Annals of Botany **112**: 545–559, 2013
doi:10.1093/aob/mct128, available online at www.aob.oxfordjournals.org

ANNALS OF BOTANY
Founded 1887

The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its evolutionary divergence from the B genome

David J. Bertioli[1], Bruna Vidigal[1,2], Stephan Nielen[2,†], Milind B. Ratnaparkhe[3,‡], Tae-Ho Lee[3], Soraya C. M. Leal-Bertioli[2], Changsoo Kim[3], Patricia M. Guimarães[2], Guillermo Seijo[4], Trude Schwarzacher[5], Andrew H. Paterson[3], Pat Heslop-Harrison[5] and Ana C. G. Araujo[2,*]

[1]*University of Brasilia, Department of Genetics, Campus Universitário, Brasília DF, Brazil,* [2]*Embrapa Genetic Resources and Biotechnology, Brasilia, DF, Brazil,* [3]*Plant Genome Mapping Laboratory, The University of Georgia, Athens, GA 30605, USA,* [4]*Plant Cytogenetic and Evolution Laboratory, Instituto de Botánica del Nordeste and Faculty of Exact and Natural Sciences, National University of the Northeast, Corrientes, Argentina and* [5]*Department of Biology, University of Leicester, Leicester LE1 7RH, UK*

[†]*Present address: Plant Breeding and Genetics Section, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, International Atomic Energy Agency, Vienna, Austria.*

[‡]*Present address: Directorate of Soybean Research, Indian Council of Agricultural Research, (ICAR), Indore, MP, India.*

** For correspondence. E-mail ana-claudia.guerra@embrapa.br*

Received: 14 December 2012 Revision requested: 22 February 2013 Accepted: 8 April 2013 Published electronically: 4 July 2013

• *Background and Aims* Peanut (*Arachis hypogaea*) is an allotetraploid (AABB-type genome) of recent origin, with a genome of about 2·8 Gb and a high repetitive content. This study reports an analysis of the repetitive component of the peanut A genome using bacterial artificial chromosome (BAC) clones from *A. duranensis*, the most probable A genome donor, and the probable consequences of the activity of these elements since the divergence of the peanut A and B genomes.

• *Methods* The repetitive content of the A genome was analysed by using *A. duranensis* BAC clones as probes for fluorescence *in situ* hybridization (BAC-FISH), and by sequencing and characterization of 12 genomic regions. For the analysis of the evolutionary dynamics, two A genome regions are compared with their B genome homeologues.

• *Key Results* BAC-FISH using 27 *A. duranensis* BAC clones as probes gave dispersed and repetitive DNA characteristic signals, predominantly in interstitial regions of the peanut A chromosomes. The sequences of 14 BAC clones showed complete and truncated copies of ten abundant long terminal repeat (LTR) retrotransposons, characterized here. Almost all dateable transposition events occurred $<3{\cdot}5$ million years ago, the estimated date of the divergence of A and B genomes. The most abundant retrotransposon is Feral, apparently parasitic on the retrotransposon FIDEL, followed by Pipa, also non-autonomous and probably parasitic on a retrotransposon we named Pipoka. The comparison of the A and B genome homeologous regions showed conserved segments of high sequence identity, punctuated by predominantly indel regions without significant similarity.

• *Conclusions* A substantial proportion of the highly repetitive component of the peanut A genome appears to be accounted for by relatively few LTR retrotransposons and their truncated copies or solo LTRs. The most abundant of the retrotransposons are non-autonomous. The activity of these retrotransposons has been a very significant driver of genome evolution since the evolutionary divergence of the A and B genomes.

Key words: *Arachis hypogaea*, *A. duranensis*, peanut, groundnut, BAC-FISH, BAC sequencing, retrotransposons, genome evolution, phylogeny, homeology.

INTRODUCTION

Peanut (*Arachis hypogaea*), also known as groundnut, is a Papilionoid legume originally from South America and is a major food crop, with a world annual production of around 38 Mt (FAOSTAT, http://faostat3.fao.org/home/index.html). It is grown throughout the tropics and sub-tropics, and is most important in Asia and Africa.

Within the Papilionoids, peanut belongs to the Dalbergioids, a clade separated from most other economically important legumes by an estimated 55 million years of evolution. The Dalbergioids are predominantly from the New World tropics (Lavin *et al.*, 2001; Lewis *et al.*, 1995). They have $2n = 20$ as an ancestral chromosome number and most of the extant *Arachis* species have $2n = 2x = 20$ chromosomes, while *A. hypogaea* is an exception in having 40 chromosomes ($2n = 4x = 40$). It is a recent allotetraploid, most probably resulting from the hybridization of two wild species followed by natural chromosome duplication (Halward *et al.*, 1991; Young *et al.*, 1996; Seijo *et al.*, 2004, 2007). The genome of *A. hypogaea* is large, being estimated at 2·8 Gb (Greilhuber, 2005), with a large repetitive fraction of approx.

64 % determined by DNA renaturation kinetics (Dhillon *et al.*, 1980).

Cytogenetic analyses in *A. hypoagaea* have revealed two types of chromosomes: ten pairs of A-type chromosomes, with strongly 4',6-diamidino-2-phenylindole (DAPI)-stained (and hence AT-rich) heterochromatin at the centromeres, including the smallest pair of all chromosomes (Husted, 1936; Smartt *et al.*, 1978), and another ten pairs of chromosomes with more weakly staining centromeric heterochromatin bands, designated B chromosomes (Smartt *et al.*, 1978; Smartt and Stalker, 1982; Seijo *et al.*, 2004; Robledo and Seijo, 2010). Studies comparing the chromosomal heterochromatic banding patterns together with evidence from positions of rDNA clusters (Seijo *et al.*, 2007; Robledo *et al.*, 2009; Robledo and Seijo, 2010) and genomic *in situ* hybridization (GISH) (Seijo *et al.*, 2007) suggest that *A. hypogaea* A chromosomes are similar to those in the wild diploid *A. duranensis* (Krapov. & W.C. Greg.), whilst the peanut B chromosomes are similar to those in the wild diploid *A. ipaënsis* (Krapov. & W.C. Greg.). Other evidence such as species geographic distribution (Robledo *et al.*, 2009; Robledo and Seijo, 2010) and molecular phylogenies (Kochert *et al.*, 1996; Burow *et al.*, 2009; Moretzsohn *et al.*, 2013) corroborates that the most probable A and B genome donors to *A. hypogaea* are *A. duranensis* and *A. ipaënsis*.

During meiosis, *A. hypogaea* chromosome pairing is almost entirely as bivalents (Smartt, 1990), an indication of genetic divergence of the A and B genomes, and potentially genetic control of chromosome pairing. Significant divergence of the repetitive DNA content of the A and B genomes is also indicated by *in situ* hybridization analyses that, using total genomic probes, are able to distinguish the two genomes (Seijo *et al.*, 2007) and show genome preferential distribution of the retroelement FIDEL (Nielen *et al.*, 2010). In contrast, evidence regarding the low copy fraction of the genome, with molecular markers [including gene and expressed sequence tag (EST) sequences], shows high homology: strong colinearity of the genetic maps shows that there have been few structural rearrangements between the A and B genomes, and gene order appears to have changed little during the evolutionary divergence of the A and B genomes (Burow *et al.*, 2001; Bertioli *et al.*, 2009; Moretzsohn *et al.*, 2009; Shirasawa *et al.*, 2013) estimated as 3–3·5 million years ago (Mya) (Nielen *et al.*, 2011; Moretzsohn *et al.*, 2013). Thus evidence points to an intriguing apparent paradox in the evolution of genome structure: the predominant repetitive DNA genome fraction is in evolutionary flux, whilst, at the same time, low copy number DNA is conserved over evolutionary time. Although this paradox has been extensively studied in grasses, it has been little studied in legumes, which are diverged from grasses by about 150 million years.

To study the divergence and evolution of the A and B genome sequences separately and in more detail, bacterial artificial chromosome (BAC) libraries were made for *A. duranensis* and *A. ipaënsis* (Guimarães *et al.*, 2008). Because of the close species relationships, BAC clones from these libraries should serve as very good proxies for the A and B genomes of peanut, respectively. In this study, we report on the use of the *A. duranensis* library to investigate the repetitive component of the A genome of *Arachis*, aiming to understand the evolutionary processes occurring during divergence of the A and B genomes.

MATERIALS AND METHODS

Selection of BAC clones and DNA isolation

For FISH (fluorescence *in situ* hybridization), clones from the *Arachis duranensis* (accession V14167) BAC library (genome A) (Guimarães *et al.*, 2008) were chosen on the basis of hybridization with probes designed from comparative genome markers derived from genes selected as likely to be unique in a diploid legume genome (Choi *et al.*, 2006; Fredslund *et al.*, 2006). In addition, one BAC clone from the same *A. duranensis* BAC library that was sequenced for a previous study (Nielen *et al.*, 2010) was analysed. Isolated DNA from a single cultured colony of each BAC clone was evaluated by length and restriction enzyme sites (*Not*I, *Hae*III, *Bam*HI, *Hin*dIII, *Xba*I and *Apa*LI). For full information on these BAC clones and their genic content, see Table 1, the Results section and the Supplementary Data Table S1.

For the comparison of homeologous sequences in the A and B genomes, *A. duranensis* and *A. ipaënsis* (accession KG30076 – genome B) (Guimarães *et al.*, 2008), and/or *A. hypogaea* ('Florunner') – genome AB (Yüksel and Paterson, 2005) BAC libraries were screened using two genic probes designed based on the sequences of genes encoding DNA gyrase (Leg128; Fredslund *et al.*, 2006) and the gene for the peanut allergen *Ara h1*. Sequences were compared by dot plots andusing the genome browser and annotator Artemis.

Fluorescence in situ *hybridization with BAC clones as probes (BAC-FISH)*

For metaphase spreads, *A. hypogaea* meristem cells from root tips were used. Samples were treated with 8-hydroxyquinoline, fixed in 100 % ethanol, glacial acetic acid (3:1, v/v) and digested with proteolytic enzymes containing cellulose and pectinase (Maluszynska and Heslop-Harrison, 1993; Schwarzacher and Heslop-Harrison, 2000). Meristem cells were isolated from other tissues on a slide and their chromosomes gently spread in 60 % acetic acid under a cover slip. Slides containing at least five complete sets of metaphase chromosomes, well spread and free of cytoplasm, were selected to be used for FISH.

For probe preparation, 200–300 ng of purified and fragmented (around 600 bp long) DNA of each selected *A. duranensis* BAC clone (genome A) was labelled with either digoxigenin-11-dUTP or biotin-11-dUTP (Roche Diagnostics) by random priming using Invitrogen Life Technologies kits (BioPrime Array CGH Genomic Labelling System and BioPrime DNA Labelling System, respectively). For the retroelement Matita, probes were a mixture of seven sub-clones spanning the whole *Matita* sequence (Nielen *et al.*, 2011). For the retroelements FIDEL (Nielen *et al.*, 2010), Feral and Curu (Ty3-*gypsy* elements, the description of which is described in the Results section), probes were obtained from small insert genomic DNA libraries previously produced and sequenced for the isolation of microsatellite markers (Moretzsohn *et al.*, 2005).

