Planta (2023) 258:81
https://doi.org/10.1007/s00425-023-04234-9

ORIGINAL ARTICLE

Extensive natural *Agrobacterium*-induced transformation in the genus *Camellia*

Ke Chen[1] · Hai Liu[1] · Todd Blevins[2] · Jie Hao[3] · Léon Otten[2]

Received: 11 July 2023 / Accepted: 30 August 2023 / Published online: 16 September 2023

Abstract

Main conclusion **The genus *Camellia* underwent extensive natural transformation by *Agrobacterium*. Over a period of 15 million years, at least 12 different inserts accumulated in 72 investigated *Camellia* species.**

Abstract Like a wide variety of other wild and cultivated plants, *Camellia* species carry cellular T-DNA sequences (cT-DNAs) in their nuclear genomes, resulting from natural *Agrobacterium*-mediated transformation. Short and long DNA sequencing reads of 435 accessions belonging to 72 *Camellia* species (representing 12 out of 14 sections) were investigated for the occurrence of cT-DNA insertions. In all, 12 different cT-DNAs were recovered, either completely or partially, called CaTA to CaTL. Divergence analysis of internal cT-DNA repeats revealed that the insertion events span a period from 0.075 to 15 Mio years ago, and yielded an average transformation frequency of one event per 1.25 Mio years. The two oldest inserts, CaTA and CaTD, have been modified by spontaneous deletions and inversions, and by insertion of various plant sequences. In those cases where enough accessions were available (*C. japonica*, *C. oleifera*, *C. chekiangoleosa*, *C. sasanqua* and *C. pitardii*), the younger cT-DNA inserts showed a patchy distribution among different accessions of each species, indicating that they are not genetically fixed. It could be shown that *Camellia* breeding has led to intersectional transfer of cT-DNAs. Altogether, the cT-DNAs cover 374 kb, and carry 47 open reading frames (ORFs). Two *Camellia* cT-DNA genes, CaTH-*orf358* and CaTK-*orf8*, represent new types of T-DNA genes. With its large number of cT-DNA sequences, the genus *Camellia* constitutes an interesting model for the study of natural *Agrobacterium* transformants.

Keywords *Camellia* taxonomy · cT-DNA · Multiple natural transformation · nGMO · Plant evolution · T-DNA proteins

Abbreviations

cT-DNA	Cellular T-DNA
HP	Hypothetical protein
LA	Left arm of the inverted repeat
LB	Left border
nGMO	Natural genetically modified organism
RA	Right arm of the inverted repeat
RB	Right border
SRR	Sequence read run
WGS	Whole genome sequencing

Communicated by Dorothea Bartels.

✉ Ke Chen
434725514@qq.com

Léon Otten
leon.otten@ibmp-cnrs.unistra.fr

[1] Shanghai Key Laboratory of Plant Functional Genomics and Resources, Shanghai Chenshan Botanical Garden, Shanghai 201602, China

[2] Institut de Biologie Moléculaire des Plantes du C.N.R.S., Rue du Général Zimmer 12, 67084 Strasbourg, France

[3] Institute of Clinical Science, Zhongshan Hospital, Fudan University, Shanghai, China

Introduction

Agrobacterium tumefaciens and *Rhizobium rhizogenes* (also called *A. rhizogenes*) induce tumors and hairy roots by genetic transformation. The mechanism for this process is well-known and has been studied in detail (Zhu et al. 2000; Nester 2015; Gelvin 2017; Hooykaas 2023). During the infection process, *Agrobacterium* transfers a well-defined part of its DNA from a large plasmid to plant cells, where this transferred DNA (T-DNA) is randomly integrated into the nuclear DNA. The T-DNA is flanked by short sequences, called left and right borders (LB and RB). Once integrated,

the T-DNA-located genes are expressed and lead to two major changes: the synthesis of specific metabolites, used by the bacterium for its growth (called opines, Petit and Tempé 1985; Dessaux et al. 1998), and induction of cell division by different mechanisms (reviewed in Nester 2015; Otten 2018). Early in *Agrobacterium* research, it was found (White et al. 1983; Furner et al. 1986; Suzuki et al. 2002) that *Nicotiana glauca* and a few other *Nicotiana* species like *N. tabacum* (tobacco) and *N. tomentosiformis* contain T-DNA sequences in their nuclear genomes (called cellular T-DNA or cT-DNA). It was proposed that this was due to the spontaneous regeneration of transformed hairy root cells into fertile plants, a process which has been experimentally demonstrated in various species, and often leads to plants with a new phenotype, called "hairy root phenotype" (Tepfer 1990; Trevenzoli Favero et al. 2022). *Nicotiana* sect. *Tomentosae* displays a series of successive transformation events (Chen et al. 2014), which led to the accumulation of four different cT-DNAs (TA to TD) in *N. tomentosiformis*, the paternal ancestor of *N. tabacum*. The order of introduction of these cT-DNAs could be reconstructed using the divergence levels of internal cT-DNA repeats. Some cT-DNA genes from *Nicotiana* have been shown to have strong morphogenic activity when overexpressed in model plant species, like *rolB*, *rolC*, and *6b* (reviewed in Chen and Otten 2017), but a role for these genes in their original hosts still remains to be demonstrated.

The cT-DNAs of *N. tabacum* (Chen et al. 2016) and *Cuscuta suaveolens* (Zhang et al. 2020) carry active genes that lead to opine synthesis in the plant. Sweet potato was the first food crop shown to be a natural transformant (Kyndt et al. 2015). Recently, it has become clear that natural *Agrobacterium* transformants are rather widespread, with 7–10% of all dicots being naturally transformed (Matveeva and Otten 2019; Matveeva 2021). Such plants were called natural *Agrobacterium* transformants or natural genetically modified organisms (nGMOs). Natural genetic transformation by *Agrobacterium* is a case of horizontal gene transfer (HGT), but differs from other HGT types (Dunning et al. 2019; Ma et al. 2022) by the fact that the transferred DNA sequences are not random fragments, but well-defined T-DNA sequences. Importantly, the transferred genes were selected over a long evolutionary period to manipulate plant growth and metabolism to the advantage of the bacterium (Petit and Tempé 1985; Nester 2015). Thus, their acquisition by nGMOs is expected to cause changes in phenotype and metabolism with respect to the nontransformed ancestor. Among the many nGMOs, the tea plant *Camellia sinensis* carries a 5.5 kb cT-DNA, called CaTA (Matveeva and Otten 2019; Chen et al. 2022). *C. sinensis* is a member of the Theaceae family and belongs to the genus *Camellia*, divided into subgenera Thea and Camellia. These are further subdivided into sections, yielding different classification systems (Sealy 1958; Wight 1962; Chang 1998; Min and Bartholomew 2007). Here we will use the Min and Bartholomew system, which proposes 97 species, placed in 14 sections. Apart from the tea plant, the genus *Camellia* contains other economically important species, like *C. japonica*, *C. azalea*, and *C. petelotii* var. *petelotii* (*C. nitidissima*), bred for their ornamental flowers, and *C. oleifera*, *C. chekiangoleosa* and *C. sasanqua*, used to produce tea seed oil. The grouping of species into sections, and the taxonomical position of several species is not clear, and requires further studies (Chen and Yamaguchi 2002; Xiao and Parks 2003; Vijayan et al. 2009; Yang et al. 2013; Zhang et al. 2021b, 2022; Lin et al. 2022; Shen et al. 2022; Wu et al. 2022). One complicating factor in *Camellia* taxonomy is the existence of interspecific and intersectional hybrids (Takeda 1990; Min and Bartholomew 2007; Hembree et al. 2019). Earlier, we explored 225 CaTA sequences from 142 *Camellia* accessions, belonging to 10 out of the 11 species from sect. *Thea* (Chen et al. 2022). All accessions contained CaTA, indicating that sect. *Thea* is derived from a single transformed plant. A phylogenetic tree of the CaTA sequences showed no clear borders between the ten species, raising the question of their taxonomic position. We now extend our cT-DNA studies to all published *Camellia* sequences, and found, besides CaTA, 11 additional cT-DNAs (CaTB to CaTL) in altogether 72 species belonging to 12 sections. Our data show that the genus *Camellia* underwent a series of independent transformation events, which can be traced over time through different sections and species.

Materials and methods

Whole genome sequencing

Fresh leaves of *C. transarisanensis*, *C. cuspidata*, *C. euryoides*, and *C. handelii* (*C. transarisanenesis* according to Min and Bartholomew 2007), grown in Chenshan Botanical Garden (Shanghai) were selected for DNA extraction using the CTAB method (Porebski et al. 1997). The extracted DNA samples were subjected to DNBSEQ-T7 and Oxford Nanopore sequencing by Wuhan Benagen Technology Co., Ltd.

For DNBSEQ-T7 DNA sequencing, after DNA extraction, 1 µg genomic DNA was randomly fragmented by Covaris, followed by fragments selection by Agencourt AMPure XP-Medium kit to an average size of 200–400 bp. Selected fragments were end repaired and 3′adenylated, then the adaptors were ligated to the ends of these 3′adenylated fragments. The products were amplified by PCR and purified by the Agencourt AMPure XP-Medium kit. The purified double stranded PCR products were heat denatured to single strand, and then circularized by the splint oligo

sequence. The single strand circle DNA (ssCir DNA) were formatted as the final library and qualified by QC. The final qualified libraries were sequenced by BGISEQ-500. ssCir DNA molecule formed a DNA nanoball (DNB) containing more than 300 copies through rolling-cycle replication. The DNBs were loaded into the patterned nanoarray by using high density DNA nanochip technology. Finally, pair-end 100 bp reads were obtained by combinatorial Probe-Anchor Synthesis (cPAS).

For Oxford Nanopore sequencing, the libraries were prepared using the SQKLSK109 ligation kit and using the standard protocol. The purified library was loaded onto primed R9.4 Spot-On Flow Cells and sequenced using a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) with 48-h runs at Wuhan Benagen Tech Solutions Company Limited, Wuhan, China. Base calling analysis of raw data was performed using the Oxford Nanopore GUPPY software (v0.3.0).

C. euryoides, *C. handelii*, and *C. cuspidata* were subjected to DNBSEQ-T7 runs yielding, respectively SRR19974293, SRR19974292, and SRR19974291. *C. transarisanensis* samples were subjected to five DNBSEQ-T7 runs, yielding SRR19974295 (tra1), SRR21420630 (tra2), and SRR22071522 (tra3), which were mixed samples from three *C. transarisanensis* plants, and SRR22937856 (tra5) and SRR24101861 (tra6), which were samples from individual *C. transarisanensis* plants. SRR22198564 (tra4) resulted from the Oxford Nanopore sequencing (ONT) of a mixed sample of three *C. transarisanensis* plants.

Recovery of cT-DNA sequences from NCBI database

cT-DNA sequence fragments were recovered from the databases of the National Center for Biotechnology Information (NCBI, Bethesda, MD, USA) and the National Genomics Data Center (NGDC, Beijing, China). Previously, 142 accessions from sect. *Thea* were studied (Chen et al. 2022), here we add 293 accessions from 11 additional sections. In brief, chromosome-scale genome assemblies and whole genome sequences (WGS) datasets, as well as sequence read run (SRR) datasets, including short and long read data, were analyzed using BLASTN (Altschul et al. 1990) with nucleotide query sequences, or using TBLASTX with protein query sequences SUPERPLAST and SUPERRHIZ (Suppl. Fig. S1). These are concatenated sequences of Plast proteins (phenotypic plasticity proteins, Otten 2018) and *R. rhizogenes* T-DNA proteins (Otten 2021). In a few cases, cT-DNA sequences without similarity to known T-DNA sequences were identified through progressive extension of already identified cT-DNA sequences, by successive rounds of BLASTN or TBLASTX searches. In order to distinguish such new sequences from the surrounding plant DNA or from plant DNA inserts within the cT-DNA, they were compared with sequences from related *Camellia* species by blastn.

Assembly of sequences from short individual reads

Recovered reads were assembled and alleles separated using the allele separation function of CodonCode Aligner (CCA, version 10.0.2, CodonCode Corporation). Apart from Default settings, a value of 100% of Minimum Percent Identity was used, in order to separate alleles with sequences as similar as possible (i.e. down to about 10 nt differences per 5000 nt). In most cases, the cT-DNA consisted of two inverted repeats. These could be separated using the allele separation function of the CAA program in case only one allele type was present. In case of insufficient coverage, the repeats of most cT-DNAs could still be assembled into various paired fragments, allowing the calculation of identity values, and the prediction of ORFs. For T-DNA types and T-DNA gene names from *Agrobacterium* we refer to Otten (2021). The comparison of different cT-DNAs from large numbers of *Camellia* accessions required a simple nomenclature. We therefore used abbreviations of the type "abcnxCaTX", where abc refers to the species name, *n* to the particular accession of the species, *x* to the allele, and *X* to the cT-DNA type, e.g., sin1aCaTA indicates CaTA from *C. sinensis* accession 1, allele a. The correspondence between our simplified accession numbers and the original accession numbers is provided in Suppl. Table S1.

Details on different *Camellia* accessions

The details on *Camellia* accessions (like geographic and genetic origin, systematic position, alternative names and so on) can be found in the International Register of *Camellia* (https://camellia.iflora.cn). Species names provided with the *Camellia* sequences in public databases may be either those chosen by Min and Bartholomew (2007), or synonyms, the latter are generally mentioned by these authors.

Phylogenetic tree construction

Protein or nucleotide sequences were aligned using MUSCLE (Edgar 2004). The aligned sequences were manually trimmed in MEGA v11.0.13 (Tamura et al. 2021) by deleting gaps with fewer than three sequences. The trimmed sequences were then used to construct a phylogenetic tree using IQtree2 (Minh et al. 2020), with the parameter "-m MFP" to automatically detect the optimal model and build a maximum likelihood tree with 1000 bootstrap replicates. The resulting tree was visualized and annotated using iTOL (Letunic and Bork 2021).

