

The potential of the WRKY gene family for phylogenetic reconstruction: An example from the Malvaceae

James W. Borrone^a, Alan W. Meerow^{a,*}, David N. Kuhn^a, Barbara A. Whitlock^b,
Raymond J. Schnell^a

^a USDA-ARS, Subtropical Horticultural Research Station, National Germplasm Repository, 13601 Old Cutler Road, Miami, FL 33158, USA

^b 29 Cox Science Center, Department of Biology, University of Miami, 13601 Memorial Drive, Coral Gables, FL 33124, USA

Received 11 October 2006; revised 15 June 2007; accepted 19 June 2007
Available online 30 June 2007

Abstract

The WRKY gene family of transcription factors is involved in several diverse pathways and includes components of plant-specific, ancient regulatory networks. WRKY genes contain one or two highly conserved DNA binding domains interrupted by an intron. We used partial sequences of five independent WRKY loci to assess their potential for phylogeny reconstruction. Loci were originally isolated from *Theobroma cacao* L. by PCR with a single pair of degenerate primers; loci-specific primers were subsequently designed. We tested those loci across the sister genera *Herrania* Goudot and *Theobroma* L., with *Guazuma ulmifolia* Lam. as the outgroup. Overall, the combined WRKY matrices performed as well or better than other genes in resolving the intrageneric phylogeny of *Herrania* and *Theobroma*. The ease of isolating numerous, independent WRKY loci from diverse plant species with a single pair of degenerate primers designed to the highly conserved WRKY domain, renders them extremely useful tools for generating multiple, single or low copy nuclear loci for molecular phylogenetic studies at lower taxonomic levels. This is the first demonstration of the potential for members of the WRKY gene family for phylogenetic reconstruction.

Published by Elsevier Inc.

Keywords: WRKY gene family; Phylogeny; *Theobroma*; *Herrania*; Low copy nuclear genes; Byttnerioideae

1. Introduction

Single or low copy nuclear genes represent a source of multiple, unlinked and independently evolving loci, the ideal data set for molecular phylogenetic inference (Cronn et al., 2002, 2003; Rokas et al., 2003) due to their high rate of synonymous substitution compared to chloroplast or mitochondrial genes (Wolfe et al., 1987; Gaut, 1998) and biparental inheritance (Hughes et al., 2006). The relative dearth of single or low copy nuclear genes successfully used for phylogenetic analysis is due, in part, to the methodological problems associated with their isolation and amplification (Small et al., 2004). In general, two approaches have been used to generate phylogenetically useful low copy

nuclear genes, each having its own advantages and disadvantages (reviewed in Hughes et al., 2006). The sequence characterized amplified region (SCAR)-based approach obtains sequence information from randomly amplified genomic regions through the use of AFLP or RAPD primers. The comparative anchor tagged sequence (CATS)-based approach compares expressed sequence tags (ESTs) and/or complete genomic sequences to identify “candidate” genes, assuming that the sequence conservation observed across evolutionarily distant taxa implies orthology.

Several gene families, including the WRKY gene family of transcription factors (named for the highly conserved amino acid motif) (Eulgem et al., 2000), show evolutionary expansion in plants (Riechmann et al., 2000; Lespinet et al., 2002; Shiu et al., 2005) and are components of plant-specific, ancient regulatory networks (Doebly and Lukens,

* Corresponding author. Fax: +1 305 969 6410.

E-mail address: ameerow@saa.ars.usda.gov (A.W. Meerow).

1998; Nishiyama et al., 2003; Gutiérrez et al., 2004; Xiong et al., 2005). WRKY transcription factors are involved in several diverse pathways (reviewed in Ülker and Somssich, 2004) including regulating starch, anthocyanin, and sesquiterpene anabolism; seed development (García et al., 2005; Luo et al., 2005); trichome development; embryogenesis; and plant responses to both abiotic and biotic stresses.

WRKY genes contain a highly conserved DNA binding domain about 60 amino acids in length composed of the conserved WRKYGQK sequence followed by a C₂H₂- or C₂HC-type zinc finger motif. WRKY genes originally were classified into groups and subgroups based upon the number and type of WRKY domains, additional amino acid motifs, and phylogeny resolved with 58 members of the *Arabidopsis thaliana* (L.) Heynh. WRKY gene family (Eulgem et al., 2000). Group 1 WRKY genes were defined by the presence of two WRKY domains, each of the C₂H₂-type zinc finger motif, with only the C-terminal WRKY domain actively binding DNA. Group 2 WRKY genes contained only a single WRKY domain, and were classified into subgroups a–e based upon additional amino acid motifs found outside the WRKY domain. Group 3 WRKY genes were defined by the presence of the C₂HC-type zinc finger motif in the DNA-binding WRKY domain. Further annotation of *Arabidopsis* WRKY genes (Dong et al., 2003) and phylogenies of *Oryza* L. and *Arabidopsis* WRKY gene families have resulted in several competing modifications to the original classification scheme (Wu et al., 2005; Zhang and Wang, 2005; Xie et al., 2005). The original group/subgroup designations proposed by Eulgem et al. (2000) are used throughout the manuscript.

A feature common to WRKY genes is interruption of the coding region of the highly conserved, DNA-binding functional WRKY domain (the C-terminal WRKY domain in Group 1 WRKY genes and the single domains of Groups 2 and 3) with an intron. The size and sequence of the intron vary in each gene, but its position is highly conserved within each group/subgroup (Eulgem et al., 2000; Dong et al., 2003; Wu et al., 2005; Xie et al., 2005; Zhang and Wang, 2005). Variability present in the intron prevents nucleotide alignment of even highly conserved regions from individual WRKY genes found within a single species (Borrone et al., 2004; Borrone, 2004). Dong et al. (2003) proposed that the WRKY gene family was ancient and had expanded via duplication events. The ancient origin of WRKY genes was confirmed by discovery of a single WRKY gene each in *Dictyostelium discoideum* Raper, *Giardia lamblia* Kofoid and Christiansen, and *Chlamydomonas reinhardtii* P.A. Dangeard and several WRKY genes from *Physcomitrella patens* (H. Crum and L.E. Anderson) B.C. Tan and *Ceratopteris richardii* Brongn. (reviewed in Ülker and Somssich, 2004). Subsequent analyses of *Arabidopsis* have shown that expansion of the WRKY gene family was due primarily to segmental duplications (Cannon et al., 2004) of entire genomic regions likely as a result of separate polyploid events (Bowers et al., 2003; Thomas et al., 2006). Genomic regions contain-

ing WRKY genes retain a great degree of microsynteny even between evolutionarily distant plant species (Rossberg et al., 2001; Grover et al., 2004).

In this paper, we use partial sequences of five independent WRKY loci to assess the potential of the WRKY gene family for phylogeny reconstruction in plants using an exemplar data set from the Malvaceae. All of the WRKY loci investigated were originally isolated from *Theobroma cacao* by amplification with a single pair of degenerate primers (Borrone et al., 2004; Borrone, 2004). Independence of the loci has been previously established by genetic mapping (TcWRKY3, TcWRKY11, and TcWRKY14) onto separate linkage groups in *Theobroma cacao* and sequence analysis (Borrone et al., 2004). We tested the phylogenetic signal of these loci with the sister genera *Herrania* Goudot and *Theobroma* L. (Table 1) for which there are two published molecular phylogenetic analyses (Whitlock and Baum, 1999; Silva and Figueira, 2005).

Twenty-two species of *Theobroma* are recognized in six sections; *Andropetalum* Cuatr., *Glossopetalum* Bernoulli, *Oreanthes* Bernoulli, *Rhytidocarpus* Bernoulli, *Telmatocarpus* Bernoulli, and *Theobroma* Bernoulli; based upon tree architecture, fruit, petal and androecial characters (Cuatrecasas, 1964). *Herrania* contains 17 spp. divided into two sections, *Herrania* and *Subcymbicalyx* R.E. Schult., based upon floral morphology (Schultes, 1958). The close relationship of these two genera has been recognized as they share similar floral and fruit morphology and are differentiated primarily by growth form and leaf morphology (Whitlock and Baum, 1999). Figueira et al. (1994) found *Herrania* nested within *Theobroma* using rDNA polymorphisms. An analysis of plastid *ndhF* sequences across Malvaceae subfamily Byttnerioideae confirmed that the two genera form a well supported clade (Whitlock et al., 2001). The two previously published molecular phylogenetic analyses, one using the nuclear gene *vicilin* (Whitlock and Baum, 1999) and the other using trypsin inhibitor sequences (Silva and Figueira, 2005), found weak support for the monophyly of *Theobroma*.