Selected slides were pre-treated, hybridized, washed and hybridization sites detected following Schwarzacher and Heslop-Harrison (2000) with minor modifications. Briefly, slides were pre-treated with 100 μg mL^{-1} RNase A and 20 U mL^{-1} pepsin (from porcine stomach mucosa), in 10 mM HCl,

TABLE 1. *Data summarizing the genic and retroelement percentage contents of the* Arachis duranensis *BACs (A genome) sequenced*

A. duranensis BAC clones ID	FIDEL/ Feral	Pipa/ Pipoka	Gordo	Curu	RE128	Mico	Matita	Grilo	All elements	No./length of contigs (bp)	Genic content
ADH180A21	34·4	16·2	18·8						69·4	1/89 966	None
ADH0051I17-83F22	46·9	9·0		6·6					62·4	5/115 680	2 putative, 1 Zn finger (1 comp. marker)
ADH123K13	36·4	16·1	0·3						52·9	2/114 820	1 WD40
ADH177M04	20·2	9·9	1·5	6·7		10·6			48·9	3/90 712	1 putative (1 comp. marker)
ADH179B13	11·1	7·8	14·6	5·1					38·7	6/92 455	3 putative genes
ADH129F24	27·6							8·8	36·4	6/99 171	FAD binding, GPDH, 2 putative
ADH167F07	10·0	19·4		3·4					32·8	9/99 579	1 putative
ADH079023-72J06	13·2	10·1	6·2						29·5	11/141 775	5 diverse functions, 8 putative (3 comp. markers)
ADH25F09		10·4			9·5		6·2		26·1	6/99 839	5 RGAs, 1 putative
ADH068E04					11·0				11·0	1/101 960	8 genes with diverse functions (4 comp. markers)
ADH18B08						7·6			7·6	3/92 084	9 genes diverse functions, 6 putative
ADH035P21							5·9		5·9	5/125 289	5 genes with diverse functions, 8 putative
Average % coverage in all clones	16·8	8·2	3·2	1·7	1·6	1·5	0·9	0·7			
No. of similarities in BAC ends	FIDEL = 88	Pipoka = 94	107	55	40	9	21	6			
	Feral = 124	Pipa = 43									

ADH0051I17-83F22 and ADH079023-72J06 are consensus sequences derived from the overlap of two BAC sequences.
Comp. marker = genome comparative markers which are derived from genes that are likely to be single copy in diploid legume genomes (Choi *et al.*, 2006; Fredslund *et al.*, 2006).

prior to fixation with 4 % (w/v) paraformaldehyde. Hybridization mixtures were prepared with one or two differently labelled probes (approx. 100 ng μL^{-1} per slide) containing 50 % (v/v) formamide, 10 % (v/v) dextran sulfate, 2 × SSC (saline-sodium citrate), 1·25 mM EDTA (ethylene diamine tetra-acetic acid) and 25 ng μL^{-1} salmon sperm DNA. To reduce unspecific hybridization from repetitive elements within the BACs, different concentrations (0·1–4 μg) of unlabelled genomic *A. hypogaea* DNA or C_0t 100 (Zwick *et al.*, 1997) were added and the hybridization mixture was incubated at 37 °C prior to applying to the slide containing denatured peanut chromosomes. Hybridization was carried out for 12–16 h at 37 °C.

Stringent post-hybridization washes were carried out at 85–95 % stringency level as estimated by Schwarzacher and Heslop-Harrison (2000). Hybridization sites were detected using anti-digoxigenin–fluorescein (Fab fragments from sheep; Roche Diagnostics), and/or Alexa Fluor 594-conjugated streptavidin (Life Technologies/Molecular Probes). Chromosomes were then counterstained with DAPI, mounted in anti-fade and observed with a Zeiss Axioscope epifluorescence microscope (Carl Zeiss, Germany), and images were captured with a CCD camera and analysed with Adobe Photoshop CS using only functions, except cropping, that affect the whole image equally.

BAC clone sequencing and assembly

The BAC clone sequencing was performed by shotgun fragmentation and Sanger and/or 454 methods. For Sanger dideoxy sequencing, a total of 768 plasmid sub-clones derived from random shearing from each BAC were sequenced. Sequence assembly was done using CAP3 (Huang and Madan, 1999). Assembled sequences were visualized and manually edited using Consed (Gordon *et al.*, 1998).

Sequencing by the Roche 454GS-FLX System with titanium chemistry was performed by GATC Biotech AG, Konstanz, Germany. Samples were sequenced on a Genome Sequencing FLX Pico-Titer plate device with GS FLX Titanium XLR70 chemistry. Sequence data were produced in Standard Flowgram Format for each read, and assembly was performed using a GS De Novo Assembler (aka Newbler v2·6, the GS FLX System Software) with default parameters.

The sequences in this publication have been deposited in The European Nucleotide Archive under study number ERP002436, project number PRJEB1745 'Exploratory sequencing of wild and cultivated peanut (*Arachis* ssp.) genomes'.

Sequence annotation

For identification of repetitive sequences, dot plots were produced with all BAC sequences vs. all BAC sequences, pairwise comparisons of BAC sequences and comparisons of the BAC sequences with known repetitive elements using the software Gepard (Krumsiek *et al.*, 2007), and also by the software LTR Finder (Xu and Wang, 2007).

For annotation of sequences, a number of publicly available programs were used; FGENESH (Salamov and Solovyev, 2000); hmm search against the pfam A library (Eddy 2011); BLAST (Altschul *et al.*, 1997) against *Arachis* ESTs and 42 000 *A. duranensis* BAC end sequences (genome survey sequences, GSS; Genbank nos FI321525–FI281689); LTR

Finder; and BLAST against local databases of soybean and arabidopsis predicted proteins. Results were visualized in the Genome browser and annotation tool Artemis (Rutherford *et al.*, 2000). To generate entries for Artemis, BLAST was used with '-m 8' option to produce table format output; various outputs from other programs were parsed and converted to GenBank format using Perl (http://www.perl.org/) as necessary. Annotated sequences were exported from Artemis into Excel, edited where necessary, and calculations as to genome coverage and others were made.

To visualize graphically the repetitive content of the BAC clones, a 'repetitive index' based on the number of similarities identified by BLASTN between the genomic sequence and the 42 000 *A. duranensis* BAC end sequences was produced. Parameters used were '-e 1e-20 -m 8'. The tabulated BLAST output was parsed using in-house Perl scripts to produce an index for each DNA base, calculated as follows: repetitive index = log10(N), where N is the number of BLASTN-detected similarities.

Tests of selection on ORFs

Tests of evolutionary selection on coding regions were done using the software Mega 5 (Tamura *et al.*, 2007) and the codon-based Z-test substitution model, based on the numbers of synonymous (dS) and non-synonymous substitutions (dN) per site. The variance of the difference dS – dN was computed using the bootstrap method (500 replicates). Analyses were conducted using the Nei–Gojobori method (Nei and Gojobori, 1986). The analysis initially involved four complete copies (obtained from the genomic sequences analysed) of an open reading frame (ORF) from Pipa (Supplementary Data File S1), a non-autonomous retrotransposon here identified and further described in the Results. The ORF is peculiar: it encodes a protein domain conserved in the different Pipa elements, but has no apparent homologue in the databases. We considered this worthy of further investigation. For a larger analysis, additionalPipa ORF sequences were determined using BLAST from *A. duranensis* BAC end sequences, retrieved and orientated using Perl scripts (Supplementary Data File S2). Muscle (Edgar, 2004) and Jalview (Waterhouse *et al.*, 2009) were used for sequence alignments, and manual editing was done using Seaview (Gouy *et al.*, 2010). The full analysis involved 82 sequences. All ambiguous positions were removed for each sequence pair. There were a total of 591 positions in the final data set. Codon-based tests of purifying, neutral and positive selection were done by averaging over all sequence pairs, and for all pairwise comparisons.

Dating transposition events

Dates of transposition were estimated for full-length long terminal repeat (LTR) retrotransposons by the LTR divergence method 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 of $1{\cdot}3 \times 10^{-8}$ per site per year described by Ma and Bennetzen (2004).

Phylogenetic analysis

For an analysis of the evolutionary relationships of FIDEL and Feral, LTR sequences were obtained from the *A. duranensis* BACs studied here. Also, for this phylogenetic analysis, FIDEL and Feral LTR sequences were extracted from some other available *A. hypogaea* BACs. Sequences were extracted from annotated sequences using the Artemis genome browser, the alignment was performed using Muscle, and the results were inspected and trimmed using Jalview.

Evolutionary analyses were conducted in MEGA5 (Tamura *et al.*, 2011) using the Minimum Evolution method (Rzhetsky and Nei, 1992). The evolutionary distances were computed using the Jukes–Cantor method (Jukes and Cantor, 1969) and are in the units of the number of base substitutions per site. The ME tree was searched using the Close-Neighbor-Interchange algorithm (Nei and Kumar, 2000) at a search level of 0. The Neighbor–Joining algorithm (Saitou and Nei, 1987) was used to generate the initial tree. The analysis involved 47 nucleotide sequences. There were a total of 1509 positions in the final data set.

The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analysed (Felsenstein, 1985). The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches (Felsenstein, 1985). The tree was drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree.

RESULTS

BAC clones and fluorescent in situ *hybridization*

In total, 27 BAC clones from the *A. duranensis* BAC library (A genome) defined to have a high gene content using comparative genome markers (Choi *et al.*, 2006; Fredslund *et al.*, 2006) were selected for use in FISH experiments. Subsequent sequencing analysis of a sub-set of these BACs showed that most, but notably not all, harboured the expected genes (see section 'Sequencing of BAC clones and annotation of repetitive elements').

Depending on the BAC used as a probe, FISH in *A. hypogaea* metaphase chromosomes spreads produced multiple and dispersed hybridization signals on several, but not all, chromosomes (e.g. Fig. 1A–C), despite the use of high concentrations of genomic or C_0t 100 blocking DNA. Double dot signals on a pair or pairs of chromosomes, as expected for FISH with BAC clones containing single or low copy DNA sequences, could not be detected. We therefore conclude that all BACs contain highly repetitive DNA elements, and these were indeed found by the sequence analysis (see below).

Overall, signals using BACs were found with variable strength mostly on proximal and interstitial regions of the chromosome arms of 20 chromosomes of A genome origin (centromeres strongly stained by DAPI). This is consistent with the BACs being derived from the A genome donor of peanut, *A. duranensis*. Centromeric and distal regions were usually excluded from hybridization (Fig. 1A, red signal), although some probes labelled centromeric (Fig 1B, green signal) or distal (not shown) chromosomal regions. In addition to labelling the 20 A genome chromosomes, some BAC probes also showed weaker hybridization signals on the remaining B genome chromosomes, indicating that they contained repetitive elements present in both genomes.

