Mol Genet Genomics (2012) 287:21–38
DOI 10.1007/s00438-011-0656-6

ORIGINAL PAPER

Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the *Arachis* A–B genome divergence

Stephan Nielen · Bruna S. Vidigal · Soraya C. M. Leal-Bertioli · Milind Ratnaparkhe · Andrew H. Paterson · Olivier Garsmeur · Angélique D'Hont · Patricia M. Guimarães · David J. Bertioli

Received: 10 June 2011 / Accepted: 20 October 2011 / Published online: 27 November 2011

Abstract Cultivated peanut is an allotetraploid with an AB-genome. In order to learn more of the genomic structure of peanut, we characterized and studied the evolution of a retrotransposon originally isolated from a resistance gene analog (RGA)-containing bacterial artificial chromosome (BAC) clone. It is a moderate copy number Ty1-*copia* retrotransposon from the *Bianca* lineage and we named it *Matita*. Fluorescent in situ hybridization (FISH) experiments showed that *Matita* is mainly located on the distal regions of chromosome arms and is of approximately equal frequency on both A- and B-chromosomes. Its chromosome-specific hybridization pattern facilitates the identification of individual chromosomes, a useful cytogenetic tool considering that chromosomes in peanut are mostly metacentric and of similar size. Phylogenetic analysis of *Matita* elements, molecular dating of transposition events, and an estimation of the evolutionary divergence of the most probable A- and B-donor species suggest that *Matita* underwent its last major burst of transposition activity at around the same time of the A- and B-genome divergence about 3.5 million years ago. By probing BAC libraries with overgos probes for *Matita*, resistance gene analogues, and single- or low-copy genes, it was demonstrated that *Matita* is not randomly distributed in the genome but exhibits a significant tendency of being more abundant near resistance gene homologues than near single-copy genes. The described work is a further step towards broadening the knowledge on genomic and chromosomal structure of peanut and on its evolution.

Communicated by M.-A. Grandbastien.

Electronic supplementary material The online version of this article (doi:10.1007/s00438-011-0656-6) contains supplementary material, which is available to authorized users.

S. Nielen · S. C. M. Leal-Bertioli · P. M. Guimarães
Embrapa Recursos Genéticos e Biotecnologia,
70770-917 Brasília, DF, Brazil

Present Address:
S. Nielen (✉)
Plant Breeding and Genetics Section, Joint FAO/IAEA Division
of Nuclear Techniques in Food and Agriculture,
International Atomic Energy Agency, Vienna, Austria
e-mail: S.Nielen@iaea.org

B. S. Vidigal · D. J. Bertioli
Universidade de Brasília, Campus Universitário,
70910-900 Brasília, DF, Brazil

B. S. Vidigal · D. J. Bertioli
Universidade Católica de Brasília, Campus II, SGAN 916,
70790-160 Brasília, DF, Brazil

M. Ratnaparkhe · A. H. Paterson
Plant Genome Mapping Laboratory, The University of Georgia,
Athens 30605, USA

O. Garsmeur · A. D'Hont
Centre de Coopération International en Recherche Agronomique
pour le Developpement (CIRAD), Montpellier, France

Keywords Peanut · Arachis · Retrotransposon · Evolution · Fluorescent in situ hybridization

Abbreviations

BAC	Bacterial artificial chromosome
BES	BAC end-sequences
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	Poly purine tract
RGA	Resistance gene analogue
RT	Reverse transcriptase
SDS	Sodium dodecyl sulphate
SSC	Standard saline citrate (1×SSC = 0.15 M NaCl; 0.015 M Na$_3$-citrate)
UTR	Untranslated region

Introduction

Cultivated peanut (*Arachis hypogaea*) is tetraploid with an AB-genome ($2n = 4x = 40$) of recent origin, arising from hybridization of two wild species and spontaneous chromosome duplication. Comparisons of the karyotypes of diploid wild species with *A. hypogaea*, together with molecular data and phylogeographic considerations, suggests that *A. duranensis* (A-genome) and *A. ipaënsis* (B-genome) (both $2n = 2x = 20$) are the extant species most closely related to the ancestors of cultivated peanut (Kochert et al. 1996; Seijo et al. 2004, 2007; Burow et al. 2009). Homoeolgous A- and B-chromosomes rarely pair during meiosis (Smartt 1990), thus the peanut genome can be characterized as being genetically diploid.

Cultivated peanut has a very limited DNA diversity (Kochert et al. 1996; Milla et al. 2005). Therefore, wild species are an attractive source of new alleles, traits, and higher DNA polymorphism which facilitates genetic map construction (Simpson et al. 1993; Moretzsohn et al. 2005; Leal-Bertioli et al. 2009; Foncéka et al. 2009). Cytogenetic markers allowing the following of individual chromosomes in hybrids derived from cultivated × wild crosses, as well as to detect genome/chromosomal rearrangements would be of great value. For studies based only on cultivated peanut, a lack of available tools for genetic mapping and a lack of knowledge of the genome in general are major obstacles. A whole genome sequence would be a major step to overcome these difficulties. The design of sequencing and assembly strategies, however, require a more thorough understanding of the repeat structure of the peanut genome, which also would contribute to understanding the genetic behaviour and biology of the peanut genome in general.

Due to small chromosome sizes and metacentric and sub-metacentric morphologies karyotype analysis in peanut is challenging. Comprehensive studies of karyotypes in *Arachis* based on classical cytogenetics have been made by Fernández and Krapovickas (1994), who described different types of SAT chromosomes and the small chromosome pair A9 as a characteristic feature of the A-genome component. The use of DAPI (4′-6-diamidino-2-phenylindole) for chromosome staining revealed distinct heterochromatic bands of A-genome chromosomes with the most prominent band in the pair A9 and absent or considerably weaker bands in the B-genome (Seijo et al. 2004). Using molecular cytogenetics with rDNA sequences as probes in FISH, the latter authors substantially contributed to the identification of the most probable ancestors of peanut.

From genetic maps, it is apparent that the order of molecular markers (which are derived predominantly from low-copy DNA) in the A- and B-genomes is mostly co-linear with only a few major rearrangements (Burow et al. 2001; Moretzsohn et al. 2009). Furthermore, comparative mapping has shown that there is detectable gene synteny between *Arachis* and *Lotus*, *Medicago* and *Phaseolus,* legumes that are separated from peanut by an estimated divergence time of 55 million years ago (Mya) (Hougaard et al. 2008; Bertioli et al. 2009; Wojciechowski et al. 2004; Cannon et al. 2010). This emphasises the slow evolution of gene order and suggests that much of the gene space of the A- and B-genome components is likely to be highly similar.

However, in most plants it is repetitive DNA and not genes that occupy most of the genome and determine large-scale structure of the chromosomes (Schmidt and Heslop-Harrison 1998). In contrast to the surprising degree of conservation observed for genes, whole genome in situ hybridization suggests that the repetitive genome fractions of the A- and B-genomes, and indeed of different species of wild peanut in general, are substantially diverged (Seijo et al. 2007). The retroelement FIDEL provides one example of a family of repetitive sequences with considerable quantitative and sequence-related divergence between the component genomes (Nielen et al. 2010). In allopolyploid species further importance is attached to the repetitive DNA as a dynamic part of the genome. The genomic shock resulting from unification of different genomes in one nucleus can initiate reactivation of transposable elements with potential consequences for genome structure and gene expression (Zhao et al. 1998; Kashkush et al. 2002, 2003; Petit et al. 2010).

As part of the repetitive DNA, transposable elements and in particular long terminal repeat (LTR) retrotransposons contribute a substantial fraction of genomes, making as much as 80% of the genome in plants such as in maize (SanMiguel and Bennetzen 1998). LTR elements can be divided into two superfamilies, the Ty1-*copia* retrotransposons (*pseudoviridae*), and the Ty3-*gypsy* retrotranspsons (*metaviridae*) (Xiong and Eickbush 1990), which differ in their reverse transcriptases, and in their structure. In the

Ty1-*copia* elements the integrase gene antecedes the reverse transcriptase gene and in the Ty3-*gypsy* elements it is located after the RNaseH gene (Kumar and Bennetzen 1999). The lifecycle of retrotransposons is complex (reviewed by Sabot and Schulman 2006) and a number of regulatory measures can interfere between initiation of transcription and successful integration of a new complete copy (Feschotte et al. 2002). Many retrotransposon copies found in the genome exhibit insertions, deletions, and frameshifts and thus, are not functional (Kumar and Bennetzen 1999, and herein mentioned references). Moreover, solo LTRs are generated by illegitimate intraelement recombination, which counteracts retrotransposon-driven genome expansion (Shirasu et al. 2000; Bennetzen et al. 2005).

Recently we described a new peanut Ty3-*gypsy* retrotransposon named FIDEL. It is more abundant in the A- than the B-genome and this uneven distribution is due to transposition activity that occurred in the A-genome, but not in the B-genome. Most likely these bursts of activity occurred after the evolutionary divergence of the A- and B-donor species (Nielen et al. 2010). Here we describe a new moderately repetitive Ty1-*copia* retrotransposon and name it *Matita*. In contrast to FIDEL, *Matita* is of similar abundance in the A- and B-genome components and has distinct chromosomal distributions. We investigate the evolution of *Matita*, suggest a place for it within the evolutionary events that gave rise to the cultivated tetraploid genome, and investigate its role in the definition of genome structure of cultivated peanut.

Materials and methods

Plant materials and DNA extraction

Leaf tissue was obtained from *A. hypogaea* cv. IAC-Runner-886, *A. duranensis* (accession V14167), and *A. ipaënsis* (accession KG30076). Originally, seed was obtained from the Active Germplasm Bank of Embrapa Recursos Genéticos e Biotecnologia. 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 agarose gel electrophoresis, comparing the fluorescence intensities of the ethidium bromide-stained samples to DNA mass standards.

