Highly variable tRNA-Cys gene clusters in *Arabidopsis thaliana*

Maciej Szymanski (✉ mszyman@amu.edu.pl)

Department of Computational Biology, Adam Mickiewicz University in Poznan

Anand Maurya

Department of Computational Biology, Adam Mickiewicz University in Poznan

Wojciech M. Karlowski

Department of Computational Biology, Adam Mickiewicz University in Poznan

Research Article

Keywords: tRNA, tRNA gene clusters, transposable elements, Helitron

Posted Date: February 8th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1314586/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Highly variable tRNA-Cys gene clusters in *Arabidopsis thaliana*

Maciej Szymanski*, Anand Maurya and Wojciech M. Karlowski*

Department of Computational Biology, Adam Mickiewicz University in Poznan, Uniwersytetu Poznanskiego 6, Poznan, Poland.

*) corresponding authors

ABSTRACT

Although most of the genes encoding tRNAs in plants are dispersed throughout the genome, a fraction of them form tRNA gene clusters. In *Arabidopsis thaliana*, the smallest of tRNA clusters on chromosome 5 consists of four tRNA-Cys(GCA) genes placed within repeating units of 0.4 kbp. A systematic analysis of the genomic sequences of syntenic regions from various ecotypes of *A. thaliana* showed that the general structure of the cluster, consisting of a tRNA-Cys pseudogene followed by repeating units containing tRNA-Cys genes, is well conserved. However, there is significant heterogeneity in the number of repeating units between different ecotypes. A unique feature of this cluster is the presence of putative transposable elements (Helitron). In addition, two further tRNA-Cys gene mini-clusters (gene pairs) in *A. thaliana* were identified. RNA-seq based evaluation of expression of tRNA-Cys-GCA genes showed positive signal for 11 out of 13 unique transcripts. An analysis of the conservation of the tRNA-Cys clusters from *A. thaliana* with the corresponding regions from four other *Arabidopsis* species suggests a sequence of events that led to the divergence of these regions.

KEYWORDS

tRNA, tRNA gene clusters, transposable elements, Helitron

INTRODUCTION

Transfer RNAs (tRNAs) play a vital role in all organisms. In the process of protein biosynthesis, aminoacylated tRNA molecules serve as an interface between the genetic information contained in the protein-coding genes and its expression in the form of functional proteins. Thus, the efficient and error-free production of proteins in the cell depends on the expression of the full set of tRNAs capable of decoding all of the codons used in messenger RNAs.

Although the tRNAs from all domains of life share all of the essential features of a common cloverleaf secondary structure, the mechanisms of their expression and gene organization differ between prokaryotes (and organelles) and eukaryotes. In Bacteria, Archaea and organelles tRNAs can be expressed either as monocistronic transcripts from individual genes or as parts of polycistronic transcripts from tRNA and ribosomal RNA operons. A common feature of the prokaryotic tRNA genes transcription is their dependence on typical bacterial promoters consisting of -35 and -10 elements ¹.

With a few exceptions, in eukaryotes, each tRNA is transcribed from an independent gene by a specialized RNA polymerase III (Pol III). Transcription of eukaryotic tRNA genes depends on the internal promoter elements, boxes A and B recognized by transcription factor IIIC. The positions of boxes A and B within the tRNA gene correspond to tRNA structure elements: D- and T-stem-loop, respectively ².

The majority of tRNA genes in higher eukaryotes are more or less evenly dispersed throughout the genome. In some instances, as a consequence of gene duplication, tRNA gene clusters can be formed. The tRNA gene clusters in eukaryotes are defined as genomic regions containing only tRNA genes. Sometimes in defining such clusters an additional criterion of the maximum size of the intergenic regions between tRNA genes (usually 1 kbp) is taken into account. In contrast to Bacteria, in which the clusters of coexpressed tRNA genes are usually composed of genes encoding tRNAs with different amino acid specificities, the eukaryotic tRNA gene clusters are often composed of closely related tRNA specific for one or two amino acids ³.

In the *Arabidopsis* genome, three clusters of tRNA-Pro genes, one cluster of tRNA-Ser and tRNA-Tyr, and one cluster of tRNA-Cys were identified ^{4,5}. In the case of tRNA-Pro clusters, the genes and pseudogenes often overlap with the intronic regions of the protein-coding and ncRNA genes on the opposite strand. On the other hand, in the largest cluster, spanning almost 40 kbp of chromosome 1 and consisting of 27 tandemly repeating units composed of one tRNA-Ser and two tRNA-Tyr genes, there are no other annotated genes. The smallest of

the tRNA gene clusters described so far in *A. thaliana* reference genome consists of four tRNA-Cys(GCA) genes and overlaps with three annotated Helitron superfamily transposable elements.

