

cT-DNA in *Linaria vulgaris* L. is multicopy, inverted and homogenized.

Ivan A. Vladimirov^{1,3}, Olga A. Pavlova^{1,2}, Dmitrii E. Polev¹ and Denis I. Bogomaz^{1,2}.

¹Beagle Ltd, 152-1-77, Bukharetskaya str., Saint-Petersburg, Russia. <http://biobeagle.com/>

²Peter the Great St.Petersburg Polytechnic University, 29, Polytechnicheskaya str., Saint-Petersburg, Russia
<http://english.spbstu.ru/>

³Saint-Petersburg State University, 7/9, Universitetskaya emb., Saint-Petersburg, Russia <https://spbu.ru/>

corresponding author: Olga A. Pavlova, olga.an.pavlova@gmail.com

Abstract

The phenomenon of evolutionary fixation of agrobacterial sequences (cT-DNA or cellular transferred DNA) in plant genomes is well known in nature. It was previously considered, that all of cT-DNA-containing species, except *Linaria vulgaris*, have multiple inverted cT-DNA repeats. Deep studying of general features of cT-DNA brings us closer to understanding the causes and mechanisms of its fixation in plants genomes. We combined multiple long-range PCR with genome walking for studying extended structure of cT-DNA. Using digital PCR method, we estimated copy number of cT-DNA elements. NGS with low covering allows us to develop a set of microsatellite markers, also used for copy number estimation. According to new data, cT-DNA elements in *L. vulgaris* form an inverted complex repeat of two simple direct repeats. After cT-DNA integration, cT-DNA sequence duplication events took place at least two times. The phenomenon of concerted evolution of cT-DNA sequences as well as some details of this process have been shown for the first time.

We have shown, that *L. vulgaris*, as well as other cT-DNA containing species, has inverted structure of repeats. This fact indicates possible existence of some general causes and mechanisms of cT-DNA fixation in plant genomes during evolution.

Keywords

cT-DNA structure; *Linaria vulgaris*; inverted repeats; microsatellite markers; concerted evolution.

Introduction

Linaria vulgaris is an example of a plant containing a vertically inherited sequence of *Agrobacterium rhizogenes* T-DNA in the genome (cT-DNA or cellular T-DNA). It is a rare phenomenon, currently known only in *Linaria* (Matveeva et al. 2012; Pavlova et al. 2014), *Ipomaea* (Kyndt et al. 2015) and *Nicotiana* genera (White et al. 1983; Intrieri and Buiatti 2001). The fixation of cT-DNA sequences in genomes carries a pathogenic program of genetic colonization (Tzfira and Citovsky 2006). The mere fact of such plant survival is of great interest. The intactness of several cT-DNA genes, as well as the observed expression of some of them, makes it necessary to discuss the function of cT-DNA in the genomes of such plants (Meyer et al. 1995; Frundt et al. 1998; Aoki and Syono 1999, 2000; Suzuki et al. 2002). During the evolution of the genus *Nicotiana* integration of cT-DNAs occurred several times in a row (Chen et al. 2014). This fact also serves as an indirect confirmation of a specific function of these cT-DNAs.

Linaria plants derived T-DNA from the mikimopine strain of *Agrobacterium rhizogenes*, which is similar to the present “1724” strain (Matveeva et al. 2012; Pavlova et al. 2014). According to the published data, the cT-DNA insert integrated into the transposon and organized a direct repeat of two cT-DNA units (Matveeva et al. 2012; Pavlova et al. 2014).

Results

This work was started to clarify the *L. vulgaris* cT-DNA structure. Earlier (Matveeva et al. 2012), it was shown that the T-DNA sequence in *L. vulgaris* is an imperfect direct repeat of two T-DNA elements. (We call the “element” a form of a

single T-DNA sequence which is found in the agrobacterial Ri-plasmid). However, the sequence EU735069.2 presented in Matveeva et al. 2012 was obtained without taking into account the possibility of high copy number of T-DNA. In this regard, we have conducted a more detailed study of the T-DNA sequence by sequencing Long-Range-PCR products obtained from unique primers. Unique primers were designed for the sites where elements join with each other and with plant DNA. The sequences of sites for elements joining were determined using GenomeWalker (Siebert et al. 1995). Six unique sites for elements joining were sequenced. Three of them (we call them β , β' , γ) are formed by direct repeats, and three (ω , ω' , ϕ) - by inverted repeats (Figure 1) shown for *L. vulgaris* for the first time. The primers selected for joining regions produced single target PCR-products, so sequences of joining regions could not be PCR-artifacts. The relative positioning of the elements was established using the Long-Range-PCR. The resulting PCR products containing sequences of individual elements were sequenced (Table 3).

