Chromosome Research (2010) 18:227–246
DOI 10.1007/s10577-009-9109-z

FIDEL—a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut

Stephan Nielen · Fernando Campos-Fonseca · Soraya Leal-Bertioli · Patricia Guimarães · Guillermo Seijo · Christopher Town · Roberto Arrial · David Bertioli

Received: 24 September 2009 / Accepted: 16 December 2009 / Published online: 2 February 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract In this paper, we describe a Ty3-*gypsy* retrotransposon from allotetraploid peanut (*Arachis hypogaea*) and its putative diploid ancestors *Arachis duranensis* (A-genome) and *Arachis ipaënsis* (B-genome). The consensus sequence is 11,223 bp. The element, named FIDEL (Fairly long Inter-Dispersed Euchromatic LTR retrotransposon), is more frequent in the A- than in the B-genome, with copy numbers of about 3,000 (±950, *A. duranensis*), 820 (±480, *A. ipaënsis*), and 3,900 (±1,500, *A. hypogaea*) per haploid genome. Phylogenetic analysis of reverse transcriptase sequences showed distinct evolution of FIDEL in the ancestor species. Fluorescent in situ hybridization revealed disperse distribution in euchromatin and absence from centromeres, telomeric regions, and the nucleolar organizer region. Using paired sequences from bacterial artificial chromosomes, we showed that elements appear less likely to insert near conserved ancestral genes than near the fast evolving disease resistance gene homologs. Within the Ty3-*gypsy* elements, FIDEL is most closely related with the *Athila*/*Calypso* group of retrovirus-like retrotransposons. Putative transmembrane domains were identified, supporting the presence of a vestigial envelope gene. The results emphasize the importance of FIDEL in the evolution and divergence of different *Arachis* genomes and also may serve as an example of the role of retrotransposons in the evolution of legume genomes in general.

Keywords peanut · *Arachis* · retrotransposon · retrovirus-like · fluorescent *in situ* hybridization

Responsible Editor: Jiming Jiang

Electronic supplementary material The online version of this article (doi:10.1007/s10577-009-9109-z) contains supplementary material, which is available to authorized users.

S. Nielen (✉) · F. Campos-Fonseca · S. Leal-Bertioli · P. Guimarães
Embrapa Recursos Genéticos e Biotecnologia,
70770-917 Brasília, DF, Brazil
e-mail: stephan@cenargen.embrapa.br

F. Campos-Fonseca · D. Bertioli
Universidade de Brasília, Campus Universitário,
70910-900 Brasília, DF, Brazil

G. Seijo
Instituto de Botánica del Nordeste,
Casilla de Correo 209,
3400 Corrientes, Argentina

C. Town
J Craig Venter Institute,
9712 Medical Center Drive,
Rockville, MD 20850, USA

R. Arrial · D. Bertioli
Universidade Católica de Brasília,
Campus II, SGAN 916,
70790-160 Brasília, DF, Brazil

Abbreviations

BAC	Bacterial artificial chromosome
DAPI	4′,6-Diamidino-2-phenylindole
EDTA	Ethylenediaminetetraacetic acid
FISH	Fluorescent in situ hybridization
FITC	Fluorescein isothiocyanate
GISH	Genomic in situ hybridization
LTR	Long terminal repeat
Mya	Million years ago
NBS	Nucleotide-binding site
NOR	Nucleolar organizer region
ORF	Open reading frame
PBS	Primer binding site
PPT	Polypurine tract
RT	Reverse transcriptase
SDS	Sodium dodecyl sulfate
SSC	Standard saline citrate (1× SSC=0.15 M NaCl; 0.015 M Na_3 citrate)
UTR	Untranslated region

Introduction

During evolution, two important factors have decisively influenced the composition and structure of plant genomes: polyploidization and active retroelements. Many of the world's most important cultivated species are polyploid, including the legume species that is the object of this study, peanut. Peanut is a member of the Dalbergioid legumes, a group that diverged from other legume clades about 55 million years ago (Mya; Schrire et al. 2005). Cultivated peanut (*Arachis hypogaea*) is tetraploid with an AB-genome ($2n = 4x = 40$) arising from hybridization of two wild species and spontaneous chromosome duplication. Its origin has been dated only about 3,500 years ago based on archaeological evidence and carbon dating (Hammons 1994). The cytogenetic and molecular data available suggest that the diploid *Arachis duranensis* and *Arachis ipaënsis* (both $2n = 2x = 20$) are the extant species most closely related to the A- and B-genome donors, respectively (Kochert et al. 1996; Seijo et al. 2004, 2007; Burow et al. 2009).

Retroelements, particularly the long terminal repeat (LTR) retrotransposons, constitute the major part of repetitive DNA of plant genomes and contribute substantially to the genome size of species with larger genomes such as maize (2.3 Gb, >75% retrotransposons; SanMiguel and Bennetzen 1998; Schnable et al. 2009). Based on phylogenetic analysis of their reverse transcriptase (RT) sequences and the order of genes in the polyprotein region (pol), LTR retrotransposons can be divided into two major groups, the Ty1-*copia* retrotransposons (*pseudoviridae*) and the Ty3-*gypsy* retrotranspsons (*metaviridae*; Xiong and Eickbush 1990; Kumar and Bennetzen 1999). The Ty3-*gypsy* group contains a subgroup of retrovirus-like elements such as *Athila4* of *Arabidopsis* and *Calypso* of soybean (Wright and Voytas 2002). These elements have an additional open reading frame (ORF) that encodes transmembrane domains and N-terminal signal sequences, which are characteristic of envelope (*env*) genes. A further feature is a primer binding site (PBS) complementary to Asp tRNA. Retrovirus-like elements can be also found, to a lesser extent, in the Ty1-*copia* group, e.g., *SIRE-1* in soybean (Laten et al. 1998). The current picture of the life cycle of LTR retrotransposons has been reviewed by Sabot and Schulman (2006). Regulation of the retrotransposon activity can occur at any step in their life cycle (Feschotte et al. 2002). However, transcriptional silencing via DNA methylation of the promoter region appears to be the most common control mechanism in plants (Liu et al. 2004; Cheng et al. 2006).

In many cases, retrotransposons suffer functional losses through insertions, deletions, and frameshifts (Kumar and Bennetzen 1999 and herein mentioned references). Some elements therefore became non-autonomous, such as *LARD*, *TRIM*, and *Morgane* (Kalendar et al. 2004; Sabot et al. 2006). Retroelements are often found as solo LTRs as shown by Vicient et al. (1999) or Devos et al. (2002). Generation of solo LTRs through illegitimate intra-element recombination is considered a possible mechanism acting against the trend for retrotransposon-driven genome expansion (Shirasu et al. 2000; Bennetzen et al. 2005). A further mechanism is the net deletion of complete elements caused by illegitimate recombination between homologous retrotransposons at different genomic locations (Vitte and Bennetzen 2006).

Previous studies of LTR retrotransposons in the legumes have concentrated on species within the Phaseoloid and Galegoid clades. Several Ty3-*gypsy* elements have been reported, such as LORE1 and

LORE2A in the model legume *Lotus japonicus* (Madsen et al. 2005; Fukai et al. 2008), the more than 22-kb element *Ogre* in *Pisum sativum* (pea; Neumann et al. 2003), or *Diaspora*, which was identified in *Glycine max* and also detected in *Lotus corniculatus* (Yano et al. 2005).

Until now, a complete description of an LTR retrotransposon in peanut, or indeed within any members of the Dalbergioid clade of legumes, has not yet been published. Chavanne et al. (1998) characterized a *gypsy*-like retrotransposon named *Cyclops* in pea and detected hybridization of a fragment of its reverse transcriptase to genomic DNA of various legumes including *A. hypogaea*. Later, Yüksel et al. (2005) screened an *A. hypogaea* bacterial artificial chromosome (BAC) library with "overgos" probes and identified one probe with a high number of hits and 71% sequence similarity to an *Arabidopsis copia* element.

In this paper, we report the isolation and characterization of a new *Athila*-like fairly long interdispersed euchromatic LTR retrotransposon in *Arachis* which we have named FIDEL. We provide a comprehensive characterization of that element and discuss its role in the evolution of the component A- and B-genomes of cultivated peanut and its wild diploid ancestors.

Material and methods

Plant materials

Leaf and root tissue were obtained from *A. hypogaea* cv. Tatu, *A. duranensis* (accession V14167), *Arachis stenosperma* (accession V10309), and *A. ipaënsis* (accession KG30076). All seed was obtained from the Brazilian *Arachis* germplasm collection (Embrapa Recursos Genéticos e Biotecnologia).

DNA extraction

Genomic DNA was extracted from young leaves using the protocol of Grattapaglia and Sederoff (1994) modified by the inclusion of an additional step for precipitation of proteins using 1.2 M NaCl. DNA concentrations were estimated by comparing fluorescence intensities of DNA dilution series with the High DNA Mass Ladder (Invitrogen) after ethidium bromide-stained agarose gel electrophoresis.

Isolation of repetitive sequences

Repetitive sequences were isolated in a dot blot survey of short insert (*Sau*3AI) libraries of genomic DNA from *A. duranensis* and *A. ipaënsis*. Clones with strong hybridization to genomic probes were sequenced and characterized. The repeats Rep-1 and Rep-2 provided the basis for the here described studies. Using the Staden Package software (Staden 1996), sequence assemblies using Rep-1 and Rep-2 together with genome survey sequences from *Arachis* were carried out creating extended pseudocontigs.

Amplification and cloning of reverse transcription sequences

The more conserved regions of the Rep-2-based reverse transcriptase encoding pseudocontig sequence were used to design PCR primers to amplify the entire *rt* sequence. In order to obtain a diverse set of *rt* sequences without bias, a set of three forward (F) and three reverse (R) PCR primers was used:

Rep-RT-F1 5′-AAGGACACACAAGACAGCTC-3′;
Rep-RT-F2 5′-GTACGCACAAGATCCTATTG-3′;
Rep-RT-F3 5′-CTAAATCCAGCCATGAAGG-3′;
Rep-RT-R1 5′-GTCAGCTACAAGGAGATTGC-3′;
Rep-RT-R2 5′-GGAGATGATAGGTGCAGAAG-3′;
Rep-RT-R3 5′-TCACACATCAGTTCAAATGG-3′.

PCR was performed in 50 µl reactions containing 100 ng of *A. duranensis*, *A. ipaënsis*, or *A. hypogaea* cv. Tatu genomic DNA, 0.06 µM of each of the six primers, 0.16 mM dNTPs, 2.5 mM MgCl, 2.5 U Taq polymerase (Invitrogen), and 1× Taq buffer. Thermocycling was 30 cycles of 92°C for 20 s, 45°C for 30 s, and 72°C for 90 s. PCR samples were cloned into pGEM-T Easy vector (Promega) according to the manufacturer's instructions. Inserts were sequenced from 92 colonies. Verified *rt* sequences are deposited in GenBank under accession numbers GU480451-77 (*A. duranensis*), GU480478-503 (*A. hypogaea*), and GU480504-31 (*A. ipaënsis*).

