Theor Appl Genet (2004) 109: 495–507
DOI 10.1007/s00122-004-1662-4

ORIGINAL PAPER

James W. Borrone · David N. Kuhn · Raymond J. Schnell

Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*

Received: 26 November 2002 / Accepted: 8 March 2004 / Published online: 18 May 2004

Abstract There is currently an international effort in improving disease resistance and crop yield in *Theobroma cacao* L., an economically important crop of the tropics, using marker-assisted selection for breeding. We are developing molecular genetic markers focusing upon gene families involved with disease resistance. One such family is the WRKY proteins, which are plant-specific transcriptional factors associated with regulating defense responses to both abiotic and biotic stresses. Degenerate PCR primers were designed to the highly conserved DNA-binding domain and other conserved motifs of group I and group II, subgroups a–c, WRKY genes. Sixteen individual WRKY fragments were isolated from a mixture of *T. cacao* DNA using one pair of primers. Of the 16 WRKY loci investigated, seven contained single nucleotide polymorphisms within the intron as detected by sequence comparison of the PCR products. Four of these were successfully converted into molecular markers and mapped in an F_2 population by capillary electrophoresis-single strand conformation polymorphism analysis. This is the first report of a pair of degenerate primers amplifying WRKY loci directly from genomic DNA and demonstrates a simple method for developing useful genetic markers from members of a large gene family.

Communicated by D.B. Neale

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00122-004-1662-4

R. J. Schnell (✉)
USDA-ARS, Subtropical Research Station,
13601 Old Cutler Rd,
Miami, FL, 33158, USA
e-mail: rschnell@ars-grin.gov
Tel.: +1-305-2543611
Fax: +1-305-9696410

J. W. Borrone · D. N. Kuhn
Department of Biological Sciences, Florida International University,
Miami, FL, 33199, USA

Introduction

Theobroma cacao L. is a small understory tree indigenous to the lowland tropical evergreen forests of South America that is cultivated throughout the tropics as an important cash crop in Africa, Southeast Asia, and South and Central America. It is the sole source of cocoa, the raw material from which chocolate is produced, as well as of cocoa butter, which is used extensively by the cosmetic industry. West Africa accounts for 70% of the world's annual production of cocoa beans with Southeast Asia and South and Central America supplying the remaining 30% (Chocolate Manufacturers Association http://www.chocolateandcocoa.org). The vast majority of cacao cultivation (estimated at 80%) occurs primarily on smallholder farms (Duguma et al. 2001).

Crop yields of *T. cacao*are limited by susceptibility to numerous pathogens (Gotsch 1997). Two diseases have devastated South and Central American production: witches' broom and frosty pod rot caused by the basidiomycetes *Crinipellis perniciosa* (Stahl) and *Moniliophthora roreri* (Cif. and Par.), respectively. Currently confined to the Americas, these diseases could threaten the world's supply of cocoa if introduced to other cacao-growing regions. Therefore, an international effort is underway to improve *T. cacao*'s disease resistance while maintaining other desirable quantitative and qualitative crop characteristics (Crouzillat et al. 2000; Marita et al. 2001).

Breeding programs for *T. cacao* have been hampered by factors common to fruit tree crops: long generation times, long initial establishment periods before fruit production, long periods before maximal fruit production is achieved, and a requirement for large planting areas. Yield trials take years to complete (Irizarry and Rivera 2002), thus traditional breeding is extremely costly and time-intensive, requiring decades to evaluate promising candidates (Toxopeus 1969). Genetic variability exists, but most breeding work has utilized only a narrow genetic base of *T. cacao* (Motamayor et al. 2000). Very little superior planting material has been made available for farmers.

The most commonly planted cultivars are mixed hybrid varieties reproduced in seed gardens, and their pedigree and genotype are often uncertain or unknown (Hunter 1990).

In 1998, the United States Department of Agriculture-Agriculture Research Service (USDA-ARS) revived a cacao genetics program based at the Subtropical Horticulture Research Station (SHRS) in Miami, Florida. The primary goal of the program is to develop a biotechnology-based approach to improve *T. cacao* for increased disease resistance, especially to witches' broom and frosty pod rot. One aim is to use marker-assisted selection (MAS) to reduce the time and cost of evaluating breeding material. Along with developing simple sequence repeats (SSRs), our laboratory has been developing molecular markers from candidate genes—genes potentially involved in the expression of desirable quantitative traits (Deng and Davis 2001; Thorup et al. 2000). Candidate gene markers provide powerful tools for evaluating plant material, thereby shortening the time required to produce improved varieties (Lamb et al. 1992). They have also been used successfully to identify loci controlling quantitative traits such as disease resistance (Byrne et al. 1996; Faris et al. 1999). Our laboratory has focused on two gene families involved in disease resistance: disease resistance gene homologs (RGHs), which are responsible for the specific recognition of pathogenic organisms (Kuhn et al. 2003), and WRKY genes, which are responsible for the regulation of a plant's response to abiotic and biotic stresses.

WRKY genes encode a family of plant-specific transcriptional factors regulating physiological responses to biotic and abiotic stresses and senescence as well as trichome development, seed development, and the biosynthesis of secondary metabolites (Alexandrova and Conger 2002; Dellagi et al. 2000; Deslandes et al. 2002; Gus-Mayer et al. 1998; Hara et al. 2000; Huang and Duman 2002; Kim et al. 2000; de Pater et al. 1996; Pnueli et al. 2002; Schenk et al. 2000; Yoda et al. 2002). WRKY proteins contain a 60-amino acid domain defined by a signature WRKYGQK motif and a distinctive zinc-finger-like motif that, in all of the WRKY proteins characterized to date, binds DNA at a sequence-specific site. All WRKY proteins contain either one or two copies of this highly conserved domain. WRKY proteins are classified into three main groups (I, II, and III), and group II is further subdivided into five subgroups (a–e) based on conserved amino acid motifs and the number of domains present in a protein. Up to 74 WRKY genes have been described in *Arabidopsis thaliana* to date (Eulgem et al. 2000).

A common element of WRKY genes is the interruption of the coding region of the C-terminal WRKY domain of group I and the single WRKY domain of groups II and III genes by an intron. When present, the position of the intron is highly conserved and aids in identifying the group/subgroup to which each gene belongs (Dong et al. 2003; Eulgem et al. 2000). Of the 74 WRKY genes described in *A. thaliana*, each exists at a single locus, and these are evenly distributed throughout the genome. Alleles of WRKY domain-containing genes have been described in the literature and are inherited and expressed co-dominantly. Alleles of AtWRKY52 are responsible for differences observed in disease susceptibility to bacterial wilt caused by *Ralstonia solanacearum* (Deslandes et al. 2002; Eulgem et al. 1999, 2000).

We hypothesized that: (1) conservation found within the coding regions of WRKY genes would allow for isolation of WRKY loci by degenerate primer PCR, (2) individual loci could be easily identified because of the high degree of conservation of the coding region coupled with extreme differences exhibited by introns, (3) alleles would be identifiable due to a relaxed evolutionary constraint within the intron, and (4) these alleles could be developed into molecular markers. This report describes the isolation of 16 *T. cacao* WRKY (TcWRKY) loci using a single pair of degenerate primers. Of the 16 TcWRKY loci isolated by degenerate PCR, seven were polymorphic by sequence comparison, and four of these were successfully converted into molecular markers and mapped within a *T. cacao* F_2population using capillary electrophoresis–single strand conformation polymorphism (CE-SSCP). Two additional TcWRKY loci were discovered to be polymorphic within the F_1 individual by sequencing but could not be mapped under the CE-SSCP conditions employed. This is the first report of a pair of degenerate primers amplifying WRKY domain-containing loci from genomic DNA and illustrates how members of a large gene family may be rapidly isolated and converted into molecular markers.

Materials and methods

Plant material

Flowers were collected and combined together from six individual trees of *Theobroma cacao* L. maintained as part of the National Germplasm Repository at the USDA-ARS, SHRS in Miami, Florida. Their accession numbers are: MIA14857, MIA15810, MIA15817, MIA29848, MIA30313, and MIA30530. For genetic mapping, leaf material from the parents [ICS1 (female) and Sca6 (male)], from an F_1individual (TSH516), and from 146 F_2 progeny (P1–P180) was obtained from CEPEC/CEPLAC, Itabuno, Brazil. The F_1individual, TSH516, is self-incompatible. Generation of the F_2 population required a mixture of TSH516 and *Herrania* sp. pollen, as mentor pollen, to overcome self-incompatibility (Ahnert 2000). Successful fertilizations resulted in an average of ten seeds per pod (U. Lopes, personal communication). The F_2 population is unrelated to the six trees used as the template for the degenerate primer PCR.

DNA isolation

For the degenerate primer PCR, DNA was isolated from 2 g fresh weight of flowers (collected from the six Miami accessions and combined together) by adding a standard sodium acetate/ethanol (NaOAc/EtOH) precipitation step (Sambrook and Russell 2001) to an established RNA isolation method (Ainsworth 1993). For the mapping population (ICS1, Sca6, TSH516, and the F_2 progeny, P1-P180) DNA was isolated from fully expanded leaves. Fresh leaf tissue (200 mg) was homogenized by disruption with beads in a FastPrep FP120 (BIO101, Savant), and DNA was isolated following the FastPrep kit procedure (Qbiogene, Carlsbad, Calif.).