2. Materials and methods

2.1. Sampling

Seven *Herrania* spp. representing both sections and 11 *Theobroma* spp. representing all six sections were sampled. Multiple individuals were sampled for a number of *Theobroma* spp. and *Herrania* spp.: two for *H. cuatrecasana*, *H. nycterodendron*, *H. purpurea*, *T. angustifolium*, *T. microcarpum*, *T. grandiflorum*, and three for *T. cacao*. Two species of *Grewia* L. and *Guazuma ulmifolia* were investigated for use as potential outgroups. *Guazuma* is in the same clade of Malvaceae (Byttnerioideae) as *Herrania* and *Theobroma* and is sister to the *Herrania/Theobroma* clade in previous phylogenetic analyses (Whitlock and Baum, 1999; Whitlock et al., 2001). *Grewia* is in the sister clade to Byttnerioideae, the Grewioideae (Alverson et al., 1999;

Table 1

Voucher specimens or USDA-ARS MIA DNA sample numbers, and GenBank accession numbers of the WRKY sequences used in the cladistic analyses

Species	Voucher or DNA sample [Source ^a]	GenBank Accession Nos. ^b				
		WRKY3	11	12	13	14
<i>Guazuma ulmifolia</i> Lam.	Whitlock & Bowser 360 (GH)	168	237	191	214	260
<i>Herrania albiflora</i> Goudot	TC01377 [CATIE]	169	238	192	215	261
<i>H. cuatrecasana</i> García-Barriga	Whitlock 312 (GH)	170	239	193	216	—
<i>H. kanukuensis</i> Schultes	Mori 24727 (NY)	171	240	194	217	262
<i>H. nitida</i> (Poepp.) Schultes	TC01371 [CATIE]	172	241	195	218	263
<i>H. nycterodendron</i> Schultes	Whitlock 315 (GH)	173	242	196	219	264
<i>H. purpurea</i> (Pitt.) Schultes	Whitlock 318 (GH)	174	243	197	220	265
<i>H. umbratica</i> Schultes	TC01376 [CATIE]	175	244	198	221	266
<i>Theobroma angustifolium</i> Moçônio & Sessé ex DC (1)	Whitlock 303 (GH)	176	245	199	222	267
<i>T. angustifolium</i> (2)	TC01368 [CATIE]	177	246	200	223	268
<i>T. bicolor</i> Humb. & Bonpl.	Hunter 1029 (WIS)	178	247	201	224	269
<i>T. cacao</i> L. cv. TSH516 (1)	TC00157 [CEPLAC]	179	248	202	225	270
<i>T. cacao</i> (2)	Whitlock 361 (GH)	180	249	203	226	271
<i>T. cacao</i> (3)	Whitlock s.n. (GH)	181	250	204	227	272
<i>T. chocoense</i> Cuatrec.	Whitlock 356 (GH)	182	251	205	228	273
<i>T. gileri</i> Cuatrec.	Whitlock 301 (GH)	183	252	206	229	274
<i>T. grandiflorum</i> (Willd. ex Spreng.) Schum.	Whitlock 305 (GH)	184	253	207	230	275
<i>T. mammosum</i> Cuatrec. & Léon	TC01367 [CATIE]	185	254	208	231	276
<i>T. microcarpum</i> C. Martius (1)	Whitlock 302 (GH)	186	255	209	232	277
<i>T. microcarpum</i> (2)	TC01369 [CATIE]	187	256	210	233	278
<i>T. simiarum</i> Donn. Smith	Whitlock 321 (GH)	188	257	211	234	279
<i>T. speciosum</i> Willd. ex Spreng	Hunter 1033 (WIS)	189	258	212	235	280
<i>T. velutinum</i> Benoist	Mori 24731 (NY)	190	259	213	236	281

^a Abbreviations: CATIE, Centro de Agronómico Tropical de Investigación Enseñanza, Costa Rica; CEPLAC, CEPLAC/CEPLAC, Itanubo, Brazil.

^b Only the last three digits are shown. All accession numbers begin with EF640. For example, the *Guazuma ulmifolia* WRKY03 GenBank Accession No. is EF640168

Bayer et al., 1999). Accessions included in the final phylogenetic data matrices are listed in Table 1.

2.2. DNA extraction

DNA for vouchered specimens (Table 1) was extracted as described in Whitlock and Baum (1999). For non-vouchered specimens, DNA was extracted using the BIO101 kit as described in Borrone et al. (2004) from leaf material of accessions maintained at the germplasm collections of the Centro de Agronómico Tropical de Investigación Enseñanza (CATIE), Turrialba, Costa Rica or CEPEC/CEPLAC, Itabuno, Brazil (Table 1). The quantity of DNA isolated was assessed with a GeneQuant pro RNA/DNA calculator (Amersham Pharmacia Biotech, Piscataway, NJ). Isolated DNA was stored at -20°C .

2.3. Prescreening

Accessions were prescreened with primer pairs designed to specifically amplify individual *Theobroma cacao* WRKY loci (previously reported in Table 2 of Borrone et al., 2004) using the amplification conditions described therein. Ten of eighteen WRKY loci were initially investigated. The eighteen *T. cacao* WRKY loci were originally isolated using a single pair of degenerate primers designed to the highly conserved WRKY domain (Borrone, 2004; Borrone et al., 2004). Five WRKY loci (Table 2)—TcWRKY3

(GenBank Accession No. AY331157), TcWRKY11 (GenBank Accession No. AY331171), TcWRKY12 (GenBank Accession No. AY331173), TcWRKY13 (GenBank Accession No. AY331174), and TcWRKY14 (GenBank Accession Number AY331177)—were further investigated because of amplification success observed across all taxa, the potential length of the fragment (>450 bp) to which specific PCR primers could be designed, and the phylogenetic signal observed by selected sequencing of the smaller products obtained from the initial prescreening amplifications. Three of the loci selected, TcWRKY3, TcWRKY11, and TcWRKY14, have been previously mapped to separate linkage groups in *T. cacao* (Borrone et al., 2004).

2.4. Primer design

PCR primers specific for WRKY3 are identical with those described in Borrone et al. (2004). PCR primers were redesigned using PRIME in GCG (Wisconsin Package Version 10.2, Accelrys, Burlington, MA) specific for TcWRKY11, TcWRKY12, TcWRKY13, and TcWRKY14 (Table 2, Fig. 1) to amplify larger fragments and differ from the PCR primers used to initially screen the accessions. The addition of 403 bp to the putative 5' end of TcWRKY11 (Borrone, 2004; Borrone et al., 2004) was obtained by resequencing the original clones isolated from the original degenerate PCR. Of the Group 1 WRKY loci (Table 2; Fig. 1), TcWRKY11, TcWRKY12, and TcWRKY13 sequences

Table 2
Primers used to amplify WRKY loci from *Guazuma ulmifolia*, seven *Herrania* spp. and 11 *Theobroma* spp.

Locus	WRKY group ^a	Primer	Sequence	T _m (°C)	Expected size ^b (bp)
TcWRKY03	1	FPW12-3	TCCTTACCCAAGGTAATGCCCTG	57.9	649
		RPW12-3	TGCTTACGGACGTTGCATCCT	56.5	
TcWRKY11	1	Tc11pF	GGTAGTGAATATCCAAGAAGC	56.8	1054
		Tc11pR	ACAGGACATCCAGGAGTTG	60.1	
TcWRKY12	1	Tc12pF	ACGCATCCTAATTGTGAAGTG	61.5	893
		Tc12pR	TTTTCTAACAGGGCAACCG	62.5	
TcWRKY13	1	Tc13pF	AAGCAAGTCAAAGGAAGTGAG	60.1	1165
		Tc13pR	TGAAAGCTCTTGGATCATCCGATGC	72.7	
TcWRKY14	2b	Tc14pF	GCCAAGGAAATCCATGTC	64.2	481
		Tc14pR	GGATTGTTCTGGTCTTCTGC	62.2	

^a WRKY group as originally defined by Eulgem et al. (2000).

^b Expected fragment size based upon *Theobroma cacao* WRKY sequences.

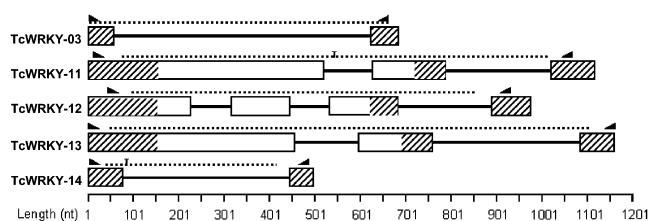


Fig. 1. *Theobroma cacao* WRKY loci used for analysis. The sequences are drawn approximately to scale. Potential exons are indicated by an open box, identified WRKY domains are crosshatched, and potential introns are indicated by solid black lines. PCR primers are indicated by the solid triangles and the sequence used for phylogenetic analysis is indicated above each locus by the dotted line. Homopolymeric regions deleted from WRKY11 and WRKY14 are indicated with a "T".

span the two conserved WRKY domains and contain several putative exons and introns, while the TcWRKY3 sequence spans a portion of the C-terminal WRKY domain and one intron (Fig. 1). The TcWRKY14 (Group 2b) sequence also contains a portion of a single WRKY domain and a single intron (Fig. 1). To verify that the target locus was amplified and to examine for potential paralogs, products for WRKY11, WRKY12, and WRKY13 were cloned and sequenced from four *Herrania* spp. and four *Theobroma* spp. (different accessions from those listed in Table 1) using the methods described previously (Borrone, 2004; Borrone et al., 2004) from amplifications independent from those used in the phylogenetic analyses.