Following the detailed sequence analysis of selected BAC clones (see below) and the identification of different families of retroelements (Table 1, Fig. 2, and Supplementary Data Table S2, File S4), FISH experiments were performed using fragments of the elements Curu (Fig. 1D), Matita (Fig. 1E), FIDEL (Fig. 1F, H) and Feral (Fig. 1G, H). They showed that each retroelement probe has a specific distribution pattern with more or less dispersion, signal strength or co-localization in certain chromosomes or chromosomal regions (compare Fig. 1D, E, F and G). Curu and Feral are predominantly present in the A genome, as is FIDEL (Fig. 1F, H; Nielen *et al.*, 2010), and in contrast to the distribution of Matita, present in both A and B genome chromosomes and including the centromere region (Fig. 1E; Nielen *et al.*, 2012). The distribution patterns of the predominant individual elements (Fig. 1D–G) and the retroelement composition of the BACs (Table 1) explain the FISH signals of the BACs (Fig. 1A–C).

Sequencing of BAC clones and annotation of repetitive elements

About half of the A genome regions sequenced were gene rich, including one resistance gene homologue (RGH) cluster (Table 1). Three *A. duranensis* BACs selected for gene content did not harbour the expected genes; presumably there was an error in the selection procedure, and therefore they were effectively randomly chosen (Table 1). In total, the *A. duranensis* BACs (A genome – predominantly sequenced using 454) spanned 1·26 Mb of unique genome sequence, in 55 contigs, all with an N50 of 55 kb (European Nucleotide Archive accession numbers HF937564-HF937576; http://www.ebi.ac.uk/ena/data/view/HF937564-HF937576). Two BAC sequences from the B genome (one from *A. ipaënsis* and one from the B genome of *A. hypogaea*) that were homeologous to two of the *A. duranensis* genome regions (ADH068E04 to AIPA147A20, and ADH035P21 to AHF417E07) were also analysed.

One of the *A. duranensis* BACs, ADHI8B08 (A genome), was sequenced independently using both random fragmentation and Sanger chemistry with paired end reads, and by the 454 GS FLX titanium method. The two different assemblies were broadly consistent, but with five small regions of inverted sequence relative to each other (Supplementary Data Fig. S1).

A dot plot of the new BAC sequences against the known retrotransposon FIDEL sequence (Nielen *et al.*, 2010) revealed both complete elements and numerous isolated LTRs. Many of the apparent 'solo LTRs' were a similar distance apart, separated by a conserved sequence encoding gag and aspartyl protease (AP) domains but no reverse transcriptase. This suggested the presence of an abundant, novel and thus non-autonomous element, with LTRs and part of the 3'-untranslated region (UTR) very similar to those of FIDEL, but with no significant similarity in the coding regions. Because of its similarity in length and only in the terminal regions, we named this element 'Feral'. It is an incomplete Athila type Ty3-*gypsy* element, most probably parasitic on the autonomous partner FIDEL (a dot plot of FIDEL *vs.* Feral sequences is available in Supplementary Data Fig. S2).

Another abundant LTR element was identified that has an open reading frame at the 3' of the 5'-LTR and coded a protein with no obvious homologies to any described protein. The element appeared to be a non-autonomous retrotransposon, and we named it "Pipa". Although Pipa's autonomous counterpart could not be found in the BACs sequenced for this study, an *A. hypogaea* BAC sequenced for another study showed two complete representatives of an autonomous retrotransposon with significant similarities to "Pipa". We named this autonomous Ty3-*gypsy* element 'Pipoka'. Pipoka encodes gag, AP, reverse transcriptase and retroviral integrase domains. Pipa and Pipoka have sequence similarities in the LTRs and the 3' half of the internal regions. However, they also have very significant differences. Because the coding regions of both Pipa and Pipoka are in the 5' halves of the non-LTR region (regions which have no significant similarity between the elements), the ORF of Pipa does not have any apparent counterpart in Pipoka, and the ORFs of Pipoka do not have any discernible counterparts in Pipa (a plot of Pipa *vs.* Pipoka sequences is available in Supplementary Data Fig. S2).

The next discovered element has distinctive, large LTRs (2337 bp), each with about seven imperfect tandem repeats, with a motif length of 116 bp. The internal region of this retrotransposon (named 'Gordo' here) codes for gag and AP domains, but again with no detectable encoded reverse transcriptase, and so is non-autonomous. Other elements discovered were named 'Curu', a Ty3-*gypsy* retrotransposon with long LTRs (3448 bp); 'RE128', a Ty1-*copia* retrotransposon; 'Mico', a Ty3-*gypsy* element; and 'Grilo', a Ty3-*gypsy* element. Complete and truncated copies of the previously described Matita and FIDEL retrotransposons (Nielen *et al.*, 2010, 2011) were also present in the 12 analysed genomic regions (representatives of the retrotransposons are in Supplementary Data File S4; also for a general overview of the repetitive structure of the A genome, see an image of an all BAC *vs* all BAC plot in Supplementary Data Fig. S5).

In addition to complete retrotransposons, pseudogenes from different classes of transposable elements and retroviruses were present, although they could not be completely characterized. These included one *Cauliflower mosaic virus* family-type, MULE transposon-types and, most frequently, 'retrotransposon-type' sequences found by homology to the Pfam and diverse annotated database sequences. For the most part, these transposon sequences were less repetitive than the elements that were completely characterized.

Evolutionary selection on the ORF in the retrotransposon Pipa

The codon-based Z-test of selection using the ORFs from four complete Pipa element sequences (three sequenced here, plus one from elsewhere) indicated purifying selection ($P = 4{\cdot}5 \times 10^{-9}$). Mindful that this test is best suited for large samples, we data-mined 79 more sequences covering the 3' region of the ORF from *A. duranensis* BAC end sequences. Using this larger sample, there were a total of 591 positions in the final data set. Performing the analysis with averaging over all sequence pairs, the test indicated purifying ($P = 4{\cdot}4 \times 10^{-6}$) and neutral selections ($P = 1 \times 10^{-5}$).

Whilst functional retrotransposons are expected to have a history of purifying selection, retrotransposons inactivated by mutation are expected to have a history of neutral selection. To investigate further, we repeated the tests with pairwise comparisons. At a significance level of $P < 0{\cdot}01$, 32 % of pairwise comparisons were significant for purifying selection whereas

