# Cloning and Sequence Analysis of Ty1-copia-like Retrotransposon RT Gene from AA Wild Peanut (Postprint)

**Authors:** Cai Tiecheng, Liu Junxian, Zhang Chong, Jing Liu, Yang Taiyi, Jiang Jing, He Liangqiong, Han Zhuqiang, Tang Ronghua, Zhuang Weijian, Xiong Faqian

**Date:** 2021-12-19T00:00:00+00:00

## Abstract

This study aimed to clone the RT gene of Ty1-copia retrotransposons, providing a sequence basis for isolating full-length sequences of Ty1-copia retrotransposons in the genus Arachis and investigating their functions. Degenerate primers were designed based on conserved regions of the RT gene, and two wild peanut species of the AA chromosome group, Arachis duranensis, were used as experimental materials. Genomic DNA was amplified by PCR, and target bands were recovered, cloned, and sequenced, followed by bioinformatics analysis of the obtained sequences. The results demonstrated: (1) The target band was approximately 260 bp in size. A total of 41 and 27 RT gene sequences were cloned from the two wild peanut materials, respectively. The 68 sequences exhibited length variations ranging from 256 to 270 bp, AT content ranging from 55.86% to 68.42%, AT/GC ratios ranging from 1.27 to 2.17, and nucleotide sequence similarities ranging from 49.8% to 99.2%, indicating high heterogeneity. (2) The 68 sequences were classified into six families, with families I and IV being the predominant components. (3) Nonsense mutations occurred in 19 of the 68 sequences, with a higher nonsense mutation rate observed in Arachis duranensis (PI219823) than in Arachis duranensis (PI262133). (4) Amino acid sequence similarities ranged from 4.7% to 100%, revealing high heterogeneity. (5) The tertiary protein structures of representative sequences from each family were consistent in overall configuration but showed substantial differences in the numbers of helices, folds, turns, and hydrogen bonds. (6) Conserved motifs among sequences were generally consistent yet exhibited certain variations, demonstrating some heterogeneity; the phylogenetic tree divided the 68 sequences into ten groups, with most sequences clustering in the two major groups A and B. (7) Additionally, some RT gene sequences from AA chromosome group wild peanuts displayed close phylogenetic relationships with RT

gene sequences from other plant species, suggesting that horizontal transfer of Ty1-copia retrotransposons may have occurred between different plant species. This study establishes a foundation for the development and application of novel molecular markers based on Ty1-copia retrotransposons in the genus Arachis.

## Full Text

# Cloning and Sequence Analysis of Reverse Transcriptase Genes of Ty1-copia-like Retrotransposons in Wild Peanut Species with AA Genome

**Cai Tiecheng[1], Liu Junxian[3], Zhang Chong[1], Liu Jing[2], Yang Taiyi[2], Jiang Jing[2], He Liangqiong[2], Han Zhuqiang[2], Tang Ronghua[2], Zhuang Weijian[1], Xiong Faqian[2]***

[1]College of Plant Protection, Fujian Agriculture and Forestry University, Fuzhou 350002, China
[2]Cash Crops Research Institute, Guangxi Academy of Agricultural Sciences, Nanning 530007, China
[3]Sugarcane Research Institute, Guangxi Academy of Agricultural Sciences/Sugarcane Research Center, Chinese Academy of Agricultural Sciences/Key Laboratory of Sugarcane Biotechnology and Genetic Improvement (Guangxi), Ministry of Agriculture and Rural Affairs/Guangxi Key Laboratory of Sugarcane Genetic Improvement, Nanning 530007, China