Genome walking

Flanking genomic regions of the two pseudo-contigs based on Rep-1 and Rep-2 were amplified using a modified Genome Walker™ (Clontech Laboratories,

Inc.) strategy. *A. hypogaea* genomic DNA (1.5 µg) was digested with *Pvu*II and ligated with a *Pvu*II adaptor consisting of 5′-ACTCGATTCTCAACCC GAAAGTATAGATCCCA-3′ (long arm) and 5′-Phosphate-TGGGATCTATACTT-H_2N-3′. PCR reactions were performed with the adapter primer AP1 (5′-ACTCGATTCTCAACCCGAAAG-3′) and genome-specific primers directed "outwards" from the known sequences. For Rep-1, primer Rep-1-Out-Right-2 (5′-GTTGCCGGGGATTGTTC-3′) was used to amplify sequences downstream; for Rep-2, the primers Rep-2-Out-Left (5′-AATCATGTCCTCAATTACGC-3′) and Rep-2-Out-Right (5′-TGGTGACTGAAGGAATT GTC-3′) were used. PCR products were cloned and sequenced.

Copy number estimation

Dot blotting

Dilution series of genomic DNA of *A. duranensis* (2*C*= 2.61 pg; Temsch and Greilhuber 2001), *A. ipaënsis* (2*C*=2.8 pg), and *A. hypogaea* (2*C*=5.93 pg; Temsch and Greilhuber 2000) were applied to nylon membranes (Hybond N+, Amersham) together with dilution series of a representative plasmid bearing the retrotransposon sequences of interest (LTR or reverse transcriptase). The plasmid DNA amounts were adjusted to represent 1–100,000 copies per microliter of the retrotransposon sequence in 500 ng of genomic DNA. The same retrotransposon sequence, which was used for calibration purposes, was used as probe: the RT clone Ah-FIDEL-9 with similarity to the other *rt* sequences of not less than 87% or the LTR clone Rep-1. Probes were labeled with digoxigenin-dUTP using the random-primed DIG DNA labeling kit (Roche) and hybridized to the filters at a concentration of 20 ng/ml. Post-hybridization washes were at 87% stringency (0.1× SSC, 60°C) for the RT probe and 82% stringency (0.25× SSC, 60°C) for the LTR probe. Chemiluminescent detection was carried out using the DIG Luminescent Detection Kit (Roche) following essentially the manufacturer's instructions. After incubation with the luminescent substrate CSPD, the filters were exposed to BioMax films (Kodak) for various times (between 10 min and 3 h). Those chemiluminographs where the signals did not fully saturate the film were selected for analysis. Films were digitalized and the gray values were measured using the Multi Gauge V3.0 software (Fujifilm). After verification of the linear or logarithmic relationship between the applied DNA amounts and the signal intensities, the data for signal intensity of the plant DNA could be compared to those of the calibration series to estimate the actual copy number of the element sequence in the plant genome.

Using BAC end sequences

BAC end sequence databases of *A. duranensis* and *A. ipaënsis* with sequence reads of on average ~700 bp were used to estimate copy numbers (*N*) of FIDEL. In Blastn searches of 500 bp queries derived from all regions of FIDEL, the number of hits better than 1e−100 (N_{hits}) was determined and applied in a formula introduced by Zhang and Wessler (2004): $N = (1/\text{genome coverage}) \times N_{hits}/2[1 + (L_{dr} - L_{eq})/(L_{dr} + L_{eq})]$, where L_{eq} is the effective query length and L_{dr} the average length of database reads. The formula considers the probability of the presence of two kinds of hits, full-length hits of the entire query and partial hits, which are truncated due to cloning and only contain part of the query.

Selection and sequencing of BAC clones

Two randomly chosen 96-well plates from BAC libraries of the *A. duranensis* and *A. ipaënsis* genomes, which represent 7.4*x* A-genome and 5.3*x* B-genome equivalents with average insert sizes of 100 and 110 kb, respectively (Guimarães et al. 2008), were screened for the presence of the Rep-2-based *rt* sequence and the LTR sequence Rep-1. Primers and PCR conditions for amplification of the *rt* sequence were as described earlier. For amplification of the LTR sequence, the primers Rep-1-Fw-1 (5′-TGAGCATCAGTGGA TAGGA-3′) and Rep-1-Rv-3 (5′-TACCATCCA TCCTTCCAGT-3′) were used. The program for amplification was 30 cycles of 92°C for 20 s, 58°C for 30 s, and 72°C for 90 s. The BAC clones positive for both *rt* and LTR amplification were selected for Southern blot analysis of *Hin*dIII-digested BAC DNA. Filters were hybridized with a DIG-labeled RT probe and rehybridized with a DIG-labeled LTR probe. Labeling, hybridization, and detection were done as described for dot blotting. The stringency of the post-hybridization washes was adjusted to 87%. One BAC clone was

shotgun-sequenced using ABI fluorescently labeled Sanger dideoxysequencing by random fragmentation, subcloning, and sequencing. Sequence assembly was done using Cap3 (Huang and Madan 1999).

Chromosome preparation and fluorescent *in situ* hybridization

Root tips from two diploid A-genome species (*A. duranensis* and *A. stenosperma*) and the tetraploid AB-genome species *A. hypogaea* were collected from young plants in the greenhouse or from rooted leaf petioles cultivated in humid Petri dishes at an average temperature of 25°C. The root tips were treated for 3 h with 8-hydroxyquinoline (2 mM) prior to fixation in 3:1 ethanol/acetic acid. Fixed root tips were incubated in a mixture of 2% cellulase Onozuka R10 (Merck) and 20% pectinase (Sigma) in enzyme buffer (4 mM citric acid, 6 mM sodium citrate, pH 4.8) for 30 min at 37°C. Chromosomes were spread in 45% acetic acid on a microscopic slide using mechanical force (Maluszynska and Heslop-Harrison 1993). Probes were labeled either with digoxigenin-11-dUTP or biotin-11-dUTP (Roche) by random primed labeling. Probes used were pTa71 containing the 18S–5.8S–25S rDNA genes from *Triticum aestivum* (Gerlach and Bedbrook 1979), pTA794, which has the complete 5S rDNA gene of *T. aestivum* (Gerlach and Dyer 1980) and pAh-E12, which is part of the FIDEL LTR sequence. Pretreatment, hybridization, washing, and detection procedures essentially followed protocols of Schwarzacher and Heslop-Harrison (2000). Denatured DNA probes (100 ng/ml) were mixed in a hybridization solution containing 50% (*v*/*v*) formamide, 10% (*w*/*v*) dextran sulfate, 0.125 mM EDTA, 0.125% (*w*/*v*) SDS, and 1 μg of salmon sperm DNA. For genomic in situ hybridization (GISH), the probe mix consisted of Dig-labeled genomic DNA of A. *duranensis* and biotin-labeled genomic DNA of *A. ipaënsis*, both at 65 ng/μl. The chromosomes and DNA were denatured together at 81.5°C for 10 min before hybridization overnight at 37°C. Post-hybridization washes were carried out at 83% stringency. Hybridization sites were detected by sheep anti-digoxigenin conjugated to fluorescein isothiocyanate (anti-dig FITC; green fluorescence) and Cy3-streptavidin (red fluorescence). FITC signals were amplified using anti-sheep antibody conjugated to fluorescein (Vector Laboratories). Chromosomes were counterstained with DAPI (4′-6-diamidino-2-phenylindole; blue fluorescence). Slides were observed under a Zeiss Axiophot epifluorescence microscope using appropriate filters. Microphotographs were taken on Kodak Ultra ISO 400 color print film, digitalized, and edited using only Adobe Photoshop functions (brightness, cropping, overlay), which affect the whole image equally.

Sequence and phylogenetic analysis

For phylogenetic studies, 132 FIDEL *rt* sequences derived from all three species were used: (a) 27 (*A. duranensis*), 23 (*A. ipaënsis*), and 25 (*A. hypogaea*) PCR-derived *rt* sequences; (b) 26 (*A. duranensis*) and 29 (*A. ipaënsis*) homologous BAC end-derived sequences; (c) *rt* sequences of the two FIDEL copies *Ad*-185P1FIDEL1 and *Ad*-185P1FIDEL2 isolated from *A. duranensis* BAC clone *ADUR*185P1 (sequences in Electronic supplementary materials (ESM) Supplement S1). The evolutionary divergence between these sequences was calculated using the Jukes–Cantor method in MEGA4 (Jukes and Cantor 1969; Tamura et al. 2007). The RT protein sequences from *Athila/Cyclops*-like elements were from supplementary information in Wright and Voytas (2002). For multiple alignments of amino acid and DNA sequences, the program ClustalW (Thompson et al. 1994) was used. The results of alignments were used for constructing neighbor-joining trees with bootstrap analysis of 1,000 replicates using MEGA4.

Using the LTR divergence method, the age of *Ad*-185P1FIDEL1 and *Ad*-185P1FIDEL2 was estimated. A molecular clock employing a substitution rate of 1.3×10^{-8} per site per year was applied in accordance with Ma and Bennetzen 2004. Ratios of transitions (Ts) to transversions (Tv; Ts/Tv) between *rt* sequences and between the LTRs from the complete elements were established using MEGA4. A search for low-complexity regions in the total FIDEL DNA sequence was implemented using the software tool "Complexity", applying a sliding window of 10 bp (http://www.mgs.bionet.nsc.ru/mgs/programs/low_complexity/; Orlov and Potapov 2004). The nucleotide identity between *Ad*-185P1FIDEL1 and *Ad*-185P1FIDEL2 was determined using the Staden Package software. A cutoff of 0.25 nucleotide identity was used because this is the random identity between DNA sequences.

Analysis of patterns of insertion of FIDEL using paired BAC end sequences

The 41,856 sequences used for analysis are from the BAC library of *A. duranensis* V14167 genomic DNA (GenBank accession nos. FI281689–FI321525). The methodology, which allows inferences of genome structure to be made, uses pairs of "forward" and "reverse" sequences from the ends of cloned DNA fragments, with an average size of about 100 kb. For instance, if FIDEL tends to cluster with a particular gene family, then if one sequence of a pair is FIDEL, then the other sequence of the same pair is more likely to be a member of that gene family than would be expected by chance. We developed a procedure implemented by a Perl script to observe these types of deviations of expected and observed frequencies.

Sequences were organized into four groups or "bins" as delineated in Fig. 1 using a Perl script. Sequence similarities were considered significant at *E* value 1×10^{-60} or below.

To test if FIDEL is clustered in the genome, the overall proportion of sequences with significant similarities to FIDEL was determined. From this, the expected number of sequences in Bin0 was calculated. A significant deviation from this number may indicate a non-random distribution of FIDEL in the genome.

