Tree Genetics & Genomes (2010) 6:783–792
DOI 10.1007/s11295-010-0291-0

ORIGINAL PAPER

Evaluating *Theobroma grandiflorum* for comparative genomic studies with *Theobroma cacao*

David N. Kuhn · Antonio Figueira · Uilson Lopes · Juan Carlos Motamayor · Alan W. Meerow · Kathleen Cariaga · Barbie Freeman · Donald S. Livingstone III · Raymond J. Schnell

Received: 22 September 2009 / Revised: 5 February 2010 / Accepted: 16 April 2010 / Published online: 3 June 2010

Abstract The seeds of *Theobroma cacao* (cacao) are the source of cocoa, the raw material for the multi-billion dollar chocolate industry. Cacao's two most important traits are its unique seed storage triglyceride (cocoa butter) and the flavor of its fermented beans (chocolate). The genome of *T. cacao* is being sequenced, and to expand the utility of the genome sequence to the improvement of cacao, we are evaluating *Theobroma grandiflorum*, the closest economically important species of *Theobroma* for its potential use in a comparative genomic study. *T. grandiflorum* differs from cacao in important agronomic traits such as flavor of the fermented beans, disease resistance to witches' broom and abscission of mature fruits. By comparing genomic sequences and analyzing viable inter-specific hybrids, we hope to identify the key genes that regulate cacao's most important traits. We have investigated the utility in *T. grandiflorum* of three types of markers (microsatellite markers, single-strand conformational polymorphism markers and single nucleotide polymorphism (SNP) markers) developed in cacao. Through sequencing of amplicons of 12 diverse individuals of both cacao and *T. grandiflorum*, we have identified new intra- and inter-specific SNPs. Two markers which had no overlap of alleles between the species were used to genotype putative inter-specific hybrid seedlings. Sequence conservation was significant and species-specific differences numerous enough to suggest that comparative genomics of *T. grandiflorum* and *T. cacao* will be useful in elucidating the genetic differences that lead to a variety of important agronomic trait differences.

Keywords Microsatellites · SSCP markers · SNPs · Population genetics · Candidate genes · Inter-specific hybrids · Comparative genomics

Abbreviations

HWE	Hardy-Weinberg equilibrium
CAE	capillary array electrophoresis
SNP	single nucleotide polymorphism

Communicated by J. Davis

D. N. Kuhn (✉) · A. W. Meerow · K. Cariaga · B. Freeman · D. S. Livingstone III · R. J. Schnell
USDA-ARS SHRS,
Miami, FL, 33158, USA
e-mail: David.Kuhn@ars.usda.gov

A. Figueira
Centro de Energia Nuclear na Agricultura,
Universidade de Sao Paulo,
CP 96, Sao Paulo SP 13400-970, Brazil

U. Lopes
Mars Center for Cocoa Science,
Caixa Postal 55,
45.630-000 Itajuípe, Brazil

J. C. Motamayor
Mars, Inc.,
800 High St.,
Hackettstown, NJ 07840, USA

Introduction

Theobroma cacao L. (cacao), an understory tree native to the upper Amazon, is cultivated globally in the humid tropics; it is a major source of currency for small farmers, as well as the main cash crop of several West African countries. Cacao is a diploid ($2n=20$) member of the *Malvaceae* with a genome size of approximately 400 Mb (Figueira et al. 1992). Its cauliflorus fruits (pods) contain the seeds (beans) that are later processed by the multi-billion dollar chocolate industry.

Since 2000, we have been involved in an international breeding program to improve cacao (Schnell et al. 2007) and in June 2008, we began the cacao genome sequencing project, a joint effort of USDA, IBM, and Mars, Inc. (the major funding source for this research) and other institutions (www.cacaogenomedb.org). The sequencing project and its associated single nucleotide polymorphism (SNP) discovery project will provide tens of thousands of molecular markers for association studies and marker-assisted selection to develop improved cacao cultivars with enhanced disease resistance and better yield.

Despite this progress in cacao breeding, there are currently few populations that segregate for cacao's most important economic traits, such as flavor and cocoa butter content. Moreover, differences among individuals in those populations are small. Phylogenetic studies of the genus *Theobroma* (Whitlock and Baum 1999; Silva and Figueira 2005; Borrone et al. 2007) place cacao in its own subgroup (Theobroma). *Theobroma grandiflorum*, the second most economically important species in the genus is in the Glossopetalum subgroup. A comparative genomics project with *T. grandiflorum* (Willd ex Spreng) Schum. (cupuassu), could allow us to take greater advantage of the cacao genome sequencing project to improve cacao. *T. grandiflorum* is native to the Brazilian Amazon and a diploid ($2n=20$). Cupuassu pulp, which surrounds the seeds, is used in juices, ice creams, jams, candies, desserts, liquors (Alves et al. 2007) and a product similar to cocoa powder (called "cupulate" in Brazil) can be obtained from fermented seeds. Cupuassu seed fat is distinct from cocoa butter (Alves et al. 2007).

Comparative genomics is an approach to the problem of identifying the genes associated with unique traits, such as cocoa butter production and flavor of cacao, by looking for significant sequence differences between the genes of *T. cacao* and *T. grandiflorum*. Cacao and *T. grandiflorum* are close genetically: microsatellite marker primers which were designed to cacao sequences work well in *T. grandiflorum* (Alves et al. 2007), and sequence comparisons have shown that there is greater than 95% sequence identity between cacao and *T. grandiflorum* for WRKY genes (Borrone et al. 2007). Sequence identity in the WRKY genes allows primers designed to intron sequences in cacao to amplify the same fragments in *T. grandiflorum*. We expect that we will be able to use the completed cacao genome as the template for the assembly of the *T. grandiflorum* genome.

