACATGTTTCAATGAAAGC	GTGTACATGGTTACTTAGATGCACTTGACTG
Cir37s112	TCTCATGGTCACGATGAAAGTGCGTGTGAA	CCATTAGCTGAACGAGCTCTACTGCGT
Cir211s1036	CTCTCTCACTCTCTCACATTCTTTTTTG	TGCTCCAGGCAGATATAGCCAATCACCT
Cir222s296	TGTAGCAGTGCCTTCAACATACTCTGTTGC	TTGTTTCGTCCTATGGTATCAACACTGTTGG
Cir222s384	TGTAGCAGTGCCTTCAACATACTCTGTTGC	TTGTTTCGTCCTATGGTATCAACACTGTTGG

NA no nested primers were designed

in Excel and genotypes from each individual were identified.

A rapid DNA extraction method was compared to the standard FastPrep DNA extraction. We examined the QuickExtract Plant DNA Extraction Solution from Epicentre in the hope that it would reduce the time and equipment needed to extract DNA under field conditions. SNP assays were performed using the TMM method, with DNA extracted by both methods serving as the template for amplification. Due to low observed fluorescence, we performed a nested pre-amplification step on the QuickExtract DNA. A second set of nested primers, outside of the normal SNP assay primers, was made for each SNP marker and used for the first round of QuickExtract amplification (Table 4). No labeled probes were present in the pre-amplification step. The product of this first PCR step was then diluted 1:50 in water and used as template for the standard (TMM) SNP assay. Different dilutions of template in both standard and nested SNP assays were also examined.

#### SSR analysis

SSR analysis was performed using 14 markers (mTc-Cir3, mTc-Cir10, mTc-Cir19, mTc-Cir29, mTc-Cir42,

mTc-Cir60, mTc-Cir87, mTc-Cir101, mTc-Cir128, mTc-Cir141, mTc-Cir146, mTc-Cir230, mTc-Cir261, and SHRSTc23), following the procedure described in Efombagn et al. (2008). This set of markers is currently used in our laboratory because of its high polymorphism level. Additional marker information can be found online (<http://www.shrs.wus.edu/>). Amplified SSR fragments were separated by capillary electrophoresis using an ABI 3730 sequencer and allele size was scored with GeneMapper (ABI, Foster City, CA, USA). The allele size of each marker was recorded in a spreadsheet.

#### Parentage analysis

The SNP genotype data collected from the 171 open pollinated seedlings, as well as the data gathered on the maternal genotypes, were used to identify the fathers of these seedlings from amongst the 125 trees genotyped from the cacao collection at the SHRS. In addition to using the 18 SNP markers, the same trees were also genotyped using 14 SSR markers. Clonal genotypes from the parent pool were combined so that only one genotype is represented for each clone during parentage analysis, and the genotypes of the

maternal trees were listed twice under differing names (addition of “\_F” suffix) to allow for the detection of possible selfing events, giving a total of 88 plants in the pool of potential paternal parents (Table 1). Genotype data from the two marker types were used separately and in conjunction to create a genotype frequency file and perform parental analysis using the Cervus 3.0 software package (Kalinowski et al. 2007). A maternal simulation was performed using 100,000 simulated progeny, the two different maternal genotypes making up the parent pool, and all of the candidate parents. Using this simulation data, a maternal selection was first performed to ensure the maternal identity of the offspring and to serve as the known parent for paternal analysis. A paternal simulation using 100,000 simulated progeny, 40 candidate parents, and a 0.75 proportion of candidate fathers sampled was performed. The results of this simulation were used to run a paternal selection with one known parent using all the potential paternal genotypes.

## Results

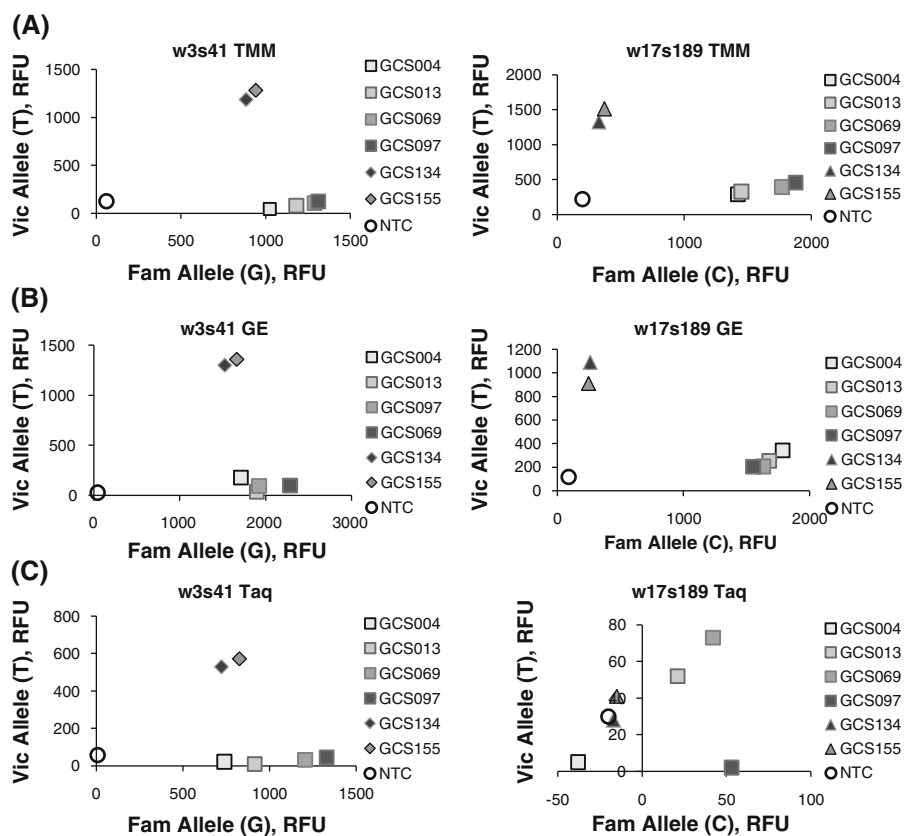
### SNP marker discovery

Of the 18 SNP markers used in this study, eight were identified by converting SSCP markers, five came from amplifying genes of interest, and the remaining five were identified by converting SSR markers into SNP markers. SNP markers were identified by sequencing the diversity panel members at loci of interest, aligning the sequencing results, and calling SNPs. For each SNP identified, the observed minor allele frequency was calculated (Table 2). The minor allele frequency for the SNP markers ranged from 3 to 38%, representing a mixture of rare and common SNPs. The large range of observed minor allele frequencies suggests a good amount of genetic diversity across the members of our SNP discovery panel, as opposed to the uniformly low minor allele frequencies expected in a closely related population. This diversity was expected based on the representation of all major cacao STRUCTURE groups in the panel, and should allow for the identification of SNPs suitable for many populations. The identified SNPs and their surrounding sequences were used to design primers and probes suitable for use in the 5' nuclease SNP assay (Table 3; Livingstone et al. 2011).

