Theor Appl Genet (2003) 106:375–383
DOI 10.1007/s00122-002-0999-9

Shiv Shankhar Kaundun · Satoru Matsumoto

Development of CAPS markers based on three key genes of the phenylpropanoid pathway in Tea, *Camellia sinensis* (L.) O. Kuntze, and differentiation between *assamica* and *sinensis* varieties

Received: 5 December 2001 / Accepted: 21 May 2002 / Published online: 17 July 2002
© Springer-Verlag 2002

Abstract The genetic diversity of tea, *Camellia sinensis* (L.) O. Kuntze, including the two main cultivated *sinensis* and *assamica* varieties, was investigated based on PCR-RFLP analysis of PAL, CHS2 and DFR, three key genes involved in catechin and tannin synthesis and directly responsible for tea taste and quality. Polymorphisms were of two types: amplicon length polymorphism (ALP) due to the presence of indels in two introns of PAL and DFR, and point mutations detected after restriction of amplified fragments with appropriate enzymes. A progeny test showed that all markers segregated in a Mendelian fashion and that polymorphisms were exclusively co-dominant. CHS2, which belongs to a multi-gene family, allowed for greater variation than the single-copy PAL gene. Based on Nei's gene diversity index, var. *sinensis* was revealed to be more variable than var. *assamica*, and that a higher proportion of overall diversity resided within varieties as compared to between varieties. Even though no specific DNA profile was found for either tea varieties following any single PCR-RFLP analysis, a factorial correspondence analysis carried out on all genotypes and markers separated the tea samples into two distinct groups according to their varietal status. This reflects the large difference between var. *sinensis* and var. *assamica* in their polyphenolic profiles. The STS-based markers developed in this study will be very useful in future mapping, population genetics and fingerprinting studies of this important crop species and other *Camellia* species, as the primers have also proven successful in the three other subgenera of this genus.

Communicated by H.C. Becker

S.S. Kaundun (✉) · S. Matsumoto
National Institute of Vegetable and Tea Science, 2769 Kanaya, Kanaya-cho, 428-8501 Shizuoka, Japan
e-mail: dkaundun@yahoo.com

Present address:
S.S. Kaundun, Lead Characterisation, Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire RG42 6EY, United Kingdom

Keywords Tea · *Camellia sinensis* · CAPS markers · Genetic diversity

Introduction

Tea, *Camellia sinensis* (L.) O. Kuntze, is indigenous to the whole South-East Asia Monsoon region, but has long found its way as an important crop species in many countries. Two basic varieties can be distinguished: var. *assamica*, a large-leaf, tall and quick-growing tree well suited to very warm tropical climates, and var. *sinensis*, the small leaf, slower-growing bush that can withstand colder climates (Wight 1962; Banerjee 1992). Tea is, however, highly outcrossing and clinal intermediates that are difficult to classify are often encountered in a wider range of conditions. Tea cultivation dates back to several centuries in most countries resulting in severe erosion of its genetic base over time. This is due to intensive selection and breeding for desirable traits and the absence of proper conservation programs in the main producing countries. In order to limit further reduction in its gene pool and to breed for new tea types that are more productive, less prone to diseases and cold, as well as satisfying consumer needs for new tea tastes, a perfect knowledge of the existing genetic diversity is a pre-requisite. Despite the agricultural and economic importance of tea in many Asian and African countries, breeding and conservation programs are still based on conventional morphological and agronomic descriptors. These are very dependent on environmental and developmental factors and, thus, poorly reflect the genetic base of the gene pool (Green 1971; Wikremaratne 1981).

Recently, several DNA markers that circumvent these above-mentioned problems have been described. RFLP markers have initially been used and have provided a wealth of information for several crop species (Tanksley et al. 1989; Zhang et al. 1992). This approach, however, requires appreciable amounts of relatively pure DNA, is time consuming, costly and is technically demanding. RAPD and AFLP, on the other hand, have the merit of re-

vealing diversity without prior knowledge of genome sequence information but have the drawbacks of generating mainly dominant and anonymous markers (Skroch and Nienhuis 1995; Shim and Jorgensen 2000). In addition RAPD has proved to be sensitive to experimental conditions and its reproducibility has been questioned (Prenner et al. 1993). SSR markers have also been extensively used to generate multi-allelic and very variable markers but have not yet been developed in tea (Saghai-Maroof et al. 1994; Diwan and Cregan 1997). Cleaved amplified polymorphic sequence (CAPS) or PCR-RFLP methodology generates another type of molecular marker combining both PCR and RFLP techniques. This requires minute amounts of genomic DNA and simple electrophoretic systems to reveal polymorphism. The only drawback is that sequence information is needed to tag the desired DNA fragment. CAPS markers have been successfully applied to a number of crop and forest species for which extensive nucleotide information is available (Chen et al. 1994; Ghareyazie et al. 1995; Tsumura et al. 1997; Perry et al. 1999). They have a definite advantage over other markers as they have often been developed from mapped cDNA clones that represent expressed genes.

In tea, three important genes, phenylalanine ammonia-lyase (PAL), chalcone synthase (CHS) and dihydroflavonol 4-reductase (DFR) involved in the phenylpropanoid pathway, have been previously cloned (Matsumoto et al. 1994; Takeuchi et al. 1994). PAL was reported as being encoded as a single-copy gene while DFR and CHS were found to exist in two and three copies respectively. PAL is ubiquitous in vascular plants and link primary and secondary metabolism by catalyzing the conversion of L-phenylanaline into cinamic acid, the initial substrate of phenylpropanoid metabolism. CHS also intervenes early in flavonoid biosynthesis while DFR occurs at a later stage prior to the synthesis of cathechin, a group of secondary metabolites that play an important role in tea taste and quality.