The quantity and quality of the isolated DNA was assessed with a GeneQuant pro RNA/DNA calculator (Amersham Pharmacia Biotech, Piscataway, N.J.) and electrophoresis on 1% agarose gels. The DNA was stored at −80°C. Working stock solutions were made by diluting the DNA samples to 2.5 ng/μl and storing them at −20°C.

Degenerate primers

Degenerate primers were designed using GENEFISHER (Giegerich et al. 1996; http://bibiserv.techfak.uni-bielefeld.de/genefisher/). Twenty-eight WRKY amino acid sequences were retrieved from the Swiss-Prot protein knowledge-base and Translated European Molecular Biology Laboratory nucleotide database (SP-TrEMBL) and aligned using the web version of DBCLUSTAL (Thompson et al. 2000; http://igbmc.u-strasbg.fr:8080/DbClustal/dbclustal.html). The query sequence was NtWRKY-9 (SP-TrEMBL: Q94IB3, GenBank: BAB61056) (Maeo et al. 2001). Further details and the sequences used for the alignment can be found in the supplementary material (ESM). Two primers were selected from the generated list: WRKY 1 FP and WRKY 2 RP. Three additional primers (WRKY 2 FP, WRKY 3 FP, and WRKY 3 RP) were designed manually based on conserved regions located outside of the WRKY domain. Primer sequences, the corresponding amino acid regions, and specificities are given in Table 1. The potential priming sites are given in Fig. 1a. Degenerate primers were supplied by (Gibco-BRL, Life Technologies, Gaithersburg, Md.). Inosine was substituted in some positions to reduce the overall degeneracy of the primers.

Degenerate primer PCR amplification conditions

PCRs were conducted with models PTC-100 or PTC-225 thermocyclers (MJ Research, Waltham, Mass.). All PCRs contained 0.2 μ*M* forward primer, 0.2 μ*M* reverse primer, 200 μ*M* dNTPs, 1× PCR buffer with 1.5 m*M* $MgCl_2$, 2.5 U AmpliTaq polymerase (Applied Biosystems, Foster City, Calif.), and 2.5 ng of DNA. PCRs with the mapping population also contained 10 ng of bovine serum albumin per reaction. PCRs were electrophoresed on a 2.5% agarose gel, 0.5× TBE, and at 150 V constant voltage and visualized by staining with ethidium bromide.

A DNA mixture from the six Miami accessions was the template for the degenerate primer PCRs. All possible primer combinations were tested. Thermocycler conditions were: (1) an annealing temperature of 50°C; (2) an initial denaturation at 94°C, pausing at 80°C to add *Taq* polymerase, and subsequent amplification at an annealing temperature of 50°C (Hot-to-80); (3) annealing temperatures ranging from 40°C to 50°C (Gradient PCR). The cycle parameters were: an initial denaturation step of 94°C, 5 min; (94°C for 30 s; annealing for 45 s; 72°C for 2.5 min) × 39; 72°C, 20 min; 4°C hold. The actual annealing temperatures tested for the Gradient PCR were 41.7°C, 44.3°C, 47.4°C, and 49.8°C.

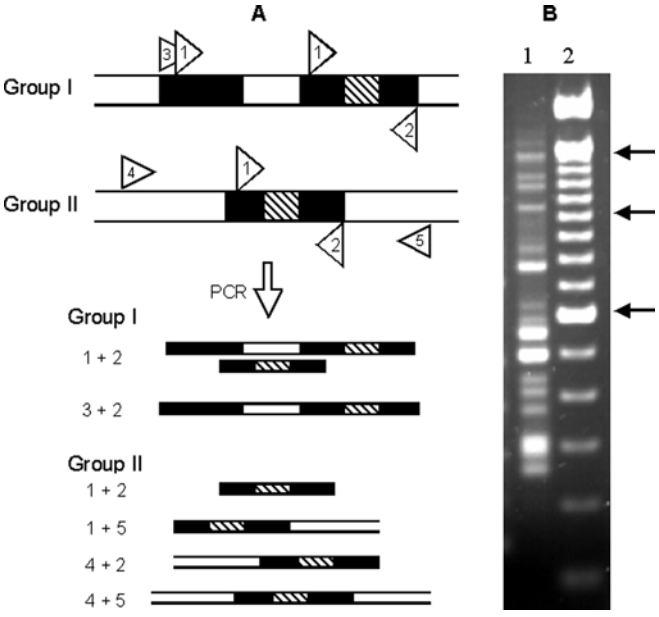

Fig. 1a, b PCR with degenerate primers on *Theobroma cacao* DNA. **a** A scheme depicting the organization of group 1 and group II, a–c, WRKY proteins and the expected results from PCR with each pairs of degenerate primers based only upon the potential binding sites for each primer. *Arrows* indicate the approximate location of potential binding sites for the degenerate primers. *Arrows at the top* indicate the forward primer binding sites; those *at the bottom*indicate reverse primer binding sites. *1*WRKY 1 FP, *2* WRKY 2 RP, *3* WRKY 2 FP, *4*WRKY 3 FP, *5* WRKY 3 RP. The WRKY domain is indicated by the *black box*, and the intron interrupting the WRKY domain is indicated by *diagonal lines*. The locations of other introns are not indicated. **b** PCR results of the Hot-to-80 Start for WRKY 1 FP + WRKY 2 RP (*Lane 1*). *Lane 2* 100-bp molecular ladder (Gibco-BRL). *Arrows* indicate, from *top* to *bottom*, the 1.5, 1.0, and 0.6 kb markers

Cloning and sequencing

PCR products were cloned using the TOPO TA Cloning kit for sequencing following the instructions of the manufacturer (Invitro-

Table 1 WRKY degenerate primers

Primer[a]	Sequence[b]	Deg[c]	Corresponding amino acids[b]	Targeted sequences
WRKY 1 FP	TGGMGIAARTAYGGNCARA	64	WRKYGQ	All groups, all WRKY domains
WRKY 2 FP	GAYGGITAYAAYTGGMGIAARTAYGG	64	DGYNWRKY	Group I, N-terminal WRKY domain
WRKY 3 FP	GCIAMYATGMGNAARGC	64	A(T/N)MRKA	Group II, subgroup b Upstream of WRKY domain
WRKY 2 RP	TGRBYRTGYTTICCYTCRTAIGTDGT	576	TTYEGKH(T/N/S/G/A/D)(H/Q)	Group I, C-terminal WRKY domain and group II, subgroups a–c
WRKY 3 RP	GTDAYIGTIGGRAANGG	48	PFPT(I/V)T	Group II, subgroup b Downstream of WRKY domain

[a] *FP*, Forward primer; *RP*, reverse primer
[b] The nucleotide sequence is given in the 5′–3′ direction using the standard IUB code where M = A or C; R = A or G; Y = C or T; B= T, C, or G; D = A, T, or G; N = A, T, C, or G; I= Inosine. Amino acids are given in the standard one-letter code
[c] *Deg*, Overall degeneracy of the primer omitting inosine

gen Life Technologies, Carlsbad, Calif.). Transformants were plated out onto LB plates with 100 μg/ml ampicillin and incubated overnight at 37°C. Individual colonies were transferred to 96-well plates and amplified overnight at 37°C in 100–200 μl SOC media (100 μg/ml ampicillin). A total of 348 clones were selected from the Hot-to-80 WRKY 1 FP/WRKY 2 RP (W12) PCR. Between 8 and 16 colonies were selected for each gene-specific PCR using TSH516 as the template.

Bacterial lysis, M13 amplification, exonuclease treatment, and cycle sequencing were as described in Kuhn et al. (2003). The ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction kit, v2.0 with AmpliTaq DNA polymerase FS (Applied Biosystems), with either T3 or T7 as the primer was used. Capillary electrophoresis was carried out using Performance Optimized Polymer 6 (POP6) on either an ABI 310 [47 cm in length, 50-μm (i.d.) capillary] or a 3100 Genetic Analyzer [50 cm in length, 50-μm (i.d.) capillary] (Applied Biosystems). Analysis was performed using the ABI Prism DNA sequencing analysis software (ver. 3.7) with the CE-2 basecaller.

Analysis of sequenced PCR products

ABI trace sequence files were imported directly into GCG (WISCONSIN PACKAGE ver. 10.2; Accelrys, Burlington, Mass.) using theFROMTRACE-RSF command and/orSEQUENCHER 4.1 (Gene Codes Corp, Ann Arbor, Mich.). Ambiguous or incorrect base calls were corrected manually using the original trace data. Cloned fragments that were 180 nucleotides long or shorter were discarded to eliminate any bias introduced by the degenerate primers. The sequences were assembled into contigs using bothSEQUENCHER 4.1 under high stringency andSEQLAB (GCG). Individual sequences and the consensus sequence of each gene fragment were putatively identified usingBLASTN, BLASTX, and TBLASTX searches (Altschul et al. 1997) of GenBank at the National Center for Biotechnology and Information (NCBI) and a local database. Polymorphisms were identified by eye and were only considered if the original base call was unambiguous and found among a representative number of sequences.