For the *T. cacao* genome sequencing project, we chose cultivar Matina 1-6 because first, it belongs to the Amelonado group which is in the genetic background of most of the world's commercially grown cacao and, second, it was determined to be ~98% homozygous after evaluation of 107 microsatellite loci (Motamayor et al. 2008). Having a highly homozygous cultivar makes assembly of the genome easier because orthologs and paralogs are more simply distinguished. A physical map of the cacao genome has been produced through fingerprinting of three bacterial artificial chromosome libraries created by Clemson University Genomics Institute (http://www.genome.clemson.edu/research/cacao). When the cacao draft genome sequence is available, it should be possible to sequence the *T. grandiflorum* genome only with Illumina GAII technology and align it to the cacao draft genome sequence, basically as a re-sequencing exercise. Nonetheless, it will still be important to identify a highly homozygous cultivar of *T. grandiflorum* for the genome sequencing project, to improve the alignment of the two genomes. Although no genetic recombinant map exists for *T. grandiflorum*, germplasm collections with over 900 accessions (Silva et al. 2004) and commercial cultivars from a variety of countries such as Brazil, Puerto Rico, and Ecuador are available to identify the best cultivar for sequencing. By aligning the sequences of the *T. cacao* and *T. grandiflorum* genomes, differences in the amino acid sequences of candidate genes involved in the traits of interest can be identified. Although trait differences may be due to differences in the expression of regulatory genes, our expectation is that comparison of the genes required for a biosynthetic pathway will show nucleotide differences (inter-specific SNPs) that lead to nonconservative amino acid changes in domains responsible for specific biosynthetic functions, which will narrow our search for candidate genes to investigate.

Another approach to the problem of low phenotypic variability in agronomic traits of commercial cacao is to produce inter-specific hybrids. Fermented *T. grandiflorum* seeds produce a flavor different from chocolate. Previous attempts at forming *T. cacao* x *T. grandiflorum* hybrids have not produced viable plants (Silva et al. 2004), but recently, putative hybrids have been produced at the Mars Center for Cocoa Science in Brazil. Thus, we wanted to identify markers that can unambiguously characterize inter-specific hybrids.

Microsatellites markers are polymerase chain reaction (PCR) based, amenable to high throughput systems, robust, polymorphic, and co-dominant; additionally, their analysis can be automated. Disadvantages of these markers are that they occur less frequently in the genome than SNPs, especially in coding regions; they can be difficult to develop; and they are platform dependent, so that it is difficult to share genotype data with other laboratories.

Single-strand conformational polymorphism (SSCP) markers, which detect SNPs in alleles by altered mobility of single-stranded conformers (electromorphs), are more widely used in animals, but they have advantages for plant systems. SSCP markers are PCR based, polymorphic, and co-dominant; they can be screened by mobility differences in a variety of low to high throughput electrophoretic techniques, including capillary array electrophoresis (Kuhn et al. 2008). They are also easier to develop than microsatellites and can easily be converted to SNP markers. However, they are highly platform dependent, which precludes sharing of genotype data with other laboratories.

SNPs, the most common DNA polymorphism, occur in coding and non-coding regions, are co-dominant and unambiguous (Rafalski 2002); they also can be assayed without electrophoresis for ultra-high throughput. More than 250,000 SNPs have already been identified by sequencing of the leaf transcriptome from genetically diverse cacao cultivars as part of the cacao genome sequencing project (Greg May, unpublished data). Thus, in cacao, SNPs occur frequently, are reasonably easy to develop and assay, and can be shared with other scientists.

Because of the potential for using *T. grandiflorum* in a comparative genomic study with cacao, we evaluated microsatellites, SSCP markers and single nucleotide polymorphism (SNP) markers developed for cacao for their utility in analyzing *T. grandiflorum* genotypes and inter-specific hybrids. We compare these three types of markers for their ability to estimate genetic diversity, their potential use for expression studies, and their utility in identifying highly homozygous individuals for use in whole genome sequencing. We also direct sequence amplicons of *T. cacao* and *T. grandiflorum* from 12 accessions to identify intra- and inter-specific SNPs. Finally, we evaluate the markers' ability to distinguish hybrids between *T. cacao* and *T. grandiflorum*.

Materials and methods

Cultivars used

Germplasm cultivars DNA was extracted from *T. grandiflorum* leaves from 36 individuals from three closely related Brazilian populations described in Alves et al. (2007). Cacao cultivars used for sequence/marker comparisons were: TSH516, PA41, IMC47, NA194, COC3335, GU124A, LasBrisas 17_17, RB_40_PL1, BE_4_PL3, CAB0331_PL4, CAB0339_PL1, SCA6, and Criollo13. Cultivar data and microsatellite genotypes are available at http://cacaodb.shrs.aphis.usda.gov/index.

Putative inter-specific hybrid leaf samples (five to seven, 0.5 mm disks per plant) were from 77 progeny from all possible crosses between ten cacao cultivars as maternal parents and six *T. grandiflorum* cultivars as paternal parents, performed at Mars Center for Cocoa Science (MCCS), Itabuna, Brazil.

DNA isolation

For *T. cacao* and *T. grandiflorum* samples, DNA was isolated from leaf disks as described in (Schnell et al. 2005) using the Fast DNA Kit (Bio101, Carlsbad, CA, USA). DNA was quantitated for microsatellite analysis using an automated SYBR Green assay (Livingstone et al. 2009) and template concentration adjusted to 4 ng/uL. For putative inter-specific hybrids and parent samples from MCCS, DNA was isolated using the Epicentre OneStep DNA isolation procedure (Epicentre, UK) as described by the manufacturer using the following modified extraction cycle: 65°C, 16 min; 98°C, 4 min.