To test for correlations between the distribution of FIDEL and genes, the sequences in Bin2 and Bin3 were used. Bin2 and Bin3 sequences do not have similarities to FIDEL, but while we know sequences in Bin2 are close to a FIDEL, sequences in Bin3 may or may not be. According to a null hypothesis of random distribution of FIDEL, the frequencies of any gene class in Bin2 would be the same as the overall frequency (in Bin2 plus Bin3). We tested this using Blastx similarity searches of Bin2 and Bin3 with four sets of proteins encoded by conserved Unique genes from *Arabidopsis*, conserved Unique genes from rice (Armisén et al. 2008), paralogous gene families in *Arabidopsis* (minus NBS encoding genes; Wortman et al. 2003), and NBS encoding resistance gene homologs from *Arabidopsis*. All databases were screened for contaminating repetitive DNA sequences using Blast similarity searches against repeat databases at GIRI (http://www.girinst.org/). The first three gene sets were chosen because *Arachis* homologs of these genes will represent conserved ancestral genes which, on average, are likely to be slow-evolving. The last gene set, NBS encoding genes, was chosen because it is well established that this gene family suffers rapid birth and death and can be considered to be fast-evolving (Michelmore and Meyers 1998). Blastx sequence similarities were considered significant at *E* value 1×10^{-10} or below.

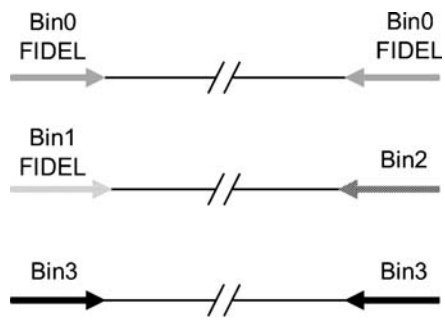

Fig. 1 Illustration of the characteristics of the four "bins" used to analyze the pattern of FIDEL insertion. *Lines* represent BAC clones with a pair of BAC end sequences each: Bin0 sequences (*gray*) have similarity to FIDLE and are located on the same BAC clone; Bin2 sequences (*black*, *striped*) have no similarity to FIDEL, but are paired to Bin1 sequences (*light gray*) that do; Bin3 sequences (*black*) have no similarity to FIDEL and are paired to sequences of the same group. *Line interruption* points out the different scales between the BAC total sequence (~100–110 kbp) and the BAC end sequences (700 bp on average)

The overall gene homolog frequencies in non-FIDEL sequences (Bin2 plus Bin3) were assumed to be a good estimate of the true gene frequencies in this type of sequence. We considered this a reasonable assumption because the number of sequences in Bin2 plus Bin3 is large. The significance of deviations in between observed and expected values was tested using the two-tailed, cumulative binomial probability function (the binomial distribution is the standard probability function that describes the probability of x successes in n trials with a given probability of success, p).

Results

Isolation and structure of FIDEL

Two repetitive sequences, Rep-1 and Rep-2, that were identified by dot blot hybridizations of genomic short insert libraries provided the basis for the isolation of the retrotransposon FIDEL. The sequence Rep-2 from the *A. ipaënsis* library had high similarity to the *rt* region of the Ty3-*gypsy* retrotransposon *Athila* as revealed by a Blastx search (Altschul et al. 1990).

Secondly, the sequence Rep-1 from the *A. duranensis* library was AT-rich and did not show any similarities to genic sequences in GenBank. Using a genome walker strategy to extend the 3′-end of Rep-1, three large inserts of 1,000, 1,200, and 2,000 bp were obtained, which showed at their 3′-ends significant similarity (5e−32) to Gag-Pol of *Athila*. Additionally, the new sequences resembled at their 5′-end the primer binding site of the retrovirus-like elements *Athila4* and *Calypso* (Nielen et al. 2009). Since these results suggested that Rep-1 and Rep-2 could be parts of the LTR and the *rt* region of the same Ty3-*gypsy* retrotransposon, we aimed to obtain a complete element using the recently generated BAC libraries of the *A. duranensis* and *A. ipaënsis* genomes. PCR screening of isolated BAC DNA from two randomly selected plates, one from each library, resulted in selection of six clones, which produced amplicons of both regions *rt* and LTR. Southern hybridizations revealed the presence of one *rt* band and multiple LTR bands (up to six) in four of these clones. One clone from the *A. duranensis* library with six LTR bands (*ADUR*185P1) was subcloned and sequenced. A total of 877 sequence reads was used for sequence assembly creating six contigs, one of 84 kb and the others all less than 6,000 bp. The analysis of the 84-kb sequence (GenBank accession no. GU480450, ESM Supplement S2) revealed two complete copies of the retroelement. Both copies had intact LTRs with flanking target site duplications (TSDs). Based on these sequences and using BAC-end- and genome survey sequences from GenBank, a consensus sequence of the retrotransposon has been established and analyzed (ESM Supplement S3). Its structure is depicted in Fig. 2a. The total length of the element is 11,223 bp. The two LTRs are 1,531 and 1,553 in length and have very similar sequences (0.023 nucleotide substitutions per site). A 5-bp target site duplication marks the beginning of the 5′- and end of the 3′LTRs. The element contains two ORFs encoding Gag-Pol and two untranslated regions (UTRs). Details on positions and characteristics of the retrotransposon components are given in Table 1. The protein domains of Gag-Pol reveal the motifs, which have been shown earlier to be characteristic of the *Athila*/*Calypso* group of retrovirus-like elements (Xiong and Eickbush 1990; Wright and Voytas 2002).

The 5′- and the 3′UTRs were of lower complexity than the coding regions and the LTR sequences and variable in size and nucleotide sequence between the two isolated FIDEL elements as shown in Fig. 2b, c. In the retrovirus-like elements, a further ORF encoding *env*-like sequences, such as membrane spanning sequences and N-terminal signal sequences, are often found in the region between *pol* and the 3′LTR. The 3′ UTR of FIDEL, however, is scattered with stop codons, and only in the BAC end sequence databases could sequences presenting ORFs >50 bp be identified. In detail, in a Blast search against the 3′UTR, from 3757 *ADUR* (*A. duranensis*) sequences, 141 were selected with a minimum *E* value of 1e−60. Thirty-nine sequences presented ORFs greater than 50 bp. Four of these were located at the end of the 3′ UTR and encoded a predicted transmembrane sequence as calculated using the transmembrane helix prediction program Membrane Protein Explorer MPEx (Snider et al. 2009; Fig. 2d). The total sequence of BAC clone *ADUR*185P1 was analyzed in order to characterize the integration loci of the two FIDEL elements. Using a Blastx search, five additional protein sequences with similarities to retrotransposon proteins, in particular *Athila*-like Ty3-*gypsy* elements, were identified (Genbank no. GU480450, ESM Supplement S4). However, none of them had similarities to FIDEL at the nucleotide level. The first FIDEL is inserted into the *gag-pol* region of an incomplete *gypsy*-like retrotransposon with similarities to *Athila* ORF1 (1e−30) that is different from FIDEL and oriented in the opposite direction. Upstream of the 5′ LTR of the second FIDEL is a sequence with a translated product similar to a retrotransposon GAG protein from *Asparagus* (5e−27) including the LCDLGS protease motif of *Athila*5-1.

Copy number estimation

Regression curves of hybridization signals of dot blot hybridizations using an RT subclone as probe against dilution series (two each) of itself and of genomic DNA from the three *Arachis* species were used for calculating the copy numbers. Only values that led to regression coefficients $R^2 > 0.95$ were considered. Figure 3 shows, as an example, the dot blot experiment used to determine the copy number in *A. ipaënsis*. The calculated copy numbers are average values including standard deviations resulting from the different possible combinations of regression curves of genomic and plasmid DNA. The representative reverse transcriptase

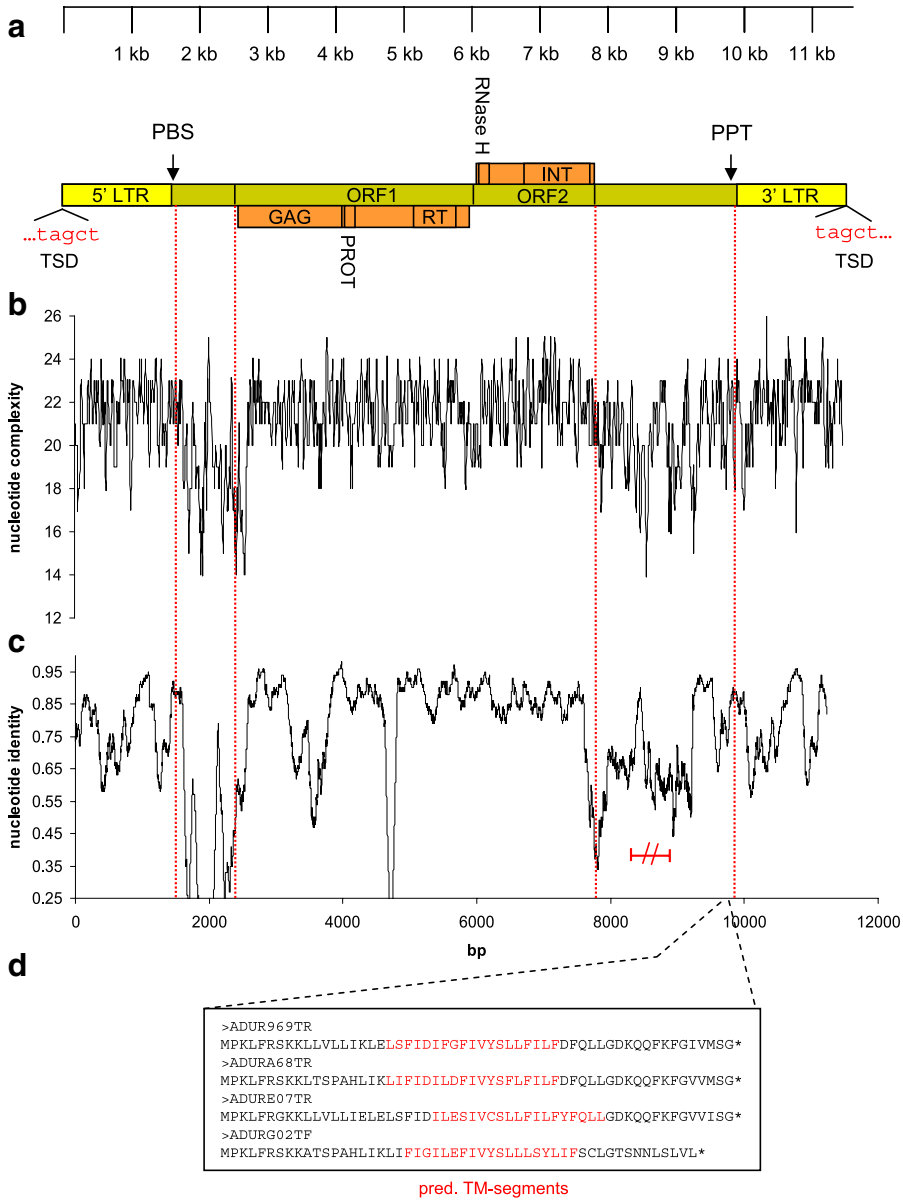

Fig. 2 Structure, composition, and characteristics of the Ty3-*gypsy* retrotransposon FIDEL, isolated from *Arachis* (FIDEL consensus sequence shown in ESM Supplement S2; GenBank accession no. xxxx). **a** Structural organization: *TSD* target site duplication, *LTR* long terminal repeat, *PBS* primer binding site, *PPT* polypurine tract used for synthesis of the second (+) DNA strand, *GAG* structural proteins for virion core, *Protease* protease for cleavage of primary translation products, *RT* reverse transcriptase, *RNase H* ribonuclease, *Integrase* endonuclease for integration in host genome. **b** Analysis of DNA complexity. Two regions corresponding to the pre-*pol*- and post-*pol* region of FIDEL (marked by *vertical red dotted lines*) are characterized by lower complexity. **c** Analysis of nucleotide identity between the two fully sequenced elements FIDEL-1 and FIDEL-2 as a measure of variability between the FIDEL copies. The *red interrupted bar* indicates a region between 8,000 and 9,000 bp in FIDEL-2 with a high number of insertions. These insertions were deleted to facilitate alignment and nucleotide comparison. **d** Amino acid sequences derived from ORFs >50 bp of four *A. duranensis* BAC end sequences highly similar to the end of the post-*pol* region (approximately 100 bp before 3′LTR). *Red* are 19 AA TM segments as predicted using the program Membrane Protein Explorer MPEx