BAC selection and sequencing and annotation

Filters of a BAC library of genomic DNA from *A. duranensis* (Guimarães et al. 2008) were probed with a radioactively labelled resistance gene analogue S1_A_36 (GenBank accession no. AY157808; Bertioli et al. 2003). A marker derived from this resistance gene analogue maps close to a QTL for resistance against late leaf spot (*Cercosporidium personatum* Berk. & M.A. Curtis) in a mapping population between the two wild A-genome species *A. duranensis* and *A. stenosperma* (Moretzsohn et al. 2005; Leal-Bertioli et al. 2009). (*A. stenosperma* is resistant against rust, late leaf spot, and root knot nematodes, all important pests of cultivated peanut; Leal-Bertioli et al. 2010; Proite et al. 2008). The selected BAC clone AD25F09 was subcloned using two shotgun strategies: cloning of restriction fragments after *Hin*dIII, or *Bam*HI digestion into pBlueScript SK-minus vector and shearing and cloning of DNA using the TOPO® Shotgun Subcloning Kit (Invitrogen) according to manufacturer's manual. Cloned inserts were sequenced using the BigDye Terminator sequencing kit on Applied Biosystems sequencers. Sequence quality was estimated using Phred (Ewing and Green 1998) and sequences assembled to contigs using Cap3 (Huang and Madan 1999) and manual assembly with Staden Package software (Staden 1996). Sequence analysis and annotation was done using RiceGAAS (Rice Genome Automated Annotation System—http://ricegaas.dna.affrc.go.jp/) (Sakata et al. 2002), BioEdit (Hall 1999), BLAST (Altschul et al. 1990), and the Spin module of the Staden sequence analysis package (Staden 1996). Based on the sequence information of the here identified retroelement *Matita*, a search for further copies of the *Matita* family was done in contigs derived from sequencing efforts of a *A. hypogaea L.* cv. Florunner *Hin*dIII BAC library, generated by Yüksel and Paterson (2005).

Chromosome preparation and fluorescent in situ hybridization

Root tips were harvested from seedlings of *A. hypogaea* cv. Tatu germinated in petri dishes on moist filter paper. The root tips were treated for 3 h with 8-hydroxyquinoline (2 mM) prior to fixation in 3:1 ethanol:acetic acid for at least 24 h. 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 used were a mixture of seven subclones spanning the whole *Matita* sequence for first hybridization and pTa71 containing the 18S-5.8S-25S rDNA genes from *Triticum aestivum* (Gerlach and Bedbrook 1979) for rehybridization. Probes were labelled either with digoxigenin-11-dUTP (*Matita* subclones) or biotin-11-dUTP (pTA71) (Roche) by random-primed labelling. Pretreatment, hybridization/rehybridization, 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 sulphate, 0.125 mM EDTA, 0.125% (w/v) SDS, and 1 μg of salmon sperm DNA. 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 85% 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 using the cooled CCD camera AxioCam MRm in combination with the AxioVison software (both Zeiss) and edited using only Adobe Photoshop functions (brightness, cropping, overlay), which affect the whole image equally.

Copy number estimation and analysis of distribution of *Matita*

Two approaches were used to estimate copy numbers (N) of *Matita.* In the first approach, 41,856 BAC end-sequences (BES) from the *A. duranensis* V14167 genomic BAC library (GenBank accession no. F1281689-F1321525) with average length of 666 bp were used. In BLASTn searches of 13 query sub-sequences (500 bp each), which together span the entire length of *Matita*, the number of BLAST detected sequence similarities with *E* values $\leq$1E$-$100 (*Nhits*) was determined and applied in the following formula: $N = [(1/\text{genome coverage}) \times N_{\text{hits}}/2][1 + (L_{\text{dr}} - L_{\text{eq}})/(L_{\text{dr}} + L_{\text{eq}})]$, where L_{eq} is the effective query length and L_{dr} the average length of database reads, in this case, the BES (Zhang and Wessler 2004). The formula considers the probability of 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.

In the second approach, high-density BAC filters with 182,784 *A. hypogaea*- and 84,096 *A. duranensis* BAC clones were screened for the presence of *Matita* using radioactively labelled "overgos" (Ross et al. 1999) based on *Matita* sequences. The high-density filters were derived from the previous experiments (Yüksel et al. 2005; Guimarães et al. 2008). The general overgos labelling, hybridization and washing procedures, and data entry and analysis were according to the methods described in Yüksel et al. (2005). Specifically, two overgos (cctagaaaggattcaaggtgatatatgtggacctattcat and caattaattgctagacccttgcttatgagaacaaatctcc) designed from the integrase domains were used for hybridization. After hybridization, the filters were blot-dried, wrapped in a sheet protector, and autoradiographed with two intensifying screens and X-ray film for 14 days at −80°C before developing. The BAC hits on the films were manually scored onto templates, scanned, and read by ABBYY FineReader ver. 5.0 with manual checking and correction. The BAC hit scores were converted to the BAC addresses with an in-house script and analysed with the BACman software (bacman.sourceforge.net) in order to assign each BAC to a specific overgos.

In order to investigate the distribution of *Matita* in the genome relative to the presence of resistance gene analogues (RGAs) and single-copy genes, the overgos probe data from *Matita* were cross-referenced with data from the previous experiments that used RGAs and low- or single-copy gene sequences as probes. In these experiments the *A. hypogaea* BAC library was probed with RGA sequences (Yüksel et al. 2005) and the *A. duranensis* BAC library with low- or single-copy gene sequences (Choi et al. 2006; and unpublished data) The number of BACs harbouring *Matita* and RGAs and *Matita* and single-copy genes, respectively, were determined and the frequency of *Matita* in RGA- and single-copy BACs calculated. The expected frequency for random distribution in the genome was derived from the quotient of all *Matita* positive BACs to the total number of BACs in the respective library. Using the two-tailed cumulative binomial distribution (in Excel), the *P* value for the null hypothesis of random distribution of *Matita* was calculated.

Evolutionary and phylogenetic analysis

Phylogenetic analyses of Matita elements

In order to further study the evolutionary history of the population of *Matita* elements that reside in the *Arachis* genomes, we amplified and cloned a 610-bp *Matita* reverse transcriptase (RT)-encoding nucleotide sequences from *A. duranensis*, *A. ipaënsis* and *A. hypogaea* cv. IAC-Runner-886. To obtain a diverse set of RT-coding sequences without bias, degenerate primers were designed using the *Matita* RT-coding sequence from BAC AD25F09 and also from highly similar DNA sequences identified in BES and genomic survey sequences (GSS). For bases differing between the retrieved sequences degenerate bases using IUPAC symbols were designated for primer synthesis. A list of the identified similar primer sequences resulting in the degenerate forward and reverse primers, and their origin can be found in Online Resource 1:

1. MatRT dFw 5′-TTGAGCYAAGAWYAGWYRAN-3′
2. MatRT dRv 5′-GCAATTATAAKGAATYCAN-3′

PCR was performed in 50 µl reactions containing 100 ng of plant DNA, 1 µM of each primer, 0.24 mM dNTPs, 2.5 mM $MgCl_2$, 2.5 units Taq polymerase (Invitrogen), and 1× Taq buffer. Thermocycling was: 35 cycles of, 94°C for 30 s, 43°C for 30 s, and 72°C for 60 s. PCR samples were cloned into pGEM-T Easy vector (Promega) according to manufacturer's instructions. Inserts were sequenced from a total of 72 colonies (36 of *A. duranensis*, 24 of *A. ipaënsis,* and 12 of *A. hypogaea*).

Translated sequences were manually edited in BioEdit and gaps were introduced where necessary to correct for frameshifts due to insertions or deletions. Alignments of these sequences were made using ClustalW (Thompson et al. 1994) and used for constructing a neighbour-joining tree of their nucleotide sequences with bootstrap analysis of 1,000 replicates, and by applying the Kimura 2-parameter method (Kimura 1980). The values of silent and non-silent nucleotide substitutions per synonymous and non-synonymous site (Ks and Ka, respectively) were calculated by pairwise analysis of the aligned sequences using the method of Nei-Gojobori (Nei and Gojobori 1986). Ratios of transitions (Ts) to transversions (Tv; Ts/Tv) between LTRs from complete elements were established using MEGA4 (Tamura et al. 2007).

Nucleotide identities between the *Matita* copy from BAC Ad25F09 and full-length copies identified in the contigs derived from *A. hypogaea* BAC sequencings were determined. Aligned sequences were scored with "1" for nucleotide identity and "0" for non-identity and the average identity values were calculated for 50 bp sliding windows. In the resulting chart, a cutoff of 0.25 nucleotide identity was applied because this is the random identity between DNA sequences.

In order to classify *Matita* within the described lineages of Ty1-*copia* elements, RT-coding sequences were derived from selected Ty1-*copia* elements from legumes, cereals, *Arabidopsis*, and from *Matita.* After manual inspection and editing in case of disrupted reading frames, the translated sequences were aligned and a phylogenetic tree was inferred using the maximum parsimony (MP) method. *Copia* sequences were retrieved from the TREP database (Triticeae) (http://wheat.pw.usda.gov/ITMI/Repeats/), Repbase (rice, sorghum, maize and *Arabidopsis*) (http://www.girinst.org/server/RepBase/), from supplementary material published by Wang and Liu (2008) (*Medicago truncatula*), and from GenBank (http://www.ncbi.nlm.nih.gov/), (*Phaseolus vulgaris, Vigna radiata*, *Lotus japonicus*, *Glycine max*). Selection of *Lotus* and *Glycine* sequences was based on supporting information of Du et al. (2010). The MP tree was obtained with bootstrap analysis of 1,000 replicates and using the Close-Neighbour-Interchange algorithm (Nei and Kumar 2000) with search level 1 in which the initial trees were obtained with the random addition of sequences (10 replicates). All phylogenetic trees and pairwise distance analysis were made using MEGA4 (Tamura et al. 2007).

Dating transposition of Matita elements

Using the full-length sequences of *Matita* obtained from BAC sequencing, the LTR divergence method was used to date the ages of insertion of three copies of the element. The number of nucleotide substitutions per site per LTR was determined using the equation $t = K/2r$, where t is the age, K is the number of nucleotide substitutions per site between each LTR pair, and r is the nucleotide substitution rate. 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).

Datamining cloning and analysis of sequences for dating the species divergence of A- and B-genomes

In order to study the genomic distribution of the retroelements against the background of species evolution, the yet unknown date of speciation of the A- and B-genome donors of cultivated peanut needed to be estimated. Therefore, we aimed to obtain orthologous sequences from *A. duranensis* and *A. ipaënsis*, and for other legumes for which the dates of evolutionary divergence are well established, the two components of the soybean genome, *L. japonicus* and *Medicago truncatula*. Sequences from soybean, *Lotus* and *Medicago* could be obtained from their genome sequence databases (see below). Sequences from *Arachis* were obtained from the databases where they were available and supplemented by experimentally obtained sequences. The *Arachis* sequences used were (1) microsomal oleate desaturase (*FAD2*) gene sequences from *A. duranensis* and *A. ipaënsis* deposited in Genbank (numbers 8980834 and 8980832), (2) partial gene sequences for the single copy gene encoding subunit A of DNA gyrase (Leg128; Fredslund et al. 2006; Bertioli et al. 2009) derived from selected BAC clones (clone numbers Adura68E04, Aipa147A20; Guimarães et al. 2008), (3) PCR amplified intron-based sequences from "anchor markers" (Fredslund et al. 2006; Bertioli et al. 2009).