2.5. Amplification and sequencing

Each WRKY locus was amplified separately from each template. Amplifications contained: 0.15 μ M forward primer, 0.15 μ M reverse primer, 200 μ M dNTPs, 15 μ g bovine serum albumin, 1 \times amplification buffer with 1.5 mM MgCl₂, 2.5 U of AmpliTaq Polymerase (Applied Biosystems, Inc., Foster City, CA) and 10 ng of DNA in a total volume of 50 μ l. Amplifications were conducted using PTC-225 thermal cyclers (MJ Research, Waltham, MA), conditions were: 95 °C, 2 min; [95 °C, 30 s; 55 °C, 60 s; 72 °C, 45 s] \times 35; 72 °C, 10 min; 4 °C, hold, and success

determined by agarose gel electrophoresis. For phylogenetic analyses, amplifications were treated with exonuclease to remove unincorporated PCR primers, ethanol precipitated, and resuspended in sterile H₂O to 50 μ l. Direct sequencing was done in both directions on 1 μ l of the treated amplification product with either the forward or reverse primer used for the initial amplification. All sequencing was done by capillary electrophoresis on an ABI 3100 Genetic Analyzer using the BigDye Terminator Cycle Sequencing Ready Reaction Kit v3.1 with AmpliTaq DNA Polymerase FS (Applied Biosystems, Foster City, CA). Contigs were assembled using Sequencher™ 4.1.4 (Gene Codes Corporation, Ann Arbor, MI), manually edited, and subjected to BLAST analyses against the non-redundant database at GenBank (Altschul et al., 1997). At positions within a sequence where double peaks of equal intensity occurred, indicating single nucleotide polymorphisms, IUPAC ambiguity codes were used. To further determine that separate loci were amplified, conceptual amino acid sequences of WRKY11, WRKY12, and WRKY13 products were aligned with *Arabidopsis thaliana* WRKY proteins using ClustalX (Thompson et al., 1997).

2.6. Phylogenetic analyses

Nucleotide sequences were aligned using Sequencher™ 4.1.4 and manually edited. For two loci, WRKY11 and WRKY14, homopolymeric regions were removed from the alignments. The aligned sequences for the five WRKY loci were analyzed separately and in every possible combination with maximum parsimony using branch and bound searches (Hendy and Penny, 1982) with simple addition conducted under the Fitch (equal) weights (Fitch, 1971) criterion in PAUP* v. 4b10 (Swofford, 2003). Each combined matrix was assessed for congruence using the incongruence length difference (ILD) test (Farris et al., 1994, 1995), implemented in PAUP* as partition homogeneity tests. One thousand heuristic searches were conducted for each test, each with 10 random addition replications, saving no more than 10 trees from each for TBR branch swapping. *p* values <0.05 were considered significant indication

of incongruence. Internal support was determined by bootstrapping (Felsenstein, 1985; 10,000 replicates with simple addition, saving 10 trees from each search for TBR branch-swapping) and calculation of Bremer (1988) decay indices (DI) using the program TreeRot v. 2.1 (Sorenson, 1996) with 1000 random addition replications in PAUP*. The cut-off bootstrap percentage is 50. A bootstrap value greater than 85% was considered strong support, 75–85% was considered good support, 65–75% was designated moderate support, and less than 65% as weak. These criteria have been used in the second author's published papers (Meerow and Snijman, 2001; Meerow et al., 2002; Meerow et al., 2003; Meerow and Clayton, 2004; Meerow et al., 2006). A minimum DI = 2 was considered to represent good support for a clade. Finally, for several of the combined matrices, partitioned Bremer indices (Baker and DeSalle, 1997) were calculated with TreeRot to determine the relative contribution of each data partition to each resolved clade. For one matrix, WRKY12+13+14, indels were included as characters using the "fifth base" gap mode in PAUP*. For two matrices (WRKY13 and all five WRKY loci combined), indels were included as characters using two methods, the "fifth base" gap mode in PAUP* and by creating a binary strict gap matrix with PAUPGAP (Anthony Cox, Wellcome Trust Sanger Institute, Cambridge, UK) which applies a strict interpretation of gaps (i.e., only gaps of equal length are considered homologous among taxa).

3. Results

3.1. Sequences

In the preliminary screening, primers for eight of 10 WRKY loci tested produced a single band migrating at the expected mobility upon gel electrophoresis for each taxon amplified. Three WRKY loci, WRKY5, WRKY8, and WRKY15, were excluded from the phylogenetic analyses because the amplified products were too short (<450 bp) and provided little phylogenetic signal (data not shown). Of the five loci further investigated, WRKY3, WRKY11, WRKY12, WRKY13, and WRKY14, direct sequencing of the amplification products gave clean, clear signals with little or no noise (with a couple of exceptions limited to individual samples noted below) and few base positions having "double peaks" of equal intensity indicating single nucleotide polymorphisms. PCR products for WRKY11, WRKY12, and WRKY13 were cloned to determine if a single locus was amplified. Nucleotide identity was >99% among the clones representing the same locus from the same species, and identical with the sequence information obtained by direct sequencing the PCR products from amplification of the same locus from the same species. WRKY3 sequences obtained from individual species were identical with those previously reported (Borrone, 2004). Sequence information for the five WRKY loci obtained from amplifications of more than one individual

representing *T. grandiflorum*, *H. cuatrecasana*, *H. nyctero-dendron*, and *H. purpurea* were identical, thus the information is presented from only one individual for these species. WRKY sequences obtained from the two *Grewia* spp. contained large insertion/deletions (indels) within putative introns of each WRKY locus. This, combined with the location of the primer binding sites close to or overlapping the putative intron/exon boundaries in the WRKY loci (Fig. 1), prevented the *Grewia* WRKY sequences from being unambiguously aligned. Thus, only *Guazuma ulmifolia* was included as an outgroup. Sequences used in the phylogenetic analyses are deposited in GenBank (Accessions EF640168–281, see Table 1) and the trees shown in this paper (Figs. 2 and 3) along with their respective NEXUS files are deposited in Treebase (study Accession No. S1813, matrix Accession Nos. M3318, M3319).

With the exception of WRKY3, alignments of the individual WRKY loci were easily accomplished. The average nucleotide identity between the outgroup, *G. ulmifolia*, and the ingroup ranged from 75% (WRKY11) to 89% (WRKY12). Within the ingroup, the average nucleotide identity for each WRKY locus was >95%. The aligned lengths ranged from 343 nt (WRKY14) to 1082 nt (WRKY13) (Table 3). Two loci, WRKY11 and WRKY14, contained homopolymeric regions within introns that could not be unambiguously aligned across all taxa, and 24 and 39 nt were trimmed, respectively, from each (Fig. 1). The percentage of ambiguous sites, including missing data, ranged from 0.13% (WRKY13) to 5.08% (WRKY14) (Table 3) and was primarily due to failure to adequately sequence through homopolymeric regions present in WRKY11 of *G. ulmifolia* (information was obtained for half of the sequence from a single direction) and the WRKY14 product from the *H. cuatrecasana* samples (in both directions). WRKY3 contained several large insertion/deletions (indels) of 39, 57, and 79 nt within the intron, in which adjacent regions, almost identical in nucleotide sequence, were duplicated/eliminated. Only one large indel, a 79 nt apparent insertion, was phylogenetically informative for members of *T. sect. Telematocarpus*. The two other large indels present in WRKY3 were each found in only a single species, a 57 nt apparent deletion in *H. kanukuensis* and a 39 nt apparent insertion in *T. bicolor*. A slight majority of the indels, 22 of 39, were autapomorphic including the only other large indel noted, 160 nt in the second intron of WRKY13. Informative indels were found in WRKY3 (4), WRKY13 (5), and WRKY14 (1). Most indels occurred within introns, except for WRKY11 in which four non-informative indels were present in exon 1. All indels occurring within exons were 3 nt in size resulting in the addition/deletion of a single amino acid.