In this study, we present a detailed analysis of the variability of the tRNA-Cys gene cluster based on the genomic sequences of various *A. thaliana* accessions. We analyzed syntenic regions of chromosome 5 from 29 genomes derived from different ecotypes. The genomic sequences from these ecotypes show a high variation of the copy number of the repeating units and the nucleotide sequences of the tRNA genes. Moreover, the survey of the genomic sequences revealed two additional mini-clusters of tRNA-Cys genes on chromosomes 1 and 2, that in *A. thaliana* consist of two genes in head-to-tail or head-to-head orientation.

In this paper we also show that although the entire chromosome 5 tRNA-Cys cluster is apparently epigenetically silenced ⁵, the transcriptional inactivation is not complete. An analysis of the tRNA expression in the reference Col-0 strain demonstrated that one of the tRNA genes located at the cluster's boundary shows a low level expression.

RESULTS

The Cys-tRNA cluster on chromosome 5 is preceded by a pseudogene and contains marks of transposable elements

In the reference sequence of the *A. thaliana* genome (TAIR10) ⁶ the Cys-tRNA cluster is located on chromosome 5 and consists of four tandemly arranged, ~420 bp long, repeating units, each of which contains a tRNA gene (Fig.1). The pairwise identity of particular repeats of the cluster ranges from 67% to 100%. All four tRNA genes annotated in the TAIR10 genome pass the tRNAScan-SE ⁷ high confidence filter which suggests that they may encode fully functional tRNA molecules. The nucleotide sequences of three of the tRNA genes are unique for the chromosome 5 cluster. For one of the genes, there is another copy encoding identical mature tRNA on chromosome 4. Moreover, in the region flanking the first repeating unit (AT5G20858) we have identified a sequence that is derived from the tRNA-Cys. This sequence was not identified in the tRNAScan-SE predictions and can be regarded as a tRNA-Cys pseudogene (see later in text).

Fig.1. Localization of the tRNA-Cys gene cluster on chromosome 5 in the reference genome of *A. thaliana* (TAIR10). The tRNA genes and transposable elements annotated within this region are shown with blue and green arrows, respectively.

A unique feature of the tRNA-Cys cluster, when compared with other tRNA gene clusters, is the presence of transposable elements (TEs) in this region (Fig.1). In the annotation of the reference genome, the entire cluster overlaps with three transposable elements belonging to the Helitron superfamily: ATREP5 (AT5TE25590), HELITRONY1D (AT5TE25595) and ATREP10A (AT5TE25560). The HELITRONY1D and ATREP5 have the same polarity as the tRNA genes and they have lengths corresponding to one or two tRNA-Cys gene cluster repeating units, respectively. Taking into account a high degree of identity between the repeats, the cluster could also be regarded as being composed of tandem repeats of these TEs. On the other hand, the 5'-terminal fragment of ATREP10A overlaps the first two repeats with the reversed polarity in respect to the tRNA genes.

Since the TAIR annotation of this region is based on the sequence similarity to the database of repetitive elements⁸, we explored further the properties of this genomic fragment using bioinformatic tools designed for identification of Helitron elements: HelitronFinder⁹, HelSearch¹⁰, HelitronScanner¹¹ and EAHelitron¹². The first tool (HelitronFinder) was not available for download at the time of the study. Analysis of both strands of the genomic fragment containing the tRNA-Cys cluster (with 500 bp extensions on both sides) as well as the entire chromosome 5 with the remaining three algorithms did not reveal any reliable prediction of Helitron transposable element that would overlap the tRNA genes located within the cluster.

The structure of the chromosome 5 tRNA-Cys gene cluster shows high variability in *A. thaliana* accessions.