With a view to develop new and informative markers for tea, the CAPS methodology was employed to detect polymorphism in phenylalanine ammonia-lyase (PAL), chalcone synthase (CHS) and dihydroflavonol 4-reductase (DFR) genes, using both var. *assamica* and var. *sinensis* tea genotypes.

Materials and methods

Plant material

Fifty tea samples collected from both cultivated and abandoned plantations were used to detect polymorphism in the three genes. Twenty four samples belonged to the var. *sinensis*, of which 10, 6 and 8 originated from Japan, China and Korea respectively. The Japanese subsample included the predominantly cultivated Yabukita cultivar as well as Yutakamidori, Kuritawase, Asagiri, Saemidori and Surugawase.

Twenty four plants belonging to the *assamica* variety originated from Sri Lanka (6), Vietnam (6) and Myanmar (12). Two other individuals resulting from a cross between a *sinensis* genotype and an *assamica* × *sinensis* hybrid were also included.

To test the inheritance of the markers, 48 progenies obtained from a cross between two Japanese tea cultivars, Saemidori × Izumi were also analyzed.

Three samples from *Camellia granthamiana*, *Camellia sasanqua* and *Camellia fraterna*, respectively belonging to the sub-genera, Protocamellia, Metacamellia and Camellia of the genus Camellia were also used to investigate cross-species transferability of the STS primers.

DNA isolation

Total DNA was extracted from fresh or frozen tissue using a slightly modified version of the CTAB method (Sul and Korban 1996). Approximately 200 mg of leaf tissue were ground to a fine powder in liquid nitrogen with a pestle and mortar. The pulverized plant material was transferred to a microtube and 600 µl of preheated 2 × CTAB (65 °C) solution was added. The tube was vortexed for a few seconds and incubated at 65 °C for 10 min. Following incubation, 600 µl of chloroform/isoamyl alcohol (24:1) was added and the mixture was shaken vigorously. The extract was centrifuged for 10 s at about 15,000 g and the supernatant transferred to a new microtube. The chloroform/isoamyl extraction step was repeated two times more and 1 ml of 100% cold ethanol was added to the final supernatant. A DNA pellet became visible upon gentle swirling. It was transferred to a new tube, rinsed two times with 70% ethanol and dissolved in 400 µl of 10 mM TE buffer solution containing 0.001 µg/l of RNASE A. DNA concentrations were estimated by agarose-gel analysis stained with ethidium bromide.

Primer design

PCR primers were designed with the Genetix software (Tokyo) and based on cDNA nucleotide sequences of PAL, CHS2 and DFR in tea (cultivar: Yabukita). Putative intron sites for each the three genes were deduced after alignment with corresponding published sequences from other plant species. Prior to thermocycling, a simulation of the PCR reaction was carried out with the Amplified software v1.3 [(Engels 1993); http://www.wisc.edu/genetics/CATG/amplify/index.html]. A schematic representation of the primer locations is shown in Fig. 1. All the primers were synthesized by the Grainer company (Japan).

PCR and restriction product analysis

The polymerase chain reaction was carried out in a 25-µl volume containing 50–100 ng of genomic DNA, 0.5 µM of each primer, 200 µM of each dNTP, 2 mM of $MgCl_2$, 50 mM of KCl, 500 ng of BSA, 10 mM of Tris-HCl, pH 8.3, and 0.5 units of *Taq* polymerase (Takara, Japan). PCR was performed on a Perkin-Elmer thermocycler model 2400 programmed for an initial denaturation step of 94 °C 5 min followed by 35 cycles for 30 s at 94 °C, 30 s at 60 °C and for 1.5 min at 72 °C. A final elongation step for 7 min at 72 °C was included. This basic cycling procedure was used for all DNA fragments except for the longest fragment DFR intron 1 + 2, for which the extension time was increased to 2 min. The PCR products were separated on a 2% agarose gel and visualized under UV to reveal Amplicon Length Polymorphism (ALP) and the Cleaved Amplified Polymorphic Sequence (CAPS). In the latter case, 2 µl of PCR products were digested with 32 different restriction endonucleases, namely *Dde*I, *Eco*RI, *Eco*RV, *Hae*III, *Hap*II, *Hind*III, *Hin*fI, *Hha*I, *Mbo*I, *Nco*I, *Rsa*I, *Alu*I, *Bam*HI, *Bgl*II, *Pst*I, *Pvu*II, *Scrf*I, *Spe*I, *Taq*I, *Xho*I, *Nla*III, *Eco*T22I, *Xmn*I, *Cfr13*I *Hinc*II, *Bsp*HI, *Dra*I, *Ban*II, *Bgl*I, *Nsp*I, *Sal*I and *Apa*I following the manufacturers' recommendations, and separated on either 2% agarose or a 3.5% metaphor gel when smaller polymorphic bands were involved. In all cases, both the gels and the buffer (1 × TBE) contained 0.5 µg/ml of EtBr.

Cloning and Sequencing of amplification products

To confirm the identity of the PCR bands generated by each of the seven primer pairs, the corresponding amplification products from Yabukita were ligated into pGEM-T easy vector (Promega) and the nucleotide sequences of the inserts determined on a Perkin-Elmer ABI 373 automated sequencer using the dye primer or dye terminator kits (Perkin Elmer).

Data analysis

Popgene v3.1 (Yeh et al. 1997) available at (http://www.ualberta.ca/~fyeh/index.htm) was used to apportion diversity into within and between tea varieties. The markers were scored as diploid data and each allele was assigned an alphabet for a particular primer set and primer set/enzyme combination, for amplicon length polymorphism and PCR/RFLP, respectively. Diversity measures were based on Nei's (1973) gene diversity index H_e. H_e was calculated for two different levels of diversity: H_{var}, the average diversity within the *sinensis* and *assamica* varieties, and H_{sp}, the diversity within the species. The proportion of genetic diversity within varieties was deduced by H_{var}/H_{sp}, while that between varieties was given by $1 - H_{var}/H_{sp}$. Statitcf v5 (ITCF, Paris) was used to perform factorial correspondence analysis on a 50 (genotypes) × 31 (alleles) matrix. For this purpose, the presence of an allele was coded as 1 and its absence as 0 for both ALP and PCR/RFLP markers.