After initial tests, four pairs of intron-amplifying primers were selected:

1. Leg083-fwd GGATCTGGGAAKGTTGGAARATGGA
2. Leg083-rev TCCCAACCATGTYTCTCTGCAAAT
3. Leg088-fwd GCTGCTGTTGGGCAAGATTGTGCTC

4. Leg088-rev GTATTGAGRTTGATTCCCATGACGC TCATG
5. Leg237-fwd ACTTGTTAACATCWCAAARCAGC GG
6. Leg237-rev ACTGGTTCACGTTCAATYGAGAGT GCAGTCCCAAG
7. Leg242-fwd GGARCATAACTATCVTGGTTCTAR TAAGC
8. Leg242-rev CACATGATGAACTGAAAMCCCCC TYGCATGCAC

PCRs were carried out with 25 ng of genomic DNA, 5 U of Taq DNA polymerase (Amersham Biosciences), 1× PCR buffer (200 mM Tris pH 8.4, 500 mM KCl), 1.5 mM $MgCl_2$, 200 µM of each dNTP, and 0.4 µM of each primer, in a final reaction volume of 50 µl. Thermocycling was as follows: 32 cycles of 30 s at 96°C; 45 s at 57°C, 48°C, 48°C or 45°C (annealing temperatures for Leg083, Leg088, Leg237 and Leg242, respectively); 1 min at 72°C, and a final extension for 10 min at 72°C. PCR products were analysed by electrophoresis on polyacrylamide gels stained with silver nitrate (Creste et al. 2001). Sequencing was performed directly from the PCR products essentially as described above. Sequences were processed using the Staden Package, with base calling using Phred (Ewing and Green 1998; Staden 1996). All single nucleotide polymorphisms between the *Arachis* species were confirmed by manual inspection.

In order to obtain orthologous sequences from the sequenced genomes, *Arachis* sequences were used in BLAST similarity searches (Altschul et al. 1990) against the genomes of *Lotus japonicus*, *Medicago truncatula* and *Glycine max*. Sequences were cropped and aligned using Spin from the Staden package (Staden 1996), Muscle (Edgar 2004), Jalview (Waterhouse et al. 2009) with manual editing using Seaview (Gouy et al. 2010). Alignments were analysed using MEGA4. Pairwise aligned sequences of a *c.* 540 bp region of Leg128 were compared for similar spans to generate dot-plots using Spin.

Results

BAC selection and sequencing and annotation

Probing of the *A. duranensis* BAC library filters with the radioactively labelled resistance gene analogue RGA S1_A_36 revealed a strong signal from a pair of spots both representing the BAC clone AD25F09, which harboured an insert of about 110 kb.

Shotgun sequencing and assembly generated eight contigs spanning together about 102 kb. Analysis of these contig sequences predicted the presence of 20 ORFs, five of them related to TIR (toll/interleukin-1 receptor) NBS (nucleotide-binding site)-, LRR (leucine-rich repeat (LRR)-resistance gene analogs, and four to transposable elements. In the largest contig, AD25F09-PO-F10.FM (27,567 bp) (Online Resource 2), the presence of a complete Ty1-*copia* retrotransposon was predicted. This retrotransposon, which we named *Matita*, is 6,179 kb long and has LTRs of 283 (5′LTR) and 236 bp (3′LTR). Its molecular structure is shown in Fig. 1a. The 5′LTR includes a $(CT)_9$ microsatellite, which at the corresponding site of the 3′LTR is $(CT)_5$. The difference in length between the two LTRs is mainly due to an additional 44 bp stretch at the end of the 5′LTR. Between both LTRs, 0.092 nucleotide substitutions per site were calculated. The internal sequence between the LTRs is scattered by many stop codons and frameshifts. A BLASTx search revealed high similarity between the translated region between 1,360 and 4,830 bp and the gag-pol region of several Ty1-*copia* elements. The conserved core domains of integrase (INT) and RT were identified in different reading frames at regions 1,907-2,251 (INT) and 3,338-4,075 (RT). The INT domain is interrupted by two stop codons.

The first start codon is located 61 bp after the end of the 5′LTR. It opens up a region of about 760 bp with several stop codons and frameshifts, which has BLASTx similarity to zinc knuckle containing proteins. An RNA-binding zinc finger motif $Cx_2Cx_3\mathit{G}Hx_4C$ is encoded at position 970-1011, if one considers a G–A mutation in the second bp, which was identified in comparison with highly similar BES from *A. duranensis*. This motif is a characteristic feature of the nucleocapsid of the Gag region in general (Vogt 1997) and the presence of a conserved glycine residue within this motif is a further characteristic of the Pseudoviridae (Peterson-Burch and Voytas 2002). Further conserved domains, for example, of the protease (PR), which marks the beginning of the Pol region, or RNase H (RH) could not be identified in the sequences available. The 1,110-bp region between position 4,830 and the beginning of the 3′LTR does not show any significant similarity in BLASTx search and therefore is designated the 3′untranslated region (3′UTR).

Based on the described *Matita* sequence, three full-length copies could be identified in contigs derived from an ongoing *A. hypogaea* BAC library sequencing project: *Matita*-Ahf303L13_c, *Matita*-Ahf74N12-Contig23_c, and *Matita*-Ahf85I19-Contig81_c. Only the first two elements exhibit defined LTRs, whereas the central sequence of *Matita*-Ahf85I19-Contig81_c is flanked by DNA of low complexity and it is not possible to identify LTRs. The annotated sequences are deposited in Online Resource 3. As shown in Fig. 1b, the elements in general are very similar with regard to their internal coding regions, where the nucleotide identities are well above 80% over a length

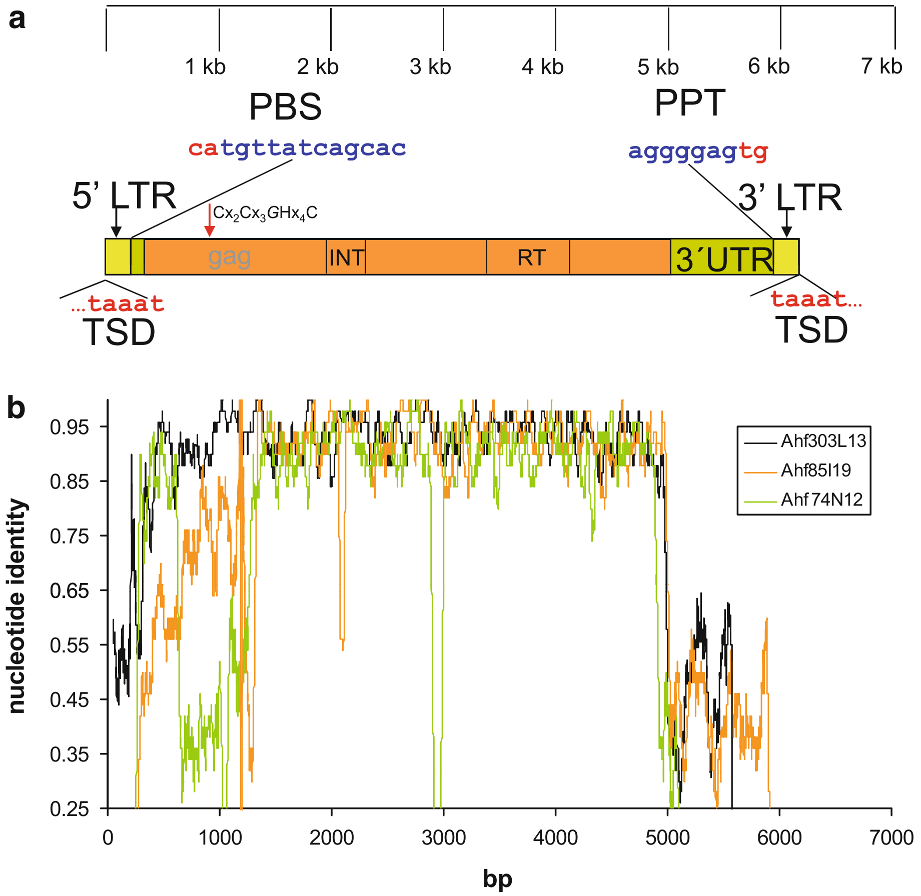

Fig. 1 Structure and composition of the Ty1-*copia* retrotransposon *Matita*, isolated from the *A. duranensis* BAC clone AD25F09. **a** Structural organization: *TSD* target site duplication; *LTR* long terminal repeat; *PBS* primer binding site; *PPT* polypurine tract used for synthesis of the second (+) DNA strand; *GAG* structural proteins for virion core*; INT* integrase (endonuclease for integration in host genome); *RT* reverse transcriptase; *UTR* untranslated region. **b** Analysis of nucleotide identity between three *Matita*-like copies identified in *A. hypogaea* BAC sequences and the initial *Matita* copy from BAC AD25F09

of 95% (*Matita*-Ahf303L13), 81% (*Matita*-Ahf85I19) and 77% (*Matita*-Ahf74N12) of the coding region. A significant drop of sequence similarity, however, is visible in the UTRs and there is also much lower similarity between the LTRs of the different *Matita* copies. According to the definition of Wicker et al. (2007), these data show that *Matita* and the elements isolated from *A. hypogaea* BAC clones are members of the same family of retrotransposons.