**Abstract:** This study aimed to clone reverse transcriptase (RT) genes of Ty1-copia-like retrotransposons to provide a sequence foundation for isolating full-length Ty1-copia-like retrotransposons and studying their functions in genus *Arachis*. Degenerate primers were designed based on conserved regions of RT genes. Using two accessions of AA-genome wild peanut *Arachis duranensis* as materials, genomic DNA was amplified via PCR. Target bands were recovered, cloned, and sequenced, followed by bioinformatics analysis of the obtained sequences. The results showed: (1) The target bands were approximately 260 bp in size. Forty-one and twenty-seven RT gene sequences were cloned from the two wild peanut materials, respectively. The length of the 68 sequences ranged from 256 to 270 bp, with AT content ranging from 55.86% to 68.42% and AT/GC ratios from 1.27 to 2.17. Nucleotide sequence similarity ranged from 49.8% to 99.2%, indicating high heterogeneity. (2) The 68 sequences were divided into six families, with families I and IV as the main components. (3) Nineteen sequences contained nonsense mutations, with *Arachis duranensis* (PI219823) showing a higher nonsense mutation rate than *Arachis duranensis* (PI262133). (4) Amino acid sequence similarity ranged from 4.7% to 100%, demonstrating high heterogeneity. (5) Representative sequences from each family showed consistent overall protein tertiary structure configurations but differed substantially in the number of helices, folds, turns, and hydrogen bonds. (6) Conserved motifs were generally consistent across sequences but exhibited some variation, showing a certain degree of heterogeneity. The phylogenetic tree divided the 68 sequences

into ten classes, with most sequences clustered in classes A and B. (7) Some RT gene sequences from AA-genome wild peanuts showed close relationships with RT gene sequences from other plant species, suggesting possible horizontal transmission of Ty1-copia-like retrotransposons between different plant species. This study lays a foundation for developing and applying novel molecular markers based on Ty1-copia-like retrotransposons in genus *Arachis*.

**Keywords:** peanut, Ty1-copia-like retrotransposons, reverse transcriptase, wild species, heterogeneity

## Introduction

Peanut, known as the "longevity nut," is cultivated in 106 countries worldwide. As a major oil crop in China, peanut production is economically significant. Current peanut breeding in China relies primarily on conventional hybridization, which suffers from long cycles, low efficiency, and weak targeting. Molecular breeding can accelerate this process, but simple, practical, and efficient DNA molecular markers are lacking. Traditional molecular marker technologies have difficulty detecting abundant polymorphism in peanut (Xiong et al., 2010; Wang et al., 2010; Xiong et al., 2011). Although recent studies have reported association and linkage analyses using SNP markers in peanut (Zhang et al., 2017; Han et al., 2018; Wang et al., 2018), SSR markers remain the most widely used. However, SSR primer pairs that can detect DNA polymorphism between any two cultivated peanut varieties are still relatively scarce (Xiong et al., 2010; Xiong et al., 2011).

LTR retrotransposons mainly include Ty1-copia and Ty3-gypsy classes (Kumar & Bennetzen, 1999; Feschotte et al., 2002; Bonchev & Parisod, 2013). RT gene sequences from both classes can be amplified and cloned using degenerate PCR technology (Voytas et al., 1992; Kumekawa et al., 1999). The ubiquity, high copy number, high heterogeneity, and insertion site polymorphism of LTR retrotransposons make them ideal for molecular marker development. However, no studies on molecular marker development based on LTR retrotransposons have been reported in peanut. Isolation and identification of LTR retrotransposons are prerequisites for their utilization in molecular markers. Research on peanut LTR retrotransposons is scarce. Nielen et al. isolated the Ty3-gypsy retrotransposon FIDEL and Ty1-copia retrotransposon Matita from peanut and analyzed their characteristics and functions (Nielen et al., 2010, 2012). We previously systematically reviewed the research status and progress of LTR retrotransposon and MITE transposon isolation and application in peanut (Xiong et al., 2017).

This study aimed to clone Ty1-copia-like retrotransposon RT genes from two AA-genome wild peanut accessions of *Arachis duranensis*, analyze their sequence characteristics and diversity, provide a sequence basis for isolating full-length Ty1-copia-like retrotransposons and studying their functions in genus *Arachis*, and lay a foundation for developing and applying novel molecular markers based on Ty1-copia-like retrotransposons.

## Materials and Methods

### 1.1 Plant Materials

Five healthy plants each of two AA-genome wild peanut accessions, *Arachis duranensis* (PI262133) and *Arachis duranensis* (PI219823), were randomly selected from the wild peanut nursery at the Wuming Lijian Research Base of Guangxi Academy of Agricultural Sciences. Top core leaves were collected and pooled from each accession.