Comparative analysis of the tRNA-Cys cluster from the reference genome with the corresponding regions from the genomic sequences of other *A. thaliana* accessions using dot-plots revealed a significant size heterogeneity of this region due to a variable number of repeats (Fig.2). Detailed analysis of these regions showed that the number of repeating units

in the genomic sequences from different *A. thaliana* accessions varies from four to six. In the majority of genomes, there are four repeats, as in the case of the reference Col-0 genome. Sequence comparison of the clusters and flanking regions suggested that the first repeating units and the sequences downstream of the last tRNA gene are homologous. Thus, the heterogeneity results from the variation of the number of copies of repeats between these regions. In most cases, the heterogeneity in the length of the cluster is a consequence of the insertion or deletion of the entire repeating units including tRNA genes and spacer regions (Fig.3). The only exception is the cluster from the Cvi accession in which there is a deletion of a 167 bp region that includes 95 base pairs of the repeat 5 and the tRNA gene of the repeat 6. This arrangement was also confirmed by the analysis of the sequence of the diploid genome from the F1 cross between Col-0 and Cvi-1 in which on each of the chromosomes 5 there are clusters identical to clusters present in parental strains. The detailed multiple sequence alignment of all analyzed regions is provided in the supplementary materials.

Fig. 2 Dot-plot analysis of the heterogeneity of the chromosome 5 tRNA-Cys clusters for selected *A. thaliana* accessions. The dot-plots represent alignments of the chromosome 5 region from the reference genome (Col-0, TAIR10) with syntenic regions from genomic sequences of (A) Ler (B) Cvi (C) C24 and (D) Kn-0 (PacBio sequence). The tRNA-Cys genes and the pseudogenes are shown as green and red boxes, respectively.

Fig. 3. (A) Evolutionary relationship between the chromosome 5 tRNA-Cys cluster sequences from the analyzed genomes and (B) the structures of the clusters. Blue lines represent sequences of the repeating units and the tRNA genes are shown as triangles pointing in the direction of transcription. The open triangles designate pseudogenes either predicted by the tRNAScan-SE or by sequence similarity. Different colors are used for tRNAs belonging to distinct groups defined based on 97% sequence identity.

Overall, pairwise identities between the tRNA-Cys clusters in all of the analyzed genomes are in the range of ~70 to 100 %. However, due to relatively high sequence similarity between repeating units, it is very difficult to establish proper alignment of the entire genomic fragments that would unambiguously reflect the relationship between particular regions in different genomic sequences. Using the phylogenetic analysis of the individual repeats, we were able to group the corresponding repeat units in all analyzed genomic sequences. The maps of the relative arrangement of these units presented in Fig. 3 were constructed to

account for the sequence similarities between corresponding repeat units from different genomes. The optimal phylogenetic tree of all the individual repeat unit sequences is presented in supplementary Fig.S1. The tree shows that the distinct groups of repeating units are very closely related. However, very weak support for most of the internal tree branches from bootstrap analysis suggests that in the case of some of the genomes the real relationship between particular repeats may be different. Interestingly, there is a difference in the structures of the cluster between the two Kn-0 strain genomic assemblies obtained in different projects. The two clusters differ in size due to the presence or absence of one repeating unit.

The tRNA-Cys genes exhibit high sequence variability within the cluster and between *A. thaliana* ecotypes.

All of the tRNA genes in the tRNA-Cys gene cluster are derived from the same variant of the tRNA-Cys(GCA). The sequence alignment of all tRNA regions is included in the supplementary material. There is, however, a notable sequence variation both between genes within one genome (90-100% identity) as well as between ecotypes (88-100% identity). Almost all of the tRNA genes in all of the analyzed regions were classified as 'high confidence' by the EukHighConfidenceFilter, shown in Fig. 3 as filled triangles. The 'high confidence set' consists of sequences best matching the canonical model of eukaryotic tRNAs. The sequences rejected by the filter do not possess all features of the secondary or tertiary structure of the ideal tRNA cloverleaf or show other unusual features like anticodon-tRNA isotype mismatch⁷. The pseudogene at the 5'-end of the repeat 1 was predicted by tRNAScan-SE only in one genomic sequence of the KBS-Mac-74 accession. However, in all genomes, in the corresponding region, there are almost identical sequences present that can also be regarded as pseudogenes derived from the tRNA-Cys(GCA) gene. Thus in all of the analyzed *A. thaliana* accessions, the cluster is flanked on one side by the well-conserved pseudogene showing, in pairwise comparison between any two sequences, 94 to 100% identity.

Two novel mini clusters of the tRNA-Cys genes are located on chromosomes 1 and 2.