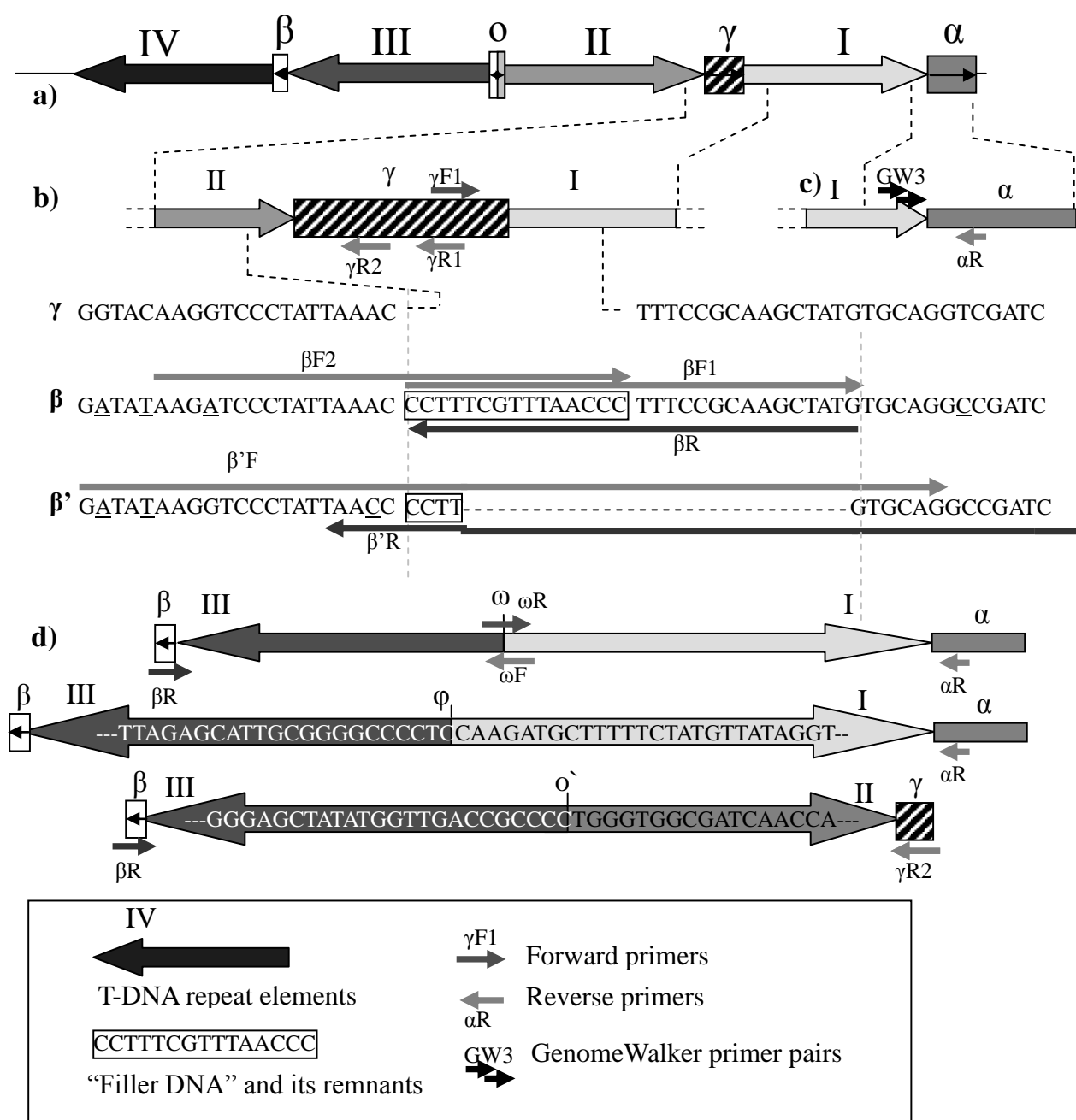


Figure 1. The construction of sites for elements joining and the positions of primers: a) Reconstructed primary structure of the T-DNA block in *L.vulgaris*; b) differences in the sequence between the sites γ , β , β' ; c) structure of α (junction of element I with plant DNA); d) The structure of joining sites formed by inverted repeats (ω , ϕ , ϕ').

Full-size blocks of fragments were sequenced (Fig.2b), as well as individual fragments (20.2, 160.1, 20.5, 18.2, 20.1, 135.11, 206.2) which could be their allelic variants.