Chromosomal distribution

The distribution of *Matita* on the chromosomes was analysed using FISH of *A. hypogaea* metaphases. The advantage of using the allotetraploid is the possibility of observing both genome components A and B in the same experiment. Centromeric DAPI-bands are distinctive in the A-genome chromosomes, allowing both genome components to be differentiated (Seijo et al. 2004). Figure 2a shows an *A. hypogaea* metaphase with 40 chromosomes, 20 of them with strong centromeric DAPI-bands. The smallest pair with the most distinctive DAPI-band is chromosomes pair A9. In order to enable detection of non-clustered copies of *Matita* we have selected as probes a mixture of seven subclones of AD25F09, which together represent the total length of the element, thus using an effective probe size that exceeds the conventional FISH detection limit. In Fig. 2b the distribution pattern of Matia is visible as green fluorescence signal, whereas the red signals represent 18S-5.8S-25S rDNA sites. Some of the chromosome pairs could be assigned according to the presence of the 18S-5.8S-25S rDNA sites, which have been mapped earlier in *A. hypogaea* (Seijo et al. 2004). By this, the SAT chromosome pair A10 with its satellites located remotely from the proximal arm segments could be clearly identified, as well as the corresponding pair B10, which does not exhibit an active 18S-5.8S-25S rDNA site in the cultivated tetraploid. The retrotransposon *Matita* is present on all chromosome pairs of the A- and B-genome. It is mainly located in the distal arm-regions, whereas centromeric regions do not exhibit significant signals of *Matita*. The pattern of its distribution is relatively distinctive with clear and distinguishable pairs of dots. In contrast to the Ty3-*gypsy* element FIDEL, *Matita* is also present on the satellites of chromosome pair A10 and on the small pair A9. Together with the results from rDNA mapping and considering the size of the chromosomes, the *Matita* pattern allowed identification of the individual chromosome pairs and construction of a karyotype of *A. hypogaea*, which is shown in Fig. 2c.

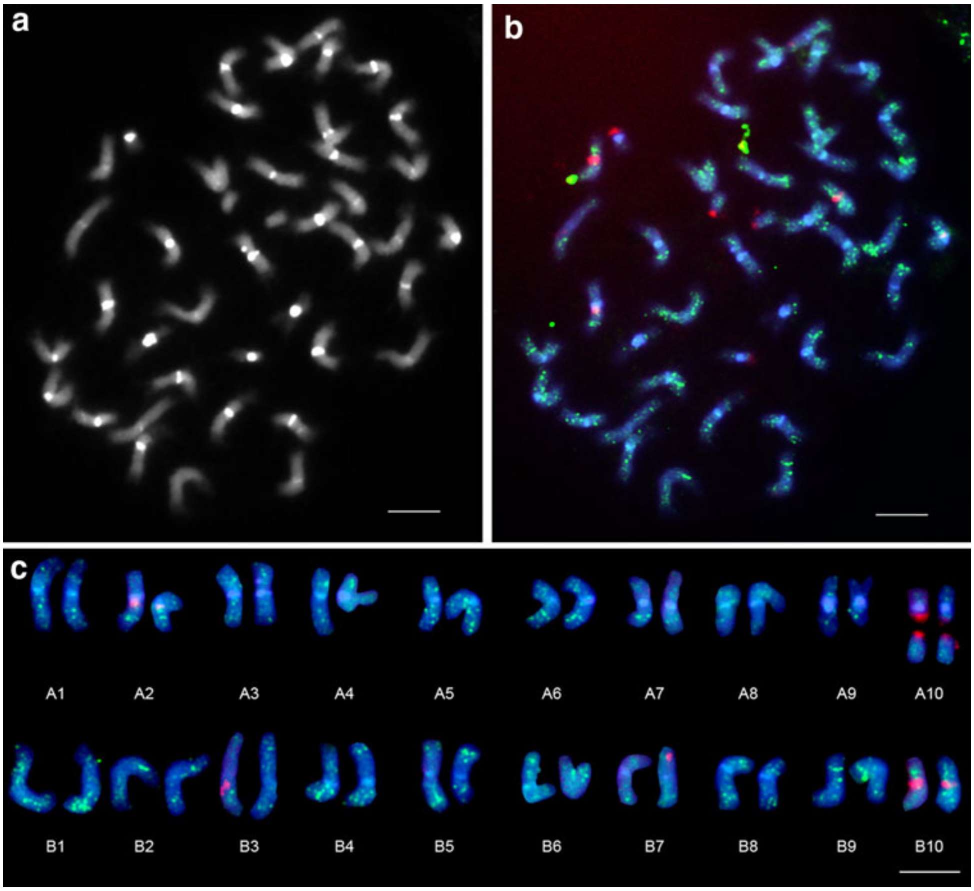

Fig. 2 Cytogenetic analysis of *Matita* using FISH. Metaphase spreads of *A. hypogaea* were probed with a mixture of seven subclones spanning the whole *Matita* sequence and with the rDNA clone pTa71. **a** DAPI counterstain showing 40 chromosomes, 20 of which with strong centromeric DAPI bands, typical for *Arachis* A-genome chromosomes. **b** FISH showing discrete chromosomal distribution of *Matita* (*green*, *light*) and the 18S-5.8S-25S sites (*red*, *medium grey*). **c** Karyogram of **b** based on chromosome size, 18S-5.8S-25S sites, and individual distribution pattern of *Matita*. *Scale bar* 5 μm (colour figure online)

Copy number estimation and analysis of distribution of *Matita*

BLAST similarity searches using the 13 equal-sized fragments from *Matita* against *A. duranensis* BES gave different numbers of significant similarities depending upon the region of the *Matita* element used as query. In the LTR region and the 3'UTR, which is expected to be the most variable region, no hits were found in the database. Five query regions spanning the more conserved gag-pol region, gave an average of 9.8 hits which, applying the formula of Zhang and Wessler (2004) is equivalent to a copy number of 260 (±74) for the diploid genome of *A. duranensis.* When searching against the less representative BES database of *A. ipaënsis*, which is about 11 times smaller than above database, only one hit was found in the gag-pol region. This, however, results in an equivalent of 280 copies (data not shown). Using FISH with *Matita* as probe confirmed that intensities of signals from the A and B component genomes of *A. hypogaea* were similar, therefore, the number of copies of *Matita* in the tetraploid genome was estimated to be about 520 (Table 1).

Overgos probes were used to detect the frequency of *Matita* elements in the BAC libraries. The *A. duranensis* and *A. hypogaea* BAC libraries had 84,096 and 182,784 clones, and represent estimated 7.4× and 6.4× genome coverage, respectively (Guimarães et al. 2008; Yüksel and Paterson 2005). Probing these libraries with overgos from the *Matita* element gave 2,141 and 3,300 positive clones, respectively, corresponding to estimated copy numbers of 289 and 516 for the genomes of *A. duranensis* and *A. hypogaea*, respectively, numbers that agree very well with the values obtained using the BES method described above.

By means of data that were derived from the previous hybridizations of overgo probes to the same BAC libraries, we have investigated the distribution of *Matita* within the genome relative to single-copy genes and RGAs. In these experiments, the *A. duranensis* library had been probed with low- or single-copy gene sequences (Choi et al. 2006; and unpublished data) and the *A. hypogaea* library with RGA sequences (Yüksel et al. 2005). The datasets from *Matita* and these two different probe sets were cross-referenced and the probability of *Matita* being distributed randomly or not was calculated. The original numbers of the calculations are listed in Table 2. The overall frequency of *Matita* containing BAC clones in the whole *A. duranensis* library was 2.5%. However, in BACs that harboured single/low-copy genes, this frequency fell to 1.3%. On the other hand, the overall frequency of

Table 1 Copy numbers of *Matita* based on a BLAST search (1E−100) of the LTR region and 500 bp units of the element against the database of 41,856 *A. duranensis* BAC end-sequences with a total length of 28,615,439 bp (genome coverage = 0.0218)

Region (bp)	Functional region	Hits (1E−100)[a]	Copies[b]
0–283	LTR	0	–
284–784		0	–
785–1,285	INT, RT	0	–
1,286–1,786		9	239
1,787–2,287		13	345
2,288–2,788		9	239
2,789–3,289		6	159
3,290–3,790		12	318
3,791–4,291		3	79
4,292–4,792		1	26
4,793–5,293	3′UTR	0	–
5,294–5,794		0	–
5,795–5,985		0	–

The average copy number of the most conserved region between 785 and 3,790 bp, which included INT, and RT, is 260 (±74)

[a] Number of similar sequences detected using BLAST with *E* value ≤1E−100

[b] Estimated number of copies based on the formula of Zhang and Wessler (2004)

Matita containing BAC clones in the whole *A. hypogaea* library was 1.9%. However, the frequency in BAC clones that harbour RGAs was 3.7%. Using the two-tailed cumulative binomial distribution, the *P* values for null hypothesis of random distribution in both cases is much less than the significance level of 0.05. These results indicate that *Matita* elements are not distributed evenly in the genome. They are less likely to be present near single-copy genes, and more likely to be present near resistance genes.

Phylogenetic analysis

Aiming at the classification of *Matita* within the evolutionary lineages of *copia* elements, a phylogenetic tree based on aligned conserved RT sequences of elements from Triticeae and other monocots, *Arabidopsis*, and the legumes *Glycine*, *Lotus, Medicago, Phaseolus and Vigna* was constructed. The tree consisting of 57 sequences with about 170 amino acids each was inferred using the maximum parsimony method. The resulting bootstrap consensus tree (Fig. 3), which was based on the previous approaches of Wicker and Keller (2007) and Du et al. (2010), resembled the presence of seven *copia* lineages, as predicted by latter authors. *Matita* is located within the *Bianca* lineage, which is supported by a bootstrap value of 100, together with Mtr13.1 from *Medicago*, an unnamed element from *Lotus* (Lj AP004976 40747 58043), *Bianca* (*Hordeum vulgare*), and *copia* elements from *Sorghum* and maize. In this clade, the elements from legumes to grasses are separated in sub-branches. The branching pattern shows that the elements from grasses are more closely related to each other than to the ones from legumes. Furthermore, *Matita* together with the elements from grasses are separated from, but equally related to the elements from *Medicago* and *Lotus*. In spite of having structural differences to *Bianca*, e.g. the lack of an 5′ ORF2 upstream the integrase gene, and the presence of a 1,110-bp 3′UTR, several characteristics support the evolutionary relation between these elements, such as the overall size, the size of the LTRs, and the copy number, which in both cases is moderate. Furthermore, aligned sequences of PBS and PPT from *Matita,* Mtr13.1, and the analogous consensus sequences of the *copia* lineages analysed by Wicker and Keller (2007) show a high degree on similarity between *Matita* and Mtr13.1 and regarding the PPT also between *Matita*, Mtr13.1, and the *Bianca* lineage (Online Resource 4). A comparison of

Table 2 Calculation of the probability of *Matita* being randomly distributed in the genome or in vicinity of resistance gene analogues or single-copy genes

	A. duranensis BAC library		*A. hypogaea* BAC library
Total no. of BAC clones	84,096		182,784
No. of *Matita* pos. BACs	2,141		3,300
Single-copy pos. BACs	26,854	RGA pos. BACs	700
Single-copy and *Matita* pos. BACs	350	RGA and *Matita* pos. BACs	26
Frequency *Matita* in single-copy BACs	0.0130	Frequency *Matita* in RGA BACs	0.0371
Expected frequency	0.0255	Expected frequency	0.0181
P value for the null hypothesis of random distribution	0.0000E+00	P null hypothesis of random distribution	0.00051

The analysis is based on hybridization of overgo probes representing the RT of *Matita* to the *A. duranensis* and *A. hypogaea* BAC libraries. The results were cross-references with data from probing the *A. hypogaea* BAC library with RGA sequences (Yüksel et al. 2005) and the *A. duranensis* BAC library with low- or single-copy gene sequences (Choi et al. 2006; and unpublished data). Using the two-tailed cumulative binomial distribution (in Excel), the *P* value for the null hypothesis of random distribution of *Matita* was calculated aligned sequences of the *Bianca* clade shows an average nucleotide identity in the coding region of about 66% between *Matita* and Mtr13.1 and of about 58% between *Matita* and *Bianca* (Online Resource 5). This demonstrates on the one hand that *Matita* is phylogenetically more closely related to the element from *Medicago* than to the one from *Hordeum*, but on the other hand that these elements are not part of the same family and thus, have not been spread among the different genera by horizontal transfer.