An analysis of the tRNAScan-SE predictions of tRNA genes in the genomic sequences of various *A. thaliana* accessions revealed additional, previously not described clusters of tRNA-Cys genes. In the TAIR10 genome, there is a short region on chromosome 2 containing two tRNA-Cys genes in the head-to-head orientation separated by a 134 bp spacer (Fig. 4A). Both genes encode the same sequence variant of tRNA which shows evidence of high expression in the high-throughput sequencing data (Table 1). The mature

tRNAs transcribed from both of these genes would have identical sequences and it is impossible to discriminate between transcripts produced from either of them.

Fig. 4. Mini clusters of tRNA-Cys genes. A) A cluster of divergently transcribed tRNA-Cys genes on chromosome 2 conserved in all *A. thaliana* accessions analyzed in this study. B) A cluster of tandemly

repeated tRNA-Cys genes overlapping in antisense orientation the second intron of the gene encoding ubiquitin domain-containing protein on chromosome 1. C) Schematic alignment of the chromosome 1 tRNA-Cys cluster region from various *A. thaliana* accessions.

A survey of the genomic sequences from different *A. thaliana* accessions demonstrated that such an arrangement of the two genes is highly conserved in all genomes analyzed in this study. There is very little variation in the tRNA-coding sequences in this mini-cluster, represented by only three tRNA sequence variants with single substitutions within the T or acceptor stems. In two instances, Can-0 and Cvi-0, in one of the genes there is a deletion of 11 base pairs within the region corresponding to the T-loop and two base pairs of the T-stem.

The second mini-cluster of tRNA-Cys genes was identified on chromosome 1 overlapping in antisense orientation the second intron of the gene encoding ubiquitin domain protein (AT1G53400). In the reference TAIR10 genome, there is only one gene, but in eight of the *A. thaliana* ecotypes genomes, we have detected two tandemly arranged tRNA-Cys genes separated by a short (120 bp) spacer (Fig. 4B and C). In the genomes in which there is only one tRNA-Cys gene, elements of the spacer and sequences flanking the missing gene are still present. The sequence of the AT1G53410 gene in TAIR10 is identical with the sequence of another copy of the tRNA-Cys gene on chromosome 1. The sequencing data show that the signal corresponding to either of these genes account for ~13% of all tRNA-Cys reads, but it is impossible to determine whether these reads are the products of transcription of both genes or only one of them.

Most of the tRNA sequences within this region can form a canonical tRNA secondary structure. In two accessions, Ws-0 and Tsu-0, in which there is a single tRNA gene copy, a small deletion of 15 nucleotides involves the 3' portion of the D-stem and 5' portion of the anticodon stem thus disrupting the canonical tRNA structure. Similar sequence variation was also observed in one of the copies in Sha and Nd-1 ecotypes in which there are two tandemly repeated tRNA-Cys genes in this region. In the accessions Ty-1, Cdm-0, Cvi and KBS-Mac-74, in which there are two tRNA-Cys gene copies, one of them shows single nucleotide substitutions resulting in mispairing within the T-stem (Ty-1, Cdm-0, Cvi) or acceptor stem (KBS-Mac-74). In these cases, the sequence of the other gene copy can form a perfect tRNA structure. For details on the arrangement of the genes and sequence variation see the supplementary file with multiple sequence alignment of the syntenic regions from all analyzed genomes.

Only 2 out of 13 the tRNA-Cys transcripts can not be detected in *A. thaliana* Col-0.

In the TAIR10 reference sequence of the *A. thaliana* genome, there are 16 predicted tRNA-Cys(GCA) genes represented by 13 unique mature tRNA transcripts¹³. Using the high throughput tRNA sequencing data (PRJNA505412) we investigated the expression levels of all the predicted tRNA-Cys variants. The comparison of the sequencing results from the libraries constructed with deacylated and control tRNAs showed detectable expression of all, except for tRNA-Cys-GCA-11 (AT5G20858) and tRNA-Cys-GCA-12 (AT5G20856) (Table 1). Both of these transcriptionally silent genes are located within the tRNA-Cys gene cluster on chromosome 5. For the other two genes, tRNA-Cys-GCA-9 (AT5G20854) and tRNA-Cys-GCA-7 (AT5G20852), low-level expression was detected. The tRNA-Cys-GCA-9 transcript can be encoded by an additional gene on chromosome 4 whose expression could account for the observed mature tRNA reads. An analysis of all the reads, including those representing unprocessed precursors that can discriminate between transcripts of these two genes, demonstrate that the tRNA-Cys-GCA-9-2 gene in the cluster is in fact silent. However, the reads matching the tRNA-Cys-GCA-7 variant can only be attributed to the expression of the last tRNA-Cys gene in the cluster.