### 1.2 Genomic DNA Extraction

High-quality genomic DNA was extracted using a modified CTAB method (Xiong et al., 2019).

### 1.3 PCR Amplification of RT Genes

The upstream primer was RTp1: 5' -ACNGCNTTYYTNCAYGG-3' , and the downstream primer was RTp2: 5'-ARCATRTCRTCNACRTA-3', where R=A/G, Y=C/T, N=A/T/C/G (Kumar et al., 1997). PCR amplification systems, programs, and product separation and detection followed previously reported methods (Xiong et al., 2019).

### 1.4 Recovery, Cloning, and Sequencing of PCR Products

Procedures followed previously reported methods (Yang et al., 2019).

### 1.5 Sequence Analysis of RT Genes

Sequence similarity searches, statistical analysis, sequence logos, protein secondary and tertiary structure prediction, counting of turns and hydrogen bonds in tertiary structures, and conserved motif prediction followed previously reported methods (Yang et al., 2019). MEGA6.0 software was used to construct phylogenetic trees using the neighbor-joining method (No. of differences model) with bootstrap values set to 1000. Information on Ty1-copia-like retrotransposon RT gene sequences from other plant species is shown in Table 1 .

## Results and Analysis

### 2.1 PCR Amplification and Sequencing of RT Genes

PCR amplification of genomic DNA from *Arachis duranensis* (PI262133) and *Arachis duranensis* (PI219823) yielded specific bands of approximately 260 bp in both accessions (Figure 1 [Figure 1: see original paper]). After recovery, cloning, and sequencing, 42 and 38 sequences were initially obtained from PI262133 and PI219823, respectively. Using DNAMAN software to remove identical sequences and NCBI database homology analysis to eliminate non-target sequences, 41 and 27 target sequences were finally obtained from PI262133 and PI219823,

respectively, and designated as AdRT1-X and AdRT2-X (Table 1). Multiple alignment of these RT gene sequences was performed (Figure 2 [Figure 2: see original paper]), and a sequence logo was generated using weblogo to display base conservation at each position (Figure 3 [Figure 3: see original paper]).

### 2.2 Analysis of RT Gene Sequences

Statistical analysis using BioEdit software showed that all sequences ranged from 256 to 270 bp in length, with deletion or insertion mutations present. Among the 41 sequences from *Arachis duranensis* (PI262133), AdRT1-47 was the shortest at 256 bp and AdRT1-11 the longest at 267 bp, with 266 bp sequences accounting for 90.24% of the total (Table 2). A, T, C, and G content ranged from 65-88, 78-94, 29-61, and 49-67, respectively. AT content ranged from 55.86% to 68.42%, with AT/GC ratios from 1.27 to 2.17. Nucleotide sequence similarity ranged from 50% to 99.2%, with the highest similarity (99.2%) between AdRT1-26 and AdRT1-25, and the lowest (50%) between AdRT1-47 and AdRT1-7/AdRT1-29. Amino acid sequence similarity ranged from 4.7% to 100%.

Among the 27 sequences from *Arachis duranensis* (PI219823), AdRT2-24 was the shortest at 257 bp and AdRT2-7 the longest at 270 bp, with 266 bp sequences accounting for 74.07% of the total. A, T, C, and G content ranged from 62-93, 71-93, 32-56, and 66-68, respectively. AT content ranged from 57.20% to 67.29%, with AT/GC ratios from 1.34 to 2.06. Nucleotide sequence similarity ranged from 49.8% to 99.2%, with the highest similarity (99.2%) between AdRT2-3 and AdRT2-4, and the lowest (49.8%) between AdRT2-13 and AdRT2-27. Amino acid sequence similarity ranged from 14.6% to 100%.

### 2.3 Cluster Analysis of RT Gene Nucleotide Sequences

Cluster analysis using MEGA6.0 software (Figure 4 [Figure 4: see original paper]) showed that the 68 sequences were divided into six families. Family I contained 28 sequences (18 from PI262133 and 9 from PI219823). Family III contained only AdRT1-29, which was distantly related to Family I and formed a separate clade, likely due to base substitution. Family IV contained 23 sequences (17 from PI262133 and 6 from PI219823), accounting for 33.82% of total sequences. Family V contained 3 sequences, and Family VI contained 4 sequences. Families V and VI were genetically distant from the other four families, with sequences in these families showing base deletions.