Microsatellite primers

Sequences for the microsatellite primers used in this study can be found in (Motamayor et al. 2008). The microsatellite loci were identified and characterized at the USDA-ARS SHRS, Miami, FL, USA. We tested SHRSTc 02, 03, 04, 05, 06, 07, 08, 11, 13, 18, 19, 21, 22, 23, 24, 25, 27, 28, 29, 31, 32, 33, 34, 36, 37, 38, 39, 41, 43, and 44.

Microsatellite PCR

For DNA isolated by the Fast DNA kit method, PCR (10 µL volume) included 5.75 µL of dH$_2$O, 1 µl of 10× buffer, 0.2 µL of 10 mM dNTP mixture, 0.05 µL of *Taq* DNA polymerase (1 U/µL), 1 µL of 10 µg/µL BSA, 0.5 µl of 10 mM forward primer, 0.5 µL of 10 mM reverse primer and 1 µL of DNA template (4 ng/µL). Samples were amplified in MJ Research (Watertown, MA, USA) tetrads. The program consisted of 94°C for 4 min, 94°C for 30 s, the annealing temperature (46°C, 48°C, or 51°C) for 1 min and 72°C for 1 min. Steps 2 through 4 were repeated 32 times followed by 72°C for 5 min then a 4°C incubation. PCR products were prepared for either the ABI3100 or ABI3730 capillary sequencers (Applied Biosystems, Inc. (ABI), Foster City, CA, USA) using 2 µL of PCR product, 20 µL of dH$_2$O and 0.2 µl of ROX500 or ROX400HD. Samples were denatured for 30 s at 95°C then put on ice. Samples were analyzed using GeneMapper software version 3.0 or version 3.5 (ABI, Foster City, CA, USA).

For DNA isolated by the Epicentre method, PCR (50 µL volume) included 32.75 µL of dH$_2$O, 5 µL of 10× buffer, 1 µL of 10 mM dNTP mixture, 0.25 µL of *Taq* DNA polymerase (1 U/µl), 5 µl of 10 µg/µL BSA, 2.5 µL of 10 mM forward primer, 2.5 µL of 10 mM reverse primer and 1 µl of DNA template using the microsatellite primers (TcSHRS19) and the SSCP primers (WRKY11), but with the WRKY11 amplicons analyzed for fragment length only, not under non-denaturing SSCP conditions. PCR products preparation and analysis on the ABI3730 was as described above.

SSCP primers

Development and use of the SSCP primers in this study were described (Kuhn et al. 2008). We tested 15 loci: CF972913, EST2545 (CA795514), EST 4445 (CA797355), EST4572 (CA797491), EST4785 (CA797698), EST5371 (CA798090), EST5446 (CA798160), RGH1S (AF402715), RGH2S (AF402738), WRKY2 (AY331153), WRKY3 (AY331157), WRKY4 (AY331159), WRKY8 (AY331165), WRKY11 (AY331171), and WRKY17 (EF173893). Primers for CF972913 and the EST loci were all designed from cDNA sequence. The primers for RGH and WRKY loci were all designed from genomic DNA sequence from the amplification of genomic DNA with degenerate primers (Kuhn et al. 2003; Borrone et al. 2004). The forward primer of the WRKY11 locus (AY331171) was designed to the intron sequence.

SSCP analysis

Amplified products were denatured at 95°C for 5 min, snap cooled and a 1:200 dilution of undenatured ROX 2500 molecular weight standards were added (Kuhn and Schnell 2005). Products were analyzed at 22°C and 28°C on the ABI 3100 and electropherograms analyzed using GeneScan and GenoTyper software (ABI, Foster City, CA, USA) as described in (Kuhn et al. 2008). Undenatured molecular weight standards allowed alignment of runs at 22°C and 28°C. Tables of allele mobilities were exported from GenoTyper into Excel for analysis of population genetic parameters and genetic distance estimation.

Population parameters and genetic distance estimation

T. grandiflorum individuals analyzed were subsets of three different populations, but the populations were close genetically, and preliminary analysis of the individual populations or the grouped populations gave essentially the same results (data not shown). Thus, we analyzed all individuals as one population. Estimation of population genetics parameters and Hardy Weinberg equilibrium (HWE) was done using GenAlEx (Peakall and Smouse 2006).

TaqMan SNP assays

TaqMan SNP assays were designed for WRKY 3, WRKY 17, and EST0050 loci in *T. cacao* using the Primer Express 3.0 software (ABI, Foster City, CA, USA) to generate the probe sequences. The WRKY 3, WRKY 17, and EST0050 SNPs had been identified from direct sequencing of amplicons from genetically diverse cultivars representing all ten genetic groups resolved in (Motamayor et al. 2008). Amplicons were sequenced on the ABI3730 using big dye terminators and sequences were analyzed with Phred (Ewing et al. 1998), Phrap (Ewing and Green 1998), Consed (Gordon et al. 1998), and Polyphred (Stephens et al. 2006) to align and identify SNPs.