Table 1 Structure of FIDEL indicating position and description of its regions, protein domains, PBS, PPT

Region/specific site/protein domain	Position (bp)	Description
5′LTR	1–1531	
PBS	1541	tggcgccgttgccggggat: complementary to 3′-end of aspartic tRNA from *A. thaliana*
5′UTR	1534–2543	1,009 bp
ORF1	2544–5603	1019 aa polyprotein
GAG		Conserved domain (position 90–176 aa); zinc finger motif $CX_2CX_{3-4}HX_4C$ (characteristic for C-terminal ends of GAG of numerous pararetroviruses and retrotransposons including *Cyclops-2* (Chavanne et al. 1998) not identified
Protease		LCDLGA motif (aa 533–539): putative active site of an aspartic acid protease (Wright and Voytas 2002)
RT		199 aa conserved domain, starts at aa position 808
ORF2	5877–7593	620 aa polyprotein
RNaseH		Conserved domain at position 51–166
Integrase		Conserved domain at position 261–473, including well-conserved zinc finger motif $HX_6HX_{30}CX_2C$
3′UTR	7594–9678	2,084 bp
PPT	9646	ttgggg: conserved core of PPT; also identified in *Athila*-like elements (Wright and Voytas 2002)
3′LTR	9670–11223	

subclone Ah-FIDEL-9 was estimated to be present in *A. ipaënsis* (2*C*=2.8 pg) in about 820 (±480) copies per haploid genome (0.68±0.39% of the genome with regard to the complete FIDEL element), in *A. duranensis* (2*C*=2.61 pg) with about 3,000 (±950) copies per haploid genome (2.65±0.84% of the genome regarding the complete FIDEL element), and in *A. hypogaea* (2*C*=5.93 pg) with about 3,900 (±1500) copies per haploid genome (1.53±0.59% of the genome regarding the complete FIDEL element). Estimation of the LTR copy number was also done with *A. duranensis* where about 6,600 (±500) copies were estimated for the 1,500-bp fragment Rep-1 per haploid genome, which is in accordance with the estimated number of complete elements in this species.

Data derived from BAC end sequences and from PCR screening of BAC plates were consulted to test the consistency with results from the dot blot. The copy numbers resulting from calculations based on the present large *A. duranensis* database (41,856 sequences) and those based on the previous small database (with 3,758 sequences) were in the same range. Thus, the use of only a small *A. ipaënsis* database (3,823 sequences) was justified. The calculated copy numbers varied along the element, and the highest values were found in the putatively more conserved regions ORF1 and ORF2. Depending on the position of the query, between 5,000 and maximal 16,000 copies were determined in these regions in *A. duranensis*, whereas in *A. ipaënsis*, the range was between 1,200 and 10,000 copies. Since the BAC library was produced by cloning *Hin*dIII fragments from partial digestion and several *Hin*dIII restriction sites are present in the ORF regions, the proportion of FIDEL sequences in the BAC ends is likely to be overrepresented. Even if the databases are not representative for the genomes, the numbers are in the same order of magnitude as the dot blot data, and moreover, the overall discovery that the copy number of FIDEL in *A. duranensis* is well above the number in *A. ipaënsis* is reflected in the above data. Relative abundance of the element was estimated by PCR screening of one randomly selected plate of BAC clones from each genome using three pairs of *rt* gene-specific primers. For the *A. duranensis* plate, 36 *rt* positive clones were identified, whereas in the *A. ipaënsis* plate, only ten positive clone were found, a ratio of 3.6:1. Assuming that a 96-well BAC plate with an average insert size of 100 kb represents about 0.7% of the genome and that the FIDEL copies are

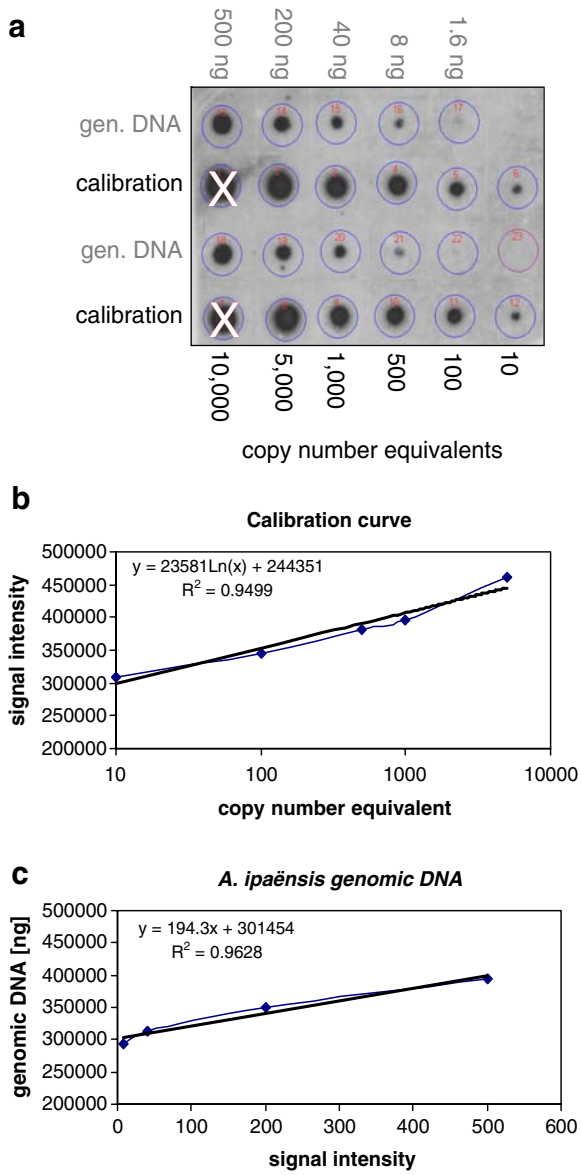

Fig. 3 Estimation of copy number. **a** Chemilumigraph of a dot blot with dilution series of RT subclone Ah-FIDEL-9 for calibration (*lanes 2* and *4*; *dots* represent 10–10,000 copy number equivalents in 500 ng *A. ipaënsis* genomic DNA) and of genomic *A. ipaënsis* DNA (*lanes 1* and *3*, 1.6–500 ng), hybridized with Dig-labeled Ah-FIDEL-9. **b** Calibration curve based on signal intensities of Ah-FIDEL-9 dots. **c** Curve based on hybridization signals of *A. ipaënsis* dots (on the same filter). The signal intensities used for the graphs are average values from the individual dots of two dilution series each. *White crosses* indicate elimination of the data due to over saturation of the film

evenly distributed in the library, the extrapolated copy numbers of FIDEL would amount to 4,720 in *A. duranensis* and 1,370 in *A. ipaënsis*. Both the experimentally determined ratio and the extrapolated copy numbers are in the range of the dot blot estimations. Considering the fact that one BAC clone can contain more than one copy of the FIDEL element, as was the case for the sequenced clone *ADUR*185P1, the extrapolated copy numbers may be underestimates.

Cytogenetic analysis and chromosomal distribution of FIDEL

We used GISH to visualize the principal distribution of repetitive sequences in the A- and B-genomes, allowing a comparison with the localization of FIDEL. At the same time, the amphidiploid nature of an interspecific hybrid between *A. ipaënsis and A. duranensis* that has been previously generated (Fávero et al. 2006; Fonceka et al. 2009) could be confirmed. Metaphase spreads were probed simultaneously with differently labeled genomic DNAs of both species without addition of blocking DNA (Fig. 4a–d). After DAPI staining, 20 chromosomes featured strong centromeric DAPI bands, which are characteristic for the A-genome but do not appear in the B-genome chromosomes of *A. ipaënsis* (Raina and Mukai 1999; Seijo et al. 2004). The GISH pattern was consistent with the DAPI results and allowed differentiation of the two genomes. It is obvious that the probes did not hybridize exclusively but predominantly to the chromosomes of their respective genome. Also, the fluorescence was not uniformly distributed along the total length of chromosomes, but mainly concentrated in the interstitial chromosome regions with absence at centromeres and chromosome ends. Additionally, the smallest pair of chromosomes with the most pronounced centromeric DAPI band, chromosome pair A9, exhibits only very weak fluorescence. The metaphase shown here exhibits a variation in the number of active 18S–5.8S–25S rDNA sites: two pairs of A-genome chromosomes reveal extended secondary constrictions, which is in contrast to the typical karyotype of *A. hypogaea* with one pair of satellite chromosomes (A10) only. The additional SAT chromosome pair is A2, which has an 18S–5.8S–25S rDNA site next to the centromere (see Fig. 4e). To determine the chromosomal distribution

