Plant Molecular Biology (2019) 101:415–437
https://doi.org/10.1007/s11103-019-00913-y

Widespread occurrence of natural genetic transformation of plants by *Agrobacterium*

Tatiana V. Matveeva[1] · **Léon Otten**[2]

Received: 18 June 2019 / Accepted: 21 August 2019 / Published online: 21 September 2019

Abstract

Key message **Naturally transgenic plant species occur on an unexpectedly large scale.**

Abstract *Agrobacterium*-mediated gene transfer leads to the formation of crown galls or hairy roots, due to expression of transferred T-DNA genes. Spontaneous regeneration of transformed cells can produce natural transformants carrying cellular T-DNA (cT-DNA) sequences of bacterial origin. This particular type of horizontal gene transfer (HGT) could play a role in plant evolution. However, the material available today is not enough for generalizations concerning the role of *Agrobacterium* in HGT from bacteria to plants. In this study, we searched for T-DNA-like genes in the sequenced genomes of dicots and monocots. We demonstrate the presence of cT-DNAs in 23 out of 275 dicot species, within genera *Eutrema, Arachis, Nissolia, Quillaja, Euphorbia, Parasponia, Trema, Humulus, Psidium, Eugenia, Juglans, Azadirachta, Silene, Dianthus, Vaccinium, Camellia,* and *Cuscuta.* Analysis of transcriptome data of 356 dicot species yielded 16 additional naturally transgenic species. Thus, HGT from *Agrobacterium* to dicots is remarkably widespread. Opine synthesis genes are most frequent, followed by *plast* genes. Species in the genera *Parasponia, Trema, Camellia, Azadirachta, Quillaja*, and *Diospyros* contain a combination of *plast* and opine genes. Some are intact and expressed, but the majority have internal stop codons. Among the sequenced monocot species, *Dioscorea alata* (greater yam) and *Musa acuminata* (banana) also contain T-DNA-like sequences. The identified examples are valuable material for future research on the role of *Agrobacterium*-derived genes in plant evolution, for investigations on *Agrobacterium* strain diversity, and for studies on the function and evolution of cT-DNA genes in natural transformants.

Keywords Naturally transgenic plants · cT-DNA · Horizontal gene transfer · Whole-genome shotgun contigs · Transcriptome shotgun assembly

Introduction

Horizontal gene transfer (HGT) is widespread in prokaryotes. Comparative and phylogenetic analyses of eukaryotic genomes show that a considerable number of eukaryotic genes also result from HGT. However, mechanisms of HGT in eukaryotic organisms are poorly understood in comparison to gene transfer among the Prokaryota. Evidence of gene transfer from bacteria to the nuclei of multi-cellular eukaryotes is rare (Richards et al. 2006; Acuna et al. 2012). The close contacts frequently found in different kinds of symbiosis could promote HGT between species (Gao et al. 2014). One of the best studied examples of natural HGT from bacteria to plants is gene transfer from *Agrobacterium sp.* to plants. *Agrobacterium*-mediated HGT relies on a highly specific DNA transfer mechanism and the introduction of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11103-019-00913-y) contains supplementary material, which is available to authorized users.

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

Tatiana V. Matveeva
radishlet@gmail.com

[1] St. Petersburg State University, University Emb., 7/9, Saint Petersburg, Russia

[2] Institut de Biologie Moléculaire des Plantes, 12 Rue du Général Zimmer, 67084 Strasbourg, France

T-DNA genes with eucaryotic promoter sequences, which can be expressed in a wide variety of plants.

Plants with cellular T-DNA (cT-DNA) sequences can be considered as natural genetically modified organisms (natural GMO's). Species of the genera *Nicotiana* (White et al. 1983; Intrieri and Buiatti 2001; Chen et al. 2014, 2018), *Linaria* (Matveeva et al. 2012, 2018; Pavlova et al. 2013), and *Ipomoea* (Kyndt et al. 2015) contain different types of cT-DNA structures. In *Nicotiana*, six cT-DNA types have been described (TA, TB, TC, TD, TE, gT), with some species having received several cT-DNAs in succession. We have argued that *Agrobacterium*-mediated HGT could create new species (Chen and Otten 2017). cT-DNA expression can confer new properties on natural GMO's, as shown by opine synthesis in roots of *Nicotiana tabacum* (Chen et al. 2016). A search for additional natural transformants in the large number of available sequence data would provide a much better idea about the frequency of *Agrobacterium*-induced HGT, on the type and variation of cT-DNA structures, and on cT-DNA-induced changes in natural transformants.

Most cT-DNA sequences identified so far seem to originate from *A. rhizogenes*, with genes already described in various *A. rhizogenes* strains. However, cT-DNAs may also contain previously unknown T-DNA genes, or unusual combinations of them. *N. tomentosiformis* contains genes distantly related to the *orf14* and agropine synthase (*ags*) genes, and typical *A. tumefaciens* and *A. vitis* genes resembling octopine synthase (*ocs*), vitopine synthase (*vis*), C-protein-like, and *6b*. It also contains a large, previously unknown T-DNA gene, *orf511* (Chen et al. 2014; Chen and Otten 2017), with unknown function. In *Ipomoea*, a gene distantly related to *rolB* and *rolC* was found (Kyndt et al. 2015). These findings suggest that *Agrobacterium* T-DNA diversity is greater than expected. The TBLASTN algorithm, which compares a protein query against a translated nucleotide sequence database, allows the detection of highly diverged gene sequences, which are undetectable at the DNA level. The aim of the present work was to search for new examples of HGT from agrobacteria to plants, using T-DNA-encoded proteins as queries against sequenced plant genomes and transcriptomes.

Results

Identification of dicot cT-DNA sequences in the Whole Genome Shotgun database

To date (April 12, 2019), data on the genomes of 275 species of dicotyledonous plants, including tobacco and sweet potato, are available in the NCBI database (O'Leary et al. 2016). This database was searched with selected T-DNA-encoded protein sequences ("Materials and methods", Table 1). Apart from the earlier studied tobacco and sweet potato genomes (the *Linaria* genome is not published), we found homologues of proteins encoded by agrobacterial T-DNA genes in 23 species belonging to 17 genera, 12 families and 10 orders (Fig. 1). The data on their T-DNA-like sequences are summarized in Table 2. As a control for *Agrobacterium* contamination, we searched for *vir* sequences, which are located outside the T-DNA. No intact *vir*-like genes were found in the genome data of most species listed in Table 2. In *Euphorbia esula*, a VirB1-like sequence was found (30% identical to YP_001967531.1). This segment is 99% identical to WP_125244063.1 from *Aquabacterium* sp. contig PJAD010111136.1. Unexpectedly, the entire contig (56 kb) shows strong identity to various *Aquabacterium* sequences. Conversely, *Euphorbia esula* contains *sus*-like and *orf18*-like T-DNA sequences, but the fully sequenced *Aquabacterium* sp. does not. Therefore, the *Euphorbia* VirB1-like sequence is most likely due to contamination of *Euphorbia* DNA with *Aquabacterium* DNA. In *Parasponia andersonii*, *virH1, virH2,* and *virF* homologs were found, close to the PaT-DNA3 region, and associated with plant sequences, suggesting abnormal co-transfer of these *vir* genes with the T-DNA (see below).

cT-DNA-positive contigs were analyzed in greater detail. Unfortunately, the quality of genome assembly was found to be variable among different species. In some cases (*Eugenia uniflora*, *Euphorbia esula*, and *Silene latifolia*), various contigs were shorter than 1000 base pairs, making it difficult to identify sequences surrounding the cT-DNA fragments, and to identify the original insertion sites. Nevertheless, the presence of homologous fragments in related species or subspecies strongly indicated their integration into the genome of the ancestral form. T-DNA-like sequences, present in groups of related species, were identified within the genera *Arachis*, *Cuscuta* and *Juglans,* in two subspecies of *Humulus lupulus*, in two subspecies of *Silene latifolia*, and in two related genera from the family Cannabaceae: *Parasponia* and *Trema.* In the following part we will describe the new natural transformants in detail, starting with the simplest cT-DNAs, which carry only opine or *plast* genes.

cT-DNAs with only opine genes

Several plant genera contain cT-DNA sequences which carry only opine genes. We will first discuss the well-known genera *Arachis* and *Juglans*, and then present the remaining ones.

Cultivated peanut (*Arachis hypogaea*) is an allotetraploid species whose ancestral genome is most likely derived from the A-genome species, *A. duranensis*, and the B-genome species, *A. ipaensis* (Kochert et al. 1996). *Arachis monticola* is a close relative of the domesticated peanut. It is the only other tetraploid species besides *A. hypogaea* in the

Table 1 Query sequences for searching proteins encoded by T-DNA and *vir* genes

Aim of search	Protein	Accession # of protein	Organism
Search for T-DNA-like sequences	RolB	CAA82552.1	*A. rhizogenes* 2659
		CAA34077.1	*A. rhizogenes* A4
		CAA27161.1	*N. glauca*
	RolC	CAA82553.1	*A. rhizogenes* 2659
		P20403.1	*A. rhizogenes* A4
		P07051.2	*N. glauca*
	RolB/C-like	AIM47586.1	*I. batatas*
	Orf13	CAB65897.1	*A. rhizogenes* 2659
		ABI54192.1	*A. rhizogenes* A4
		BAB85946.1	*N. glauca*
		CAA07584.1	*N. tabacum*
	Orf14	CAB65899.1	*A. rhizogenes* 2659
		ABI54193.1	*A. rhizogenes* A4
		BAB85948.1	*N. glauca*
		AIM40184.1	*N. tomentosiformis* (TD)
	6b	CAB44643.1	*A. tumefaciens* C58
		WP_032488313.1	*A. vitis* Tm4
	B protein	AAD30482.1	*A. tumefaciens* C58
	C' protein	AAD30484.1	*A. tumefaciens* C58
	D protein	AAD30485	*A. tumefaciens* C58
	E protein	ASK49488.1	*A. larrymoorei*
		ARU12465.1	*A. tumefaciens* Chry5
	Gene 5	AAD30487.1	*A. tumefaciens* C58
		AAB41867.1	*A. vitis* CG474
	Fungal Plast proteins	XP_001884963.1 XP_001881215.1 XP_001884861.1 XP_001884964.1 XP_001884962.1	*Laccaria bicolor*
	Nos	CAB44644.1	*A. tumefaciens* C58
	Mis	NP_066601	*A. rhizogenes* 1724
		WP_010900210.1	*N. glauca*
	Ags	ASK40986.1	*A. rhizogenes* CBFP2692
	Mas2′	AIM40180.1	*N. tomentosiformis* (TB)
	Acs	AAK20401.1	*A. tumefaciens* Chry5
	Sus	ARU12438.1	*A. tumefaciens* Chry5
	Ocs	NP_059680.1	*A. tumefaciens* Ach5
	Vis	WP_080855286.1	*A. deltaense*
	Cus	BAB13344.1	*A. rhizogenes* 2659
Study of possible contamination of candidate plant species with *Agrobacterium*	VirB1	WP_080855255.1 ACM39672.1 NP_066734.1 YP_001967531.1	*A. deltaense* *A. vitis* S4 *A. rhizogenes* 1724 *A. tumefaciens* Bo542
	VirB2	NP_066735.1 ACM39671.1 BAA28696.1	*A. rhizogenes* 1724 *A. vitis* S4 *A. tumefaciens*
	VirD2	YP_001967546.1 WP_032488282.1 ACM39658.1	*A. tumefaciens* Bo542 *A. rhizogenes* 15834 *A. vitis* S4
	VirE2	AAA98372.1 GAJ95556 ACM39679.1	*A. tumefaciens* C58 *A. rhizogenes* 13257 *A. vitis* S4