In their phylogenetic analysis of plant *copia* elements Du et al. (2010) found no soybean elements belonging to the *Bianca* lineage. This contrasts to the situation in *Medicago* and *Lotus* and, as shown here, to *Arachis* as well. In order to find out if Phaseolid legumes other than soybean do have *Bianca* like elements, a tBLASTn search with the conserved amino acid sequence of *Bianca* as query was done against the Phaseoleae taxon. The search detected in 14 sequences from *Phaseolus vulgaris* including sequences from the *Tpv2* (Garber et al. 1999) and *pva1* families (*E* values *c*.1E−40) and one sequence of *Vigna radiata* (*E* value 1E−40). However, all these sequences showed primary affinities to non-*Bianca* Ty1-*copia* lineages.

The neighbour-joining tree of 67 aligned nucleotide sequences of the *Matita* RT region formed a large single clade containing sequences from *A. duranensis, A. ipaënsis* and *A. hypogaea,* including the sequences from the full-length elements derived from *A. duranensis* and *A. hypogaea* BACs, with no statistically supported substructure (Fig. 4). The average number of base substitutions per site between all aligned sequences was 0.092 ± 0.01, showing that only copies of the *Matita* family have been amplified (the DNA-sequences from the RT region are listed in Online Resource 6). Thus, in contrast to a previous tree based on FIDEL RT-coding sequences (Nielen et al. 2010), species-specific clades were not detected.

Using the LTR divergence method, the estimated age for transposition of the element present in the BAC clone AD25F09 was 3.54 Mya. In case of the copies derived from the *A. hypogaea* BAC library the age was 1.15 Mya for the Ahf303L13 copy and 3.08 Mya for the Ahf74N12 copy. All sequences versus all sequences pairwise Ks values were calculated, and Ks frequency distributions plotted for *A. duranensis* and *A. ipaënsis*. Ks values were grouped in bins of 0.025. These graphs were very similar with peaks around 0.15 (Online Resource 7). The average Ks value for the RT from BAC AD25F09 was 0.126 ± 0.069 suggesting that the insertion event for this copy of *Matita* was fairly typical if somewhat older than the average insertion age for *Matita* elements in the A- and B-genomes of *Arachis.* The average ratio between non-synonymous substitution rates (Ka) and Ks was 0.435 (±0.196), indicating that *Matita* was under functional selective pressure.

Since 5-methyl cytosine can be replaced by thymine at a high frequency during replication, the occurrences of transitions as compared to transversions are higher in methylated DNA sequences (SanMiguel et al. 1998; Ma and Bennetzen 2004; Vitte and Bennetzen 2006). The Ts/Tv ratio in the LTR sequences of the complete copies available was 19.0 for *Matita* AD25F09, 2.33 for *Matita* Ahf303L13, and 4.2 for *Matita* Ahf74N12. This is an indication that at least these three *Matita* copies might be in epigenetic-silenced states.

Cloning and datamining and analysis of sequences for dating the species divergence of A- and B-genomes

Arachis sequences obtained from Leg083 appeared to be from closely related multiple sequences, therefore were not used in further analysis.

High-quality sequences for *A. duranensis* and *A. ipaënsis* were obtained for Leg088, Leg237, Leg242. BLAST similarity searches identified clearly homologous regions to these *Arachis* Leg introns, the BAC-derived Leg128 regions and the *FAD*-2 genes from GenBank as follows: Leg088, single sequences from *L. japonicus* and *M. truncatula*, and a pair of sequences from *G. max*; Leg237, *L. japonicus* and *M. truncatula*, and *G. max*; Leg242, a single sequence from *L. japonicus*, and a pair of sequences from *G. max*; Leg128 region, a single sequence from *L. japonicus* and a pair of sequences from *G. max*; *FAD*-2, single sequences from *L. japonicus* and *M. truncatula*, and a pair of sequences from *G. max.* Satisfactory multi-sequence alignments could be obtained for Leg088, Leg242, Leg128 (coding sequence for a DNA gyrase subunit and two introns) and *FAD*-2, but not for Leg237. Single-copy sequences are expected from the *Arachis* species used, *L. japonicus,* and *M. truncatula* because they are diploid. Pairs of sequences are expected from *G. max* since it is a paleotetraploid, whose genomes diverged 13 Mya (Schmutz et al. 2010). Since homologues of some of the sequences could not be found in *M. truncatula* this species was eliminated from the analysis. Sequence alignments are provided as Online Resource 8.

We have combined the rates of nucleotide substitutions per site in intron regions and silent nucleotide substitutions per synonymous site in coding regions (=Ks values) to calculate average substitution rates (SR in this manuscript). SR values were calculated for the four different species divergences with previously known time estimates, and for the unknown *A. duranensis* and *A. ipaënsis* divergence (Table 3). The Ks values for *FAD-2* between *Lotus* and *Glycine* (Ks = 1.749) and between *Lotus* and *Arachis* (Ks = 1.372) were not considered in the average SRs, because the high Ks suggests a saturation in synonymous sites (Van de Peer 2004). For the former four known

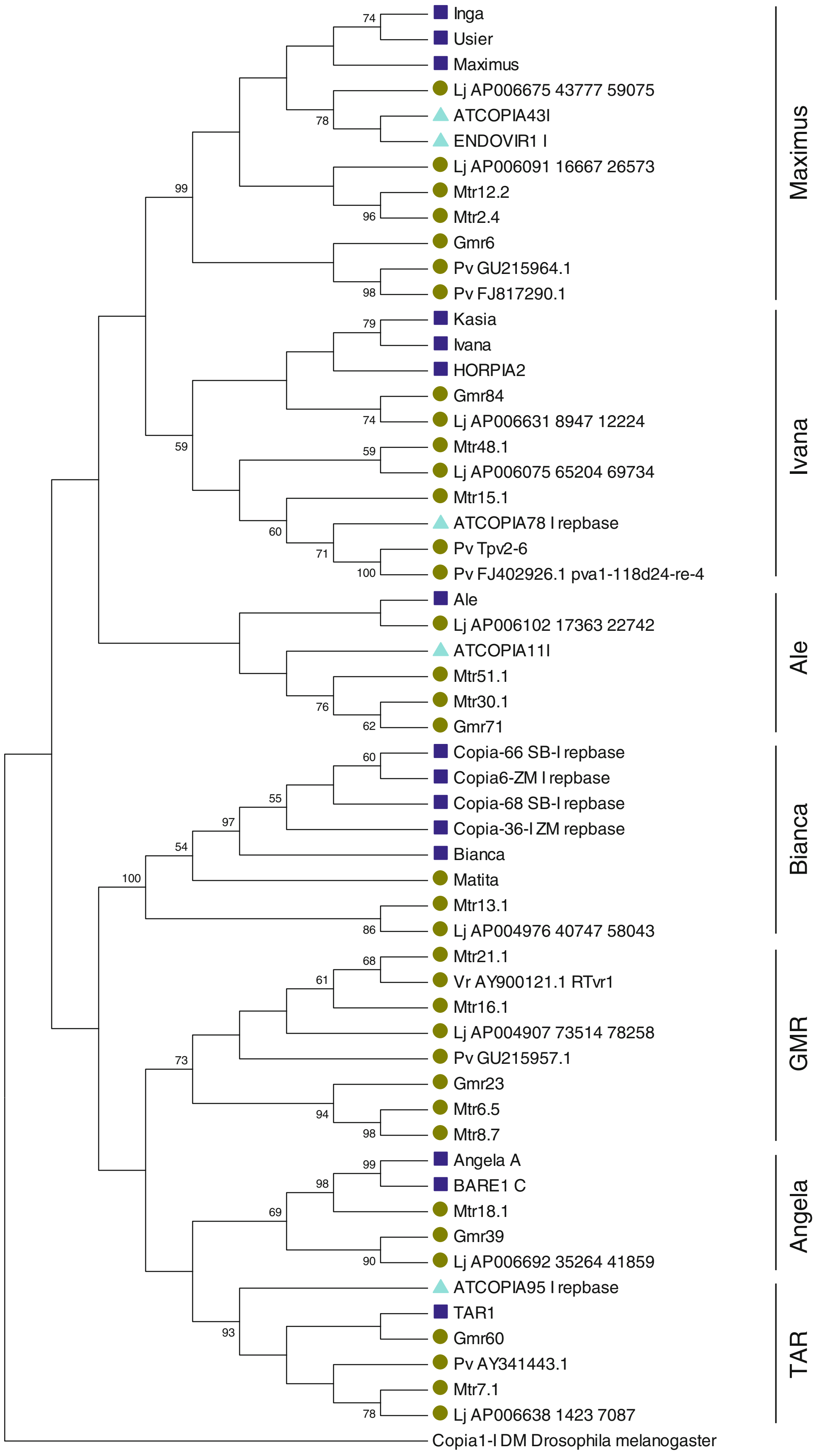

Fig. 3 Maximum parsimony phylogenetic tree of typical *copia* families from Triticeae, Panicoidae (labelled with *squares*), *Arabidopsis* (*triangles*), and Leguminosae including *Matita* (*circles*), based on alignments of amino acid sequences of the RT region. The major evolutionary lineages are indicated by *vertical lines* in accordance with Wicker and Keller (2007) and Du et al. (2010). The elements were selected mainly on the basis of these publications and their sequences were derived from the databases TREP (Triticeae), RepBase (*Arabidopsis*, Panicoidae), SoyTE (*Glycine*), GenBank (*Lotus* (Lj), *Phaseolus* (Pv), and *Vigna* (Vr)—accession numbers as indicated in the figure, and from Wang and Liu (2008) (*Medicago*). The tree was rooted using the RTsequence of *DM* from *Drosophila melanogaster*. Numbers adjacent to branches indicate the percentage values (≥50) from 1,000 bootstrap replicates supporting a particular clade