Results

Investigation of the available *Camellia* short and long read WGS datasets, as well as chromosome-scale genome assemblies (up to March 1, 2023) (Materials and methods, Suppl. Table S1) yielded 12 different cT-DNA inserts, which we will call CaTA to CaTL (see below). Their sequences are provided in Suppl. Fig. S2, maps are shown in Fig. 1. Various properties (size of insert, size of left and right arms, size of flanking plant sequences, percentage of identity of repeats) of these inserts are listed in Table 1, and their distribution among the various *Camellia* accessions is given in Suppl. Table S1. Protein sequences predicted for ORFs are listed in Suppl. Fig. S3, and their properties are shown in Table 2, and discussed under "cT-DNA proteins from *Camellia*". For some inserts, mainly CaTA and CaTD, different subtypes were identified (see below).

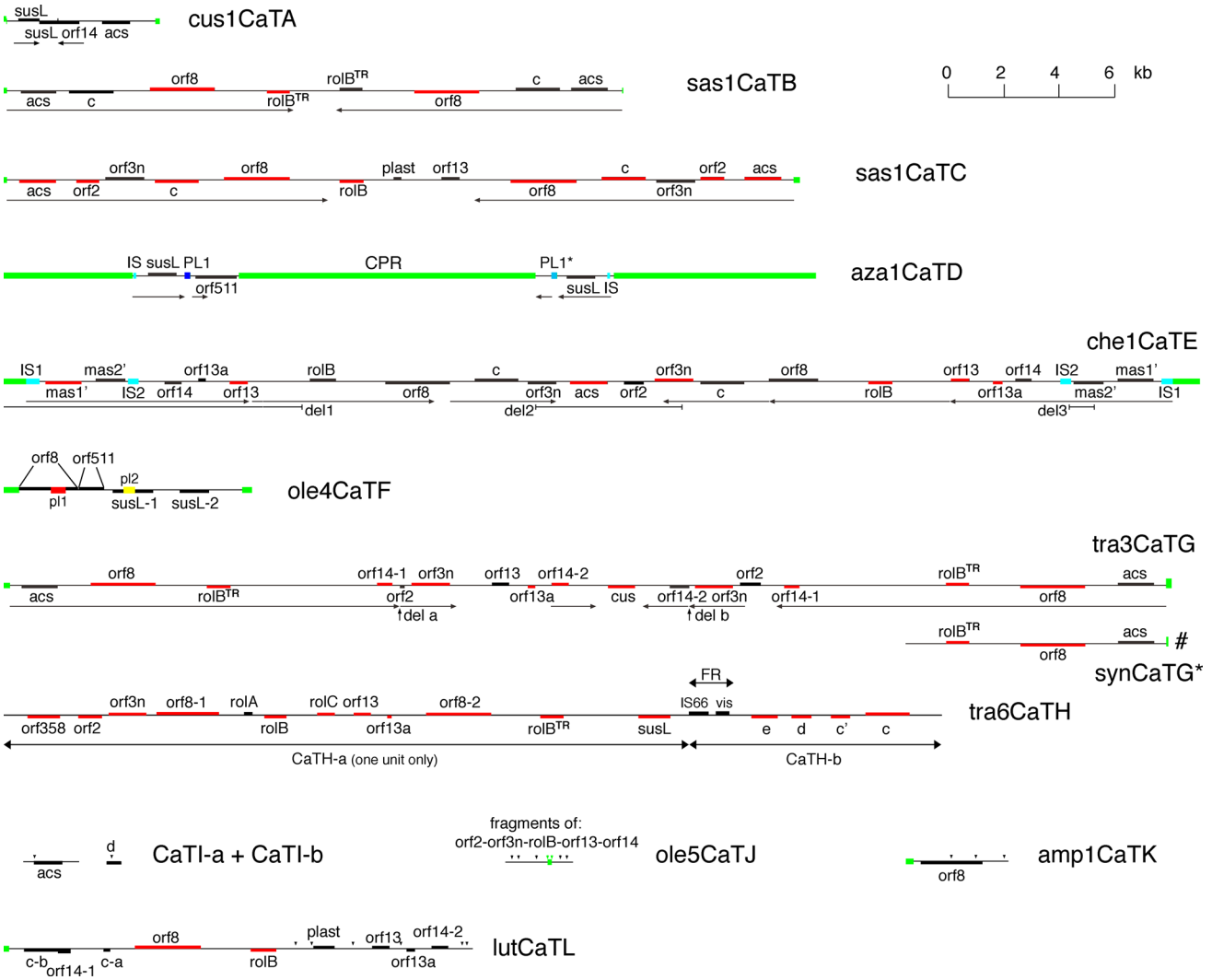

Fig. 1 Representative maps of CaTA to CaTL. Horizontal arrows: inverted repeats. Small vertical arrows: gaps in the assembly caused by insufficient coverage with small reads. Genes are marked in red (open reading frame) or black (no open reading frame). Green: plant DNA outside cT-DNA. Names of genes as in Otten (2021). PL1 and PL1* in aza1CaTD: similar but not identical plant insertion sequences (see also Fig. 4). CPR in aza1CaTD: central plant region. del1-3 in che1CaTE: deletions found in sas1CaTE. IS1 and IS2 in che1CaTE: bacterial insertion sequences. pl1 and pl2 in ole4CaTF: plant insertion sequences. *susL-1* and *susL-2* in ole4CaTF: two different *susL* genes. del a and del b in tra3CaTG: deletions in the left and right arm, respectively. synCaTG*: sequence very similar to right part of tra3CaTG, but inserted in another chromosomal location (marked by #). *orf14-1* and *orf14-2* in tra3CaTG and in lutCaTL: two different *orf14* sequences on the same cT-DNA. *orf8-1* and *orf8-2* in tra6CaTH: two different *orf8* genes. FR: fragment with similarity to *IS66* and *vis* from the right end of the CG412 TB-region

Table 1 Properties of different cT-DNA sequences from *Camellia*

cT-DNA	Type of assembly	Size (kb) In Fig. 1	cT-DNA (kb)	Left arm (kb)	Right arm (kb)	Plant seq left (bp)	Plant seq right (bp)	Equivalent ole1 left	Equivalent ole1 right	Deleted (bp)	% identity of the cT-DNA repeats	Remarks
sin1CaTA	WGS	5.687	5.287	4.327	0.960	200	200	X05:112,515,619	X05:112,515,690	70	90%	Chen et al. (2022)
cus1aCaTA	SRA	5.171	5.019	0.925	4.095	110	42	X05:112,515,619	X05:112,515,690	70	92%	
sas1CaTB	SRA	22.224	22.080	11.835	10.288	99	45	X10:79,580,748	X10:79,580,754	5	97.2%	
sas1CaTC	SRA	16.888	28.4	16.9	11.5	114	105	X13:122,883,895	X13:122,948,557	64,661	97%	
aza1CaTD	WGS	25.831	17.132	3.880	2.714	1324	7374	NCLE	NCLE	ND	91%	10,650 nt plant insert
ole1CaTD	WGS	18.125	13.687	3.831	1.471	200	200	NCLE	NCLE	ND	89%	
che1CaTD	WGS	5.506	5.106	5.106	deleted	200	200	NCLE	NCLE	ND	-	
amp1CaTD	SRA	2.410	> 2.333	ND	> 2.333	ND	77	NCLE	NCLE	ND	ND	Additional plast gene
che1CaTE	WGS	42.934	41.143	19.020	22.123	815	977	X13:68,157,945	X13:68,157,917	27	98%	
ole4CaTF	SRA	8.914	8.012	–	–	546	356	X11:38,014,189	X11:38,002,474	11,714	–	
tra3CaTG	SRA	41.917	41.501	22.711	18.789	227	190	X02:59,033,764	X02:59,033,796	31	99%	
synCaTG*	SRA	9.280	9.221	–	9.221	ND	59	ND	X12:37,287,561	ND	–	
tra6CaTH	SRA	33.7 (one unit)	194	–	–	14,700	33,000	X01:180,303,540	X01:180,303,039	500	99%	See suppl. fig. S4(a)
CaTIa	SRA	2.010	> 2.010	–	–	–	–	ND	ND	ND	–	Incomplete
CaTIb	SRA	0.522	> 0.522	–	–	–	–	ND	ND	ND	–	Incomplete
ole5CaTJ	SRA	2.270	> 2.270	–	–	–	–	ND	ND	ND	–	Incomplete
amp1CaTK	SRA	3.681	> 3.404	–	–	278	–	ND	ND	ND	–	Incomplete
lutCaTL	SRA	16.850	> 16.667	–	–	184	–	ND	ND	ND	–	Incomplete

Listed are size of sequences shown in Fig. 1, size of cT-DNAs, size of left and right flanking plant sequences, the equivalent of the surrounding plant sequences (left and right) in *C. oleifera* cv. CON1 (X = JAKJMZ0100000), and the region deleted during the insertion process, and the similarity of the repeats. NCLE: no cT-DNA-less equivalent. ND: not determined. WGS: Whole Genome Sequence, SRA: Sequence Read Archive, sin: *sinensis*, cus: *cuspidata*, sas: *sasanqua*, aza: *azalea*, ole: *oleifera*, che: *chekiangoleosa*, amp: *amplexicaulis*, tra: *transarisanensis*, syn: *synaptica*, lut: *lutchuensis*. For details on the origin of the cT-DNA sequences see Suppl. Table 1. kb: kilobase, bp: base pairs

Table 2 Properties of predicted proteins from intact ORFs from *Camellia* cT-DNAs

CaTX type	T-DNA protein	Size AA	Accession number of homolog	%Sim	*Agrobacterium* strain (Otten 2021)	Prot type
sas1CaTB	NOrf8-L	671	NTF91497.1	48	MCX	plast
	RolB^{TR}	260	P15397.1	72	LMG152	plast
	NOrf8-R	510	NTF91497.1	48	MCX	plast
sas1CaTC	Acs-L + R	444	KEA04447.1	80	LMG152	OS
	Orf2-L + R	273	NTF91494.1	80	MCX	Orf2
	C-L + R	524	NTF91496.1	81	MCX	C
	COrf8-L + R	715	NTF91497.1	78	MCX	Orf8
	RolB	284	NTF91498.1	83	MCX	plast
che1CaTE	Mas1'	430	NTI85400.1	81	LMG152	OS
	Orf13-L	213	NTF91500.1	67	MCX	plast
	Orf13-R	213	NTF91500.1	69	MCX	plast
	Acs	443	KEA04447.1	83	LMG152	OS
	Orf3n	454	NTF91495.1	88	MCX	Orf3n
	RolB	280	NTF91498.1	74	MCX	plast
	Orf13a	119	NTF91501.1	64	MCX	Orf13a
tra3CaTG	Orf8-L + R	779	NTF91497.1	47	MCX	Orf8
	RolB^{TR}-L + R	261	P15397.1	72	LMG152	plast
	Orf14-1-L + R	185	AIM40184.1	68	N-TOF*	plast
	Orf3n-L + R	454	NTF91495.1	92	MCX	Orf3n
	Orf13a	118	NTF91501.1	52	MCX	Orf13a
	Orf14-2	200	NTF91502.1	74	MCX	plast
	Cus	363	MVA69244.1	78	CG108	OS
tra6CaTH	Orf358	390	MCZ7448279.1	64	ST15-13.057	OS
	Orf2	275	MCZ7448278.1	84	ST15-13.057	Orf2
	Orf3n	448	TRB05100.1	88	LMG63	Orf3n
	Norf8	281	P09178.1	34	N-TOF*	plast
	RolB	259	NTF72659.1	89	MCW	plast
	RolC	184	NTG71384.1	86	NCIB8196	plast
	Orf13	198	NTF72661.1	90	MCX	plast
	Orf13a	90	CAB65898.1	86	2659	Orf13a
	Orf8-2	779	NTF91497.1	46	MCX	Orf8
	RolB^{TR}	274	P15397.1	71	LMG152	plast
	SusL	379	MCZ7448270.1	68	ST15-13.057	OS
	E	320	ASK74226.1	73	EU6	plast
	D	239	ASK41783.1	66	EU6	plast
	C'	230	MVA63187.1	69	CG474	plast
	C	524	ASK46911.1	86	CFBP2407	C
lutCaTL	Orf8	764	NTF91497.1	78	MCX	Orf8
	RolB	265	NTF91498.1	75	MCX	plast

Shown are the name of the CaTX insert, the T-DNA protein (L, R: left and right arm), size in amino acids (AA), accession number of the nearest homolog in *Agrobacterium/Rhizobium*, % similarity between *Camellia* cT-DNA protein and *Agrobacterium/Rhizobium* protein, the corresponding *Agrobacterium/Rhizobium* strain (according to Otten 2021), and the T-DNA protein type. N-TOF*: nGMO *N. tomentosiformis*. OS: opine synthase, C, Orf2, Orf3n, and Orf13a are unique protein types, Orf8 belongs to the tryptophan 2-monooxygenase (IaaM) group. sas: *sasanqua*, che: *chekiangoleosa*, tra: *transarisanensis*, lut: *lutchuensis*

CaTA

A 5.5 kb cT-DNA insert (CaTA) was initially identified in *C. sinensis* cv. Shuchazao (Matveeva and Otten 2019) from sect. *Thea*. Subsequently, the CaTA insert was identified in 142 accessions belonging to 10 out of 11 species of this section. No other cT-DNA-like sequences were found in these species (Chen et al. 2022). CaTA from *C. sinensis*