### 2.4 Analysis of RT Gene Amino Acid Sequences

Analysis of amino acid sequences (Figure 5 [Figure 5: see original paper]) revealed that 19 of the 68 sequences contained nonsense mutations. Nine sequences from PI262133 showed nonsense mutations (21.95% of its sequences), while ten sequences from PI219823 showed nonsense mutations (37.04% of its sequences), indicating a higher nonsense mutation rate in PI219823. Specific

nonsense mutations included: AdRT2-21 had 8 mutations at amino acid positions 31, 34, 65, 66, 71, 75, 76, and 77; AdRT2-5 had 7 mutations at positions 51, 58, 74, 75, 76, 77, and 87; AdRT1-47 had 6 mutations at positions 18, 26, 30, 37, 79, and 82; AdRT1-11 had 5 mutations at positions 41, 71, 75, 76, and 77; AdRT1-16 had 5 mutations at positions 54, 58, 80, 81, and 84; AdRT2-10 had 5 mutations at positions 54, 58, 81, 82, and 84. Some sequences showed consecutive nonsense mutations, which can affect retrotransposon transcriptional activity.

### 2.5 Protein Structure Prediction of RT Genes

The 68 Ty1-copia-like retrotransposon RT gene sequences were translated into amino acids. Based on nucleotide clustering results, representative sequences from each family in both accessions were selected for protein secondary and tertiary structure prediction using the online program Phyre2 (Table 3 , Figure 6 [Figure 6: see original paper], Figure 7 [Figure 7: see original paper]). The highest-coverage templates for representative sequences were d1hara, c4rs7R, d1sqwa1, and c5xvnM, with confidence values of 16.5-86.2. Only d1hara belonged to the reverse transcriptase family; other proteins matched retrotranscription protein templates. Secondary structures contained 2-3 $\alpha$-helices and 5 $\beta$-strands. Tertiary structures contained 2-6 turns and 9-30 hydrogen bonds, with one obvious helical structure and two less obvious folded structures.

### 2.6 Conserved Motif Prediction of RT Genes

Conserved motif prediction revealed 11 motifs among the 68 sequences. Fifty-seven sequences simultaneously contained motifs 1, 2, and 3 (83.82% of total), indicating these are the main conserved motifs in AA-genome wild peanut Ty1-copia-like retrotransposon RT genes. This also demonstrates high conservation and similarity between the two AA-genome wild peanut accessions. AdRT2-27 and AdRT2-38 contained motifs 3, 4, and 6, with motifs 4 and 6 occupying the same positions as motifs 1 and 2 but having different amino acid arrangements and compositions. AdRT1-11 contained motifs 1, 3, and 7. AdRT1-16 and AdRT2-10 contained motifs 1 and 5; these two sequences grouped together in the phylogenetic tree. AdRT2-19 contained motifs 1 and 10, AdRT2-24 contained motifs 8, 9, and 10, and AdRT2-21 contained motifs 1 and 7; these three sequences each formed separate groups in the phylogenetic tree. AdRT1-47 contained motifs 8 and 9, AdRT2-5 contained motifs 7, 8, and 11, and AdRT1-23 contained only motif 11; these three sequences grouped together in the phylogenetic tree, with motifs 8 and 11 being short and located in the upstream region. Some motifs occurred at low frequency and were short, suggesting mutations occurred during evolution (Figure 8 [Figure 8: see original paper]).