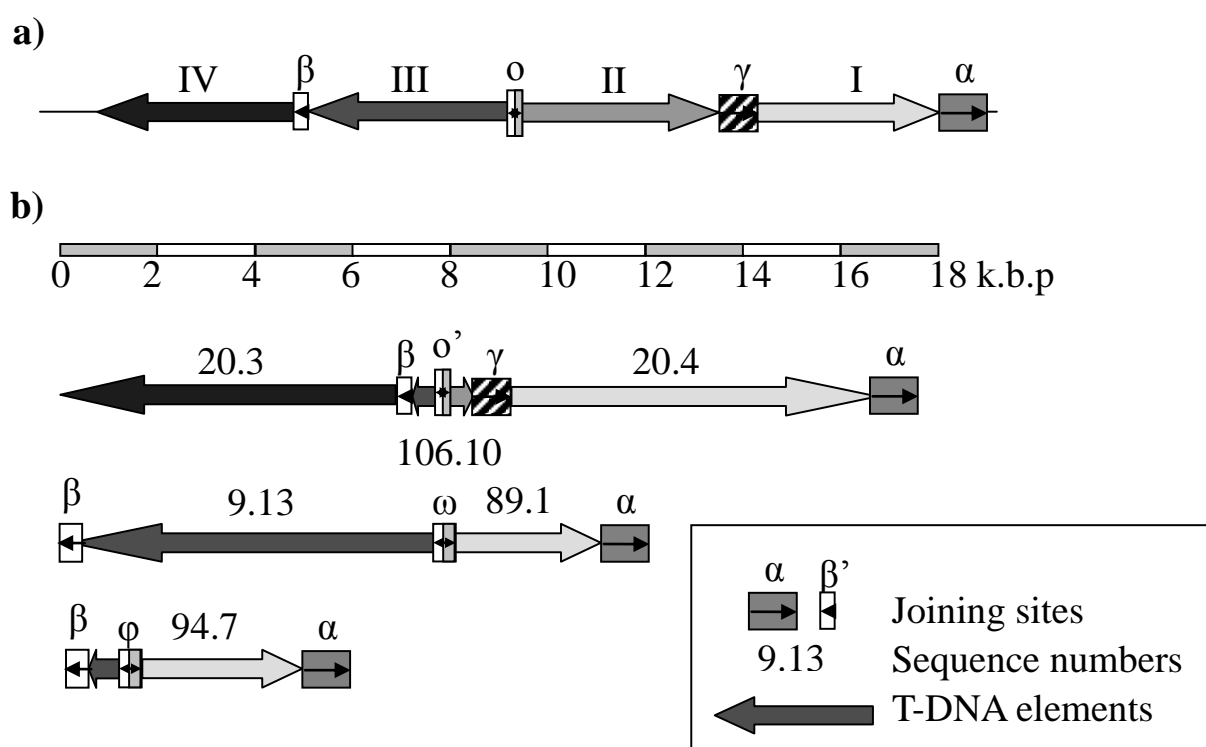


Figure 2. a) Reconstructed primary structure of the T-DNA block; b) Structure of the sequenced T-DNA blocks.

Analysis of the sequences gave us an opportunity to investigate T-DNA structure in *L. vulgaris* (Fig. 2a). The block of T-DNA elements is represented by a complex inverted repeat consisting of two simple direct repeats, with variants of deletions of various sizes. The positions of the elements are marked with Roman numerals (I-IV from right to left), and the sites of joining between them with Greek letters. The most intact of all joining sites is γ between elements I and II (Fig. 1b). β sequence is located between III and IV elements and it differs from γ by a large deletion. Instead of the deleted fragment, there is a sequence of 15 bp long, which is not present in γ and looks like a typical "filler DNA" (Neve et al. 1997), which formed directly upon the transformation during integration of T-DNA into the genome. The β sequence has a short allelic variant with nested deletion (β') which also contains traces of "filler DNA" (Fig. 1b). Docking sites formed by inverted repeats (ω , ϕ , ϕ') do not contain filler DNA. These sites seem to be a result of big deletions which occurred after transformation act (Fig. 1d).

In addition, we found out that:

1) The number of variants of the sequences for almost all positions is more than two (Fig. 2, b). As far as *L.vulgaris* is diploid (Tandon, Bali, 1957), and the number of alleles of a single-copy sequence cannot exceed two, this means that there are several such T-DNA blocks in the *L.vulgaris* genome. In position III five variants of individual sequences corresponding to at least 3 blocks (94.7, 106.10, 20.1, 9.13, in Fig. 2b and 206.2) were detected.

2) Four different variants of the sequences containing the site of joining of T-DNA with plant DNA were found (20.4, 89.1, 94.7- Fig. 2b and 160.1), and the integration site (α) in all four cases was identical.

The analysis of the state of ORF's from each element was carried out. It was found that all sequenced T-DNA genes except *orf13a* were pseudogenized. Premature stop codons and deletions leading to a reading frame shift were detected in all genes but not in *orf13a*. Gene *orf13a* was found between the *orf13* and *orf14* genes, and it is potentially intact in the sequences 20.1, 9.13 and 206.2.

To clarify cT-DNA copy number digital PCR was carried out. The test system for the α -sequence from cT-DNA was designed. Since *Linaria* genome is not well studied and single copy genes are unknown we decided to use microsatellite markers (previously selected) as reference sequences. Genome sequencing with low coverage was performed using Ion Torrent System for searching candidate sequences for microsatellite markers. Twenty three suitable reads with repeats were selected and primers were designed. All primers were tested on plants DNA from geographically distant regions (Peterhof, Russia and Hakassiya, Russia). Then fragment analysis of PCR products was performed to determine suitability of microsatellite markers. Suitability criteria were: 1) the presence of no more than two peaks of different lengths in each plant studied (indirect confirmation of single copy number); 2) the presence of a at least one peak of the sequence of predicted length in plants sequenced by Ion Torrent System. The list of markers that satisfied all requirements is given in Table 2. The TaqMan probes were designed for markers M006, M009, M010, M014, M015 and M019., Effectiveness of the test systems was estimated by real-time PCR. The M006 marker was excluded since it was multi-copy. The most efficient system based on the M009 marker was selected from the remaining ones. Using a digital PCR the copy number of the α -sequence from T-DNA was compared to the M009 marker. A ratio of 2.19 (1528 copies of the α sequence in 4590 cells and 697 copies in 4590 cells in M009) was obtained.

Discussion

We have shown an inverted structure of repeats for *L. vulgaris* cT-DNA. Such orientation of the elements during the integration into the genome is often observed (but not necessary). Neve (1997) has shown that the frequency of inverted repeats occurrence depends on the type of T-DNA and varies from 25 to 50% of all cases of multiple elements integration. In addition, there is also a possibility of single T-DNA elements integration. However, the vast majority of known cT-DNAs (TA, TB, TE, TC, TD in *Nicotiana* plants (Chen et al. 2014) and IbT-DNA1 from *Ipomoea batatas* (Kyndt et al. 2015) form exactly inverted repeats. In this work the same organization of cT-DNA (as inverted repeat) in *Linaria vulgaris* genome was first demonstrated.

Given the digital PCR data which demonstrate at least two insertions and taking into account the number of variants of the elements, we can conclude that there are three cT-DNA loci per haploid genome.