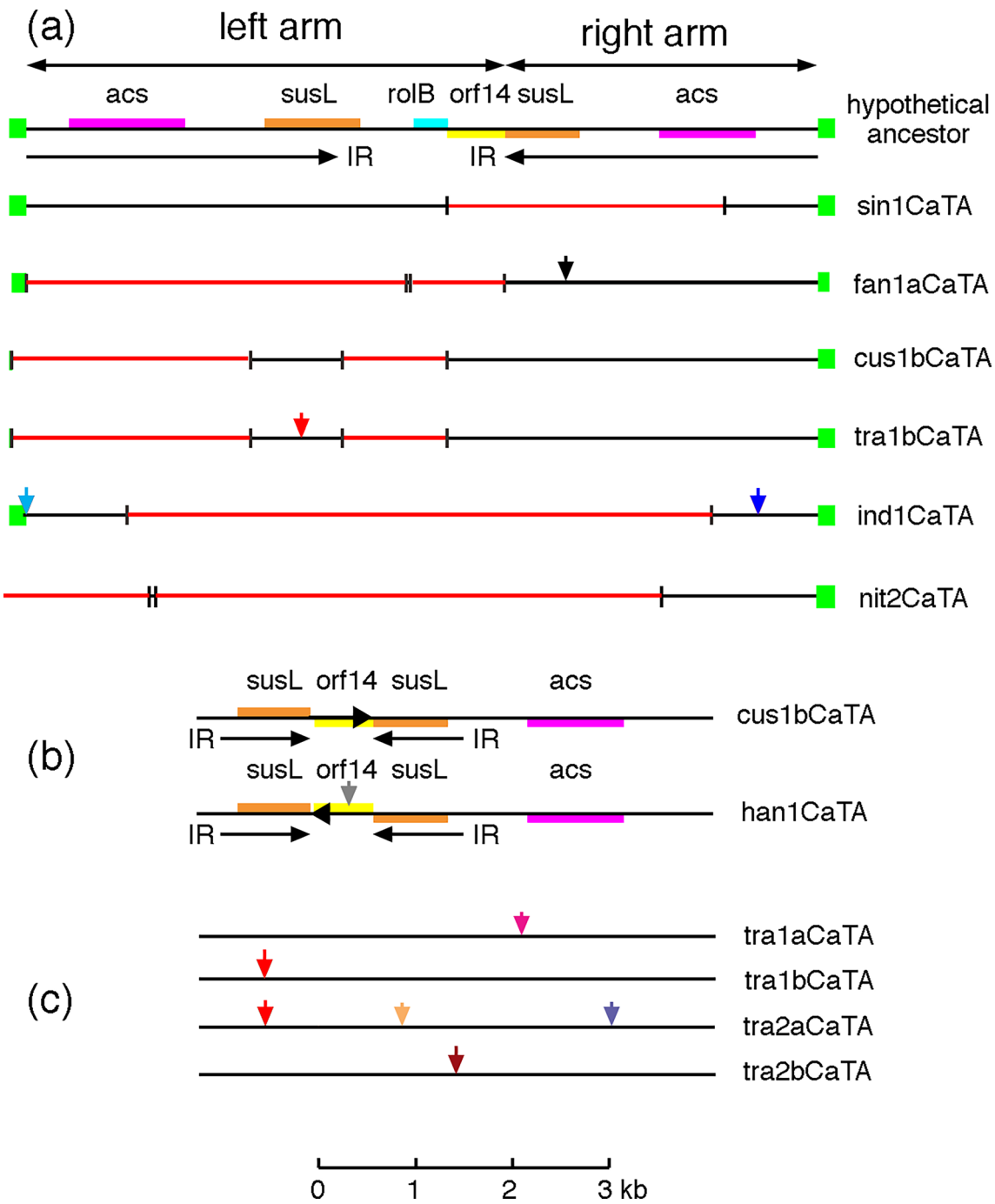

Fig. 2 **a** Maps of the hypothetical ancestor CaTA, sin1CaTA, fan1aCaTA, cus1bCaTA, tra1bCaTA, ind1CaTA and nit2CaTA. Plant sequences surrounding the CaTA insert are marked in green. The various CaTA sequences are aligned with the hypothetical ancestor CaTA, and deleted regions are indicated in red. Each map can be derived from the ancestor CaTA by one or more deletions. IR inverted repeat. Different plant insertions are marked by vertical arrows with different colors. **b** cus1bCaTA and han1CaTA differ by an inversion of the *orf14* region and the insertion of a plant sequence (grey vertical arrow) in *orf14* in han1CaTA. **c** Four different traCaTA regions. They differ by the insertion of different plant sequences, indicated by vertical arrows of different color

is a partial inverted repeat with a left and right arm (LA, RA), carrying sequences related to the *acs*, *susL*, and *rolB* T-DNA genes (Chen et al. 2022). The repeats show 10% DNA sequence divergence. fan1CaTA of *C. fangchangensis* (sect. *Thea*) is different from sin1CaTA and was interpreted as a CaTA derivative with an inversion between the two inverted repeats, and two deletions (Chen et al. 2022). In this work, we investigated additional *Camellia* species in

other sections, and detected further CaTA types. One was found in *C. cuspidata* (cus1) from sect. *Theopsis* (Figs. 1, 2, Table 1, Suppl. Fig. S2). Two CaTA alleles could be assembled (Materials and methods): cus1aCaTA (5171 nt) and cus1bCaTA (5349 nt), these are colinear and 95% identical. Compared with sin1CaTA and fan1CaTA, cus1CaTA contains an additional sequence with *orf14* homology. Analysis of the three different CaTA types indicated that they are probably derived from a 9.5 kb ancestor CaTA, by different deletions (Fig. 2a).

CaTAs from three other species of sect. *Theopsis*, *C. euryoides*, *C. transarisanensis*, and *C. handelii*, have the same overall structure as cus1CaTA (eur1CaTA, tra1CaTA, tra2CaTA, han1CaTA). Although han1CaTA is similar to cus1CaTA, the central part between the inverted repeats is inverted and carries a plant insert (Fig. 2b). *C. transarisanensis* shows four CaTA subtypes, with different plant inserts (Fig. 2c).

Additional CaTA types (Fig. 2a) were found in *C. indochinensis* (four accessions) and *C. nitidissima* (10 accessions). They are very short and most likely result from a large deletion of the central part of the original inverted repeat.

Interestingly, all 156 accessions from sect. *Camellia* and *Paracamellia* (Suppl. Table S1) lack CaTA. Five *C. japonica* (sect. *Camellia*) hybrids contain CaTA sequences. One (cv. XF) is a *C. japonica* 'Tiffany' x *C. lutchuensis* (sect. *Theopsis*) hybrid. The four others ('Meigui Chun', 'Meiyu', 'Xiao Fenyu', and 'Jinye Fenyu') are *C. japonica* 'Kuro-tsubaki' x *C. synaptica* hybrids. *C. lutchuensis* and *C. synaptica* belong to sect. *Theopsis* and the CaTA sequences of the hybrids are most similar to *Theopsis* CaTAs, strongly suggesting that the CaTA sequences from these hybrids originate from the *Theopsis* partner. The lack of CaTA in sect. *Camellia* and *Paracamellia* could result from deletion or segregation in case the insert was not yet fixed, or from the fact that these sections descend from ancestors which never contained this insert. *C. oleifera* (ole1) and *C. sasanqua* (sas1) from sect. *Paracamellia* have an intact CaTA insertion site, including a 70 nt sequence which was lost upon insertion (Chen et al. 2022). *C. azalea* (aza1) and *C. chekiangoleosa* (che1), both from sect. *Camellia*, showed the same intact site. We therefore assume that the ancestor of sect. *Camellia* and *Paracamellia* lacked CaTA, whereas the ancestor of the other investigated sections carried the CaTA insert. Fourteen out of 37 accessions of the closely related sections *Theopsis* and *Eriandria* also lack CaTA (Suppl. Table S1). In these sections, this is probably due to secondary loss of the CaTA sequences (see below). None of the CaTA sequences shows an open reading frame (ORF). Thus, this cT-DNA appears to be degenerated.

CaTB

When probed for CaTA sequences, several *Camellia* accessions were found to have an unusual *acs* sequence, suggesting the presence of an additional cT-DNA. Individual reads from *C. sasanqua* (sas1, sect. *Paracamellia*) were recovered with different query sequences (Materials and methods) and assembled into a single contig, called sas1CaTB (Fig. 1, Table 1, Suppl. Fig. S2). This cT-DNA is a 22 kb partial inverted repeat of the LB-RB-RB-LB type. The region between the repeats lacks SNPs. Thus, there is only one type of sas1CaTB allele in this accession, allowing separation of the two repeats by the CCA program (Materials and methods), they are 97.2% identical.

sas1CaTB carries sequences with homology to *acs*, *c*, *orf8* and $rolB^{TR}$. Among the four sas1CaTB genes, gene *c* and $rolB^{TR}$ on the left repeat are intact. The predicted CaTB-Orf8 proteins are truncated at the C-terminus, but their RolB-like N-termini are intact. The possible roles of the intact genes and their corresponding proteins are described below.

Interestingly, only five of 11 *C. sasanqua* accessions carry CaTB (Suppl. Table S1). Thus, the CaTB is not fixed in this species. The same is true for *C. oleifera* (23 of 68 accessions), *C. fluviatilis* (1/5), *C. brevistyla* (3/5), *C. kissii* (1/2). Other species with CaTB are: *C. vietnamensis* (3/3), *C. osmantha* (similar to *C. oleifera*, 5/5), and *C. meiocarpa* (2/2). All belong to sect. *Paracamellia*. A single accession from sect. *Heterogenea* (*C. furfuracea* var. *furfuracea*) contains CaTB sequences, eight other ones do not. Only *C. osmantha* (osm1), *C. brevistyla* cv. DYDZC (bre2), and *C. oleifera* cv. Xianglin210 (ole4) had enough reads for assembly, they contain the full CaTB sequence, like sas1CaTB. In these three accessions, multiple SNPs in the unique region revealed the presence of several alleles.

CaTC

Assembly of sas1CaTB led to the discovery of a further cT-DNA, called sas1CaTC (Fig. 1, Table 1, Suppl. Fig. S2). sas1CaTC consists of a partial, inverted repeat of 28.4 kb of the LB-RB-RB-LB type. Owing to incomplete coverage, only about 80% of the two repeats (each about 11.5 kb) could be separated, they are 3% diverged. CaTC carries sequences with homology to *acs*, *orf2*, *orf3n*, *c*, *orf8*, *rolB*, gene *plast* (truncated at its 5' end, the predicted protein has 87% identity to NTF91499.1 from the *R. rhizogenes* T155/95 TL-region), and *orf13*. The *rolB*, *plast*, and *orf13* genes are part of the unique central region and, according to typical *R. rhizogenes* T-DNA structures (Otten 2021), belong to the left arm (Suppl. Fig. S2). Several ORFs are intact: two copies each of *acs*, *orf2*, *c*, *orf8*, and the unique *rolB* copy. CaTC was found in eight species of sect. *Paracamellia*

(Suppl. Table S1). As in the case of CaTB, the CaTC insert could be either present or absent in some species.

CaTD

When searching for CaTA sequences in different *Camellia* accessions, we detected a new type of *susL* sequence, suggesting an additional cT-DNA. The corresponding cT-DNA (17.3 kb) was found in the genome assembly of *C. azalea* cv. Z01, at coordinates GWHAOSQ00000004:6,615,474–6,632,762. In order to simplify the coordinate numbers, we will transpose 6,610,000 to 1. This cT-DNA (Fig. 1, Table 1, Suppl. Fig. S2) was designated aza1CaTD. aza1CaTD is a partial, inverted repeat, separated by a 10.6 kb plant sequence (central plant region, CPR). The repeats are 91% identical. The CPR region was most likely linked to the T-DNA sequences during the T-DNA insertion process, as it is unrelated to the CaTD flanking sequences. aza1CaTD 5894–7039 is 72% identical to the *susL* gene from nGMO *Jasminum sambac*. The CaTD region shows several unusual features. 8046–8926 has 65% identity to the *orf511* cT-DNA gene from nGMO *N. tomentosiformis* (Chen et al. 2014). 5498–5596 is 80% identical to part of a bacterial insertion sequence (IS)-like sequence of the *IS110* family, found in several *Rhizobium* and *Agrobacterium* strains, like CP049220.1. LA also contains a 190 nt plant sequence (7361–7551, called PL1) between *susL* and *orf511* with 91% identity to *C. sinensis* cv. Shuchazao XM_028212253.1:144–333. RA is a shorter version of LA, it contains *susL*, the bacterial IS fragment and a PL1-like sequence. The latter is 27 nt shorter than PL1 (PL1*, 20,530–20,685) and a 66 nt sequence adjacent to PL1 in LA (7552–7618) is found on the other side of PL1* in RA (20,686–20,756). The presence of PL1 and PL1* in both CaTD arms seems to indicate that this plant DNA fragment was inserted in a T-DNA fragment, which was then duplicated and integrated into the plant genome. However, the LA homolog LA* from che1CaTD and other CaTDs (see below) lacks PL1.

The overall structure of aza1CaTD was confirmed by the WGS sequence JAKJMZ (compressed haplotype) from *C. oleifera* cv. Nanyongensis (CON1). ole1CaTD (Fig. 3) is 13,687 nt long and has the same overall structure as aza1CaTA, but its CPR carries a plant sequence not found in aza1CaTD, and a 5802 nt deletion removes part of the CPR and the adjacent right arm. This shorter right arm was called shRA. An assembled CaTD sequence from *C. fluviatilis* var. *megalantha* (*C. lanceoleosa*, JANUSC) is fully colinear with ole1CaTD and 99.6% identical.

Another modified CaTD sequence was found in accession *C. chekiangoleosa* Hu (che1, Shen et al. 2022). A 18799 nt deletion has removed most of the CPR and the entire right arm (Fig. 3). The che1CaTD LA sequence (called LA*) lacks PL1, but carries another plant sequence, PL2. *C. handelii*, *C. euryoides*, and *C. cuspidata* (all sect. *Theopsis*), carry a modified CaTD (Fig. 3) with a 16,426 nt deletion with respect to aza1CaTD (6879–23,305), removing RA and part of LA (leaving a shorter LA, shLA).