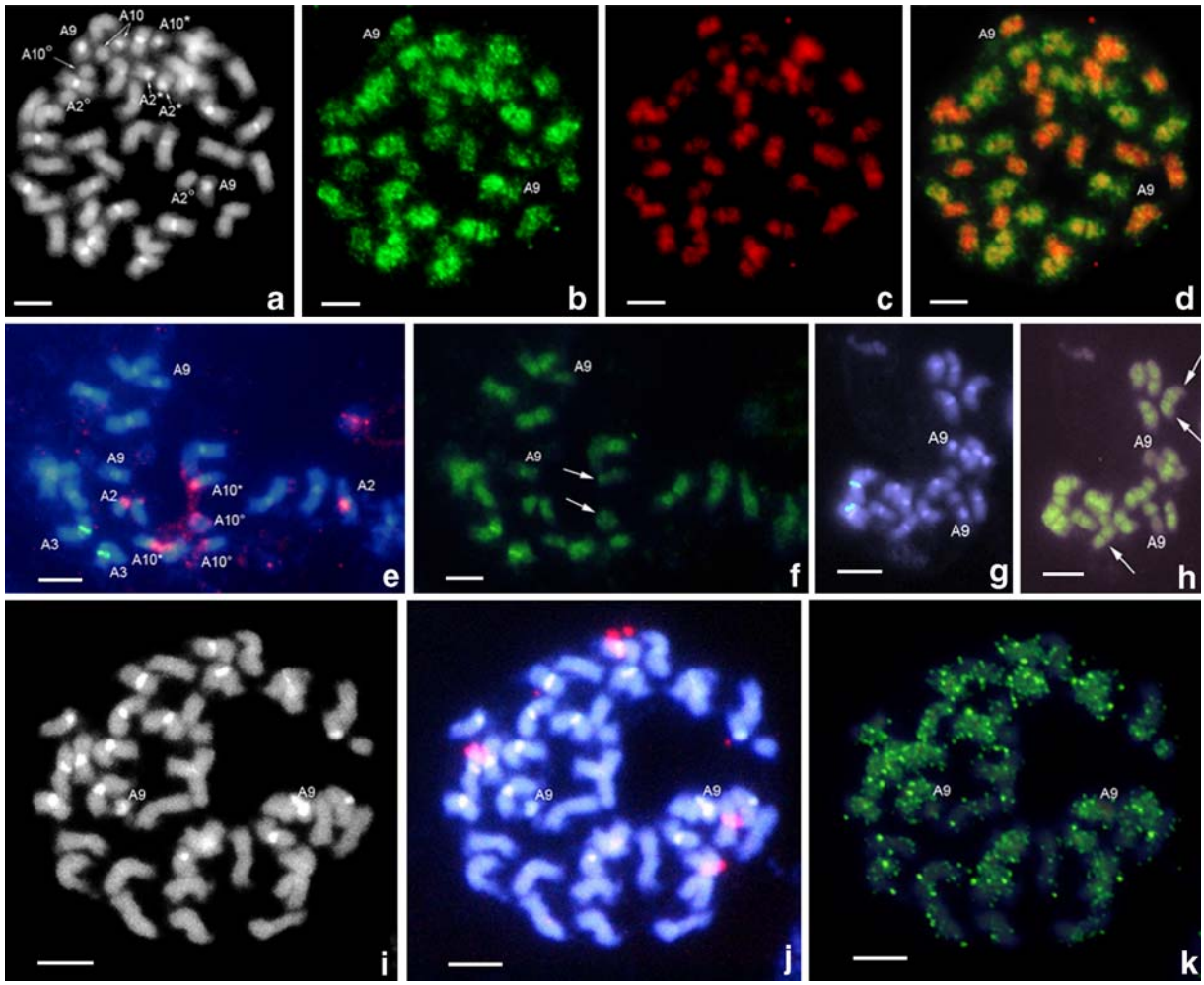

Fig. 4 Cytogenetic analysis using GISH and FISH.**a–d** Genomic in situ hybridization of a metaphase preparation of an amphidiploid hybrid of *A. duranensis* × A. *ipaënsis* probed with genomic DNA of both parents. **a** DAPI counterstain showing 40 chromosomes, 20 of which with centromeric bands typical for *Arachis* A-genome chromosomes. **b** *A. duranensis* genomic probe. **c** *A. ipaënsis* genomic probe. **d** Superposition of **b** and **c**. Note chromosome pair A9 without significant hybridization signal. In contrast to the typical *A. hypogaea* karyotype, two A-genome chromosome pairs (A10 and A2) exhibit (predominantly) extended secondary constrictions (see text). The short arm and the proximal segment of the long arm are indicated by an *asterisk*, and the separated satellites are marked by a *degree sign*. **e–k** FISH using rDNA and LTR sequences of the FIDEL element as probes. **e**, **f** Chromosome spreads of *A. duranensis*. **e** Localization of one sub-centromeric 5S rDNA sites (*green*) and two 18S–5.8S–25S sites (*red*). Note the distance from satellites in A10 from the proximal arm segment (chromosome annotation according to Seijo et al. 2004). **f** Rehybridization with a Dig-labeled LTR-probe. Chromosome A10a bearing a 18S–5.8S–25S site shows only weak hybridization (*arrows*). **g**, **h** FISH of chromosome spread of *A. stenosperma* (A-genome). **g** Localization of one sub-centromeric 5S rDNA sites (*green*). **h** Rehybridization with Dig-labeled LTR probe. Note the lack of or less intensive hybridization to centromeres and telomeric regions (*arrows*) as well as to chromosome pair A9. **i–k** FISH of chromosome spread of *A. hypogaea*. **i** DAPI staining enables discrimination between A-genome (DAPI bands) and B-genome chromosomes. **j** Localization of two 5S rDNA sites (*red*). **k** Dig-labeled LTR probe (*green*) showing hybridization preferentially to A-genome chromosomes. B-genome chromosomes with 5S rDNA site exhibiting stronger signals as compared to the others. *Scale bar*, 5 μm

of the FIDEL element, metaphase spreads of *A. duranensis*, *A. stenosperma* (both A-genome), and *A. hypogaea* were hybridized with Dig-labeled LTR and reverse transcriptase probes. In order to enable identification of more chromosome pairs, spreads were initially hybridized with heterologous 5S and/or 18S–5.8S–25S rDNA probes and subsequently rehybridized with the retrotransposon probes. In *A.*

duranensis, one paracentromeric 5S rDNA site and two 18S–5.8S–25S rDNA were detected (Fig. 4e). Rehybridization of this preparation with the LTR probe showed its dispersed distribution in euchromatic regions of chromosomes and absence from certain heterochromatic regions, such as the centromere, telomeric region, and nucleolar organizer region, which is indicated by arrows in Fig. 4f. The fluorescence signals on chromosome pair A9 were significantly lower, thereby resembling the GISH results. In Fig. 4h, which shows hybridization to an *A. stenosperma* metaphase, the absence of the element from centromeres and chromosome A9 is even more apparent. Also, FIDEL signals are absent (see arrows) or less intense on telomeric regions. When hybridized to metaphase spreads of *A. hypogaea* (Fig. 4i–k), the probe hybridized preferably to the A-genome chromosomes (which becomes apparent by comparison with the chromosomes bearing the characteristic DAPI band). However, faint signals can be also found on B-genome chromosomes, and one pair of chromosomes bearing the 5S rDNA site exhibited relative strong signals. Among the A-genome chromosomes, the pair A9 again has only weak signals. Preferential hybridization to the A-genome was also detected when using a clone representing the reverse transcriptase as a probe (unpublished observation).

Phylogenetic analysis

Considering the significant differences in copy numbers and chromosomal distribution of FIDEL in *A. duranensis* and *A. ipaënsis*, we were interested in elucidating the molecular evolution of the element in both parental species and in *A. hypogaea*. Therefore, a phylogenetic analysis was made of 132 FIDEL *rt* sequences derived from all three species. The average genetic distance values (substitutions per site) between the *A. duranensis* sequences were 0.095 (±0.036), between the *A. ipaënsis* sequences 0.104 (0.029), and between *A. ipaënsis* and *A. duranensis* sequences 0.110 (±0.018). Thus, the distances of the individual FIDEL elements between both species are in the same range as within the species. Similar results were achieved when comparing *A. duranensis* with *A. hypogaea* sequences (0.106±0.043), whereas *A. ipaënsis* sequences appeared to be slightly more distant from *A. hypogaea* (0.144±0.031). The phylogenetic tree (Fig. 5) reveals a more detailed picture. The *A. duranensis rt* sequences (labeled with green dots) emerge mainly in two distinct clades, which are supported by bootstrap values of 90% and 93%, respectively. They are separated from the majority of the *A. ipaënsis* sequences (red dots) which form one bootstrap-supported (89%) clade and a number of smaller subgroups. The tree reveals no clades in which *A. hypogaea* sequences predominate. This was expected due to the evolutionary short time after the origin of the tetraploid species. Instead, the *A. hypogaea* sequences (yellow dots) are distributed within the *A. duranensis* and *A. ipaënsis* clades and subgroups with a majority of 68% in the *A. duranensis* clades, corroborating copy number and fluorescent in situ hybridization (FISH) results.

The two *A. duranensis* whole BAC-derived sequences (dark green dots in Fig. 5) group into the two different *A. duranensis* clades and have estimated ages of 1.7 (±0.3) Mya and 2.8 (±0.4) Mya. The age was estimated based on LTR divergence, which was 0.022 (±0.004) nucleotide substitutions per site for *Ad*-185P1FIDEL1 and 0.036 (±0.005) substitutions per site for *Ad-185P1FIDEL2*.

The ratio between Ts and Tv (Ts/Tv) for the LTRs of *Ad*-185P1FIDEL1 was 1.6:1 and for those of *Ad*-185P1FIDEL2 2.9:1. This is of interest for the question of the potential activity of FIDEL since high C to T transition rates are a result of extensive cytosine-5-methylation. Ts/Tv values >1.5:1 indicate that the elements are in an epigenetically silenced state, which does not allow activation (SanMiguel et al. 1998; Ma and Bennetzen 2004; Vitte and Bennetzen 2006). The average ratio in *A. duranensis rt* sequences was 4.073 (±0.745):1, in *A. ipaënsis* 4.021 (±0.722):1, and in *A. hypogaea* 3.473 (±0.605):1. In spite of these indications for methylation of FIDEL, particularly in the promoter region, significant hits were found in a Blast search of the whole element against the *Arachis* EST database, mostly to sequences from a normalized cDNA library of cotyledons and young leaves of *A. hypogaea*.

A phylogenetic analysis of Ty3-*gypsy* elements including FIDEL was carried out in order to obtain further evidence for its position within the group of retrovirus-like retrotransposons. The neighbor-joining tree of reverse transcriptases from FIDEL and a diverse set of *gypsy* elements (Fig. 6) is divided into two major clades, one with the LORE elements (red bars), which are generally characterized by lack of