BLAST results always identified the appropriate *T. cacao* WRKY loci as the top hit with *E* values <−100 for nucleotide matches. Alignments of the conceptual amino acid translations of WRKY11, WRKY12, and WRKY13 sequences with *Arabidopsis thaliana* WRKY sequences produced results identical with those previously reported for

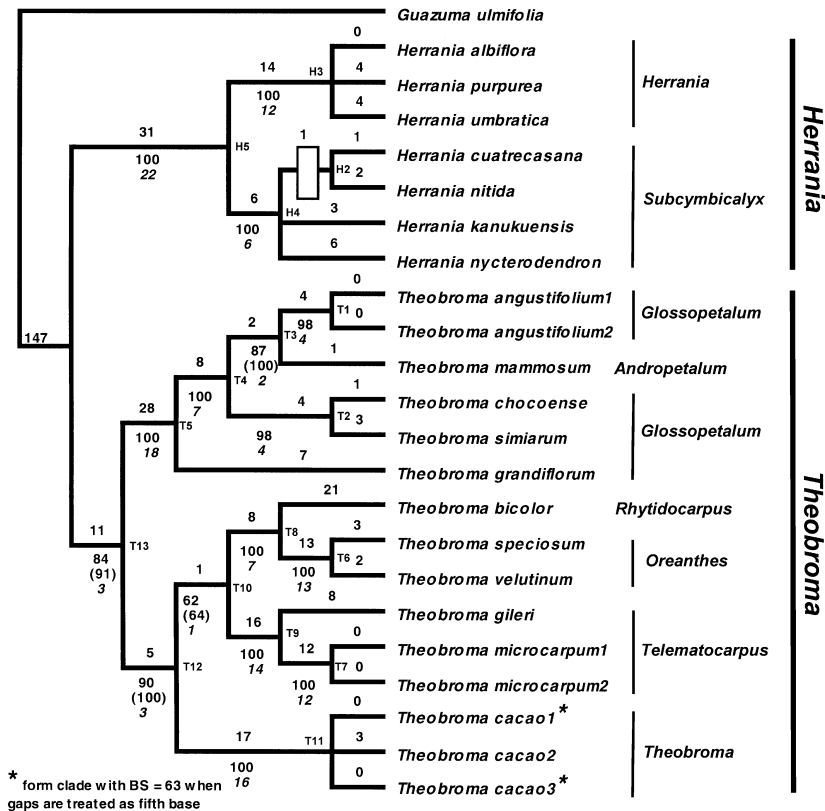


Fig. 2. One of two equally most parsimonious trees found by branch and bound cladistic analysis of combined DNA sequences of WRKY12, 13, and 14 loci across 18 *Herrania* and *Theobroma* spp. with *Guazuma ulmifolia* used as outgroup. Numbers above branches are branch lengths; numbers below are BS and DI (italic). BS percentages are only shown if ≥ 50 . BS percentages in parentheses are from analysis with gaps coded as a “fifth base” and are shown only if different from BS percentages with gaps coded as missing. White bar across a branch indicates its collapse in the strict consensus of both trees.

alignments of the *T. cacao* WRKY domains (Borrone et al., 2004). WRKY11 did not group with strong support with any individual AtWRKY protein, whereas WRKY12 putative translations associated with AtWRKY20 protein and the WRKY13 putative translations associated with AtWRKY44 with strong bootstrap support (data not shown). Uncorrected pairwise differences were calculated for each WRKY locus. WRKY3 and WRKY14 showed the highest evolutionary rates, 13–18% and 9–12%, respectively, between *G. ulmifolia* and the ingroup and <1–7% for both loci between *Herrania* and *Theobroma*. WRKY11, WRKY12, and WRKY13 each showed similar rates of evolution: 7–11% between *G. ulmifolia* and the ingroup and <1–6% between *Herrania* and *Theobroma*. These rates are comparable with those reported for the two previous phylogenetic analyses with *vicilin* (Whitlock and Baum, 1999) and trypsin inhibitor (Silva and Figueira, 2005) sequences.

3.2. Phylogenetic reconstruction

Excluding gaps, the range of phylogenetically informative characters was 22 (WRKY14) to 95 (WRKY13) for a total of 308 informative characters (7.9% of the aligned length) for all five loci combined (Table 4). Consistency and retention indices were consistently above 0.9 (Table

4), and never less than 0.85 when uninformative characters were excluded, for all the matrices investigated.

In the individual analyses, eight trees were found with WRKY3, three with WRKY11 and WRKY13, two with WRKY14, and one with WRKY12 and WRKY13 (gaps coded as a fifth base). None of the individual analyses were able to completely resolve the infrageneric relationships, but generally placed taxa within subclades consistent with their sectional classifications (Table 5) (Schultes, 1958; Cuatrecasas, 1964).

In all trees, *Herrania* was resolved as monophyletic with BS = 96–100 (Table 4). The monophyly of *Theobroma* was resolved (Table 4) only by WRKY12 (BS = 69) and WRKY13 (BS < 50, 1 of 3 trees); when gaps were coded as a fifth base with WRKY13, support for a monophyletic *Theobroma* rose from <50 to 68 (Table 4). The single trees found with WRKY12 and WRKY13 (gaps coded as a fifth base) were the best resolved and shared identical topologies. The two trees found by WRKY14 were the least resolved and placed various subclades of *Theobroma* in a polytomy with *Herrania* (Table 5). WRKY3, WRKY11 (2 of 3 trees), and WRKY13 (2 of 3 trees) resolved *Theobroma* as paraphyletic with differing topologies (not shown). For WRKY3, *T. speciosum* and *T. velutinum* (sect. *Oreanthes*) were sister with *Herrania* in a clade that that included *T. cacao* (sect. *Theobroma*) (BS = 97) and *T. gran-*

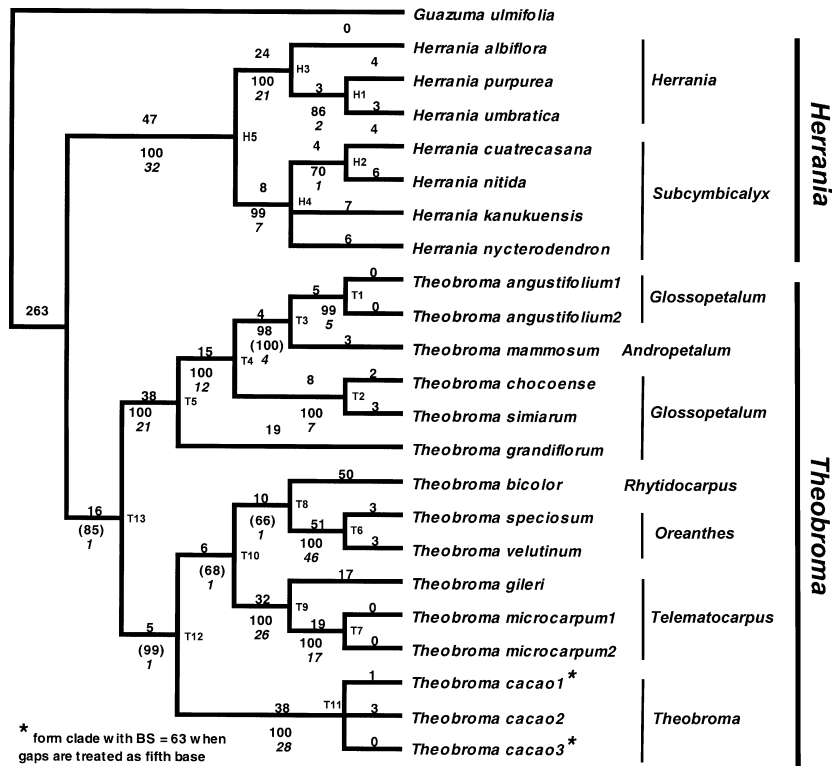


Fig. 3. Single most parsimonious trees found by branch and bound cladistic analysis of combined DNA sequences of all five WRKY loci across 18 *Herrania* and *Theobroma* spp., with *Guazuma ulmifolia* used as outgroup. Numbers above branches are branch lengths; numbers below are BS and DI (italic). BS percentages are only shown if ≥ 50 . BS percentages in parentheses are from analysis with gaps coded as a “fifth base” and are shown only if different from BS percentages with gaps coded as missing.

diflorum (5 of 8 trees, BS < 50). WRKY11 resolved *T. bicolor* (sect. *Rhytidocarpus*) sister to the remaining *Theobroma* subclades (BS = 58) that either formed a polytomy (1 of 3 trees) or formed a grade in which *T. cacao* (1 of 3 trees) or *T. cacao* and a subclade of *T.* sects. *Glossopetalum* and *Andropetalum* (1 of 3 trees) were grouped (BS < 50) with *Herrania*. WRKY13 (2 of 3 trees) also grouped with weak support (BS < 50) a subclade of *T.* sects. *Glossopetalum* and *Andropetalum* with *Herrania*.

All combinations of WRKY loci found 1–4 trees, except the combined WRKY3+13 matrix for which nine trees were found. Two *Herrania* subclades were resolved (BS > 65), by all matrices except WRKY3+14. *Herrania* spp. were fully resolved by matrices containing WRKY3+11, except those that also included WRKY12. The weakest support (BS < 65) was for the placement of *H. nycterodendron* sister to a *H. cuatrecasana*/*H. nitida* clade. The monophyly of *Theobroma* was resolved (Table 4) by five duo combinations [WRKY3+13 (BS < 50, three of nine trees), 11+12 (BS = 64), 12+13 (BS = 84), 12+14 (BS = 72), 13+14 (BS < 50, one of three trees)], four trios [WRKY3+12+13 (BS = 58), 11+12+13 (BS = 73), 11+12+14 (BS = 70), 12+13+14 (BS = 84)], three quad combos [WRKY3+11+12+13 (BS < 50, 1 of 2 trees), 3+12+13+14 (BS = 55), 11+12+13+14 (BS = 76)], and all five in combination (BS < 50). When gaps were coded as a fifth base, bootstrap support increased from 84 to 91 with

WRKY12+13+14 (Fig. 2) and from <50 to 85 with all five loci combined (Fig. 3, Table 4). When a strict gap matrix was used, there was no increased support for a monophyletic *Theobroma* over the nucleotide matrix alone in the combined matrices (Table 4).