Some accessions from *C. nitidissima* var. *nitidissima* (sect. *Archecamellia*) show a different type of CaTD deletion (Figs. 3, 4, nit2, nit5), removing half of the LA (the rest being called sh*LA) and the complete RA. Other *C. nitidissima* accessions (Fig. 4, nit1, nit3, nit4) do not (Suppl. Table S1). Inspection of long reads of sect. *Theopsis* species

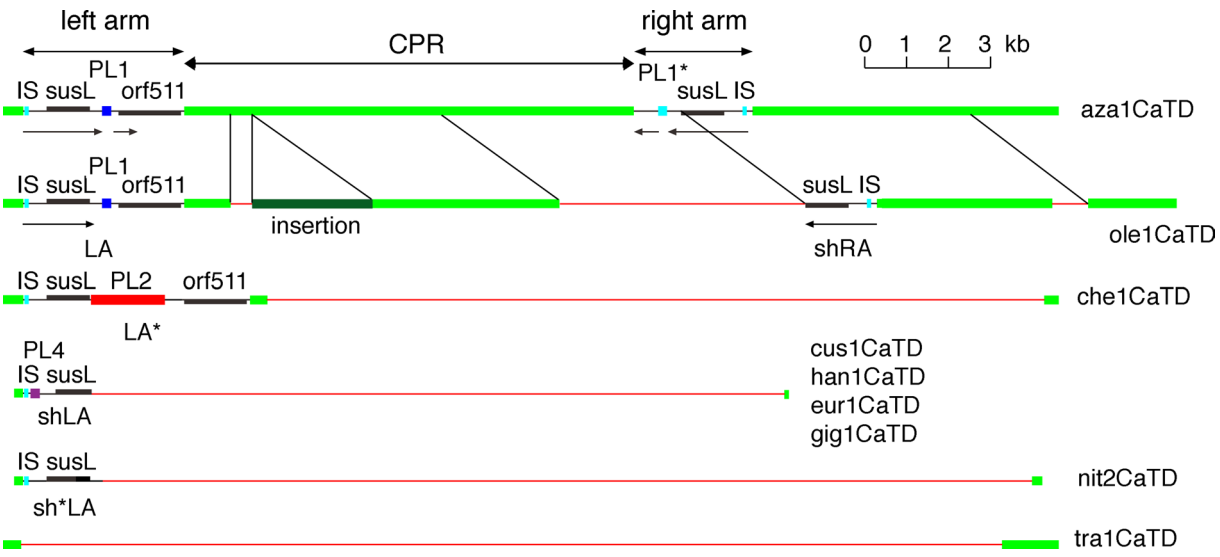

Fig. 3 Different forms of CaTD. shLA, sh*LA, and shRA are short forms of LA and RA. Red lines: deletions as compared to aza1CaTD. IS (in blue): bacterial IS sequence. PL1, PL1*, PL2 and PL4: different plant insertion sequences, see also Fig. 4

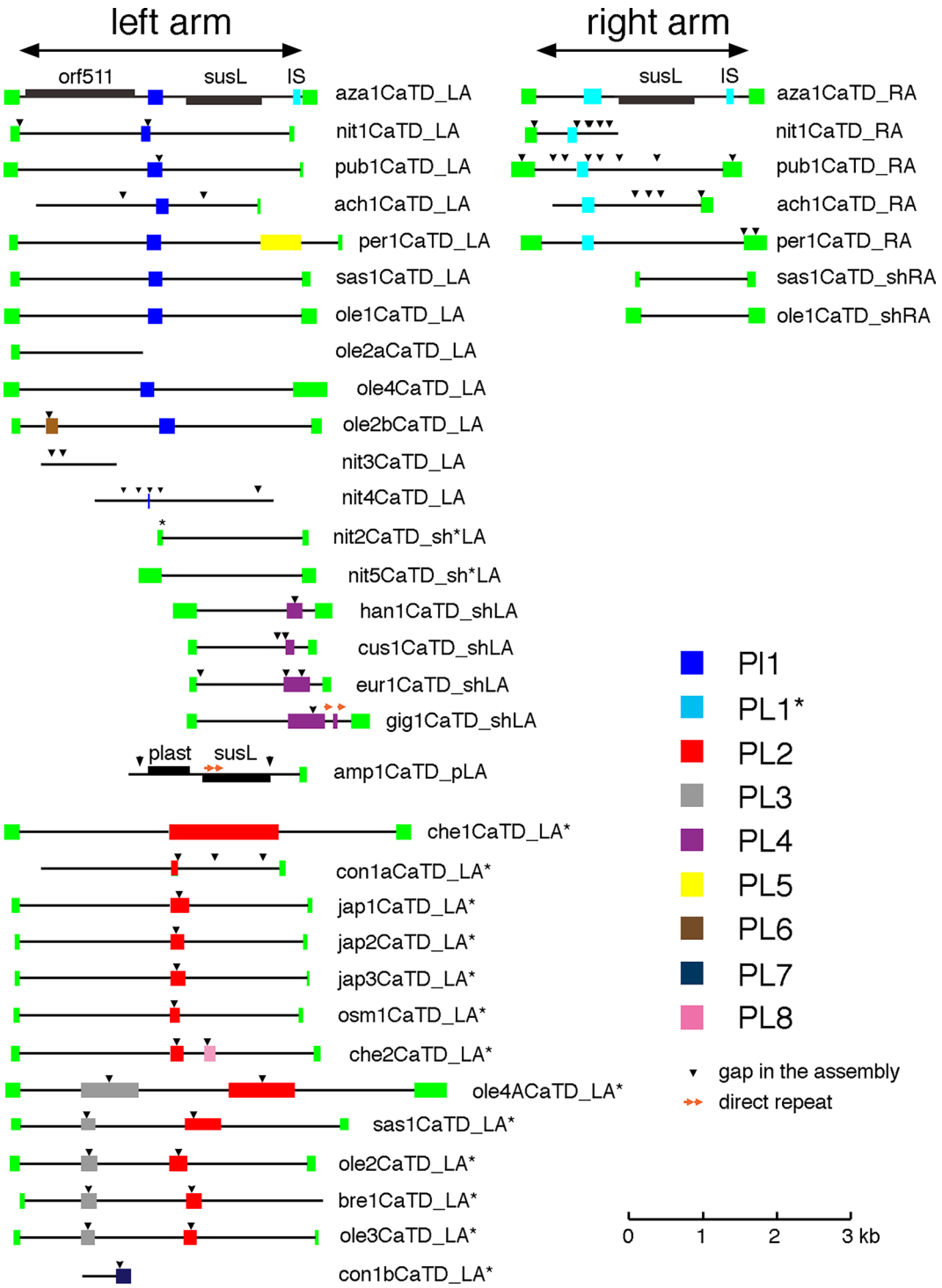

Fig. 4 CaTD maps based on the complete or partial assembly from 32 accessions. The left and right arm are shown in the same orientation to facilitate comparison. Plant inserts are indicated with different colors. Vertical small arrows: gaps in the assembly. Small horizontal arrows in amp1CaTD_pLA: direct repeat. IS: bacterial IS sequence (in blue)

C. transarisanensis (tra1) revealed a 23.3 kb deletion as compared to aza1CaTD 5409–28,675, removing the entire CaTD. In sect. *Theopsis*, 15 out of 30 accessions lack the CaTD region, in sect. *Eriandria*, five out of six.

C. crapnelliana (sect. *Heterogenea*) has a CaTD sequence with a similar overall structure as han1CaTD, cus1CaTD, and eur1CaTD, but the sequence is only 89–90% identical (see below). A further type of CaTD was found in *C.*

amplexicaulis. amp1CaTD lacks *orf511*, but contains an additional *plast* gene sequence of about 0.6 kb (Fig. 4), the predicted Plast protein has 31% similarity to the unusual T-DNA Plast protein NTF91499.1 from *R. rhizogenes* strain T155/9 (Otten 2021). We assume that this *plast* gene was part of the left arm of the original CaTD insert. In addition, the amp1CaTD region shows a direct repeat in *susL*.

To study the distribution of the various CaTD types among *Camellia* species, reads from 32 accessions with sufficient coverage were assembled into 39 contigs (Fig. 4) covering the LA and RA regions. The CPR regions of these accessions could not be assembled because they are composed of highly repeated sequences. Altogether, eight different plant inserts (PL1–PL8) were found in different CaTD contigs, PL3–PL8 could not be fully reconstructed from the short reads, because they are highly repeated in the *Camellia* genome. The gigCaTD insert shares PL4 with han1CaTD, cus1CaTD, eur1CaTD, confirming their close relationship. Although the ancestral insert carries LA and RA, we found 13 cases (Fig. 4, lower part) which only contained LA* with PL2, as in che1CaTD (Fig. 3). *C. sasanqua* sas1CaTD is an exception, as it contains both LA*, but also LA and RA. This could be due to the presence of two different alleles. ole2CaTD from *C. oleifera* (sample TO) has LA* and LA, but no RA (the extent of the deletion around RA could not be established).

A phylogenetic tree was constructed with the 39 assemblies (Fig. 5). The tree shows a clear separation between the LA, RA, shLA, shRA, sh*LA, and LA* sequences. It indicates five independent losses of RA (red arrows), partial deletions of LA and RA, and insertion of different plant sequences in various branches.

Although CaTD is widespread, some *Camellia* accessions lack CaTD. These include all accessions from sect. *Thea* and *Tuberculatae*, two out of seven *C. indochinensis* accessions (sect. *Archecamellia*), and the one accession of *C. connata* (sect. *Calpandria*). The lack of CaTD in sect. *Thea* was further investigated by comparing sequences around aza1CaTD with JACBKZ of *C. sinensis* cv. Shuchazao. When compared with aza1, the *C. sinensis* cv. Shuchazao sequence shows a 23.2 kb gap, between aza1CaTD:5420 and 28,617.

CaTE

We detected a further cT-DNA in *C. chekiangoleosa* Hu (CCH) in GWHBGBN00000017, called che1CaTE (41.2 kb, Fig. 1, Table 1, Suppl. Fig. S2). The che1CaTE map (Fig. 1) starts at coordinate 1,458,180, which we redefine as 1. che1CaTE consists of an inverted repeat with 98.6% identity between the two arms. The general structure of the repeat is RB-LB-LB-RB and the gene order of the original T-DNA is *acs*, *orf2*, *orf3n*, *c*, *orf8*, *rolB*, *orf13*, *orf13a*, *orf14*, *mas2'*, and *mas1'*. The *acs*, *orf2*, and *orf3n* genes are part of the unique region.

Between *orf14* and *mas2'*, and in both arms (che1CaTE:4613–4788 and 37,983–38,158), we found a small 178 nt fragment (called IS1) with 81% identity to a TnpA family transposase gene from *Rhizobium* sp. JDWI01000005.1:83,390–83567. Another fragment (called IS2) with 79% identity to a part of an *IS5*-like transposon gene from *R. rhizogenes* A4 (VCBD01000008:167,938–168,233) was found immediately adjacent to the left and right border (che1CaTE:987–1277 and che1CaTE:41,495–41,785). IS1 and IS2 flank the *mas1'-mas2'* region and were probably part of the original T-DNA. The central part of che1CaTE is most similar to JAAMCX01000012 from *R. rhizogenes* T155-95, whereas *acs* and *mas* are more similar to sequences from *R. rhizogenes* A4. This reflects the well-known chimeric nature of the T-DNAs (Otten 2021).

The *acs* and *orf3n* genes, one copy of *rolB* and *mas1'*, and two copies of *orf13* and *orf13a* are intact. che1CaTE was used as a query to investigate its presence in other accessions. Significant parts of the CaTE insert were found in *C. sasanqua* sas1. *C. japonica* cv. 'Huaheling' jap4CaTE is very similar to sas1CaTE. The *C. sasanqua* reads are not sufficient for complete reconstruction, but the simplest model shows three deletions in sas1CaTE as compared to che1CaTE (Fig. 1, del1-3). The left part of LA is missing (del1), with a shift of the left border to the left (now corresponding to JAKJMZ010000013.1:68,158,951). The central part is also modified. An unusual CaTE read in *C. japonica* 'Naidong' (SRR17085750.18983087) suggests a 5193 nt internal deletion (a region which contains *acs*, *orf2* and *orf3n*), probably due to homologous recombination between the inverted repeats, which removes the unique region and 700 nt each of the adjacent inverted repeat (del2). Another deletion in sas1CaTE occurs in the RA *mas2'* gene (del3).

CaTE traces were found in *C. sasanqua* cv. 'Xiaomeigui' and 'Zhaohe Zhi Rong', and in *C. japonica* cv. 'HXZ', 'Naidong 768', 'red,' 'Chidan', 'Fendan' and 'Yudan'. Ten *C. chekiangoleosa* accessions (Suppl. Table S1) contain large numbers of CaTD reads, but no CaTE reads. Thus, the presence of CaTE in *C. chekiangoleosa* Hu (che1CaTE) is an exception in this species. It may be derived from another *Camellia* species by hybridization.