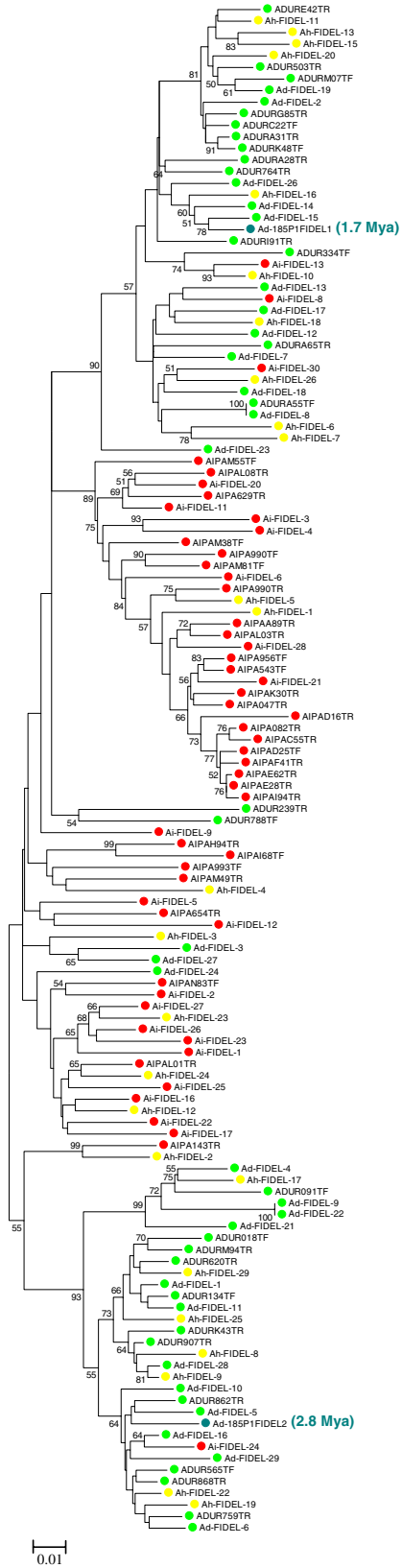

Fig. 5 Neighbor-joining tree based on nucleic acid sequences of 132 FIDEL *rt* sequences derived from cloned PCR products from *A. duranensis*, *A. ipaënsis*, and *A. hypogaea* (Ad-FIDEL-x, Ai-FIDEL-x, Ah-FIDEL-x) and from homologous BAC end sequences (ADURxxxTF/TR, AIPAxxxTF/TR). Also included are the *rt* sequences from the isolated complete elements *Ad*-185P1FIDEL1 and -2, which have an estimated age of insertion 1.7 and 2.8 Mya (sequences in ESM Supplement S1). Prior to phylogenetic analysis, the sequences were aligned using ClustalW. Bootstrap values from 1,000 replicates are shown next to the branches as percentages (values lower than 50% are hidden). Sequence names were color-labeled to facilitate discrimination of the host species: *A. duranensis* (*green dots*), *A. ipaënsis* (*red dots*), and *A. hypogaea* (*yellow dots*)

significant sequence between the end of the integrase and the 3′LTR, and the other with members of the *Athila/Calypso* (blue bars) and *Tat* group (green bars), which are further separated into two distinct branches with 98% bootstrap support. FIDEL is clearly grouped into the *Athila/Calypso* branch, close to the retrovirus-like elements *Cyclops*-1 and *Cyclops*-2 from *P. sativum* and Fababean-1 from *Vicia faba*. The tree also shows that FIDEL is distinct from *Cyclops*-1, which was earlier reported on the basis of Southern hybridization to be present in various legume species including *A. hypogaea* (Chavanne et al. 1998). We identified an overlapping sequence of 569 bp with 72% identity between the 864 *Sph*I–*Nhe*I fragment of the *Cyclops*-1 *rt* sequence used as a probe in that study and the FIDEL *rt* sequence (data not shown). Blasting the total sequence of *Cyclops* against FIDEL and all available BAC end sequences did not show any further significant similarities. Therefore, it is highly possible that in the described experiment, actually FIDEL was detected in *A. hypogaea* rather than a version of the *Cyclops* element.

Patterns of insertion of FIDEL on the approximately 100-kb scale observed using paired BAC end sequences

The observed frequencies of BAC end sequence pairs where both of a sequence pair were FIDEL were significantly higher than expected at a statistically significant level ($P=1.6\times10^{-12}$). This may imply that FIDEL clusters within the *Arachis* genome or may be caused by an overrepresentation of FIDEL in BAC end sequences because of the presence of *Hind*III sites within the element.

The observed frequencies of all gene classes in Bin0 and Bin1 were very low; this is to be expected

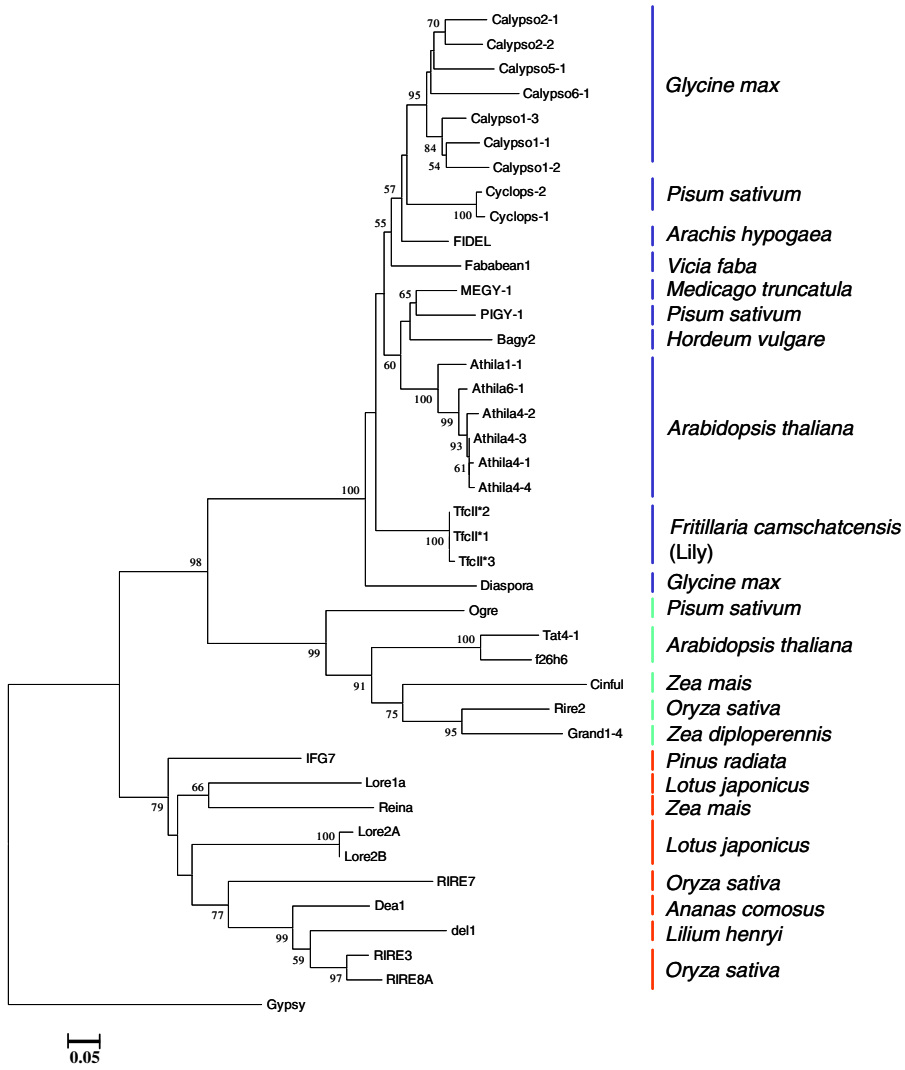

Fig. 6 Neighbor-joining tree based on amino acid sequences of reverse trancriptases of Ty3-*gypsy* class LTR retrotransposons from higher plants. Bootstrap values from 1,000 replicates are shown next to the branches as percentages (values lower than 50% are hidden). The tree is rooted to *gypsy* (P10401). Apart from MEGY-1 (AC146683), PIGY-1 (AY299398, nucleotides 26–13419), Ogre (AY299398, nucleotides 14501–37253; the RT sequences of the *Athila/Calypso* and *Tat* group were derived from supplemental material to Wright and Voytas (2002). The GenBank accession nos. of other *gypsy*-like elements included in the analysis (lower clade) are as follows (from *top to bottom*): IFG7: AJ004945; Lore1a: AJ966990; Reina: U69258; Lore2A: AB430320; Lore2B: AB430231; RIRE7: BAA89466; Dea1: Y12432; del1: X13886; RIRE3: AB014738; RIRE8A: AB014746. Phylogenetic analyses were conducted in MEGA4

because they represent FIDEL sequences. The presence of some gene classes in this group suggests that in a few cases, FIDEL is inserted into a genic sequence. The remainder of the analysis hinges on Bin2 and Bin3. The observed frequencies of conserved Unique genes from *Arabidopsis* and rice and paralogous gene families in *Arabidopsis* (minus NBS encoding genes) were significantly lower in sequences paired with FIDEL (Bin2) than expected from overall frequencies (Bin2 plus Bin3: $P=8.5\times10^{-7}$; $P=2.8\times10^{-13}$; $P\approx0$, respectively). This suggests that overall, FIDEL tends to insert in positions that are distant from ancestral conserved genes, which can be presumed to be mostly slowly evolving. However, the observed frequency of the fast-evolving homologs NBS encoding genes in Bin2 is very close to that expected from overall frequencies. A summary of the results is shown in Table 2.

Table 2 Summary results of the analysis of tendencies of genome insertion positions for the retrotransposon FIDEL

	Bin0 Seq1f , Seq1r Both FIDEL	Bin1 Seq2f FIDEL	Bin2 Seq2r Non-FIDEL	Bin3 Seq3f, Seq3r Both non-FIDEL
FIDEL bin groupings Total no. of sequences	3,102 expected 2596 $P=1.6\times10^{-12}$	7,094	7,094	22,748
No. of *At* Unique gene hits	2	0	34 Expected 72 $P=8.5\times10^{-7}$	269
No. of *Os* Unique gene hits	9	6	132 Expected 233 $P=2.8\times10^{-13}$	851
No. of *At* paralogous (minus NBS encoding) gene hits	1	3	163 Expected 229 $P \approx 0$	1052
No. of *At*-NBS protein encoding genes hits	0	0	30 Expected 34 $P=0.62$	111

The first row of the table gives the total number of sequences in the "bins," the definition of which is detailed in "Material and methods" and is summarized in the column headings. Rows represent the number of significant similarities to different gene groups in the bins. Bin2 sequences (sequences which are paired with FIDEL) show significant underrepresentation of homologs of *At* and *Os* Unique genes and of *At* paralogous genes, but not of NBS encoding resistance gene homologs

Discussion

The presented work evolved from our efforts in isolating repetitive elements from the A- and B-genome of *Arachis* aimed at finding out more about the genomic relationships between the ancestors of cultivated peanut. The isolated retroelement FIDEL to our knowledge is the first completely isolated and comprehensively characterized retrotransposon in *Arachis* species. Analysis of the translated sequence revealed that FIDEL shares the typical structure of a Ty3-*gypsy* LTR retrotransposon. Phylogenetic analysis of the reverse transcriptase sequences of FIDEL and other plant Ty3-*gypsy* elements allows two major conclusions: Firstly, FIDEL belongs to the *Athila*/*Calypso* group of retrovirus-like elements, which together form a strongly supported clade, and secondly, it is a distinct element within this group, clearly separated from other elements. FIDEL reveals the characteristic features of other *Athila*/*Calypso* group elements, namely, long LTRs >1.2 kb, a PBS complementary to the 3′-end of Asp tRNA, overall sequence length of more than 11 kb, and strong similarities to the conserved amino acid motifs of Gag-Pol. With the exception of *Diaspora*, the members of the *Tat* and the *Athila/Calypso* branch are further characterized by substantial sequences between the stop of the integrase gene and the 3′ LTR. The members of the *Athila/Calypso* branch, wherever full-length sequence is available, exhibit an additional ORF with *env*-like characteristics, although with various degrees of sequence degeneracy. In FIDEL, it was not possible to identify an additional ORF encoding *env*-like amino acid sequences. However, the sequences available did give some indications that in the past, an *env*-like ORF may have existed in FIDEL. It cannot be ultimately decided if the 3′UTR carries remnants of any cellular gene or the vestiges of a degraded region that once encoded transmembrane regions.