Analysis of *T. cacao*WRKY putative loci

We attempted to align, usingPILEUP (GCG), the complete nucleotide sequence, the nucleotide sequence of WRKY domain region with and without the intron, and the putative open reading frames (ORFs) of the TcWRKY putative loci. TcWRKY putative loci were further characterized by aligning putative ORFs of the *T. cacao* WRKY domains with *A. thaliana* WRKY (AtWRKY) domains obtained from the Munich Information Center for Protein Sequences *A. thaliana* Database (MAtDB). PROTDIST was used for distance analysis using the Dayhoff's Pam 001 matrix on 1,000 datasets generated by SEQBOOT inPHYLIP 3.5 (Felsenstein 1989). Neighbor-joining trees were created using NEIGHBOR and CONSENSE (PHYLIP 3.5). TREEEXPLORER ver. 2.12, was used to display and edit the tree (Kumar et al. 2001). Further details, a complete list of the AtWRKY sequences used, and the complete tree are presented in the ESM.

Specific primers

Specific primers for CE-SSCP were designed to flank the identified polymorphisms in six of the 16 TcWRKY loci, for potentially polymorphic regions in two other loci (TcWRKY-4 and -14), and for two loci where no polymorphisms were detected (TcWRKY-5 and -9) using PRIME (GCG). Primers were not designed for TcWRKY loci represented by fewer than five clones with no potential polymorphisms. The design of PCR primers specific for individual members of gene families for SSCP analysis can be successful provided each set of primers is specific for a single target and stringent PCR conditions are employed (Schneider et al. 1999; Slabaugh et al. 1997). Candidate primers were analyzed by BLASTN searches on GenBank and a locally created database. If the last four base pairs of the 3′ end matched other sequences or more than one TcWRKY locus, the primer was rejected. These constraints, combined with a high A/T content, prevented the design of suitable primers for TcWRKY-7 and -12. Unlabeled primers were supplied by Operon (Alameda, Calif.), and 5′ fluorescently end-labeled specific primers were supplied by Sigma-Genosys (The Woodlands, Tex.). The forward primer in each pair was labeled with 6-FAM and the reverse

Table 2 TcWRKY locus specific primers

Locus	Direction[a]	Primer sequence[b]	Position[c]	Expected product size (bp)	T_a[d] (°C)
TcWRKY-2	F	CGAGCTTACTACCGTTGCACCAT	28–50	169	55
	R	CTGCACACCTTTGCACCTGTTG	175–196		
TcWRKY-3	F	TCCTTACCCAAGGTAATGCCCCTG	17–40	649	55
	R	TGCTTACGGACGTTGCATCCT	645–665		
TcWRKY-4	F	CCAAGGTAAATTTCAGCTTCATCG	24–47	431	50
	R	CCTGTAAGCACCAGCAGT	437–454		
TcWRKY-5	F	CGAGCATATTATCGTTGCACTGTTGCACC	27–55	206	55
	R	GCACATCGTTGCACCTGAAATTCC	209–232		
TcWRKY-8	F	GGATGCCCTGTCAAAAAGAAGGTACTG	57–83	284	55
	R	CGCTGAACCTAGAGCCAGATGA	319–340		
TcWRKY-9	F	AACAGCCCTTATCCCAGGTATCAA	12–35	183	55
	R	GCTGGTGCAACGATAATAGCTCCT	171–194		
TcWRKY-10	F	CCCTTCACCTAATTGTTCAGGA	622–643	160	50
	R	CCCTCAAATCATGGGATGCT	762–781		
TcWRKY-11	F	CTCTCTTTCTTGCTCCATCAC	410–431	270	50
	R	CTTTCCACATGCTTTCTCACAG	659–680		
TcWRKY-13	F	AGCTGATAGGACCTCCTTTAGGC	799–821	325	55
	R	CGCACATTGCACTTAAGACTCGT	1,101–1,123		
TcWRKY-14	F	GCCCAGTTAGAAAACAGGTACTCCCA	61–86	437	50
	R	GGTGGTGATGAGGATTGTTCGGT	475–497[e]		

[a]Direction of the primer in relation to the sense strand of each gene. F, Forward; R, reverse

[b]Sequences are given in the 5′–3′ direction

[c]Positions of TcWRKY-13-specific primers are given for TcWRKY-13.1, accession no. AY331174 (see Table 4). Gene-specific primers were not designed for TcWRKY-1, -6, -7, -12, -15, and -16 (see text for details)

[d]T_a, Annealing temperature

[e]The reverse primer for TcWRKY-14 extends six nucleotides into the WRKY 2 RP degenerate primer binding site

primer was labeled with HEX. Primer sequences and the expected fragment sizes are given in Table 2.

All PCR primers were evaluated using OLIGO ANALYZER ver. 2.5 (http://biotools.idtdna.com/Analyzer) (Integrated DNA Technologies Coralville, Iowa) andOLIGONUCLEOTIDE PROPERTIES CALCULATOR ver. 3.02 (http://www.basic.northwestern.edu/biotools/oligocalc.html) (Northwestern University, USA).

Specific primer amplification conditions

The parents (Sca6, ICS1), F_1(TSH516), and F_2 (P1–P180) were the DNA templates for the TcWRKY-specific primers, both labeled and unlabeled. Thermocycler conditions were: 94°C, 5 min; (94°C for 30 s; annealing for 1 min; 72°C for 1 min) × 35; 72°C, 10 min; 4°C hold. Annealing temperatures were either 50°C or 55°C as empirically determined for each primer pair (Table 2).

CE-SSCP of the F_2population

CE-SSCP was performed on the ABI 310 using the protocol described by the manufacturer (Applied Biosystems) and on the ABI 3100 Genetic Analyzer as described in Kuhn et al. (2003). The PCR reactions were diluted 1:10 with deionized H_2O. ROX-labeled GeneScan-500 or GeneScan-1000 (Applied Biosystems) was used as the internal lane size standards. Samples were denatured at 95°C for 5 min and placed on ice for 10 min prior to electrophoresis. Electrophoresis was performed using 5% GeneScan Polymer (Applied Biosystems) in 1× TBE with 10% glycerol and 1× TBE with 10% glycerol as the running buffer. For the ABI 310, samples were injected electrokinetically at 10.0 kV for 5 s and electrophoresed at 13 kV, 30°C for 25–30 min on a 47-cm-long, 50-μm (i.d.) capillary. For the ABI 3100, the run parameters were as described in Kuhn et al. (2003). Data files were imported into GENESCAN 3.7 (Applied Biosystems), and individuals were manually scored as heterozygous or homozygous. Controls consisted of cloned sequences representing each identified allele.

Mapping

JOINMAP ver. 3.0 software (Van Ooijen and Voorrips 2001) was used to produce the genetic map using the Kosambi (1944) mapping function. Sixty-six markers were mapped to ten linkage groups (LGs) as has been previously described (Kuhn et al. 2003). The markers were: 37 SSR markers (Risterucci et al. 2000) provided by Dr. Claire Lanaud (CIRAD-Biotrop, Montpellier, France), 18 SSR markers, and seven RGHs developed at the USDA-ARS, SHRS (Kuhn et al. 2003). Novel markers mapped were four WRKY genes —TcWRKY-3, 10, 11, and -14—as well as an additional RGH—RGH3a. Markers were tested using a Chi-squared test at two degrees of freedom for goodness of fit to the expected 1:2:1 segregation ratio and placed at LOD≥8. LGs were identified by shared SSR markers between this map and a genetic map of UPA402×UF676 (Risterucci et al. 2000).

Table 3 Sequencing results of WRKY 1 FP/WRKY 2 RP PCR

GenBank accession no.	Locus	Potential alleles[ad]	Clones[b] (number)	Fragment length[c](bp)	WRKY domains[c]	WRKY group	Top BLASTX[d]	BLASTX accession
AY331152	TcWRKY-1	1	2	239	1	IIb	AtWRKY9	NP176982
AY331153–6	TcWRKY-2	4	5	217	1	IIb	AtWRKY6	AAK28312
AY331157–8	TcWRKY-3	2	10	702	1	I	AtWRKY4	NP849658
AY331159	TcWRKY-4	ND	2	455	1	I	PcWRKY2	S724444
AY331161	TcWRKY-5	1	51	254	1	IIb	AtWRKY72	NP197017
AY331162	TcWRKY-6	1	1	236	1	IIc	AtWRKY23	NP182248
AY331163–4	TcWRKY-7	2	39	241	1	I	NtWRKY4	BAA86031
AY331165–6	TcWRKY-8	2	17	368	1	IIa	WIZZ (Nt)	BAA87058
AY331167	TcWRKY-9	1	5	254	1	IIc	AtWRKY48	NP199763
AY331168–70	TcWRKY-10	3	52	202	1	I	SPF1 (Ib)	S51529
			1	789	2			
AY331171–2	TcWRKY-11	2	17	342	1	I	AtWRKY2	NP200438
			2	701a	1a			
AY331173	TcWRKY-12	1	16	317	1	I	RrWRKY	AAL32033
			1	976	2			
AY331174–6	TcWRKY-13	3	78	436	1	I	AtWRKY44	NP181263
			1	1166	2			
AY331177	TcWRKY-14	ND	19	491	1	IIb	AtWRKY42	NP192354
AY331178	TcWRKY-15	1	4	749	1	IIb	AtWRKY72	NP197017
AY331179	TcWRKY-16	1	1	566	1	IIc	StWRKY1	CAB97004