### Field-suitable SNP assay

Reactions set up using FastPrep-extracted DNA with the TMM and GE methods consistently demonstrated tight cluster formation for all six markers examined (Fig. 1a, b). However, reactions using the Taq method produced inconsistent results with only two out of six markers (33%) showing moderate cluster formation (Fig. 1c). A rapid DNA extraction method was also compared to the standard FastPrep DNA extraction in the hope that it would reduce the time and equipment needed to extract DNA under field conditions. A comparison of these DNA extraction methods on a subset of trees demonstrated tight cluster formation with the FastPrep samples (Fig. 2a) but low fluorescence levels and no cluster formation with the QuickExtract samples (Fig. 2b). The nested SNP assay with the QuickExtract samples appeared to have increased the observed fluorescence (Fig. 2c); however, cluster formation is not as well defined as with the FastPrep samples (Fig. 2a). Different dilutions of QuickExtract template, in an effort to reduce the amount of potential PCR-inhibiting cellular contaminants, in both standard and nested SNP assays showed no discernable change in cluster formation consistency (data not shown). The experiment was expanded from a subset to all greenhouse-grown plants and examined under the most promising QuickExtract conditions and compared to the FastPrep samples (Fig. 3). The FastPrep samples (Fig. 3a) show strong, well defined cluster formation, while nested SNP assays with the QuickExtract samples produced an indistinguishable arch making cluster identification impossible (Fig. 3b). To determine whether our problems with the QuickExtract samples were limited to SNP markers, we used QuickExtract to isolate DNA from 18 cacao trees and assayed 22 SSR markers. Many of the QuickExtract samples produced very low fluorescence signals (Fig. 4b) and had many extra peaks present. For some loci, alleles could be called; however, the alleles identified using FastPrep samples, shown as black-filled peaks in Fig. 4a, were not always concordant with the alleles identified with QuickExtract samples (Fig. 4b, shaded gray). For the markers assayed, 122 out of 396 reactions (31%) either failed to produce identifiable alleles due to low signal or produced alleles that were not identical with results using FastPrep extractions.

**Fig. 1** Comparison of SNP assay reagents. SNP assays for the w3s41 and w17s189 markers were performed on a subset of greenhouse-grown cacao plants using three different sets of PCR reagents. The PCR reagents included: (a) 2× TaqMan Genotyping Master Mix (TMM) from ABI; (b) Ready-2-Go plates (GE) from GE Healthcare; and (c) standard Taq Polymerase (Taq) from Sigma. *Squares* represent homozygous FAM-labelled alleles, *diamonds* are heterozygous for both alleles, and *triangles* are homozygous VIC-labelled alleles. *Empty circles* are no-template controls. *RFU* relative fluorescence units



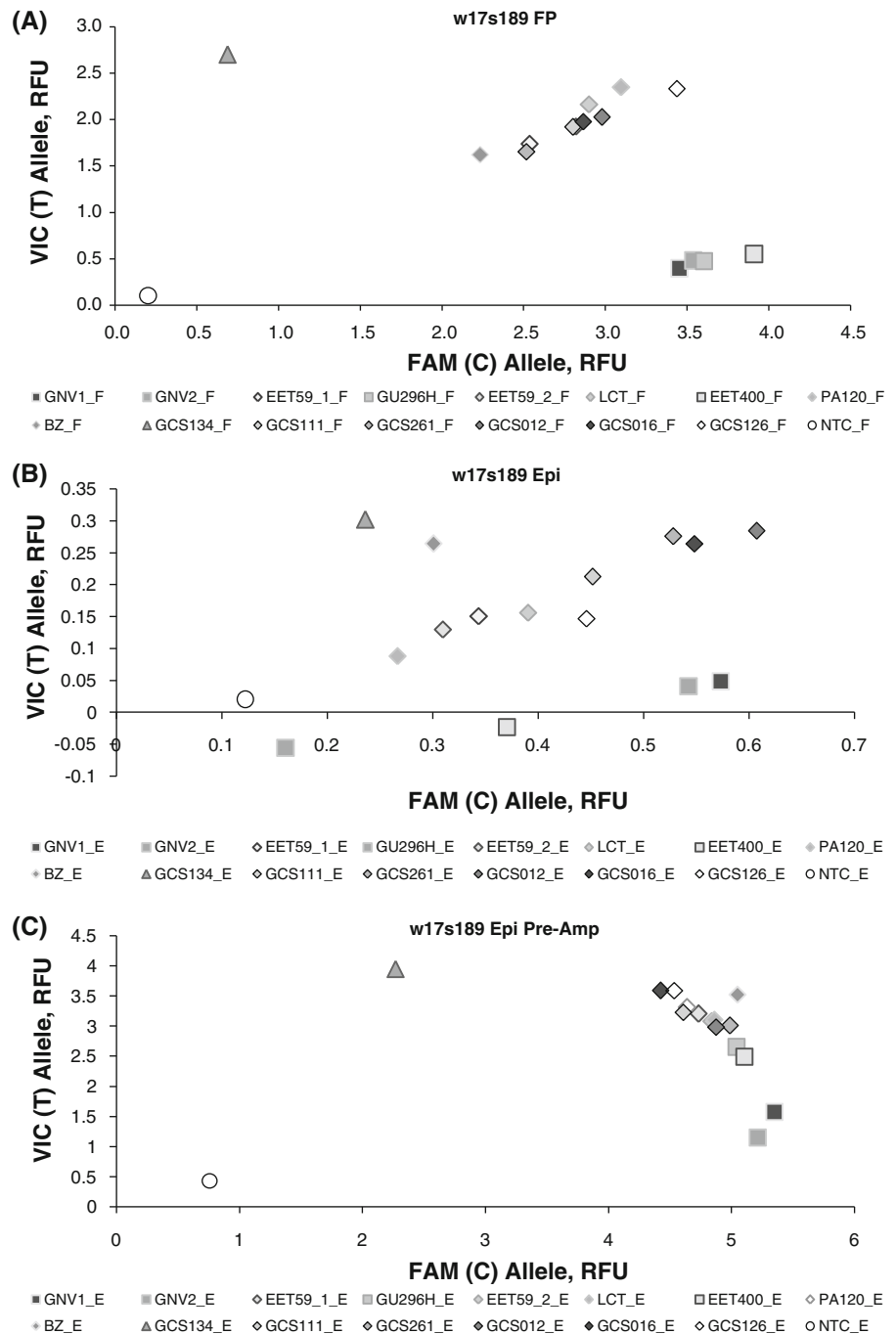
Ultimately, FastPrep-extracted DNA amplified using the TMM method was chosen as the standard for SNP analysis. Using the method with the plate-reader, 171 greenhouse grown seedlings, the four maternal plants, and 125 possible paternal trees were successfully genotyped with SNPs (Online Resource 1). Some 6.5% of the total possible data points are missing from the SNP genotyping results, mostly because of amplification failures (no signal) likely due to DNA degradation. SSR genotyping was previously performed on the same samples and the resulting fragment size data were presented in a spreadsheet (Online Resource 2). SSR data have only 0.4% missing data for the 14 loci that were examined. These genotyping data were used for parental analysis.