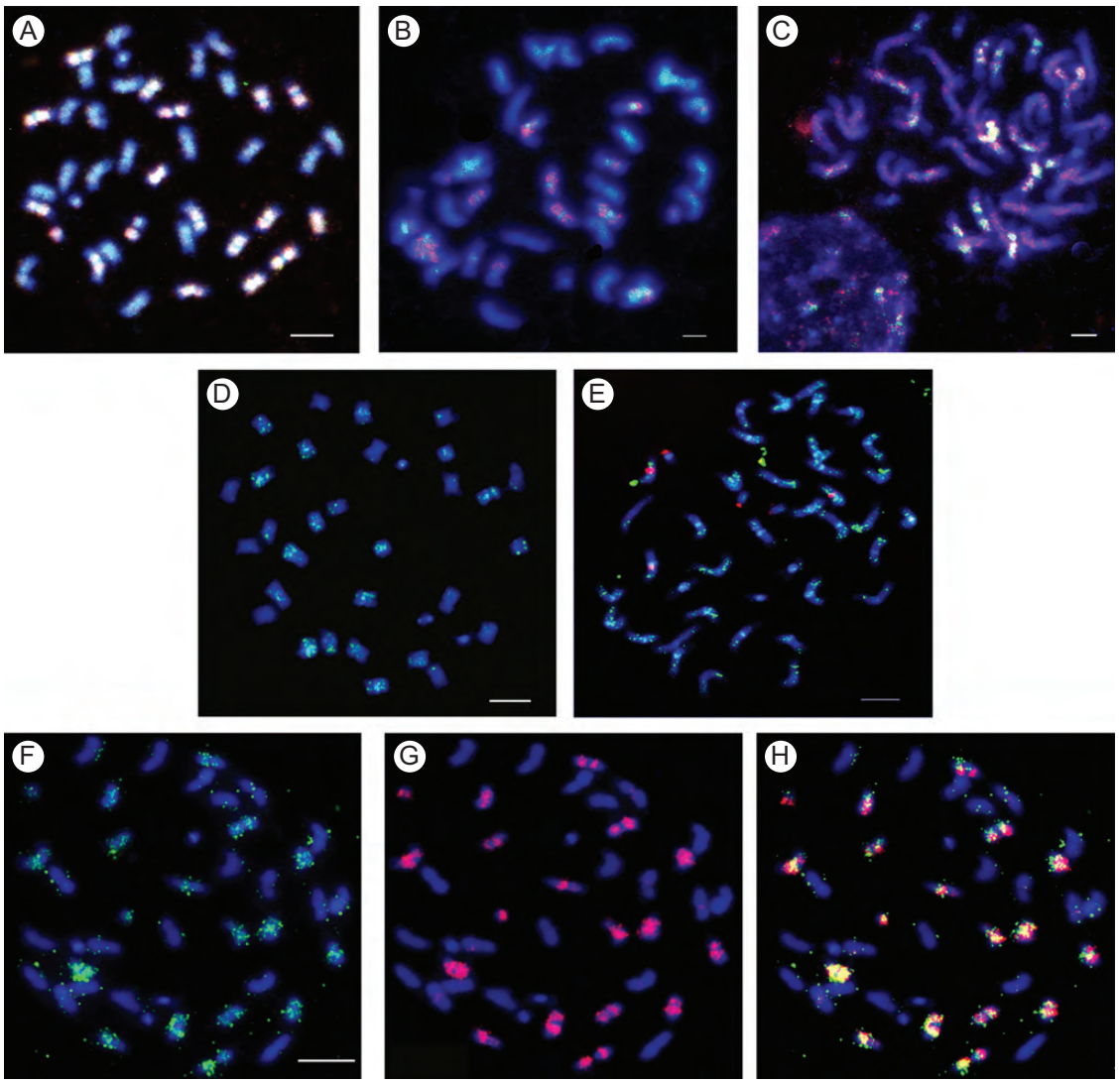

FIG. 1. *In situ* hybridization showing the genomic distribution of sequences from *Arachis duranensis* BACs (A genome) on metaphase chromosomes (stained blue with DAPI) of *A. hypogaea*. Scale bars correspond to 5 μm. (A) Hybridization with the ADH179B13 probe (labelled with digoxigenin and observed in green) shows strong dispersed signals along the entire A genome chromosome arms, and weaker, more dotted signals along the B genome chromosomes arms. The ADH177M04 probe (labelled with biotin and observed in red) shows signals only from the A genome chromosomes, and strongly labels the proximal and interstitial regions. The red and green signals overlap to give the yellowish colour. With both these probes, centromeric and distal parts of chromosomes are not labelled. Overlap of signals ('orangish') is evident in all A chromosomes, presumably because both clones have the retrotransposons FIDEL and Feral (11 and 20 %, respectively); Pipoka and Pipa (8 and 10 %); and Gordo (15 and 1 %) and Curu (5 and 6 %) (Table 1). On the other hand, B chromosomes were only labelled by the ADH179B13 (green) probe. Since the element Gordo was detected in the sequence of ADH179B13 (14 %) and much less in ADH177M04 (1 %), these green signals on B chromosomes could correspond to part of Gordo's distribution. The exclusive element presence in ADH177M04 of the retrotransposon Mico (10 %) could not be differentiated here perhaps due to a dispersed distribution on the A genome which is overwhelmed by other signals. (B) Hybridization with the ADH79023 probe (labelled with digoxigenin and observed in green) shows dispersed and weak signals along the arms of many but not all chromosomes, with some centromeres strongly labelled. The hybridization with the probe ADH51I17 (labelled with biotin and observed in red) is detected in about half the chromosomes of A genome origin in a weak and dispersed pattern excluding the centromeres and with a little overlap to the green signal. This overlap is seen in all A chromosomes as a 'yellowish' colour; presumably it occurs because both clones have FIDEL and Feral elements (13 and 46 %, respectively) and Pipoka and Pipa (10 and 9 %). Interestingly ADH51I17 red signals (pericentromeric region) are similar to the Feral element distribution (compare with G), but probably added to some FIDEL signals. The green signals observed in B chromosomes could correspond to a part of the Gordo element distribution, only detected in the sequence of the ADH79O23 clone (6 %). The exclusive element present in ADH51I17, Curu (6 %), could not be distinguished by FISH, possibly due to a dispersed distribution on the A genome, which is overwhelmed by other signals. (C) Hybridization of the probe ADH129F24 (observed in green) was only detected in A genome chromosomes, with strong signals at the proximal regions. Different chromosomes show different intensities of signals. With ADH167F07 as the probe (observed in red), hybridization sites were detected in A and B chromosomes, with a weak diffuse dotted pattern along all chromosomes arms, but concentrated in proximal and interstitial portions. In the A chromosomes, both probes showed almost the same hybridization pattern. Centromeric and distal regions of chromosomes did not show detectable signals from either of the probes. Overlap of signals ('yellowish') observed in all A chromosomes may be due to FIDEL and Feral elements, present in both ADH129F24 and ADH167F07 clones (27 and 10 %, respectively). The signals at the proximal region with ADH129F24 (green) correspond mostly to the Feral element (compare with G), but also to FIDEL signals, that could be causing the difference in the signal intensity among the A chromosomes. The red signals (ADH167F07 probe) present in all chromosomes but not overlapping the green signals (ADH129F24 probe) might correspond to the presence of Pipa and Pipoka elements, exclusive to clone ADH167F07 with coverage of almost 20 % and lacking in ADH129F24. The element Grilo, only detected in the sequence of the ADH129F24 clone (8 %), and the element Curu (3 %), only present in ADH167F07, were probably overwhelmed by the other signals. (D) Curu: hybridization signals present only in the A genome with strong signals at the

21 % were significant for neutral selection. Sequences and alignments are available in Supplementary Data Files S1–3.

Estimated ages of transposition and abundance of elements

In total, 20 complete LTR retroelements were present in the *A. duranensis* BACs, comprising seven Feral, three FIDEL, three Pipa, two Gordo, two Mico, two Matita and one RE128. The median age of transposition was 1·38 million years (Fig. 3), with only two of the estimated transposition ages older than 3·5 Mya. Complete retrotransposons cover 14·5 % of the total analysed *A. duranensis* genome sequence, almost all of this (14 %) being retrotransposons with an estimated age of insertion of >3·5 million years.

In addition, truncated sequences and solo LTRs were very common; together the ten elements covered >30 % of the BACs analysed (Table 1). The relative coverage of the analysed genome regions of these different elements is consistent with the number of BLAST-detected sequence similarities of the different retrotransposons, with 42 000 *A. duranensis* BAC end sequences with E-values $\leq 1 \times 10^{-40}$ (Table 1).

Observations using the Artemis genome browser with annotations of the genome regions including retrotransposons together with plots of 'repetitive index' indicated that these retrotransposons explained almost all the very highly repetitive DNA content in the 1·29 Mb of sequences analysed here (Fig. 4, Supplementary Data Fig. S3, and ENA sequence accession numbers HF937564–HF937576).

Phylogenetic analysis of Feral and FIDEL LTRs

In total, 60 LTR sequences of Feral/ FIDEL could be retrieved from the *A. duranensis* BAC sequences. Of these, 13 were highly divergent, or too small to be properly aligned, and were removed from the analysis. The 47 remaining LTRs were aligned. Some of these LTRs were from complete retrotransposons or from recognizable fragments, and so could be identified as being LTRs of FIDEL/Feral (Supplementary Data Files S5 and S6). Others were solo LTRs and so could not be assigned. The phylogeny shows that the LTRs of FIDEL and Feral form two related but distinct lines of evolution (Fig. 5).

Annotation of genes

FGENESH predicted numerous genes within and overlapping retrotransposons and their truncated fragments. These predicted genes often encoded Pfam domains with retrotransposon-related functions, but they also often encoded protein regions of no annotated function. FGENESH predicted inappropriate exon/intron structures for these genes that are, apparently, pseudogenes of polyprotein-encoding genes or artefacts of the FGENESH algorithms. For the annotation of BACs, we included pseudogenes in copies of retrotransposons with all their characteristic domains (FIDEL, Feral, Pipoka, Pipa, etc.), but not for fragments of them. Prominent pseudogenes of uncharacterized transposable elements were also annotated. It is notable that there are numerous *Arachis* ESTs with similarity to both the coding and non-coding regions of the described retrotransposons and other repetitive genomic regions, showing their transcriptional activity.

Genes predicted by FGENESH in non-transposable element/ non-repetitive regions were supported by varying amounts of secondary evidence. Some predicted genes were well supported by *Arachis* ESTs, encoded Pfam domains and similarity to predicted genes in arabidopsis and soybean, and therefore could be assigned putative functions. Other genes were annotated as putative proteins. In non-transposable element/non-repetitive genome regions, gene models predicted by FGENESH were used for annotation with, in a few cases, manual editing. For instance, in two cases, separate genes were predicted for a Toll-interleukin receptor domain (TIR)-encoding ORF and for adjacent nucleotide-binding site (NBS)- and leucine-rich repeat (LRR)-encoding ORFs, and the annotation was changed to indicate the TIR–NBS–LRR-encoding genes, corresponding to the well-characterized class of disease resistance genes (Meyers *et al.*, 1999). The annotated sequences are available as ENA sequence accession numbers HF937564–HF937576.

Searching EST data for evidence that FIDEL/Feral have promoted transcription or formed chimeric genes after transposition, hundreds of ESTs with sequence similarity to FIDEL were identified. Of these, four had sequence similarities to non-retrotransposon protein-encoding genes (the EST GenBank numbers are gi-296598828, gi-224930886, gi-207478594 and gi-149648595) and they were similar to: shaggy-related protein kinase; isoflavonoid glucosyltransferase; calcineurin-like phosphoesterase; and chloroplast nucleoid DNA-binding protein. Of these, three were homologous to the 3' end of the FIDEL LTR, suggesting that, indeed, the LTR of Feral/FIDEL can act as an active promotor, at least in some cases.

Comparison of homeologous sequences in the A and B genomes

Three homeologous BAC clones were obtained containing the peanut allergen gene *Ara h1*: two from the *A. hypogaea* library and one from the *A. duranensis* library. One of the sequences from the *A. hypogaea* library was almost identical to the *A. duranensis* clone, and thus A and B genome representatives from the *A. hypogaea* library could be assigned. The assembly for the *A. duranensis* produced a single contig, whereas the assembly for the *A. hypogaea* A genome was fragmented. Therefore, the comparison was based on the A genome from *A. duranensis* and the B genome from the *A. hypogaea* clones (ADH035P21 and AHF417E07, respectively).

A dot plot of the two clones showed regions of microsynteny of approx. 57 kb in AHF417E07 (B genome) and 53 kb in ADH035P21 (A genome). The microsynteny between these regions is delimitated at the 5' and 3' ends by regions with no

proximal regions but excluded from the centromeres and distal regions of the chromosomes. (E) Matita: hybridization sites (green) observed in both A and B chromosomes as dots and bands, mostly in the centromeric and distal regions of chromosomes arms. The hybridization signals in red correspond to the 45S rDNA loci. Photomicrograph published by Nielen *et al.* in *Molecular Genetics and Genomics* (2012) **287**: 21–38. (F) FIDEL: hybridization present mostly in the A genome, with dispersed and dotted signals in the interstitial regions, excluding the centromere and with stronger signals in two pairs of chromosomes. (G) Feral: hybridization signals present only in the A genome, with strong presence in the proximal and interstitial regions, excluding the centromere and distal region. (H) Fidel and Feral hybridization images after overlay. Although the majority of hybridization sites are co-localized, there are signals specific for each of the probes. Note the two pairs of chromosomes with stronger overlapping signals.

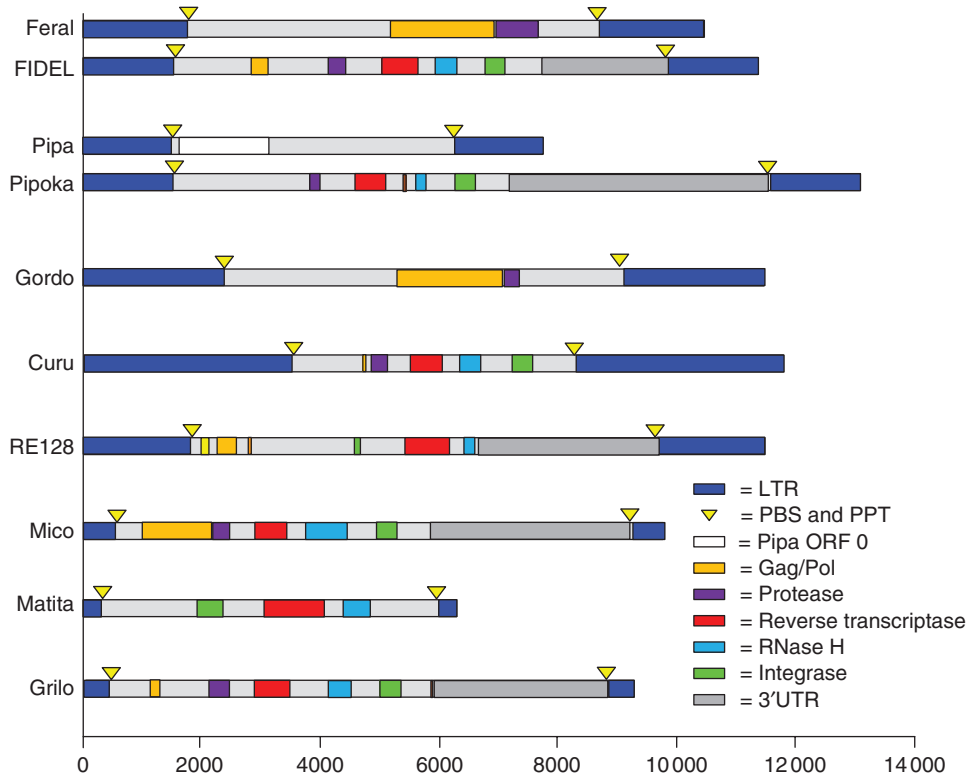

FIG. 2. Schematic diagram of LTR retrotransposons from the *Arachis duranensis* A genome. The elements and their components are drawn to scale. DNA sequences encoding conserved protein domains are colour labelled according to the legend. Pipa ORF 0 encodes a protein of unknown function.