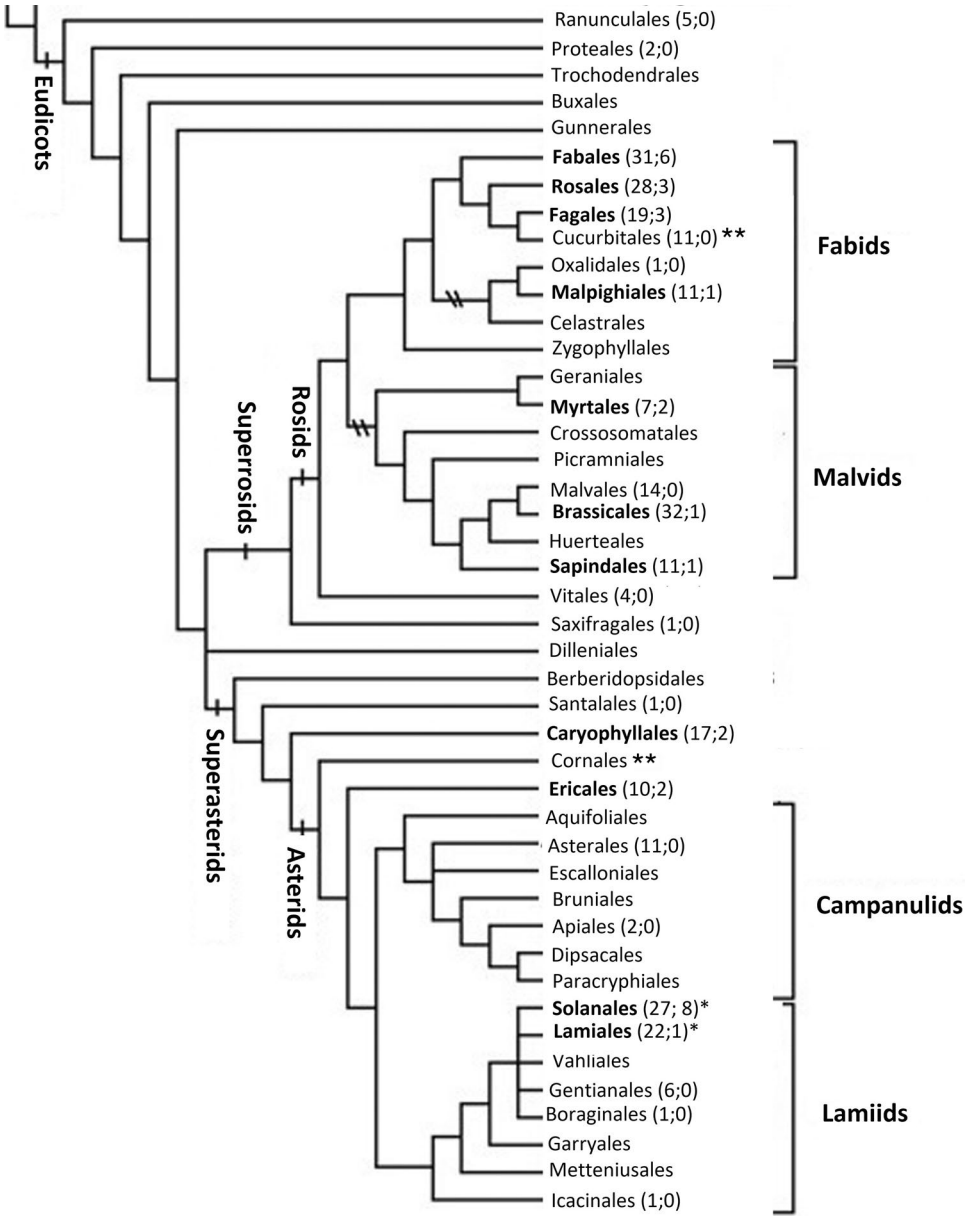

Fig. 1 Number of species with sequenced genomes in different plant orders of Eudicots and number of naturally transgenic species among them. Relation of orders is after APG IV (Angiosperm Phylogeny Group 2016). Orders with naturally transgenic plants are in bold. The numbers in parentheses denote the number of species with sequenced genomes and the number of species with cT-DNA. *Solanales and Lamiales include data on natural transformants described earlier. **Cucurbitales and Cornales only show T-DNA-like sequences in the TSA transcriptome database

genus *Arachis*. *A. monticola* might be an immediate wild ancestor of cultivated peanut (Guillermo et al. 2007), or a weedy form, segregated from cultivated peanuts (Pattee et al. 1998). *A. duranensis* contains two copies of a cucumopine synthase (*cus*)-like gene, one complete, the other truncated. The common parts are 99% identical. The full size *cus* homologs from *A. duranensis* cultivars P1263393 (Table 2) and V14167 are 99% identical. The partially deleted copies differ in length and localization (upstream or downstream) relative to the intact gene. Probably, the A ancestor carried two full-size genes, which diverged over time. *A. ipaensis* contains a strongly rearranged *cus*-like gene, showing a deletion, a replacement of part of the gene, and insertion of a large DNA fragment (0.5 Mb). Analysis of plant sequences next to the cT-DNA suggests that they result from different integration events. Apart from a *cus* gene, *A. ipaensis* also contains a mannopine synthase 2′ (*mas2′*) homolog. Both parents' cT-DNAs are present in the peanut crop genome. However, the cT-DNA derived from *A. duranensis* contains only one full-size copy of the *cus*-like gene (Supplementary Fig. 1). The cT-DNA from *A. hypogaea* cultivar Shitouqi (CP030990.1) is 99% identical to that of cultivar Tifrunner (Table 2).

Juglans cathayensis (Chinese walnut), *J. manshurica* (Manchurian walnut), and *J. sigillata* (iron walnut) also contain cT-DNA sequences with intact, succinamopine synthase (*sus*)-like opine genes. These species are closely related (Stanford et al. 2000; Dong et al. 2017), and their cT-DNAs

Table 2 cT-DNA genes detected by analysis of WGS database

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
Malpighiales	Euphorbiaceae	*Euphorbia esula* (cultivar 1984-ND001 c294543_g1_i1)	PJAE01404665.1	*sus*-like	1	+?	533–1396	67	*A. rhizogenes* WP_034521016.1	n/a
			PJAD010880283.1	*sus*-like	1	+?	743–3	67	*A. rhizogenes* WP_034521016.1	n/a
			PJAE01404667.1	*sus*-like	1	+?	1144–539	68	*A. rhizogenes* WP_034521016.1	n/a
			PJAD011597089.1	*orf18*-like	1	?	1–231	57	*A. rhizogenes* CAB46634.1	n/a
Fabales	Fabaceae	*Arachis duranensis* (cultivar PI475845)	MAMN01003931.1	*cus*-like	2	+	7349–6387	71	*A. vitis* WP_071201425.1	99%
						–	8059–7499 (short)	78	*A. vitis* WP_071201425.1	
		Arachis duranensis (cultivar V14167 Adur16_3)	JQIN01001286.1	*cus*-like		–	2661549–2661427	65	*A. vitis* WP_071201425.1	92%
						+	2667636–2666674	71	*A. vitis* WP_071201425.1	
		Arachis ipaensis (cultivar K30076)	JQIO01000321.1	*cus*-like (interrupted)	1	–	5265248–5264662 5776277–5776119	63	*A. vitis* WP_071201425.1	n/a
			JQIO01000351.1	*mas2′*-like	1	–	6282379–6281948	47	*A. rhizogenes* AIM40180.1	n/a
		Arachis monticola (isolate PI 263393)	QBTX01000020.1	*cus*-like	1	+	24042114–24043064	71	*A. vitis* WP_071201425.1	n/a
			QBTX01000007.1	*cus*-like	1	–	23709857–23709271 137208979–137208908	63	*A. vitis* WP_071201425.1	n/a
				mas2′-like	1	–	106506829–106507260	47	*A. rhizogenes* AIM40180.1	n/a
		Arachis hypogaea (cultivar Tifrunner)	PIVG01000008.1	*cus*-like	1	+	37350716–37351648	71	*A. vitis* WP_071201425.1	n/a
			PIVG01000018.1	*cus*-like (interrupted)	1	–	18314323–18313737 18883031–18882873	63	*A. vitis* WP_071201425.1	n/a
				mas2′-like	1	–	97262939–97262508	47	*A. rhizogenes* AIM40180.1	n/a

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
		Nissolia schottii (isolate NF-2018-10)	QANU01003176.1	*mis*-like	1	–	1123891–1125027	60	*N. glauca* BAB85949.1	100%
			QANU01104125.1	*mis*-like	1	–	1063–1851	56	*N. glauca* BAB85949.1	
			QANU01020083.1	*mis*-like	1	–	29363-30437	64	*N. glauca* BAB85949.1	92% to others
	Quillajaceae	*Quillaja saponaria*	PVLG01028938.1	*rolB*-like	1	–	2976–3443	37	*A. rhizogenes* WP_077768174.1	n/a
			PVLG01027119.1	*orf13*-like	1	–	2–313	41	*A. rhizogenes* GAJ95531.1	n/a
				sus-like	1	–	1194–2071	55	*A. rhizogenes* WP_034521016.1	n/a
			PVLG01028247.1	*e*-like	1	–	1996–2145	31	*A. vitis* WP_060718539.1	n/a
				IS3-like	1	–	999–1518	72	*R. mesoamericanum* CCM79798.1	n/a
Rosales	Cannabaceae	*Parasponia andersonii* (isolate WU1-14)	JXTB01000070.1 (PaT-DNA1) (*orf3-orf8-rolB-orf14-orf14-rolB-orf8-orf3*[a])	*orf3*-like	2	–	786120–787462	53	*A. rhizogenes* CAB65892.1	76%
						–	805921–805195	47	*A. rhizogenes* WP_034521034.1	
				orf8-like	2	–	788784–790884	39	*A. rhizogenes* WP_034521028.1	77%
						–	803946–801979	40	*A. rhizogenes* WP_034521028.1	
				rolB-like	2	–	794071–793757	37	*A. rhizogenes* P09178.1	79%
						–	797794–798417	35	*A. rhizogenes* P09178.1	
				orf14-like	2	–	795016–795448	47	*N. tomentosiformis* AIM40184.1	78%
						–	796755–796222	47	*N. tomentosiformis* AIM40184.1	