#### Parental analysis

The maternal identity analysis revealed that two of the 171 progeny plants, one from pod 2 and one from pod 3, have the opposite maternal genotype to that expected. This is most likely due to mislabeling

during planting and neither of these two individuals was used subsequently. Using the different marker types and given a known mother, the Cervus 3.0 software package generated a summary of the paternal assignments (Table 5). The assignments and assignment rate represent the number of progeny that were assigned a paternal parent at either the strict (95%) or relaxed (80%) confidence level as either a raw number or a percentage, respectively. Observed values are those generated from the actual genotype data, while the expected values are those estimated by the simulations before parentage analysis was completed. The critical delta represents the difference in LOD (logarithm of odds) scores between the two most likely candidate parents needed to obtain a given level of confidence. At a 95% confidence level SNPs identified fewer assigned fathers than SSRs; however, the fathers were most effectively assigned when both marker types were utilized (Table 5). In all, 124 progeny (out of 169) were assigned a paternal parent with 95% confidence when both marker types were used (Table 6). LOD scores identify the

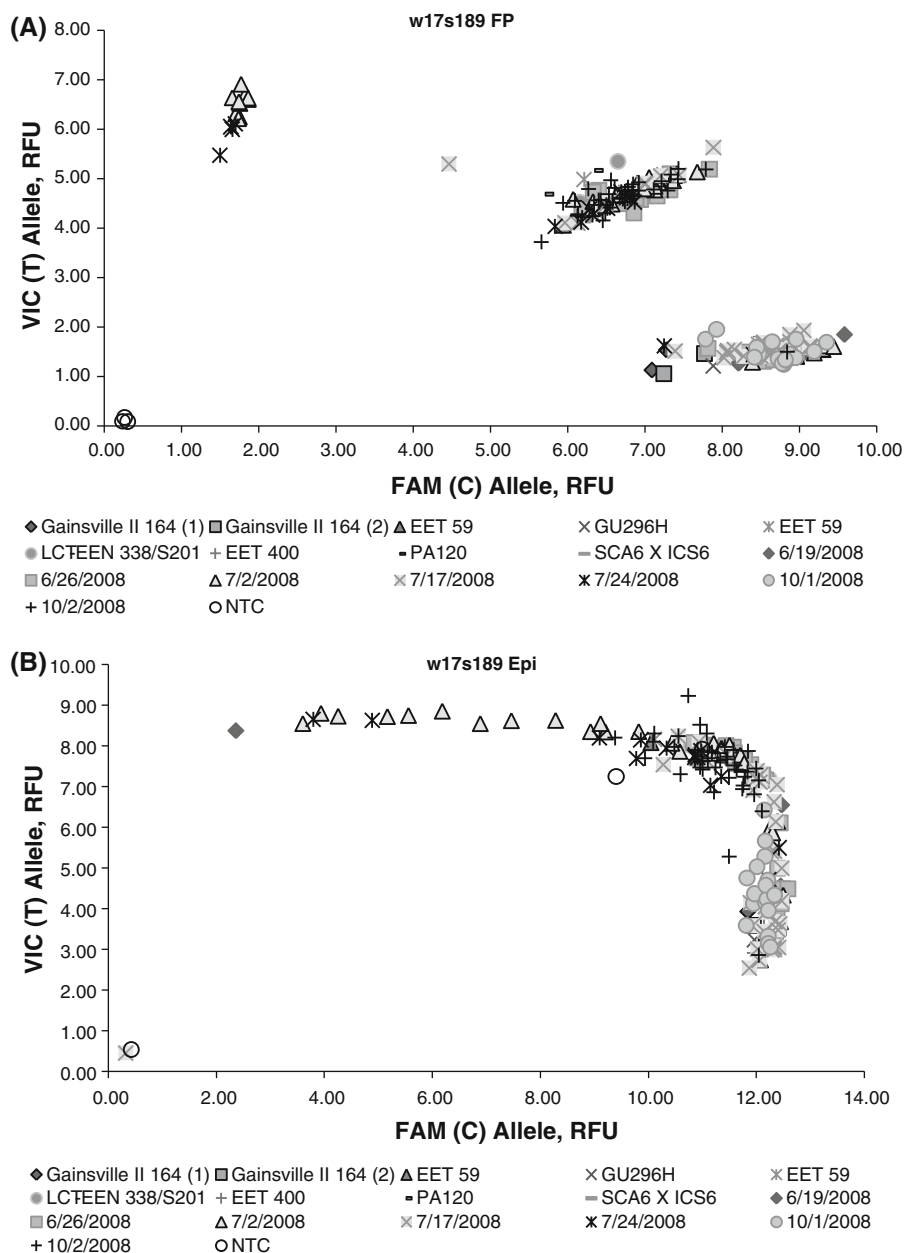
**Fig. 2** Comparison of extraction methods on a subset of plants. SNP assays for the w17s189 marker were performed on a subset of samples, the DNA for which was extracted with (a) Bio101 FastPrep or (b) Epicentre QuickExtract. Epicentre QuickExtract samples were also pre-amplified with unlabeled primers, and then re-amplified with SNP assay primers (c). *Squares* represent homozygous FAM-labelled alleles, *diamonds* are heterozygous at the locus, and *triangles* are homozygous VIC-labelled alleles. *Empty circles* are no-template controls. *RFU* relative fluorescence units



likelihood that a particular plant is more likely than chance to be the parent of a given offspring, scores above zero being more likely than chance. The pair LOD scores represent this value when looking at a given parent and the offspring, while the trio LOD scores represent the likelihood related to both parents and the offspring. The values presented in Table 6 are

the average LOD scores for all individuals assigned a given parent within a pod. All the progeny within a given pod were assigned to either a single father or to multiple different paternal genotypes, suggesting that pollination can occur as either a single or multiple event. Furthermore, some amount of self-fertilization is possible as observed in three progeny from pod 4.