tRNA-Cys(GCA) variant	GtRNAdb gene symbol	tRNA gene cluster	Deacylated libraries		Non-deacylated libraries	
			average	% tRNA- Cys	average	% tRNA- Cys
tRNA-Cys-GCA-1	tRNA-Cys-GCA-1-1 tRNA-Cys-GCA-1-2	chr1 –	21519.8	13.36	10108.7	26.91
tRNA-Cys-GCA-6	tRNA-Cys-GCA-6-1	–	26932.0	16.73	6835.3	18.19
tRNA-Cys-GCA-2	tRNA-Cys-GCA-2-1 tRNA-Cys-GCA-2-2	chr2 chr2	21524.7	13.37	4820.1	12.83
tRNA-Cys-GCA-13	tRNA-Cys-GCA-13-1	–	68481.1	42.54	11337.9	30.18
tRNA-Cys-GCA-4	tRNA-Cys-GCA-4-1	–	2514.3	1.56	495.9	1.32
tRNA-Cys-GCA-5	tRNA-Cys-GCA-5-1	–	2981.4	1.85	466.9	1.24
tRNA-Cys-GCA-9	tRNA-Cys-GCA-9-1 tRNA-Cys-GCA-9-2	– chr5	266.9	0.16	76.3	0.20
tRNA-Cys-GCA-8	tRNA-Cys-GCA-8-1	–	8196.1	5.09	2200.7	5.85
tRNA-Cys-GCA-3	tRNA-Cys-GCA-3-1	–	7379.1	4.58	1174.2	3.13
tRNA-Cys-GCA-7	tRNA-Cys-GCA-7-1	chr5	1157	0.71	54.8	0.15
tRNA-Cys-GCA-12	tRNA-Cys-GCA-12-1	chr5	0.9	0	0.0	0
tRNA-Cys-GCA-11	tRNA-Cys-GCA-11-1	chr5	2.6	0	0.7	0
tRNA-Cys-GCA-14	tRNA-Cys-GCA-14-1	–	0.0	0	0.0	0
CysGCA.pseudo*	–	chr5	0.0	0	0.0	0

Table 1. Expression of tRNA-Cys(GCA) variants in *A. thaliana* Col-0. For both deacylated and non-deacylated samples, averages represent normalized read count values from two libraries.

*) the pseudogene from chromosome 5 tRNA-Cys genes cluster is not predicted by tRNAScan-SE in TAIR10 genomic sequence but was identified by sequence similarity with other tRNA-Cys genes.

The tRNA-Cys gene clusters are very variable in other *Arabidopsis* species.

To check if the clustered arrangement of the tRNA-Cys genes is a conserved feature, we surveyed the available genomes from four other species belonging to the genus *Arabidopsis*: *A. halleri*, *A. lyrata*, *A. suecica*, and *A. arenosa*. We could identify the entire clusters or

individual repeat elements in different assemblies of genomic sequences from all four species.

The single repeating units of the chromosome 5 cluster could be found in all genomes. In *A. lyrata*, *A. halleri*, and *A. arenosa* the fragments consisting of the spacer region with one or two flanking tRNA genes were found. The largest cluster, consisting of four repeating units, was identified in the *A. suecica* genome (GCA_905175345.1). The sequence of this region shows a high sequence identity with the sequence from *A. thaliana*. This finding suggests, that the amplification of the cluster occurred in *A. thaliana* genome before the origin of *A. suecica* which is a hybrid between *A. thaliana* and *A. arenosa*

The mini-cluster of tRNA-Cys genes on chromosome 2 is conserved in *A. halleri*, *A. suecica*, and *A. arenosa* where, apart from the differences in the nucleotide sequences, there are no structural variations between different species. We could not identify the entire cluster in either of the available genomic sequences from *A. lyrata* in which a single gene is present.

The second mini-cluster overlapping the intronic sequence of the *A. thaliana* AT1G53400 gene is also present in all analyzed genomes and consists of at least two genes as in the case of eight *A. thaliana* ecotypes discussed above. However, in *A. halleri*, *A. lyrata*, and *A. arenosa*, the cluster is expanded and includes an additional tRNA-Cys gene and a pseudogene (Fig. 5). The alignment comparing sequences of this cluster from different *Arabidopsis* species is a part of the supplementary material. The sequence similarity, including the conserved anticodon loop, suggests that the pseudogene is derived from the tRNA-Cys(GCA) gene. The distances between the additional copies also indicate that the larger cluster was a result of the duplication of a fragment containing the entire pair of tRNA-Cys genes (the min-cluster preserved in some *A. thaliana* ecotypes).