Results

This study was conducted with the aim of developing CAPS markers, and subsequently estimating the level of polymorphism and differentiation between *assamica* and *sinensis* tea types based on the PAL, CHS and DFR genes. PAL, CHS and DFR that have been cloned in other species, respectively show one, one and five introns at conserved positions. The primers were designed in such a way as to cover the maximum length of the corresponding genes, including both coding and non-coding regions for

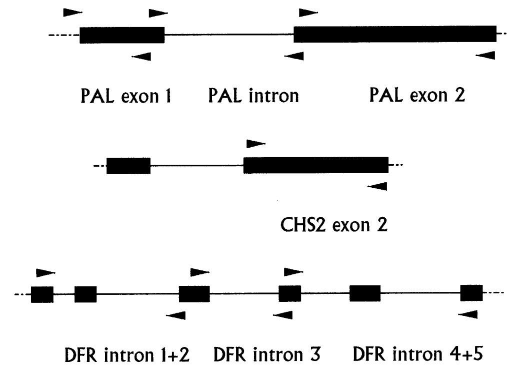

Fig. 1 Schematic representation of the structures of PAL, CHS and DFR, and primer locations

PAL and DFR. Three CHS genes having been described in tea showing between 85 and 90% homologies between them, a primer pair was designed aimed at targeting exon 2 of CHS2 only, as this region showed an appreciable sequence difference compared to CHS1 and CHS3. This was done in order to avoid amplification of multiple fragments which would in turn make scoring and diversity studies impossible. In this respect, once a primer pair was selected for CHS2, PCR reactions were simulated using the Amplify software to ensure that no corresponding PCR product would be generated with CHS1 and CHS3. As for PAL and DFR no such strategy was adopted as they were respectively found to exist as a single copy and two copies with 100% homologies in tea. Seven primer pairs in all were sufficient to amplify all the three genes (Fig. 1). Three primer pairs were used to tag PAL exon 1, PAL intron and PAL exon 2, respectively. One primer pair amplified CHS2 exon 2 while three other primer pairs were sufficient to target DFR intron 1 + 2, DFR intron 3 and DFR intron 4 + 5.

Table 1 Sequence and position of STS primers, PCR fragment size and informative enzymes

Fragment	Primer sequence (forward and reverse)	Location on cDNA[a] clones	Size (bp) in genomic DNA	Informative enzymes or ALPs
PAL exon 1	5′-TCCATCAATCTATACACCTACCTG-3′ 5′-CCTTCTTTGGTCCTCCTATGTGA-3′	3–28 499–477	497	*Hpa*II, *Nla*III
PAL intron	5′-CACATAGGAGGACCAAAGAAGG-3′ 5′-GGCAATGTAAGATAGGGGGACT-3′	478–499 764–743	1,818 Intron: 1,531	ALP
PAL exon 2	5′-AGTCCCCCTATCTTACATTGCC-3′ 5′-AACAGATAGGAAGAGGAGCACCATTC-3′	743–764 2,274–2,249	1,532	*Taq*I
CHS2 exon 2	5′-AAACCCAAATGTGTGTGCCTAC-3′ 5′-AGGATAAACAACACACAAGCGC-3′	308–328 1,108–1,037	801	*Rsa*I, *Hae*III, *Eco*RI, *Bsp*HI
DFR intron 1 + 2	5′-CGCGGCTATATTGTTCGTGCA-3′ 5′-GTTGATTGTCGGCTTGATTACC-3′	222–242 455–434	3,262 Intron1: 126; Intron 2: 2,902	ALP
DFR intron 3	5′-CCAGGAACACCAACAACCCGT-3′ 5′-CCATGCTGCTTTCTCTGCCAA-3′	542–561 665–645	958 Intron3: 834	*Hin*dIII
DFR intron 4 + 5	5′-AACATTCCCACCAAGCCTAATC-3′ 5′-ATGAGAACGACACAACTGGCAA-3′	740–761 1,020–999	1,365 Intron4: 101; Intron5: 983	*Hpa*II, *Nla*III, *Xmn*I

[a] Based on cDNA sequences of *C. sinensis* (cultivar: Yabukita) of PAL (Accession: D26596), CHS2 (Accession: D26594) and DFR (Accession: AB018685) in GenBank

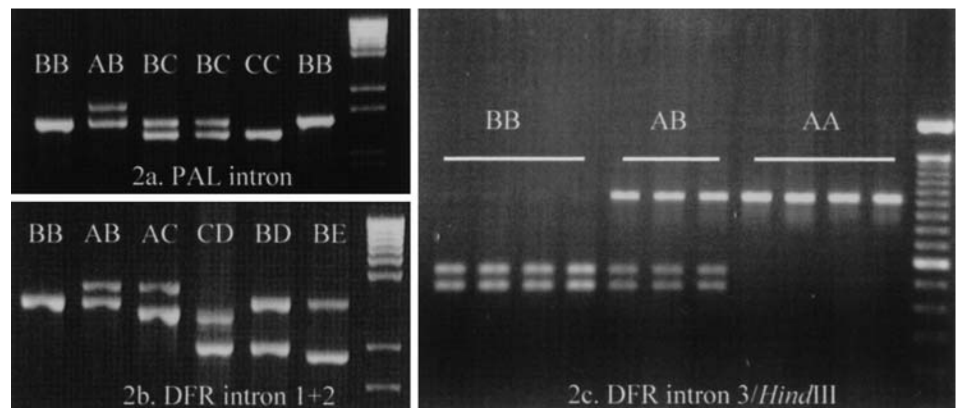

Fig. 2 Amplified length polymorphisms observed with PAL intron (2a) and DFR intron 1 + 2 (2b); PCR/RFLP markers obtained with DFR intron 3 digested by *Hin*dIII (2c): homozygote for a polymorphic restriction site (BB), homozygote for the absence of the site (AA) and a heterozygote state (AB) with only one of the two alleles being cut