**Fig. 3** Comparison of extraction methods on all greenhouse seedlings and maternal plants. SNP assays were performed for all samples extracted with either (a) Bio101 FastPrep or (b) Epicentre QuickExtract. Epicentre QuickExtract samples were pre-amplified with unlabeled primers, and then re-amplified with SNP assay primers. Maternal samples are identified by their pedigree, while seedlings are differentiated by date of harvest. *Empty circles* are no-template controls. *RFU* relative fluorescence units



## Discussion

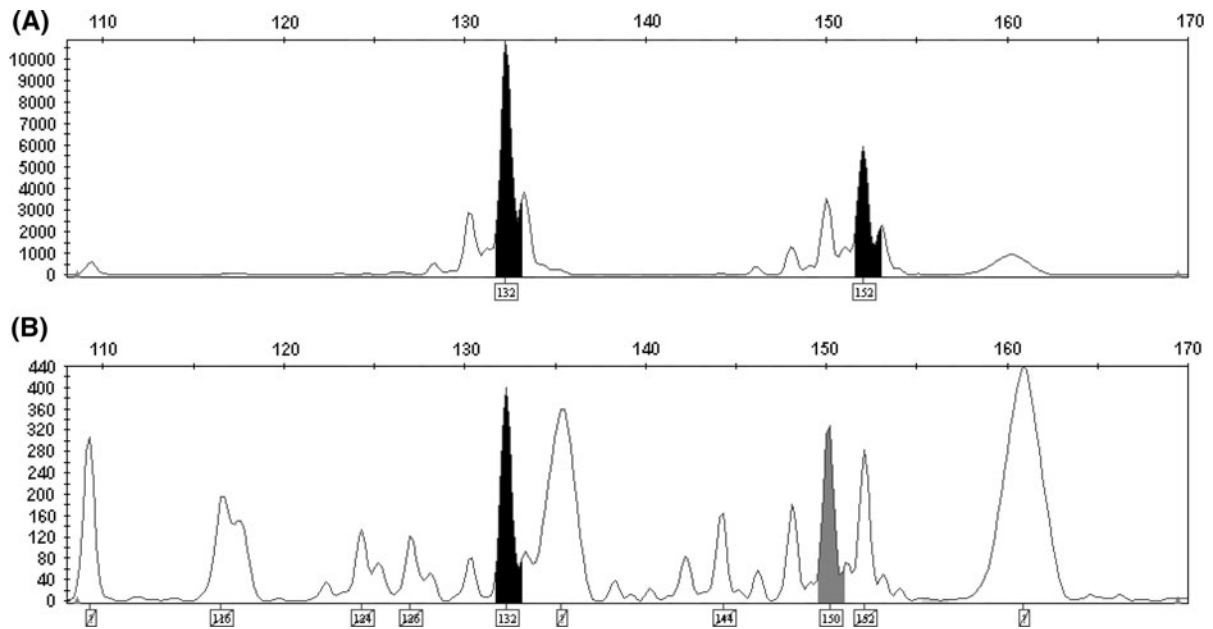
### SNP marker discovery

Using a variety of means, we identified 18 SNP markers that have been successfully used in a 5' nuclease assay. These markers have an average minor allele frequency of 20.5%, which is within the norm for what is usually reported for other SNP discovery projects (Lijavetzky et al. 2007; Van et al. 2005; Van

Tassell et al. 2008). The calculated minor allele frequency of our SNPs is a reflection of the genetic diversity found within the members of our SNP identification (diversity) panel (Table 2).

### Field-suitable SNP assay

Previously, we have demonstrated that SNP-based markers perform equally as well as SSR markers for the purpose of clonal determination and have the



**Fig. 4** Comparison of FastPrep and QuickExtract isolated DNA in SSR analysis. Alleles were identified using SSR markers and cacao DNA isolated using either the FastPrep

(a) or the QuickExtract (b) methods. Alleles identified using FastPrep samples are shown as *black-filled peaks*. Alleles unique to analysis with QuickExtract samples are *shaded gray*

**Table 5** Summary of paternal assignment given known mother with SNP markers, SSR markers, and a combination of both marker types

Marker type	Level	Confidence (%)	Critical delta	Assignments		Assignment rate	
				Observed	Expected	Observed (%)	Expected (%)
SNP	Strict	95	4.33	5	11	3	7
	Relaxed	80	2.05	12	35	7	21
	Unassigned			158	135	93	79
SSR	Strict	95	0	116	131	68	77
	Relaxed	80	0	116	131	68	77
	Unassigned			55	40	32	23
SNP + SSR	Strict	95	0	124	130	73	76
	Relaxed	80	0	124	130	73	76
	Unassigned			47	41	27	24

added benefits of unambiguous data and less expensive equipment and training costs (Livingstone et al. 2011). In this manuscript, we examine modifications to our 5' nuclease cacao SNP assay that will allow this assay to be performed in most cacao production areas. The first modification made to the previously described assay was the replacement of a real-time PCR system with a standard thermocycler and fluorescence microplate reader, which reduces equipment costs without sacrificing result quality.

The inconsistency of the Taq reagents (Fig. 1c) compared to the TMM (Fig. 1a) and GE (Fig. 1b) reagents is surprising. The PCR product is formed in the reactions, as determined by agarose gel electrophoresis (data not shown); however, the lack of fluorescence signal suggests that the probe is not being digested. It is possible that adjustment of the salt concentration in the reaction mix may produce more reliable results. In fact, collaborators in Ghana have had success using standard Taq polymerase in

**Table 6** Summary results of parentage analysis with SSR and SNP markers at a 95% confidence level

Pod ID	Number of offspring	Maternal genotype	Average pair LOD score	Putative candidate paternal genotype	Average pair LOD score	Pair delta	Average trio LOD score	Trio delta
1	6	GainesvilleII164	0.9	LCTEEN162_S1010	17.71	14.96	25.87	23.06
2	3	GainesvilleII164	-0.89	POUND12	13.28	2.00	20.90	5.19
	3	GainesvilleII164	1.12	ICS1	11.56	2.06	17.87	2.24
	2	GainesvilleII164	2.22	UF676	10.09	0.31	17.52	0.00
	1	GainesvilleII164	0.40	LCTEEN338_S201	7.07	0.37	13.98	0.37
	14	GainesvilleII164	-	Unassigned father	-	-	-	-
3	11	EET59	4.64	Bz000221_SCA6xICS6	3.62	3.05	6.21	5.49
	2	EET59	5.94	SCA6	-3.05	0.00	0.99	0.78
	21	EET59	-	Unassigned father	-	-	-	-
4	2	GainesvilleII164	3.30	Bz000250_SCA6xICS6	7.32	2.68	12.65	5.31
	7	GainesvilleII164	3.69	Bz000265_SCA6xICS6	4.97	1.21	9.06	3.69
	3	GainesvilleII164	10.19	GainesvilleII164_F	8.39	8.39	7.09	7.09
	2	GainesvilleII164	6.15	EQXZ	6.91	4.50	6.61	4.10
	1	GainesvilleII164	5.67	Bz000219_SCA6xICS6	2.74	0.00	5.58	0.36
	5	GainesvilleII164	5.11	Bz000241_SCA6xICS6	0.66	0.86	3.50	2.59
	4	GainesvilleII164	5.02	Bz000214_SCA6xICS6	0.87	0.00	3.00	2.67
	3	GainesvilleII164	5.27	Bz000220_SCA6xICS6	1.53	0.00	2.54	2.54
	1	GainesvilleII164	5.94	Bz000283_SCA6xICS6	-0.17	0.00	2.54	2.54
	2	GainesvilleII164	4.32	Bz000234_SCA6xICS6	-0.38	0.16	1.99	1.99
	1	GainesvilleII164	6.06	Bz000208_SCA6xICS6	-1.23	0.00	0.07	0.07
	2	GainesvilleII164	-	Unassigned father	-	-	-	-
5	9	EET59	5.10	IMC11	14.58	13.43	21.73	18.09
	4	EET59	4.79	AMAZ1515	14.56	14.36	20.51	20.51
	1	EET59	4.99	SPA16_11	10.08	0.00	20.22	2.06
6	18	GainesvilleII164	2.29	U26	15.59	14.08	21.17	21.17
7	30	GainesvilleII164	1.53	GAINESVILLEII360	13.01	11.41	20.67	20.68
	2	GainesvilleII164	6.08	EET59	4.52	1.14	8.58	5.21
	1	GainesvilleII164	1.60	SPEC138_15	6.40	3.09	1.01	1.01
	7	GainesvilleII164	-	Unassigned father	-	-	-	-