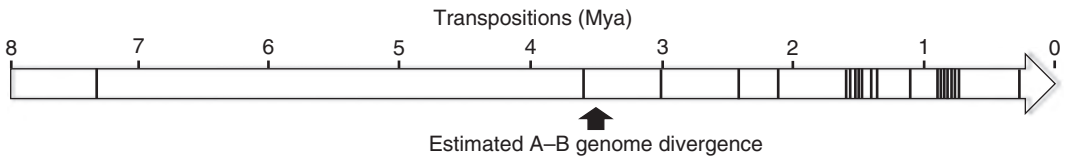

FIG. 3. Timing of the 20 datable transposition events present in the 12 *Arachis duranensis* genomic regions represented in a time line. Vertical lines within the arrow are transposition events. The estimated date of evolutionary divergence of the A and B genomes, about 3·5 million years ago (Mya), is represented by a solid black arrow.

significant sequence similarity between the A and B genomes (Fig. 6). Within this region of microsynteny, there are segments of sequence with very high identity, the longest region of 100 % identity being 290 bp long. These highly similar segments are punctuated by smaller segments (most frequently indels), with no significant sequence similarity. It is notable that the regions delimitating the microsyntenic regions are repetitive in both A and B genomes, but the nature of the repetitive sequences is totally different in the two genomes. At the 5' boundary of the microsyntenic region, the A genome harbours an insertion of the retrotransposon Matita (estimated age of insertion 1·1 million years); at the same boundary, the B genome harbours an insertion of the retrotransposon Mico (estimated age of insertion 3·8 million years). At the 3' boundary of the microsyntenic region, both genomes harbour repetitive DNA sequences that are completely different in nature and difficult to define. In the B genome, the region harbours some fragments of a retrotransposon named 'Yara' (Supplementary Data File S4).

It is notable that the segments without significant sequence similarity within the microsyntenic region also show a tendency to be repetitive. One of the regions, an insertion in the B genome relative to the A genome, is a complete retrotransposon we named 'Joka' (Supplementary Data File S4); it has an estimated insertion date of 423 000 years. Many of the other smaller segments are detectably repetitive, but do not have an obvious origin.

The genic content of the two genomes within the microsyntenic region is predicted to be the same, in the same order and orientation. These predicted genes, encoded, in order from 5' to 3', are: two putative proteins; a transmembrane BT1 family protein; one putative protein; two lipid transfer/seed storage/trypsin-alpha amylase inhibitor proteins; one proteosome protein; and one seed storage Ara h1 protein. It is notable that all the predicted genes do not include detectably repetitive DNA, and are well conserved between the two genomes. The base substitution rates (single nucleotide polymorphism rates) for the genes are 1·7, 2·0, 1·6, 2·9, 6·1, 2·4, 3·3 and 1·8 %, respectively. Base differences accumulated through indels are surprisingly high, at 0, 1·6, 2·1, 23·8, 6·1, 1·6, 20·6 and 11·1 %, respectively.

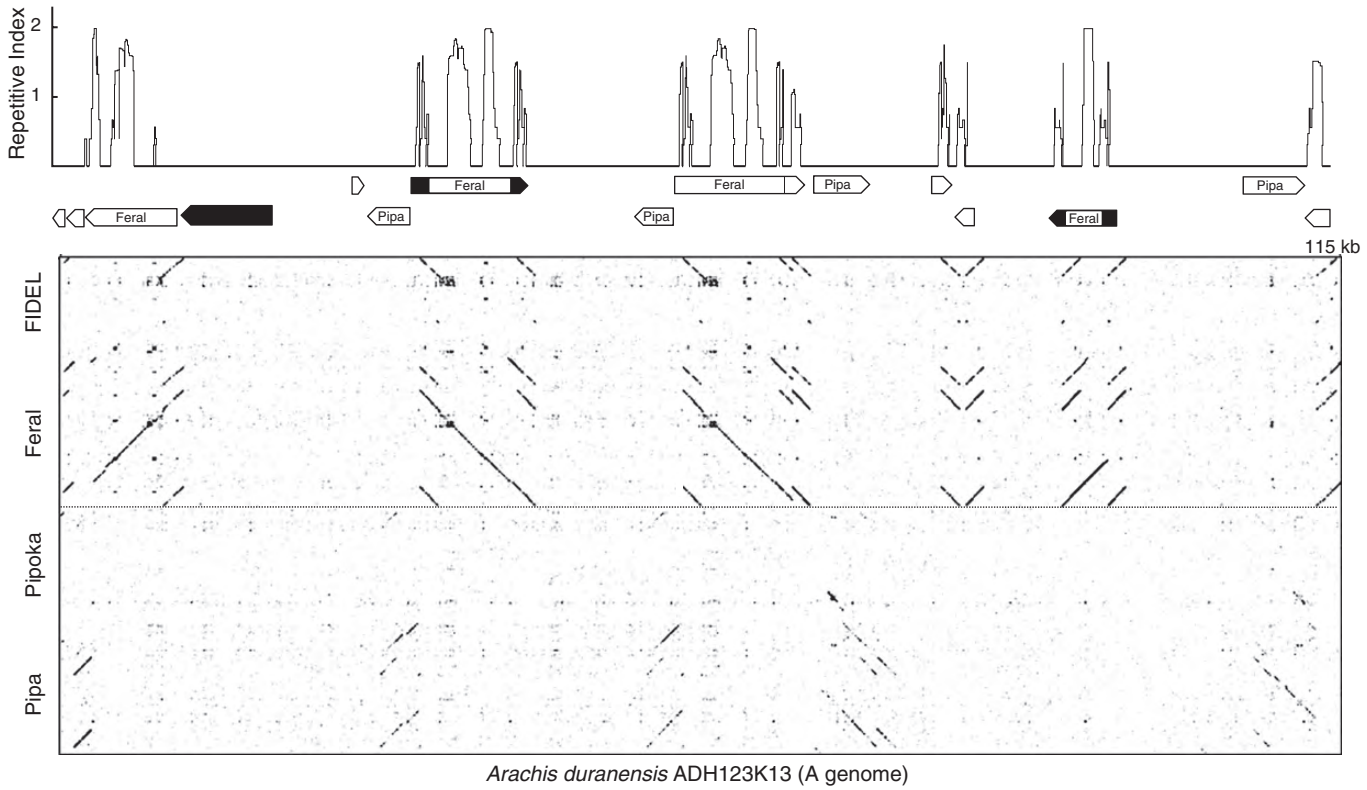

FIG. 4. Representation of genic and repetitive content of one of the *A. duranensis* BAC clones (A genome) analysed, ADH123K13. Top, repetitive index graph; middle, annotation scheme; and bottom, dot plot. The repetitive index is a score for repeat content based on BLASTN against 41 856 *A. duranensis* BAC end sequences. The score is calculated using the formula repetitive index = log10(N), where N is the number of BLASTN homologies with an E-value of ≤1e-20. The highest peak represented here is 2, which is equivalent to 100 BLASTN homologies; the lowest peak represented is equivalent to two BLASTN homologies. The annotation scheme represents complete retrotransposons by white arrows with long terminal repeats (LTRs) in black. The single predicted gene encoding a WD40 domain is represented by a black arrow. The dot plot is of the BAC sequence (horizontal) against whole representative sequences of the retrotransposons FIDEL, Feral, Pipoka and Pipa. More than half the sequence (Table 1, Fig. 2) and all the highly repetitive DNA is accounted for by two retrotransposons. Annotation schemes for the other BACs are available in Supplementary Data.

Also within the dot plot, the patterns of 'granularity' or 'dots' in the background off-axis are worthy of comment. The distribution of the dots can form a granular pattern that apparently represents the presence of short low complexity sequences that have random short matches with low complexity sequences in the other sequence being compared. It is notable that the genes, complete retrotransposons and their truncated fragments 'trace out' clear lanes on the dot plot, and those non-repetitive intergenic sequences tend to trace out a granular path.

The two other homeologous genome regions that could be identified were harboured in the BAC clones *A. duranensis* ADH068E04 (A genome) and *A. ipaënsis* AIPA147A20 (B genome) that contained a DNA gyrase gene (Leg 128, Fredslund *et al.*, 2006). These two regions showed microsynteny over about 43 and 47 kb, respectively (for a dot plot of these BACs see Supplementary Data Fig. S4). The microsyntenic regions were situated at the 5' and 3' ends of the *A. duranensis* and *A. ipaënsis* BAC clones, respectively. The microsyntenic regions consist almost entirely of highly similar sequence segments interrupted by regions with no discernible sequence similarity (most frequently indels). The longest segment of sequence with 100 % identity is 331 bp. There are 15 distinct interruptions, four being detectably repetitive. In one AT-rich region of low complexity, there has been a small sequence inversion and DNA identity has been degraded in a qualitative manner. The 5' border of the microsyntenic region is delimited by repetitive DNA of a completely different nature in the two genomes. In the B genome, the region encodes a fragment of a Mutator transposon protein and, in the opposite orientation, a plant mobile domain protein. In the A genome, the repetitive DNA is not attributable and has no sequence similarity to the homeologous B genome region. The observable 3' border of the microsyntenic region is delimitated by the end of the BAC sequences.

The predicted genes of the microsyntenic regions are almost the same: glycosyl phosphatidyl inositol transamidase; oligosaccharide biosynthesis protein; mitochondrial carrier protein; two fatty acid elongases; and DNA gyrase genes are in common. The sequence of *A. ipaënsis* AIPA147A20 (B genome) contains an extra putative gene before the first fatty acid elongase gene and an extra C2 domain-containing protein before the gyrase gene. In AIPA147A20, the gyrase gene is truncated by the end of the BAC clone. The genic sequences (exons and introns) are highly similar; the base substitution rates (single nucleotide polymorphism rates) for the genes that are in common are 2·3, 1·7, 2·3, 1·7 and 3·4 %, respectively. Nucleotide differences accumulated by indels are 3·9, 1·6, 1·1, 0·9 and 0·1 %, respectively.