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
			JXTB01000448.1 (PaT-DNA2) (*rolB-sus-orf14-IS630-IS630-orf14-sus*)	*rolB*-like	1	–	330011–329197	47	*A. rhizogenes* WP_034521028.1	n/a
				sus-like	2	+	330171–331190	64	*A. rhizogenes* WP_034521016.1	87%
						–	339517–338500	63	*A. rhizogenes* WP_034521016.1	
				orf14-like	2	–	332345–331836	46	*A. rhizogenes* WP_042474756.1	89%
						–	337359–337867	49	*A. rhizogenes* WP_042474756.1	
				IS630-like	2	–	334379–333395	50	Alphaproteobacteria bacterium PA3 OYU74375.1	77%
						–	335835–336116	50	Alphaproteobacteria bacterium PA3 OYU74375.1	
			JXTB01000642.1 (PaT-DNA3) (*c'-rolC-orf13-sus-acs-sus*)	*c'*-like	1	–	128737–129210	29[b]	*Agrobacterium sp.* ASK41782.1	n/a
				rolC-like		–	130259–130348	47	*A.rhizogenes* WP_077768173.1	n/a
				orf13-like	1	–	143125–143619	38	*N. glauca* BAB85946.1	n/a
				sus-like	2	–	145165–144266	41	*A. rhizogenes* WP_034521016.1	88%
						–	151027–151857	39	*A. rhizogenes* WP_034521016.1	
				acs-like	1	–	148051–149046	30	*S. meliloti* WP_088199097.1	n/a
			Sequences, homologous to *Agrobacterium* DNA outside T-DNA	*virH1*-like	1	–	154775–154326	73	*Mesorhizobium amorphae* WP_006204707.1	n/a
				virF-like	1	–	155583–154954	85	*A. rhizogenes* ASK41115.1	n/a
				virH2-like	1	–	156908–155866	70	*A. tumefaciens* OCJ40236.1	n/a

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
			JXTB01000079.1 (PaT-DNA4) (*orf2-orf3n-orf8-rolB-rolC-sus1-sus2-sus2-sus1-rolC–rolC-orf8-orf3n-orf2*)	*orf2*-like	2	–	864853–864039	59	*A. rhizogenes* GAJ95538.1	90%
						–	944836–945167	59	*A. rhizogenes* GAJ95538.1	
				orf3n-like	2	–	865102–866436	58	*A. rhizogenes* WP_034521034.1	81%
						–	910409–909306 944853–944683	58	*A. rhizogenes* WP_034521034.1	
				orf8-like	2	–	873819–875939	42	*A. rhizogenes* WP_034521028.1	81%
						–	907430–905922	50	*A. rhizogenes* WP_034521028.1	
				rolB-like	1	–	878840–878293	36	*A. rhizogenes* P49409.1	n/a
				rolC-like	3	–	881736–882002	42	*A. rhizogenes* XP_016495712.1	1 and 2 or 3 -84% 2 and 3 - 100%
						–	892939–892691	48	*A. rhizogenes* ART94485.1	
						–	899881–899807	48	*A. rhizogenes* ART94485.1	
				sus-like 1	2	–	884634–883678	38	*A. rhizogenes* WP_034521016.1	76%
						–	889476–890000	31	*A. rhizogenes* WP_034521016.1	
				sus-like 2	2	–	886379–885446	56	*A. rhizogenes* WP_034521016.1	78%
						–	888141–888934	42	*A. rhizogenes* WP_034521016.1	

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
			JXTB01000406.1 (PaT-DNA5) (*rolC-IS66-orf2orf3-orf8-iaaM-IS66*)	*rolC*-like	1	–	195235–195552	45	*A. rhizogenes* P20403.1	n/a
				IS66-like	2	–	218288–217887	59	*A. vitis* WP_071205530.1	n/a
						–	263183–262547	67	*A. vitis* WP_071205530.1	n/a
				orf2-like	1	–	220050–219363	62	*A. rhizogenes* GAJ95538.1	n/a
				orf3-like	1	–	235228–236565	51	*A. rhizogenes* WP_034521034.1	n/a
				orf8-like	1	–	241782–242509 250720–251199 253145–254150	42	*A. rhizogenes* WP_034521028.1	n/a
			JXTB01000435.1 (PaT-DNA6) (*orf14-IS630-sus1-acs-sus2-sus1-IS630-orf14*)	*orf14*-like	2	–	152389–152946	58	*N. tomentosiformis* AIM40184.1	87%
						–	163215–162659	53	*N. tomentosiformis* AIM40184.1	
				IS630-like	2	–	153796–153368	67	*Rhizobium* sp. WP_107106966.1	87%
						–	161163–161998	67	*Rhizobium* sp. WP_107106966.1	
				sus-like1	2	–	154860–153872	50	*A. rhizogenes* WP_034521016.1	81%
						–	159662–160597	46	*A. rhizogenes* WP_034521016.1	
				acs-like	1	–	156106–156492	26[b]	*S. meliloti* ASP89596.1	n/a
				sus-like2	1	–	157714–158707	51	*A. rhizogenes* WP_034521016.1	n/a

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
			JXTB01000069.1 (PaT-DNA7) (*orf14-orf14-IS630-sus-IS630-orf14*)	*sus*-like	1	–	544819–545793	39	*A. rhizogenes* ARU12438.1	n/a
				IS630-like	2	–	544339–543648	66	*S. fredii* WP_097588257.1	82%
						–	545803–546275	66	*S. fredii* WP_097588257.1	
				orf14-like	3	–	542268–542537	50	*N. tomentosiformis* AIM40184.1	1 and 2 – 94% 2 and 3 – 87% 1 and 3 – 84%
						–	542658–543214	52	*N. tomentosiformis* AIM40184.1	
						–	547244–546834	55	*N. tomentosiformis* AIM40184.1	
			JXTB01000289.1 (PaT-DNA8) (*sus-orf511-orf14-sus-d*)	*orf511*-like	1	–	215885–215651	56	*N. tomentosiformis* AIM40183.1	n/a
				orf14-like	1	–	216446–216664	50	*A. rhizogenes* AIM40184.1	n/a
				sus-like	2	–	214332–215179	41	*A. rhizogenes* WP_034521016.1	78%
						–	220139–219475	36	*A. rhizogenes* WP_034521012.1	
				d-like	1	–	221129–221534	30	*A. tumefaciens* WP_099086244.1	n/a
			JXTB01000192.1 (PaT-DNA9)	*vis*-like	1	+	284083–285156	53	*A. deltaense* WP_080855286.1	n/a
		Trema orientalis isolate RG33-2	JXTC01000104.1 (ToT-DNA1) *(d-sus-orf14-orf14-sus-d)*	*d*-like	2	–	858368–858009	30	*A. tumefaciens* WP_099086244.1	69%
						–	866532–866939	30	*A. tumefaciens* WP_099086244.1	
				orf8-like	2	–	858332–857733	30	*A. rhizogenes* WP_034521028.1	87%
						–	866532–867000	31	*A. rhizogenes* WP_034521028.1	
				orf14-like	2	–	862759–863301	50	*N. tomentosiformis* AIM40184.1	86%
						–	859804–859528	51	*N. tomentosiformis* AIM40184.1	

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
				sus-like	2	–	865644–864644	52	*A. rhizogenes* WP_034521016.1	91%
						–	859297–859515	42	*A. rhizogenes* WP_034521016.1	
			JXTC01000290.1 (TɷT-DNA2)	*vis*-like	1	+	15220–16293	52	*A. vitis* WP_012649057.1	n/a
		Humulus lupulus var. *lupulus*	BBPC01123852.1 (HllT-DNA1)	*vis*-like	2	+	8242–9315	54	*A. deltaense* WP_080855286.1	99%
						+	10444–11517	54	*A. deltaense* WP_080855286.1	
			BBPC01047313.1 (HllT-DNA2)	*vis*-like	1	+	7250–8323	54	*A. deltaense* WP_080855286.1	n/a
		Humulus lupulus var. *cordifolius*	BBPB01123852.1 (HlcT-DNA1)	*vis*-like	2	+	8224–9315	54	*A. deltaense* WP_080855286.1	99%
						+	10534–11520	54	*A. deltaense* WP_080855286.1	
			BBPB01047313.1 (HlcT-DNA2)	*vis*-like	1	+	7250–8323	54	*A. deltaense* WP_080855286.1	n/a
Fagales	Juglandaceae	*Juglans cathayensis*	QEOU01000974.1	*sus*-like	1	+	212576–213586	66	*A. rhizogenes* WP_034521016.1	n/a
		Juglans mandshurica (isolate M4)	PKSJ01002911.1	*sus*-like	1	+	1584–574	66	*A. rhizogenes* WP_034521016.1	n/a
		Juglans sigillata (isolate DJUG951.04)	QEOY01004675.1	*sus*-like	1	+	7654–8670	66	*A. rhizogenes* WP_034521016.1	n/a
Brassicales	Brassicaceae	*Eutrema yunnanense*	PKML01043526.1	*ocs*-like	1	+	238572–239651	82	*A. vitis* WP_060718528.1	n/a
Myrtales	Myrtaceae	*Psidium guajava*	NTGF01001690.1	*mas1′*	1	–	72971–71705	74	*A. rhizogenes* WP_034520976.1	n/a
				mas2′	1	–	73467–73742	66	*A. rhizogenes* WP_032488585.1	n/a
		Eugenia uniflora	RQIG01001211.1	*rolB*-like	1	–	23–457	66	*A. rhizogenes* CAA34077.1	n/a
				e-like	1	–	801–1103	45	*A. rhizogenes* ASK44378.1	n/a