Different variants of the sequences containing the same joining site of cT-DNA with plant DNA (20.4, 89.1, 94.7, Fig. 2, b and 160.1 as allelic variant) were found. It is known that during the agrobacterial transformation T-DNA was integrated in random regions of the plant genome (Mayerhofer 1991; Neve 1997), because the integration mechanism is not site-specific. Thus, in this case the multicopy of cT-DNA is not associated with successive transformation acts, as in *Nicotiana* species containing several cT-DNA, but with the duplication of cT-DNA together with the border regions of plant DNA after integration. Since cT-DNA in *Linaria vulgaris* was integrated into the transposon (Matveeva et al. 2012 and MH298319 sequence), a possible reason for cT-DNA duplications could be a temporal activation of this type of transposon which caused a transposition of its cT-DNA-containing copy. In our opinion the formation of extensively distant variants, can be explained by an unequal crossing-over which leads to the loss of individual elements (Fig. 3)

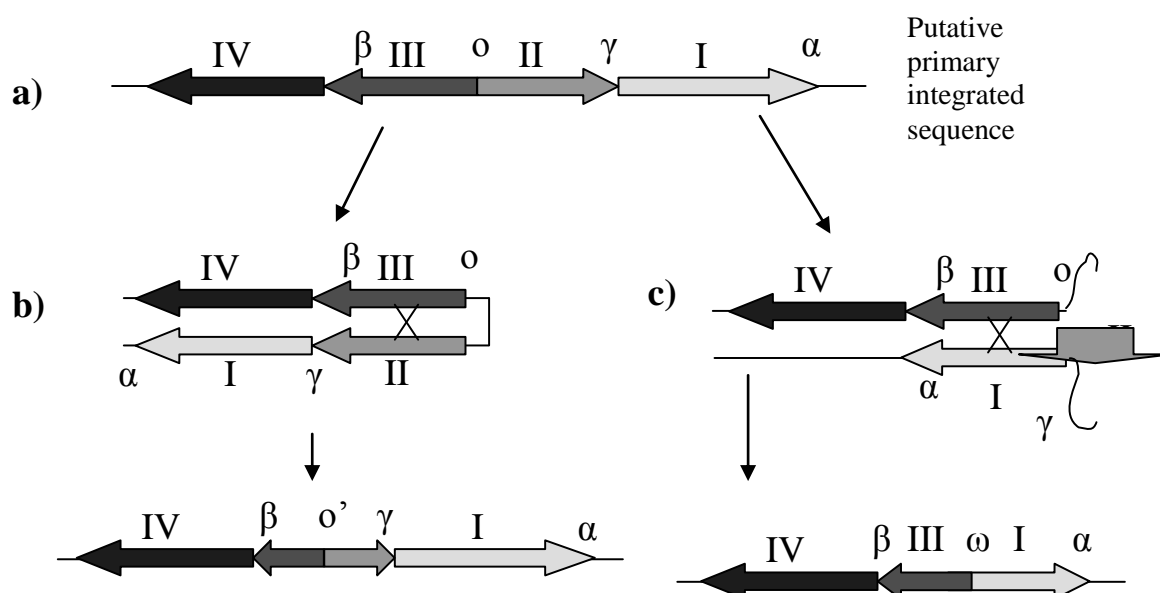


Figure 3. Possible unequal crossing-over can explain the occurrence of nested deletions in the region of o (b) and the formation of ω (c).

Potentially intact copies of the *orf13a* gene were detected in sequences 20.1, 9.13 and 206.2. The *orf13a* may encode the transcription factor (Hansen et al. 1994). Unfortunately, its function is not sufficiently studied.

Comparison of the sequences of different elements reveals, that a number of elements has the same large deletion of *rolC-mis*. This deletion exists in elements located in different positions, for example, I (20.4), II (20.5). Also we have found out, that this deletion is absent in the elements located in similar position (160.1, 18.2). The random occurrence of such deletions with accurately matching boundaries several times seems to be too great assumption. Explanation of the phenomenon due to illegitimate recombination with displacement requires double cross-over which is unlikely at such a short distance.

We assume that the cT-DNA elements undergo (or at least underwent earlier) the mechanism of homogenization (or "concerted evolution"). This process is similar to the mechanism, which aligns repeats of ribosomal RNA and other multicopy sequences (Pavelitz and Rusche 1995; Liao 1999) and maintains their high identity. Distribution of the variants with large deletions this way has also been described (Pavelitz and Rusche 1995).

Another evidence in favor of this hypothesis is the distribution of SNP-polymorphism in T-DNA elements, which cannot be explained by divergent evolution. Phylogenetic trees, constructed on the base of different SNP-sites, contradict each other (Fig. 4). The alignment looks as if elements sequences were copied by fragments of 25-50 bp in size. This picture cannot also be explained by PCR-artifacts, such as chain change, since the elongation time in amplification programs was enough (see Materials and methods). Also mosaic patterns are observed at very short intervals (25-50 bp). However if, we assume, that this is a chain change, then it would have to occur in 25-50 bp steps, which is impossible. Thus, it is impossible to give any alternative hypotheses explaining the observed phenomenon, except concerted evolution.