[a]The number of potential alleles for each locus was determined by sequence comparison among the representative clones
[b]Clones indicate the number of clones sequenced from the degenerate PCR found to represent each locus
[c]For TcWRKY loci containing two WRKY domains and TcWRKY-11, the fragment length and number of clones representing both WRKY domains and only the C-terminal domain are indicated separately. Lengths do not include the degenerate primer binding sites. The number of alleles for TcWRKY-14 could not be determined. An "a" following the value denotes that TcWRKY11 has not been sequenced in its entirety
[d]nt, Nucleotide; ND, not determined; At, *Arabidopsis thaliana*; Ib, *Ipomoea batatas*; Nt, *Nicotiana tabacum*; Pc, *Petroselinum crispum*; Rr, *Retama raetam*; St, *Solanum tuberosum*; Tc, *Theobroma cacao*

Results

Degenerate primer PCR

Of the six possible primer combinations, only the WRKY 1 FP + WRKY 2 RP combination (the primers designed byGENEFISHER) gave discrete PCR products consistently (Fig. 1b). WRKY 1 FP + WRKY 2 RP produced the same electrophoretic patterns for all the thermocycler parameters tested. A number of bands ranging in size from 250 bp to more than 1.5 kb were observed, as was expected, because of the two potential binding sites for the degenerate WRKY 1 FP (Fig. 1a). The other degenerate primer combinations either failed to amplify or produced a broad smear of products.

The WRKY 1 FP + WRKY 2 RP primer combination was specific. Of the 348 colonies sequenced, 324 (93%) identified with WRKY genes/proteins using BLASTX, BLASTN, and TBLASTX(Table 2). Only four PCR products (1%) identified with sequences other than WRKY genes at any level of significance. Twenty sequences (6%) were eliminated because they were either too small (<180 bps), composed of poor-quality sequence, or did not have any significant BLAST hits against GenBank. Of the clones identifying with the WRKY genes, the vast majority—304 of the 324—ranged in size from 202 bp to 491 bp.

Analysis of TcWRKY putative loci

The 324 sequenced clones represented fragments of 16 unique WRKY loci (Table 3). Of these 16 TcWRKY putative loci, seven were represented by five or fewer clones, while the other nine were represented by ten or more clones. The TcWRKY putative loci ranged in size from 217 bp (TcWRKY-2) to 1,166 bp (TcWRKY-13). Three loci—TcWRKY-10, -12, and -13—contained two WRKY domains, while the other 13 contained a single WRKY domain. Only one clone of the 53 sequenced for TcWRKY-10, one of the 17 sequenced for TcWRKY-12, and one of the 79 sequenced for TcWRKY-13 encompassed both WRKY domains; the remaining clones for these three fragments were 300 bp in length or smaller and represented only the C-terminal WRKY domain for each of these genes. Of the 19 clones representing TcWRKY-11, two extended the sequence information in the 5′ direction of the sense strand beyond the single identified WRKY domain. As sequencing reactions in the opposite direction of these two clones failed, the presence of a second, N-terminal WRKY domain for TcWRKY-11 has not been confirmed.

Size differences among the 16 TcWRKY putative loci were primarily due to variations in the length of the intron contained within the WRKY domain. These introns ranged in size from 93 bp (TcWRKY-10) to 636 bp (TcWRKY-15). When multiple clones representing a single TcWRKY putative locus were compared, the nucleotide sequences of the overlapping segments were more than 99% identical, and the nucleotide sequences of the intron were 98% or more identical. Alignments of the nucleotide sequences of the TcWRKY putative loci or of the WRKY domains alone, when the introns were included, were not possible due to the extreme variation in length and sequence of the introns. This same difficulty in aligning nucleotide sequences of WRKY domains and genes has been noted in separate investigations of *Arabidopsis* WRKY genes (Dong et al. 2003; Eulgem et al. 2000). Nucleotide sequences of two putative loci, TcWRKY-5 (group IIb) and -9 (group IIc), of the same overall length and containing introns of approximately the same size were aligned and were only 67% identical (data not shown). The coding regions of the WRKY DNA-binding domain of all the TcWRKY putative loci could be aligned using either nucleotide or translated amino acid sequences. The alleles identified for each TcWRKY putative locus were 100% identical for the coding portion of the DNA-binding domain. None of the 16 putative loci were identical with one another based on a comparison of the nucleotide or amino acid sequence of the coding region of the WRKY domain (see ESM). The closest nucleotide percentage identity between two TcWRKY putative loci for the WRKY domain coding region was 86% for TcWRKY-1 and TcWRKY-5. As the DNA-binding WRKY domain is expected to be the most conserved portion of the gene, it was concluded that each identified TcWRKY putative locus represented an individual gene locus.

The introns within the domains were at positions identical to those described in *Arabidopsis* for each group and subgroup (Eulgem et al. 2000). In TcWRKY loci with two domains, only the C-terminal domain was interrupted by an intron, which is consistent with group I AtWRKY genes. The position of the putative intron for six of the TcWRKY loci (-1, -2, -5, -8, -14, and -15)—following the conserved CX_5C motif of the zinc-finger—identified these as belonging to group II, subgroups a or b, WRKY genes. Further assignments of the group/subgroup for each of the 16 TcWRKY loci were made by comparisons with the top BLASTN,BLASTX, and TBLASTX hits for each sequence (Table 3).

Group assignments were confirmed by aligning the putative ORFs of each TcWRKY domain (19 in total) with AtWRKY domains (83 in total) obtained from MAtDB (Fig. 2 and ESM). The WRKY domains separated into distinct clusters representing each WRKY group and subgroup. The tree generated was consistent with trees from previous analyses (Dellagi et al. 2000; Eulgem et al. 2000). The association of individual TcWRKY domains with one another received little bootstrap support (<60%). Instead, the TcWRKY domains associated with individual AtWRKY domains or groups of AtWRKY domains (≥60% bootstrap support). The domains of seven TcWRKY loci (TcWRKY-3, -4, -7, -10, -11, -12, and -13) grouped with AtWRKY group I domains. Both of the N- and C-terminal domains of TcWRKY-10, -12, and -13 associated independently with the N- and C-terminal domains of the same AtWRKY gene (Fig. 2a, b). Of the six TcWRKY loci classified as group II, subgroups a or b,

only the TcWRKY-8 domain grouped with AtWRKY group IIa domains. The remaining loci were placed within group IIb domains (Fig. 2c). Three loci, TcWRKY-6, -9, and -16, grouped with AtWRKY group IIc domains (Fig. 2d). The TcWRKY domains were all located in the same groups and subgroups identified by sequence analysis.

Alleles were identified for seven of the 16 TcWRKY loci (44%) by sequence comparison of the clones comprising each TcWRKY locus produced by the degenerate PCR (Table 4). Twenty-four polymorphisms were detected, and all were single nucleotide polymorphisms (SNPs): 22 were substitutions and two were single nucleotide insertion/deletions (indels). All were located within the intron interrupting the conserved DNA-binding domain. The number of polymorphisms within each fragment ranged from a high of six SNPs in TcWRKY-3, denoting two alleles, to one SNP defining each of the four, three, and two alleles of TcWRKY-2, -10, and -11 respectively. No sequence differences were detected in the multiple clones representing each of seven TcWRKY loci. For two loci, TcWRKY-4 and -14, polymorphisms were noted, but the potential alleles could not be determined from the degenerate PCR. TcWRKY-4 was represented by only two sequences, and TcWRKY-14 contained a 21-nucleotide-long T-rich region interspersed with cytosines, which proved difficult to sequence. Sequences for each TcWRKY locus and identified allele have been deposited in GenBank (AY331152–AY331179).