Table 2 Segregation of polymorphic alleles in 48 progenies from the cross Saemidori × Izumi

STS/RFLP	Parent genotypes	Genotypic frequencies	Chi^2 value	Significance[a] (P = 0.05)
PAL exon 1/*Hpa*II	AB × AB	AA/AB/BB 15/23/10	1.33	NS
PAL exon 1/*Nla*III	BB × BB	BB 48	0	NS
PAL intron (ALP)	AB × AB	AA/AB/BB 12/26/10	0.50	NS
PAL exon 2/*Taq*I	BB × BB	BB 48	0	NS
CHS2 exon 2/*Rsa*I	BB × AB	AB/BB 26/22	0.33	NS
CHS2 exon 2/*Bsp*HII	AA × BB	AB 48	0	NS
CHS exon 2/*Hae*III	BB × BB	BB 48	0	NS
CHS exon 2/*Eco*RI	BB × BB	BB 48	0	NS
DFR intron 1 + 2 (ALP)	AB × BC	AB/BB/AC/BC 13/11/14/10	0.83	NS
DFR intron 3/*Hin*dIII	AB × AB	AA/AB/BB 11/24/13	0.17	NS
DFR intron 4 + 5/*Xmn*I	BB × BB	BB 48	0	NS
DFR intron 4 + 5/*Hpa*II	AB × AB	AA/AB/BB 11/23/14	0.46	NS
DFR intron 4 + 5/*Nla*III	AB × BB	AB/BB 27/21	0.75	NS

[a] NS: non significant

The primers amplifying the exons generated bands corresponding to the expected sizes as deduced from cDNA sequences, while those flanking the introns were greater in length (Table 1). Sequencing reactions carried out on the seven amplification products for cultivar Yabukita confirmed that they corresponded to the targeted genes. With the exception of DFR intron 4, found 20 bp further downstream in tea, all the other intron positions were the same as those predicted by a homology search with corresponding genes in other species. Notable differences were also observed with DFR intron 2 whose length is 2,902 bp in Yabukita while being less than 300 bp in *Arabidopsis thaliana* and *Ipomea nil*. The other PAL and DFR intron sizes for Yabukita were in the range described in other species.

Five of the seven primer pairs produced a single band while those flanking the PAL intron and DFR intron 1 + 2 gave one or two bands with a total of three and five amplicon-length polymorphisms respectively (Fig. 2a and b). For DFR intron 1 + 2, the amplicon length polymorphism was located in intron 2 only as confirmed by targeting the two introns separately with two different pairs of primers (data not shown). Moreover, amplification was also obtained at all loci with the three Camellia samples belonging to three other subgenera of the genus, thus demonstrating the complete transferability of the STS primers. The band sizes of the three other Camellia samples were comparable to those observed in tea, but with a relatively lower fragment length for the PAL intron. These three samples

were not included in a further search of CAPS polymorphism in tea.

In the absence of ALP, the PCR products were digested with a set of 32 restriction enzymes (Table 1). Four base pair cutters were the most successful at revealing polymorphism among the samples due to their higher number of recognition sites. The higher proportions of AT in the introns were responsible for multiple restriction sites for enzymes with AT in their recognition sites, while most of those with GC sequences were more suitable for exons. Characteristic DNA patterns obtained with DFR intron 3 digested by *Hin*dIII are shown on Fig. 2c. Three types of profiles could be observed: a homozygote for a restriction site, a homozygote for the absence of the site and a heterozygous state due to the co-dominant nature of the polymorphism. An exception to this rule was observed when PAL exon 1 was digested with *Hpa*II. This enzyme cut the PAL exon 1 fragment at two positions, both of which were polymorphic. Only three alleles were however observed, generating four genotypes only, namely the AA, AB, BB and CC genotypes. PAL exon 1 was also found to be polymorphic with *Nla*III, while PAL exon 2, though greater in length, showed only one polymorphic site with *Taq*I. CHS2 exon 2 was by far the most variable with four polymorphic sites in all.

Table 3 Allele frequencies at three phenylpropanoid gene loci in *assamica* and *sinensis* varieties

Locus	Allele	var. *sinensis*	var. *assamica*
PAL exon 1/*Hpa*II	a	0.5208	0.0208
	b	0.4792	0.9375
	c	0.0000	0.0417
PAL exon 1/*Nla*III	a	0.4583	0.6458
	b	0.5417	0.3542
PAL intron (ALP)	a	0.8125	0.7917
	b	0.1667	0.1667
	c	0.0208	0.0417
PAL exon 2/*Taq*I	a	0.5417	0.9167
	b	0.4583	0.0833
CHS2 exon 2/*Rsa*I	a	0.3750	0.0625
	b	0.6250	0.9375
CHS2 exon 2/*Bsp*HII	a	0.4167	0.0833
	b	0.5833	0.9167
CHS2 exon 2/*Hae*III	a	0.0000	0.1667
	b	1.0000	0.8333
CHS2 exon 2/*Eco*RI	a	0.0000	0.0625
	b	1.0000	0.9375
DFR intron 1 + 2 (ALP)	a	0.3125	0.0000
	b	0.6250	0.6042
	c	0.0000	0.0625
	d	0.0417	0.2500
	e	0.0208	0.0833
DFR intron 3/*Hin*dIII	a	0.7083	0.1042
	b	0.2917	0.8958
DFR intron 4 + 5/*Xmn*I	a	0.0000	0.7292
	b	1.0000	0.2708
DFR intron 4 + 5/*Hpa*II	a	0.3750	0.0833
	b	0.6250	0.9167
DFR intron 4 + 5/*Nla*III	a	0.3125	0.0000
	b	0.6875	1.0000