SNP analysis was performed using a 5′ nuclease (Taqman) assay (Holland et al. 1991; Livak 1999). In brief, a total reaction volume of 25 µL included 2.25 µL of each primer (10 mM), 0.5 µL of 6-carboxyfluorescein (FAM)-labeled probe (10 mM), 0.5 µL of VIC-labeled probe (10 mM), 6 µL water, 12.5 µL Taqman Genotyping Master Mix with ROX (2×; ABI, Foster City, CA, USA), and 1 µL of template DNA (4 ng/µl). Allelic determination was performed by first recording background fluorescence using a BioTek FLx 800 microplate fluorescence reader (BioTek Instruments, Winooski, VT, USA) with 485/528 nm excitation/emission wavelength filters for FAM-labeled probes and 540/575 nm excitation/emission wavelength filters for VIC-labeled probes. Then, PCR amplification was 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 microplate fluorescence reader. The VIC fluorescence values (relative fluorescence units) were plotted against the FAM fluorescence for each sample. Samples homozygous for the VIC-labeled allele will cluster along the *x*-axis, while those homozygous for the FAM-labeled allele cluster along the *y*-axis. Heterozygous samples cluster between the two. Allelic data for each SNP marker was collected in Excel and genotypes from each individual were identified.

Allele sequencing

Unlabeled primers designed to amplify the intron portion of WRKY8 and WRKY11 (Borrone et al. 2004, 2007) were used to amplify genomic DNA from 12 genetically diverse cacao cultivars and 12 selected *T. grandiflorum* trees that were diverse by microsatellite and SSCP analysis (TGR0023, 24, 30, 31, 32, 37, 38, 40, 46, 52, 53, and 56). Amplified DNA was treated with exonuclease to remove PCR primers, ethanol precipitated and direct sequenced with BigDye dideoxyterminators on the ABI3730. ABI electropherogram files were analyzed with Phred, Phrap, Consed and Polyphred as above to locate SNPs. SNP genotypes of *T. grandiflorum* for the WRKY8 and WRKY11 loci were compared with SSCP electromorph mobility data. A third locus (TIR1a; Kuhn et al. 2006) was sequenced for the same cacao and *T. grandiflorum* individuals described above, although no corresponding SSCP analysis was conducted. TIR1a is a representative of the NBS/LRR resistance gene homologues and the region amplified does not contain an intron, unlike the two WRKY loci. TIR1a was included to determine the frequency of SNPs in a coding region in both cacao and *T. grandiflorum*.

Results

Genotyping of *T. grandiflorum* with microsatellites and SSCP markers

Microsatellites We tested 30 microsatellite markers that had been developed for *T. cacao* in *T. grandiflorum*. All but four of the primer pairs amplified *T. grandiflorum* DNA (87%) and gave fragments similar in size to those in cacao. Ten of those loci (SHRSTc18, 19, 24, 28, 29, 32, 37, 38, 39, 43) were monomorphic across all *T. grandiflorum* individuals tested. Data for the 16 polymorphic markers are listed in Table 1. Observed heterozygosity for these loci ranged from 0.056 to 1 (mean ± SD=0.513±0.236). The number of alleles ranged from two to 12 (SHRSTc34) (mean ± SD=4.267±3.035).

Ten of the 16 loci were not in HWE (Table 1). Of loci with only two alleles that were not in HWE (SHRSTc5, 25, 27, 36), three (SHRSTc5, 27, 36) had much higher observed heterozygosity than expected. The SHRSTc27 locus had only two alleles and every individual analyzed was heterozygous, suggesting that in *T. grandiflorum*, these primers may amplify a duplicated locus. For SHRSTc25, heterozygotes were infrequently observed, which may be due to a null allele in *T. grandiflorum.*

SSCP markers We tested 15 cacao SSCP markers in *T. grandiflorum*; all of them amplified *T. grandiflorum* DNA. We selected 12 of these because three loci (CF972913, EST4445, and EST4785) gave fragments too large for SSCP analysis. After SSCP analysis, the EST5446 locus gave more than two electromorphs per individual, suggesting that the primers had amplified more than one locus in *T. grandiflorum*; this marker was not analyzed further. WRKY4 and WRKY8 were monomorphic across all *T. grandiflorum* samples, which were very different from *T. cacao*, where WRKY4 had two alleles and WRKY8 had four alleles (Kuhn et al. 2008). The marker data for the nine polymorphic loci are listed in Table 2. Observed heterozygosity for these markers ranged from 0.088 to 0.645 (mean ± SD=0.246±0.059). The number of alleles per locus ranged from two to five (mean ± SD=3.000± 0.333). Two of the nine loci were not in Hardy-Weinberg equilibrium (Table 2).

The microsatellite markers had a greater number of alleles (mean=4.3 alleles per locus) and effective alleles (mean=2.2 effective alleles per locus) per locus than the SSCP markers (mean=3.0 alleles per locus, mean=1.7 effective alleles per locus) (Tables 1 and 2). For both the microsatellite and SSCP markers, there were loci that were

Table 1 Microsatellite marker (locus), sample size (*N*), no. alleles (Na), no. effective alleles (Ne), information index (I), observed heterozygosity (Ho), expected (He) and unbiased expected heterozygosity (UHe), and fixation index (F), chi-square tests for Hardy-Weinberg equilibrium