CaTF

Reads in several *C. oleifera* (sect. *Paracamellia*) accessions showed homology to CaTD, but differed from CaTD. Those from *C. oleifera* cv. XL210 (ole4) were assembled and extended into an 8 kb cT-DNA sequence, ole4CaTF (Fig. 1, Table 1, Suppl. Fig. S2). ole4CaTF contains two plant DNA inserts, called pl1 and pl2. The overall CaTF structure is:

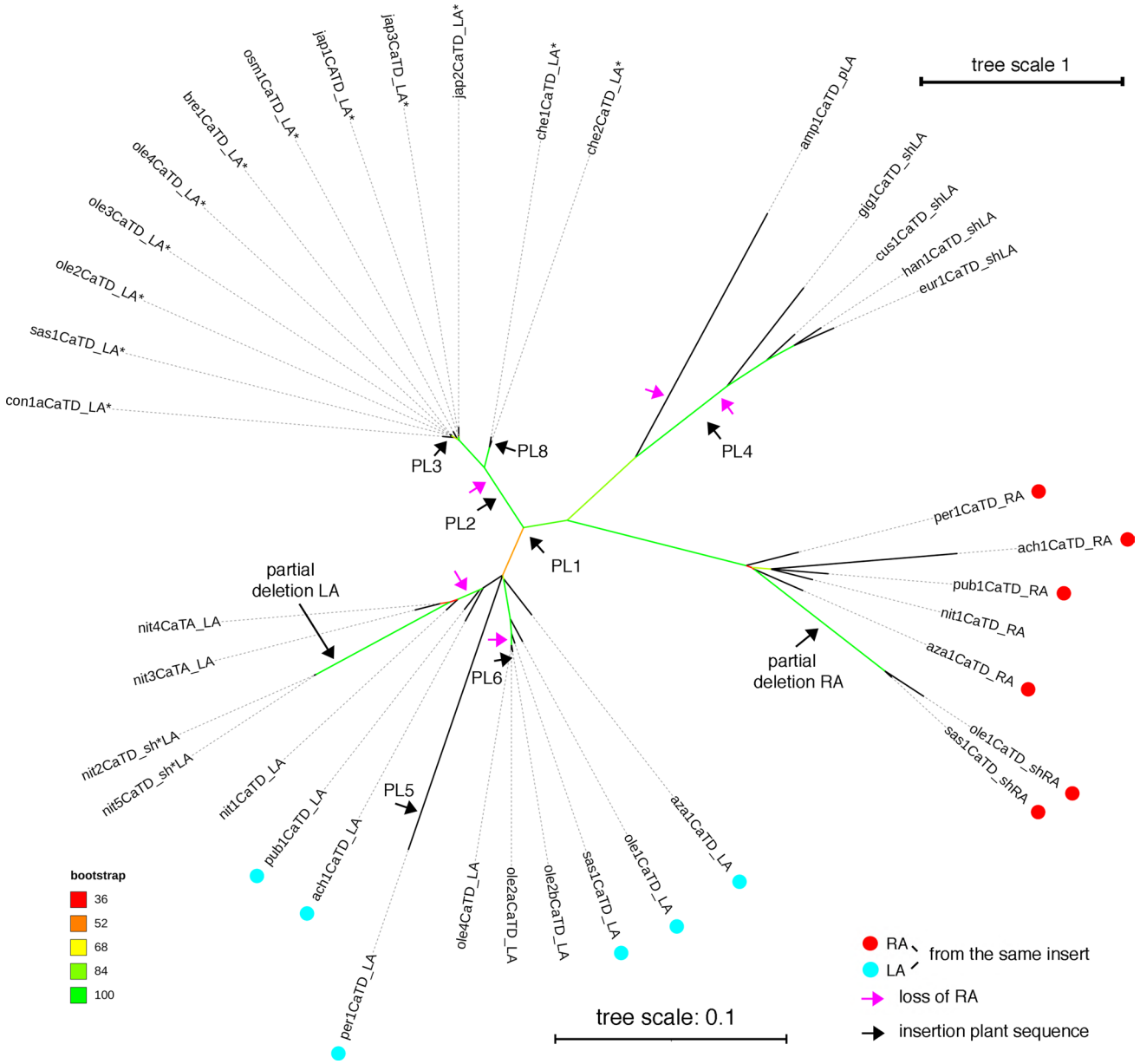

Fig. 5 Phylogenetic tree of the left and right arm of CaTD from different accessions. The original CaTD insert contains the LA and RA sequences, this LA-RA combination is conserved in the sequences indicated with blue (LA), and red dots (RA). The other LA sequences lost the accompanying RA sequences by five different deletion events, indicated with pink arrows. The different plant insertion sequences (PL1 to PL8) are also indicated by arrows

orf8 interrupted by pl1, *orf511*, *susL-1*, interrupted by pl2, and a second *susL*-like gene, *susL-2* (72% identity to *susL-1*). The *R. rhizogenes* LMG152 TL-DNA (Otten 2021) also carries two dissimilar *susL* genes in the same order, but their sequences are different. The ole4CaTF *orf8* and *orf511* genes are truncated at both ends. A search for CaTF in a large collection of cultivated and wild *C. oleifera* accessions showed that CaTF was present in 14 out of 68 accessions (Suppl. Table S1). In addition, CaTF-like sequences were found in one out of two *C. kissii* accessions, but not in other *Camellia* species. *C. kissii* has been considered as a variant of *C. oleifera* by Sealy (1958).

CaTG

In order to search for cT-DNAs in sect. *Theopsis*, we sequenced different accessions of *C. transarisanensis* (Materials and methods). In one of these (tra3, SRR22071522), we found a new cT-DNA, designated tra3CaTG (41.5 kb, Fig. 1, Table 1, Suppl. Fig. S2). It

consists of a partial, inverted repeat of the LB-RB-RB-LB type. Part of LA and RA could be aligned with *R. rhizogenes* T155/95T-DNA. Another region is similar to the nGMO *N. tomentosiformis* TD region (KJ599829.1). The remaining regions showed no detectable DNA homology to either plant or *Agrobacterium* sequences, but BLASTX analysis showed homology to various T-DNA proteins.

tra3CaTG carries the following T-DNA genes: *acs*, *orf8*, *rolB*^{TR}, *orf14-1, orf2*, *orf3n*, *orf13*, *orf13a*, *orf14-2*, and *cus*. This represents a new type of T-DNA with two different *orf14* genes (30% identity at the protein level), and an *orf2-orf3n* fragment in an unusual position. For a discussion on the origin of the *orf2-orf3n* region, see below. It should be noted that only ten different *R. rhizogenes* T-DNAs have so far been described (Otten 2021). The variability of T-DNA structures and sequences in *Agrobacterium* and in nGMOs, clearly suggests that many *R. rhizogenes* strains remain to be discovered. Two copies each of *orf8*, *rolB*^{TR}, *orf14-1* and *orf3n*, and one copy of *orf13*, *orf13a*, *orf14-2* and *cus* are intact and could code for active proteins (see below). The CaTG structure varied in different *C. transarisanensis* accessions, these variants are shown in Suppl. Fig. S4.

CaTG was also found in other species from sect. *Theopsis* (Suppl. Table S1), but is lacking in one accession of *C. cuspidata* and the two accessions of *C. lutchuensis* (sect. *Theopsis*). *C. euryoides* eur1CaTG is similar to tra3CaTG. Its repeats are 99.3% identical. The high similarity between tra3CaTG and eur1CaTG on the one hand, and between LA and RA from each of the two species on the other hand, show that CaTG is of recent origin. CaTG is also found in sect. *Eriandria*, but not in other sections. Thus, CaTG must have been introduced in the immediate ancestor of these two sections. This corresponds well to the fact that sect. *Theopsis* and *Eriandria* are considered to be closely related (Vijayan et al. 2009; Zhang et al. 2021a).

Five *C. japonica* hybrids also have CaTG-like sequences (Suppl. Table S1). They are over 99% identical to each other and 98.3% identical to tra1CaTG. Each hybrid contains DNA from *C. synaptica* var. *parviovata* (sect. *Theopsis,* International *Camellia* Register; Zhang et al. 2021a). We therefore, assume that these CaTG sequences originated from the sect. *Theopsis* partner. The CaTG-like reads from the *C. japonica* hybrids were assembled into 16 contigs. They aligned with the *rolB*^{TR}*-orf8-acs* part of the right arm of tra1CaTG (Fig. 1). The remaining sequence is missing. Inspection of the right-most contig showed that the *C. synaptica* sequence diverges from tra1CaTG at tra1CaTG:41,764, which represents the right border. The *C. synaptica* CaTG-like sequence is thus inserted in another plant sequence (see below). It was therefore called synCaTG* (Fig. 1).

CaTH

A new cT-DNA was found in *C. transarisanensis* tra4 and tra6. Two different units were found: CaTH-a (24.5 kb, several copies) and CaTH-b (9 kb, one copy), assembled together in one long fragment (Fig. 1, Table 1, Suppl. Fig. S2). The gene order of CaTH-a is *orf358* (a new type of T-DNA gene, see below), *orf2, orf3n, orf8-1, rolA, rolB, rolC, orf13, orf13a, orf8-2, rolBTR,* and *susL.* All, except *orf8-1* and *rolA*, are intact ORFs. The *orf8-1* gene shows a stop codon at 6327 in all copies of the CaTH-a unit, which was therefore present in the original T-DNA. The truncated *orf8-1* gene would still allow the synthesis of a 281 AA N-Orf8 protein with potential biological activity (see below). The predicted Orf8-1 and Orf8-2 proteins show only 45% similarity. Long read SRR22198564.1153256 from tra4 carries the single CaTH-b unit, linked on the left to the RB of a CaTH-a unit, and on the right to the LB of the next CaTH-a unit. As the long reads are imprecise, the sequence of CaTH-b was assembled using the tra6 short reads (Fig. 1, Suppl. Fig. S5). CaTH-b shows a 1.5 kb fragment on the left (FR, Fig. 1) with 68% identity to the right end of the *A. vitis* CG412 TB region (Group IIa, Otten 2021). This fragment includes a partial *IS66*-like sequence, and two *vis* gene fragments. It is followed by the intact genes *e-d-c'-c*. The *e-d-c'-c* gene combination is found (in the reverse order) on the T-DNA of Group IIb and IIf (Otten 2021) of *A. tumefaciens* strains C58 (which includes the well-known nopaline strain C58), but has not been reported in *R. rhizogenes* or nGMOs so far. The structure of the complete insert is shown in Suppl. Fig. S5a. To the right of the *e-d-c'-c* region, a 2 kb junction region (intercal 1, Suppl. Fig. S5b), composed of two short CaTH-a fragments, links this unique fragment to the LB of the neighboring CaTH-a unit. A second rearranged sequence composed of three short CaTH-a fragments (intercal 2, Suppl. Fig. S5b) is found more to the right. The CaTH-a unit from tra6 shows only 14 SNPs, showing the very recent origin of this insert. A short, partial CaTH fragment from the *susL*-RB-LB region was found in tra5. No other *Camellia* accessions contained CaTH sequences.

CaTI

A new type of cT-DNA, CaTL, was found in *C. impressinervis*, *C. huana*, *C. pingguoensis*, *C. fascicularis*, and *C. aurea* (all from sect. *Archecamellia*) and *C. insularis* (not yet attributed to a section) (Fig. 1, Table 1, Suppl. Table S1). Only few reads were found per species, but they were over 99% similar. Two separate fragments could be assembled (called CaTI-a and CaTI-b, Fig. 1, Suppl. Fig. S2). CaTI-a (2010 nt) contains an *acs*-like region (129–1367) with 77% identity to the *acs* region of a cT-DNA from nGMO *Ailanthus altissima* (OX327691.1). CaTI-b (522 nt) contains a

gene *d* sequence (78% identity to gene *d* from *A. tumefaciens* strain CFBP5499). No flanking plant sequences could be detected and further sequencing is required to complete this cT-DNA. The limited distribution of CaTI in sect. *Archecamellia* (only five out of 63 accessions) suggests the existence of a subgroup of at least five *Archecamellia* species derived from an ancestor carrying CaTI. The occurrence of CaTI sequences in *C. insularis* suggests that the latter may belong to sect. *Archecamellia*.

CaTJ

A new, incomplete cT-DNA fragment (called CaTJ) was found in reads of *C. oleifera* 'Qionghai No.1'. This is a wild accession from Hainan, China. Assembly yielded a 2270 nt fragment with several gaps (ole5CaTJ, Fig. 1, Table 1, Suppl. Fig. S2), due to insufficient coverage. The sequence shows homology to *orf2*, *orf3n*, *rolB*, *orf13*, and *orf14*. The CaTJ sequences are provisionally arranged in the same order as they are normally found in *R. rhizogenes*. None of the other 67 *C. oleifera* accessions were found to contain CaTJ sequences. The CaTJ insert needs further sequencing.

CaTK

A DNA fragment with an *orf8*-like sequence could be assembled from one *C. amplexicaulis* and two *C. azalea* x *amplexicaulis* hybrid sequences. *C. amplexicaulis* has been placed in sect. *Cylindraceae*, *C. azalea* belongs to sect. *Camellia*. The assembled 3.7 kb sequence (Fig. 1, Table 1, Suppl. Fig. S2) was called amp1CaTK. BLASTN analysis did not detect similarity to *Agrobacterium* and *Rhizobium* sequences. However, BLASTX analysis shows that amp-1CaTK:543–2753 potentially encodes an Orf8-like protein. No CaTK was found in *C. azalea* (Suppl. Table S1). This indicates that CaTK originates from *C. amplexicaulis*. A third *C. amplexicaulis* accession lacks CaTK sequences. Additional sequencing is required to complete CaTK.

CaTL

C. lutchuensis (sect. *Theopsis*) was found to contain CaTC-like sequences. However, these differed from sas1CaTC by 10%, which was higher than expected on the basis of the sas1CaTC internal repeat divergence (3%). We therefore assumed that these *C. lutchuensis* sequences belonged to another cT-DNA. Very similar sequences (over 99% identical) were also found in *C. japonica* accession XF, which results from hybridization with *C. lutchuensis* (see above, CaTA subchapter). Thus, the CaTC-like sequences of *C. japonica* XF are most likely derived from *C. lutchuensis*. By combining all reads, we obtained a partial, 16.8 kb assembly (lutCaTL, Fig. 1, Table 1, Suppl. Fig. S2). It contains gene *c*, interrupted by an *orf14*-like sequence, *orf14-1*, followed by *orf8*, *rolB*, *plast*, *orf13*, *orf13a*, and *orf14-2*. The predicted Orf14 proteins are only 29% similar. Orf14-1 most resembles TD-Orf14 from nGMO *N. tomentosiformis* (70%, AIM40184.1), Orf14-2 is similar to Orf14 from *R. rhizogenes* T155/95 (71%, NTF91502.1). CaTL is incomplete and requires further sequencing.