their reactions (personal communication, J. Takrama). Each of these reagents (GE, TMM, and Taq) offers different benefits. The GE reagents are present in 96-well plates as a freeze-dried pellet. Each pellet contains polymerase, dNTPs, and reaction buffer so that only primers, probes and water need to be added for reaction setup. This method is the simplest to set up and provides for cluster formation at least equal to that achieved using specialized genotyping master mixes. The GE reagents are also thermostable and can be stored at ambient temperature indefinitely. This is particularly useful in areas where reagent degradation due to power failures affecting proper

storage is common. The simplified setup is also advantageous for regions that lack highly trained personnel. However, the GE reagents are the most expensive of the three tested and may be cost-prohibitive for high-throughput analysis. The TMM reagents resulted in consistent, identifiable cluster formation and offer a good compromise of simple setup and cost. When choosing between the GE and TMM method, one must balance the need for thermostable storage with cost. The Taq reagents are the least expensive and may provide a cheap alternative to ready-made master mixes for select loci. However, setup is not as simple as with the other

two methods because the user is required to prepare master mix for each PCR run and cluster formation results were the least reliable across different loci.

While some cacao breeding programs in cocoa-producing countries have the capability to extract DNA by a variety of methods, we wanted to examine quick extraction procedures that could be applied to our 5' nuclease SNP assay for less equipped programs. The samples that were extracted with the FastPrep method reliably demonstrated good cluster formation and allowed for simple identification of alleles (Figs. 2a, 3a). Samples extracted with QuickExtract, however, showed very low fluorescence with a standard SNP assay (Fig. 2b) and indistinguishable clustering with a nested SNP assay (Fig. 3b). The large fluorescence signal of several of the no-template controls in the QuickExtract reaction may be the result of over-amplification of trace contaminants; however, repeated attempts with less template and less cycling still produced similar results (data not shown). The failure of the SNP assay with QuickExtract samples is not well understood. It is possible that the concentration of DNA isolated by QuickExtract is simply too low to produce enough fluorescence, although this would not explain the failure of the nested SNP assay. Another possibility is the presence of cellular compounds that are inhibiting PCR. PCR inhibitors could cause low fluorescence in standard SNP assays and inconsistent amplification during the first step of a nested SNP assay. Attempts to dilute out any inhibitors did not succeed and may have compounded the problem of low sample concentration. Both of these possibilities are supported by the SSR data, as these QuickExtract samples consistently showed low signal intensities. Finally, it is possible that our difficulties are cacao-specific, as QuickExtract samples have been successfully utilized in our laboratory for other species (Kuhn et al. 2010). Perhaps further optimization of reaction conditions may improve the performance of these quick extraction methods in genotyping assays for cacao. Certainly, a method that decreased the time and material needed to extract DNA, and yet allowed for accurate assaying of SNP markers, would greatly improve throughput and allow the timely analysis of thousands of seedlings. However, since many cacao germplasm collections currently have facilities that are capable of performing DNA extraction, a QuickExtract method is not essential to the success of the method.

With a standard thermocycler, a fluorescence microplate reader, and using a genotyping master mix we were able to successfully genotype FastPrep-extracted DNA from 300 cacao trees. Genotype data were collected for 18 different SNP markers representing 12 loci (Online Resource 1). Less than 7% missing data were reported for the entire data set, most of which were caused by poor amplification due to template degradation. Template degradation can be overcome by preparing additional DNA samples, but we did not have any remaining tissue for many of our 300 trees. Thus, by using the modified 5' nuclease SNP assay, genotyping data can be successfully generated under field conditions.

#### Application to parental analysis

Early papers have discussed the use of SNP genotyping data, including its successful application to clonal determination in cacao (Livingstone et al. 2011; Rafalski 2002; The Bovine HapMap Consortium 2009; Yoon et al. 2007). We attempted to evaluate the performance of our SNP markers when performing a parental analysis. Breeding programs carefully control and document all crosses; however, occasionally one of the parents of a cross, usually the father, is uncertain. Also, seed gardens are often planted in the field as an array of maternal and paternal genotypes in such a way so as to avoid manual pollinations. Through parental analysis of the resulting seedlings it is possible to determine the success of these crosses. Here the genotypes of the known parent and the progeny can be compared to the genotypes of the unknown potential parents to identify the true parent of the progeny. Using only the SSR markers, 116 paternal assignments (68% of offspring) were determined with 95% confidence (Table 5). This is in comparison to using only the SNP markers which assigned fathers to only five of the offspring (3%) with a 95% confidence (Table 5). This demonstrates the drawbacks of a biallelic marker system for which more marker information is needed to confidently exclude individuals as potential fathers, as opposed to the multiallelic nature of SSRs which allows for the presence of more unique alleles with which to identify fathers. The use of more SNP markers, especially markers specific to particular varieties such as those being developed as part of the cacao genome project, may help to

alleviate the problem, but the number of additional SNPs required to carry out parental determination in cacao has yet to be determined. However, when a combination of both SSR and SNP markers was used the results were improved, as 124 assignments (73%) were made. It is not surprising that the addition of the SNP markers improved the number of parental assignments, as additional markers provide greater resolving power. This increase in assignments is greater than the number of assignments made when using SNPs alone, suggesting that additional SNP markers, especially SNP markers at additional loci, will allow for increased reliability in parental determinations (Table 5).