Fig. 5. The structures of the tRNA-Cys clusters corresponding to the *A. thaliana* cluster on chromosome 1. The clusters from *A. suecica*, *A. halleri*, *A. lyrata*, and *A. arenosa* are compared with corresponding clusters from *A. thaliana* Col-0 and Cvi. The pseudogenes are shown as open arrows.

DISCUSSION

In this study, we analyzed the clustered arrangement of the tRNA-Cys genes in the genomic sequences of various *A. thaliana* accessions. In addition to the previously described cluster on chromosome 5, we identified two additional mini-clusters on chromosomes 1 and 2. Two of the analyzed regions, on chromosome 5 and chromosome 1 show heterogeneity of the number of tRNA gene copies between different ecotypes.

The tRNA-CysGCA cluster on chromosome 5 is the smallest of tRNA clusters initially identified in the *A. thaliana* genome with a relatively small size of the repeating units (~420 bp). The comparison of regions identified by tRNAscan-SE showed that the majority of tRNA genes within the tRNA-Cys clusters may represent functional genes, classified as 'high confidence' with only a small fraction classified as pseudogenes or genes encoding potentially defective tRNAs in terms of the secondary or tertiary structures. The comparison of the syntenic regions from the complete genomic sequences of various *A. thaliana* accessions suggested that there is significant structural heterogeneity between different ecotypes. In genomic sequences, we identified variants consisting of four to six repeating units. A common feature in all of the analyzed genomes is the presence of a tRNA-Cys pseudogene in the 5'-flanking region (relative to the direction of transcription). In all cases, the regions flanking the cluster are almost identical. A striking feature of the chromosome 5 cluster is a regular pattern of variation. The genomic sequences always differ by the presence or absence of the entire repeating units.

Earlier studies on the expression of clustered tRNA genes in *A. thaliana* suggested that all of them are transcriptionally silent. The silencing of all four tRNA clusters was attributed to increased DNA methylation and the presence of histone modifications associated with transcriptionally repressed heterochromatic regions⁵. Thus, despite the presence of intact promoter elements and Pol III termination signals, the genes within the chromosome 5 tRNA-Cys cluster were presumed to be transcriptionally inactive. Our analysis of the high-throughput tRNA sequencing data confirmed that three of the four genes in the reference strain Col-0 are not transcribed, but we observed sequencing reads corresponding to the last tRNA-Cys gene in the cluster. The presence of sequencing reads in both deacetylated and non-deacetylated tRNA libraries also suggests that the product of this gene is correctly processed and aminoacylated. However, when compared with other tRNA-Cys loci in the genome, the expression of this gene is relatively low. These results suggest that despite epigenetic modifications of DNA and histones associated with transcriptionally silent

chromatin in this region, the silencing effects may be less efficient closer to the boundary of the cluster.

An interesting feature of the chromosome 5 tRNA-Cys cluster is the presence of the putative transposable elements. According to the TAIR10 annotation, the tRNA-Cys cluster overlaps with three transposable elements belonging to the Helitron family. It has been suggested that the TEs from this group transpose using a rolling circle mechanism via a single-stranded DNA intermediate, analogous to the rolling-circle mechanism involved in the propagation of bacterial transposons. Helitrons encode a Y2-type tyrosine recombinase responsible for catalytic activity associated with transposition, and at the 5'- and 3'-ends they have conserved TC and CTRR sequences, respectively. In addition, the Helitron TEs frequently have a short palindromic sequence located very close to the 3' end ⁸.

Helitrons have been reported to carry genomics fragments from different chromosomal locations. However, in most of the studies, only protein-coding genes and gene fragments were reported as potential cargo. If the Helitron transposition were in fact involved in the amplification of the tRNA-Cys cluster it would represent the first example of the non-coding RNA captured by a Helitron family transposon.