Possible further cT-DNAs

C. caudata (sect. *Eriandria*) (PRJNA 934752) showed 18 reads (totalling 2.7 kb) with homology to *acs*, *orf3n*, *orf8*, *orf14* and *mas1'*. These could not be attributed to other cT-DNAs. This indicates that *C. caudata* contains at least one other cT-DNA, which remains to be investigated.

Insertion sites of *Camellia* cT-DNAs

Due to the conservation of large parts of nuclear DNA sequences in different *Camellia* species, it is possible to identify sequences corresponding to the unmodified insertion site of a cT-DNA, by using long reads or assembled genomes from related species which lack the cT-DNA. For this, we used the assembled WGS sequence of *C. oleifera* cv. CON1 (JAKJMZ), which contains only one cT-DNA, ole1CaTD. JAKJMZ010000001 refers to chromosome 1, and for simplification, JAKJMZ0100000 was abbreviated by "x". ole1CaTD from JAKJMZ is located at × 02:187,476,098–187462012, i.e. on chromosome 2. The JAKJMZ equivalents of the insertion sites of the other cT-DNAs, and the size of the fragments deleted upon insertion are listed in Table 1. The results clearly show that the inserts are located in different parts of the *Camellia* genome, further confirming that they result from independent transformation events.

Divergence of internal repeats

In six of the 12 cT-DNAs we found repeated T-DNA sequences. These likely result from insertion of two copies of the same sequence at the time of insertion. After each insertion, the cT-DNA sequences diverged over time. Divergence values for the *Camellia* cT-DNAs are 10%, 9%, 3%, 2.8%, 1.4%, 0.7% and 0.05% for CaTA, CaTD, CaTC, CaTB, CaTE, CaTG, and CaTH, respectively. These values indicate that CaTA and CaTD are the oldest inserts, followed by CaTC and CaTB, and then by CaTE and CaTG. The CaTH insert is very recent. Based on the

estimated 1% divergence in 1.5 Mio yrs (Chen et al. 2022), CaTH would be 75.000 yrs old. CaTF has no repeats, apart from two *susL* sequences (72% identity). These two *susL* genes were undoubtedly part of the T-DNA in the *Rhizobium/Agrobacterium* strain that carried the CaTH sequences. The TL region of *R. rhizogenes* strain LMG152 also contains two *susL* genes (Otten 2021). These are oriented in the same way as on CaTH, but their sequences are different. In both LMG152 and CaTH, the two *susL* sequences result from *susL* gene divergence during *Agrobacterium* evolution. The remaining sequences (CaTI, CaTJ, CaTK and CaTL) require more data to establish whether they contain repeats or not.

cT-DNA proteins from *Camellia*

The *Camellia* cT-DNA-encoded proteins can be grouped according to their function into four different groups (Table 2). The largest is the Plast group, they comprise N-Orf8, RolB^{TR}, RolB, Orf13, Orf14, E, D, and C'. RolB and RolB^{TR} have different growth effects (Lemcke and Schmülling 1998). N-Orf8 (the 250 AA N-terminal part of Orf8) has homology to RolB (Levesque et al. 1988) and induces starch accumulation in leaves (Otten and Helfer 2001; Umber et al. 2002, 2005). These Plast proteins are expected to affect growth in different ways (Otten 2018), but further studies are required to show their role in *Camellia*. Even if the genes are intact, they could be expressed in a limited set of cell types, be silenced, or carry mutations that abolish the protein activity. The Plast proteins from *Camellia* significantly extend the Plast protein sequence range.

A second group is formed by the opine synthases: Acs, Mas1', Cus, SusL, and Orf358. A third group contains several unrelated T-DNA proteins with unknown function: Orf2, Orf3n, Orf13a, Orf511, and protein C. The fourth group contains several intact Orf8 proteins. These code for tryptophan monooxygenase, and could lead to indoleacetamide synthesis. The amp1CaTK Orf8 protein is highly unusual: its N-terminal part is only 40% similar to RolB from *R. rhizogenes* T5/73, the C-terminal part is 45–48% similar to bacterial IaaM sequences, which do not carry the N-terminal RolB part, like ATO31311.1 from *Dickeya dianthicola* or the bacterial (non-T-DNA-encoded) IaaM protein QTG17270.1 from *A. tumefaciens* Q15/94. Thus, amp1CaTK *orf8* is a combination of an unusual *rolB*-like gene and an unsual bacterial *iaaM* gene. Its evolutionary origin seems different from the other *Agrobacterium orf8* and *iaaM* genes. tra6CaTH Orf8-2 is also unusual, with only 46% similarity to Orf8 (NTF91497.1) from *R. rhizogenes* strain T155/95.

The high diversity of the T-DNA proteins points to their high antiquity. The presence of *Agrobacterium* T-DNA sequences in a member of the Pteridophytes, *Adiantum nelumboides* (JAKNSL020005917.1:2,140,000–2,225,000), indicates that *Agrobacterium* may have transformed plants well before the emergence of the Angiosperms.

Discovery of new T-DNA protein homologs

The role of several T-DNA proteins remains unknown, due to the lack of homologs with a known function. However, searches carried out in the present study led to the discovery of homologs outside the T-DNA context, which can provide new possibilities to study their function. One example is Orf358. Its gene was found close to the left border of tra6CaTH. The predicted protein has 64% similarity to protein MCZ7448279.1 from the recently sequenced *R. rhizogenes* strain ST15.13.057 (accession JAPYZZ010000006.1, 358 AA), annotated as "hypothetical protein" (HP). We propose to call it Orf358. The *orf358* gene of ST15.13.057 is followed by the classical T-DNA sequence *orf2-orf3n-orf8-rolB-rolC-orf13-orf14-susL* (the corresponding proteins were all annotated as HP). On tra6CaTH, *orf358* is also followed by *orf2-orf3n-orf8-rolB-rolC-orf13*. Thus, both on tra6CaTH and the ST15.13.057 T-DNA, *orf358* replaces the usual *acs* gene. No functional domains were found in Orf358. Similar genes are present in four other members of the Rhizobiaceae (Fig. 6). The corresponding proteins were annotated as HP. In these four strains, the genes are not accompanied by other T-DNA genes, suggesting a bacterial function. Remarkably, Orf358-like proteins are also found in eight different fungi (Fig. 6). PSI-BLAST analysis using MCZ7448279.1 as query, revealed several phosphodiester glycosidases in the second iteration, for example WP_150671122.1 from the bacterium *Pandoraea anhela* (20% similarity). Thus, *orf358* from ST15.13.057 and CaTH may encode a new opine synthase. The T-DNA protein Acs (agrocinopine phosphodiesterase) is also a phosphodiester glycosidase, but Orf358 and Acs show no similarity.

Another enigmatic T-DNA protein is protein C, the gene is found on CaTB, CaTC, CaTE, CaTH, and CaTL. We noted previously that gene *c* also occurs in two fungi, *Pestalothiopsis fici* and *Melampsora larici-populina*, possibly introduced through natural transformation by *Agrobacterium* (Chen et al. 2014). A recent homology search found that gene *c* sequences not only occur on T-DNAs or cT-DNAs, but also in several free-living *Rhizobium* strains outside a T-DNA context. Remarkably, in those cases, gene *c* is associated with an opine synthase gene. *R. leguminosarum* JHI54 gene *c* is linked to an *acs* gene. *R. leguminosarum* Cermik105A and *Sinorhizobium meliloti* M270 also carry gene *c* and an *acs* gene. In *R. nepotum* 39/7, gene *c* is linked to a vitopine synthase (*vis*) gene. Thus, gene *c* could have an function in opine synthesis, degradation or transport, in free-living bacteria. A phylogenetic tree of these C protein sequences is shown in Suppl. Fig. S6.

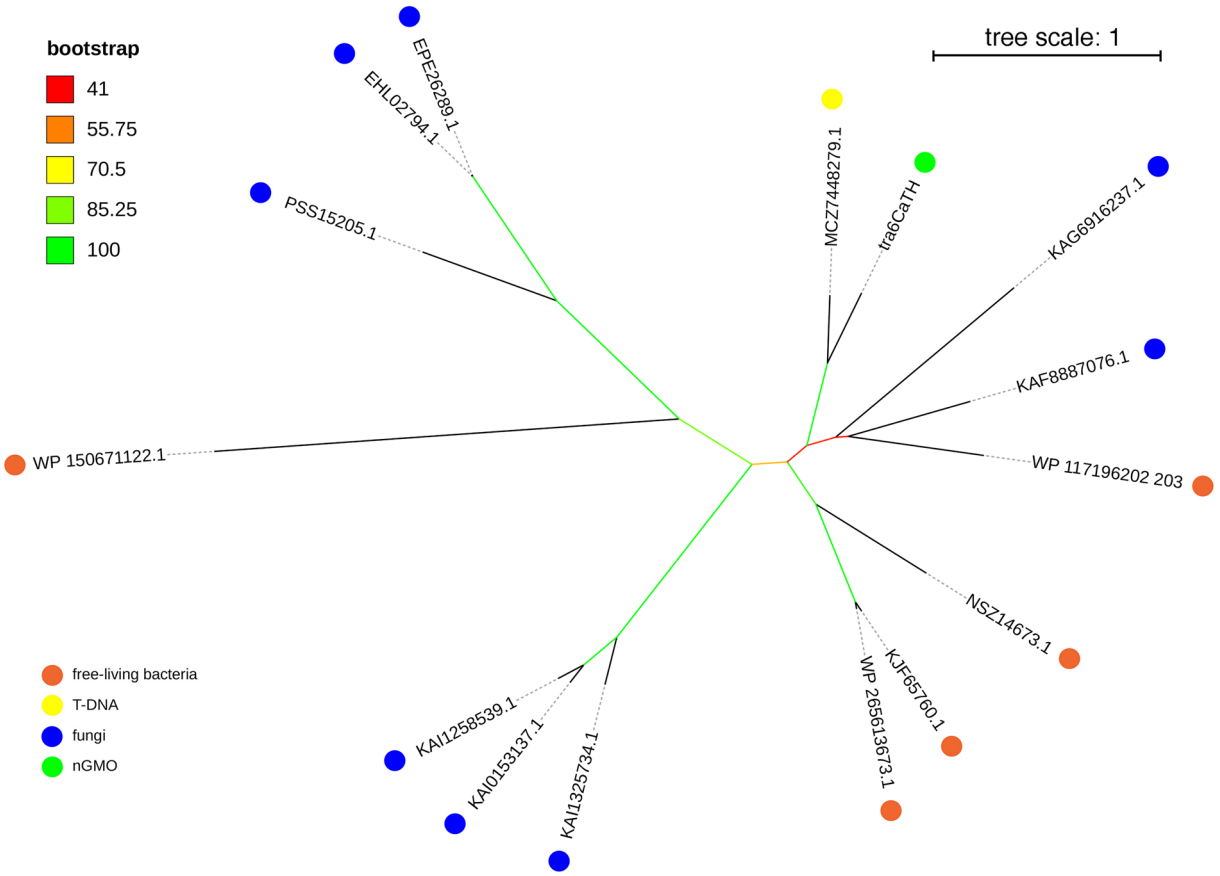

Fig. 6 Phylogenetic tree of Orf358-like proteins. Genes coding for Orf358-like proteins occur in fungi (*Amorphotheca resinae*, PSS15205.1; *Glarea lozoyensis* 74030, EHL02794.1; *Glarea lozoyensis* ATCC 20868, EPE26289.1; *Infundibulicybe gibba*, KAF8887076.1; *Xylariaceae* sp. FL0255, KAI1325734.1; *Xylariaceae* sp. FL1272, KAI0153137.1; *Xylariaceae* sp. FL1019, KAI1258539.1; *Tephrocybe rancida*, KAG6916237.1), in free-living bacteria (*Pandoraea anhela*, WP_150671122.1; *Rhizobium terrae*, WP_117196202 + 203; *R. nepotum*, NSZ14673.1; *A. fabrum* ID132, WP_265613673.1; *A. tumefaciens* MD_2022a, KJF65760.1), on one T-DNA (*R. rhizogenes* ST15.13.057, MCZ7448279.1), and in one nGMO (tra6CaTH, *C. transarisanensis*)

Another rare T-DNA protein gene, found in CaTD and CaTF, is *orf511*. It was first identified in one of the cT-DNAs of nGMO *N. tomentosiformis* (Chen et al. 2014), and later in cT-DNAs from nGMOs *Diospyros lotus*, *Parasponia andersonii* (Matveeva and Otten 2019), and *Quillaja saponaria* (KAJ7971940.1). This gene also occurs in Paracoccaceae bacteria (MCU0909928, 751 AA, annotated as HP, 30% similarity), suggesting a bacterial function. It has not yet been found in *Agrobacterium/Rhizobium*.

Finally, we recently found the first *orf2-orf3* sequences outside a T-DNA context, in *Mesorhizobium* (SAOB01000027.1). Although DNA sequence identity is undetectable, the predicted proteins RWI50086.1 and RWI50087.1 (both annotated as HP) show low but significant similarity to Orf3n (30%) and Orf2 (28%).

Distribution of cT-DNAs among *Camellia* sections

Analysis of all available *Camellia* sequences, using CaTA to CaTL as queries, allowed us to study their distribution among 12 out of 14 *Camellia* sections (Suppl. Table S1, Table 3). CaTA and CaTD are more widely distributed than the other cT-DNAs. However, sect. *Camellia* and *Paracamellia* lack CaTA, whereas sect. *Thea* (Chen et al. 2022 and this study), *Tuberculatae*, *Calpandria*, and *Longipedicellatae* lack CaTD. The distribution of the other cT-DNAs is much more limited. No data are available for sect. *Corallinae* and *Piquetia*. Figure 7 shows the relationship between the divergence of the internal repeats in the cT-DNAs (indicative for their order of introduction and age) and the cT-DNA distribution over different sections.