Locus	*N*	Na	Ne	I	Ho	He	UHe	F	DF	ChiSq	Prob	Significance
SHRSTc2	35	3.000	2.114	0.803	0.486	0.527	0.535	0.078	3	35.000	0.000	$P<0.001$
SHRSTc3	35	7.000	1.890	1.054	0.343	0.471	0.478	0.272	21	41.807	0.004	$P<0.01$
SHRSTc4	36	7.000	3.277	1.384	0.417	0.695	0.705	0.400	21	40.377	0.007	$P<0.01$
SHRSTc5	36	2.000	1.946	0.679	0.667	0.486	0.493	−0.371	1	4.967	0.026	$P<0.05$
SHRSTc6	36	7.000	2.642	1.170	0.528	0.622	0.630	0.151	21	86.148	0.000	$P<0.001$
SHRSTc7	36	4.000	2.054	0.806	0.667	0.513	0.520	−0.299	6	76.513	0.000	$P<0.001$
SHRSTc11	36	2.000	1.246	0.349	0.167	0.198	0.200	0.156	1	0.879	0.349	ns
SHRSTc13	33	2.000	1.800	0.637	0.545	0.444	0.451	−0.227	1	1.705	0.192	ns
SHRSTc23	35	5.000	4.529	1.559	0.543	0.779	0.790	0.303	10	15.391	0.118	ns
SHRSTc25	36	2.000	1.246	0.349	0.056	0.198	0.200	0.719	1	18.598	0.000	$P<0.001$
SHRSTc27	35	2.000	2.000	0.693	1.000	0.500	0.507	−1.000	1	35.000	0.000	$P<0.001$
SHRSTc31	33	3.000	1.412	0.565	0.333	0.292	0.296	−0.143	3	1.320	0.724	ns
SHRSTc33	35	2.000	1.849	0.652	0.486	0.459	0.466	−0.058	1	0.117	0.732	ns
SHRSTc34	35	12.000	5.235	1.967	0.743	0.809	0.821	0.082	66	139.635	0.000	$P<0.001$
SHRSTc36	34	2.000	1.841	0.649	0.706	0.457	0.464	−0.545	1	10.116	0.001	$P<0.01$
SHRSTc44	36	7.000	3.064	1.430	0.556	0.674	0.683	0.175	21	32.388	0.053	ns
Mean and SE over loci												
Mean	35.125	4.313	2.384	0.922	0.515	0.508	0.515	−0.019				
SE	0.256	0.734	0.283	0.115	0.057	0.045	0.046	0.103				

Table 2 SSCP marker (locus), sample size (*N*), no. alleles (Na), no. effective alleles (Ne), information index (I), observed heterozygosity (Ho), expected (He) and unbiased expected heterozygosity (UHe), and fixation index (F), chi-square tests for Hardy-Weinberg equilibrium

Locus	*N*	Na	Ne	I	Ho	He	UHe	F	DF	ChiSq	Prob	Significance
2545_1	36	3.000	1.866	0.774	0.278	0.464	0.471	0.401	3	12.099	0.007	*P*<0.01
5371_1	31	3.000	1.103	0.225	0.097	0.093	0.095	−0.039	3	0.080	0.994	ns
4572_1	31	4.000	2.707	1.095	0.323	0.631	0.641	0.488	6	29.741	0.000	*P*<0.001
RGH1s_1	34	2.000	1.092	0.181	0.088	0.084	0.086	−0.046	1	0.072	0.788	ns
RGH2s_1	33	2.000	1.502	0.517	0.242	0.334	0.339	0.275	1	2.491	0.115	ns
WRKY2_1	30	2.000	1.105	0.199	0.100	0.095	0.097	−0.053	1	0.083	0.773	ns
WRKY3_1	25	3.000	1.334	0.501	0.280	0.250	0.256	−0.118	3	0.663	0.882	ns
WRKY11_1	26	5.000	3.539	1.367	0.654	0.717	0.732	0.089	10	8.358	0.594	ns
WRKY17_1	33	3.000	1.167	0.319	0.152	0.143	0.145	−0.061	3	0.222	0.974	ns
Mean and SE over loci												
Mean	31.000	3.000	1.713	0.575	0.246	0.312	0.318	0.104				
SE	1.202	0.333	0.288	0.141	0.059	0.081	0.082	0.075				

not in Hardy-Weinberg equilibrium, as expected for a study involving relatively few individuals (*N*=25-36, depending on locus). None of the *T. grandiflorum* individuals was less than 12.5% heterozygous (87.5% homozygous) by the combined microsatellite and SSCP data (Tables 1 and 2). The average heterozygosity was 29% (71% homozygosity) when all microsatellite and SSCP marker loci were included.

SNP markers

When the four TaqMan SNP markers developed for *T. cacao* from three loci (WRKY3 (two SNPs in intron), WRKY17 (one SNP in intron) and EST0050 (one SNP in exon)) were tested against all *T. grandiflorum* individuals, all were monomorphic (Table 3), even though the SSCP results for WRKY3 and 17 had three alleles at each locus in *T. grandiflorum* (Table 2).

Sequencing of *T. cacao* and *T. grandiflorum* amplicons

Primers for three loci (WRKY8, WRKY11, and TIR1a) were used to amplify the DNA of 12 genetically diverse individuals of both *T. cacao* and *T. grandiflorum* and the resulting amplicons were direct sequenced. When the WRKY8 locus in *T. grandiflorum* was sequenced, no sequence differences were observed and no mobility differences were seen for electromorphs in the SSCP assay (data not shown). All *T. grandiflorum* electromorphs for WRKY8 were of identical mobility and none shared the mobility of any *T. cacao* electromorph (data not shown), which is consistent with the inter-specific sequence differences in Table 4. Comparing the WRKY8 locus (368 nt total, 113 exon+255 intron) sequences between *T. cacao* and *T. grandiflorum*, there were four cacao intra-specific SNPs, all in the intron region (one SNP per 64 nt) and 11 inter-specific SNPs, all in the intron region (one SNP per 23 nt).