According to the dot blot analysis, FIDEL is present in about 3,000 (±950) copies in the *A. duranensis* genome (~2.7%), about 820 (±480) copies in the *A. ipaënsis* genome (~0.7%), and about 3,900 (±1,500) copies in the *A. hypogaea* genome (1.5%). This essentially means an at least threefold higher copy number in the A-genome ancestor as compared to the B-genome ancestor. Calculation of absolute copy numbers requires knowledge of the actual genome sizes in question, and often, *C* values vary

depending on the methodology and standards used for their determination (Greilhuber 2005). The values of DNA contents of the *Arachis* species studied here diverge drastically in the literature. The 2*C* DNA contents of *A. duranensis* and *A. hypogaea* reported by Temsch and Greilhuber (2000, 2001) are two times smaller than the data described earlier by Singh et al. (1996). Since Greilhuber (2005) convincingly demonstrated the correctness of their numbers for these two species, we concluded that previously published data on *A. ipaënsis* given by Singh et al. (1996) are also a twofold overestimation of the actual reality. This conclusion was further supported by cytological data. In addition, copy number estimations, especially in the case of higher copy numbers, often are controversial since they are influenced by several experimental factors depending on the method applied. We have shown here that it is advisable to refer to more than one source of evidence for the quantitative data. In the presented case, the data obtained by dot blot analysis are supported by the relative estimations based on BAC end sequences and PCR screening of BAC plates. However, it also became clear that BAC end sequences from a library produced with only one restriction enzyme could lead to a biased result if used for quantitative matters. The copy numbers and relative contributions to the total genomes of various *Gypsy* elements are compiled in Table 3. The numbers for FIDEL are in the range of other Ty3-*gypsy* retrotransposons, in particular the three pea elements *Cyclops*, *Ogre*, and *Pigy*. Table 3 also shows that retrotransposons in plants with smaller genomes are present in smaller copy number, independent from their potential activity. This could be an indication for posttranscriptional silencing or effective elimination of newly integrated elements. The general tendency that large plant genomes expand by having a few retrotransposon families in high copy numbers was already described by Vitte and Bennetzen (2006).

Through GISH, it was possible to differentiate both genomes in metaphase preparations of an amphidiploid derived from a hybrid of *A. ipaënsis* with *A. duranensis*, which resembles the genome of *A. hypogaea*. Certain characteristics in the GISH hybridization pattern, such as weak hybridization to centromeres, chromosome ends, and to chromosome pair A9, were also found when using sequences of FIDEL as a probe in FISH experiments. The fact that multiple insertions of the element were found in four different *Arachis* species indicates that it is an ancient component of the genus. The distribution pattern of FIDEL in *A. duranensis* seems to be typical for the A-genome, as the results achieved with *A. stenosperma* have shown. Comparing the quantitative and qualitative distribution of FIDEL with the GISH results suggests that it critically contributes to the differences in the *Arachis* genomes. Trying to elucidate the fate of FIDEL in *A. ipaënsis*, one possible regulation mechanism, the generation of solo LTRs, appears to be unlikely in light of the FISH results. Since the probe used for FISH was an LTR sequence, fluorescence signals stand for complete FIDEL elements and also for solo LTRs. However, if the low FIDEL copy number in *A. ipäensis* would have resulted from genome-specific drastic reduction of a previously higher copy number by means of unequal intra-element recombination, the remaining solo LTRs should result in stronger fluorescence signals.

Localization in the euchromatic region of the chromosomes and absence from centromeric and telomeric regions and NOR is a characteristic feature often found with retrotransposons, particularly Ty1-*copia* elements (Brandes et al. 1997; Heslop-Harrison et al. 1997; Pearce et al. 1996), whereas the Ty3-*gypsy* *Athila* element in *Arabidopsis* is mainly associated with centromeric and pericentromeric regions (Pélissier et al. 1995; Fransz et al. 1998; Tabata et al. 2000). An *Athila*-like gypsy element was also found in *Brassica oleracea* to be distributed along the chromosomes with concentration in the broad centromeric region (Alix et al. 2005). Within the Fabaceae, Ty3-*gypsy* elements such as the *Ogre* element in *Vicia pannonica* or PIGY-1 in pea are dispersed over the entire genome and evenly distributed along the chromosomes (Neumann et al. 2005, 2006). In contrast, the *Calypso* element of soybean is preferentially located in heterochromatic and/or pericentromeric regions (Lin et al. 2005). Hence, FIDEL exhibits a pattern of distribution that is different from these other Ty3-*gypsy* elements, even those most closely related to it. Presence in the euchromatic regions means that at the time of insertion, the chance to modify expression of genes by silencing or activation is likely to be high compared to elements that preferably insert into heterochromatin, as are the chances that FIDEL influences gene evolution by promoting unequal crossing over and/or other genome restructuring events. Such events are likely to be

Table 3 Overview of selected Ty3-*gypsy* retroelements and their copy numbers

Element	Type, size	Active (+/−)	Species (DNA content [1C])	Method of copy number estimation	Copy number/(1*C*)	Percent of genome	Reference
Athila	Ty3-*gypsy*, ~10.5 kb	+ (only in strain with decreased ddm1)	*A. thaliana*, 0.2 pg[a]		150	1.2	Pélissier et al. (1995)
Tat1	Ty3-gypsy, ~4,000 bp	Not determined	*A. thaliana*, 0.2 pg[a]	Southern blot	2–10	0.006–0.03	Wright and Voytas (1998)
RIRE2	Ty3-gypsy, ~11 kb	Not determined	*O. sativa* 0.4 pg[a]	Southern, dot blot	≤5,000 complete elements	14	Ohtsubo et al. (1999)
Lore1	Ty3-*gypsy*, 5,041 bp	+	*L. japonicus*, 0.48 pg[b]	Southern blot analysis	10	0.01	Madsen et al. (2005)
Diaspora	Ty3-*gypsy*, 11,737 bp	−	*G. max*, 1.1 pg[c]	Hybridization of BAC clone filter	500	0.5	Yano et al. (2005)
Cyclops	Ty3-*gypsy*, 12,314 bp	−	*P. sativum*, 4.4 pg[c]	Hybridization of genomic library	5,000	0.9	Chavanne et al. (1998)
Ogre	Ty3-*gypsy*, ~22 kb	+	*P. sativum*, 4.4 pg[c]	Hybridization of phage and cosmid clones from pea	10,000	4.2	Neumann et al. (2003)
Pigy	Ty3-*gypsy*, 13,645 bp	+	*P. sativum*, 4.4 pg[c]	Colony/plaque blot hybridizations	1,000–5,000	0.3–14	Neumann et al. (2005)
Ogre	Ty3-*gypsy*, ~25 kb	+	*V. pannonica*, 6.75 pg[b]	(a) Dot blot; (b) hybridization to short insert libraries	~100,000	38	Neumann et al. (2006)
FIDEL	Ty3-*gypsy* 11,223 bp	−	*A. duranensis* 1.305 pg	(a) Dot blot	~3,000	~2.7	This manuscript
FIDEL	Ty3-*gypsy* 11,223 bp	−	*A. ipaënsis* 1.4 pg	(a) Dot blot	~820	~0.7	This manuscript
FIDEL	Ty3-*gypsy* 11,223 bp	−	*A. hypogaea* 2.96 pg	(a) Dot blot	~3,900	~1.5	This manuscript

The elements are ordered according to the genome size of their host species. Listed are as well observed activity, the percentage on the respective genome, and the method applied by the authors to determine the copy number. In the case of the *Arachis* species, the values refer to original references (Temsch and Greilhuber 2000, 2001)

[a] 1*C* values were retrieved from Bennett et al. (2000)

[b] 1*C* values were retrieved from Bennett and Leitch (2004)

[c] 1*C* values were retrieved from Bennett and Leitch (1997)

deleterious when they affect most genes, especially those that are evolutionarily conserved. However, these events may be less deleterious or perhaps even advantageous when the genes are less conserved or fast-evolving. Supporting this idea, an analysis of paired BAC end sequences suggests that FIDEL does indeed show a tendency to insert at sites distant from most ancestrally conserved genes, but that it shows no such tendency with regard to the fast-evolving NBS-encoding disease resistance gene homologs (Table 2). This pattern of FIDEL insertion sites is consistent with the model for gene space in papilionoids recently proposed by Bertioli et al. (2009). In this model, it is suggested that gene space may be divided into two broadly defined components: more conserved regions which tend to have low retrotransposon densities are relatively stable during evolution and have higher densities of evolutionary conserved genes and variable regions that tend to have high retrotransposon densities and, in some cases, higher densities of certain fast-evolving genes such as resistance gene homologs.

Phylogenetic analysis of *rt* sequences from *A. duranensis*, *A. ipaënsis*, and *A. hypogaea* revealed that FIDEL experienced a distinct evolution in the two parental genomes of cultivated peanut. The sequences from *A. duranensis* and *A. ipaënsis* emerged almost

exclusively in separate clades. Furthermore, the presence of two well-supported *A. duranenis* clades, each with one sequence dated 1.7 and 2.8 Mya, respectively, suggests that there were two eras of amplification of FIDEL in *A. duranensis*. On the other hand, in *A. ipaënsis*, the element suffered a different evolutionary history. The values for synonymous and non-synonymous substitutions (K_s and K_a) were similar in both species, indicating that the elements have suffered conservative selection (data not shown). It is possible that particular species-specific mutations might have contributed to different amplification patterns. Conceivable also are pre- and/or posttranscriptional processes in the B-genome that repressed further amplification of the element or the deletion of complete existing copies. The phenomenon of differential amplification of specific *gypsy* elements has been described in the different lineages of cotton (Hawkins et al. 2006). It is also of interest if the allotetraploid hybrid *A. hypogaea* exhibits any signs of reactivation of FIDEL due to "genomic shock," as it was described for retroelements in other allopolyploid species such as cotton and wheat (Zhao et al. 1998; Kashkush et al. 2002). Both copy number estimations and FISH could not demonstrate any substantial increase in FIDEL numbers or a change in genome distribution. In the phylogenetic tree of *rt* sequences, FIDEL sequences from *A. hypogaea* were distributed within the ancestral clades with a similar proportion as was expected from copy number estimations. It seems that before the unification of both genomes in tetraploid cultivated peanut, most if not all copies of the element were already in a mutated and/or epigenetic silenced state that did not allow generation of new copies. This is suggested by the scattered stop codons found in the coding regions of the PCR-cloned and BAC-derived *rt* sequences and supported by the Ts/Tv ratio >1.5:1 found in the LTR sequences of FIDEL-1 and FIDEL-2 from *ADUR*185P1. Searching EST data from GenBank resulted in a few hits suggesting FIDEL does show some activity. However, these data do not permit quantitative conclusions on the level of activity because most of the EST sequences identified were from normalized cDNA libraries. These findings imply that perhaps some of the FIDEL copies could be in a state that allows transcription. However, of the mRNAs that are transcribed, most are probably not translated to fully functional proteins due to indels, substitutions, and interspersed stop codons. Even if non-autonomous activation cannot be excluded, it rather appears that FIDEL elements, at least at the moment, rest as a more or less silent part of the different *Arachis* genomes, where in the past, they played a role in the development and differentiation of the genomes within the *Arachis* species complex.