Table 3 Distribution of cT-DNAs in different *Camellia* sections

Section	A	B	C	D	E	F	G	H	I	J	K	L
Archecamellia (63)	61			61					5			
Cylindraceae (4)	4			4							3	
Theopsis (30)	19			13			25	3				3
Eriandria (7)	4			1			7					
Thea (148)	148											
Longipedicellatae (1)	1											
Calpandria (1)	1											
Tuberculatae (10)	10											
Heterogenea (9)	8	1		3								
Stereocarpus (4)	4			4								
Camellia (49)				49	8							
Paracamellia (107)		43	17	107	3	15				1		

Normal: subgenus Thea, bold: subgenus Camellia. Numbers indicate the numbers of accessions for each section and cT-DNA. Between parentheses: total number of accessions per section. Two accessions (PRJNA861867, *C. insularis* and *C. minima*) could not be attributed to a section. Total number of accessions: 435

Discussion

Various species of the genus *Camellia* carry 12 different cT-DNAs, CaTA to CaTL, which result from multiple transformation events. All 72 investigated *Camellia* species are natural transformants, showing the prevalence of genetic transformation in this genus. Although we do not have a complete overview because of insufficient coverage in different *Camellia* species, the overall pattern is fairly clear. CaTA and CaTD were inserted early, as shown by the divergence of their repeats (about 10%) and wide distribution. However, CaTA is lacking in sect. *Camellia* and *Paracamellia*, and the corresponding insertion sites are intact. Previously, we estimated that the CaTA insert is about 15 Mio yrs old. As the *Camellia* group arose some 14.3 Mio yrs ago (Wu et al. 2022), this suggests that sect. *Camellia* and *Paracamellia* separated at a very early stage from the other sections. The case of CaTD seems different. It is absent from sect. *Thea*, *Longipedicellatae*, *Calpandria*, and *Tuberculatae*. However, in sect. *Thea*, this is not due to the absence of transformation, but to a CaTD deletion event (Chen et al. 2022 and this work). More data are needed to establish whether the same is true for the other sections. In some accessions of sect. *Archecamellia*, *Theopsis*, *Eriandria*, and *Heterogenea*, CaTA or CaTD are missing, but present in others, again suggesting deletion events.

Compared with CaTA and CaTD, the other cT-DNAs show less divergence between their internal repeats, indicating the following ages: 4.6 Mio yrs for CaTB and CaTC (about 3% divergence), 2.0 Mio yrs for CaTE (1.4%) and 1.1 Mio yrs for CaTG (0.7%). CaTH is about 75.000 yrs old (0.05% divergence), and 14 of its 16 genes are intact. Its sequence is still very similar to the T-DNA from the *R. rhizogenes* strain that introduced this cT-DNA in a *C. transarisanensis* plant, and the original CaTH-a sequence could be entirely reconstructed from the consensus sequence of its repeats. *C. transarisanenesis* is endemic to Taiwan (Internal *Camellia* Register), and it might be possible to study the geographic origin and distribution of plants containing CaTH.

Our data provide new insight in the origin and evolution of nGMOs. We have argued (Chen and Otten 2017; Chen et al. 2022) that the appearance of an nGMO requires different steps: the induction of hairy roots, the spontaneous regeneration of a hairy root into a fertile nGMO plant, and the survival of the descendants over evolutionary time. Transformed plants could rapidly replace nontransformed siblings if they are sufficiently different to cause reproductive isolation. The present study shows that new nGMO plants might not always be able to rapidly establish a population with a genetically fixed cT-DNA insert. Indeed, the age and distribution of the CaTB and CaTC inserts indicate that it may take millions of years for a cT-DNA to become fixed. The rapidity of fixation probably depends on the selective value of the cT-DNA. Crosses between closely related accessions containing either a cT-DNA or not, may yield more information on this. Another conclusion from these findings is that several accessions must be investigated before one can state that a given species never underwent genetic transformation.

From our *Camellia* results, it appears that the approximate average natural transformation frequency in *Camellia* is about one event every 1.25 Mio yrs (12 cT-DNAs in 15 Mio yrs), but this figure could change as more *Camellia* species are sequenced. The *Camellia* data hint at the possibility that many plant species could be mixtures of transformed and nontransformed individuals. This means that if nGMO detection is based on one or a few accessions (as in

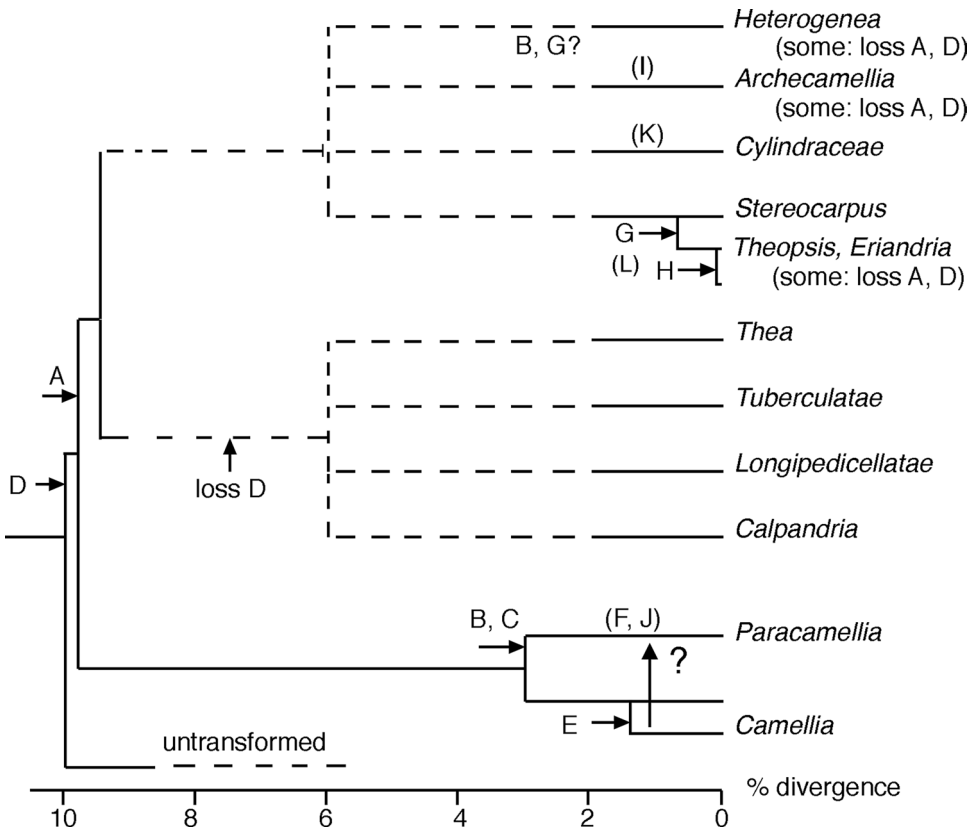

Fig. 7 Hypothetical order of cT-DNA insertion events in the genus *Camellia*. The cT-DNA sequences with internal repeats (CaTA, CaTB, CaTC, CaTD, CaTE, CaTG, and CaTH) can be ordered according to the divergence values of their repeats, the other cT-DNAs, shown within parentheses (CaTF, CaTI, CaTJ, CaTK, CaTL) cannot. Starting on the left, the initial *Camellia* species were non-transformed. After insertion of CaTD (upper branch, arrow), the remaining untransformed plants (lower branch) were lost (or their descendants have not been found yet). One CaTD-containing plant was then re-transformed with the CaTA insert (upper branch), while other CaTD-containing plants (lower branch) were not, generating sect. *Paracamellia* and *Camellia*. These two sections acquired further cT-DNAs: CaTB, CaTC, CaTF, and CaTJ for sect. *Paracamellia*, and CaTE for sect. *Camellia*. Possibly, CaTE inserts in sect. *Paracamellia* were obtained from sect. *Camellia* by hybridization (arrow with question mark). After acquisition of CaTA, sect. *Thea*, *Tuberculatae*, *Longipedicellatae*, and *Calpandria* lost CaTD, while retaining CaTA (lower branch). In the upper branch, five of six sections acquired further cT-DNAs. CaTB and CaTG (*Heterogenea*), CaTI (*Archecamellia*), CaTK (*Cylindraceae*), CaTG (*Theopsis* and *Eriandria*), and CaTH and CaTL (*Theopsis*). Some accessions from sect. *Archecamellia*, *Heterogenea*, *Theopsis*, and *Eriandria* lost CaTA or CaTD

Matveeva and Otten 2019), the overall frequency of species with nGMOs could be severely underestimated. Also, if a single accession of a given species is found to contain a cT-DNA, this species should (strictly speaking) not be designated as an nGMO, but as a species "with nGMOs among its members".

Our data show the dynamic nature of the transformed state, as the genus *Camellia* has both acquired and lost cT-DNAs. Species carrying CaTA and CaTD have persisted for millions of years, with sufficient time for re- transformation, leading to plants with two to four inserts; 35 *Camellia* accessions carry three cT-DNAs (in different combinations), *C. sasanqua* accession CM-1 has four.

A study on *Camellia* cT-DNA distribution and evolution should also consider the problem of interspecific and intersectional hybridization, both natural and man-made, as this will lead to abnormalities in the phylogenetic tree. We detected three cases of cT-DNA transmission by intersectional breeding. The first involves transfer of synCaTG* from *C. synaptica* to five *C. japonica* × *C. synaptica* hybrids, the second, transfer of CaTL from *C. lutchuensis* to a *C. japonica* × *C. lutchuensis* hybrid, and the third, transfer of CATK from *C. amplexicaulis* to a *C. azalea* × *C. amplexicaulis* hybrid. Other cases might be more difficult to detect, for example the replacement of a CaTA variant in species A by another CaTA variant from species B. However, the strong correlation between cT-DNA types and *Camellia* sections indicates that cT-DNA transfer by intersectional hybridization is not a widespread phenomenon.

The oldest cT-DNAs, CaTA and CaTD, have been modified by insertion of different plant sequences. These plant sequences likely originate from transposition, with

different inserts having been introduced at different times and in different evolutionary branches (Fig. 5). They could not be fully reconstructed from short reads, because of their highly repetitive nature. Altogether, we found 12 different plant inserts within CaTA (six of which in sect. *Thea*, Chen et al. 2022), seven in CaTD, two in CaTF, and one in CaTG, none were found in the other cT-DNAs. Their further study could provide some indications which *Camellia* elements are involved, how they modify the insertion sites, and whether they display any target-site specificity.

With the further accumulation of *Camellia* cT-DNA sequence data, more detailed phylogenetic trees can be constructed, using the cT-DNA sequences already available. Analysis of cT-DNA distribution patterns and sequences, including their plant inserts, could be useful to improve the grouping of species and sections. For example, the CaTD sequences of *C. handelii*, *C. cuspidata* and *C. euryoides* (sect. *Theopsis*) are similar to those from *C. crapnelliana* (sect. *Heterogenea*), indicating a close relationship among the four species. Also, the presence of CaTA in sect. *Archecamellia*, *Cylindraceae*, *Theopsis*, *Eriandria*, *Thea*, *Longipedicellatae*, *Calpandria*, *Tuberculatae*, *Heterogenea* and *Stereocarpus* indicates a common ancestor for these sections. Another interesting finding concerns *C. amplexicaulis*. This species carries the unique CaTK sequence with its highly unusual *orf8* gene, and a unique CaTD sequence with a *plast* gene, lacking in all other CaTDs. The precise taxonomic status of this species is not clear (Vijayan et al. 2009; Zhang et al. 2019; Zhao et al. 2019; International *Camellia* register). Further studies on its two unusual cT-DNAs and their occurrence in other *Camellia* species may help clarify its affinity to other species.

The *Camellia* cT-DNAs contain altogether 47 ORFs (Fig. 1). These need to be further investigated for their possible influence on metabolism (as expected for the opine synthesis genes) and development (as expected for the *plast* genes, including *orf8*). The functions of several cT-DNA genes remain unknown. These include *orf358*, *orf8*, *orf13a*, *orf14*, gene *c*, gene *c'*, gene *d*, gene *e* and *orf511*, all found in *Camellia*. The new T-DNA gene *orf358* is found at a position normally occupied by an *acs* gene and might code for an enzyme, as the predicted protein is weakly related to phosphodiester glycosidases. The *orf358* gene occurs on the T-DNA of a single *Agrobacterium* strain, in one nGMO (*C. transarisanensis*), and strikingly, in several fungi. Thus, this gene could have been transferred from fungi to a T-DNA in *Agrobacterium*, and then to *Camellia*, or from *Agrobacterium* to fungi and *Camellia*. The possibility that some fungi were transformed by *Agrobacterium* (Chen and Otten 2017) and represent another group of nGMOs, is also indicated by the remarkable distribution of gene *c* among bacteria and fungi. This distribution is compatible with the possibility that the corresponding genes originated from free-living bacteria, were later incorporated into T-DNAs, and then transferred into various plant nGMOs, including *Camellia*, and into at least two fungal species. Sofar, the role of the fungal cT-DNA-like genes is unknown.

Camellia cT-DNAs encode a number of proteins which are only weakly related to known T-DNA proteins, like CaTH-Orf8-2, amp1CaTK-Orf8, and various Plast proteins. The considerable divergence of these T-DNA protein sequences indicates that *Agrobacterium* strains and their T-DNAs have a very long evolutionary history, and that many T-DNA types (either from *Agrobacterium* or from nGMOs) remain to be discovered (Otten 2021). The genus *Camellia* by itself contains more T-DNAs (12) than are presently known from *R. rhizogenes* (10). The ancient origin of the T-DNA implies that *Agrobacterium* may have generated nGMOs for hundreds of millions of years. In spite of this, none of the nGMOs studied so far shows more than 10% repeat divergence (equivalent to 15 Mio yrs). The lack of very old cT-DNA sequences may be due to their progressive loss, as found in this study.