Table 3 Primer and probe sequences for cacao SNP markers for WRKY3, WRKY17, and EST0050

SNP locus	Pos	Primer seq. (5′-3′)	Probe seq. (5′-3′)	Fl label	Nuc	No. all	Gen
WRKY3	s41	F:AAAGGCAATCCTTACCCAAGGT	ATGCCCCTGgTTGT	FAM	G	1	GG
		R:AAGAATGAACCACTTTGCAGTAGATAGT	ATGCCCCTGtTTGT	VIC	T		
	s558	F:GTTGTTGTTCTGTTCAATTCGTATGA	TGACTaCCTTTTATGTGATCT	FAM	A	1	AA
		R:ATCAGGAATGCTCCAAAATAATCAA	TGACTgCCTTTTATGTGAT	VIC	G		
WRKY17	s189	F:TGATTACACTGTTACACCAACTTTAGACG	TCTTGCtGAGATATC	VIC	T	1	CC
		R:ACGTGTAAAGAAAGGAGGAAAACTTT	TCTCTTGCcGAGATAT	FAM	C		
EST0050	s274	F:CTCAGGTTCCAACCATTGATTTAA	AAGCTGCCAcGGAGT	FAM	C	1	TT
		R:CCGAGATCCCATGGTTAACAA	AAGCTGCCAtGGAGT	VIC	T		

Pos position of SNP in relation to the reference sequence, *Fl label* fluorescent label, *nuc* nucleotide, *No. all* number of alleles, *Gen* genotype of all *T. grandiflorum* samples

Table 4 Position of intra- and inter-specific SNPs from intron regions for WRKY8 and WRKY11

	WRKY8													
	100	119	157	175	201	204	235	261	264	266	267	269	288	318
T. cacao	A	C/T	A	A	T	C/T	C	T	A	A	A	A	A/G	A
T. grandiflorum	G	C	G	G	C	T	G	C	G	G	G	G	G	G
	WRKY11													
	847	867	892	898	905	921	932	937	973	978	987			
T. cacao	G	C/T	C	T	A	T	A	C	G	G	T			
T. grandiflorum	A/G	T	T	G	G	C	G	T	A	C	C/T			

In the WRKY11 locus, we sequenced a 270 nt portion that began in the second intron (216 nt) and ended 54 nt into the third exon of the reference sequence. There was one cacao intra-specific SNP, two *T. grandiflorum* intra-specific SNPs, eight inter-specific SNPs in the intron region (one inter-specific SNP per 42 nt) and no SNPs in the exon (Table 4).

For WRKY11, both cacao and *T. grandiflorum* had SNPs that distinguished individual trees of each species (intra-specific SNPs; Table 4). The SSCP electromorph mobilities also differed among individuals (data not shown) and thus were consistent with these sequences. Since amplicons were direct sequenced, rather than first cloned and then sequenced, it was not possible to assign haplotypes to individual SSCP electromorphs, but all sequence differences were captured as mobility differences in the SSCP analysis of WRKY11. In addition, sequencing identified an 11 nt insertion in the intron region of WRKY11 in *T. grandiflorum* that made the amplicons of WRKY11 in cacao and *T. grandiflorum* distinguishable by their length without need for further SSCP analysis.

Analysis of TIR1a by SSCP for cacao demonstrated that it was polymorphic (Kuhn et al. 2008). However, direct sequencing of the cacao and *T. grandiflorum* TIR1a amplicon disclosed numerous nucleotide positions that were heterozygous in all individuals. This may be the result of the amplification of more than one locus, as was noted for TIR1a when mapping was attempted in cacao (Kuhn et al. 2006). One interspecific SNP was identified for TIR1a (one SNP per 250 nt; data not shown) with the polymorphism in the third codon position and no change in the amino acid sequence.

Analysis of putative inter-specific hybrids

Putative inter-specific hybrids were created by all possible crosses of ten *T. cacao* cultivars as the maternal parent with six *T. grandiflorum* cultivars as the paternal parent. We received leaf disk material from 77 putative hybrids, a subset of the total progeny of all crosses, and from the parents and genotyped them using two marker loci, SHRSTc19 and WRKY11, which are ~0.9 cM apart on the genetic map (Brown et al. 2008). We chose SHRSTc19 because it was monomorphic in *T. grandiflorum* (allele size=173) and did not share any alleles with *T. cacao* (allele sizes≥178). We chose WRKY11 because all *T. grandiflorum* samples had an insertion of 11 nucleotides in the intron (size=282) when compared to *T. cacao* (size=271). Because we had limited amounts of leaf material, we used the Epicentre DNA isolation method, which worked well providing strong signal for the amplicon peaks on the ABI3730 and very little missing data. Genotype data for the parents and putative hybrids are shown in Table 5. For the WRKY11 marker, all *T. cacao* parents were homozygous for the 271 allele; all *T. grandiflorum* parents were homozygous for the 282 allele. For TcSHRS19, there were three genotypes observed in the *T.cacao* parents (188, 188; 180, 188; 178, 178) while all *T. grandiflorum* parents were homozygous for the 173 allele. Of the 77 putative hybrid seedlings, 48 were identified as hybrids by our markers and 29 were identified as *T. cacao*, the maternal parent. In two hybrids, results for WRKY11 were ambiguous because of poor amplification, but results for SHRSTc19 were strong, so the genotype was called by only one marker. For all other genotypes, both markers were congruent.