Among combined matrices, the greatest degree of incongruence is observed when WRKY3 is combined with either WRKY12 or 13, or when WRKY11 is combined with WRKY12 (Table 4). Among the duo combinations, WRKY12+14 and WRKY13+14 have the highest degree of congruence (ILD, $p = 1$). Among the trio combinations, the most congruent was WRKY12+13+14 (ILD, $p = 0.765$); the least congruent were any trios including WRKY3 except WRKY3+11+14 (ILD, $p = 0.115$) (Table 4). Combined matrices of four or all five loci produced 1–4 trees, but were all significantly incongruent with WRKY11+12+13+14 the least so (ILD, $p = 0.025$). Nonetheless, there is a striking degree of congruence among the individual loci in terms of the subclades resolved within both *Herrania* and *Theobroma* (Table 5).

The two trees produced by the highly congruent WRKY12+13+14 combination (Fig. 2), and the single tree resolved by all five WRKY loci combined (Fig. 3), despite the incongruence of the partitions, are the most fully resolved and best supported of all the trees produced by the various iterations of our five WRKY sequences. These will be used to provide the framework in our discussion.

Table 3
Alignment results of WRKY loci

Locus	WRKY3	WRKY11	WRKY12	WRKY13	WRKY14
<i>Number of taxa</i>					
Ingroup	22	22	22	22	21 ^a
Outgroup	1	1	1	1	1
<i>Nucleotide (bp)</i>					
Mean length	650.1	908.3	751.9	1049.7	328.9
Range	584–727	630–923	726–755	886–1069	302–340
Aligned length	776	933	763	1082	343
GC content	34%	45%	42%	42%	38%
Ambiguous sites ^b	0.55%	1.60%	0.34%	0.13%	5.08%
<i>Exons, aligned length^c</i>					
Total	57	629	421	595	31
Exon 1	12 ^P	470 ^P	132 ^P	413 ^P	31 ^P
Exon 2	45 ^P	159	133	159	—
Exon 3	—	—	156	23 ^P	—
<i>Introns, aligned length^c</i>					
Total	719	304	342	487	312
Intron 1	719	89	92	138	312
Intron 2	—	215 ^P	85	349	—
Intron 3	—	—	165 ^P	—	—
<i>Variable sites</i>					
Total	163 (21.0%)	108 (11.6%)	108 (14.2%)	173 (16.0%)	61 (17.8%)
Exons	3	55	43	71	2
Introns	160	53	75	102	59
<i>Informative sites</i>					
Total	71 (9.1%)	61 (6.5%)	59 (7.7%)	95 (8.8%)	22 (6.4%)
Exons	1	29	24	34	0
Introns	70	32	35	61	22
Indels/informative indels	8/4	7/0	6/0	11/5	7/1
<i>Amino acids</i>					
Mean length	16.9	204.8	138.4	195.1	9.59
Range	3–18	179–208	133–139	182–198	1–10
Aligned length	18	209	139	198	10
Variable sites	2 (11.1%)	29 (13.9%)	11 (7.9%)	29 (14.6%)	0
PIC	0	16 (7.7%)	6 (4.3%)	14 (7.1%)	0

^a Sequence information was included as missing data for the *H. cuatrecasana* sample.

^b Includes missing data.

^c A “P” indicates only a partial sequence was obtained for this intron/exon.

Two main subclades are resolved in *Herrania*: one (H3) consisting of *H. sect. Herrania* (BS = 100, DI = 12, 21), and a second (H4) consisting of four species of *H. sect. Subcymbicalyx* (BS = 100, 99; DI = 6, 7). In *Theobroma*, two main subclades are resolved (Figs. 2 and 3). One (T5) represents *T. sect. Glossopetalum* with a single species of *T. sect. Andropetalum* (*T. mammosum*) as sister to *T. angustifolium*. The second larger clade (T12) includes a monophyletic (T11) *T. sect. Theobroma* (*T. cacao*) as sister to a larger subclade (T10) consisting of *T. sect. Rhytidocarpus* and its sister *T. sect. Oreanthes* (T8), and a monophyletic *T. sect. Telematocarpus* (T9). In the tree resolved by all five loci (Fig. 3), WRKY3 contributes the greatest amount of incongruence, with a negative DI at seven of the 18 nodes, followed by WRKY11 at five nodes (Table 6). WRKY13 contributes positively (DI ≥ 1) at 17 nodes. All five loci contribute positive support at five of the 18 nodes (28%): H3, H5, T6, T9, and T11. The most weakly

supported nodes are H1 and H2 in *Herrania* and T8, T10, T12, and T13 in *Theobroma*.

4. Discussion

The intent of this study was to assess the utility of the WRKY gene family for phylogenetic reconstruction. Part of the difficulty with generating low copy nuclear genes for phylogenetic reconstruction has been the methodology associated with isolating informative loci that can be readily amplified across taxa (Small et al., 2004; Hughes et al., 2006). The isolation and development of WRKY loci is straightforward. The ease and specificity of isolating numerous, independent WRKY loci using a single pair of degenerate primers has already been demonstrated from species representing the three major clades of the angiosperms (Borrone, 2004; Borrone et al., 2004). Eighteen separate WRKY loci from *Theobroma cacao* (eudicot), 21

Table 4

Results of branch and bound phylogenetic analyses of *Theobroma* and *Herrania* with WRKY loci alone and in combination^a

WRKY locus	# Characters	# PIC	# Trees	Length	CI ^b	RI	Clades ^c	Bootstrap	ILD (<i>p</i>) ^d
WRKY3	776	71	8	190	0.9526	0.9561	<i>Herrania</i>	97	—
WRKY11	933	61	3	125	0.9280	0.9521	<i>Herrania</i>	100	—
WRKY12	763	59	1	136	0.9338	0.9563	<i>Herrania</i>	100	—
							<i>Theobroma</i>	69	
WRKY13	1082	95	3	192	0.9583	0.9781	<i>Herrania</i>	100	—
							<i>Theobroma</i> (in one tree)	<50	
WRKY13-gap as 5th base	1082	122	1	434	0.9677	0.9743	<i>Herrania</i>	98	—
							<i>Theobroma</i>	68	—
WRKY13-strict gap matrix	1100	100	1	210	0.9619	0.9790	<i>Herrania</i>	100	—
							<i>Theobroma</i>	<50	
WRKY14	343	22	2	68	0.9853	0.9861	<i>Herrania</i>	96	—
WRKY3+11	1709	132	2	318	0.9340	0.9466	<i>Herrania</i>	100	0.088
WRKY3+12	1539	130	1	336	0.9167	0.9319	<i>Herrania</i>	100	0.001
WRKY3+13	1858	166	9	389	0.9383	0.9579	<i>Herrania</i>	100	0.005
							<i>Theobroma</i> (in 3 trees)	<50	
WRKY3+14	1119	93	3	259	0.9575	0.9603	<i>Herrania</i>	100	0.735
WRKY11+12	1696	120	2	266	0.9135	0.9416	<i>Herrania</i>	100	0.002
							<i>Theobroma</i>	64	
WRKY11+13	2015	156	2	318	0.9434	0.9675	<i>Herrania</i>	100	0.724
WRKY11+14	1276	83	3	194	0.9433	0.9577	<i>Herrania</i>	100	0.485
WRKY12+13	1845	154	2	329	0.9453	0.9685	<i>Herrania</i>	100	0.557
							<i>Theobroma</i>	84	
WRKY12+14	1106	81	1	204	0.9510	0.9640	<i>Herrania</i>	100	1
							<i>Theobroma</i>	72	
WRKY13+14	1425	117	2	260	0.9654	0.9794	<i>Herrania</i>	100	1
							<i>Theobroma</i> (in one tree)	<50	
WRKY3+11+12	2472	191	4	465	0.9188	0.9316	<i>Herrania</i>	100	0.001
WRKY3+11+13	2791	227	3	516	0.9322	0.9538	<i>Herrania</i>	100	0.004
WRKY3+11+14	2052	154	2	387	0.9406	0.9505	<i>Herrania</i>	100	0.115
WRKY3+12+13	2621	225	4	533	0.9231	0.9472	<i>Herrania</i>	100	0.001
							<i>Theobroma</i>	58	
WRKY3+12+14	1882	152	2	405	0.9259	0.9379	<i>Herrania</i>	100	0.001
WRKY11+12+13	2778	215	1	458	0.9323	0.9592	<i>Herrania</i>	100	0.01
							<i>Theobroma</i>	73	
WRKY11+12+14	2039	142	2	334	0.9281	0.9485	<i>Herrania</i>	100	0.007
							<i>Theobroma</i>	70	
WRKY11+13+14	2358	178	2	387	0.9483	0.9680	<i>Herrania</i>	100	0.617
WRKY12+13+14	2188	176	2	397	0.9521	0.9705	<i>Herrania</i>	100	0.765
							<i>Theobroma</i>	84	
WRKY12+13+14 gaps as 5th base	2188	205	2	712	0.9649	0.9698	<i>Herrania</i>	100	0.723
							<i>Theobroma</i>	91	
WRKY3+11+12+13	3554	286	2	662	0.9184	0.9440	<i>Herrania</i>	100	0.001
							<i>Theobroma</i> (in one tree)	<50	
WRKY3+11+12+14	2815	213	4	534	0.9195	0.9359	<i>Herrania</i>	100	0.001
WRKY3+11+13+14	3134	249	3	585	0.9368	0.9554	<i>Herrania</i>	100	0.007
WRKY3+12+13+14	2964	247	2	601	0.9301	0.9505	<i>Herrania</i>	100	0.001
							<i>Theobroma</i>	55	
WRKY11+12+13+14	3121	237	1	526	0.9392	0.9615	<i>Herrania</i>	100	0.025
							<i>Theobroma</i>	76	
All 5	3897	308	1	730	0.9247	0.9469	<i>Herrania</i>	100	0.001
							<i>Theobroma</i>	<50	
All 5-gap as 5th base	3897	442	1	1310	0.9450	0.9511	<i>Herrania</i>	100	0.001
							<i>Theobroma</i>	85	
All 5-strict gap matrix	3977	334	1	823	0.9174	0.9382	<i>Herrania</i>	100	
							<i>Theobroma</i>	<50	