In order to test the inheritance of the markers, we examined the progenies of a cross between Saemidori and Izumi, two local tea cultivars in Japan (Table 2). In this respect the PAL intron loci showed a 1:2:1 segregation for which both parents were heterozygotes. On the other hand, a 1:1 segregation was obtained with one parent being a heterozygote and the other a homozygote as in the case of CHS2 exon 2 digested with *Rsa*I. In general, the alleles segregated in a Mendelian fashion at all informative loci and were consistent with the parental genotypes. Therefore the markers represented different alleles of a unique locus rather than different loci. They could subsequently be used in total confidence at estimating the different diversity parameters within and between *sinensis* and *assamica* teas.

The basic statistics given by Popgene reveals that 12 and ten of the 13 loci were repectively polymorphic in the *assamica* and *sinensis* subgroups (Table 3). Two alleles which were fixed in *sinensis* tea, namely CHS2 exon 2 digested by *Hae*III and *Eco*RI, were, however, present at low frequencies in the *assamica* subgroup. It should also be noted that only four alleles were detected for DFR intron 1 + 2 in both *assamica* and *sinensis* teas, with a predominance of allele 'B' in the two groups and the absence of allele 'C' and 'A' in *sinensis* and *assamica* teas respectively. The diversity measures determined by the Nei's (1973) gene diversity indices

Table 4 Apportionment of genetic diversity within and between *sinensis* and *assamica* tea varieties

Item	H_0*sinensis*	H_0*assamica*	H_{var}	H_{sp}	H_{var}/H_{sp}	$1 - H_{var}/H_{sp}$
PAL exon 1/*Hpa*II	0.499	0.119	0.309	0.425	0.728	0.272
PAL exon 1/*Nla*III	0.497	0.458	0.477	0.495	0.964	0.036
PAL intron (ALP)	0.312	0.344	0.328	0.328	0.999	0.001
PAL exon 2/*Taq*I	0.497	0.153	0.325	0.395	0.822	0.178
CHS2 exon 2/*Rsa*I	0.469	0.117	0.293	0.342	0.857	0.143
CHS2 exon 2/*Bsp*HII	0.486	0.153	0.319	0.375	0.852	0.148
CHS exon 2/*Hae*III	0.000	0.278	0.139	0.153	0.909	0.091
CHS exon 2/*Eco*RI	0.000	0.117	0.059	0.061	0.969	0.031
DFR intron 1 + 2 (ALP)	0.510	0.562	0.536	0.573	0.935	0.065
DFR intron 3/*Hin*dIII	0.413	0.187	0.300	0.482	0.622	0.378
DFR intron 4 + 5/*Xmn*I	0.000	0.395	0.198	0.463	0.426	0.574
DFR intron 4 + 5/*Hpa*II	0.469	0.153	0.311	0.353	0.880	0.120
DFR intron 4 + 5/*Nla*III	0.430	0.000	0.215	0.264	0.815	0.185
Mean	0.352	0.233	0.293	0.362	0.829	0.171

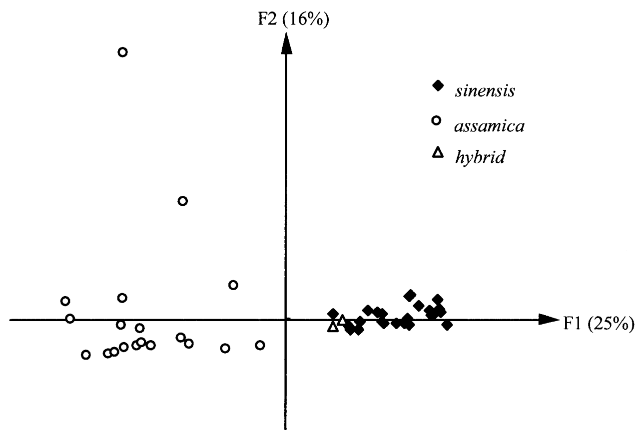

Fig. 3 Principal plane (F1 × F2) of a factorial correspondence analysis based on 50 tea genotypes and 31 polymorphic alleles

and genotypic frequencies varied from zero at three loci for *sinensis* tea and one locus for *assamica* tea, to 0.562 and 0.510 for DFR intron 1 + 2 in *assamica* and *sinensis* teas respectively (Table 4). The average unbiased heterozygosities were higher for *sinensis* (H_e = 0.352) as compared to *assamica* (H_e = 0.233) teas. This is due to the fact that even though more loci were polymorphic in the *assamica* subgroup, the heterozygosity values were in many cases relatively lower, while higher in *sinensis* teas. Partitioning the diversity into within (H_{var}/H_{sp}) as compared to between ($1 - H_{var}/H_{sp}$) tea groups revealed that 83% of the total diversity resided within as compared to 17% between groups. The loci that were the most discriminating between the two varieties was DFR intron 4 + 5 digested by *Xmn*I followed by PAL exon 1 cut by *Hpa*II, and the least discerning, the ALPs of PAL intron and CHS2 exon 2 cut by *Eco*RI.

In a view of obtaining additional information regarding the genetic relationships between the two main cultivated teas we analyzed the samples at an individual level through a factorial correspondence analysis (Fig. 3). This multivariate analysis was carried out on a 50 × 31 (genotypes × alleles) data matrix including the two hybrids between *assamica* and *sinensis* teas. The first factorial plane accounted for 41% of total variability with a complete separation of *assamica* from *sinensis* teas. The two hybrid teas grouped closer with the *sinensis* group, though at the extremity facing *assamica* samples. The fact that *assamica* teas appeared more scattered on the FCA plot is not in contradiction with the heterozygosity values, and do not imply that they are more diverse than *sinensis* tea. This is because the first plane of the multivariate analysis explains only a portion of the total diversity featuring *assamica* tea as diverse samples, but the contrary is true when other planes are taken into consideration.