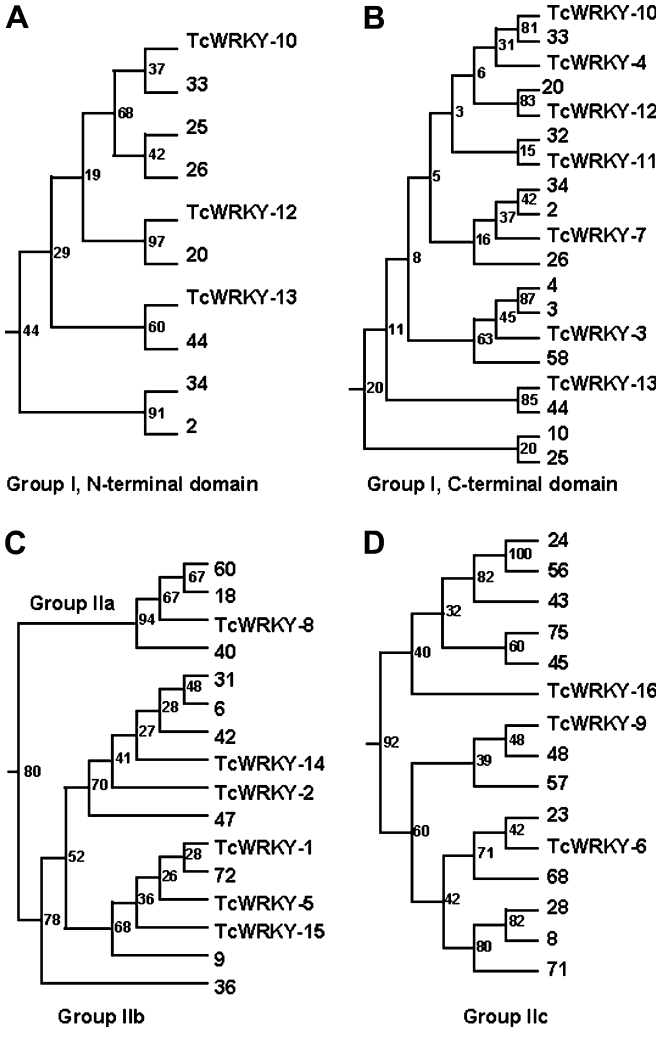

Fig. 2a–d Phylograms depicting the relationship of TcWRKY domains with *Arabidopsis thaliana*WRKY (*AtWRKY*) domains. AtWRKY domains are indicated by number only, i.e. 36 = AtWRKY36. Bootstrap support is given as a percentage of 1,000 datasets at each node, and groups/subgroups are designated as has been previously described for AtWRKY sequences (Eulgem et al. 2000). The complete tree and a list of AtWRKY accession numbers used are provided in the ESM

CE-SSCP

Primer sets, 5′ end-labeled with fluorescent nucleotides, were initially tested against the F_1 individual, TSH516, and a small sample of the F_2 population. Each amplified a single PCR product of the expected size. Four of the ten targeted loci, TcWRKY-3, -10, -11, and -14, were scored as heterozygous in the F_1 based on the pattern generated by CE-SSCP. The peaks and patterns representing each allele were easily distinguishable from one another. Six loci, TcWRKY-2, -4, -5, -8, -9, and -13, were judged to be homozygous in the F_1 and F_2individuals. For each locus, the forward strand and reverse strand gave simple patterns that were indistinguishable among the F_1 and F_2 individuals and clone controls.

For TcWRKY-10 (Fig. 3) and TcWRKY-14 (not shown), both the forward strand (blue) and reverse strand (green) for each allele produced a single major peak, indicating a single, stable conformation for each strand. For TcWRKY-10, Sca6 was homozygous, ICS1 was heterozygous, and TSH516 was heterozygous for two easily distinguished alleles (Fig. 3). For TcWRKY-11 (Fig. 3) and TcWRKY-3 (not shown), the patterns were more complicated. The TcWRKY-11 forward strand (blue) of each allele gave two major peaks and one minor peak, representing two equally stable conformations and one minor conformation, respectively. The TcWRKY-11 reverse strand (green) was uninformative. Electropherograms of clones 3A11 and 3E1, each a single allele of known sequence for TcWRKY-11, demonstrate that the SSCP patterns are from a single allele (Fig. 3). The parents, Sca6 and ICS1, were each homozygous for a different allele at this locus. TSH516 and F_2 individuals heterozygous for TcWRKY-11 contained four peaks of equal intensity, representing the two most stable conformations assumed by the forward strand for each allele.

The gene-specific PCR products of the F_1 individual, TSH516, were sequenced and the specificity of each primer set for the TcWRKY locus it was designed to amplify verified. Sequences for TcWRKY-3, -10, and -11 were heterozygous in the F_1 individual, thereby confirming the CE-SSCP results, and the polymorphisms present were identical with alleles already characterized. Four of the six loci judged homozygous by CE-SSCP—TcWRKY-2, -5, -9, and -13—were also homozygous based on

Table 4 TcWRKY alleles. Polymorphisms and alleles discovered for each TcWRKY locus

Locus	Allele[a]	Polymorphisms[a]	Position:type	TSH516[b]	Accession no.
TcWRKY-2	1	Indel; SNP	125: t/-; 135: g/a		AY331153
	2	Indel	125: t/-		AY331154
	3	–	–		**AY331155**
	4	SNPs	123: c/t; 124; g/t	X	AY331156
TcWRKY-3	**1**	–	–	X	**AY331157**
	2	SNPs	41: t/g; 159: a/t; 215: g/a; 406: c/t; 463: c/g; 558: a/g	X	AY331158
TcWRKY-4	**1**	–	–	X	**AY331159**
	2	SNPs	234: g/a; 277: c/g; 416: t/c	X	AY331160
TcWRKY-7	**1**	–	–		**AY331163**
	2	SNPs	64: g/a; 104: a/t		AY331164
TcWRKY-8	**1**	–	–	X	**AY331165**
	2	SNPs	119: t/c; 204: t/c		AY331166
	3	SNPs	131: t/c; 204: t/c	X	
TcWRKY-10	**1**	–	–		**AY331168**
	2	SNP	695: a/g	X	AY331169
	3	SNPs	695: a/g; 700: a/c	X	AY331170
TcWRKY-11	**1**	–	–	X	**AY331171**
	2	SNP	460: c/t	X	AY331172
TcWRKY-13	1	Indel; SNPs	847:t/-; 930: a/c; 962: c/t		AY331174
	2	Indel	117:t/- (= 847 in allele 1)		AY331175
	3	–	–	X	**AY331176**

[a]The reference sequence for the polymorphisms is given in bold. The clone encompassing both WRKY domains for TcWRKY-13 represents only allele 1; alleles 2 and 3 for TcWRKY-13 are represented only by the C-terminal WRKY domain. Novel alleles were detected by sequencing the TSH516 locus-specific products for TcWRKY-4 and -8. The polymorphisms for TcWRKY-14, although mapped in the F_2 population by SSCP, could not be determined (see text) and therefore are not represented

[b]Alleles present in the F_1 individual, TSH516, are indicated (X). Because suitable PCR primers could not be designed for TcWRKY-7, their presence were not verified in TSH516

sequencing and 100% identical with cloned sequences representing a fragment (TcWRKY-5 and -9) or cloned sequences representing an identified allele of a fragment (TcWRKY-2 and -13) (Table 4). TcWRKY-4 and -8, judged as homozygous by CE-SSCP at a single temperature, were discovered to be polymorphic in TSH516 by sequencing. In both cases, an additional allele (allele 2 for TcWRKY-4 and allele 3 for TcWRKY-8) containing novel SNPs was identified (Table 4).

Placement onto linkage groups

The genetic map was constructed from the Brazilian F_2 population of *T. cacao* using 60 markers (Fig. 4). Six markers of the 66 originally developed could not be placed reliably and were eliminated, including one marker, mTcCIR3, previously mapped on LG2 (Kuhn et al. 2003). The 60 markers formed ten LGs corresponding to the ten chromosomes of *T. cacao*, with nine of the ten LGs formed at LOD≥10. LGs ranged in size from 2.1 cM (LG3) to 80.1 cM (LG1). The map covered a total distance of 420 cM, which is 47% of the genetic map for UPA402×UF676 comprised of 424 markers and covering 885 cM (Risterucci et al. 2000). All shared markers (SSRs) between the two maps were colinear, except for LG3 in which the positions of mTcCIR40 and mTcCIR62 were inverted.

TcWRKY-3, -10, -11, and -14 were placed at LOD≥10, while RGH3a was placed on the map at a LOD≥8. TcWRKY-10 (35:79:29, one missing individual) and -11 (38:80:28) were both located on separate ends of LG5. TcWRKY-10 was on the terminal end of LG5 distal to SHRSTc37 extending LG5 out by 5.83 cM. TcWRKY-3 (40:71:30, five missing individuals) extended LG2 by 16.6 cM from mTcCIR48. TcWRKY-14 (43:79:16, eight missing individuals) mapped to LG1 within 1.7 cM of mTcCIR29. RGH3a (27:77:39) was placed between mTcCIR6 and mTcCIR53 on LG6 and was not associated with any of the other RGHs mapped previously (Kuhn et al. 2003). TcWRKY-3, -10, -11, and RGH3a segregated at the expected Mendelian ratio of 1:2:1 for codominant markers ($\chi^2 \leq 2.9$, $P > 0.05$). TcWRKY-14 showed skewed segregation at 3:5:1 ($\chi^2 = 13.5$, $P < 0.0001$). Its placement on LG1 is within a group of markers (SHRSTc3, mTcCIR29, and mTcCIR15) also showing skewed segregation.

Of the 60 markers mapped, 13 (22%) showed skewed segregation at $P < 0.05$. One SSR marker, mTcCIR15 ($\chi^2 = 47.9$, $P < 0.001$), was mapped to LG1 at a LOD≥3. The seven other markers mapped at this LG with LOD≥10. As the placement and recombination distance of mTcCIR15 agreed with previous *T. cacao* genetic maps from separate populations (Crouzillat et al. 2000; Flament et al. 2001; Risterucci et al. 2000), it was retained on the map. Most of the markers showing skewed segregation were clustered in four regions on separate LGs: four markers (mTcCIR15, TcWRKY-14, mTcCIR29, and SHRSTc3) were on LG1, three (SHRSTc44, mTcCIR25, and mTcCIR9) on LG6, three (mTcCIR8, RGH2, and mTcCIR58) on LG9, and two (mTcCIR10 and mTcCIR42) on LG5.