The requirement of the Helitron replication process for the hairpin structure at the 3' termini and the presence of tRNA in this location, open a tempting opportunity for speculation about the possible role of this tRNA in its replication process. As a first option, the tRNA structure could be used as a strong termination signal during amplification. The alternative scenario would assume that the presence of a strong tRNA structure close to the 3' end of the transposon disturbs the hairpin terminator. The absence of positive results from Helitron annotation tools seems to support the second alternative. In addition, we have conducted a manual search for TC/CTRR signals (data not shown) and found at least one CTAG motif followed by T or C for each of the repeating units containing one tRNA gene. Such an arrangement of the Helitron conserved signals supports the tandem copy duplication mode of the tRNA-Cys containing TEs. It has been reported for bacterial *IS91* elements that only one end is necessary to initiate transposition. Such one-ended transposition, in the absence of a precise termination signal, results in the generation of tandem copies of the donor plasmid ¹⁴. Hence, failure to recognize the termination signal for Helitron transposition may result in local duplication of the element. In such a case, the tRNA-Cys cluster could be regarded as an array of short Helitron elements with a captured tRNA-Cys gene preventing the formation of the correct termination signal for transposition.

The structural variation of the chromosome 1 tRNA-Cys cluster can be explained by two alternative evolutionary models. The first one assumes the presence of the four-gene cluster

in the common ancestor of all tested *Arabidopsis* species that is preserved in the *A. halleri*, *A. lyrata*, and *A. arenosa* lineages. In this model, the cluster would have undergone a reduction in the *A. thaliana* lineage leading to the origin of the two- or one-gene arrangements. In some cases, such reduction may have even included only a part of the tRNA gene. A second model would assume the presence of the two-gene cluster in the common ancestor that underwent a reduction in the *A. thaliana* genome and expansion of the whole ancestral cluster in the *A. halleri*, *A. lyrata*, and *A. arenosa* lineage.

In addition, regardless of the evolutionary pathways and molecular mechanisms involved in the creation of the observed variability of the tRNA-Cys clusters, the presented study expands the repertoire of the reference tRNA genes and transcripts that can be assigned to the *A. thaliana*.

MATERIALS AND METHODS

Sources of genomic data.

The genomic sequences for various accessions of *A. thaliana* were downloaded from the NCBI Assembly database (<https://www.ncbi.nlm.nih.gov/assembly/>)¹⁵ and the 1001 Genomes Project website (<https://1001genomes.org/>)^{16–18}. The syntenic regions of chromosome 5 corresponding to the Cys-tRNA gene cluster on chromosome 5 in the reference TAIR10 genome were identified by BLAST¹⁹. Only the unique sequences for particular accessions were retained. The data were further filtered to eliminate low-quality sequences based on the content of ambiguous (non-A/T/C/G) positions. The sequences containing over 1% of ambiguous positions were removed from the data set. The summary of the sources of genomic sequence in the final dataset used for particular accessions is shown in Supplementary Table 1.

Predictions of tRNA genes.

The regions of the tRNA genes were predicted using tRNAscan-SE 2.0 software⁷. The results of predictions were validated using the EukHighConfidenceFilter included in the tRNAscan-SE package to identify regions that can be regarded as high confidence tRNA genes and regions potentially encoding tRNAs that do not entirely conform to the canonical tRNA structure (secondary or tertiary) or are likely tRNA pseudogenes.

Prediction of Helitron transposons.

For the prediction of Helitron transposable elements, we have used the genomic sequence fragment corresponding to the tRNA-Cys cluster extended with 500bp flanking regions on each site. The annotation procedures were also conducted on the whole chromosome 5. In each case, both strands were searched for the presence of the Helitrons. The predictions were generated with HelSearch ¹⁰, HelitronScanner ¹¹, and EAHelitron ¹² using in each case the default parameters. The results of the annotation were compared with a gff3 file containing complete annotation (genes and pseudogenes) of the tRNA-Cys cluster.

Sequence analysis.

For the comparative analysis of the genomic regions and the generation of the dot plots, we used the FlexiDot software ²⁰. Clustering of the tRNA gene sequences based on the identity was performed with USEARCH *cluster_fast* command ²¹.

The sequence alignments were created using ClustalW ²². For all phylogenetic analyses of the genomic sequences and generation of phylogenetic trees, MEGA X software was used ²³. The pairwise identity values between sequences were calculated from the global alignments using the *needle* program from the EMBOSS package ^{24,25}.

Assessing the tRNA expression levels from the high-throughput sequencing data.