Discussion

PAL, CHS and DFR are three key genes that play an important role in tea taste and quality through catechin synthesis, a main component in tea astringency. The activities of PAL, CHS and DFR and catechin concentrations have been found to increase with increasing light intensities (Annan and Nakagawa 1974; Takeuchi et al. 1994). Catechin occurs in higher concentration in young leaves and in var. *assamica* as compared to older leaves and var. *sinensis* respectively (Iwasa 1977). Variations in catechin contents have also been detected among Japanese tea cultivars used in different tea products (Maeda-Yamamoto et al. 2001). It was thus legitimate to determine the genetic relationships within and between the two main cultivated *assamica* and *sinensis* teas based on polymorphism revealed in these three genes. For this purpose we have employed the CAPS methodology which is cheap, easy to develop and very reproducible.

The most important factor in the establishment of CAPS markers is the careful choice of PCR primers targeting a single locus at a time, in order to ensure reliable results. This was facilitated here because PAL and DFR respectively exist in a single copy and two copies in tea. CHS is present in three copies but with sufficient sequence difference among them, allowing specific amplification of CHS2 only. Primers could also be designed at the 5′-untranslated and variable region of CHS, but with the risk of low PCR amplification or failure due to polymorphism at priming sites among tea samples within or between varieties. The second important factor increasing the chance of revealing polymorphism is the choice of restriction enzymes due to the relatively short length and sometimes conserved nature of the amplification products. In this respect, four base cutters should be preferred to six base-pair cutters, and restriction enzymes recognizing AT and GC sequences should be dedicated to introns and exons respectively. The percentage of AT was as high as 73%, 70% and 66% for the PAL intron and the second and fourth introns of DFR in Yabukita respectively, and is in agreement with high percentages of purines in introns generally.

Polymorphisms were of two types: readily observable following PCR due to mutation deletions at intron sites and, most frequently, after restriction with specific restriction enzymes within both introns and exons. Appreciable length differences existed between the PAL ALPs and could easily be allocated to one of the three detected alleles. On the other hand, although the DFR intron 1 + 2 locus appeared very polymorphic with several co-dominant alleles, its use may be limited as they differed only slightly in size, thus making scoring difficult. That the different bands were, nevertheless, alleles of the same gene was confirmed by digesting them with several restriction enzymes. The electrophoregram of the restriction products revealed that the different samples shared common fragments and differed from each other by the corresponding band length difference observed between them following PCR only. Suitable enzymes for this purpose were *Hae*III for DFR intron 1 + 2 and *Bsp*HI for the PAL intron. Confirmation of allelism at most loci was also obtained through inheritance analysis, in which case all informative markers showing polymorphism among the parents segregated in a ratio as predicted by the Mendelian law of segregation.

The STS primers designed for *C. sinensis* directed amplification in three species; *C. granthamiana*, *C. sasanqua* and *C. fraterna* which respectively belong to the three other subgenera Protocamellia, Camellia and Metacamellia of the genus Camellia. The transferability of STS markers can easily be explained by the fact that the DNA sequences for expressed genes are highly conserved and, thus, can be efficiently used at low cost in close species as well. In previous studies, STS primers developed in black spruce could be used in Norway spruce (Perry et al. 1999), while those in *Cryptomeria japonica* successfully amplified corresponding DNAs from other genera in a proportion matching the phylogenetic relationship with the pioneer species (Tsumura et al. 1997).

PAL exon 1 appeared more variable than PAL exon 2 with as many as three polymorphic sites for a fragment length of less than 500 bp, while PAL exon 2 exhibited a single polymorphic site, though more than 1,500 bp in length. This appears to be a characteristic of the PAL gene with a highly conserved second exon and a relatively more variable exon1 (Wanner et al. 1995). It is also noteworthy that if restriction sites are normalized, and the number expressed per kilobase of PCR products analyzed, then the coding regions of PAL would appear less polymorphic than that of CHS2. This is to be expected as CHS belongs to a small multi-gene family in tea which can vary more than the single-copy PAL gene in tea (Matsumoto et al. 1994).

A larger number of loci were found to be polymorphic among *assamica* samples as compared to that of *sinensis*. The allelic frequency and heterozygosities at several loci were, however, low in *assamica* tea and higher in *sinensis* tea. Overall, China tea was found to be more variable than Assam tea. This result is in agreement with two former studies, using RAPD (Wachira et al. 1995) and AFLP (Paul et al. 1997) markers, though the origins of the tea samples used were not the same. The higher level of diversity within *sinensis* tea would support the opinion that the center of origin of tea is the Southern region of China, and that migration and isolation has generated Assam tea in the recent past rather than the other way round (Hasimoto and Takasi 1978). However, a larger number of samples need to be collected and analyzed with additional classes of molecular markers in order to confirm this result.

When diversity was apportioned into within and between varieties, a predominantly higher proportion of diversity (83%) was found within varieties as compared to between varieties. This is a characteristic of highly outcrossing species, maintaining high diversity within groups as compared to between-groups diversity (Hamrick 1990). Also, *assamica* and *sinensis* teas can cross-hybridize very easily and over time it is thought that pure tea archetypes no longer exist (Visser 1969). Classification of many hybrids and introgressants are determined based on their morphological proximity with either varieties, and it would not be surprising that some of the samples considered as pure *assamica* genotypes in this study contain a percentage of the *sinensis* genome and vice versa.