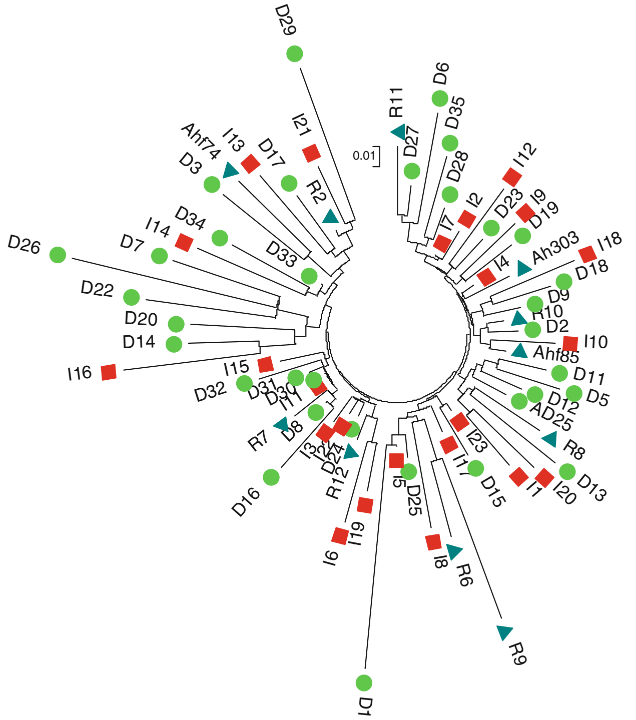

Fig. 4 Neighbour-joining phylogenetic tree of 67 aligned nucleotide sequences of the RT region of *Matita*. The sequences were derived by PCR cloning from DNA of *A. duranensis* (*circles*), *A. ipaënsis* (*squares*) and *A. hypogaea* (*triangles*) using degenerate primers based on the *Matita* copy present in BAC clone AD25F09 and BAC end-sequences. Included are also the RT sequences from *Matita*-AD25F09 and from the three *A. hypogaea* BAC contics *Matita*-Ahf303L13, *Matita*-Ahf85I19, and *Matita*-Ahf74N12

species divergences, the rates of change of SR values were similar, averaging 0.00723 substitutions per site per million years. The observed average SR value between *A. duranensis* and *A. ipaënsis* is 0.0252. The above numbers expressed a decreasing proportion of similar sequence spans with increasing evolutionary distance between the species being compared. This is clearly visible in dot-plots, e.g. of the pairwise-aligned *c*. 540 bp coding and intron sequences of Leg128 (see Online Resource 9). Assuming the substitution rate per time unit (SR/Mya) for the *Arachis* A–*Arachis* B divergence is the same as for the other species divergences, this gives an approximate date of divergence of the *Arachis* species of 3.5 million years (Fig. 5).

Discussion

The sequenced *Matita* elements had accumulated mutations since transposition, but were conserved enough for an adequate annotation. *Matita* has a typical structure for a Ty1-*copia* element, and also other characteristics of the Pseudoviridae. For phylogenetic analysis of the relation of *Matita* to other Ty1-*copia* families, the parsimony method was chosen as it was reported to be the most appropriate approach for the analysis of gene families (Thornton and DeSalle 2000). Phylogenetically *Matita* fits clearly within the *Bianca* clade, one of seven *copia* lineages found in grasses, *Arabidopsis* and legumes (Wicker and Keller 2007; Du et al. 2010; Fig. 3). The latter authors also described the surprising absence of any element in the *Bianca* lineage in the soybean genome. Since the *Bianca* lineage is found in diverse monocots and dicots including the model legumes *Medicago* and *Lotus*, they concluded that the *Bianca* lineage was lost during the evolution of soybean some time after its divergence from *Medicago* and *Lotus* about 50 Mya. This study now shows the presence of the *Bianca*-like *Matita* in *Arachis,* a legume that diverged from soybean, *Medicago* and *Lotus* about 55 Mya (Wojciechowski et al. 2004; Schrire et al. 2005). This adds further weight to the hypothesis of loss within the soybean lineage, because it strongly suggests that the common ancestor of *Arachis*, soybean, *Medicago* and *Lotus* harboured *Bianca*-like elements.

The evolutionary divergence of soybean from *Medicago* and *Lotus* defines the formation of the Warm Season Legumes, a group that includes soybean, common bean, cowpea and pigeon pea. Interestingly, BLAST similarity searches of all available nucleotide sequences from this group did not reveal any *Bianca* clade elements. It seems possible that the *Bianca* lineage was lost from the Warm Season Legumes early in their radiation. At whichever exact stage it happened, the fact that loss did occur emphasizes that the rate of degradation and deletion suffered by retrotransposons is sufficiently high that whole lineages of elements may become extinct even though they are normally present in high copy numbers.

Using FISH, Matita's positions with regard to the individual chromosomes and the two genomes were determined. The distribution patterns were quite distinctive for different chromosome pairs with several cases of distinguishable pairs of dots on the chromosome arms. We have used as probes a mixture of cloned sequences spanning the whole *Matita* element. Therefore, in principle it is possible that FISH signals arose from complete or partial-disrupted elements, and also from solo-LTRs, remnants of old *Matita* elements generated by unequal intrastrand recombination (Devos et al. 2002). However, due to the small size and the low nucleotide identity demonstrated here (Fig. 1b), solo-LTRs might not have been efficiently detected by FISH. *Matita* signals, together with the results from rDNA mapping and chromosome sizes, allowed the identification of chromosome pairs of *A. hypogaea* (Fig. 2c). The development of karyotypes based on classical parameters such as chromosome length, centromeric index, arm ratio, and inter- and intrachromosomal asymmetry index is still used

Table 3 Estimation of the *Arachis* A B genome divergence time

Evolutionary divergence	Subst./site			Ks	Ks	Average (SR)	Divergence (SR/Mya)	Divergence (Mya)
	Leg088 Intron	Leg242 Intron 1 + 2	Leg128 Coding	Leg128	*FAD*-2			
Lotus–Glycine	0.34	*0.224*	0.295	0.3855	(1.749)[a]	0.311125	0.006223	50
Lotus–Arachis	0.424	0.334	0.413	0.3485	(1.372)[a]	0.379875	0.006907	55
Arachis–Glycine	0.375	0.419	0.541	0.4463	0.6445	0.48516	0.008821	55
Glycine–Glycine	0.113	0.043	0.093	0.077	0.127	0.0906	0.006969	13
Arachis–Arachis	0.035	0.012	0.004	0.045	0.030	0.0252	**0.00723**[b]	**3.49**

Substitution rates (SR) for total substitutions in the Leg-introns and silent substitutions of synonymous sites in *FAD*-2 and the Leg128 coding sequence for the four different species divergences with previously known time estimates, and for the unknown divergence of *A. duranensis* and *A. ipaënsis*. Numbers in bold are inferred values

[a] These Ks values were not considered in the average SR, since they indicate saturation in synonymous sites (Van de Peer 2004)

[b] Average of the four known divergences

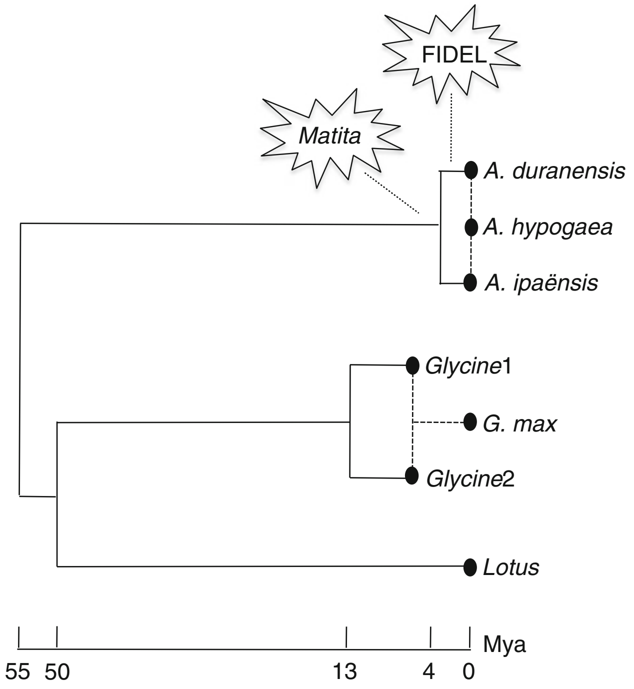

Fig. 5 Schematic phylogeny of *Arachis duranensis*, *A. hypogaea*, *A. ipaënsis*, *Glycine max* and *Lotus japonicus*. With time scale calculated using mutation rates from the legume anchor markers Leg088, Leg242, Leg128 and microsomal oleate desaturase genes. The dates of divergence of the ancestral diploid species of *A. hypogaea* is calculated to be about 3.5 million years ago. *Dashed lines* are used to represent the polyploidy events because their dates are unknown. The most recent datable bursts of transposition for the retrotranspons FIDEL and *Matita* are positioned on the phylogeny. The most recent burst for FIDEL is after the date of divergence of the ancestral species, and accordingly FIDEL is located primarily in the A-genome. The most recent burst for *Matita* predates the divergence of the ancestral species, and therefore *Matita* occurs in roughly equal numbers in both A- and B-genomes

for species and variety characterization in *Arachis* (Inés and Fernández 2004; Lavia et al. 2009). Using FISH with rDNA sequences as probes has added valuable information to cytogenetics in *Arachis* and allowed the analysis of relations among wild species (Robledo et al. 2009). Repetitive sequences such as *Matita,* which are present in a distinctive manner on all chromosomes has the potential to be a useful tool for the identification of individual chromosomes and chromosomal parts. This will also be valuable for the analysis of synthetic amphiploids, their hybrids with cultivated peanut and their progeny, that are produced with the aim of introgressing desired traits from wild species into cultivated peanut (Simpson et al. 1993; Fávero et al. 2006; Foncéka et al. 2009). As to the genomes, the intensities of FISH signals from the A- and B-genomes were not detectably different. This supported copy number estimations, in which two different methods gave very similar estimates for copy numbers of *Matita* in *Arachis*, about 270 in the diploid A-genome and twice as much in the tetraploid genome.