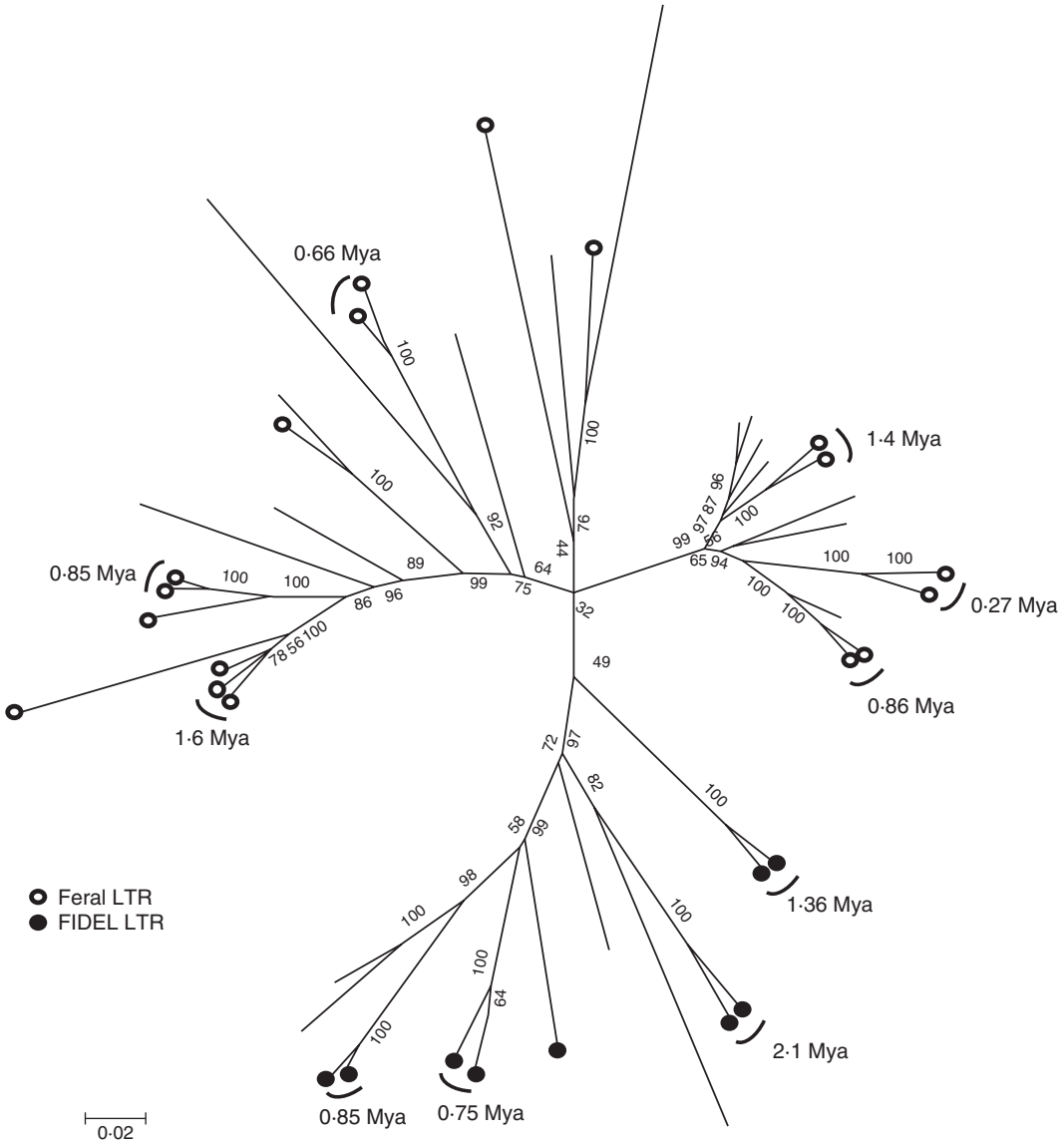

FIG. 5. Reconstructed phylogeny of the retrotransposons FIDEL and Feral using long terminal repeat (LTR) sequences from *A. duranensis* and *A. hypogaea* BACs. FIDEL LTRs are represented by filled circles; Feral LTRs by open circles; branches without circles are of solo or truncated LTRs that could not be assigned. The estimated ages of intact retrotransposons are indicated on their LTR pairs. FIDEL and Feral represent related, but distinct, lines of evolution. Bootstrap values are given in %.

DISCUSSION

It is an intriguing feature of eukaryotic genomes that genes which have the greatest functional significance occupy only a small fraction of the whole; repetitive DNA occupies most of the genome and determines the large-scale structure of the chromosomes (Schmidt and Heslop-Harrison, 1998). The repetitive fraction of plant genomes has perhaps been most studied in members of the Poaceae family, especially the cereal genomes. The importance of retrotransposon activity in the observed variation of cereal genome sizes is now well established (Bennetzen and Kellogg, 1997; Sanmiguel and Bennetzen, 1998; Estep *et al.*, 2012). Retrotransposon transposition results in genome expansion, and their removal by illegitimate or unequal homologous recombination leads to genome contraction. Retrotransposon-driven recombination is also thought to be important in gene duplication and deletion (e.g. Ragupathy and Cloutier, 2008). Although the role of transposable elements in gene creation has been discussed in the literature for more than a decade (Bennetzen, 2000, 2005), concrete examples have been relatively few (e.g. Elrouby and Bureau, 2010), especially considering their extraordinary abundance.

Here we investigated the repetitive DNA of the A genome of the legume peanut. Cultivated peanut is an allotetraploid with an AB-type genome (Smartt *et al.*, 1978; Smartt and Stalker, 1982). In terms of investigation of legume genome structure, peanut is very informative because phylogenetically it is an outgroup to most other economically important legumes (Lavin *et al.*, 2001; Lewis *et al.*, 2005). Also, it is an interesting subject of study because its component A and B genomes are closely related. Previously we have estimated that these component genomes diverged about 3–3·5 Mya (Nielen *el al.*, 2011;

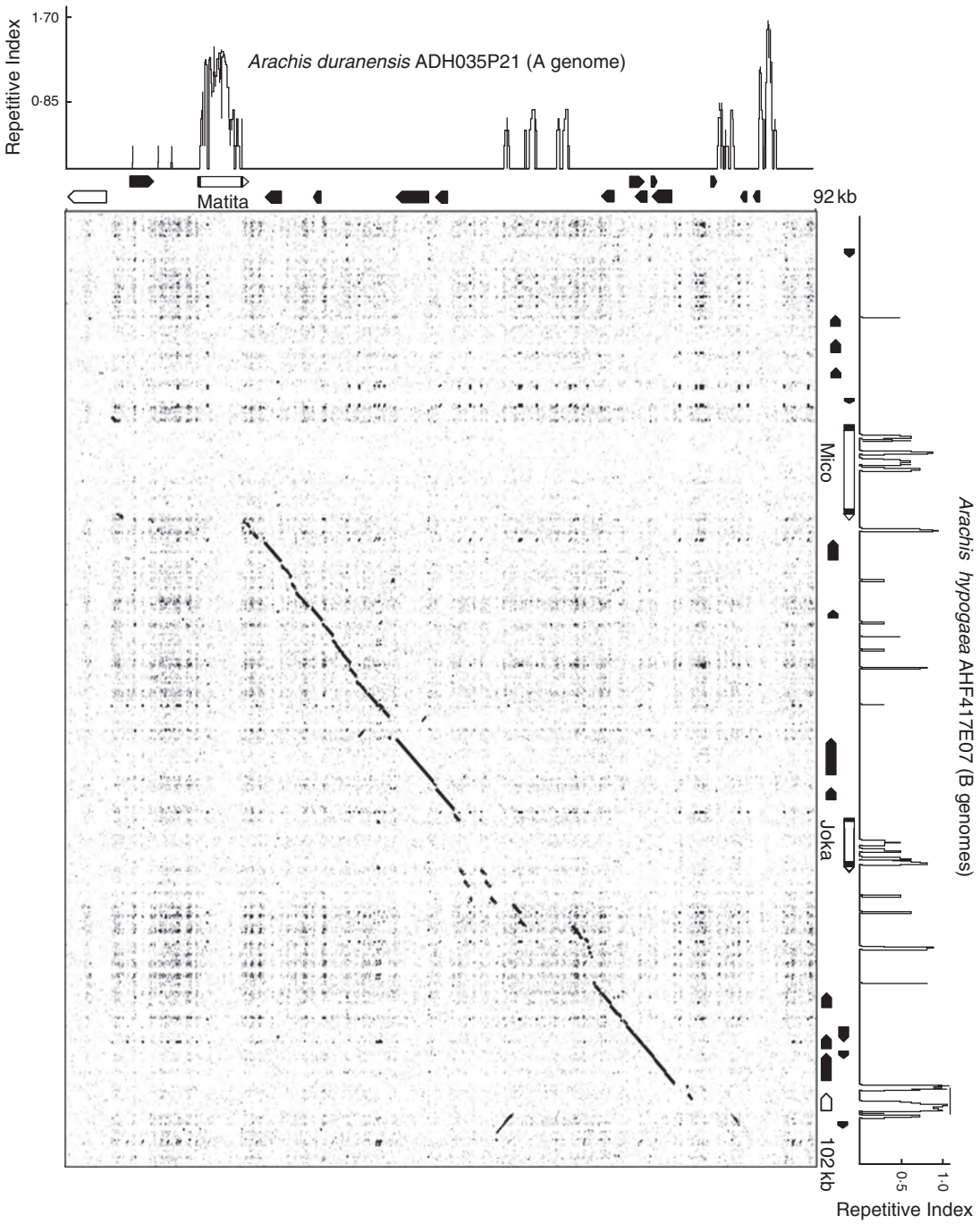

FIG. 6. A comparison of two homeologous A-B genome regions from *Arachis*, the BAC clones ADH035P21 (A genome) and AHF417E07 (B genome). The main area of the figure is a dot plot generated by the software Gepard, and annotation schemes and repetitive index plots are aligned with *x*- and *y*-axes. The annotation scheme represents non-transposon genes as black arrows, and retrotransposons and their truncated fragments as white arrows. The LTRs of complete retrotransposons are shown in black. The repetitive index plots represent a score for repeat content based on BLASTN against *A. duranensis* BAC end sequences. The score is calculated using the formula repetitive index = log10(N), where N is the number of BLASTN homologies with an E-value of $\leq$1e-20. The dot plot shows a complete micro-syntenic region flanked by repetitive DNAs in both A and B genomes; note that these flanking repetitive DNAs in the A and B genomes are completely different in nature. Within the microsyntenic region, segments of very high sequence identity are broken by segments with no significant sequence similarity, most frequently indels. The microsyntenic region harbours eight predicted genes which reside on the segments of high sequence identity; these genes encode: two putative proteins, a transmembrane BT1 family protein, one putative protein, two lipid transfer/seed storage/trypsin-alpha amylase inhibitor proteins, one proteosome protein and one seed storage Aha h1 protein. Segments with no significant sequence similarity are frequently repetitive. Note also the patterns in the background granularity of the plot. This granular signal represents the presence of short low complexity sequences that have random short matches with low complexity sequences in the other BAC sequence being compared. It is notable that the genes, complete retrotransposons and their truncated fragments 'trace out' a clear 'path' on the dot plot, indicating strong purifying for the genes, and a recent origin for the transposon sequences. Non-repetitive intergenic sequences tend to trace out a granular path, indicating their lack of selective pressure and ancient evolutionary origin.

Moretzsohn *et al.*, 2013). Only very recently were they bought together by a polyploidy event, probably in pre-historical times (Bertioli *et al.*, 2011).