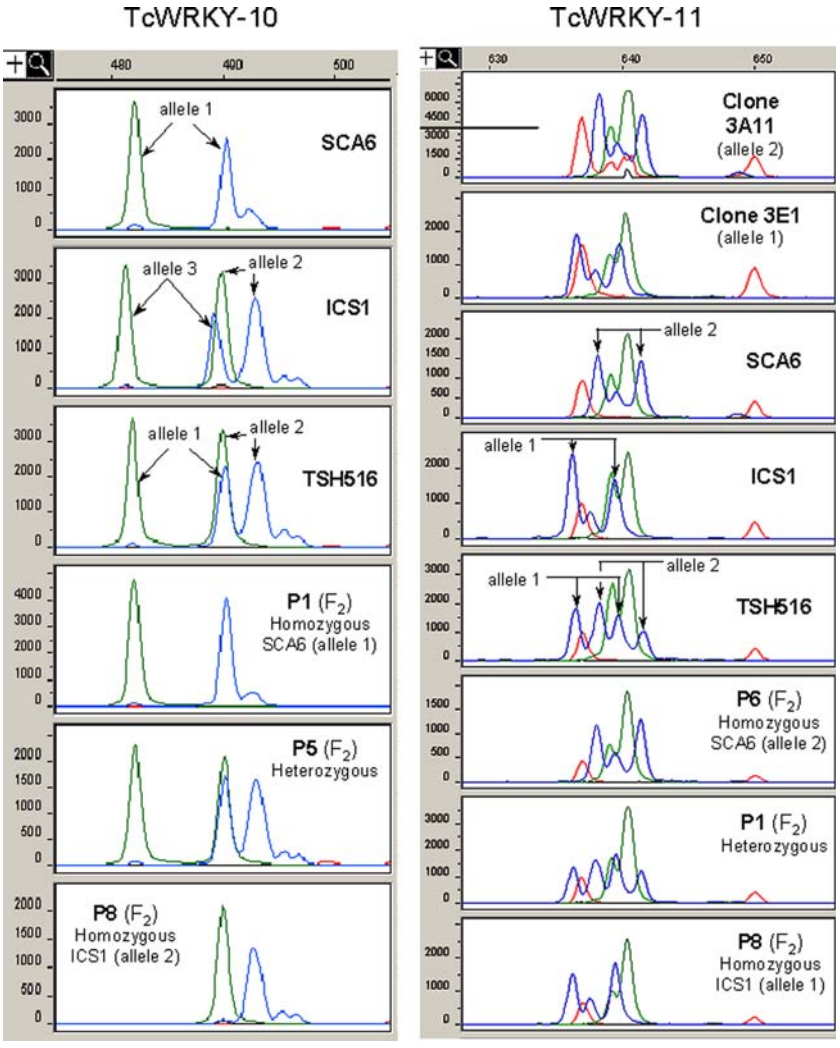

Fig. 3 SSCP electropherograms of TcWRKY-10 and -11 amplicons. The forward strand is shown in *blue* (FAM), the reverse strand in *green* (HEX). Peaks in *red* are ROX-labeled molecular-weight standards. *3A11* (AY331172) and *3E1* (AY331171) are clones of TcWRKY-11, each representing an individual allele. *Sca6* and *ICS1*, and *TSH516* are the parents and F_1individual, respectively. *P1*, *P5*, *P6*, and *P8* are F_2individuals

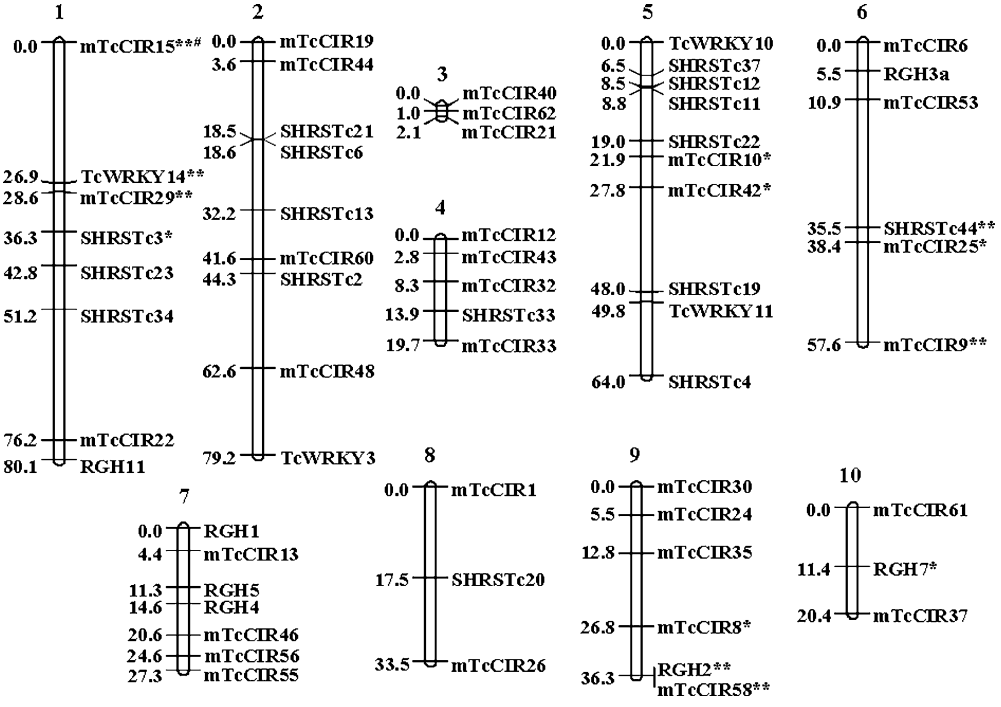

Fig. 4 Genetic linkage map of TSH516. CIRAD SSR markers are designated mTcCIR followed by a number, and USDA-ARS, SHRS SSR markers are designated SHRSTc followed by a number. *RGH*, Resistance gene homologs. Recombination distances are given in centi-Morgans (Kosambi 1944) and designate the position of the marker. *,**Markers with skewed segregation at $P<0.05$ and $P<0.01$ or lower, respectively. # designates TcCIR15 as the marker linked at a LOD≥3.0

Discussion

WRKY genes have been identified by two methods: sequence analysis of the entire genomes of *A. thaliana* (The Arabidopsis Genome Initiative 2000) and *Oryza sativa* (Goff et al. 2002; Yu et al. 2002), and by the isolation of differentially expressed cDNA (Alexandrova and Conger 2002; Hara et al. 2000; Hinderhofer and Zentgraf 2001; Huang and Duman 2002; de Pater et al. 1996; Yoda et al. 2002). Degenerate PCR primer pairs designed to specifically amplify WRKY gene fragments directly from genomic DNA have not been previously reported. Two publications have reported the use of degenerate primers to amplify WRKY genes, but in both cases, only a single degenerate primer designed from the gene sequence was used for amplification. Degenerate forward primers for the conserved WRKYGQK motif, similar in sequence with WRKY 1 FP which was independently developed, were used for domain-specific RT-PCR of expressed WRKY genes (Chen and Chen 2000). More recently, ligation-mediated-PCR (LM-PCR) was used to amplify WRKY genes from genomic DNA of potato (Trognitz et al. 2002). For LM-PCR, three degenerate primers designed to regions of the WRKY domain were used separately, in combination with a linker-specific primer, to generate what could be described as domain-specific amplified fragment length polymorphisms. The method described here allows the isolation and conversion of WRKY genes into sequence-specific molecular markers.

Ninety-three percent of the cloned PCR products from one pair of degenerate primers, WRKY 1 FP + WRKY 2 RP, identified with known WRKY genes and represented 16 unique loci. The size and sequence variability of the intron, the position of the intron within the WRKY domain, the number of WRKY domains, and conserved amino acid motifs within the WRKY domain were used to classify the cloned PCR products into 16 groups, and each group of cloned sequences represented an individual TcWRKY locus. In nucleotide and amino acid alignments of the most conserved portion of the TcWRKY loci, the DNA-binding WRKY domain, none of the TcWRKY loci were identical with one another. When TcWRKY domains were aligned with AtWRKY domains, the TcWRKY domains associated with individual AtWRKY genes or groups of AtWRKY genes with significant bootstrap support.

Of the 16 TcWRKY loci, nine (56%) contained multiple alleles based upon sequence comparison and CE-SSCP. All characterized polymorphisms were SNPs located within the intron interrupting the WRKY domain. Six of the 16 loci were polymorphic within an F_2 population of *T. cacao*. Four of the loci (TcWRKY-3, -10, -11, and -14) segregated within the F_2 population as detected by CE-SSCP, further confirming that the isolated TcWRKY fragments were separate loci. The placement of the four TcWRKY loci at disparate positions on the genetic map is consistent with the even distribution of WRKY genes found within the genomes of both *A. thaliana* and *Oryza sativa* (Eulgem et al. 2000; Goff et al. 2002). Two loci, TcWRKY-4 and -8 that were detected as monomorphic under the CE-SSCP conditions employed were heterozygous in the F_1 individual as detected by sequencing. Many factors contribute to the resolution of alleles by CE-SSCP, including pH of the buffer, temperature, size of the PCR product, and the number and location of SNPs (Kourkine et al. 2002; Ren and Ueland 1999). We are currently investigating electrophoresis conditions to resolve the alleles of TcWRKY-4 and -8 within the F_2 population.