^a PIC, phylogenetically informative characters; CI, consistency index; RI, retention index; BS, bootstrap support; ILD, incongruence length difference.^b CI and RI include autapomorphies. Excluding uninformative characters the CI and RI were all ≥ 0.85 .^c Resolution of *Herrania* or *Theobroma* as monophyletic.^d *p* values <0.05 were considered significant indication of incongruence.

from *Cocos nucifera* L. (monocot), and nine WRKY loci from *Persea americana* P. Mill. (a basal angiosperm) were identified. In this particular study, primers specifically

designed to selectively amplify 10 of 18 *T. cacao* WRKY loci were tested across related species. Eight of 10 were successful, although three loci—WRKY5, WRKY8, and

Table 5
Comparative resolution of *Herrania* and *Theobroma* subclades from Fig. 3 in trees produced by separate cladistic analyses of each WRKY locus^a

Node	WRKY3	WRKY11	WRKY12	WRKY13	WRKY14
H1	nr	86	nr	nr	nr
H2	nr	63	nr	63	n/a
H3	99	98	63	100	1/2
H4	nr	86	99	64 (67)	nr
H5	97	100	100	100 (99)	96
T1	nr	63	60	95	nr
T2	91	nr	87	64 (63)	64
T3	91	nr	nr	87 (100)	nr
T4	79	96	84	99 (97)	nr
T5	3/8	94	93	100	99
T6	100	98	95	100	63
T7	nr	92	100	98	nr
T8	89	nr	99	nr	64
T9	100	98	94	100	84
T10	nr	nr	59	nr	nr
T11	99	100	100	87 (99)	98
T12	nr	nr	nr	1/3 (100)	nr
T13	nr	nr	69	1/3 (68)	nr

^a Integers are bootstrap support; fractions are number of trees in which a node was resolved; nr, not resolved. BS in parentheses indicates value >50 when gaps were coded as fifth base and are shown only if different from BS percentages with gaps coded as missing.

Table 6
Partitioned DI for each node in the phylogenetic tree resolved from the combined matrix of all five WRKY loci (Fig. 3)

Node	WRKY3	WRKY11	WRKY12	WRKY13	WRKY14
H1	0	2	0	0	0
H2	0	1	-1	1	0
H3	5	3	1	10	2
H4	-1	2	5	1	0
H5	7	3	8	11	3
T1	0	1	1	3	0
T2	3	0	2	1	1
T3	2	0	0	2	0
T4	2	3	2	5	0
T5	-1.5	2	4	10.5	6
T6	21	3	10	10	2
T7	-5	-1	17	5	1
T8	-7	-3	9	1	1
T9	5	3.5	8	7	2.5
T10	-7	-3	9	1	1
T11	8	4	11	1	4
T12	-7	-3	9	1	1
T13	-7	-3	9	1	1

Negative DI indicates incongruent resolution for that locus at that node.

WRKY15—did not provide enough phylogenetic signal (data not shown) within the size fragments amplified. *Theobroma cacao* WRKY specific primers successfully amplified WRKY loci in the two *Grewia* spp. and *Adansonia digitata* L. (data not shown). Thus, there is merit in experimenting with these highly conserved WRKY primers across other genera of the Malvaceae. Similar results have been observed in the Arecaceae where PCR primers designed specifically to *Cocos nucifera* WRKY loci are successfully amplifying targets in 12 other genera of the tribe Cocoeae (Meerow et al., in prep.). The use of specific,

non-degenerate primers greatly reduces the likelihood of amplifying multiple loci.

Paralogy is the leading concern when using nuclear genes, especially members of multigene families. Paralogous sequences from gene duplication events due to unequal crossing over, replicative transposition or ancient polyploidization events, when unrecognized, may lead to erroneous phylogenetic inferences (Hughes et al., 2006; Martin and Burg, 2002). Several lines of evidence, both direct and indirect, argue against paralogy as an issue with our WRKY loci, although it cannot be ruled out conclusively. For all loci, including WRKY3, there was no indication, either by direct sequencing multiple individuals from a single species or from cloning (WRKY11, WRKY12, and WRKY13), that paralogous copies for any of the individual WRKY locus were amplified within a single species. The two WRKY loci, WRKY12 and WRKY13, not mapped to a single locus in *T. cacao* are analogous with individual *Arabidopsis* WRKY genes. Approximately one-third of the *Arabidopsis* WRKY loci found to be paralogous resulted from a polyploidization event (the alpha event) subsequent to the divergence of the Malvaceae family (Bowers et al., 2003); the next most recent polyploid event detected in *Arabidopsis* (the beta event) is estimated to have occurred just after the divergence of monocots from the rest of the angiosperms (Bowers et al., 2003). Naturally occurring hybridization events within *Theobroma* are rare, and intergeneric hybridizations with *Herrania* fail to produce viable offspring (see Section 4 below). *Theobroma* and *Herrania* share the same chromosome number, $2n = 20$, thus any polyploidization event most likely occurred prior to the divergence of these genera from *Guazuma*.

That the individual WRKY loci gave differing tree topologies was not unexpected, as many loci commonly used to infer phylogenetic relationships often result in incongruent topologies (Hughes et al., 2006; Rokas et al., 2003). WRKY3 was the most notably incongruent of the five loci included in this study (Table 6). For WRKY3, the same primer pair used for mapping WRKY3 onto the *T. cacao* genetic linkage map (Borrone et al., 2004) was used in this study. That WRKY3 has experienced a faster rate of evolution is evidenced by the number of large insertion/deletion events present (Table 3), and is supported further by the uncorrected pairwise distances (see Section 3). Alignments of WRKY3 sequences were difficult due to these large indels. Thus, the estimated rate of evolution and the incongruence observed are most likely artifacts of the uncertain alignment (Planet, 2006). WRKY3, therefore, may not be the best candidate to include in the phylogenetic analyses for this particular sample set. Inclusion of WRKY3, however, had no significant effect upon the overall tree topology found with all five loci combined (Figs. 2 and 3), and combinations of WRKY3+11 were the only ones sufficient to fully resolve the intraspecific relationships in *Herrania*, albeit with weak support at some nodes (Tables 5 and 6).

Two recent phylogenetic analyses of *Theobroma* and *Herrania* using nuclear DNA sequences provide good arbiters of the performance of WRKY loci for phylogenetic inference (Table 7). Both used proteins found as major constituents of seed protein, Whitlock and Baum (1999) used the nuclear gene *vicilin*, while Silva and Figueira (2005) analyzed trypsin inhibitor sequences. Our sampling is more similar to that of Whitlock and Baum (1999), and contains a number of the same genomic templates (Table 1). Both previous studies resolved *Theobroma* as monophyletic with only weak support (BS = 56, 55; DI = 1, 2; Table 7), while we obtained BS as high as 91 (DI = 3) for a monophyletic *Theobroma* clade in the combined analysis of WRKY12+13+14 with gaps included as a fifth base (Table 4; Fig. 2). WRKY12 alone had BS = 69 for a monophyletic *Theobroma* (Table 4). In Whitlock and Baum's (1999) *vicilin* study sections *Oreanthes* and *Rhytidocarpus* are sister to the rest of the genus. In contrast, our two large subclades are more similar to the results reported by Silva and Figueira (2005) and suggest that *T. sect. Theobroma* and sect. *Telematocarpus* are more closely related to sects. *Rhytidocarpus* and *Oreanthes* than to sect. *Glossopetalum* with the most support from the four of five loci combined WRKY11+12+13+14, (BS = 75) (data not shown). The weaker support for nodes of *Theobroma* in the five loci combined is due to WRKY3 placing *T. sect. Oreanthes* as sister to *Herrania*. Silva and Figueira (2005) found inconsistent resolution for *Telematocarpus* section using trypsin inhibitor sequences; this was corroborated by seed fatty acid profile (Silva et al., 2001) and purine alkaloids (Silva and Figueira, 2005, citing unpublished data).