In contrast to *Matita*, other retroelements in allopolyploids such as cotton or tobacco exhibited differential genome-specific amplification (Hawkins et al. 2006; Melayah et al. 2004). This was also the case for an *Arachis* A-genome preferential Ty3-*gypsy* element named FIDEL that we recently described. FIDEL's distribution was due to a period of transposition activity that occurred in the A-, but not the B-genome, most probably after the evolutionary divergence of the two genome donors, but before the polyploidy event that gave rise to the cultivated species (Nielen et al. 2010). We reasoned that *Matita* may be evenly distributed between the two sub-genomes because the most recent major burst of activity pre-dated the evolutionary divergence of the A- and B-donor species.

This hypothesis was tested by a phylogenetic analysis, by estimating the dates of transposition of the *Matita* elements isolated and by estimating the time of evolutionary divergence of the A- and B-donor species. Using the LTR mutational divergence method, we showed that the sequenced copies of *Matita* had transposition dates of about 3.54, 3.08 and 1.15 Mya, and another two elements were too degraded (perhaps too old) to date. To estimate the divergence of peanut's ancestral species a mutational "molecular clock" consisting of orthologous intron and genic sequences was used. This clock was calibrated using the known divergence dates of the subgenomes of soybean (13 Mya), the divergence dates of the phylogenetic clades of *Lotus* (Cool Season Legumes; 50 Mya), soybean (Warm Season Legumes; 50 Mya) and *Arachis* (Dalbergioid Legumes; 55 Mya). By this method, the date of the *Arachis* A–B divergence was estimated at 3.5 Mya, which is comparable with the two older *Matita* transposition dates. Furthermore, using Ks distribution curves and reconstructing a phylogeny from PCR-cloned RT sequences from both diploid and the tetraploid species, it was clear that the evolutionary history of *Matita* was not distinct in the A- and B-genomes. Overall, it seems most likely that the most recent large-scale expansion of *Matita* through the *Arachis* genome occurred in the common ancestor of *A. duranensis* and *A. ipaënsis* around 3.5 Mya (Fig. 5). As such, it may be that most *Matita* elements will be among the oldest of the easily classifiable transposons in the peanut genome because it is thought that most transposons in plant genomes are less than 3 million years old due to elimination by unequal crossing-over, illegitimate recombination and mutational degradation (Vicient et al. 1999; Devos et al. 2002; Pereira 2004).

Our study provides data suggesting that Matita's copy number in *A. hypogaea* is the sum of the copies from *A. ipaensis* and *A. duranensis*. From this point of view, it seems that the genomic shock after spontaneous hybridization of the two donor species, which was dated only about 3,500 years ago (Hammons 1994), did not induce a new very substantial burst of transposition of *Matita*. Reports on other natural allotetraploids such as *Brassica napus*, *B. carinata*, and *B. juncea* have shown that amplification of retrotransposons is not necessarily a consequence of genomic shock (Alix and Heslop-Harrison 2004). In general, the number of retrotransposons being amplified after interspecific hybridization seems to be low (reviewed in Parisod et al. 2010). However, as has been described for the retroelement Wis2-1A in synthetic allopolyploid wheat (Kashkush et al. 2003), it is conceivable that the transcription of *Matita* was activated as an immediate response to the genomic shock, but no, or a few functional copies were generated and inserted. This could be due to post-transcriptional silencing and/or through mutations in the transposon genes (Feschotte et al. 2002). For Wis2-1A it was shown that transcriptional activation of the promoter region located in the LTRs can have an effect on adjacent genes, a scenario, which still remains to be demonstrated for *Matita* or any other retroelement in peanut.

By detecting Ts/Tv ratios >1.5:1 in the LTRs of three *Matita* copies, we have found some indirect indications for epigenetic silencing of *Matita*. However, BLAST searches against the ESTs available for *Arachis* reveal a few *Matita* transcripts, suggesting that some elements can be activated. A dynamic evolutionary process, consisting of partial and whole deletions and occasional transpositions of *Matita* during the divergence of the A- and B-genomes and after the polyploidization event, also would explain why the FISH pattern of *Matita* is not identical between homoeologous chromosomes.

It is a well-known paradox that the part of the eukaryotic genome, to which function is most easily ascribed, the genes usually occupy only a small fraction of the total genome. Most of the genome is occupied by repetitive DNA, which as a consequence, largely defines genome structure (Schnable et al. 2009). The repetitive and genic fractions of the plant genome have very different evolutionary dynamics. Detectable macro synteny in legumes remains even after more than 50 million years of evolutionary divergence (Choi et al. 2004; Bertioli et al. 2009; unpublished data). In contrast, evidence indicates that the repetitive fractions of the genomes of *Arachis* species are fast evolving, being substantially diverged even between closely related species (Seijo et al. 2007; Nielen et al. 2010). An intriguing question is how these distinct genomic fractions are distributed within the genome, are they intercalated more or less at random, or do structural tendencies exist? By probing BAC libraries with overgos representing the *Matita* element, and cross-referencing this data with results from probing with single- or low-copy genes (unpublished data) and resistance gene analogues (a multi-copy gene class; Yüksel et al. 2005) we show that a *Matita* element is about half as likely to be present in a BAC harbouring a single- or low-copy gene as it is in a randomly chosen BAC. In contrast, a *Matita* element is about twice as likely to be present in a BAC that harbours an RGA than it is in a randomly chosen BAC clone. This statistically significant tendency was also reflected in the BAC AD25F09 contigs, which together exhibited the presence of five RGA and four transposon sequences. RGA S1_A_36 from BAC AD25F09 and other RGAs were genetically mapped in the more distal region of linkage groups of the Arachis A-genome. (Leal-Bertioli et al. 2009, unpublished data). Also in other species, such as barley and wheat, RGAs are located in distal chromosomal regions (Madsen et al. 2003; Spielmeyer et al. 2000). These regions have been shown to exhibit higher frequencies of recombination (Dvorak et al. 2004; Ott et al. 2011). Recombination events in turn play an important role in the evolution of resistance genes (Michelmore and Meyers 1998; Hulbert et al. 2001). An association between RGAs and retrotransposons has been documented before in *Medicago*

(Ameline-Torregrosa et al. 2008). The fact that FISH detected *Matita* predominantly in the same chromosomal regions also supports the observed tendency of physical proximity between *Matita* and RGAs. Furthermore, Bertioli et al. (2009) showed in an analysis of synteny of *Arachis* with *Lotus* and *Medicago* that, for the later two species, single-copy genes tend to be more frequent in evolutionary more conserved regions (also called Synteny blocks) and that these regions also tend to be retrotransposon poor. In contrast, evolutionary more variable regions tend to be retrotransposon-rich and often harbour clusters of resistance gene homologues. In soybean, Graham et al. (2002) found retrotransposons being inserted in RGA clusters and, in this context, it is also interesting that a retrotransposon was identified carrying a full-length plant disease-resistance gene belonging to the NBS-LRR family (Wawrzynski et al. 2008). This led to the suggestion that in case of activation of this element by pathogen infection, a link between pathogen infection and creation of new disease-resistance genes could arise. In other species, a linkage between RGA and retrotransposon density has not been observed. In *Vitis vinifera*, Moisy et al. (2008) have compared the genomic distribution of retrotransposons with mapped RGA markers (Di Gaspero et al. 2007) and suggested that their densities on the chromosomes are independent from each other. However, it seems that for various legumes, including *Arachis*, that more evolutionarily conserved genes tend to reside in retrotransposon poor regions and that the faster evolving resistance genes tend to reside with the fast-evolving repetitive fraction of the genome.

Conclusion

After the retrotransposon FIDEL, *Matita* is the second thoroughly characterized LTR retrotransposon in *Arachis*. While *Matita* is a Ty1-*copia* element, FIDEL is a Ty3-*gypsy* element. Both elements are less likely to be present near single- or low-copy genes than near disease-resistance gene homologues. In other aspects they are more distinct. *Matita* is present in much lower copy number and is not more present on one subgenome than the other. It also has a distinct distribution on the chromosomes. These specific characteristics harmonize with the observation that *Matita's* last era of amplification dates back considerably longer than FIDEL's, namely to the time of the divergence of the most probable ancestors of cultivated peanut.

The detailed knowledge of transposon characteristics is giving us insight into genome- and chromosome structure of peanut, which is expected to improve our ability to manipulate and control genes for crop improvement, and is likely to be key for the assembly of the whole genome in a future genome-sequencing project.

Acknowledgments Stephan Nielen would like to thank the National Council for Scientific and Technological Development of Brazil (CNPq) for a post-doctoral fellowship. Bruna Vidigal is grateful for post-graduate grants from the Brazilian Ministry of Education (CAPES—*Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*). David Bertioli would also like to thank CNPq for a productivity fellowship. We also thank the Generation Challenge Program (TLI) for supporting this work. We thank José Valls for providing *Arachis* germplasm.