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
Sapindales	Meliaceae	*Azadirachta indica*	AMWY02033922.1	*orf8*-like	2	–	534–2108	34	*A. rhizogenes* WP_034521028.1	72%
						–	2689–2177	36	*A. rhizogenes* WP_034521028.1	
			AMWY02058681.1	*orf8*-like	1	–	986–393	33	*A. rhizogenes* P09178.1	81%[c]
			AMWY02033921.1	*orf14*-like	1	–	1250–930	35	*N. tomentosiformis* AIM40184.1	n/a
			AMWY02012435.1	*cus*-like	1	–	6435–5620	53	*A. vitis* WP_071201425.1	n/a
Caryophyllales	Caryophyllaceae	*Silene latifolia*	FMHP01041951.	*cus*-like	?	+	37–966	64	*A. vitis* WP_071208191.1	n/a
			(similar sequences in FMHP01031079.1)				1–237	64	*A. vitis* WP_071208191.1	
		Silene latifolia subsp. *alba* cultivar K-line	LHUT01012347.1	*cus*-like	2	+	917–1846	64	*A. vitis* WP_071208191.1	86%
						–	5266–6154	63	*A. vitis* WP_071208191.1	
			(similar sequences in LHUT01032243.1 LHUT01035309.1 LHUT01034187.1 LHUT01087374.1 etc.)	*cus*-like		+	1220–2149 2396–1524 1623–2555 1300–371	63	*A. vitis* WP_071208191.1	n/a
		Silene latifolia isolate Sa984	QBIE01063662.1	*cus*-like	?	–	1041–2020	63	*A. vitis* WP_071208191.1	n/a
			(similar sequences in QBIE01113379.1 QBIE01011535.1 QBIE01027236.1 QBIE01037485.1 etc.)	*cus*-like		–	1387–460 11877–10926 13268–12410 1307–522	63	*A. vitis* WP_071208191	
		Dianthus caryophyllus	BAUD01000269.1	*cus*-like	1	+	77778–78743	64	*A. vitis* WP_071201425.1	n/a

Table 2 (continued)

Order	Family	Species (cultivar)	Accession#	cT-DNA gene (homolog)	Copy number	Intact	Positions	Identity level to proteins from NCBI		Similarity level between 2 copies of the gene
								% of identity	Organism and protein ID	
Ericales	Ericaceae	*Vaccinium macrocarpon* (cultivar Ben Lear (CNJ99-125-1 clone)	JOTO01169953.1	*plast*-like	1	+	2069–1320	40	*Laccaria bicolor* XP_001881215.1	n/a
	Theaceae	*Camellia sinensis*	SDRB01002054.1 (*acs1-rolB-sus-acs2*)	*acs*-like	2	–	1834755–1835240	25[b]	*A. larrymoorei* WP_084631721.1	n/a
						–	1838742–1837927	34	*S. meliloti* ASP89596.1	n/a
				rolB-like	1	–	1835551–1835210	42	*A. rhizogenes* AAA22095.1	n/a
				sus-like	1	–	1837085–1836100	53	*A. rhizogenes* WP_034521016.1	n/a
Solanales	Convolvulaceae	*Cuscuta australis*	NQVE01000054.1	*mis*-like	1	+	1050893–1049952	62	*N. glauca* BAB85949.1	n/a
		Cuscuta campestris	OOIL01000121.1	*mis*-like	1	+	353475–352534	62	*N. glauca* BAB85949.1	n/a

n/a not applicable

[a]Order of the gene homologs in extended cT-DNAs

[b]In some cases fragments with low identity to ref. sequences are also shown to give more information about the insert structure

[c]Similar to AMWY02033922.1 fragment

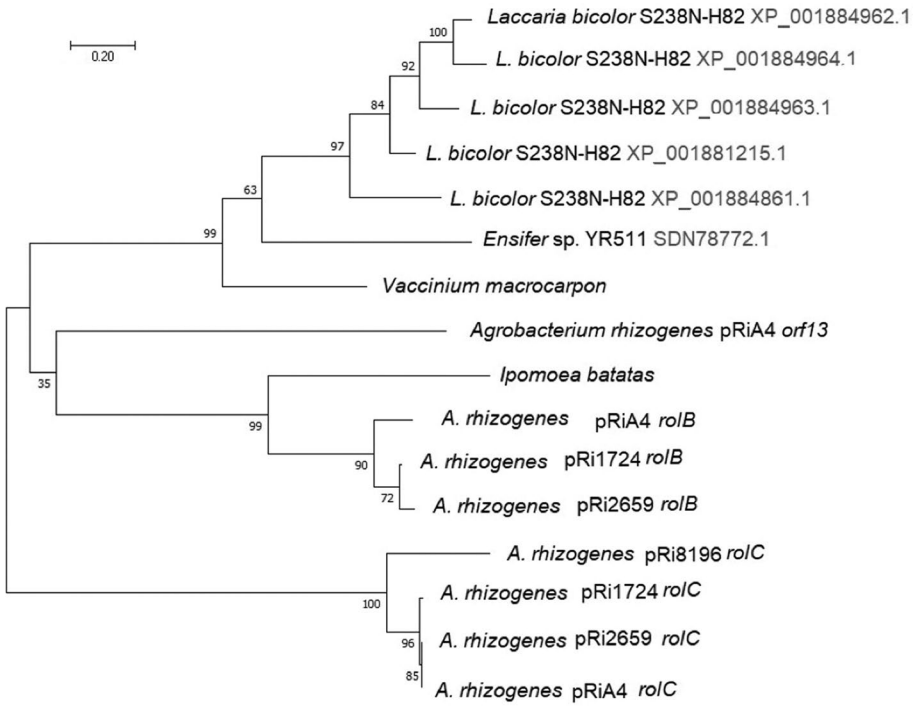

Fig. 2 *Vaccinium* Plast protein clusters with fungal Plast proteins. Molecular phylogenetic analysis of Plast proteins from *Agrobacterium*, Fungi and *Vaccinium* was done with the Maximum Likelihood method. The tree with the highest log likelihood (− 3418.91) is shown. The percentage of trees in which the associated taxa clustered together is shown next to the branches. The tree is drawn to scale, with branch lengths measured in number of substitutions per site. The analysis involved 16 amino acid sequences. All positions containing gaps and missing data were eliminated. There were a total of 144 positions in the final dataset. Evolutionary analyses were conducted in MEGA7 (Kumar et al. 2016)

are 99% similar. *J. regia* is closely related to *J. sigillata*, but their exact relationships are not yet fully established. Interestingly, cT-DNA sequences were found in *J. sigillata*, but not in *J. regia*. The phylogenetically more distant species *J. microcarpa,* and *J. hindsii* do not contain cT-DNA (Supplementary Fig. 2).

Apart from *Arachis* and *Juglans*, homologs of opine genes were also found in *Euphorbia esula* (green spurge), *Nissolia schottii* (yellowhoods), *Humulus lupulus* (hops, used in beer production), *Eutrema yunnanense* (japanese horseradish or wasabi, used for seasoning), *Psidium guajava* (goyave), *Silene latifolia* (white campion), *Dianthus caryophyllus* (carnation), *Cuscuta australis*, and *Cuscuta campestris* (dodder). *Humulus*, *Eutrema*, *Silene*, *Dianthus* and *Cuscuta* contain intact opine genes. The opine genes of *N. schottii* and *P. guajava* are interrupted by stop codons. In *Euphorbia esula*, five contigs with *sus*-like sequences were found. The three longest sequences are mentioned in Table 2. However, none of the contigs covers the gene completely. No stop codons were detected within these fragments. Overlapping parts are 97–99% similar. Additional studies are required to determine the precise copy numbers and extent of these genes.

In *Nissolia schottii*, three contigs contain mikimopine synthase (*mis*)-like sequences. The similarity of the sequences surrounding the first two copies of the *mis*-like gene (Table 2) is about 80%. One *mis* copy is full-length, the other truncated, and 100% identical to the complete copy. The third copy is located in a different region, and 92% identical to the first and second copy. *vis*-like sequences from *Humulus* will be considered below, in comparison with genes from other Cannabaceae.

In the case of *Eutrema*, an intact *ocs*-like gene was found in *E. yunnanense*, but no T-DNA-like sequences were detected in *E. salsugineum* or *E. heterophyllum*. *Psidium* contains a full-size mannopine synthase 1′ (*mas1′*)-like sequence and a *mas2′*-like fragment. *Silene* is represented in the database with three genotypes belonging to two subspecies. They contain a large amount of *cus*-like genes or gene fragments with small sequence differences. One contig of *S. latifolia* subsp. *alba* (LHUT01012347.1) contains two copies of the *cus* gene in direct orientation, one full-size, the second with deletions. The relatively poor quality of the genome assembly may cause an artifical increase in cT-DNA copies. To clarify the situation about these multiple *cus*-like copies, molecular studies are required. *Dianthus caryophyllus* carries an intact *cus*-like gene. It is interesting to note, that carnation plants with blue flowers are a well-known example of a man-made GMO ornamental plant (Tanaka et al. 2009). The two *Cuscuta* species carry highly similar *mis* sequences (99% identity), surrounded by similar plant sequences. This indicates a common and recent origin.

cT-DNAs with only *plast* genes

Two species contain only *plast*-like genes (Otten 2018) on their cT-DNAs, *Eugenia* and *Vaccinium*. *Eugenia uniflora* (pitanga or Suriname cherry) (Nascimento e Santos et al. 2015) is a plant of the Myrtaceae family. It is native