The sister relationship of *T. sects. Rhytidocarpus* and *Oreanthes* was well supported in the *vicilin* phylogeny

(BS = 96, DI = 7). The congruent three locus combination WRKY12+13+14 strongly supported this clade with a BS = 100 and DI = 7 (Table 7, Fig. 2). In the combined five locus WRKY tree (Table 7, Fig. 3), this relationship is moderately supported only when gaps are included as a fifth base (BS = 66, DI = 2), and was primarily due to WRKY3 placing *T. sect. Oreanthes* as sister to *Herrania* and, less so, WRKY11 resolving *T. bicolor* as sister to the remaining *Theobroma* (Table 6). *Theobroma mammosum* (sect. *Andropetalum*) is well-supported as sister to *T. angustifolium* (BS = 87, 98, DI = 2, 4) in both of the combined trees (Table 7, Figs. 2 and 3), several other combined matrices (not shown), and by WRKY3 and WRKY13 alone (Table 5), rendering sect. *Glossopetalum* paraphyletic. This same resolution was supported by *vicilin* (BS = 92, DI = 1; Whitlock and Baum, 1999) and more weakly by trypsin inhibitor sequences (BS = 70, DI = 1; Silva and Figueira, 2005).

Like Whitlock and Baum's (1999) phylogeny, the monophyly of *Herrania* and resolution of *Herrania* into two main clades were strongly supported (Tables 4 and 7; Figs. 2 and 3). However, even with multiple WRKY loci, the relationships within *H. sect. Subcymbicalyx* remain unresolved. Unlike the *vicilin* phylogeny, in instances where *Theobroma* was rendered non-monophyletic, *Herrania* was always found on a terminal branch. Additionally, the weakly supported sister relationship found with *vicilin* between *H. cuatrecasana* and *H. nycterodendron* was resolved only by one locus, WRKY12, with weak support. All other matrices in which the relationships of *Herrania* were fully resolved placed *H. nitida* and *H. cuatrecasana* as sisters, albeit, also with moderate or weak support (Table 7; Figs. 2 and 3).

Table 7

Comparative resolution of combined analyses of WRKY12+13+14 (Fig. 2) and five WRKY loci (Fig. 3) with previous phylogenetic analyses of *Herrania* and *Theobroma*^a

Node	This study		Whitlock and Baum (1999)	Silva and Figueira (2005)
	WRKY12-13-14	All five		
H1	nr	86/2	100/7	n/a
H2	<50/0	70/1	nr	n/a
H3	100/12	100/21	100/7 (less <i>H. albiflora</i>)	n/a
H4	100/6	99/7	98/5	n/a
H5	100/22	100/32	100/27	nr
T1	98/4	99/5	n/a	n/a
T2	98/4	100/7	70/1	n/a
T3	87/2	98/4	92/1	70/1
T4	100/7	100/12	98/4	nr
T5	100/18	100/21	100/10	nr
T6	100/13	100/46	96/4	n/a
T7	100/12	100/17	99/5	n/a
T8	100/7	(66)/1	96/7	<50/1 (<i>T. sylvestre</i> instead of <i>T. velutinum</i>)
T9	100/14	100/26	100/23	n/a
T10	62/1	(68)/1	nr	<50/1
T11	100/16	100/28	100/13	100/13
T12	90/3	(99)/1	nr	<50/1 (<i>T. sylvestre</i> instead of <i>T. velutinum</i>)
T13	84/3	(85)/1	56/1	55/2

^a Numbers are BS/DI. nr, not resolved; n/a, not applicable. BS in parentheses indicates value >50 only when gaps were coded as fifth base.

Three morphological characters, branching pattern, tree architecture, and germination mode, divide *Theobroma* into two main groups (Cuatrecasas, 1964). Their evolutionary ambiguity, especially that of branching pattern (reviewed in Whitlock and Baum, 1999), was noted by Cuatrecasas (1964) in that they "...like other characters... are not sufficient to give taxonomic recognition to the two groups separated by those characters." The resolution of *Theobroma* by several combined matrices of the WRKY loci into two main clades is inconsistent with the two morphological characters that can be scored: tree architecture and germination mode. Trunks of *Theobroma* are described as sympodial, with sections *Andropetalum*, *Glossopetalum*, and *Telematocarpus* having "pseudoterminal" growth, wherein a new orthotropic shoot is produced from a bud above the whorl of plagiotropic branches. The seeds of these sections have hypogeal, cryptocotylar germination. Sections *Oreanthes*, *Theobroma*, and *Rhytidocarpus* have "subterminal" growth in which the orthotropic shoot is produced from a bud below the whorl of plagiotropic branches, and have epigeal, phaerocotylar seed germination. *Herrania* is monopodial, unbranching, and two species, *H. mariae* (C. Martius) Decne. ex Goudot (Addison and Tavares as reported by Cuatrecasas, 1964) and *H. purpurea* (N. Garwood unpublished data as reported by Whitlock and Baum, 1999), are described as having hypogeal, cryptocotylar germination whereas *G. ulmifolia* has epigeal, phaerocotylar germination (Burger, 1972). The phylogeny obtained with the WRKY sequences suggests that seed germination and tree architecture are homoplasious as a result of at least two separate evolutionary events.

The WRKY phylogeny for *Theobroma* is consistent with evidence from both intra- and intergeneric hybridization attempts and intrageneric grafting attempts (Addison and Tavares, 1952; Cuatrecasas, 1964; Silva et al., 2004). Very few natural hybrids of the genus *Theobroma* have been described. Hybridization attempts between *T. cacao* and *Herrania* (*H. mariae* or *H. balaensis* P. Preuss) and *T. cacao* with *Theobroma* spp. (*T. angustifolium*, *T. bicolor*, *T. grandiflorum*, *T. mammosum*, *T. microcarpum*, *T. obovatum* Klotzsch ex Bernoulli, *T. speciosum*) were negative, produced abnormal seeds that failed to germinate, or produced few seedlings that did not survive. Within *Theobroma*, hybridizations have produced reproductively viable adult trees only between species within the *Glossopetalum/Andropetalum* clade (*T. grandiflorum* × *T. obovatum*, *T. grandiflorum* × *T. subinicanum* Mart., *T. obovatum* × *T. subinicanum*, *T. mammosum* × *T. angustifolium*, *T. mammosum* × *T. simiarum*), and between species of *T.* sect. *Oreanthes* (*T. speciosum* × *T. sylvestre* Mart.). Only two compatible groups have been described from grafting: one confined to members of *T.* sect. *Glossopetalum* and the other composed of members of the strongly supported *Rhytidocarpus/Oreanthes* clade—*T. bicolor*, *T. speciosum*, and *T. sylvestre*. Although *T. obovatum* and *T. subinicanum* (sect. *Glossopetalum*) and *T. sylvestre* (sect. *Oreanthes*) were not included in the current analysis, their inclusion within

these sections was strongly supported in the trypsin inhibitor study (Silva and Figueira, 2005).

Overall, the combined WRKY matrices performed as well or better than *vicilin* or trypsin inhibitor sequences in resolving the intrageneric phylogeny of *Herrania* and *Theobroma*. The relative information content of even the shortest of the five sequences (WRKY14) was surprisingly high. Both CI and RI were consistently above 0.9, and ≥0.85 when uninformative characters were excluded, for all of the sequence matrices. The phylogenetic signal among three of our five loci in combination was highly congruent (WRKY12, 13, and 14) with two other published single gene phylogenies. The utility of WRKY loci for determining infraspecific relationships has been demonstrated by genetic mapping in *T. cacao* (Borrone et al., 2004) and by differentiating individuals from one another within *T. cacao* (Borrone, 2004) and *C. nucifera* germplasm collections (Mauro-Herrera et al., 2006). Thus, the WRKY gene family is a potentially rich source of unlinked, single and low copy nuclear genes for phylogeny reconstruction of plants at lower taxonomic levels. Moreover, they may be applicable for resolving infraspecific relationships, including hybridization, that are difficult to discern with more commonly used gene sequences.

References

- Addison, G., Tavares, R., 1952. Hybridization and grafting in species of *Theobroma* which occur in Amazonia. *Evolution* 6, 380–386.
- Alverson, W.S., Whitlock, B.A., Nyffeler, R., Bayer, C., Baum, D.A., 1999. Phylogeny of the core Malvales: evidence from *ndhF* sequence data. *Am. J. Bot.* 86, 1474–1486.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. GappedBLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Syst. Biol.* 46, 654–673.
- Bayer, C., Fay, M.F., De Bruijn, A.Y., Savolainen, V., Morton, C.M., Kubitzki, K., Alverson, W.S., Chase, M.W., 1999. Support for an expanded family concept of Malvaceae within a recircumscribed order Malvales: a combined analysis of plastid *atpB* and *rbcL* DNA sequences. *Botan. J. Linnean Soc.* 129, 267–303.
- Borrone, J.W., 2004. The isolation, characterization, and application of WRKY genes as useful molecular markers in tropical trees. Ph.D. dissertation.
- Borrone, J.W., Kuhn, D.N., Schnell, R.J., 2004. Isolation, characterization, and development of WRKY genes as useful genetic markers in *Theobroma cacao*. *Theor. Appl. Genet.* 109, 495–507.
- Bowers, J.E., Chapman, B.A., Rong, J., Paterson, A.H., 2003. Unraveling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 422, 433–438.
- Burger, D., 1972. Seedlings of some tropical trees and shrubs mainly of South East Asia. Plenum Press, New York.
- Bremer, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42, 198–213.
- Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., May, G., 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4, 10. doi:10.1186/1471-2229-4-10.
- Cronn, R.C., Small, R.L., Haselkorn, T., Wendel, J.F., 2002. Rapid diversification of the cotton genus (*Gossypium*: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am. J. Bot.* 89, 707–725.