The *Arabidopsis* and *Oryza* genome sequencing projects have identified 74 and 76 WRKY domain-containing genes, respectively (The Arabidopsis Genome Initiative 2000; Stephen A. Goff, personal communication). In *A. thaliana*, there are 16 group I, three group IIa, eight group IIb, and 17 group IIc WRKY genes (Eulgem et al. 2000). Assuming this is generally representative of the number and distribution of WRKY genes to be found in *T. cacao*, the single WRKY 1 FP + WRKY 2 RP PCR isolated 18–60% of each group and subgroup, and 20% of the total number of WRKY genes present within the genome. The same pair of degenerate primers, WRKY 1 FP and WRKY 2 RP, has also amplified putative WRKY gene fragments from *Cocos nucifera* and *Persea americana* in this laboratory (data not shown) and *Sorghum bicolor*, *Pennisetum glaucum*, and *Zea mays* independently in another laboratory (Randy Wisser, personal communication). In each case, the numbers of individual WRKY genes amplified and the types characterized (group I and group IIa-c) have been similar (data not shown). This pair of degenerate primers, WRKY 1 FP and WRKY 2 RP, should be useful for the amplification of groups I and II, subgroups a–c, WRKY genes across a wide range of plant species.

The reverse primer WRKY 2 RP determined the specificity of the WRKY 1 FP + WRKY 2 RP combination for group I and group II, subgroups a–c, WRKY genes. WRKY domains of group II, subgroups d and e, and group III WRKY proteins differ in their amino acid sequence at the C-terminal portion of the zinc-finger motif (Eulgem et al. 2000). It was expected that WRKY genes belonging to these groups would not be amplified, and none were identified. This suggests that degenerate primers could be designed to specifically amplify WRKY genes from each group or subgroup.

The association of the TcWRKY domains with the AtWRKY domains (Fig. 2 and ESM) is similar to that described in another large gene superfamily, the MADS-box genes (Theissen et al. 1996). Individual MADS-box gene homologs from different species are evolutionarily more related to one another than MADS-box gene paralogs present within a species. The duplication and divergence of the MADS-box gene family occurred early in the evolutionary history of plants (Alvarez-Buylla et al. 2000; Johansen et al. 2002; Purugganan 1997; Theissen et al. 1996). The evolution of the WRKY gene superfamily has also been proposed to be via a series of duplication events (Dong et al. 2003; Eulgem et al. 2000). The success of the degenerate PCR with templates from evolutionarily

distantly related plants (a basal angiosperm, monocots, and a Rosid), the size and sequence variability observed in WRKY genes within an individual species, and the association of individual WRKY genes from different species with one another when aligned suggest that the duplication events and divergence of WRKY genes into groups (at least for group I and group II, subgroups a–c) also occurred early in the evolution of plants.

WRKY genes may prove useful for comparative mapping between species or for the introgression of genes from closely related wild species. Synteny over a 50-kb region has been described between*Arabidopsis*, *Capsella*, and *Lycopersicon* for one WRKY gene (Rossberg et al. 2001), and genes involved in defense responses have been closely linked with WRKY domain-derived molecular markers in potato (Trognitz et al. 2002). Primers (different than those reported here) specific for TcWRKY-3, -8, and -13 were tested against several species related to *T. cacao* and successfully amplified a single PCR product whose nucleotide sequence was 85% or more identical to the appropriate *T. cacao*nucleotide sequence—including the non-coding regions (data not shown).

The method employed in this report could be considered a "shotgun approach" for the discovery of alleles because the DNA of six cacao cultivars/hybrids were bulked as template for the degenerate PCR and a large number of clones were sequenced. Other methods have been just as effective for isolating WRKY sequences and identifying alleles. For the *Cocos nucifera* PCR, two individual trees, each one representing a different variety, were used separately as the DNA template for the degenerate WRKY 1 FP + WRKY 2 RP PCR. Sixteen WRKY fragments were isolated and the alleles were identified by sequence comparison between WRKY genes from each tree and subsequently confirmed by CE-SSCP (data not shown). The size variation exhibited by WRKY gene PCR products can also be exploited to minimize the costs of screening the degenerate PCR. To determine the initial success of the *T. cacao*WRKY 1 FP + WRKY 2 RP PCR, we selected 62 of the 348 cloned sequences on the basis of insert size for sequencing. Of the 16 TcWRKY loci ultimately discovered, 15 were identified from these 62 clones. The products of the *T. cacao* WRKY 1 FP + WRKY 2 RP PCR were cloned in a single reaction, and 94% of the cloned sequences were shorter than 500 bp. Several products of the degenerate PCR were larger than 1.2 kb (Fig. 1b) and not represented in the library of clones. Several of the group I TcWRKY loci isolated—TcWRKY-3, -4, and -7—are only represented by a single WRKY domain, the C-terminal domain. As group I WRKY genes typically contain two domains (for example, TcWRKY-10, -12, and -13), the larger fragments may be representative of PCR amplification of TcWRKY-3, -4, -7 across both WRKY domains.

The WRKY gene superfamily offers several advantages for developing molecular markers. A single pair of degenerate primers, WRKY 1 FP and WRKY 2 RP, allows for the rapid isolation of several individual loci with a single PCR. The presence of an intron interrupting the WRKY domain is common and aids in identifying individual loci from one another. The introns for several of the WRKY genes identified in this project contained SNPs, oftentimes more than one, which allowed for the detection of alleles from individual loci. The application of SSCP to capillary array electrophoresis (CE-SSCP) (Kuhn et al. 2003; Kukita et al. 2002) allows for the high-throughput screening of hundreds of individuals for mapping populations or detecting novel alleles (i.e., germplasm screening). WRKY genes have been correlated with disease resistance and regulate plant responses to abiotic stresses, including salinity, mechanical damage, drought, and cold (Cheong et al. 2002, Cormack et al. 2002; Schenk et al. 2000; Seki et al. 2002). The *T. cacao* F_2 population used in this study is segregating for disease resistance for *Crinipellis perniciosa*. QTL mapping for disease resistance awaits the collection of further phenotypic data from this population. The TcWRKY markers generated in this report are now available for QTL mapping in this population as well as in other cacao populations to develop cultivars resistant to witches' broom, frosty pod, and black pod diseases. Because of the high degree of conservation of the WRKY domain, the pair of degenerate primers developed in this study can be applied across an evolutionary wide range of plant species. Therefore, they allow the application of a rapid method for developing sequence-specific markers in plants with undefined genomes, as is the case for most tropical trees and crops.

Acknowledgements The support of the USDA for research is gratefully acknowledged. J.B. was supported by NIH/NIGMS R25GM61347.

References

Ahnert DE (2000) Use of QTLs for Witches' broom resistance in cocoa breeding. In: Proc Int Workshop New Technologies Cocoa Breed. Kota Kinabalu, Sabah, pp 116–119

Ainsworth C (1993) Isolation of RNA from floral tissue of *Rumex acetosa* (Sorrel). Plant Mol Biol Rep 12:198–203

Alexandrova KS, Conger BV (2002) Isolation of two somatic embryogenesis-related genes from orchardgrass (*Dactylis glomerata*). Plant Sci 162:301–307

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) GappedBLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Alvarez-Buylla ER, Pelaz S, Liljegren SJ, Gold SE, Burgeff C, Ditta GS, de Pouplana LR, Marti'nez-Castilla L, Yanofsky MF (2000) An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci USA 97:5328–5333

Byrne PF, McMullen MD, Snook ME, Musket TA, Theuri JM, Widstrom NW, Wiseman BR, Coe EH (1996) Quantitative trait loci and metabolic pathways: genetic control of the concentration of maysin, a corn earworm resistance factor, in maize silks. Proc Natl Acad Sci USA 93:8820–8825

Chen C, Chen Z (2000) Isolation and characterization of two pathogen- and salicylic acid-induced genes encoding WRKY DNA-binding proteins from tobacco. Plant Mol Biol 42:387–396

Cheong YH, Chang HS, Gupta R, Wang X, Zhu T, Luan S (2002) Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*. Plant Physiol 129:661–677

Cormack RS, Eulgem T, Rushton PJ, Kochner P, Hahlbrock K, Somssich IE (2002) Leucine zipper-containing WRKY proteins widen the spectrum of immediate early elicitor-induced WRKY transcription factors in parsley. Biochim Biophys Acta Gene Struct Expr 1576:92–100

Crouzillat D, Phillips W, Fritz PJ, Pétiard V (2000) Quantitative trait loci analysis in *Theobroma cacao* using molecular markers. Inheritance of polygenic resistance to *Phytophthora palmivora* in two related cacao populations. Polygenic resistance to *Phytophthora palmivora* in cacao. Euphytica 114:25–36

Dellagi A, Helibronn J, Avrova AO, Montesano M, Palva ET, Stewart HE, Toth IK, Cooke DE, Lyon GD, Birch PR (2000) A potato gene encoding a WRKY-like transcription factor is induced in interactions with *Erwinia carotovora* subsp. *atroseptica* and *Phytophthora infestans* and is coregulated with class I endochitinase expression. Mol Plant Microbe Interact 13:1092–1101