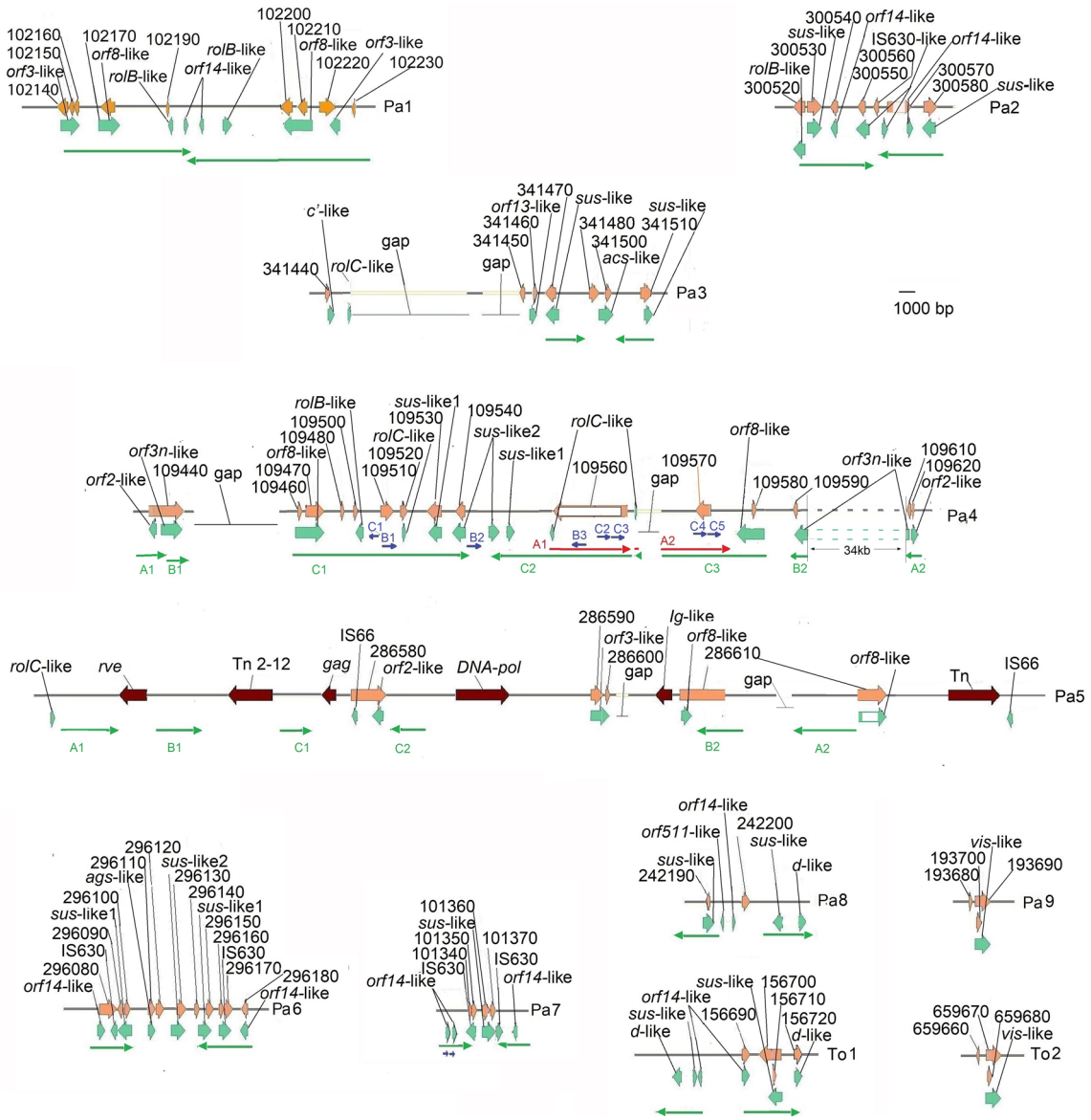

Fig. 3 cT-DNAs in *Parasponia andersonii* (Pa1 to Pa9) and *Trema orientalis* (To1 and To2). Dark brown arrows: plant genes, green arrows: T-DNA-like sequences, thin green arrows: cT-DNA inverted repeats. Orange arrows with numbers: genes found by database annotation. Thin blue and red arrows show different direct repeats

to tropical South America's east coast, but is also grown in the West Indies and Florida (Morton 1987). *Eugenia uniflora* contains *rolB*-like and *e*-like sequences; both are truncated and contain stop codons. Unfortunately, the assembly quality of this genome is not very good. Therefore, the structure and localization of the cT-DNA are unclear. *Vaccinium macrocarpon* (American cranberry) belongs to the Ericaceae family, and forms a symbiotic relationship with ericoid mycorrhizal fungi. It has an single, apparently intact *plast* gene. Interestingly, the predicted protein sequence is more related to Plast proteins from the fungus *Laccaria bicolor* than to agrobacterial Plast sequences. This group also contains a protein from the bacterium *Ensifer* sp. YR511 (Fig. 2). In the following paragraphs we will present natural transformants with more complex cT-DNA structures, carrying both opine and *plast* genes.

cT-DNAs with *plast* and opine genes in Cannabaceae

cT-DNAs were detected in three genera from the Cannabaceae family: *Parasponia*, *Trema* and *Humulus*. Whereas *Humulus* only contains opine gene homologs, *Parasponia*

Table 3 Expressed cT-DNA genes

Order	Family	Species, cultivar	Expressed genes	Acc #
Fabales	Fabaceae	*Arachis hypogaea* Ahy.Unigene39445	*cus* (from A genome)	JR576408.1
Rosales	Cannabaceae	*Humulus lupulus* var. *lupulus*, cv. Shinshu Wase	*vis*-like	LA344641.1
Caryophyllales	Caryophyllaceae	*Silene conica*	*cus*-like[a]	GDJG01032840.1 GDJG01032843.1 GDJG01032841.1 GDJI01014094.1 GDIV01034313.1 GDIS01022199.1 GDJG01032839.1 GDIV01034315.1 GDIX01022641.1 GDIV01034317.1 GDIR01030138.1 GDIV01034312.1 GDIT01036057.1 GDJF01032276.1 GDJF01032275.1 GDIV01034316.1 GDIT01036064.1 GDJD01001444.1 GDIT01036069.1 GDIT01036056.1 GDIT01036067.1 GDIT01036061.1 GDJF01032277.1
		S. dioica		GFCH01054977.1 GFCH01054978.1 GFCG01034229.1 GFCH01054976.1
		S. vulgaris		JL416294.1
		S. undulata		GEYX01050718.1 GEYX01050715.1
		S. sartorii		GEZA01127296.1
Ericales	Ericaceae	*Vaccinium vergatum*	*plast*-like[b]	GGAB01084477.1
	Theaceae	*Camellia sinensis*	*sus*-like	GHEL01136624.1
			acs-like1	GHEL01140452.1 GGTM01069769.1
			acs-like2	GGTM01066813.1
Solanales	Convolvulaceae	*Cuscuta gronovii*	*mis*-like	GDKE01020856.1 GDKE01010421.1 GDKE01024882.1
		Cuscuta pentagona		GAON01250168.1 GAON01250169.1 GAON01250170.1 GAON01250171.1 GAON01250175.1 GAON01250165.1
		Cuscuta suaveolens		GAQC01002147.1

[a]There are numerous *cus*-like fragments of *Silene* species in TSA, they are not identical, but additional molecular studies are needed to verify all of them

[b]GGAB01084477.1 is 96% identical to part of *V. macrocarpon* (JOTO01169953.1)

and *Trema* contain both *plast* and opine gene homologs and have complex multiple cT-DNA structures. *Prunus andersonii* contains no less than nine cT-DNAs, each being part of a long contig (Fig. 3). All PaT-DNA regions, except PaT-DNA9, are imperfect inverted repeats. Based on the degree of divergence between the repeats (indicated between brackets), we propose that PaT-DNA1 (78%), 4 (78%) and 8 (78%) were the first to enter the plant genome, followed by PaT-DNA5 (80%), 7 (82%), 6 (85%), 3 (88%) and 2 (89%). PaT-DNA9 cannot be dated by this method. The PaT-DNA3 and PaT-DNA4 sequences display large gaps, which may carry additional cT-DNA sequences.

PaT-DNA1 contains *orf3*, *orf8, rolB,* and *orf14* homologs on each of the two arms. PaT-DNA2 contains *sus* and *orf14* homologs on the two arms of its inverted repeat, the left part of PaT-DNA2 also contains a *rolB* homolog. The right arm *orf14*-like sequence is shorter than the left arm one. In addition to these T-DNA-like sequences, PaT-DNA2 contains two *IS630*-like transposase genes from Alphaproteobacteria, a complete gene on the left arm, and a partial one on the right arm. Bacterial insertion sequences have been detected in T-DNAs from various Agrobacteria (Machida et al. 1984; Otten et al. 1992; Fournier et al. 1993). The PaT-DNA2 *IS630*-like

Table 4 New cT-DNAs in the genomes of dicots, based on TSA data

Order	Family	Species, cultivar	Accession	cT-DNA gene (homolog)	Nucleotides
Fabales	Fabaceae	*Aeschynomene evenia* var. *serrulata*	GFVI01019205.1	*mis*-like	3–389
Malpighiales	Salicaceae	*Salix purpurea*	GGGM01142278.1	*ags*-like	143–624
			GGGM01084155.1	*nos*-like	148–1061
			GGGM01092451.1	*mas2′*-like	57–1001
Cucurbitales	Cucurbitaceae	*Luffa aegyptiaca*	GDHR01052107.1	*ags*-like	160–1376
			GDHR01016077.1	*mas2′*-like	201–1150
Fagales	Juglandaceae	*Cyclocarya paliurus*	GEUI01000098.1	*sus*-like	40–1050
Sapindales	*Rutaceae*	*Citrus maxima*	GGUJ01012965.1	*ags*-like	254–1543
			GGWA01030381.1	*mas2′*-like	79–1155
Cornales	Cornaceae	*Camptotheca acuminata*	GACF01083404.1	*orf8*-like	3–827
Ericales	Ebenaceae	*Diospyros lotus*	GBSJ01178504.1	*orf8*-like	148–921
				orf3-like	1512–2861
				orf2-like	3133–3954
				acs-like	4554–5778
				Plant retrotransposon	6496–end
			GBSJ01376739.1	*cus*	668–1558
				orf14-like	3058–3618
			GBSJ01097499.1	IS5 family transposase	473–1176
				orf8-like (partial)	1846–2834
			GBSJ01021238.1	*orf14*-like	11–364
			GBSJ01120913.1	*orf13*-like	312–629
			GBSJ01097825.1	*orf511*-like	1–1141
				orf14-like	1905–2453
			GBSJ01101836.1	*orf2*-like	1–212
				acs-like	870–2168
			GBSJ01376992.1	*orf511*-like	3–371
				sus-like	799–1652
			GBSJ01098934.1	*orf3*-like	1–1181
			GBSJ01115705.1	*orf2*-like	26–718
			GBSJ01020915.1	*sus*-like	52–273
			GBSJ01374851.1	*orf8*-like	2463–3602

Table 5 Sequence identity (%) between fragments of *Diospyros lotus* contigs and genes, located in overlapping fragments

	GBSJ01178504.1	GBSJ01376739.1	BSJ01097499.1	GBSJ01101836.1
GBSJ01098934.1	85 (*orf13*)			100 (81 bp of *orf2*)
GBSJ01115705.1	83 (*orf2*)			
GBSJ01101836.1	69 (*acs, orf2*)			
GBSJ01097499.1	72 (*orf8*)			
GBSJ01021238.1		81 (*orf14*)		
GBSJ01376992.1		74 (*orf511*)		
GBSJ01020915.1		73 (*sus*)		
GBSJ01374851.1			74 (*orf8*)	

sequences were most likely part of the original T-DNA insert.