Fluorescence *in situ* hybridization of 27 BAC clones from *A. duranensis* (A genome), selected for the presence of genes, showed hybridization signals distributed at multiple dispersed sites on interstitial regions mostly of the chromosome of the peanut A genome (Fig. 1). Depending on the probe, hybridization was also weakly present in peanut B chromosomes. The diffused pattern of the hybridization signals observed is similar to those observed with *Arachis* repetitive elements (Nielen *et al.*, 2010, 2011), clearly different from the distinct dotted signals obtained with the genes *ara h 2* (peanut allergen) and *ara h 6* (a related conglutin) in *Arachis* spp. (Ramos *et al.*, 2006).

These signals were characteristic of repetitive DNA, and we investigated the nature of the repetitive elements and its genome specificity by sequencing selected *A. duranensis* BACs. For the *A. duranensis* BAC sequencing we used the standard methods of Sanger and 454 sequencing and for one BAC we compared assemblies using both methods. Although they were broadly consistent, there were some inversions and regions of poor match (see Supplementary Data Fig. S1). This supports the very great difficulty of generating a completely correct representation of genomic sequences that contain repetitive elements using both these sequencing technologies (see discussion by Kuhn *et al.*, 2012). For (especially early generation) Illumina sequencing, which generates shorter sequence reads, these problems are likely to be more acute; indeed many recent assemblies of whole-genome shotgun sequences are discarding large proportions of the most structurally dominant component of plant genomes even before assembly.

In total, 1·26 Mb of genomic sequence from 12 genomic regions of the A genome of *A. duranensis* were analysed. To identify repeated sequences, we used dot plots and BLAST sequence similarity searches against a database of BAC end sequences. Dot plot comparisons of BAC sequences against themselves and against each other showed that many sequences were repetitive and indicated the presence of LTR retrotransposons (an image of all *A. duranensis* BAC sequences *vs.* all *A. duranensis* BAC sequences is available in Supplementary Data Fig. S5). In total, we identified ten complete different types of retrotransposons. These were, FIDEL and Matita, peanut retrotransposons already identified (Nielen *et al.*, 2010, 2011), and eight other new elements (Fig. 2; representative sequences of retrotransposons are available in Supplementary Data File S4). It was notable that the most abundant of these elements, named Feral, had high sequence similarity to FIDEL in the LTRs and the 3'-UTR but completely different 5'-UTR and coding regions. Furthermore, whilst FIDEL is an autonomous retrotransposon encoding all essential proteins, Feral does not encode reverse transcriptase and is non-autonomous. It would seem most likely that Feral is parasitic on FIDEL. In addition to complete elements, we found many fragmentary sequences, especially solo LTRs. In total, complete elements and fragments of FIDEL and Feral make up about one-sixth of the analysed *A. duranensis* genome regions.

The third most prominent element was a retrotransposon named Pipa. This element is notable in that it does not harbour coding regions for any retrotransposon proteins that can easily be identified, and so must be non-autonomous. However, it does harbour an ORF close to the 5'-LTR. An autonomous retrotransposon with significant similarity to Pipa was also identified. We named it Pipoka.

Feral and Pipa share an interesting characteristic: they both encode proteins that are completely different from those encoded by their probable autonomous pairs. Feral does have ORFs that encode retrotransposon proteins, but they are derived from a retrotransposon different from FIDEL, with the ORF for reverse transcriptase absent (Fig. 2). Pipa and Pipoka have detectable sequence similarity in all regions, except for their ORFs. Pipoka's 3' ORF has been deleted precisely from Pipa, and Pipoka's 5' ORF has been replaced with an unrelated ORF encoding a protein with no apparent homology to any known protein (Fig. 2, and Supplementary Data File S1). Intriguingly, this ORF shows significant signs of evolutionary selection, indicating that it encodes a protein with a biological function. Previous studies have shown that DNA transposons and retrotransposons can 'capture' gene sequences (e.g. Alix *et al.*, 2008) and the amplification of such coding sequences may play an important role not only in gene amplification but also in genome divergence. Although we cannot assign any higher biological significance to the non-autonomous elements discovered here, it is clear that Feral and Pipa are not derived from their autonomous pairs simply by mutational degradation and deletion, but that they must have evolved by complex mechanisms not typical for non-autonomous elements.

The next most abundant element found, Gordo, has tandem repeats within its LTRs and is also non-autonomous, followed by less abundant autonomous retrotransposons such as Curu that has long LTRs of 3448 bp, RE-128, Mico and Grilo.

These different retrotransposons cover a surprising proportion of the analysed genomic regions. FIDEL and Feral, together with Pipa and Pipoka, make up a quarter, and these four added to Gordo and Curu make up a third of the sequence space. In the sampling of *A. duranensis* BACs for sequencing there were a number of types of BAC clones: gene rich, including one harbouring a resistance gene cluster; gene poor; and entirely repetitive (Table 1). That the sequenced regions are quite representative of the A genome as a whole is supported by the consistency of relative abundances of the retrotransposons as determined from the sequenced regions and as estimated by BLAST searches against a database of BAC end sequences. We can be confident that these retrotransposons constitute a very significant proportion of the peanut genome and are probably the most abundant A genome elements.

The BAC-FISH results are consistent with the sequence analysis. The signals from the interstitial regions of the peanut A genome chromosomes arms reflects the distribution of the most abundant autonomous/non-autonomous retrotransposon pair. Minor differences in hybridization patterns can be accounted for by the exact mixture of retrotransposons in each *A. duranensis* BAC sequence (Fig. 1). The A-genome contigs consisted of 5·9–69·4 % retrotransposon sequences (Table 1). Even when abundant unlabelled DNA or re-annealed high-copy DNA was used to pre-hybridize with either the probe in solution or chromosome spreads on the slides, none of the A genome BACs tested gave a distinct double pair of point signals from any pair of chromosomes. The repetitive signals were so strong that they rendered the relatively small signals from single-copy hybridization of the non-repetitive parts of the BAC

undetectable, even in the peanut B chromosomes where there was less homology of the repetitive sequences.

Although the frequency of repetitive elements in genomes should take into account sequencing/assembly strategies used and the size of the genome, the elevated presence of Class I LTRs elements that we observed in *Arachis* is well documented in many monocot and dicotyledonous plants such as *Sorghum bicolor* (55 %; Paterson *et al.*, 2009), maize (79 %; Meyers *et al.*, 2001), rice (22 %; Ma *et al.*, 2004), *Medicago truncatula* (26·5 %; *Medicago truncatula* Genome Project, http://jcvi.org/cgi-bin/medicago/annotation.cgi?page=repeats); soybean (42 %; Schmutz *et al.*, 2010); and *Lotus japonicus* (19·23 %; Sato *et al.*, 2008). Also, abundant non-autonomous elements of recent origin, solo LTRs and nested retroelements have been observed in other plant genomes (Wawrzynski *et al.*, 2008; Schmutz *et al.*, 2010).

That the A and B genome sequences have rapidly diverged in repetitive components is supported by the observation that the most abundant A genome retrotransposons are predominantly located in the peanut A genome chromosomes (Fig. 1). It is also supported by the fact that almost all datable transposition events were $<3{\cdot}5$ million years old (the estimated date of evolutionary divergence of the A and B genomes, Fig. 3). In total, 14 % of the sequenced A genome regions is occupied by complete retrotransposons that are $<3{\cdot}5$ million years old. Evidence that the amount of evolutionary new sequences is greater than this can be observed in the overall granularity of dot plots. This granular distribution of dots represents short low complexity sequences that accumulate by DNA polymerase replication slippage over long evolutionary time scales. It is absent from dot plot signals of gene exons, presumably because mutations in exons tend to be eliminated by natural selection. It was also notable that this granularity was largely absent, not only from complete retrotransposons, but also from their truncated fragments, presumably because they are of recent origin (Fig. 6, and Supplementary Data). The predominant location of the A genome retrotransposons in the A genome chromosomes of tetraploid peanut also shows that they have not undergone a very large-scale activity since the allopolyploidy event that gave rise to the cultivated species.

The software FGENESH predicted the presence of genes throughout the BAC sequences in both high copy and low copy regions, with further evidence from searches against the Pfam databases, protein sequences from arabidopsis or soybean, and *Arachis* ESTs. Notably, ESTs provide strong evidence for transcription of retrotransposons in both polyprotein-encoding and non-genic regions, but mostly corresponding to pseudogenes or artefacts, while well-supported genes were confined to non-retrotransposon regions (reviewed by Bennetzen, 2000). If left unidentified, retrotransposons and their truncated fragments are likely to be major confusing factors for annotation of the peanut genome that is currently being sequenced. Furthermore, as noted by Wang *et al.* (2012), retrotransposon-related genes may be scored in the 'gene' fraction in some annotation pathways, thus underestimating the repetitive element content of the genome. Although predicted genes in and overlapping with retrotransposon regions appear to have support from ESTs, it must be considered that the transcripts are produced from very large numbers of retrotransposons spread across the genome. Therefore, the activity of individual retrotransposons must be low on average. However, there are several documented examples of transposable elements capturing gene fragments and evolving into functional genes (e.g. Elrouby and Bureau, 2010; Barbaglia *et al.*, 2012). It is possible that the *Arachis* retrotransposons characterized here have played a role in the evolution of new genes. In accordance with this we could identify hundreds of ESTs with sequence similarity to FIDEL; of these, four apparently encode non-transposon proteins.

For two of the *A. duranensis* BAC sequences (A genome), we were able to compare their homeologous regions in the *Arachis* B genome (Fig. 6). In both cases, the microsyntenic regions are flanked by repetitive DNA regions that were completely different in the A and B genomes. In both cases, within the microsyntenic regions, highly conserved segments (with about 95 % identity) are punctuated by segments with no significant homology. There was a distinct tendency for these segments to be detectably repetitive. This indicates a key role for repetitive DNA in the structural divergence of the A and B genomes. Notably, this divergence is not evenly distributed; it is concentrated in intergenic regions. Therefore, gene sequences and gene orders remain highly conserved. This provides a resolution to the apparent paradox of dynamic repetitive and conservative genic fractions in genome structure – the action of repetitive DNA accumulates predominantly in intergenic regions.

However, over longer evolutionary time frames (55 million years) retrotransposons have been associated with the erosion of genome synteny (Bertioli *et al.*, 2009): there is a negative correlation between retrotransposon density and degree of synteny between *Lotus* and *Medicago* and *Arachis*. This is most probably due to the facilitating action of repetitive sequences on the pairing of non-homologous chromosome regions, a process that can lead to unequal crossovers and chromosomal rearrangements (inversions, deletions, duplications, additions and translocations).

Conclusions

In this study we have shown that a substantial proportion of the highly repetitive component of the A genome of peanut is accounted for by relatively few LTR retrotransposons. Three of the most abundant elements are non-autonomous, and two of these appear to harbour 'hitchhiking' ORFs, in one case with retrotransposon-related function, and in the other with a biological function that remains to be identified.

During our studies, it became apparent that these retrotransposons and their truncated fragments would be a major confusing factor in gene annotation if not properly identified. The retrotransposons described are all transcribed, although, considering their copy numbers, transcription levels are low.