Even though no characteristic DNA profile was obtained for either *sinensis* or *assamica* teas after PCR or PCR-RFLP at any loci, a multivariate analysis based on all samples and markers clearly separated these tea types into two groups. This finding is in accordance with anonymous DNA markers (Wachira et al. 1995; Paul et al. 1997) and chloroplast DNA sequences (Kato 2001). It also highlights the overall physiological and chemical, namely polyphenolic, differences between these two tea varieties. The two hybrid teas included in this study clustered with the *sinensis* subgroup due to the fact that three quarters of their genomes were of *sinensis* types as they resulted from a cross between a *sinensis* tea with a *sinensis* × *assamica* tea. These hybrids contain larger amounts of catechins and tannins, and are used to produce semi-fermented and fermented teas in Japan.

The most discriminating loci between *sinensis* and *assamica* teas is DFR intron 4 + 5 digested by *Xmn*I. This polymorphic site is located in the fifth intron of the DFR gene as deduced from the banding patterns following RFLP and nucleotide sequences in Yabukita. This is quite unexpected because it concerns a non-coding rather than a coding region. On a more general basis it would be interesting through sequencing reactions to investigate the nature of the different polymorphisms, especially at exons, in order to determine whether they cause a change in amino-acid sequence between genotypes or not. The biological value of classifying tea based on such polymorphism would be greatly increased and would be of particular interest in breeding for new tea tastes.

RFLP analysis using PAL cDNA as a probe has already been employed for the classification of Japanese tea cultivars (Matsumoto et al. 1994). Three different alleles were identified, the combination of which generated five genotypes in all. Yabukita the main tea cultivar in Japan was found to be a heterozygote with the BD genotype. In our study, Yabukita was also a heterozygote when PAL exon 1 was digested with *Hpa*II, but a homozygote for PAL intron and PAL exon 2 cut by *Taq*I. The conflicting results between classical RFLP and PCR-RFLP in some cases are easily explained by the fact that RFLPs screen polymorphism over a large region upstream of and downstream from the actual gene, whereas CAPS markers detect polymorphism within the gene itself.

The caffeine synthase and polyphenol oxidase genes, respectively directing the synthesis of caffeine and catalyzing the fermentation process in black tea, have recently been cloned in tea, and their cDNA sequences are readily available in the GenBank. These genes are as important as PAL, CHS and DFR regarding the organoleptic properties of tea. Using the same procedure described here, CAPS markers can easily be devised based on the polymorphism of these genes and contribute to clarifying the genetic relationships among and within the varieties of this poorly studied crop species. This work also opens the way for correlative analyses between newly established tea genotypes based on phenylpropanoid genes and chemotypes, mainly polyphenolic.

A first genetic map of *assamica* tea has been published (Hackett et al. 2000) and another one using a *sinensis* mapping population is in progress (Tanaka 2001). Both maps have been constructed using RAPD and AFLP markers. Though these two marker types are undoubtedly the most suited for mapping studies, requiring no prior genome information for marker development and having a wide genome coverage, they are however dominant and anonymous. The newly established CAPS markers are not only informative and may be linked to important QTLs associated with tea taste and quality, and due to their co-dominant and sometimes multi-allelic nature, they could be used with others to bridge the individual tea maps and produce a synthetic map with an increased number of markers.

Several studies aimed at investigating the genetic diversity in different tea germplasms have revealed a relatively low variability in each case (Wachira et al. 1995; Kaundun et al. 2000; Mewan et al. 2001) This is particularly true for Japanese tea with a leading Yabukita cultivar covering 75% of the total tea cultivated-area in the country. Yabukita has also been used as a parent in the breeding of a number of new cultivars, with a consequence of limiting tea types and tastes in Japan. The CAPS markers will be useful in screening for new tea genotypes from abandoned tea populations from Japan, Korea and China and, thus, facilitate enrichment of the existing germplasm collection. Since CAPS markers are PCR-based, requiring minute amounts of DNA template as well as being very reproducible, they may also be applied in fingerprinting tea and provide a means for discouraging the fraudulent commercialization of processed tea.

The development of STS-based markers may initially be time-consuming, as sequence information is required for primer design and sequencing reactions are necessary to confirm that the right DNA fragments have been amplified. In addition, a number of restriction enzymes have to be tested to reveal polymorphism. Nonetheless, once markers are established they can not only find wide application in mapping, population genetics and fingerprinting studies in the species in question, but also in a number of closely related genera also.

Acknowledgements The authors are grateful to Mr. A. Nesumi (NIVTS) for providing the plant material used in inheritance analysis and to Dr. N. Mizuno (NIVTS) for fruitful discussions about statistical calculations. The first author was supported by a fellowship and grant from the Science and Technology Agency (STA, Japan) for the period 1999–2001.

References

Annan N, Nakagawa M (1974) Effect of light on chemical constituents in the tea leaves. Nippon Nogei Kagaku Kaishi 48:91–96

Banerjee B (1992) Botanical classification of tea. In: Wilson KC, Clifford MN (eds) Tea: cultivation to consumption. Chapman and Hall, London, pp 25–51

Chen HB, Martin JM, Lanvin M, Talbert LE (1994) Genetic diversity in hard red spring wheat based on Sequence-Tagged-Site PCR markers. Crop Sci 34:1628–1632

Diwan N, Cregan PB (1997) Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. Theor Appl Genet 95:723–733

Engels WR (1993) Contributing software to the Internet: the Amplify program. Trends Biochem Sci 18:448–450

Ghareyazie B, Huang N, Second G, Khush GS (1995) Classification of rice germplasm. I. Analysis using ALP and PCR-based RFLP. Theor Appl Genet 91:218–227

Green MJ (1971) An evaluation of some criteria used in selecting large yielding tea clones. J Agric Sci 76:43–156

Hackett CA, Wachira FN, Paul S, Powell W, Waugh R (2000) Construction of a genetic linkage map for *Camellia sinensis* (tea). Heredity 85:4–355