PaT-DNA3 contains *c′*-gene, *rolC*, *orf13*, *sus*, and agrocinopine synthase (*acs*) homologs on its left part. The right part is smaller and contains a *sus*-like sequence. About 2.5 kb to the right, a plant retrotransposon sequence is found, followed by a 3 kb region with *Agrobacterium virF*-, *virH1*- and *virH2*-like sequences, and a plant sequence with unknown function. Thus, the PaT-DNA3 insert is closely linked to a Ti plasmid fragment from outside the T-DNA

borders. Another *Parasponia* contig, JXTB01000142.1 (850 kb), also contains a bacterial, non-T-DNA pTi sequence (492–5522), 69% identical to several fragments of pTi_Tun151 (KY000068). These code for an ABC transporter permease, a phosphodiesterase, and an arabinose phosphate phosphatase. This pTi-like sequence is connected to a plant sequence beyond nucleotide 5522. No T-DNA-like sequences were found on JXTB01000142.1.

The PaT-DNA4 left and right arm contain *orf2*-, *orf3n*-, *orf8-, rolB*-, and *rolC*-like sequences, and two different *sus*-like sequences (*sus*-like1 and *sus*-like2). The *orf3n* copy on the right is interrupted by a 34 kb plant sequence (910411–944678). Apart from its inverted repeat, PaT-DNA4 also shows several direct repeats (Fig. 3). The 6 kb long direct repeat A1 (892861–898343) is 99% identical to A2 (1899807–905285). The A repeats contain smaller, direct repeats (C2 and C3 for A1, C4 and C5 for A2, 94% identity), an isolated copy (C1) is found on the left part of PaT-DNA4. Repeat B2 (886476-887492), localized between the two cT-DNA arms, is 82% identical to B1 (880331–881415). B3 (893372–894466) is inverted with respect to B1 and B2.

PaT-DNA5 also has a complex structure, in which *rolC*-, *orf2*-like, *orf3*-like, and *orf8*-like sequences alternate with plant genes of unknown function, one coding for DNA polymerase, and one coding for an Ig-like domain-containing protein. In addition, PaT-DNA5 carries two *IS66*-like fragments (unrelated to *IS630*). Fragments A1, B1, C1 and A2, B2, C2 form an inverted repeat with 80% identity between the two arms, but do not seem to contain T-DNA-like genes.

PaT-DNA6 contains *orf14*- and *sus*-like genes on both arms of its inverted repeat. *IS630*-like sequences (one full-size, the other partial) are found between the *orf14*- and *sus*-like sequences and belong to the initial T-DNA structure. The PaT-DNA6 *IS630*-like sequences are different from the PaT-DNA2 *IS630*-like sequences (50% protein identity). An *acs*-like sequence and an additional *sus*-like gene are localized between the two arms, the latter has 63% and 66% identity to the *sus*-like genes of the left and right arm. PaT-DNA7 is a shortened version of PaT-DNA6, and is surrounded by similar plant sequences. It contains a *sus*-like gene surrounded by partial *IS630*-like elements in opposite orientation, 82% identical to each other. These sequences are surrounded by two inverted *orf14*-like sequences and an additional, partial *orf14*-like sequence.

PaT-DNA8 carries an unusual *plast* gene which encodes a protein with weak homology to protein D and other Plast proteins, one *orf14*-like gene, and two copies of a *sus*-like gene. It also carries a remnant of an *orf511*-like gene. An *orf511* gene has sofar only been found in the *N. tomentosiformis* TD cT-DNA (Chen et al. 2014), its function is unknown. PaT-DNA9 contains an intact *vis*-like sequence.

Another naturally transgenic species from the Cannabaceae family is *Trema orientalis*, closely related to *Parasponia*. cT-DNA sequences were found in two contigs. ToT-DNA1 is similar to PaT-DNA8, the average identity of their matching fragments being 84%. ToT-DNA1 is organized as an imperfect inverted repeat, containing two copies of a truncated *sus*-like gene, and two copies of an *orf14*-like and *d*-like gene. Average similarity between the two T-DNA arms is 85%. ToT-DNA2 is 95% identical to PaT-DNA9 and contains a *vis*-like sequence. Most likely, the insertion events which gave rise to ToT-DNA1/PaT-DNA8 and ToT-DNA2/PaT-DNA9 predate the *Trema*/*Parasponia* separation. Nearly all cT-DNA genes of *Parasponia* and *Trema* are degenerated. However, the left PaT-DNA2 *sus*-like gene and the *vis*-like genes of ToT-DNA2 and PaT-DNA9 are intact, and may be functional. Additional studies are needed to confirm this.

Within the Cannabaceae family, *Humulus* contains a *vis*-like gene, like *Parasponia* and *Trema*, but *Cannabis sativa* does not. In two subspecies of *Humulus*, the *vis*-like gene is present in three copies. Two of them form a direct repeat, the third one is located in another contig. The *vis*-like genes of *Parasponia, Trema* and *Humulus* are highly similar (Supplementary Fig. 3). However, in *Humulus* the surrounding sequences differ from those in *Parasponia* and *Trema.* This indicates independent acquisition of the gene by *Humulus* and the *Parasponia*/*Trema* ancestor. The 6854-8448 fragment of *Humulus* contig BBPC01047313.1 is similar to *Parasponia* PaT-DNA7, the intact *vis*-like ORF (7250-8323) is located within this fragment. The 6854-8448 fragment most likely delimits the cT-DNA insert. The high similarity of the cT-DNA sequences in these species may be due to recent transformation by similar *Agrobacterium* strains.

cT-DNAs with *plast* and opine genes in *Azadirachta indica*, *Quillaja saponaria*, and *Camellia sinensis*

Azadirachta indica, or neem tree, has been used in folk medicine in India for over 2000 years (Kausik et al. 2002). Four *Azadirachta indica* contigs contain cT-DNA sequences. They are relatively short, but also contain plant sequences, demonstrating integration of T-DNA into the plant genome. These contigs contain sequences similar to *orf8*, *orf14*, and *cus.* Because the contigs are small, it is not possible to tell whether they are part of the same cT-DNA or located on different cT-DNAs. To elucidate the fine structure of the cT-DNA in *Azadirachta indica*, additional experiments are required.

Quillaja saponaria (soap bark tree) is a medicinal plant native to South America (Muravieva 1983). It contains *rolB*-like, *orf13*-like, *e*-like, and *sus*-like sequences. The *e*-like sequence is associated with a bacterial *IS3*-like sequence. The *rolB*-like gene on contig PVLG01028938.1 is situated at the right border of the contig, and therefore partial. A

TBLASTN search with RolB did not yield additional *Quillaja* contigs, indicating that coverage may be incomplete. The quality of the assembly does not allow us to draw conclusions about the structure of the *Quillaja* cT-DNA(s).

Camellia sinensis is a species of evergreen shrub or small tree, and has been used for thousands of years to make tea. Its genome has been sequenced (Wei et al. 2018) and our search showed that it contains homologs of *rolB, sus,* and two *acs*-like genes. These sequences are located in the same contig, and organized as an imperfect inverted repeat of 5.3 kb, the common parts are 90% identical. No intact ORFs are found on this fragment.

Summarizing the analysis of the plant genomes, we note that opine genes are more common and better preserved. Most of the *plast* genes acquired stop codons and probably lost their function. The presence of intact *Agrobacterium*-derived ORFs in several natural transformants suggests that these genes may be expressed. This led us to analyze the TSA database to search for expressed cT-DNA genes.

Identification of dicot cT-DNA sequences in the Transcriptome Shotgun Assembly database

The Transcriptome Shotgun Assembly (TSA) database was searched for cT-DNA-like sequences with the BLASTX option, using the T-DNA protein query sequences (Table 1), and with BLASTN using the nucleotide sequences of natural transformants listed in Table 2. It should be noted that the lack of TSA sequences for a given species does not mean a lack of such sequences in the genome, because transcripts could be missing in the libraries if expression is limited to certain tissues or stages. For this reason, the results of this section should not be viewed as definitive, but as starting material for future research. At this stage, it is not possible to match the data of the fully sequenced genomes and the transcriptomes. Within some genera, some species have only sequenced genomes, while others have only sequenced transcriptomes. The TSA library was found to contain transcript sequences from some of the cT-DNA genes of the above-mentioned species, or from closely related species, these are summarized in Table 3.

The TSA database analysis yielded representatives of seven additional dicot genera with cT-DNAs (Table 4). These are *Aeschynomene evenia* (shrubby jointvetch), *Salix purpurea* (purple willow), *Luffa aegyptiaca* (sponge gourd), *Cyclocarya paliurus* (sweet tea), *Citrus maxima* (pomelo), and *Diospyros lotus* (Caucasian persimmon, one of the oldest cultivated plants). Among these candidates, *Diospyros lotus* is especially worth mentioning. Its transcriptome contains opine genes and *plast* genes. They show similarity to *acs, cus, sus, orf2, orf3n, orf8, orf13, orf14, orf17n and orf511* sequences. Some of these are combined into longer sequences, which may result from abnormal read-through transcription. Table 5 shows the identity percentages of extended areas among contigs. The TSA data require verification by additional methods, since they cannot exclude contamination with *Agrobacterium* DNA, and do not provide information on the location and structure of the inserts. The TSA analysis confirmed the predominance of opine genes, already noticed for the WGS data.

Search for monocot cT-DNA sequences in the WGS and TSA databases

As of June 2019, 73 monocot sequences were available in the WGS database. We searched these sequences in the same way as for dicots. T-DNA-like sequences were found in *Dioscorea alata* (greater yam), but not in *D. rotundata* (white Guinea yam). Greater yam is an important and geographically widely distributed staple food (Cormier et al. 2019). Although sweet potato (*Ipomoea batatas*) is also called yam, it is unrelated to *Dioscorea alata*. *Dioscorea alata* contigs CZHE02045212.1 (3009 nt) and CZHE02050078.1 (2550 nt) are 96% similar. CZHE02045212.1 potentially encodes an intact Cus-like protein, 85% identical to WP_071201425.1 from *Agrobacterium vitis*, CZHE02050078.1 encodes a truncated Cus-like protein. Another *Dioscorea* species, *Dioscorea bulbifera* (potato yam) was among the first monocot species to be transformed (Schäfer et al. 1987).