References

Alix K, Heslop-Harrison JS (2004) The diversity of retroelements in diploid and allotetraploid *Brassica* species. Plant Mol Biol 54:895–909

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

Ameline-Torregrosa C, Wang B-B, O'Bleness MS, Deshpande S, Zhu H, Roe B, Young ND, Cannon SB (2008) Identification and characterization of Nucleotide-Binding Site-Leucine-Rich Repeat genes in the model plant *Medicago truncatula*. Plant Physiol 146:5–21

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

Bertioli DJ, Leal-Bertioli SCM, Lion MB, Santos VL, Pappas G Jr, Cannon SB, Guimarães PM (2003) A large scale analysis of resistance gene homologues in *Arachis*. Mol Genet Genomics 270:34–45

Bertioli D, Moretzsohn M, Madsen LH, Sandal N, Leal-Bertioli SCM, Guimarães PM, Hougaard BK, Fredslund J, Schauser L, Nielsen AM, Sato S, Tabata S, Cannon SB, Stougaard J (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

Burow MD, Simpson CE, Starr JL, Paterson AH (2001) Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.): Broadening the gene pool of a monophyletic polyploid species. Genetics 159:823–837

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

Cannon SB, Ilut D, Farmer AD, Maki SL, May GD et al (2010) Polyploidy did not predate the evolution of nodulation in all legumes. PLoS ONE 5(7):e11630

Choi HK, Mun JH, Jin Kim DJ, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR (2004) Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci USA 101:15289–15294

Choi HK, Luckow MA, Doyle J, Cook DR (2006) Development of nuclear gene-derived molecular markers linked to legume genetic maps. Mol Genet Genomics 276:56–70

Creste S, Tulmann Neto A, Figueira A (2001) Detection of single sequence repeat polymorphisms in denaturing polyacrylamide sequencing gels by silver staining. Plant Mol Biol Rep 19:299–306

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

Di Gaspero G, Cipriani G, Adam-Blondon AF, Testolin R (2007) Linkage maps of grapevine displaying the chromosomal locations of 420 microsatellite markers and 82 markers for R-gene candidates. Theor Appl Genet 114:1249–1263

Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, Shoemaker RC, Ma J (2010) Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J 63:584–598

Dvorak J, Yang ZL, You FM, Luo MC (2004) Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. Genetics 168:665–675

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred II. Error probabilities. Genome Res 8:186–194

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 Sci 46:1546–1552

Fernández A, Krapovickas A (1994) Cromosomas y evolucíon en *Arachis* (Leguminosae). Bonplandia 8:187–220

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, Faye I, Sall MN, Ndoye O, Fávero AP, Bertioli DJ, Glaszmann JC, Courtois B, Rami JF (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

Fredslund J, Madsen LH, Hougaard BK, Nielsen AM, Bertioli D, Sandal N, Stougaard J, Schauser LA (2006) A general pipeline for the development of anchor markers for comparative genomics in plants. BMC Genomics 7:207

Garber K, Bilic I, Tohme J, Bachmair A, Schweizer D, Jantsch V (1999) The Tpv2 family of retrotransposons of *Phaseolus vulgaris*: structure, integration characteristics, and use for phenotypic classification. Plant Mol Biol 39:797–807

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

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224

Graham MA, Marek LF, Shoemaker RC (2002) Organization, expression and evolution of a disease resistance gene cluster in soybean. Genetics 162:1961–1977

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

Guimarães PM, Garsmeur O, Proite K, Leal-Bertioli SCM, Seijo G, Chaine C, Bertioli DJ, D'Hont A (2008) BAC libraries construction from the ancestral diploid genomes of the allotetraploid cultivated peanut. BMC Plant Biol 8:14

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acid S 41:95–98

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 JF, Kim HR, 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

Hougaard BK, Madsen LH, Sandal N, Moretzsohn MC, Fredslund J, Schauser L, Nielsen AM, Rohde T, Sato S, Tabata S, Bertioli DJ, Stougaard J (2008) Legume anchor markers link syntenic regions between *Phaseolus vulgaris*, *Lotus japonicus*, *Medicago truncatula* and *Arachis*. Genetics 179:2299–2312

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

Hulbert SH, Webb CA, Smith SM, Sun Q (2001) Resistance gene complexes: evolution and utilization. Annu Rev Phytopathol 39:285–312

Inés LG, Fernández A (2004) Karyotypic studies in *Arachis hypogaea* L. varieties. Caryologia 57:353–359

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

Kashkush K, Feldman M, Levy AA (2003) Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. Nat Genet 33:102–106

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

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

Lavia GI, Ortiz AM, Fernández A (2009) Karyotypic studies in wild germplasm of *Arachis* (Leguminosae). Genet Resour Crop Evol 56:755–764

Leal-Bertioli SCM, José ACFV, Alves-Freitas DMT, Moretzsohn MC, Guimarães PM, Nielen S, Vidigal B, Pereira RW, Pike J, Fávero AP, Parniske M, Varshney R, Bertioli DJ (2009) Identification of candidate genome regions controlling disease resistance in *Arachis*. BMC Plant Biol 9:112

Leal-Bertioli SCM, de Farias MP, Silva PIT, Guimarães PM, Brasileiro ACM, Bertioli DJ, Guerra de Araújo AC (2010) Ultrastructure of the initial interaction of *Puccinia arachidis* and *Cercosporidium personatum* with leaves of *Arachis hypogaea* and *Arachis stenosperma*. J Phytopathol 158:792–796

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

Madsen LH, Collins NC, Rakwalska M, Backes G, Sandal N, Krusell L, Jensen J, Waterman EH, Jahoor A, Ayliffe M, Pryor AJ, Langridge P, Schulze-Lefert P, Stougaard J (2003) Barley disease resistance gene analogs of the NBS-LRR class: identification and mapping. Mol Genet Genomics 269:150–161

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

Melayah D, Lim KY, Bonnivard E, Chalhoub B, de Borne FD, Mhiri C, Leitch AR, Grandbastien M-A (2004) Distribution of the *Tnt1* retrotransposon family in the amphidiploid tobacco (*Nicotiana tabacum*) and its wild *Nicotiana* relatives. Biol J Linn Soc 82:639–649

Michelmore RW, Meyers BC (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res 8:1113–1130

Milla SR, Isleib TG, Stalker HT (2005) Taxonomic relationships-among *Arachis* sect. *Arachis* species as revealed by AFLP markers. Genome 48:1–11

Moisy C, Garrisonc KE, Meredith CP, Pelsy F (2008) Characterization of ten novel Ty1/*copia*-like retrotransposon families of the grapevine genome. BMC Genomics 9:469

Moretzsohn MC, Leoi L, Proite K, Guimarães PM, Leal-Bertioli SCM, Gimenes MA, Martins WS, Valls JFM, Grattapaglia D, Bertioli DJ (2005) Microsatellite based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). Theor Appl Genet 111:1060–1071

Moretzsohn MC, Barbosa AVG, Alves-Freitas DMT, Teixeira C, Leal-Bertioli SCM, Guimarães PM, Pereira RW, Lopes CR, Cavallari MM, Valls JFM, Bertioli DJ, Gimenes MA (2009) A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. BMC Plant Biol 9:40

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

Nielen S, Campos-Fonseca F, Leal-Bertioli S, Guimarães P, Seijo JG, Town C, Cook D, Arrial R, Bertioli D (2010) FIDEL—a retrovirus-like retrotransposon and its distinct evolutionary histories in the A and B-genome components of cultivated peanut. Chromosom Res 18:227–246

Ott A, Trautschold B, Sandhu D (2011) Using microsatellites to understand the physical distribution of recombination on soybean chromosomes. PLoS ONE 6:e22306

Parisod C, Alix K, Just J, Petit M, Sarilar V, Mhiri C, Ainouche M, Chalhoub B, Grandbastien M-A (2010) Impact of transposable elements in organization and functioning of allopolyploid genomes. New Phytol 186:37–45

Pereira V (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome. Genome Biol 5:R79

Peterson-Burch BD, Voytas DF (2002) Genes of the Pseudoviridae (Ty1/*copia* retrotransposons). Mol Biol Evol 19:1832–1845

Petit M, Guidat C, Daniel J, Denis E, Montoriol E, Lim KY, Kovarik A, Leitch AR, Grandbastien M-A, Mhiri C (2010) Mobilization of retrotransposons in synthetic allotetraploid tobacco. New Phytol 186:135–147

Proite K, Carneiro R, Falcão R, Gomes A, Leal-Bertioli S, Guimarães P, Bertioli D (2008) Post-infection development and histopathology of *Meloidogyne arenaria* race 1 on *Arachis* spp. Plant Pathol 57:974–980

Robledo G, Lavia GI, Seijo G (2009) Species relations among wild *Arachis* species with the A genome as revealed by FISH mapping of rDNA loci and heterochromatin detection. Theor Appl Genet 118:1295–1307

Ross MT, LaBrie T, McPherson J, Stanton VM (1999) Screening large-insert libraries by hybridization. Wiley, New York

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

Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, Watanabe W, Shimizu Y, Horiuchi I, Matsumoto T, Sasaki T, Higo K (2002) RiceGAAS: an automated annotation system, database for rice genome sequence. Nucleic Acids Res 30:98–102

SanMiguel P, 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

Schmidt T, Heslop-Harrison JS (1998) Genomes, genes and junk: the large scale organization of plant chromosomes. Trends Plant Sci 3:195–199

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Myron Peto, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernethy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen H, Wing R, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacy G, Shoemker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

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. Kew Royal Botanic Gardens, Kew, pp 21–54

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

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

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

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

Simpson CE, Starr JL, Nelson SC, Woodard KE, Smith OD (1993) Registration of TxAG6 and TxAG7 peanut germplasm. Crop Sci 33:1418

Smartt J (1990) The groundnut, *Arachis hypogaea* L. In: Smartt J (ed) Grain legumes: evolution and genetic resources. Cambridge University Press, Cambridge, pp 30–84

Spielmeyer W, Moullet O, Laroche A, Lagudah ES (2000) Highly recombinogenic regions at seed storage protein loci on chromosome 1DS of *Aegilops tauschii*, the D-genome donor of wheat. Genetics 155:361–367

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

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

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

Thornton JW, DeSalle R (2000) Gene family evolution and homology: genomics meets phylogenetics. Ann Rev Genomics Hum Genet 1:41–73

Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. Nat Rev Genet 5:752–763

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

Vogt VM (1997) Retroviral virions and genomes. In: Coffin J, Hughes SH, Varmus HE (eds) Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp 27–70

Wang H, Liu J-S (2008) LTR retrotransposon landscape in *Medicago truncatula*: more rapid removal than in rice. BMC Genomics 9:382

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wawrzynski A, Ashfield T, Chen NWG, Mammadov J, Nguyen A, Podicheti R, Cannon SB, Thareau V et al (2008) Replication of nonautonomous retroelements in soybean appears to be both recent and common. Plant Physiol 148:1760–1771

Wicker T, Keller B (2007) Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res 17:1072–1081

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982

Wojciechowski MF, Lavin M, Sanderson MJ (2004) A phylogeny of legumes (Leguminosae) based on analysis of the plastid MatK gene resolves many well-supported subclades within the family. Am J Bot 91:1846–1862

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

Yüksel B, Paterson AH (2005) Construction and characterization of a peanut *Hin*dIII BAC library. Theor Appl Genet 111:630–639

Yüksel B, Estill JC, Schulze SR, Paterson AH (2005) Organization and evolution of resistance gene analogs in peanut. Mol Genet Genomics 274:248–263

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, Crane CF, Price HJ, Stelly DM, Wendel JF, Paterson AH (1998) Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res 8:479–492