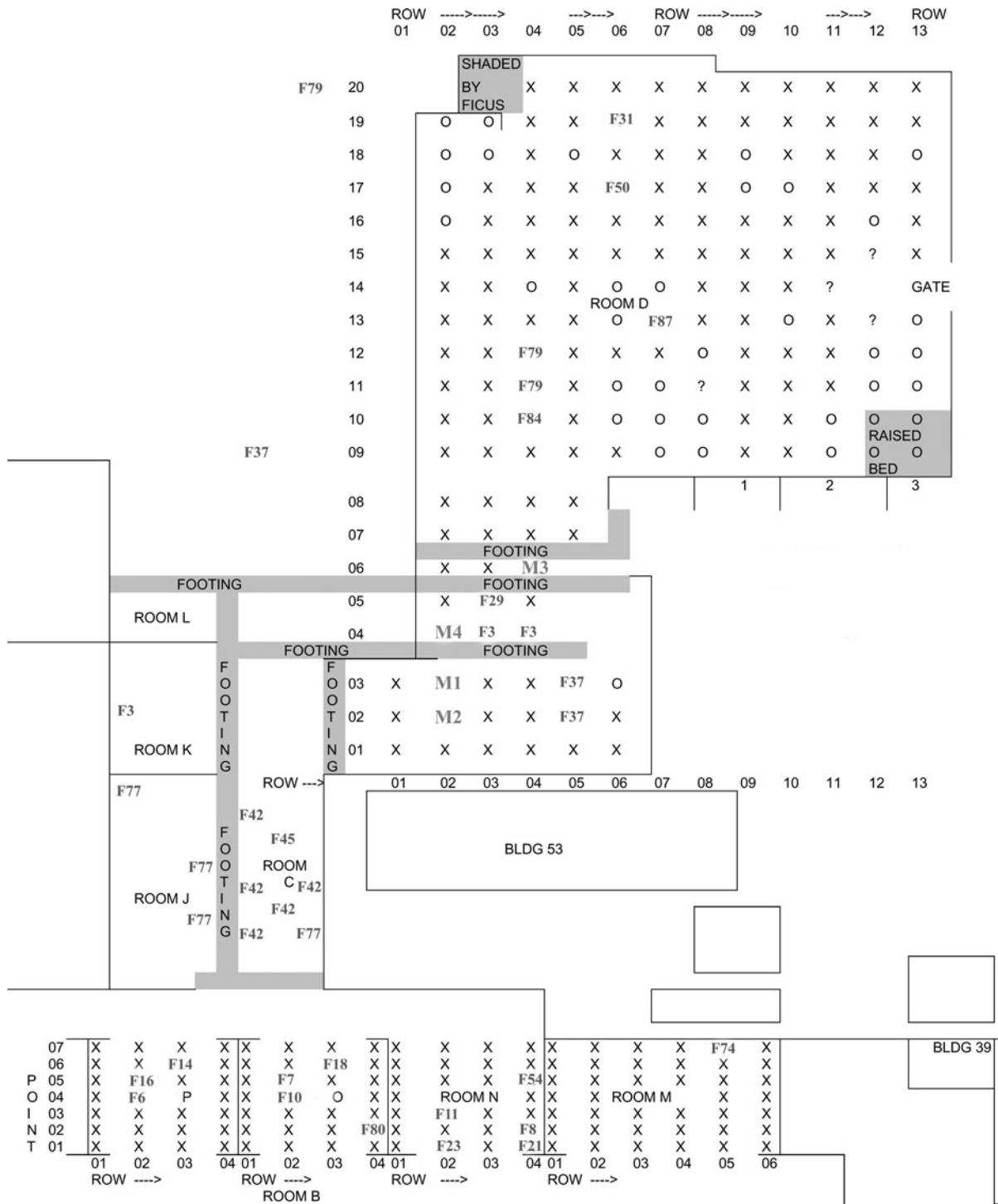
Based on the fathers assigned to the seedlings of a given pod (Table 6), some conclusions can be made about the pollination events leading to the generation of the pods examined. From this data we can see that pod #1 and pod #6 both show only one assigned father for all of the offspring from that pod, suggesting that a single pollination event occurred in these pods. However, the other pods sampled show strong assignment for multiple fathers, which indicates that multiple non-exclusive pollination events took place in each. Thus, apparently, either single or multiple pollen donors can fertilize a single cacao pod.

To take into account any possible self-fertilizations, the maternal genotypes were listed twice in the parent pool with one having “\_F” appended to its name. Cervus 3.0 then treats this second genotype as a separate possible father; otherwise the genotype of the known mother is removed by the software as a possible father. Using this method, we found strong support for three offspring from pod #4 from the clone Gainesville II 164 (M2) having been either self-pollinated or pollinated by another genetic clone of Gainesville II 164 (M1). Although cacao is generally thought to be self-incompatible, this suggests the possibility of a self-pollination event in the context of a multiple pollination event. The percentage of self-fertilized seedlings (8%) within a pod is consistent with previously observed self-pollination rates (Glendinning 1960). Furthermore, certain methods utilizing multiple pollen donors have been used in the past to help to circumvent self-incompatibility mechanisms when selfing cacao. One of these methods is the use of mentor-pollen, i.e., a mixture of self-incompatible and compatible pollen in different ratios

and usually containing phenotypic markers or enzymatic markers to distinguish selfs from non-selfs (Glendinning 1960; Lanaud et al. 1987; Opeke and Jacob 1967; Posnette 1940). Another method is pollination with *Herrania* pollen followed by pollination with cacao self-pollen. The resulting seeds of this cross-fertilization are recognized by their abnormal shape (Bartley 1969, 2005; Opeke and Jacob 1967). A third method that also results in viable seeds is the mixing of irradiated compatible pollen and normal self-incompatible pollen (Adu-Ampomah et al. 1990).

From a field map showing the location of the maternal and assigned paternal plants (Fig. 5), we can see that the assigned father is often separated from the maternal plants not only by distance but also by manmade objects. Because cacao pollen is very sticky, wind pollination is unlikely. In agricultural or agroforestry settings, pollination seems to be performed primarily by midges within two families: the ceratopogonid midges [especially *Forcipomyia* (*Euprojoannisia*)] and the cecidomyiid midges. The ceratopogonid midges are less abundant than the cecidomyiid midges, but they are more effective pollinators (Frimpong et al. 2009; Young and Severson 1994). However, studies have shown that multiple visits of midges are necessary (about 20 pollen grains are deposited after a single midge visit) to obtain the highest number of seeds per pod (Falque et al. 1995). The midges can only fly up to 5–6 m, but they can reach longer distances with higher wind speeds (Decazy and Coulibaly 1981). There are many species of these two families of midges present in Florida, some of which have been documented as pollinators (Chan and Linley 1989; personal communication, P. Kendra and N. Epsky), and it is certainly possible that these midges served as the pollinators for the cacao plants in this study.