The TSA database contains 132 monocot sequences. We found *cus*-like sequences in *Musa acuminata* AAA group (dessert banana), in accession numbers JV331205.1 (157–1071), JV353951.1 (3–554), and JV360234.1 (8–316), with identity values to WP_071201425.1 of 64%, 57%, and 63% respectively. JV331205.1 is intact, the two other ones are truncated. Banana is one of the oldest cultivated plants, found on all continents.

Discussion

In 2012, we reported the identification of a new natural transgenic plant, *Linaria vulgaris* (toadflax) by analyzing more than a hundred species of dicotyledonous plants, using PCR primers designed from typical T-DNA gene sequences (Matveeva et al. 2012). The current study shows that a search for HGT sequences by bioinformatic methods is an order of magnitude more efficient. This is not surprising, since such an approach allows detection of highly diverged sequences, not possible with PCR primers. Our analysis of the WGS and TSA databases (in all, 631 dicot and 205 monocot species) revealed numerous examples of naturally transgenic plants. New T-DNA-like dicot sequences were found in representatives of 39 (23 + 16) species, 24 (17 + 7) genera, 17 (12 + 5) families, and 12 (10 + 2) orders. Previously, six species

from these databases (*N. tabacum*, *N. tomentosiformis*, *N. otophora*, *N. noctiflora*, *Ipomoea batatas* and *I. trifida*) were already found to be transgenic. Thus, our data indicate that about 7% of the sequenced dicot species are naturally transformed. With an estimated number of 175,000–200,000 dicot species, this yields a minimum of 10,000 naturally transformed dicot species. Out of 205 available WGS and TSA monocot sequences, only those of *Dioscorea alata* and *Musa acuminata* contain T-DNA-like sequences.

Many of the cT-DNA sequences appear to be relatively small T-DNA fragments. They may result from incomplete transfer of T-DNA, or from partial cT-DNA deletion subsequent to the initial insertion event. In some cases, these fragments may have been amplified. Partial deletions of cT-DNA sequences are known for various cultivars of *N. tabacum* (Chen et al. 2014), and a cT-DNA duplication was detected in *N. otophora* (Chen et al. 2018). Many cT-DNA structures are inverted repeats (Chen et al. 2014), with typical LB-associated sequences at both ends. PaT-DNA1, 4, 6, and 7 from *Parasponia* are organized in the same way. Such structures are consistent with the proposed mechanism of T-DNA integration by polymerase theta (PolQ), which can link two T-DNA molecules with their 3′ ends (LB ends) to plant DNA breaks (van Kregten et al. 2016).

Apart from *Agrobacterium* and plants, various Fungi (Mohajjel-Shoja et al. 2011) and *Rhizobium* species (Chen et al. 2014; Chen 2016) also contain genes encoding Plast-like sequences. In the case of Fungi, these sequences could result from *Agrobacterium* transformation. In the case of the Rhizobia, these bacteria do not contain pTi/pRi sequences, and their Plast-like sequences cluster separately from the *Agrobacterium* ones (Chen 2016). Thus, their origin remains unclear. In most plant cases described here, the T-DNA-like sequences were phylogenetically closer to *Agrobacterium* sequences than to sequences from other taxons. However, in two species of the genus *Vaccinium*, a cT-DNA *plast*-like gene was found which is closer to fungal (*Laccaria bicolor*) and bacterial (*Ensifer* sp.) *plast* sequences. We have proposed (Mohajjel-Shoja et al. 2011; Chen and Otten 2017) that fungal *plast* genes are derived from transformation with an unknown *Agrobacterium strain*, because of their rare and patchy distribution among fungal groups. A similar *Agrobacterium* strain could then also be the source of the *Vaccinium plast*-like genes. Additional unusual types of T-DNA-like genes include a *c'*-like gene in PaT-DNA3, a *d*-like gene in PaT-DNA8 and a highly diverged *acs*-like gene in PaT-DNA6. The corresponding *Agrobacterium* strains remain to be identified.

Among the newly described naturally transgenic plant species, those containing only opine genes predominate (16 out of 23 species from the WGS database and 14 out of 16 from the TSA database). This might be due to several reasons. The first one is related to the T-DNA transfer mechanism. T-DNA transfer starts from the right T-DNA border, and is not always complete. Because opine genes like *mas2′* (in *A. rhizogenes* strain 8196, Hansen et al. 1991), nopaline synthase (*nos*), *cus*, *sus*, *mis*, *vis*, and *ocs* are situated close to the right border, they are more likely to be transferred in case of incomplete transfer. Since opine genes are not known to favour regeneration, regeneration of such incompletely transformed cells most likely occurred spontaneously. The second possibility is that the opine genes were initially located on a T-DNA fragment together with *plast* genes like *rolB* and *rolC*, allowing formation of hairy roots with a high potential for regeneration. If *plast* genes reduced growth or fertility of the regenerant plants, they may have been lost by negative selection. Alternatively, if the initial transformation event involved the insertion of different T-DNA fragments, some with *plast* genes, the others with opine genes, the *plast* genes could have been lost by segregation. Some of the potential opine enzymes reported here, are only distantly related to known sequences and could produce new types of opines. Thus, their properties should first be extensively investigated in vitro, before searching the corresponding opines in the natural transformants. In addition, it will be necessary to determine where these genes are expressed by using reporter genes. In several *N. tabacum* cultivars, the TB-*mas2′* gene is expressed in roots and leads to measurable amounts of opines (Chen et al. 2016).

Longer cT-DNAs with a combination of opine and *plast* genes, usually have an imperfect repeat structure. This was already noted for *Nicotiana* and *Ipomoea*, and could reflect a basic property of the T-DNA transfer system, or a higher stability of such repeats over longer periods.

Of particular interest are the complex cT-DNA sequences from *Parasponia, Trema*, *Quillaja*, *Camellia* and *Azadirachta*. Most of their genes are degenerated and carry stop codons. However, two opine genes of *Parasponia* have an intact ORF and may still encode opine synthesis, in spite of the rather large divergence between the repeats. In *Parasponia*, these divergence rates range from 12 to 22%. In *Linaria*, *Nicotiana*, and *Ipomoea* these are 8%, 1–6% and less than 1% respectively (Chen and Otten, 2017). As in the case of the TA, TB, TC and TD regions of *N. tomentosiformis*, the different *Parasponia* cT-DNA sequences are footprints of multiple transformation events, that took place well before the transformation of *Linaria*, *Nicotiana* and *Ipomoea*. PaT-DNA6 and 7 result from duplication of the same original insert, like the TE region of *N. otophora* (Chen et al. 2018), adding to the complexity of cT-DNA structure and function in natural transformants. The ancestor sequences of ToT-DNA1/PaT-DNA8 and ToT-DNA-2/PaT-DNA9 were inserted before the *Parasponia-Trema* divergence. PaT-DNA8 shows 22% divergence between the two arms, ToT-DNA1 only 15%. Thus, the divergence rate between the repeats was faster in *Parasponia* as in *Trema*.

The insertion of ToT-DNA2/PaT-DNA9 seems to be more recent, with only 5% divergence. *Parasponia andersonii* and *Trema orientalis* are non-legume plants, which evolved nitrogen fixation capability through symbiosis with a large range of *Rhizobium* species (Op den Camp et al. 2012). The strong association between *Parasponia/Trema* and Rhizobiaceae and the close relationship between *Rhizobium* and *Agrobacterium* may have favored multiple T-DNA transformation events. In this scenario, *Agrobacterium* strains would have transmitted their Ri plasmids to *Rhizobium* strains, which then transferred T-DNA sequences to their host plants. It would be interesting to test whether some Rhizobia associated with *Parasponia andersonii* and *Trema orientalis* carry Ri plasmids, and whether such strains can lead to transformation and regeneration of their hosts. More generally, natural transgenic species may have a higher spontaneous regeneration capacity than non-transformed species.

PaT-DNA2, and PaT-DNA6 and 7 of *Parasponia* carry different *IS630*-like bacterial insertion elements, PaT-DNA5 shows an *IS66*-like sequence. In *Quillaja*, an *IS3*-like element is found close to an *e*-like *plast* gene sequence. IS elements can be easily transferred between bacteria, and their frequent insertion in the *Agrobacterium* genome, including the Ti plasmids, strongly contributes to *Agrobacterium* evolution, pTi structure and modification of T-DNA function (Otten et al. 1992). The presence of *IS630*-, *IS66*- and *IS3*-like sequences in natural transformants shows that *Agrobacterium* can transfer bacterial elements to plants in a two-step HGT process. In the first step, IS elements from other bacteria insert into a T-DNA (probably in a random way). In the second step, they are transferred to a plant as part of the T-DNA. IS elements are not expected to function in plants, as they lack plant expression signals. We could not find free IS-like elements in *Parasponia*, indicating that they did not transpose. These elements may have been transferred by chance, without influence on the plant, but they could have played a role in allowing efficient regeneration, by inactivating a T-DNA gene interfering with that process. *Agrobacterium* strain A66 for example, carries an *IS66* insertion element in the auxin synthesis gene *iaaH*, which leads to shooty tumors (Binns et al. 1982; Machida et al. 1984). Apart from IS elements, we also found other agrobacterial sequences in *Parasponia*. One fragment carried *vir* genes and was situated close to a T-DNA, the other fragment carried Ti plasmid genes, unconnected to T-DNA sequences. These two cases represent another type of HGT, with a more or less random transfer of non-T-DNA sequences, probably through abnormal activity of the *Agrobacterium* DNA transfer system.

Transformed plants carrying non-T-DNA sequences like *vir* region DNA and vector backbone sequences have been reported by several authors (Ooms et al. 1982; Ramanathan and Veluthambi 1995; Kononov et al. 1997; Gelvin 2017). The more or less frequent occurrence of such abnormal structures may depend on the properties of the virulence genes. In the case of the natural GMOs, the original Ri plasmids and their *vir* genes are unknown. Since *Agrobacterium* can transfer large fragments of its chromosomal DNA via the T-DNA transfer system (Ulker et al. 2008), these sequences may also be found in natural GMOs. This interesting possibility has not been tested sofar.

The present findings expand the list of natural GMOs to a much larger number of plant taxa and suggest some directions for further research. One of these is the functional study of intact cT-DNA genes. Another concerns the variability and evolution of cT-DNA sequences in natural plant populations and in cultivated species. Among the new natural GMOs are several plants used for food, drinks and medicine, with large collections of accessions and cultivars already available, and rapidly increasing amounts of sequence data. Such data will provide excellent material for studies on cT-DNA evolution. Based on our results, it is clear that throughout their history, almost all human cultures have encountered natural transformants, which they adopted for food, drinks, medicine or decorative purposes.