### 2.7 Phylogenetic Tree Construction of RT Gene Sequences

The phylogenetic tree (Figure 9 [Figure 9: see original paper]) divided all RT gene sequences into ten classes. Most sequences clustered in classes A and B,

indicating high conservation and similarity among AA-genome wild peanut RT gene sequences. Class A contained 20 sequences from PI262133 and 15 from PI219823, showing high similarity to sequences from grape (*Vitis vinifera*, CAN67451.1), mung bean (*Vigna radiata*, AAT90460.1, AAT90479.1), chickpea (*Cicer arietinum*, CAD59770.1), potato (*Solanum tuberosum*, CAA13067.1), American ginseng (*Panax quinquefolius*, ABU94811.1), tea (*Camellia sinensis*, CAJ09751.1), plum (*Prunus salicina*, AGX45518.1), apple (*Malus domestica*, ABS11062.1), and tobacco (*Nicotania tabacum*, AAA03507.1). Class B contained 17 sequences from PI262133 and 5 from PI219823, showing high similarity among these 22 sequences. Class C contained only AdRT1-29, which showed high similarity to sequences from rapeseed (*Brassica napus*, AAA32987.1), plum (*Prunus mume*, ABF57071.1), soybean (*Glycine max*, E47759), quinoa (*Chenopodium quinoa*, AEX61031.1), tomato (*Lycopersicon esculentum*, AAC34611.1), large crabgrass (*Setaria faberi*, AAL36472.1), rice (*Oryza sativa*, AAA33902.1), and maize (*Zea mays*, AAK84849.1). Class D contained three sequences from mung bean, Norway spruce, and red spider lily, which were distantly related to most AA-genome wild peanut and other plant sequences. Class E contained only two sequences: AdRT2-24 from AA-genome wild peanut and S71291 from *Arabidopsis thaliana*, showing the closest relationship. Classes F-J contained 3 and 7 sequences from PI262133 and PI219823, respectively, with no sequences from other species.

## Discussion and Conclusion

The Ty1-copia-like retrotransposon RT genes cloned from the same wild peanut germplasm using the same degenerate primers showed substantial differences and polymorphisms in sequence length, composition, and similarity, indicating high heterogeneity of the same retrotransposon group within a single wild peanut species. Additionally, RT gene sequences from both wild peanut materials were AT-rich, which contributed to high heterogeneity. Nucleotide sequence similarity showed high heterogeneity. Nineteen amino acid sequences contained nonsense mutations, contributing to high heterogeneity. Amino acid sequence similarity showed high heterogeneity. Conserved motifs among the 68 RT gene sequences also showed heterogeneity. Representative sequences from families V and VI differed substantially from other families in helix number, fold number, turn number, and hydrogen bond number, showing high heterogeneity and polymorphism. Overall, RT gene sequences from the two wild peanut materials showed heterogeneity in AT content, nucleotide sequence similarity, amino acid sequence similarity, nonsense mutation rate, conserved motifs, and protein secondary and tertiary structures.

Representative sequences from families I-IV showed similar overall protein tertiary structure configurations but differed substantially in helix number, fold number, turn number, and hydrogen bond number. AdRT1-16 from family V had only two obvious helical structures with significantly fewer turns and hydrogen bonds than other sequences. AdRT1-47 from family VI had far fewer

hydrogen bonds and turns than other sequences, with shorter helical structures. AdRT2-27 had two obvious folded structures and one less obvious folded structure, with fewer turns and hydrogen bonds than other sequences. These differences may affect copy number, transcriptional activity, and transposition efficiency of Ty1-copia-like retrotransposons.

The phylogenetic tree showed that families I and IV were the main families, indicating high conservation and similarity among AA-genome wild peanut Ty1-copia-like retrotransposon RT gene sequences. More complex families with higher sequence similarity are more likely to contain transcriptionally active retrotransposons that may have transposed more recently (Tang et al., 2005). Therefore, families I and IV likely contain transcriptionally active Ty1-copia-like retrotransposons and have a longer evolutionary history.

The phylogenetic analysis showed that 35 AA-genome wild peanut RT gene sequences in class A had high similarity to sequences from grape, mung bean, chickpea, potato, American ginseng, tea, plum, apple, and tobacco. AdRT1-29 in class C showed high similarity to sequences from rapeseed, plum, soybean, quinoa, tomato, large crabgrass, rice, and maize. AdRT2-24 in class E showed high similarity to *Arabidopsis* S71291, suggesting that AA-genome wild peanut Ty1-copia-like retrotransposons may have undergone horizontal transmission with these species. RT gene sequences in classes F-J were all from AA-genome wild peanuts and were most genetically distant from other sequences, suggesting these classes may represent ancient, highly specific sequences unique to *Arachis duranensis* (PI262133) and *Arachis duranensis* (PI219823).