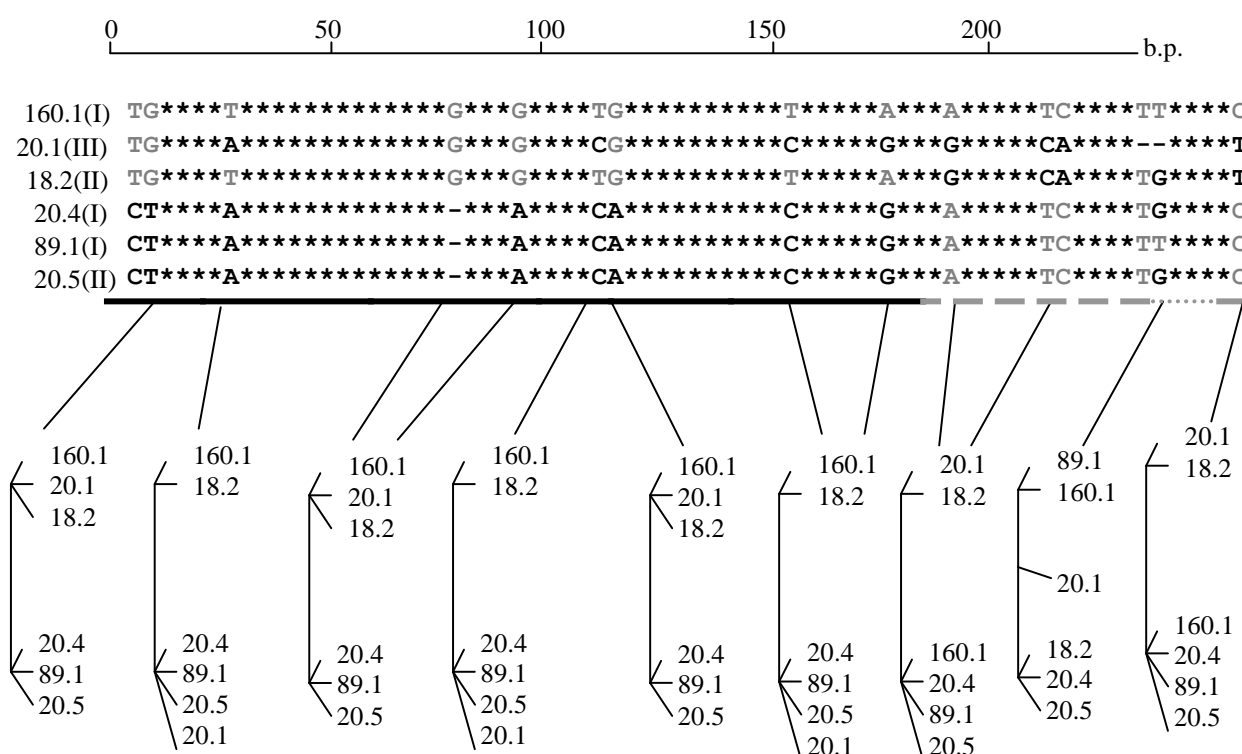


Figure 4. Alignment of *mis* gene fragment in various studied elements (only polymorphic sites are shown). Phylogenetic trees, constructed on the base of different polymorphic sites, contradict each other. Bold lines (solid and dotted) underline polymorphic positions, which give phylogenetic trees of different types.

It seems that the mechanism of concerted evolution works efficiently on high-copy and not very divergent sequences (Liao 1999). Typical examples of such sequences (rDNA, U2 RNA, histone genes) used as models of concerted evolution usually satisfy these requirements. Consequently, for this reason the homogenization process is often observed at final stages, when the alignment has already been completed. In case of *Linaria vulgaris* cT-DNA, individual elements of cT-DNA have undergone less degree of homogenization and it is possible to observe an incomplete state of the process. Particularly, the discreteness of the process is visible.

According to our data, homogenization is progressing due to the replacement of short (25-50 bp) fragments of the sequence, but not whole repeat elements. *Linaria vulgaris* cT-DNA can become a valuable model object for studying the process of concerted evolution.

Conclusions

According to the results, the following conclusions were made:

The structure of cT-DNA in *Linaria vulgaris* has been studied at a deeper level. According to new data, cT-DNA elements form inverted repeats (which is now shown for all cT-DNA-containing plants). The cT-DNA unit is organized as a complex inverted repeat of two simple tandem repeats. This block was duplicated at least twice after integration into the genome. Elements of cT-DNA are subject to homogenization process.

Materials and methods

DNA was isolated from *L. vulgaris* leaves by CTAB-method (Doyle and Doyle 1987).

Long-range PCR (with amplicon size >2 kb) were made with a set of "D-polymerase" (Beagle Ltd., Russia) with a mixture of *Taq* and *Pfu* polymerases.

A set of the specific primers for junction sites was designed (see Fig. 1, Table 1) to exclude the possibility of their non-targeted annealing to other elements. Also, 10 forward (Set of universal forward primers for T-DNA elements, (see Table 1) and 26 reverse universal primers (Set of universal reverse primers for T-DNA elements, (see Table 1) were designed for the internal non-unique regions of the elements (Fig. 5). They were used to test specific primers, and to determine the approximate boundaries of the elements.