The expression data were downloaded from the NCBI SRA database ²⁶ project number PRJNA505412. The download files included two replicates of control and deacylated RNAs isolated from 2-weeks old Arabidopsis seedlings. The data set was reduced to unique sequences and reads' counts were normalized using the "Trimmed Mean of M-values" (TMM) method ²⁷. The fraction of sequences representing transcripts of 55 - 110nt was used for the filtration procedure that included two major steps: i) contamination and ii) non-specific fragments cleaning. The first stage consists of a BLAST-based similarity search of all sequences versus the NCBI non-redundant nucleotide database (as for December 2020). At this step, all fragments that did not match any sequences from species classified in the mustard family (*Brassicaceae*) were removed. Subsequently, the remaining sequences were searched with BLAST using the evaluated set of tRNAs sequences as reference. At each step of data analysis only Linux text processing tools, in-house generated Perl scripts and AWK language commands were used. Exclusively only the best matching (based on cumulative BLAST score), uniquely mapped fragments were assigned to corresponding tRNA genes.

ACKNOWLEDGMENTS

This work was supported by the grant from the National Science Centre (NCN) 2017/27/B/NZ2/00467 and the computational grant number 312 from the Poznań Supercomputing and Networking Centre (PCSS).

AUTHOR CONTRIBUTIONS

M.S. and W.M.K conceived the study and wrote the manuscript. M.S., A.M., and W.M.K. performed all computational analyses.

DATA AVAILABILITY

All the analyses were performed on the genomic sequences of *A. thaliana* available from public databases: the NCBI Assembly Database (<https://www.ncbi.nlm.nih.gov/assembly/>) and the 1001 Genomes Project (<https://1001genomes.org>). The sources of particular sequences and the NCBI Assembly database accession numbers are given in the Supplementary material Table 1. Genomic sequences from the 1001 Genomes Project were downloaded from the <http://mtweb.cs.ucl.ac.uk/mus/www/19genomes/fasta/UNMASKED/> website.

The data used for the analysis of tRNA expression were downloaded from the NCBI SRA database (<https://www.ncbi.nlm.nih.gov/sra>) project number PRJNA505412.

REFERENCES

1. Söll, D. & RajBhandary, U. L. *TRNA: Structure, Biosynthesis, and Function*. (1995).
2. White, R. J. Transcription by RNA polymerase III: more complex than we thought. *Nat. Rev. Genet.* **12**, 459–463 (2011).
3. Bermudez-Santana, C. *et al.* Genomic organization of eukaryotic tRNAs. *BMC Genomics* **11**, 270 (2010).
4. Michaud, M., Cognat, V., Duchêne, A.-M. & Maréchal-Drouard, L. A global picture of tRNA genes in plant genomes. *Plant J.* **66**, 80–93 (2011).
5. Hummel, G. *et al.* Epigenetic silencing of clustered tRNA genes in Arabidopsis. *Nucleic Acids Res.* **48**, 10297–10312 (2020).
6. Initiative, T. A. G. & The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* vol. 408 796–815 (2000).
7. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic

- Sequences. *Methods Mol. Biol.* **1962**, 1–14 (2019).
8. Kapitonov, V. V. & Jurka, J. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8714–8719 (2001).
 9. Du, C., Caronna, J., He, L. & Dooner, H. K. Computational prediction and molecular confirmation of Helitron transposons in the maize genome. *BMC Genomics* **9**, 51 (2008).
 10. Yang, L. & Bennetzen, J. L. Structure-based discovery and description of plant and animal Helitrons. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 12832–12837 (2009).
 11. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10263–10268 (2014).
 12. Hu, K. *et al.* Helitron distribution in Brassicaceae and whole Genome Helitron density as a character for distinguishing plant species. *BMC Bioinformatics* **20**, 354 (2019).
 13. Chan, P. P. & Lowe, T. M. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.* **44**, D184–9 (2016).
 14. Mendiola, M. V., Bernales, I. & de la Cruz, F. Differential roles of the transposon termini in IS91 transposition. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 1922–1926 (1994).
 15. Kitts, P. A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–80 (2016).
 16. Weigel, D. & Mott, R. The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biology* vol. 10 107 (2009).
 17. Jiao, W.-B. & Schneeberger, K. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).
 18. Gan, X. *et al.* Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419–423 (2011).
 19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* vol. 215 403–410 (1990).

20. Seibt, K. M., Schmidt, T. & Heitkam, T. FlexiDot: highly customizable, ambiguity-aware dotplots for visual sequence analyses. *Bioinformatics* **34**, 3575–3577 (2018).
21. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
22. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
23. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
24. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
25. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
26. Leinonen, R., Sugawara, H., Shumway, M. & on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. *Nucleic Acids Research* vol. 39 D19–D21 (2011).
27. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplement.pdf](#)