Materials and methods

Identification of cT-DNA

In order to detect new cT-DNA sequences, we performed a 4-step blast search. In the first step, representative protein sequences of *A. rhizogenes* oncogenes, their homologs from *Ipomoea* and *Nicotiana* plants, from the fungus *Laccaria bicolor* and protein sequences of opine genes of different strains of *Agrobacterium* sp. (Table 1) were recovered, and used as queries to search the National Center for Biotechnology Information (NCBI) Whole-Genome Shotgun (WGS) contigs of all plant genomes sequenced to date, using the TBLASTN algorithm. In the second step, Vir protein sequences (Table 1) were used to search for possible *Agrobacterium* contaminations in those genomes, where T-DNA-like sequences were detected. In case homologs of *vir* genes were detected, the surrounding sequences were studied. When plant genes were found to be linked to *vir* genes, the hypothesis of contamination was rejected. In the third step, contigs that potentially encoded T-DNA-like protein sequences with identity levels 30% or higher, were analyzed further. They were used as queries in BLASTX to detect the closest protein homologs and to identify proteins encoded by plant genes surrounding the cT-DNA. The resulting cT-DNA maps (based on sequence similarities) were mapped to annotated sequences from the same plant species, wherever

possible. The Vector NTI Advance™ software was used to build the combined maps. In the fourth step, the TSA database was used to search for expressed new cT-DNA genes, using sequences described in the third step as a query. In addition to this, the TSA database was used to search for cT-DNA transcripts, as described for step one.

Phylogenetic analysis of cT-DNA sequences

Phylogenetic analysis of cT-DNA sequences was done in MEGA 7.0 (Kumar et al. 2016). Evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei 1993). The bootstrap consensus tree inferred from 500 replicates is taken to represent the evolutionary history of the taxa analyzed (Felsenstein 1985). Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value.

Acknowledgements This work was partially carried out using the software of the St. Petersburg State University Resource Center "Development of molecular and cellular technologies". We would like to dedicate this work to the memory of Rob Schilperoort, one of the pioneers in *Agrobacterium* research and founder of Plant Molecular Biology.

Author contributions TM found new naturally transgenic plants, TM and LO characterized cT-DNA structures and prepared the manuscript.

Funding Funding for T.M. was obtained from the Russian Science Foundation (Grant No. 16-16-10010).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Acuna R, Padilla BE, Florez-Ramos CP, Rubio JD, Herrera JC, Benavides P, Lee SJ, Yeats TH, Egan AN, Doyle JJ, Rose JK (2012) Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc Natl Acad Sci USA 109:4197–4202

Angiosperm Phylogeny Group (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181(1):1–20

Binns AN, Sciaky D, Wood HN (1982) Variation in hormone autonomy and regenerative potential of cells transformed by strain A66 of *Agrobacterium tumefaciens*. Cell 31:605–612

Chen K (2016) Sequencing and functional analysis of cT-DNAs in *Nicotiana*. PhD Thesis University of Strasbourg, France

Chen K, Otten L (2017) Natural *Agrobacterium* transformants: recent results and some theoretical considerations. Front Plant Sci 8:e1600

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

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

Chen K, Dorlhac de Borne F, Sierro N, Ivanov NV, Alouia M, Koechler S, Otten L (2018) Organization of the TC and TE cellular T-DNA regions in *Nicotiana otophora* and functional analysis of three diverged TE-*6b* genes. Plant J 94:274–287

Cormier F, Lawac F, Maledon E, Gravillon MC, Nudol E, Mourmet P, Vignes H, Chaïr H, Arnau G (2019) A reference high-density genetic map of greater yam (*Dioscorea alata* L.). Theor Appl Genet 132:1733–1744

Dong W, Xu C, Li W, Xie X, Lu Y, Liu Y, Jin X, Suo Z (2017) Phylogenetic resolution in *Juglans* based on complete chloroplast genomes and nuclear DNA sequences. Front Plant Sci 8:e1148

e Santos DN, de Souza L, Nilson JF, de Oliveira AL (2015) Study of supercritical extraction from Brazilian cherry seeds (*Eugenia uniflora* L.) with bioactive compounds. Food Bioprod Process 94:365–374

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Fournier P, Paulus F, Otten L (1993) IS870 requires a 5′-CTAG-3′ sequence to generate the stop codon for its large ORF1. J Bact 175:3151–3160

Gao C, Reno X, Mason AS, Liu H, Xiao M, Li J, Fu D (2014) Horizontal gene transfer in plants. Funct Integr Genomics 14(1):23–29

Gelvin SB (2017) Integration of *Agrobacterium* T-DNA into the plant genome. Annu Rev Genet 51:195–217

Guillermo S, Lavia GI, Fernandez A, Krapovickas A, Ducasse DA, Bertioli DJ, Moscone EA (2007) Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double *GISH*. Am J Bot 94(12):1963–1971

Hansen G, Larribe M, Vaubert D, Tempé J, Biermann B, Montoya AL, Chilton M-D, Brevet J (1991) *Agrobacterium rhizogenes* pRi8196: mapping and DNA sequence of functions involved in mannopine synthesis and hairy root function. Proc Natl Acad Sci USA 88:7763–7767

Intrieri MC, Buiatti M (2001) The horizontal transfer of *Agrobacterium rhizogenes* genes and the evolution of the genus Nicotiana. Mol Phylogenet Evol 20:100–110

Kausik B, Chattopadhyay I, Banerjee RK, Bandyopadhyay U (2002) Biological activities and medicinal properties of Neem (*Azadirachta indica*). Curr Sci 82(11):1336–1345

Kochert G, Stalker HT, Gimenes M, Galgaro L, Lopes CR, Moore K (1996) RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). Am J Bot 83:1282–1291

Kononov ME, Bassuner B, Gelvin SB (1997) Integration of T-DNA binary vector 'backbone' sequences into the tobacco genome: evidence for multiple complex patterns of integration. Plant J 11:945–957

Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874

Kyndt T, Quispe D, Zhai H, Jarret R, Ghislain M, Gheysen Liu Q, Kreuze JF (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(18):5844–5849

Machida Y, Sakurai M, Kiyokawa S, Ubasawa A, Suzuki Y, Ikeda J-E (1984) Nucleotide sequence of the insertion sequence found in the

T-DNA region of mutant Ti plasmid pTiA66 and distribution of its homologues in octopine Ti plasmid. Proc Natl Acad Sci USA 81:7495–7499

Matveeva TV, Bogomaz DI, Pavlova OA, Nester EW, Lutova LA (2012) Horizontal gene transfer from genus *Agrobacterium* to the plant *Linaria* in nature. Mol Plant Microbe Interact 25(12):1542–1551

Matveeva TV, Bogomaz OD, Golovanova LA, Li YuS, Dimitrov D (2018) Homologs of the *rolC* gene of naturally transgenic toadflaxes *Linaria vulgaris* and *Linaria creticola* are expressed in vitro. Vavilovskii Zhurnal Genetiki i Selektsii 22(2):273–278

Mohajjel-Shoja H, Clément B, Perot J, Alioua M, Otten L (2011) Biological activity of the *Agrobacterium rhizogenes*-derived *trolC* gene of *Nicotiana tabacum* and its functional relationship to other *plast* genes. Mol Plant Microbe Interact 24:44–53

Morton J (1987) Surinam cherry. In: Fruits of warm climates. Miami, p 386–388

Muravieva DA (1983) Tropical and subtropical medicinal plants. Moscow: Medicine (**in Russian**)

O'Leary NA, Wright MW, Brister JR et al (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745

Ooms G, Bakker A, Molendijk L, Wullems GJ, Gordon MP, Nester EW, Schilperoort RA (1982) T-DNA organization in homogeneous and heterogeneous octopine-type crown gall tissues of *Nicotiana tabacum*. Cell 30:589–597

Op den Camp RHM, Polone E, Fedorova E, Roelofsen W, Squartini A, Op den Camp HJM, Bisseling T, Geurts R (2012) Nonlegume *Parasponia andersonii* deploys a broad *Rhizobium* host range strategy resulting in largely variable symbiotic effectiveness. Mol Plant Microbe Interact 25:954–963

Otten L (2018) The *Agrobacterium* phenotypic plasticity (*plast*) genes. Curr Top Microbiol Immunol 418:375–419

Otten L, Canaday J, Gérard JC, Fournier P, Crouzet P, Paulus F (1992) Evolution of agrobacteria and their Ti plasmids: a review. Mol Plant Microbe Interact 5:79–87

Pattee HE, Stalker HT, Giesbrecht FG (1998) Reproductive efficiency in reciprocal crosses of *Arachis monticola* with *A. hypogaea* subspecies. Peanut Sci 25:7–12

Pavlova OA, Matveeva TV, Lutova LA (2013) *Linaria dalmatica* genome contains a homologue of *rolC* gene of *Agrobacterium rhizogenes*. Ecol Genet 11:10–15

Ramanathan V, Veluthambi K (1995) Transfer of non-T-DNA portions of the *Agrobacterium tumefaciens* Ti plasmid from the left terminus of TL-DNA. Plant Mol Biol 28:1149–1154

Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW (2006) Evolutionary origins of the eukaryotic shikimate pathway: gene fusions horizontal gene transfer and endosymbiotic replacements. Eukaryot Cell 5(9):1517–1531

Schäfer W, Görz A, Kahl G (1987) T-DNA integration and expression in a monocot crop plant after induction of *Agrobacterium*. Nature 327:529–532

Stanford AM, Harden R, Parks CR (2000) Phylogeny and biogeography of *Juglans* (Juglandaceae) based on *matK* and ITS sequence data. Am J Bot 87:872–882

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tanaka Y, Brugliera F, Chandler S (2009) Recent progress of flower colour modification by biotechnology. Int J Mol Sc 10:5350–5369

Ulker B, Li Y, Rosso MG, Logemann E, Somssich IE, Weisshaar B (2008) T-DNA-mediated transfer of Agrobacterium tumefaciens chromosomal DNA into plants. Nat Biotech 26:1015–1017

van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas PJJ, Tijsterman M (2016) T-DNA integration in plants results from polymerase-θ—mediated DNA repair. Nat Plants 2:16164

Wei C, Yang H, Wang S et al (2018) Draft genome sequence of *Camellia sinensis* var. sinensis provides insights into the evolution of the tea genome and tea quality. Proc Natl Acad Sci USA 115(18):E4151–E4158

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