Unfortunately, not all of the possible fathers in Miami were genotyped, resulting in a large number of offspring without confidently assigned fathers. Additionally, the presence of clones in the potential paternal pool and the need for additional markers to distinguish a few non-clonal trees from one another prevented the identification of the actual pollen donor plants in our analysis. These factors would need to be addressed in any future parental analysis experiments. Currently, it appears that the multiallelic nature of SSRs provides more resolution than SNPs when inferring a parental genotype. However, the



**Fig. 5** Field map of maternal and assigned paternal trees. This map represents the cacao trees planted at the SHRS in Miami, Florida. Maternal trees and paternal trees assigned to the offspring with a 95% confidence level during parentage

analysis are identified according to Table 1. In all, 125 individual cacao trees radiating out from the maternal trees were sampled for parental analysis

vast number of SNPs within the genome and careful SNP selection in the future may help to overcome some of the disadvantages created by the biallelic nature of SNPs. Furthermore, breeding programs utilizing MAS often perform manual crosses between known maternal and paternal trees. In these instances, clonal determination and identification of particular alleles within progeny are of greater import than parental determination. In these two particular roles, SNP markers perform well and should serve as a suitable alternative to SSRs, especially when SNP analysis can be performed on site as opposed to outsourcing SRR analysis.

## Conclusions

The continuation of the cacao genome sequencing project will result in the discovery of hundreds of additional SSR markers and tens of thousands of SNP markers. This will allow for the saturation of the cacao genetic map and increased precision of identified QTL regions. SNP-based markers, while individually less polymorphic than SSR markers, promise to provide a genotyping system that will allow on-site analysis in producer regions. This will allow for the creation of an international database of unambiguous SNP genotypes that can be available to breeders worldwide. Similar databases currently exist that contain SSR marker genotypes (<http://www.icgd.rdg.ac.uk/index.php>) but, since SSR markers require an electrophoretic step, the resulting data can vary from machine to machine, thus compromising the transfer of genotypes between laboratories. SNPs using the 5' nuclease assay, on the other hand, do not require any electrophoresis steps and produce genotypes that can be easily transferred (Kuhn et al. 2008; Rafalski 2002). A SNP-based assay appears to provide a nice balance of throughput, simplicity, cost savings, and data transferability that would be of vital use to breeders in cacao-producing regions.

We set out to adapt a SNP marker assay that requires sophisticated laboratory equipment so that it can be performed in any region where cacao is grown. Most cacao-breeding centers throughout the world have limited resources for new specialized equipment. By replacing the need for a real-time PCR system with a fluorescence microplate reader and standard thermocycler, the cost of entry for on-site

SNP genotyping is reduced. Many of these centers currently have a thermocycler and the means to extract DNA. The microplate reader can also be used to quantify DNA samples (Leggate et al. 2006; Livingstone et al. 2009; Vitzthum et al. 1999), thereby providing multiple uses for a single piece of equipment. The use of thermostable reagents is also a possibility for regions where power supply is inconsistent. Even though additional markers may be required for parental determination, SNP markers have been shown to be very useful for clonal (off-type) determination (Livingstone et al. 2011) and, with the identification of new markers linked to desirable traits, SNPs promise to be a useful tool for the cacao breeder. In addition, novel SNP assays are now being developed that reduce reagent costs and facilitate analysis (Hirotsu et al. 2010). By demonstrating that seedlings can be successfully genotyped under simulated field conditions, we believe we have provided cacao-producing regions with a viable alternative to outsourcing genotyping analysis. This basic system has been established at the Cocoa Research Institute of Ghana, where it is being used successfully for genotyping and clonal determination and promises to streamline their cacao breeding efforts.

**Acknowledgments** The authors wish to thank Dr. Nancy Epsky and Dr. Paul Kendra for their discussions about midges and cacao pollination. The authors are also grateful to Mike Winterstein, Carol Lee, and Paul Kuhn for their assistance maintaining and planting the cacao seedlings used in this study. The authors would also like to thank MARS, Inc. for their continued financial support.

## References

- Adu-Ampomah Y, Novak FJ, Klu GYP, Lamptey TVO (1990) Use of irradiated pollen as mentor pollen to induce self-fertilization of two self-incompatible Upper Amazon cacao clones. *Euphytica* 51:219–225
- Bartley BGD (1969) Selfing of self-incompatible trees. *Ann Rep Cacao Res* (1968), Trinidad, pp 22–23
- Bartley BGD (2005) The Genetic diversity of cacao and its utilization. CABI Publishing, Wallingford, p 341
- Borrone J, Kuhn D, Schnell R (2004) Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. *Theor Appl Genet* 109: 495–507
- Borrone JW, Brown JS, Kuhn DN, Motamayor JC, Schnell RJ (2007) Microsatellite markers developed from *Theobroma*