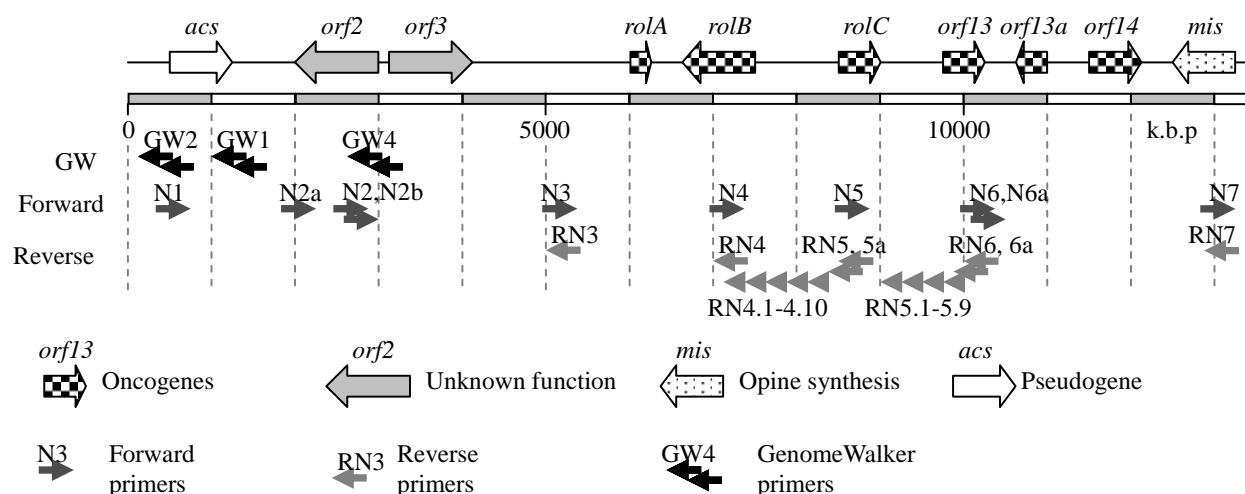


Figure. 5. Positions of GemomeWalker primers and primers for non-unic regions of T-DNA element.

Reaction mixture consisted of 1x Buffer for D-polymerase, 0.3 mM dNTP, 0.5 μ M forward primer, 0.5 μ M reverse primer, 0.125 U / μ l D-polymerase. The reaction was performed in a volume of 20 μ l with 1 μ l of plant DNA in PTC-100 thermocycler (MJ Research) with step-down protocol: 94 $^{\circ}$ C - 3 min, 7 cycles with high annealing / elongation temperature (94 $^{\circ}$ C - 5 s, 70 $^{\circ}$ C - 10 min), 38 main cycles (94 $^{\circ}$ C - 5 s, 65 $^{\circ}$ C - 10 min), and final elongation at 72 $^{\circ}$ C for 7 minutes. PCR results were evaluated by electrophoresis in a 0.5% agarose gel on 1x SB buffer (Goedhart et al. 2005).

Sequencing was performed on a MegaBACE 1000 sequencer (Molecular Dynamics) using the manufacturer's protocol.

Primer design was made for specific cT-DNA sites of junction between elements of repeats with each other or with the adjacent plant DNA (Fig. 1). Previously unknown junction sites were sequenced using the GenomeWalker approach (Siebert et al. 2005). The designed specific primers were checked for operability with a set of non-unique primers of the opposite direction, also the position of the adjacent element boundary was approximately determined.

The long-range PCR with various combinations of specific primers were conducted. Thus, the location of unique sites was studied (by the presence or absence of a PCR product), in the case of a positive result, a pure sequence of single elements was obtained after sequencing of a LongRange-PCR products.

To search for microsatellite markers *L. vulgaris* genomic DNA was sequenced with Ion Torrent System (TermoFisher Scientific, sequencing performed at Azco BioTech, USA). Obtained reads were used only for microsatellite motifs search, since a low coverage does not allow using this data for genome assembly or studying the structure of T-DNA. Digital PCR was performed by Azco BioTech (USA).

Table 1.
Primers list.

Primer name	5'-3' sequence	Purpose
GW1_Ext	TTCGGGATGTTCAACTGGGCAAACAC	Genome Walker external and internal primers
GW1_Int	CACAGCGATACATACACCGTCACGAAAGACT	
GW2_Ext	GCCTTCAACGCCGATCACGCTGCTGTTC	
GW2_Int	TCCTGAGATGCGAGCGATATGTCCTTACC	
GW3_Ext	GCCGCCAACGTGCCAATATGCTAATTTAC	
GW3_Int	ACCTGCCTTCAGACCTACTGTTTGCTGTGCT	
GW4_Ext	ATCGCTGCTTCGCCCTTCAAAGACATTCC	
GW4_Int	CGGAACGGTCACAATCTTCAAAGGTCAACG	
α R	GGTAATGTCGTTTGGACTGAAAAATGCGG	Reverse primer for α region
β R	CATAGCTTGCGGAAAGGGTTAAACGAAAGG	Reverse primer for β region
β F1	CCTTTCGTTTAACCCCTTCCGCAAGCTATG	Forward primer for β region
β F2	AAGATCCCTATTAAACCCCTTTCGTTTAACCC	Forward primer for β region
β^+ F	GATATAAGGTCCCTATTAAACCCCTTGTGCAG	Forward primer for β^+ region
β^+ R	CCTCATAAGATCGGCCTGCACAAGGGGTTA	Reverse primer for β^+ region
γ F	TACATCAACTGCTAGGTGGGGTGCATCG	Forward primer for γ region
γ R1	GCACCCACCTAGCAGTTGATGTAATTTTCG	Reverse primers for γ region
γ R2	CCCACAATATCAAATATCCGCAGACAACCA	
OmegaF	TCCGTGTTCAACAACCTCTAGGAGATGCTCTA	Forward primer for ω region
OmegaR	TCCTAGAGCATCTCCTAGAGGTTGTGAACACG	Reverse primer for ω region
N1	GCAACTCTTCAGGGAATAGGACTGGCGTAAC	Set of universal forward primers for T-DNA elements
N2	CACCGAATTTGCTCTCCATAACCATGTGAAC	
N2a	ACCGATCGGTCAGTTGCTGCGATCTG	
N2b	CTCTGATGTCACCGAATTTGCTCTCCATAAC	
N3	AATAACTCAAGTGTGTGTCCCGTTCGTTGTG	
N4	AAGTGAATGAACAAGGAACTTGCAGAAATGG	
N5	TGCCTCACGATGATGTGTTGCTTCCATAGA	