Further analysis of cT-DNAs in *Camellia* will require the completion of several partial cT-DNAs (CaTI, CaTJ, CaTK, and CaTL). Transcriptional data and promoter analysis are needed to provide insight in the activity of the cT-DNA genes, especially those of the most recent cT-DNAs like CaTH. A large effort will be required to elucidate their biological effects, both on the level of the individual genes, and on the level of the species that carry them. Analysis of rarely studied *Camellia* species will help us to trace these inserts through evolutionary time, and thereby provide new tools to study *Camellia* evolution. We believe that the genus *Camellia* with its numerous cT-DNAs represents an excellent model for the origin, function and evolution of plant nGMOs.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00425-023-04234-9.

Acknowledgements This work was supported by the National Natural Science Foundation of China (82170045 to JH and KC, 31800253 to KC and JH, 32370382 to KC). It was also supported by a special fund for scientific research of Shanghai landscaping and city appearance administrative bureau (G222410), and 中科院青年英才培育计划 (Zhongkeyuan qingnian yingcai peiyu jihua) grant from the Chinese Academy of Sciences to KC. TB was supported by the Interdisciplinary Thematic Institute IMCBio (ITI 2021-2028 program), including funds from IdEx Unistra (ANR-10-IDEX-0002), SFRI-STRAT'US (ANR 20-SFRI-0012) and EUR IMCBio (ANR-17-EURE-0023) in the framework of the French Investments for the Future Program.

Author contribution statement KC and LO conceived and designed research. KC provided material. LO, KC and HL conducted experiments. JH, HL and TB provided bioinformatics support. LO, HL, TB analyzed data. The first draft of the manuscript was written by LO. The manuscript was revised by LO, KC, TB, HL, and JH. All authors read and approved the final document.

Data availability Data generated and analyzed during this work, can be obtained from Léon Otten on request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Human research and animal participant The authors declare that no human and/or animal material, data, or cell lines were used in this study.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Cai Y, Meng J, Cui Y, Tian M, Shi Z, Wang J (2022) Transcriptome and targeted hormone metabolome reveal the molecular mechanisms of flower abscission in *Camellia*. Front Plant Sci 13:1076037. https://doi.org/10.3389/fpls.2022.1076037

Chang HT (1998) Genus *Camellia*. In: Chang HT, Ren SX (eds) Theaceae, Flora Republicae Popularis Sinicae. Sci Press, Beijing, pp 6–194

Chen K, Otten L (2017) Natural *Agrobacterium* transformants: recent results and some theoretical considerations. Front Plant Sci 8:e1600. https://doi.org/10.3389/fpls.2017.01600

Chen L, Yamaguchi S (2002) Genetic diversity and phylogeny of tea plant (*Camellia sinensis*) and its related species and varieties in the section *Thea* genus *Camellia* determined by randomly amplified polymorphic DNA analysis. J Hortic Sci Biotech 77:729–732. https://doi.org/10.1080/14620316.2002.11511564

Chen K, Dorlhac de Borne F, Szegedi E, Otten L (2014) Deep sequencing of the ancestral tobacco species *Nicotiana tomentosiformis* reveals multiple T-DNA inserts and a complex evolutionary history of natural transformation in the genus *Nicotiana*. Plant J 80:669–682. https://doi.org/10.1111/tpj.12661

Chen K, Dorlhac de Borne F, Julio E, Obszynski J, Pale P, Otten L (2016) Root-specific expression of opine genes and opine accumulation in some cultivars of the naturally occurring GMO *Nicotiana tabacum*. Plant J 87:258–269. https://doi.org/10.1111/tpj.13196

Chen K, Zhurbenko P, Danilov L, Matveeva T, Otten L (2022) Conservation of an *Agrobacterium* cT-DNA insert in *Camellia* section *Thea* reveals the ancient origin of tea plants from a genetically modified ancestor. Front Plant Sci 13:997762. https://doi.org/10.3389/fpls.2022.997762

Dessaux Y, Petit A, Farrand S, Murphy P (1998) Opines and opine-like molecules involved in plant-Rhizobiaceae interactions. In: Spaink HP, Kondorosi A, Hooykaas PJJ (eds) The Rhizobiaceae. Kluwer Academic Publishers, Dordrecht, pp 173–197. https://doi.org/10.1007/978-94-011-5060-6_9

Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y et al (2019) Lateral transfers of large DNA fragments spread functional genes among grasses. Proc Natl Acad Sci USA 116:4416–4425. https://doi.org/10.1073/pnas.1810031116

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

Furner IJ, Huffman GA, Amasino RM, Garfinkel DJ, Gordon MP, Nester EW (1986) An *Agrobacterium* transformation in the evolution of the genus *Nicotiana*. Nature 319:422–427. https://doi.org/10.1038/319422a0

Gelvin SB (2017) Integration of *Agrobacterium* T-DNA into the plant genome. Annu Rev Genet 51:195–217. https://doi.org/10.1146/annurev-genet-120215-035320

Hembree WG, Ranney TG, Jackson BE, Weathington M (2019) Genetics, ploidy and genome sizes of *Camellia* and related genera. Hort Sci 7:1124–1142. https://doi.org/10.21273/HORTSCI13923-19

Hooykaas P (2023) The Ti plasmid, driver of *Agrobacterium* pathogenesis. Phytopathology 113:594–604. https://doi.org/10.1094/PHYTO-11-22-0432-IA16

International *Camellia* Register. https://Camellia.iflora.cn. Accessed 10 March 2023

Kyndt T, Quispe D, Zhai H, Jarret R, Ghislain M, Liu Q et al (2015) The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: an example of a naturally transgenic food crop. Proc Natl Acad Sci USA 112:5844–5849. https://doi.org/10.1073/pnas1419685112

Lemcke K, Schmülling T (1998) A putative *rolB* homologue of the *Agrobacterium rhizogenes* TR-DNA has different morphogenetic activity in tobacco than *rolB*. Plant Mol Biol 36:803–808. https://doi.org/10.1023/a:1005905327898

Letunic I, Bork P (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301

Levesque H, Delepelaire P, Rouzé P, Slightom J, Tepfer D (1988) Common evolutionary origin of the central portion of the Ri TL-DNA of *Agrobacterium rhizogenes* and the Ti T-DNAs of *Agrobacterium tumefaciens*. Plant Mol Biol 11:731–744. https://doi.org/10.1007/BF00019514

Lin P, Wang K, Wang Y et al (2022) The genome of oil-Camellia and population genomics analysis provide insights into seed oil domestication. Genome Biol 23:14. https://doi.org/10.1186/s13059-021-02599-2

Ma J, Wang S, Zhu X, Sun G, Chang G, Li L, Hu X, Zhang S, Zhou Y, Song CP, Huang J (2022) Major episodes of horizontal gene transfer drove the evolution of land plants. Mol Plant 15:857–871. https://doi.org/10.1016/j.molp.2022.02.001

Matveeva TV (2021) New naturally transgenic plants: 2020 update. Biol Commun 66:36–46. https://doi.org/10.21638/spbu03.2021.105

Matveeva TV, Otten L (2019) Widespread occurrence of natural transformation of plants by *Agrobacterium*. Plant Mol Biol 101:415–437. https://doi.org/10.1007/s11103-019-00913-y

Min TL, Bartholomew B (2007) Theaceae. In: Wu ZY, Raven PH, Hong DY (eds) Flora of China. Beijing-St Louis Science Press & Missouri Botanical Garden Press, pp 367–478

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015

Nester EW (2015) *Agrobacterium*: nature's genetic engineer. Front Plant Sci 5:730. https://doi.org/10.3389/fpls.2014.00730

Otten L (2018) The *Agrobacterium* phenotypic plasticity (*plast*) genes. Curr Top Microbiol Immunol 418:375–419. https://doi.org/10.1007/82_2018_93

Otten L (2021) T-DNA regions from 350 *Agrobacterium* genomes: maps and phylogeny. Plant Mol Biol 106:239–258. https://doi.org/10.1007/s11103-021-01140-0

Otten L, Helfer A (2001) Biological activity of the *rolB*-like 5' end of the A4-*orf8* gene from the *Agrobacterium rhizogenes* TL-DNA. Mol Plant Microbe Interact 14:405–411. https://doi.org/10.1094/MPMI.2001.14.3.405

Petit A, Tempé J (1985) The function of T-DNA in nature. In: van Vloten-Doting L, Groot G, Hall T (eds) Molecular form and function of the plant genome. Plenum Press, New York, pp 625–636

Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and

polyphenol components. Plant Mol Biol Rep 15:8–15. https://doi.org/10.1007/BF02772108
Sealy JR (1958) A revision of the genus *Camellia*. The Royal Horticultural Society Press
Shen TF, Huang B, Xu M, Zhou PY, Ni ZX, Gong C et al (2022) The reference genome of *Camellia chekiangoleosa* provides insights into *Camellia* evolution and tea oil biosynthesis. Hortic Res 9:uhab083. https://doi.org/10.1093/hr/uhab083
Suzuki K, Yamashita I, Tanaka N (2002) Tobacco plants were transformed by *Agrobacterium rhizogenes* infection during their evolution. Plant J 32:775–787. https://doi.org/10.1046/j.1365-313x.2002.01468.x
Takeda Y (1990) Cross compatibility of tea (*Camellia sinensis*) and its allied species in the genus *Camellia*. Jpn Agric Res Quart 24:111–116
Tamura K, Stecher G, Kumar S (2021) MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38:3022–3027. https://doi.org/10.1093/molbev/msab120
Tepfer D (1990) Genetic transformation using *Agrobacterium rhizogenes*. Physiol Plant 79:140–146. https://doi.org/10.1111/j.1399-3054.1990.tb05876
Trevenzoli Favero B, Tan Y, Chen X, Müller R, Lütken H (2022) *Kalanchoe blossfeldiana* naturally transformed with *Rhizobium rhizogenes* exhibits superior root phenotype. Plant Sci 321:111323. https://doi.org/10.1016/j.plantsci.2022.111323
Umber M, Voll L, Weber A, Michler P, Otten L (2002) The *rolB*-like part of the *Agrobacterium rhizogenes orf8* gene inhibits sucrose export in tobacco. Mol Plant Microbe Interact 15:956–962. https://doi.org/10.1094/MPMI.2002.15.9.956
Umber M, Clément B, Otten L (2005) The T-DNA oncogene A4-*orf8* from *Agrobacterium rhizogenes* A4 induces abnormal growth in tobacco. Mol Plant Microbe Interact 18:205–211. https://doi.org/10.1094/MPMI-18-0205
Vijayan K, Zhang W, Tsou C (2009) Molecular taxonomy of *Camellia* (Theaceae) inferred from nrITS sequences. Am J Bot 96:1348–1360. https://doi.org/10.3732/ajb.0800205
White FF, Garfinkel DJ, Huffman GA, Gordon MP, Nester EW (1983) Sequence homologous to *Agrobacterium rhizogenes* T-DNA in the genomes of uninfected plants. Nature 301:348–350. https://doi.org/10.1038/301348a0
Wight W (1962) Tea classification revised. Curr Sci 31:298–299
Wu Q, Tong W, Zhao H, Ge R, Li R, Huang J et al (2022) Comparative transcriptomic analysis unveils the deep phylogeny and secondary metabolite evolution of 116 *Camellia* plants. Plant J 111:406–421. https://doi.org/10.1011/tpj.15799
Xiao TJ, Parks CR (2003) Molecular analysis of the genus *Camellia*. Int Camellia J 35:57–65
Yang JB, Yang SX, Li HT, Yang J, Li DZ (2013) Comparative chloroplast genomes of *Camellia* species. PLoS ONE 8:e73053. https://doi.org/10.1371/journal.pone.0073053
Zhang M, Tang YW, Qi J, Liu XK, Yan DF, Zhong NS et al (2019) Effects of parental genetic divergence on gene expression patterns in interspecific hybrids of *Camellia*. BMC Genomics 20:828. https://doi.org/10.1186/s12864-019-6222-z
Zhang Y, Wang D, Wang Y, Dong H, Yuan Y, Yang W et al (2020) Parasitic plant dodder (*Cuscuta* spp.): a new natural *Agrobacterium*-to-plant horizontal gene transfer species. Sci China Life Sci 63:312–316. https://doi.org/10.1007/s11427-019-1588-x
Zhang YL, Du C, Hu YH (2021a) Resources of *Camellia* sect. *Theopsis* and sect. *Eriandria* and germplasm innovation. Subtrop Plant Sci 50:323–332. https://doi.org/10.3969/j.issn.1009-7791.2021.04.013
Zhang M, Tang YW, Xu Y, Yonezawa T, Shao Y, Wang YG et al (2021b) Concerted and birth-and-death evolution of 26S ribosomal DNA in *Camellia* L. Ann Bot 127:63–73. https://doi.org/10.1093/aob/mcaa169
Zhang Q, Zhao L, Folk RA, Zhao JL, Zamora NA, Yang SX et al (2022) Phytotranscriptomics of Theaceae: generic-level relationships, reticulation and whole-genome duplication. Ann Bot Lond 129:457–471. https://doi.org/10.1093/aob/mcac007
Zhao DW (2019) New synonyms in *Camellia* (Theaceae): *Camellia cucphuongensis*, *C. cylindracea* and *C. vidalii*. Phytotaxa 419:100–104. https://doi.org/10.1164/phytotaxa.419.1.7
Zhu J, Oger PM, Schrammeijer B, Hooykaas PJJ, Farrand SK, Winans SC (2000) The bases of crown gall tumorigenesis. J Bact 182:3885–3895. https://doi.org/10.1128/JB.182.14.3885-3895.2000