- Cronn, R.C., Small, R.L., Haselkorn, T., Wendel, J.F., 2003. Cryptic repeated genomic recombination during speciation in *Gossypium gossypoides*. *Evolution* 57, 2475–2489.
- Cuatrecasas, J., 1964. Cacao and its allies: a taxonomic revision of the genus *Theobroma*. *Contrib. U.S. Natl. Herb.* 35, 379–614.
- Doebly, J., Lukens, L., 1998. Transcriptional regulators and the evolution of plant form. *The Plant Cell* 10, 1075–1082.
- 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.
- Eulgem, T., Rushton, P.J., Robatzek, S., Somssich, I.E., 2000. The WRKY superfamily of plant transcription factors. *Trends Plant Sci.* 5, 199–206.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1994. Testing significance of incongruence. *Cladistics* 10, 315–319.
- Farris, J.S., Källersjö, M., Kluge, A.G., Bult, C., 1995. Constructing a significance test for incongruence. *Syst. Biol.* 44, 570–572.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Figueira, A., Janick, J., Goldbrough, P.B., 1994. Re-examining the classification of *Theobroma cacao* L. using molecular markers. *J. Am. Soc. Hort. Sci.* 119, 1073–1082.
- Fitch, W.M., 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst. Zool.* 20, 406–416.
- Garcia, D., Gerald, J.N.F., Berger, F., 2005. Maternal control of integument cell elongation and zygotic control of endosperm growth are coordinated to determine seed size in *Arabidopsis*. *Plant Cell* 17, 52–60.
- Gaut, B.S., 1998. Molecular clocks and nucleotide substitution rates in higher plants. *Evol. Biol.* 30, 93–120.
- Grover, C.E., Kim, H., Wing, R.A., Paterson, A.H., Wendel, J.F., 2004. Incongruent patterns of local and global genome size evolution in cotton. *Genome Res.* 14, 1474–1482.
- Gutiérrez, R.A., Green, P.J., Keegstra, K., Ohlrogge, J.B., 2004. Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? *Genome Biol.* 5, R53.
- Hendy, M.D., Penny, D., 1982. Branch and bound algorithms to determine minimal evolutionary trees. *Math. Biosci.* 59, 277–290.
- Hughes, C.E., Eastwood, R.J., Bailey, C.D., 2006. From feast to famine? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Phil. Trans. Soc. B* 361, 211–225.
- Lespinet, O., Wolf, Y.I., Koonin, E.V., Aravind, L., 2002. The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res.* 12, 1048–1059.
- Luo, M., Dennis, E.S., Berger, F., Peacock, W.J., Chaudhury, A., 2005. MINISEED3 (MINI3), a WRKY family gene, and HAIKU2 (IKU2), a leucine-rich repeat (LRR) KINASE gene, are regulators of seed size in *Arabidopsis*. *Proc. Natl. Acad. Sci. (USA)* 102, 17531–17536.
- Martin, A.P., Burg, T.M., 2002. Perils of paralogy: using HSP70 genes for inferring organismal phylogenies. *Syst. Biol.* 51, 570–587.
- Mauro-Herrera, M., Meerow, A.W., Borrone, J.W., Kuhn, D.N., Schnell, R.J., 2006. Ten informative markers developed from WRKY sequences in coconut (*Cocos nucifera*). *Mol. Ecol. Notes* 6, 904–906.
- Meerow, A.W., Snijman, D.A., 2001. Phylogeny of Amaryllidaceae tribe Amaryllideae based on nrDNA ITS sequences and morphology. *Am. J. Bot.* 88, 2321–2330.
- Meerow, A.W., Guy, C.L., Li, Q.-B., Clayton, J.R., 2002. Phylogeny of the tribe Hymenocallideae (Amaryllidaceae) based on morphology and molecular characters. *Ann. Missouri Bot. Gard.* 89, 400–413.
- Meerow, A.W., Lehmiller, D.J., Clayton, J.R., 2003. Phylogeny and biogeography of *Crinum* L. (Amaryllidaceae) inferred from nuclear and limited plastid non-coding DNA sequences. *Bot. J. Linn. Soc.* 141, 349–363.
- Meerow, A.W., Clayton, J.R., 2004. Generic relationships among the baccate-fruited Amaryllidaceae (tribe Haemantheae) inferred from plastid and nuclear non-coding DNA sequences. *Pl. Syst. Evol.* 244, 141–155.
- Meerow, A.W., Francisco-Ortega, J., Kuhn, D.N., Schnell, R.J., 2006. Phylogenetic relationships and biogeography within the Eurasian clade of Amaryllidaceae based on plastid *ndhF* and nrDNA ITS sequences: lineage sorting in a reticulate area? *Syst. Bot.* 31, 42–60.
- Nishiyama, T., Fujita, T., Shin-I, T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., Kohara, Y., Hasebe, M., 2003. Comparative genomics of *Physcomitrella patens* gametophytic transcriptome and *Arabidopsis thaliana*: implication for land plant evolution. *Proc. Natl. Acad. Sci. (USA)* 100, 8007–8012.
- Planet, P.J., 2006. Tree disagreement: measuring and testing incongruence in phylogenies. *J. Biomed. Inform.* 39, 86–102.
- Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., Yu, G.-L., 2000. *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290, 2105–2110.
- Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, 798–804.
- Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., Schmidt, R., 2001. Comparative sequence analysis reveals extensive microlinearity in the *Lateral Suppressor* regions of the tomato, *Arabidopsis*, and *Capsella* genomes. *The Plant Cell* 13, 979–988.
- Schultes, R.E., 1958. A synopsis of the genus *Herrania*. *J. Arnold Arboretum* 39, 216–295.
- Shiu, S.-H., Shih, M.-C., Li, W.-H., 2005. Transcription factor families have much higher expansion rates in plants than animals. *Plant Physiol.* 139, 18–26.
- Silva, C.R.S., Figueira, A.V.O., Souza, E.C.A.S., 2001. Diversidade no gênero *Theobroma*. In: Dias, L.A.S. (Ed.), *Melhoramento genético do cacauero*. FUNAPE—UFG, Goiânia, pp. 49–80.
- Silva, C.R.S., Eastwood, R.J., Figueira, A., 2004. Description of Amazonian *Theobroma* L. collections, species identification, and characterization of interspecific hybrids. *Acta Bot. Bras.* 18, 333–341.
- Silva, C.R.S., Figueira, A., 2005. Phylogenetic analysis of *Theobroma* (Sterculiaceae) based on Kunitz-like trypsin inhibitor sequences. *Plant Syst. Evol.* 250, 93–104.
- Small, R.L., Cronn, R.C., Wendel, J.F., 2004. Use of nuclear genes for phylogenetic reconstruction in plants. *Aust. Syst. Bot.* 17, 145–170.
- Sorenson, M.D., 1996. *TreeRot*. University of Michigan, Ann Arbor.
- Swofford, D.L., 2003. *PAUP**. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Thomas, B.C., Pedersen, B., Freeling, M., 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.* 16, 934–946.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucl. Acids Res.* 24, 4876–4882.
- Ulker, B., Somssich, I.E., 2004. WRKY transcription factors: from DNA binding towards biological function. *Curr. Opin. Plant Biol.* 7, 491–498.
- Whitlock, B.A., Baum, D.A., 1999. Phylogenetic relationships of *Theobroma* and *Herrania* (Sterculiaceae) based on sequences of the nuclear gene *Vicilin*. *Syst. Bot.* 24, 128–138.
- Whitlock, B.A., Bayer, C., Baum, D.A., 2001. Phylogenetic relationships and floral evolution of the Byttnerioideae (“Sterculiaceae” or Malvaceae s.l.) based on sequences of the chloroplast gene, *ndhF*. *Syst. Bot.* 26, 420–437.
- Wolfe, K.H., Li, W.-H., Sharp, P.M., 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. (USA)* 84, 9054–9058.

- Wu, K.-L., Guo, Z.-J., Wang, H.-H., Li, J., 2005. The WRKY family of transcription factors in rice and *Arabidopsis* and their origins. *DNA Res.* 12, 9–26.
- Xie, Z., Zhang, Z.-L., Zou, X., Huang, J., Ruas, P., Thompson, D., Shen, Q.J., 2005. Annotations and functional analyses of the rice WRKY gene superfamily reveal positive and negative regulators of abscisic acid signaling in aleurone cells. *Plant Physiol.* 137, 176–189.
- Xiong, Y., Liu, T., Sun, S., Li, J., Chen, M., 2005. Transcription factors in rice: a genome-wide comparative analysis between monocots and eudicots. *Plant Mol. Biol.* 59, 191–203.
- Zhang, Y., Wang, L., 2005. The WRKY transcription factor superfamily: its origin in eukaryotes and expansion in plants. *BMC Evol. Biol.* 5, 1. doi:10.1186/1471-2148-5-1.