Acknowledgments We are grateful to Douglas Cook for useful discussions and for BAC end data generated within the National Science Foundation Project number 0605251. We thank Marc D. Burow for critical reading of the manuscript. We also thank José Valls for providing *Arachis* germplasm. The presented work was supported by EU INCO-DEV (ARAMAP: ICA4-2001-10072) and the Generation Challenge Program (Project G3005.05 and TLI).

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Alix K, Ryder CD, Moore J, King GJ, Pat Heslop-Harrison JS (2005) The genomic organization of retrotransposons in *Brassica oleracea*. Plant Mol Biol 59:839–851

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Armisén D, Lecharny A, Aubourg S (2008) Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol Biol 8:280

Bennett MD, Leitch IJ (1997) Nuclear DNA amounts in angiosperms—583 new estimates. Ann of Bot (Lond) 80:169–196

Bennett MD, Leitch IJ (2004) Plant DNA *C*-values database (release 3.0, Dec. 2004). http://www.kew.org/cvalues/homepage.html

Bennett MD, Bhandol P, Leitch IJ (2000) Nuclear DNA amounts in angiosperms and their modern uses—807 new estimates. Ann Bot (Lond) 86:859–909

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot (Lond) 95:127–132

Bertioli D, Moretzsohn M, Madsen LH et al (2009) An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 10:45

Brandes A, Heslop-Harrison JS, Kamm A, Kubis S, Doudrick RL, Schmidt T (1997) Comparative analysis of the chromosomal and genomic organization of Ty1-*copia*-like retrotransposons in pteridophytes, gymnosperms and angiosperms. Plant Mol Biol 33:11–21

Burow MD, Simpson CE, Faries MW, Starr JL, Paterson AH (2009) Molecular biogeographic study of recently described B- and A-genome *Arachis* species, also providing

new insights into the origins of cultivated peanut. Genome 52:107–119

Chavanne F, Zhang DX, Liaud MF, Cerff R (1998) Structure and evolution of *Cyclops*: a novel giant retrotransposon of the *Ty3*/*Gypsy* family highly amplified in pea and other legume species. Plant Mol Biol 37:363–375

Cheng C, Daigen M, Hirochika H (2006) Epigenetic regulation of the rice retrotransposon Tos17. Mol Genet Genomics 276:378–390

Devos KM, Brown JKM, Bennetzen JL (2002) Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12:1075–1079

Fávero AP, Simpson CE, Valls JFM, Vello NA (2006) Study of the evolution of cultivated peanut through crossability studies among *Arachis ipaënsis*, *A. duranensis*, and *A. hypogaea*. Crop Science 46:1546–1552

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3:329–341

Fonceka D, Hodo-Abalo T, Rivallan R et al (2009) Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid. BMC Plant Biol 9:103

Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G (1998) Cytogenetics for the model system *Arabidopsis thaliana*. Plant J 13:867–876

Fukai E, Dobrowolska AD, Madsen LH et al (2008) Transposition of a 600 thousand-year-old LTR retrotransposon in the model legume *Lotus japonicus*. Plant Mol Biol 68:653–663

Gerlach WL, Bedbrook JR (1979) Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res 7:1869–1885

Gerlach WL, Dyer TA (1980) Sequence organization of the repeating units in the nucleus of wheat, which contain 5S rRNA genes. Nucleic Acids Res 8:4851–4865

Grattapaglia D, Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137:1121–1137

Greilhuber J (2005) Intraspecific variation in genome size in angiosperms: identifying its existence. Ann Bot (Lond) 95:91–98

Guimarães PM, Garsmeur O, Proite K et al (2008) BAC libraries construction from the ancestral diploid genomes of the allotetraploid cultivated peanut. BMC Plant Biol 8:14

Hammons RO (1994) The origin and early history of the peanut. In: Smartt J (ed) The peanut crop: a scientific basis for improvement. Chapman and Hall, London, pp 24–42

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16:1252–1261

Heslop-Harrison JS, Brandes A, Taketa S et al (1997) The chromosomal distributions of Ty1-*copia* group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100:197–204

Huang X, Madan A (1999) CAP3: a DNA sequence assembly program. Genome Res 9:868–877

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic, New York, pp 21–132

Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166:1437–1450

Kashkush K, Feldman M, Levy AA (2002) Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. Genetics 160:1651–1659

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

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Laten HM, Majumdar A, Gaucher EA (1998) *SIRE-1*, a *copia*/*Ty1*-like retroelement from soybean, encodes a retroviral envelope-like protein. Proc Natl Acad Sci U S A 95:6897–6902

Lin JY, Jacobus BH, SanMiguel P et al (2005) Pericentromeric regions of soybean (*Glycine max* L. Merr.) chromosomes consist of retroelements and tandemly repeated DNA and are structurally and evolutionarily labile. Genetics 170:1221–1230

Liu ZL, Han FP, Tan M et al (2004) Activation of a rice endogenous retrotransposon *Tos17* in tissue culture is accompanied by cytosine demethylation and causes heritable alteration in methylation pattern of flanking genomic regions. Theor Appl Genet 109:200–209

Ma JX, Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Madsen LH, Fukai E, Radutoiu S et al (2005) LORE1, an active low-copy-number TY3-*gypsy* retrotransposon family in the model legume *Lotus japonicus*. Plant J 44:372–381

Maluszynska J, Heslop-Harrison JS (1993) Physical mapping of rDNA loci in *Brassica* species. Genome 36:774–781

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 11:1113–1130

Neumann P, Pozarkova D, Macas J (2003) Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plant Mol Biol 53:399–410

Neumann P, Pozarkova D, Koblizkova A, Macas J (2005) PIGY, a new plant envelope-class LTR retrotransposon. Mol Genet Genomics 273:43–53

Neumann P, Koblizkova A, Navratilova A, Macas J (2006) Significant expansion of *Vicia pannonica* genome size mediated by amplification of a single type of giant retroelement. Genetics 173:1047–1056

Nielen S, de Assis C, Fonseca F, Guimarães P, Leal-Bertioli SC, Bertioli D (2009) Isolation and characterization of retrotransposons in wild and cultivated peanut species. In: Shu QY (ed) Induced plant mutations in the genomics era. Food and Agriculture Organization of the United Nations, Rome, pp 499–502

Ohtsubo H, Kumekawa N, Ohtsubo E (1999) RIRE2, a novel *gypsy*-type retrotransposon from rice. Genes Genet Syst 74:83–91

Orlov YL, Potapov VN (2004) Complexity: an Internet resource for analysis of DNA sequence complexity. Nucleic Acids Res 32:W628–W633

Pearce SR, Harrison G, Li D, Heslop-Harrison J, Kumar A, Flavell AJ (1996) The Ty1-*copia* group retrotransposons in *Vicia* species: copy number, sequence heterogeneity and chromosomal localisation. Mol Gen Genet 250:305–315

Pélissier T, Tutois S, Deragon JM, Tourmente S, Genestier S, Picard G (1995) *Athila*, a new retroelement from *Arabidopsis thaliana*. Plant Mol Biol 29:441–452

Raina SN, Mukai Y (1999) Genomic in situ hybridization in *Arachis* (*Fabaceae*) identifies the diploid wild progenitors of cultivated (*A. hypogaea*) and related wild (*A. monticola*) peanut species. Plant Syst Evol 214:251–262

Sabot F, Schulman AH (2006) Parasitism and the retrotransposon life cycle in plants: a hitchhiker's guide to the genome. Heredity 97:381–388

Sabot F, Sourdille P, Chantret N, Bernard M (2006) Morgane, a new LTR retrotransposon group, and its subfamilies in wheats. Genetica 128:439–447

SanMiguel B, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot (Lond) 82(Supplement A):37–44

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43–45

Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schrire BD, Lewis GP, Lavin M (2005) Biogeography of the Leguminosae. In: Lewis G, Schrire B, Mackinder B, Lock M (eds) Legumes of the world. Royal Botanic Gardens, Kew, Kew, pp 21–54

Schwarzacher T, Heslop-Harrison JS (2000) Practical in situ hybridization. Springer, New York

Seijo JG, Lavia GI, Fernandez A, Krapovickas A, Ducasse D, Moscone EA (2004) Physical mapping of the 5S and 18S–25S rRNA genes by fish as evidence that *Arachis duranensis* and *A. ipaënsis* are the wild diploid progenitors of *A. hypogaea* (*Leguminosae*). Am J Bot 91:1294–1303

Seijo G, Lavia GI, Fernandez A et al (2007) Genomic relationships between the cultivated peanut (*Arachis hypogaea*, *Leguminosae*) and its close relatives revealed by double GISH. Am J Bot 94:1963–1971

Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P (2000) A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Res 10:908–915

Singh KP, Raina SN, Singh AK (1996) Variation in chromosomal DNA associated with the evolution of *Arachis* species. Genome 39:890–897

Snider C, Jayasinghe S, Hristova K, White SH (2009) MPEx: a tool for exploring membrane proteins. Protein Sci 18:2624–2628

Staden R (1996) The Staden sequence analysis package. Mol Biotechnol 5:233–241

Tabata S, Kaneko T, Nakamura Y et al (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature 408:823–826

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol Biol Evol 24:1596–1599

Temsch EM, Greilhuber J (2000) Genome size variation in *Arachis hypogaea* and *A. monticola* re-evaluated. Genome 43:449–451

Temsch EM, Greilhuber J (2001) Genome size in *Arachis duranensis*: a critical study. Genome 44:826–830

Thompson JD, Higgins DG, Gibson TJ (1994) 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

Vicient CM, Suoniemi A, Anamthawat-Jonsson K et al (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. Plant Cell 11:1769–1784

Vitte C, Bennetzen JL (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Wortman JR, Haas BJ, Hannick LI et al (2003) Annotation of the *Arabidopsis* genome. Plant Physiol 132:461–468

Wright DA, Voytas DF (1998) Potential retroviruses in plants: *Tat1* is related to a group of *Arabidopsis thaliana* Ty3/*gypsy* retrotransposons that encode envelope-like proteins. Genetics 149:703–715

Wright DA, Voytas DF (2002) *Athila4* of *Arabidopsis* and *Calypso* of soybean define a lineage of endogenous plant retroviruses. Genome Res 12:122–131

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353–3362

Yano ST, Panbehi B, Das A, Laten HM (2005) *Diaspora*, a large family of Ty3-*gypsy* retrotransposons in *Glycine max*, is an envelope-less member of an endogenous plant retrovirus lineage. BMC Evol Biol 5:30

Yüksel B, Bowers JE, Estill J, Goff L, Lemke C, Paterson AH (2005) Exploratory integration of peanut genetic and physical maps and possible contributions from *Arabidopsis*. Theor Appl Genet 111:87–94

Zhang XY, Wessler SR (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proc Natl Acad Sci USA 101:5589–5594

Zhao XP, Si Y, Hanson RE et al (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res 8:479–492