- cacao* L. expressed sequence tags. *Mol Ecol Notes* 7:236–239
- Brown JS, Schnell RJ, Motamayor JC, Lopes U, Kuhn DN, Borrone JW (2005) Resistance gene mapping for witches' broom disease in *Theobroma cacao* L. in an F2 population using SSR markers and candidate genes. *J Am Soc Hort Sci* 130:366–373
- Brown J, Sautter R, Olano C, Borrone J, Kuhn D, Motamayor J, Schnell R (2008) A composite linkage map from three crosses between commercial clones of cacao, *Theobroma cacao* L. *Trop Plant Biol* 1:120–130
- Chan KL, Linley JR (1989) A new Florida species of *Forcipomyia* (*Euprojoannisia*) (*Diptera: Ceratopogonidae*) from leaves of the water lettuce, *Pistia stratiotes*. *Fla Entomol* 72:252–262
- Decazy B, Coulibaly N (1981) Behaviour of cacao cultivars with respect to biting-sucking insects: possibility of early selection of tolerant cacao trees. In: Proceedings of 8th international Cocoa Res conference, Cartagena. COPAL, Nigeria
- Duguma B, Gockowski J, Bakala J (2001) Smallholder cacao (*Theobroma cacao* Linn.) cultivation in agroforestry systems of West and Central Africa: challenges and opportunities. *Agrofor Syst* 51:177–188
- Efombagn I, Motamayor J, Sounigo O, Eskes A, Nyassé S, Cilas C, Schnell R, Manzanares-Dauleux M, Kolesnikova-Allen M (2008) Genetic diversity and structure of farm and GenBank accessions of cacao (*Theobroma cacao* L.) in Cameroon revealed by microsatellite markers. *Tree Genet Genom* 4:821–831
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8:186–194
- Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8:175–185
- Faleiro FG, Queiroz VT, Lopes UV, Guimaraes CT, Pires JL, Yamada MM, Araujo IS, Pereira MG, Schnell R, de Souza GA, Ferreira CF, Barros EG, Moreira MA (2006) Mapping QTLs for witches' broom (*Crinipellis pernicioso*) resistance in cacao (*Theobroma cacao* L.). *Euphytica* 149:227–235
- Falque M, Vincent A, Vaissiere B, Eskes A (1995) Effect of pollination intensity on fruit and seed set in cacao (*Theobroma cacao* L.). *Sex Plant Reprod* 8:354–360
- Frimpong EA, Gordon I, Kwapong PK, Gemmill-Herren B (2009) Dynamics of cocoa pollination: tools and applications for surveying and monitoring cocoa pollinators. *Int J Trop Insect Sci* 29:62–69
- Glendinning DR (1960) Selfing of self-incompatible cocoa. *Nature* 187:170
- Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. *Genome Res* 8:195–202
- Hirotsu M, Murakami N, Kashiwagi T, Ujiie K, Ishimaru K (2010) Protocol: a simple gel-free method for SNP genotyping using allele-specific primers in rice and other plant species. *Plant Methods* 6:12
- ICCO (2010) Latest quarterly bulletin of cocoa statistics. <http://www.icco.org/about/press2.aspx?Id=onl12753>. Accessed 8 Feb 2011
- Irish BM, Goenaga R, Zhang D, Schnell R, Brown JS, Motamayor JC (2010) Microsatellite fingerprinting of the USDA-ARS Tropical Agriculture Research Station cacao (*Theobroma cacao* L.) germplasm collection. *Crop Sci* 50:656–667
- Kalinowski ST, Taper ML, Marshall TC (2007) Revising how the computer program cervus accommodates genotyping error increases success in paternity assignment. *Mol Ecol* 16:1099–1106
- Kuhn DN, Heath M, Wisser RJ, Meerow A, Brown JS, Lopes U, Schnell RJ (2003) Resistance gene homologues in *Theobroma cacao* as useful genetic markers. *Theor Appl Genet* 107:191–202
- Kuhn DN, Motamayor JC, Meerow AW, Borrone JW, Schnell RJ (2008) SSCP markers provide a useful alternative to microsatellites in genotyping and estimating genetic diversity in populations and germplasm collections of plant specialty crops. *Electrophoresis* 29:1–14
- Kuhn D, Figueira A, Lopes U, Motamayor J, Meerow A, Cariaga K, Freeman B, Livingstone D, Schnell R (2010) Evaluating *Theobroma grandiflorum* for comparative genomic studies with *Theobroma cacao*. *Tree Genet Genom* 6:783–792. doi:10.1007/s11295-010-0291-0
- Lanaud C, Sounigo O, Amefia YK, Paulin D, Lachenaud P, Clément D (1987) New data on the mechanisms of incompatibility in cocoa and its consequences on breeding. *Café Cacao Thé* 31:278–282
- Lanaud C, Risterucci AM, Pieretti I, Falque M, Bouet A, Lagoda PJL (1999) Isolation and characterization of microsatellites in *Theobroma cacao* L. *Mol Ecol* 8:2141–2143
- Lanaud C, Fouet O, Clément D, Boccara M, Risterucci A, Surujdeo-Maharaj S, Legavre T, Argout X (2009) A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol Breed* 24:361–374
- Leggatt J, Allain R, Isaac L, Blais B (2006) Microplate fluorescence assay for the quantification of double-stranded DNA using SYBR Green I dye. *Biotechnol Lett* 28:1587–1594
- Lijavetzky D, Cabezas J, Ibanez A, Rodriguez V, Martinez-Zapater J (2007) High-throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8:424
- Lima L, Gramacho K, Carels N, Novais R, Gaiotto F, Lopes U, Gesteira A, Zaidan H, Cascardo J, Pires J, Micheli F (2009) Single nucleotide polymorphisms from *Theobroma cacao* expressed sequence tags associated with witches' broom disease in cacao. *Genet Mol Res* 8:799–808
- Livingstone D III, Freeman B, Tondo CL, Cariaga KA, Oleas NH, Meerow AW, Schnell RJ, Kuhn DN (2009) Improvement of high-throughput genotype analysis after implementation of a dual-curve Sybr Green I-based quantification and normalization procedure. *Hort Sci* 44:1228–1232
- Livingstone D, Motamayor J, Schnell R, Cariaga K, Freeman B, Meerow A, Brown J, Kuhn D (2011) Development of single nucleotide polymorphism markers in *Theobroma cacao* and comparison to simple sequence repeat markers for genotyping of Cameroon clones. *Mol Breed* 27: 93–106

- Motamayor JC, Risterucci AM, Lopez PA, Ortiz CF, Moreno A, Lanaud C (2002) Cacao domestication I: the origin of the cacao cultivated by the Mayas. *Heredity* 89:380–386
- Motamayor JC, Lachenaud P, e Mota JW, Loor R, Kuhn DN, Brown JS, Schnell RJ (2008) Geographic and genetic population differentiation of the Amazonian chocolate tree (*Theobroma cacao* L.). *PLoS ONE* 3:e3311
- Opeke LK, Jacob VJ (1967) Studies on methods of overcoming self-incompatibility in *Theobroma cacao* Linn. In: 2e Conférence Internationale sur les Recherches Cacaoyères, pp 356–359
- Posnette AF (1940) Self-incompatibility in cocoa (*Theobroma* spp.). *Trop Agric* 17:67–71
- Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 5:94–100
- Rice RA, Greenberg R (2000) Cacao cultivation and the conservation of biological diversity. *Ambio* 29:167–173
- Schnell RJ, Kuhn DN, Brown JS, Olano CT, Phillips-Mora W, Amores FM, Motamayor JC (2007) Development of a marker assisted selection program for cacao. *Phytopathology* 97:1664–1669
- Schnell RJ, Motamayor JC, Brown JS, Kuhn DN, Tondo CL, Livingstone D, III, Royaert S, Nagai C, Phillips W, Amores FM, Suarez-Capello C, Lopes U, Takrama J, Padi F, Opoku S, Efombagn IB, Aikpokpodion P, Pokou D, Epaina P, Marfu J (in press) The international marker-assisted selection program for cacao. In: Proceedings of 16th international cocoa research conference, Bali, 2009. COPAL, Nigeria
- Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* 23:1087–1088
- The Bovine HapMap Consortium (2009) Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* 324:528–532
- Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF, Schnabel RD, Lawley CT, Haudenschild CD, Moore SS, Warren WC, Sonstegard TS (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth* 5:247–252
- Van K, Hwang EY, Kim MY, Park HJ, Lee SH, Cregan PB (2005) Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *J Hered* 96:529–535
- Vitzthum F, Geiger G, Bisswanger H, Brunner H, Bernhagen J (1999) A quantitative fluorescence-based microplate assay for the determination of double-stranded DNA using SYBR Green I and a standard ultraviolet transilluminator gel imaging system. *Anal Biochem* 276:59–64
- Yoon M, Song Q, Choi I, Specht J, Hyten D, Cregan P (2007) BARCSoySNP23: a panel of 23 selected SNPs for soybean cultivar identification. *Theor Appl Genet* 114:885–899
- Young A, Severson D (1994) Comparative analysis of steam-distilled floral oils of cacao cultivars (*Theobroma cacao* L., Sterculiaceae) and attraction of flying insects: implications for a *Theobroma* pollination syndrome. *J Chem Ecol* 20:2687–2703
- Zhang D, Mischke S, Johnson E, Phillips-Mora W, Meinhardt L (2009) Molecular characterization of an international cacao collection using microsatellite markers. *Tree Genet Genom* 5:1–10