Hamrick JL (1990) Isozymes and the analysis of genetic structure in plant populations. In: Soltis ED, Soltis PS (eds) Isozymes in plant biology. Chapman and Hall, London, pp 87–90

Hasimoto M, Takasi S (1978) Morphological studies on the Origin of the tea plant. V. A proposal of one place of origin by cluster analysis. Jap J Trop Agric 21:93–101

Iwasa K (1977) Biosynthesis of catechins in the tea plant. Bull Natl Res Tea 13:101–126

Kato M (2001) Analyses of differentiation of teas using DNA markers in evergreen forests. International Conference on O-Cha Culture and Science: Exploring New Possibilities for O-Cha (tea) in the 21st Century, Shizuoka, Japan, ICOS (in press)

Kaundun SS, Zhyvoloup A, Park YG (2000) Evaluation of the genetic diversity among elite tea (*Camellia sinensis* var. *sinensis*) genotypes using RAPD markers. Euphytica 115:7–16

Maeda-Yamamoto M, Sano M, Matsuda N, Miyase T, Kawamoto K, Suzuki N, Yoshimura M, Tachibana H, Hakamata K (2001) The change of epigallocatechin-3-*O*-(3-*O*-methyl) gallate content in tea of different varieties, tea seasons of crop and processing method. Nippon Shokuhin Kakaku Kaishi 48:64–68

Matsumoto S, Takeuchi A, Hayatsu M, Kondo S (1994) Molecular cloning of phenylalanine ammonia-lyase cDNA and classification of varieties and cultivars of tea plants (*Camellia sinensis*) using a tea Pal cDNA probe. Theor Appl Genet 89:671–675

Mewan K, Liyanage AKG, Jayamanne E, Fernando U, Karunanayke E (2001) Estimation of genetic diversity of tea (*Camellia sinensis* L.) genotypes in Sri Lanka using isozyme polymorphisms and RAPD markers. International Conference on O-Cha Culture and Science: Exploring New Possibilities for O-Cha (tea) in the 21st Century, Shizuoka, Japan, ICOS (in press).

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Paul S, Wachira FN, Powell W, Waugh R (1997) Diversity and genetic differentiation among populations of Indian and Kenyan tea [*Camellia sinensis* (L.) O. Kuntze] revealed by AFLP markers. Theor Appl Genet 94:255–263

Perry DJ, Isabel N, Bousquet J (1999) Sequence-tagged-site (STS) markers of arbitrary genes: the amount and nature of variation revealed in Norway spruce. Heredity 83:239–248

Prenner GA, Bush A, Wise R, Kim W, Dommier L, Kasha K, Laroche A, Scoles G, Molnar SJ, Fedak G (1993) Reproducibility of Random Amplified Polymorphic DNA (RAPD) analysis among laboratories. PCR Methods Applic 2:341–345

Saghai-Maroof MA, Biyahev RM, Yang GP, Zhang Q, Allard RW (1994) Extraordinarily polymorphic microsatellite DNA in barley: species diversity, chromosomal locations, and population dynamics. Proc Natl Acad Sci USA 91:5466–5470

Shim SI, Jorgensen RB (2000) Genetic structure in cultivated and wild carrots (*Daucus carota* L.) revealed by AFLP analysis. Theor Appl Genet 101:227–233

Skroch PW, Nienhuis J (1995) Qualitative and quantitative characterization of RAPD variation among snap bean (*Phaseolus vulgaris*) genotypes. Theor Appl Genet 91:1078–1085

Sul IW, Korban SS (1996) A highly efficient method for isolating genomic DNA from plant tissues. Plant Tissue Cult Biotechnol 2:113–116

Takeuchi A, Matsumoto S, Hayatsu M (1994) Chalcone synthase from *Camellia sinensis*: isolation of the cDNA and the organ-specific and sugar-responsive expression of the genes. Plant Cell Physiol 35:1011–1018

Tanaka J (2001) Relationship between tea breeding and genetic analysis by DNA markers. International Conference on O-Cha Culture and Science: Exploring New Possibilities for O-Cha (tea) in the 21st Century, Shizuoka, Japan, ICOS (in press)

Tanksley SD, Young ND, Paterson AH, Bonirbale MW (1989) RFLP mapping in plant breeding: new tools for an old science. Biotechnology 7:257–264

Tsumura Y, Suyama Y, Yoshimura K, Shirato N, Mukai Y (1997) Sequence-tagged-sites (STSs) of cDNA clones in *Cryptomeria japonica* and their evaluation as molecular markers in conifers. Theor Appl Genet 94:764–772

Visser (1969) Tea, *Camellia sinensis* (L.) O. Kuntze. In: Ferwarda FP, Wit F (eds) Outline of perrenial crop breeding in the tropics. Veenman H and Zonen NV, Wageningen, The Netherlands, pp 459–493

Wachira FN, Waugh R, Hackett CA, Powell W (1995) Detection of genetic diversity in tea (*Camellia sinensis*) using RAPD markers. Genome 38:201–210

Wanner LA, Li G, Ware D, Somssich IE, Davis KR (1995) The phenylalanine ammonia-lyase gene family in *Arabidopsis thaliana*. Plant Mol Biol 12:227–234

Wight W (1962) Tea classification revised. Curr Sci 31:298–299

Wikremaratne MR (1981) Variation in some leaf characteristics in tea (*Camellia sinensis*) and their use in identification of clones. Tea Quart 50:183–189

Yeh FC, Young RC, Timothy B, Boyle TBJ, Ye ZH, Mao JX (1997) POPGENE, the user-friendly shareware for population genetics analysis, Molecular Biology and Biotechnology Center, University of Alberta, Canada

Zhang Q, Saghai Maroof MA, Lu TY, Shen BZ (1992) Genetic diversity and differentiation of indica and japonica rice detected by RFLP analysis. Theor Appl Genet 83:495–499