In summary, this study successfully cloned Ty1-copia-like retrotransposon RT gene sequences from AA-genome wild peanuts, which is important for developing LTR retrotransposon-based molecular markers and peanut molecular breeding. This work provides a sequence foundation for future isolation of full-length sequences and studies of transcriptional activity, transposition activity, and function.

## References

BONCHEV G, PARISOD C, 2013. Transposable elements and microevolutionary changes in natural populations[J]. Mol Ecol Resour, 13(5): 765-775.

FESCHOTTE C, JIANG N, WESSLER SR, 2002. Plant transposable elements: where genetics meets genomics[J]. Nat Rev Genet, 3(5): 329-341.

HAN SY, YUAN M, CLEVENGER JP, et al., 2018. A SNP-based linkage map revealed QTLs for resistance to early and late leaf spot diseases in peanut (*Arachis hypogaea* L.)[J]. Front Plant Sci, 9: 1012.

KUMAR A, BENNETZEN JL, 1999. Plant retrotransposons[J]. Annu Rev Genet, 33(1): 479-532.

KUMEKAWA N, OHTSUBO E, OHTSUBO H, 1999. Identification and phylo-

genetic analysis of gypsy-type retrotransposons in the plant kingdom[J]. Genes Genet Syst, 74(6): 299-307.

NIELEN S, CAMPOS-FONSECA F, LEAL-BERTIOLI S, et al., 2010. FIDEL-a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut[J]. Chrom Res, 18(2): 227-246.

NIELEN S, VIDIGAL BS, LEAL-BERTIOLI SCM, et al., 2012. Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the *Arachis* A-B genome divergence[J]. Mol Genet Genom, 287(1): 21-38.

TANG YM, MA YZ, LI LC, et al., 2005. Identification and characterization of reverse transcriptase domain of transcriptionally active retrotransposons in wheat genome[J]. J Integr Plant Biol, 47(5): 604-612.

VOYTAS DF, CUMMINGS MP, KONIECZNY A, et al., 1992. Copia-like retrotransposons are ubiquitous among plants[J]. Proc Natl Acad Sci USA, 89(15): 7124-7128.

WANG LF, ZHOU XJ, REN XP, et al., 2018. A major and stable QTL for bacterial wilt resistance on chromosome B02 identified using a high-density SNP-based genetic linkage map in cultivated peanut Yuanza 9102 derived population[J]. Front Genet, 9: 652.

WANG Q, ZHANG XY, TANG FS, et al., 2010. Construction of genetic linkage map of peanut (*Arachis hypogaea* L.) based on SRAP markers[J]. Chinese journal of oil crop science, 32(3): 374-378.

XIONG FQ, JIANG J, ZHONG RC, et al., 2010. Application of SCoT molecular markers in genus arachis[J]. Acta Agronomica Sinica, 36(12): 2055-2061.

XIONG FQ, LIU JX, HE LQ, et al., 2017. Recent advances on the development and utilization of molecular markers based on LTR retrotransposons peanut(*Arachis hypogaea* L.)[J]. Mol Plant Breed, 15(2): 640-647.

XIONG FQ, LIU JX, LIU J, et al., 2019. Comparative analysis and application of five improved CTAB extraction methods for peanut DNA[J]. Mol Plant Breed, 17(7): 2207-2216.

XIONG FQ, ZHONG RC, HAN ZQ, et al., 2011. Start codon targeted polymorphism for evaluation of functional genetic variation and relationships in cultivated peanut (*Arachis hypogaea* L.) genotypes[J]. Mol Biol Rep, 38(5): 3487-3494.

YANG TY, LIU JX, LIU J, et al., 2019. Cloning and analysis of reverse transcriptase of Ty1-copia-like retrotransposons in *Arachis monticola*[J]. Shandong Agric Sci, 51(9): 9-20, 27.

ZHANG XG, ZHANG JH, He XY, et al., 2017. Genome-wide association study of major agronomic traits related to domestication in peanut[J]. Front Plant Sci, 8: 1611.