Discussion

Marker development and identification of *T. cacao* and *T. grandiflorum* inter-specific hybrids

Our evaluation in *T. grandiflorum* of three types of PCR-based genetic markers that had initially been developed in cacao (microsatellites, SSCP markers, and SNPs) showed that sequence conservation between *T. grandiflorum* and *T. cacao* is high, which was reflected in the successful amplification of almost all the cacao markers in *T. grandiflorum*. Twenty six of the 30 microsatellite loci (87%) amplified, as compared to 60% amplification success in a previous survey of cacao microsatellite markers in *T. grandiflorum* (Alves et al. 2007). Failure to

Table 5 Genotype results for cacao and *T. grandiflorum* parents and putative inter-specific hybrids assayed with markers SHRSTc19 and WRKY11

			SHRSTc19		WRKY11	
Sample type	ID	No.	All 1	All 2	All 1	All 2
Cacao maternal parent	Cacao	7	188	188	271	271
Cacao maternal parent	Cacao	1	178	178	271	271
Cacao maternal parent	Cacao	2	180	188	271	271
T. gr Paternal parent	T gr	6	173	173	282	282
Putative hybrid	Hybrid	45	173	188	271	282
Putative hybrid	Hybrid	2	173	178	271	282
Putative hybrid	Hybrid	1	173	180	271	282
Putative hybrid	Cacao	2	178	178	271	271
Putative hybrid	Cacao	29	188	188	271	271

T. gr, *T. grandiflorum*; *ID* determined genotype, *No.* number of samples of each category, *All* allele

amplify is not surprising as the majority of microsatellite loci are in non-coding regions where DNA sequence is much less conserved, which may explain the greater numbers of alleles per locus. All of the SSCP loci amplified, probably because the primers were designed to coding region sequences where sequence is more conserved and which usually have fewer alleles per locus because of sequence conservation. The microsatellite markers and SSCP markers were polymorphic in the population of *T. grandiflorum* individuals we studied, whereas the cacao SNP markers tested were not. Our sequencing of three loci (WRKY8, WRKY11, and TIR1a) did not identify any nucleotide positions that were intra-specific SNPs for both *T. cacao* and *T. grandiflorum*. The *T. grandiflorum* individuals we genotyped were not highly diverse, as measured by any of the markers. However, there also were no *T. grandiflorum* individuals that were as highly homozygous as the Matina 1-6 cacao cultivar (98% homozygous) that was chosen for the cacao genome sequencing. In general, in *T. grandiflorum* microsatellite markers had more alleles than SSCP markers. Microsatellite markers should thus be more informative for studying *T. grandiflorum* populations or for searching for highly homozygous individuals to use for sequencing the *T. grandiflorum* genome.

We compared allele profiles for six of the microsatellite loci (SHRSTc3, 4, 5, 7, 11, and 19) for which both *T. grandiflorum* and *T. cacao* data were available (Motamayor et al. 2008). We found that four of the loci (SHRSTc 3, 5, 7, 11) shared alleles across species, and one (SHRSTc4, an AG dinucleotide repeat) had alleles in *T. cacao* that were very close to those in *T. grandiflorum* (e.g., 126 bp in *T. grandiflorum*, 125 and 127 bp in *T. cacao*). Only SHRSTc19, which was monomorphic (173 bp allele) in the 42 *T. grandiflorum* individuals of this study, shared no alleles with *T. cacao*. For SHRSTc19, *T. cacao*'s closest allele was at 178 bp.

For SSCP markers, there was no overlap of alleles between *T. cacao* and *T. grandiflorum*: electromorphs from cacao and *T. grandiflorum* were always distinguishable, even when they were monomorphic in both species. However, SSCP is not a widely used technique and requires specialized polymers for high-throughput analysis on the ABI3100; analysis time of SSCP markers is greater than for microsatellites, and genotyping data cannot be readily shared between laboratories. At one time, an efficient use of SSCP was for initial screening of amplicons to determine heterozygosity, followed by sequencing to identify intra and inter-specific SNPs, but as it is now possible to rapidly sequence entire genotypes, the role of SSCP will be limited in SNP discovery.

The most useful marker identified in this study for analyzing putative inter-specific hybrids was WRKY11, because the amplicons of all of the individuals of the two species analyzed in this study differed by 11 nucleotides in length and thus would not require SSCP analysis. This size difference should be found in any putative hybrid analyzed in this study, regardless of individuals used in the cross. The most useful microsatellite marker we found was SHRSTc19, because it had no allele overlap between the individuals of the two species analyzed in this study. For any other microsatellite marker, the parents of the cross would also have to be tested to determine that no allele overlap occurred, and markers would have to be excluded if there was significant allele overlap.

We used SHRSTc19 and WRKY11 to analyze some putative inter-specific hybrids between *T. cacao* and *T. grandiflorum*. Previously, when such hybrids have been created, they did not grow beyond the seedling stage (15 cm tall; Silva et al. 2004). The hybrids we genotyped were also seedlings, so very little leaf material was available to test. We took advantage of a rapid one-step DNA isolation method that can be done from a single leaf disk and were able to genotype all of the parents and putative hybrids by this method. By chance, the markers that we used were only 0.9 cM apart on our genetic map, so markers that covered all the linkage groups would be preferable to fully analyze the putative inter-specific hybrids.

We identified interspecific SNPs for WRKY8, WRKY11 (Table 4), and TIR1a that allow us to distinguish the contribution of each species in an inter-specific hybrid. Interspecific SNPs are defined here as all cacao individuals homozygous for one nucleotide and all *T. grandiflorum* individuals homozygous for a different nucleotide at the inter-specific SNP position. These SNPs provide completely informative sites for the analysis of inter-specific hybrids. In all loci studied, which included both intron and exon regions, there were distinct species-specific sequence differences, with both inter- and intra-specific SNPs occurring more frequently in intron regions.