We also show that these elements are predominantly of recent evolutionary origin, most apparently post-dating the evolutionary divergence of the A and B genomes of cultivated peanut. It is clear that these elements have contributed very substantially to the divergence of the peanut A and B genomes. These genomes are likely to consist of mosaics of highly similar segments interrupted by segments of repetitive DNA with no corresponding sequence in the homeologous genome. Furthermore, observations on two pairs of homeologous A-B genome segments indicate that the retrotransposons we have identified here and other repetitive DNAs have played an important part in

genome remodelling, especially in intergenic regions, over evolutionary time.

SUPPLEMENTARY DATA

Supplementary data are available online at www.aob.oxfordjournals.org and consist of the following. Figure S1: plot of two *A. duranensis* BAC ADHI8B08 (A genome) sequences, one obtained by random fragmentation and Sanger chemistry with paired end reads, and the other by 454 GS FLX titanium chemistry. Figure S2: dot plots of retrotransposon sequences; first FIDEL vs. Feral, and secondly Pipoka vs. Pipa (autonomous vs. non-autonomous). Figure S3: annotations of *A. duranensis* and *A. hypogaea* sequences of BAC clones showing genes, and complete and incomplete retrotransposons. Figure S4: dot plot showing homeologous genome regions of the *A. duranensis* BAC clone ADH068E04 (A genome) × *A. ipaënsis* AIPA147A20 (B genome) containing a DNA gyrase gene, with a microsyntenic region situated at the 5' and 3' ends and over about 43 and 47 kb, respectively. Figure S5: dot plot of all *A. duranensis* BAC sequences vs. all *A. duranensis* BAC sequences. Table S1: list of BAC clones used as probes for fluorescent *in situ* hybridization. Table S2: list of ten *Arachis duranensis* (A genome) retroelements indicating their superfamily, total length and LTR length in bp. File S1: sequences of ORFs from four complete Pipa retrotransposons. File S2: Pipa ORFs datamined from BAC end sequences. File S3: alignment of Pipa ORFs in fasta format. File S4: text file with sequences of representatives of each of the retrotransposons. File S5: FIDEL and Feral LTR sequences in fasta format. File S6: Multiple alignments of FIDEL and Feral LTRs in fasta format.

ACKNOWLEDGEMENTS

S.N. and D.B. thank the National Council for Scientific and Technological Development of Brazil (CNPq) for fellowships. B.V. is grateful for a post-graduate grant from the Brazilian Ministry of Education (CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). We thank Igor Bacon for help with scripting in Perl.

LITERATURE CITED

Alix K, Joets J, Ryder C, *et al.* 2008. The CACTA transposon Bot1 played a major role in Brassica genome divergence and gene proliferation. *The Plant Journal* **56**: 1030–1044.

Altschul SF, Madden TL, Schäffer AA, *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Barbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC, Lal SK. 2012. Gene capture by helitron transposons reshuffles the transcriptome of maize. *Genetics* **190**: 965–975

Bennetzen JL. 2000. Transposable element contributions to plant gene and genome evolution. *Plant Molecular Biology* **42**: 251–269.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics and Development* **15**: 1–7.

Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *The Plant Cell* **9**: 1509–1514.

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.

Bertioli DJ, Seijo G, Freitas FO, Valls JFM, Leal-Bertioli SCM, Moretzsohn MC. 2011. An overview of peanut and its wild relatives. *Plant Genetic Resources: Characterization and Utilization* **9**: 134–149.

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.

Choi H-K, Luckow MA, Doyle J, Cook DR. 2006. Development of nuclear gene-derived molecular markers linked to legume genetic maps. *Molecular Genetics and Genomics* **276**: 56–70.

Dhillon SS, Rake AV, Miksche JP. 1980. Reassociation kinetics and cytophotometric characterization of peanut (*Arachishypogaea* L.) DNA. *Plant Physiology* **65**: 1121–1127.

Eddy SR. 2011. Accelerated profile HMM searches. *PLoS Computational Biology* **7**: e1002195.

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

Elrouby N, Bureau TE. 2010. Bs1, a new chimeric gene formed by retrotransposon-mediated exon shuffling in maize. *Plant Physiology* **153**: 1413–1424.

Estep MC, DeBarry JD, Bennetzen JL. 2013. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity* **110**: 194–204.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Fredslund J, Madsen LH, Hougaard BK, *et al.* 2006. A general pipeline for the development of anchor markers for comparative genomics in plants. *BMC Genomics* **7**: 207.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195–202.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* **27**: 221–224.

Greilhuber J. 2005. Intraspecific variation in genome size in angiosperms: identifying its existence. *Annals of Botany* **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 Biology* **8**: 14.

Halward T, Stalker HT, Laure EA, Kochert G. 1991. Genetic variation detectable with molecular markers among unadapted germplasm resources of cultivated peanut and related wild species. *Genome* **34**: 1013–1020.

Huang X, Madan A. 1999. CAP3: a DNA sequence assembly program. *Genome Research* **9**: 868–877.

Husted L. 1936. Cytological studies on the peanut, *Arachis*. II. Chromosome number, morphology and behavior, and their application to the problem of the origin of the cultivated forms. *Cytologia* **7**: 396–423.

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

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). *American Journal of Botany* **83**: 1282–1291.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics* **23**: 1026–8.

Kuhn GCS, Kuttler H, Moreira-Filho O, Heslop-Harrison JS. 2012. The 1·688 repetitive DNA of *Drosophila*: concerted evolution at different genomic scales and association with genes. *Molecular Biology and Evolution* **29**: 7–11.

Lavin M, Pennington RT, Klitgaard BB, Sprent JI, de Lima HC, Gasson PE. 2001. The dalbergioid legumes (Fabaceae): delimitation of a pantropical monophyletic clade. *American Journal of Botany* **88**: 503–533.

Lewis G, Schrire B, Muackinder B, Lock M. 2005. *Legumes of the world*. Kew: Royal Botanic Gardens.

Ma JX, Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proceedings of the National Academy of Sciences, USA* **101**: 12404–12410.

Ma J, Devos MK, Bennetzen JL. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Research* **14**: 860–869.

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

Meyers BC, Dickerman AW, Michelmore RW, Sivaramakrishnan S, Sobral BW, Young ND. 1999. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. *The Plant Journal* **20**: 317–332.

Meyers BC, Tingey SV, Morgante M. 2001. Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Research* **11**: 1660–1676.

Moretzsohn MC, Leoi L, Proite K, *et al*. 2005. Microsatellite based, gene-rich linkage map for the AA genome of *Arachis* (Fabaceae). *Theoretical and Applied Genetics.* **111**: 1060–1071.

Moretzsohn M, Barbosa A, Alves-Freitas D, *et al*. 2009. A linkage map for the B-genome of Arachis (Fabaceae) and its synteny to the A-genome. *BMC Plant Biology* **9**: 40.

Moretzsohn MC, Gouvea EG, Inglis PW, Leal-Bertioli SCM, Valls JFM, Bertioli DJ. 2013. A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Annals of Botany* **111**: 113–126.

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

Nielen S, Campos-Fonseca F, Leal-Bertioli S, *et al*. 2010. FIDEL – a retrovirus-like retrotransposon and its distinct evolutionary histories in the A- and B-genome components of cultivated peanut. *Chromosome Research* **18**: 227–246.

Nielen S, Vidigal BS, Leal-Bertioli SCM, *et al*. 2011. Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the Arachis A–B genome divergence. *Molecular Genetics and Genomics* **287**: 21–38.

Paterson AH, Bowers JE, Bruggmann R, *et al*. 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.

Ragupathy R, Cloutier S. 2008. Genome organisation and retrotransposon driven molecular evolution of the endosperm Hardness (Ha) locus in Triticum aestivum cv Glenlea. *Molecular Genetics and Genomics* **280**: 467–481.

Ramos ML, Fleming G, Chu Y, Akiyama Y, Gallo M, Ozias-Akins P. 2006. Chromosomal and phylogenetic context for conglutin genes in Arachis based on genomic sequence. *Molecular Genetics and Genomics* **275**: 578–592.

Robledo G, Seijo G. 2010. Species relationships among the wild B genome of *Arachis* species (section Arachis) based on FISH mapping of rDNA loci and heterochromatin detection: a new proposal for genome arrangement. *Theoretical and Applied Genetics* **121**: 1033–1046.

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. *Theoretical and Applied Genetics* **118**: 1295–1307.

Rutherford K, Parkhill J, Crook J, *et al*. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944–945.

Rzhetsky A, Nei M. 1992. A simple method for estimating and testing minimum evolution trees. *Molecular Biology and Evolution* **9**: 945–967.

Saitou N, Nei M. 1987. The neighbor–joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406–425.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* **10**: 516–522.

Sanmiguel P, Bennetzen JL. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Annals of Botany* **82**: 37–44.

Sato S, Nakamura Y, Kaneko T, *et al*. 2008. Genome structure of the legume, *Lotus japonicus*. *DNA Research* **15**: 227–239.

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

Schmutz J, Cannon SB, Schlueter J, *et al*. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* **463**: 178–183.

Schwarzacher T, Heslop-Harrison JS. 2000. *Practical in situ hybridization*. Oxford, UK: BIOS Scientific Publishers.

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. ipaensis* are the wild diploid progenitors of*A. hypogaea* (Leguminosae). *American Journal of Botany* **91**: 1294–1303.

Seijo JG, Lavia GI, Fernández A, *et al*. 2007. Genomic relationships between the cultivated peanut (*Arachis hypogaea*, Leguminosae) and its close relatives revealed by double GISH. *American Journal of Botany* **94**: 1963–1971.

Shirasawa H, Bertioli DJ, Varshney RK, *et al*. 2013. Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of Arachis and divergences with other legume genomes. *DNA Research* **20**: 173–184.

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

Smartt J, Stalker HT. 1982. Speciation and cytogenetics in *Arachis*. In: Pattee HE, Young CT. eds. *Peanut science and technology*. Yoakun: American Peanut Research Education Society, 21–49.

Smartt J, Gregory WC, Gregory MP. 1978. The genomes of *Arachis hypogaea*. 1. Cytogenetic studies of putative genome donors. *Euphytica* **27**: 665–675.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4·0. *Molecular Biology and Evolution* **24**: 1596–1599.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**: 2731–2739.

Wang H, Penmetsa RV, Yuan M, *et al*. 2012. Development and characterization of BAC-end sequence derived SSRs, and their incorporation into a new higher density genetic map for cultivated peanut (*Arachis hypogaea* L.). *BMC Plant Biology* **12**: 10.

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, *et al*. 2008. Replication of nonautonomous retroelements in soybean appears to be both recent and common. *Plant Physiology* **148**: 1760–1771.

Xu Z, Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35** (suppl 2): W265–W268.

Young ND, Weeden N, Kochert. 1996. Genome mapping in legumes. In: Paterson A. ed. *Genome mapping in plants*. Austin, TX: Landes, 212–227.

Yüksel B, Paterson AH. 2005. Construction and characterization of a peanut HindIII BAC library. *Theoretical and Applied Genetics* **111**: 630–639.

Zwick MS, Hanson RE, McKnight TD, *et al*. 1997. A rapid procedure for the isolation of Cot-1 DNA from plants. *Genome* **40**: 138–142.