N6	ATATCGTGCCTCGTTCCTCCTTATCAGACA	Set of universal reverse primers for T-DNA elements
N6a	CTGGTTTAAGGGAGATCCACAATTCCATTTCG	
N7	CAATGCTCTAAATTATTTCAGTCGCCAACGTGC	
RN3	CACAACGAACGGGACACACACTTGAGTTATTG	
RN4	CGCTATTTTTTGCAAGTTCCTTGTTCAATTCA	
RN4_1	CCCTCTAGGAGATGCTCCGTGTTCAACAACC	
RN4_2	CAGGCTTGAGAAAAGGACCAGCATGTATTTG	
RN4_3	CCAAGTCTACAGCAAAACCCTTATGGATTTC	
RN4_4	CCAGCTCACATCACAATGGCTCCCAAATTG	
RN4_5	TTGTAAGGCCGCCATGTCTTGAGCATCG	
RN4_6	CACGAAACCCAATAGTCTGCCAGCTTTAACC	
RN4_7	AGTCCCCAAACAACATGTATTGTTCAATTTGGA	
RN4_8	GCGGAAAGCCACTACTTTGTCAAGCTAAACC	
RN4_9	TGTGGCTTGTATCCAGAACTCGAAGCAACC	
RN4_10	AAGGGTTTCCGTTCATTTGAGGCGTTCTTC	
RN5	CTATGGAAGCAACACATCATCGTGAGGCAGT	
RN5_1	AACCTGATGCCTGCTTGAGCCTTCTGTTGAC	
RN5_2	ACTTTCCTTTGTGCGAAGTTAGCTCCATCTGC	
RN5_3	CTTCTCCTGACATCAAACCTCGTCAGCCCATC	
RN5_4	AGAACGCGAGTTGAAATCAACGTATCCAGTT	
RN5_5	TGAAACAAGAATAACGCAAATTTCAAAGCAG	
RN5_6	GTCGGCCCTTCTTTTCAGTAGACCGATAG	
RN5_7	TGACACGTTGTCCAATAGCTCCTATCATGTCC	
RN5_8	GCCATTCGACTGCCACTTTAGGTTTGTTCAA	
RN5_9	CGGATGCAGCTCCATTTCTGCATTGTTGAC	
RN6	TGTCTGATAAGGAGGAACGACGCACGATAT	
α _RT_F	ATTAAACGAATTTAGGACAAACA	Real-Time test system for α region
α _RT_probe	FAM -CGTTGGTATGTGGYGCCCGCATTTT -BHQ1	
α _RT_R	GGTAATGTCGTTTGGACTGA	

Table 2.

Microsatellites markers list

Nº	Predicted length	Motif	All detected allelic variants (in different plants)	Primers	Tm	Acc #
M002	91	4 x GAA	91	GGCAAACAACCCAGAGAATC	60	MH298301

				CGATCGGGAGTCACATGAT		
M003	81	7 x CA	81, 85, 91	TTCATCCATCTCGCCCTAC TATAGCTGTAATATACTGAC TAACG	60	MH298302
M005	156	6 x GGA	156, 165, 171, 180	CTAACAAGAAATGGGTGGA TGA TATAGGGACGAGTTGGTCA GAT	55	MH298303
M006	215	3 x CCAA, 3 x GGA, 12 x GT	210, 213, 215, 218	CGATGGAAATGTGATAGAA CCA GCGTGTACTCCTTCCACTTT C	55	MH298304
M007	172	10 x AT	168, 170, 172, 176, 180, 186	CTAACTGCACACTTAGAAA TAAGATG TTGTGAAGTAAGTTGTGAA GATGAA	55	MH298305
M009	144	5 x GTT	141, 144, 147, 153	GCAGAAGATTCATGTTTAC GA ATCTGCACCCTATGAGATTT TG	55	MH298306
M010	218	11xGT, 6xGA	194, 214, 218, 220,	CACTTATGGGGCACACTT CCAAGTCCACAGTGCTCAA T	60	MH298307
M012	191	4xGGA	167, 176, 182, 191	GCCTATCACTAGGAGCTACT TCA TTAGGGTCTCTTCCACAAAT GA	60	MH298308
M013	86	7xCCT	83, 86	CTTTTACGGCGACCCACAT AAGAGAAGCGAAGCGAGCA	60	MH298309
M014	101	6xGA	101	GGGTTTAGGCAACTTGTC GT TTCCTATGTGGGATAGTGAA	55	MH298310

				GC		
M015	105	9xGA	105, 107, 109, 111	TTTAACTTCGCCATCGTGA GCTCAAAGTTCTGGCTGTAT TAC	60	MH298311
M016	179	9xTA	179	TGCCTTTGGTTTGATTGCT TTGTATGGCACCACCATTAT C	55	MH298312
M017	173	4xAAT	152, 167, 173, 179	GGGTTTGCACAATAGTTGTA TTAG GTGTTTGAGAGTTTATGTGT CCA	55	MH298313
M018	268	11xATT	266, 268, 274, 280	TTGTACTGGCACGTAGGAA AT GAGGACTTCAAAATAAATG CCTT	60	MH298314
M019	183	3xCA, 9xCT	181, 183, 185	CGTCTGTAGTGTCTGGTTA CA GTAGAAATACTTGTGCTCA ATAGAA	55	MH298315
M020	191	11xAT	191, 209	CACCTTCTACAGAGCCTTCCA GA AGCTTAACAGCCCACAAAC AT	60	MH298316
M021	202	10xAGT	200, 202	GTCGGCTCGTGAGATTCTAA T AGAGTCACAACCTGGTGTAT CGTA	60	MH298317
M023	193	5xCTC, 5xAAG	193	AGACTATCATGGAGACGCT GAA TTATCACTCGAATGGCTCGT A	55	MH298318

Table 3.