Deng C, Davis TM (2001) Molecular identification of the yellow fruit color ©) locus in diploid strawberry: a candidate gene approach. Theor Appl Genet 103:316–322

Deslandes L, Olivier J, Theulieres F, Hirsch J, Feng DX, Bittner-Eddy P, Beynon J, Marco Y (2002) Resistance to *Ralstonia solanacearum* in *Arabidopsis thaliana* is conferred by the recessive *RRS1-R* gene, a member of a novel family of resistance genes. Proc Natl Acad Sci USA 99:2404–2409

Dong J, Chen C, Chen Z (2003) Expression profiles of the *Arabidopsis* WRKY gene superfamily during plant defense response. Plant Mol Biol 51:21–37

Duguma B, Gockowski J, Bakala J (2001) Smallholder cacao (*Theobroma cacao* L.) cultivation in agroforestry systems of West and Central Africa: challenges and opportunities. Agrofor Syst 51:177–188

Eulgem T, Rushton PJ, Schmelzer E, Hahlbrock K, Somssich IE (1999) Early nuclear events in plant defence signalling: rapid gene activation by WRKY transcription factors. EMBO J 18:4689–4699

Eulgem T, Rushton PJ, Robatzek S, Somssich IE (2000) The WRKY superfamily of plant transcription factors. Trends Plant Sci 5:199–206

Faris JD, Li WL, Liu DJ, Chen PD, Gill BS (1999) Candidate gene analysis of quantitative disease resistance in wheat. Theor Appl Genet 98:219–225

Felsenstein J (1989)PHYLIP—phylogeny inference package (version 3.2). Cladistics 5:164–166

Flament MH, Kebe I, ClTment D, Pieretti I, Risterucci AM, N'Goran JAK, Cilas C, Despreaux D, Lanaud C (2001) Genetic mapping of resistance factors to *Phytophthora palmivora* in cocoa. Genome 44:79–85

Giegerich R, Meyer F, Schleiermacher C (1996) GENEFISHER—software support for the detection of postulated genes. Proc Int Conf Intell Syst Mol Biol 4:68–77

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gotsch N (1997) Cacao crop protection: an expert forecast on future progress, research priorities and policy with help of the Delphi survey. Crop Prot 16:227–233

Gus-Mayer S, Naton B, Hahlbrock K, Schmelzer E (1998) Local mechanical stimulation induces components of the pathogen defense response in parsley. Proc Natl Acad Sci USA 95:8398–8403

Hara K, Yagi M, Kusano T, Sano H (2000) Rapid systemic accumulation of transcripts encoding a tobacco WRKY transcription factor upon wounding. Mol Gen Genet 263:30–37

Hinderhofer K, Zentgraf U (2001) Identification of a transcription factor specifically expressed at the onset of leaf senescence. Planta 213:469–473

Huang T, Duman JG (2002) Cloning and characterization of a thermal hysteresis (antifreeze) protein with DNA-binding activity from winter bittersweet nightshade, *Solanum dulcamara*. Plant Mol Biol 48:339–350

Hunter JR (1990) The status of cacao (*Theobroma cacao*, Sterculiaceae) in the Western Hemisphere. Econ Bot 44:425–439

Irizarry H, Rivera E (2002) Early yield of five cacao families at three locations in Puerto Rico. J Agric Univ P R 82:163–171

Johansen B, Pedersen LB,Skipper M, Frederiksen S (2002) MADS-box gene evolution—structure and transcription patterns. Mol Phylogenet Evol 23:458–480

Johnson CS, Kolevski B, Smyth DR (2002) TRANSPARENT TESTA GLABRA2, a trichome and seed coat development gene of Arabidopsis, encodes a WRKY transcription factor. Plant Cell 14:1359–1375

Kim CY, Lee SH, Park HC, Bae CG, Cheong YH, Choi YJ, Han CD, Lee SY, Lim CO, Cho MJ (2000) Identification of rice blast fungal elicitor-responsive genes by differential display analysis. Mol Plant Microbe Interact 13:470–474

Kosambi DD (1944) The estimation of map distance from recombination values. Ann Eugen 12:172–175

Kourkine IV, Hestekin CN, Barron AE (2002) Technical challenges in applying capillary electrophoresis-single strand conformation polymorphism for routine genetic analysis. Electrophoresis 23:1375–1385

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

Kukita Y, Higasa K, Baba S, Nakamura M, Manago S, Suzuki A, Tahira T, Hayashi K (2002) A single-strand conformation polymorphism method for the large-scale analysis of mutations/polymorphisms using capillary array electrophoresis. Electrophoresis 23:2259–2266

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Lamb CJ, Ryals JA, Ward ER, Dixon RA (1992) Emerging strategies for enhancing crop resistance to microbial pathogens. Biotechnology 10:1436–1445

Maeo K, Hayashi S, Kojima-Suzuki H, Morikami A, Nakamura K (2001) Role of conserved residues of the WRKY domain in the DNA-binding of tobacco WRKY family proteins. Biosci Biotechnol Biochem 65:2428–2436

Marita JM, Nienhuis J, Pires JL, Aitken WM (2001) Analysis of genetic diversity in *Theobroma cacao* with emphasis on witches' broom disease resistance. Crop Sci 41:1305–1316

Motamayor JC, Risterucci AM, Lopez AP, Lanaud C (2000) Domestication du cacoyer cultives par les Mayas. I. Preuve d'une origine sud americaine des cacaoyers cultives par les mayas. In: Proc 13th Int Cacao Res Conf. Kota Kinabalu, pp 22–44

de Pater S, Greco V, Pham K, Memelink J, Kijne J (1996) Characterization of a zinc-dependent transcriptional activator from *Arabidopsis*. Nucleic Acids Res 24:4624–4631

Pnueli L, Hallak-Herr E, Rozenberg M, Cohen M, Goloubinoff P, Kaplan A, Mittler R (2002) Molecular and biochemical mechanisms associated with dormancy and drought tolerance in the desert legume *Retama raetam*. Plant J 31:319–330

Purugganan M (1997) The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. J Mol Evol 45:392–396

Ren JC, Ueland PM (1999) Temperature and pH effects on single-strand conformation polymorphism analysis by capillary electrophoresis. Hum Mutat 13:458–463

Risterucci AM, Grivet L, N'Goran JAK, Pieretti I, Flament MH, Lanaud C (2000) A high-density linkage map of *Theobroma cacao* L. Theor Appl Genet 101:948–955

Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, *Arabidopsis*, and *Capsella* genomes. Plant Cell 13:979–988

Sambrook J, Russell DW (2001) Molecular cloning: a laboratory manual, 3rd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, Manners JM (2000) Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc Natl Acad Sci USA 97:11655–11660

Schneider K, Borchardt DC, Schafer-Pregl R, Nagl N, Glass C, Jeppsson A, Gebhardt C, Salamini F (1999) PCR-based cloning and segregation analysis of functional gene homologues in *Beta vulgaris*. Mol Gen Genet 262:515–524

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A, Sakurai T, Satou M, Akiyama K, Taji T, Yamaguchi-Shinozaki K, Carninci P, Kawai J, Hayashizaki Y, Shinozaki K (2002) Monitoring the expression profiles of 7000 *Arabidopsis* genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 31:279–292

Slabaugh MB, Huestis GM, Leonard J, Holloway JL, Rosato C, Hongtrakul V, Martini N, Toepfer R, Voetz M, Schell J, Knapp SJ (1997) Sequence-based genetic markers for genes and gene families: single-strand conformational polymorphisms for the fatty acid synthesis genes of Cuphea. Theor Appl Genet 94:400–408

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Theissen G, Kim JT, Saedler H (1996) Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene subfamilies in the morphological evolution of eukaryotes. J Mol Evol 43:484–516

Thompson JD, Plewniak F, Thierry J, Poch O (2000) DBCLUSTAL: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 28:2919–2926

Thorup TA, Tanyolac B, Livingstone KD, Popovsky S, Paran I, Jahn M (2000) Candidate gene analysis of organ pigmentation loci in the Solanaceae. Proc Natl Acad Sci USA 97:11192–11197

Toxopeus H (1969) Cacao. In: Ferwerda FP, Wit F (eds) Outlines of perennial crop breeding in the tropics. H. Veenman & Zonen, Wageningen, pp 79–109

Trognitz F, Manosalva P, Gysin R, Ninio-Liu D, Simon R, del Herrera MR, Trognitz B, Ghislain M, Nelson R (2002) Plant defense genes associated with quantitative resistance to potato late blight in *Solanum phureja* × dihaploid *S. tuberosum* hybrids. Mol Plant Microbe Interact 15:587–597

Van Ooijen JW, Voorrips RE (2001) JOINMAP version 3.0, software for the calculation of genetic linkage maps. Plant Research International, Wageningen

Yoda H, Ogawa M, Yamaguchi Y, Koizumi N, Kusano T, Sano H (2002) Identification of early-responsive genes associated with the hypersensitive response to tobacco mosaic virus and characterization of a WRKY-type transcription factor in tobacco plants. Mol Genet Genomics 267:154–161

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92