Marker development and *T. cacao* and *T. grandiflorum* comparative genomics

For *T. grandiflorum* to be useful in a comparative genomics program with *T. cacao*, it would be best if the majority of inter-specific SNPs that occurred in exons did not lead to amino acid changes. If this were the case, it would be easier to identify candidate genes in either *T. cacao* or *T. grandiflorum* where amino acid changes occurred and to investigate their effect on phenotype. For example, in anthocyanin biosynthesis in developing seeds, we could compare RNA sequences from *T. cacao* and *T. grandiflorum* and identify which of the expressed genes involved in anthocyanin synthesis differed in their amino acid sequence. Such an expression study would also determine if differential expression of the candidate genes played a role in the phenotypic differences.

In our WRKY sequence data, species-specific sequence differences were found, but they were solely in the intron regions of WRKY8 and WRKY11. Cacao and *T. grandiflorum* reference sequences that had more than 50 nt of exon sequence were also available for WRKY11, 12, and 13 (Borrone et al. 2007); we analyzed these for the occurrence of SNPs in the exons. By aligning the cacao and *T. grandiflorum* reference sequences for WRKY11, seven interspecific SNPs were detected in the coding region (622 nt, one SNP per 89 nt of exon) with three occurring in the third codon position and causing no change in the amino acid, and four occurring in either the first or second codon position and causing a change in the amino acid (T→A, F→V, L→S, D→N, cacao→*T. grandiflorum*). Of these four changes, only F→V was conservative. Aligned reference sequences for WRKY12 showed four inter-specific SNPs in an exon region of 421 nt (one SNP per 105 nt) that produced no amino acid changes. Aligned reference sequences for WRKY13 had six inter-specific SNPs in an exon region of 589 nt (one SNP per 98 nt) with two SNPs leading to no amino acid change, three SNPs to a conservative amino acid change (V→A, I→V, D→E) and only one non-conservative change (S→P). Direct comparison of the *T. cacao* genome to the *T. grandiflorum* genome in association with phenotypic trait data may provide an indication of which amino acid changes are important for altered activity of enzymes. In addition, sufficient inter-specific SNPs were found in exons so that expression studies of the inter-specific hybrids will allow determination of differential expression of the alleles of both species.

Summary and conclusions

We have demonstrated that SSCP markers developed for cacao and used to analyze *T. grandiflorum* are more sensitive than cacao microsatellites in identifying species-specific differences, although they are not widely used due to technical constraints. Microsatellites developed for cacao will be more useful in identifying highly homozygous *T. grandiflorum* individuals for genome sequencing. We have also observed that SNP markers designed for *T. cacao* were not useful in estimating diversity in *T. grandiflorum*. Finally, we have analyzed several loci by sequencing amplified fragments from individuals of *T. grandiflorum* and *T. cacao* and identified both intra-specific SNPs that can be used to assess heterozygosity in *T. grandiflorum* and species-specific SNPs that will aid in characterizing inter-specific hybrids. Sequence conservation was significant and species-specific differences numerous enough to suggest that comparative genomics of *T. grandiflorum* and *T. cacao* will be useful in elucidating the genetic differences that lead to a variety of important agronomic trait differences.

Acknowledgments This research was funded by USDA-ARS and Mars, Inc.

References

Alves RM, Sebbenn AM, Artero AS, Clement C, Figueira A (2007) High levels of genetic divergence and inbreeding in populations of cupuassu (*Theobroma grandiflorum*). Tree Geneti Genom 3:289–298

Borrone JW, Kuhn DN, Schnell RJ (2004) Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. Theor Appl Genet 109:495–507

Borrone JW, Meerow AW, Kuhn DN, Whitlock BA, Schnell RJ (2007) The potential of the WRKY gene family for phylogenetic reconstruction: an example from the Malvaceae. Mol Phylogenet Evol 44:1141–1154

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 Bio 1:120–130

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

Figueira A, Janick J, Goldsbrough P (1992) Genome size and DNA polymorphism in *Theobroma cacao*. J Am Soc Hortic Sci 117:673–677

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Holland PM, Abramson RD, Watson R, Gelfand DH (1991) Detection of specific polymerase chain reaction product by utilizing the 5′—3′ exonuclease activity of *Thermus aquaticus* DNA polymerase. Proc Natl Acad Sci USA 88:7276–1180

Kuhn DN, Schnell RJ (2005) Use of capillary array electrophoresis single-strand conformational polymorphism analysis to estimate genetic diversity of candidate genes in germplasm collections. Meth Enzymol 395:238–258

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, Narasimhan G, Nakamura K, Brown JS, Schnell RJ, Meerow AW (2006) Identification of cacao TIR-NBS-LRR resistance gene homologues and their use as genetic markers. J Am Soc Hortic Sci 131:806–813

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 29:4096–4108

Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Gen Anal: Biomol Eng 14:143–149

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. HortScience 44:1228–1232

Motamayor JC, Lachenaud P, da Silva EMJW, 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

Peakall R, Smouse PE (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Mol Ecol Notes 6:288–295

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Schnell RJ, Olano CT, Brown JS, Meerow AW, Cervantes-Martinez C, Nagai C, Motamayor JC (2005) Retrospective determination of the parental population of superior cacao (*Theobroma cacao* L.) seedlings and association of microsatellite alleles with productivity. J Am Soc Hortic Sci 130:181–190

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

Silva CRS, Figueira A (2005) Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. Plant Syst Evol 250:93–104. doi:10.1007/s00606-004-0223-2

Silva C, Venturieri GA, Figueira A (2004) Description of Amazonian *Theobroma* L. collections, species identification, and characterization of interspecific hybrids. Acta Bot Bras 18:333–341

Stephens M, Sloan JS, Robertson PD, Scheet P, Nickerson DA (2006) Automating sequence-based detection and genotyping of SNPs from diploid samples. Nat Genet 38:375–381

Whitlock BA, Baum DA (1999) Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene Vicilin. Syst Bot 24:128–138