Accession numbers.

Sequence name	Acc. number in NCBI
γ sequence	KR612321.1
9.13	MH107074
18.2	MH107075
20.1	MH119454
20.2	MH119451
20.3	MH119450
20.4	MH119453
20.5	MH119452
89.1	MH119458
94.7	MH119455
106.10	MH119459
135.11	MH119456
160.1	MH119460
206.2	MH119457
right border	MH298319

Authors' contributions

Conceived and designed the experiments: BD, IV OP.

Conducted the experiments: BD, IV OP.

Performed the analysis: BD, IV OP, DP.

Wrote the paper: IV, BD, OP.

All authors read and approved the final manuscript.

Competing interests

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank Prof. Aleksandr V. Rodionov, Dr. Elena A. Andreeva, Dr. Yuliya V. Sopova, Dr. Mikhail Liskovych, Dr. Maria S. Vishnevskaya, the Editor and Anonymous reviewers whose comments greatly improved this manuscript.

References

1. Aoki S, Syono K. (1999) Horizontal gene transfer and mutation: *Ngrol* genes in the genome of *Nicotiana glauca*. PNAS 96:13229–13234.
2. Aoki S, Syono K. (2000) The roles of *Rirol* and *Ngrol* genes in hairy root induction in *Nicotiana debneyi*. Plant Sci 159:183–189.
3. 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(4):669–682. <https://doi.org/10.1111/tpj.12661>

4. Doyle, J.J., Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11–15.
5. Frundt C, Meyer AD, Ichikawa T, Meins FJ (1998) A tobacco homologue of the Ri-plasmid *orf13* gene causes cell proliferation in carrot root discs. *Mol Gen Genet* 259:559–568.
6. Goedhart J, Theodorus WJ, Gadella J (2005) Analysis of oligonucleotide annealing by electrophoresis in agarose gels using sodium borate conductive medium. *Analytical Biochemistry* 343:186–187. [https://doi: 10.1016/j.ab.2005.04.048](https://doi.org/10.1016/j.ab.2005.04.048)
7. Hansen G, Vaubert D, Cl  rot D, Temp   J, Brevet J (1994) A new open reading frame, encoding a putative regulatory protein, in *Agrobacterium rhizogenes* T-DNA. *C R Acad Sci* 317(1):49-53.
8. Intrieri MC, Buiatti M (2001) The horizontal transfer of *Agrobacterium rhizogenes* genes and the evolution of the genus *Nicotiana*. *Mol Phylogenet Evol.* 20(1):100-110. [https://doi: 10.1006/mpev.2001.0927](https://doi.org/10.1006/mpev.2001.0927)
9. Kyndt T, Quispe D, Zhai H, Jarret R, Ghislain M, Liu Q, Gheysen G, Kreuzeb JF (2015) The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: An example of a naturally transgenic food crop. *PNAS* 112(18):5844-5849. [https://doi: 10.1073/pnas.1419685112](https://doi.org/10.1073/pnas.1419685112)
10. Liao D (1999) MOLECULAR EVOLUTION Concerted Evolution: Molecular Mechanism and Biological Implications *Am J Hum Genet* 64:24–30.
11. 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. [https://doi: 10.1094/MPMI-07-12-0169-R](https://doi.org/10.1094/MPMI-07-12-0169-R).
12. Mayerhofer R, Koncz-Kalman Z, Nawrath C, Bakkeren G, Cramer A, Angelis K, Redei GP, Schell J, Hohn B, Koncz C (1991) T-DNA integration: a mode of illegitimate recombination in plants. *EMBO J* 10:697-704.
13. Meyer AD, Ichikawa T, Meins F (1995) Horizontal gene transfer: regulated expression of a tobacco homologue of the *Agrobacterium rhizogenes* *rolC* gene. *Mol Gen Genet* 249:265–273. doi: 10.1007/BF00290526
14. Neve MD, Buck SD, Jacobs A, Van Montagu M, Depicker A (1997) T-DNA integration patterns in cotransformed plant cells suggest that T-DNA repeats originate from co-integration of separate T-DNAs. *The Plant Journal* 11(1):15–29.
15. Pavelitz T, Rusche L (1995) Concerted evolution of the tandem array encoding primate U2 snRNA occurs in situ, without changing the cytological context of the RNU2 locus. *The EMBO Journal* 14(1):169-177.
16. Pavlova OA, Matveeva TV, Lutova LA (2014) Genome of *Linaria dalmatika* contains *Agrobacterium rhizogenes* *rolC* gene homolog. *RJGAR* 4(5):461–465. [https://doi 10.1134/S2079059714050116](https://doi.org/10.1134/S2079059714050116)
17. Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved method for walking in uncloned genomic DNA. *Nucleic Acids Res* 23:1087–1088.
18. Suzuki K, Yamashita I, Tanaka N (2002) Tobacco plants were transformed by *Agrobacterium rhizogenes* infection during their evolution. *Plant J* 32:775–787.
19. Tandon SL, Bali PN. (1957) Morphological and cytological studies of the diploid and the colchicines induced tetraploid in *Linaria vulgaris*. *Genetica* 29:101–109. doi: 10.1007/BF01535703.
20. Tzfira T, Citovsky V (2006) *Agrobacterium*-mediated genetic transformation of plants: biology and biotechnology. *Curr Opin Biotechnol* 17:147–154. doi: 10.1016/j.copbio.2006.01.009
21. White FF, Garfinkel DJ, Huffman GA, Gordon MP, Nester EW (1983). Sequence homologous to *Agrobacterium rhizogenes* TDNA in the genomes of uninfected plants. *Nature* 301:348–350.