Universidade de Brasília

Instituto de Ciências Biológicas

Departamento de Biologia Celular

Estudo dos genomas A e B de *Arachis*

Bruna Vidigal dos Santos

Brasília, 2014

Universidade de Brasília
Instituto de Ciências Biológicas
Departamento de Biologia Celular

Estudo dos genomas A e B de *Arachis*

Bruna Vidigal dos Santos

Tese apresentada ao Curso de Pós-Graduação em Biologia Molecular oferecido pelo Departamento de Biologia Celular – Instituto de Ciências Biológicas – Universidade de Brasília, como requisito parcial para obtenção do Grau de Doutor em Biologia Molecular.

Brasília, 2014

Trabalho realizado no Laboratório de Interação Planta-Praga III – Embrapa Recursos Genéticos e Biotecnologia

Banca Examinadora:

Membro interno: Prof. Dr. Robert Neil Gerard Miller, Universidade de Brasília - UnB
Membro interno: Prof. Dr. Renato Oliveira Resende, Universidade de Brasília – UnB
Membro externo: Dr. Guillermo Seijo, Instituto de Botánica del Nordeste, Corrientes, Argentina
Membro externo: Dr^{a}. Ana Cláudia Guerra Araújo, Embrapa Recursos Genéticos e Biotecnologia
Membro externo: Dr. Márcio de Carvalho Moretzsohn, Embrapa Recursos Genéticos e Biotecnologia

Suplente: Dr**.** Artur Fellipe de Andrade Fonsêca, Embrapa Recursos Genéticos e Biotecnologia

Aos meus pais Jorge e Márcia, meu Príncipe Iugo e familiares pela dedicação, incentivo e apoio constante para a minha formação.

Dedico.

Agradecimentos

É durante a realização de um trabalho que sentimos a necessidade de termos amigos que nos auxiliem, nos guiem e nos orientem. Aqui, expresso minha gratidão:

À minha família, em especial aos meus pais Jorge e Márcia, por todo o amor, incentivo, apoio e por acreditarem em mim. Amo vocês.

Ao meu príncipe encantado Iugo, por todo o companheirismo, dedicação, amor e amizade. Te amo.

Ao meu Orientador Prof. Dr. David John Bertioli, por me acolher em sua equipe, por confiar em meu trabalho, pela orientação, pelos conhecimentos passados que ajudaram em meu crescimento profissional, e que sem dúvida levarei por toda a minha vida. Obrigada por ter apostado em mim.

Ao grupo de pesquisa em *Arachis*: Dr[a]. Ana Cláudia Guerra, Dr[a]. Patrícia Guimarães, Dr[a]. Soraya Bertioli, Dr. Márcio Moretzsohn, pelo imenso apoio, orientação, atenção e sugestões. E ainda, Dr[a]. Ana Brasileiro, Dr[a]. Simone Ribeiro e o analista Leonardo Nunes pelo apoio e disponibilidade.

Aos meus queridos amigos Igor, Eliza, Kaká, Raquel, Uiara e Larissa M., por todo carinho e colaboração neste estudo. Sem vocês este trabalho não seria o mesmo.

Aos meus queridos amigos Andressa, Amanda, Natália, Cris, Thaís, Larissa A., Aninha, Raul, Ana Zotta, Paulinha, Gabi, Mário, Ana Laura, Dione e Lílian, por tudo que me ensinaram e por terem tornado o LPPIII uns dos lugares mais divertidos que já conheci.

A todos que já passaram pelo LPPIII, Pedro, Vânia, Aline, Bárbara, Carol, Marcelo, Rafa, Tati, Talles e muitos outros... a todos que já me emprestaram soluções e equipamentos...

Aos meus queridos amigos Fábio, Lorena e Lua, por todo carinho e encontros animados.

Aos professores do Programa de pós-graduação da UnB, pela contribuição para minha formação acadêmica.

Aos membros da banca de defesa, pela disponibilidade e contribuição.

À CAPES, pelo apoio financeiro que viabilizou esse estudo.

A todos que direta ou indiretamente contribuíram para a realização desse trabalho.

"Penso que só há um caminho para a ciência ou para a filosofia: encontrar um problema, ver a sua beleza e apaixonar- se por ele."

Karl Popper

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

O amendoim (*Arachis hypogaea* L.) é um alotetraploide com origem recente e cujo genoma tem aproximadamente 2,8 Gb, composto majoritariamente por sequências repetitivas. Este estudo relata uma investigação do componente repetitivo presente nas espécies parentais do amendoim, *A. duranensis*, provável doador do genoma A e *A. ipaënsis*, provável doador do genoma B, por meio de análises das suas sequências genômicas completas, bem como de clones selecionados da biblioteca BAC de *A. duranensis*. Nos clones, foram identificados dez retrotransposons LTR distintos, enquanto que nas sequências genômicas completas, 81 famílias de retrotransposons LTR foram identificadas em *A. duranensis* e 89 em *A. ipaënsis*, ocupando aproximadamente 28,5% e 27,6% do genoma A e B, respectivamente. Dessas famílias, 37 representam a maior parte do conteúdo repetitivo nos dois genomas, sendo que os elementos FIDEL e Feral são os mais frequentes. Esses resultados mostram que uma parte substancial do componente altamente repetitivo desses genomas é explicada por um número relativamente pequeno de retrotransposons LTR, seus fragmentos e LTRs-solo. A maioria das datas de transposição estimadas para esses retrotransposons foi posterior a 3,5 milhões de anos atrás, data estimada da divergência dos genomas A e B, indicando que esses retrotransposons LTR tiveram um papel notável na organização desses genomas. Análises de hibridização in situ por fluorescência (FISH), utilizando sondas obtidas a partir das sequências dos genes que codificam a transcriptase reversa de cada família de retrotransposon LTR, mostraram sinais de hibridização detectáveis múltiplos e dispersos em vários, mas não em todos os cromossomos dos subgenomas A e B de amendoim, com marcação predominantemente ao longo dos braços dos cromossomos. Comparações entre sequências homeólogas dos genomas A e B indicaram alta semelhança no conteúdo gênico, porém grandes diferenças no conteúdo repetitivo, mostrando que os retrotransposons identificados neste estudo, juntamente com outros elementos repetitivos têm desempenhado um papel importante na remodelação do genoma ao longo da evolução, especialmente em regiões intergênicas. A construção e validação de *pools* 3-D construídos para os clones das bibliotecas BAC representativas dos genomas A (*A. duranensis*) e B (*A. ipaënsis*) foram realizadas. Essa ferramenta possibilitou a identificação e isolamento de genes de interesse em *Arachis*, tais como, expansina e dessaturase de ácidos graxos.

Palavras-chave: *Arachis*, genoma, biblioteca BAC, retrotransposon LTR, FISH.

Abstract

Peanut (*Arachis hypogaea* L.) is an allotetraploid of recent origin with a genome of about 2.8 Gb and high repetitive content. This study reports an analysis of the repetitive component present of the progenitor species from peanut, *A. duranensis*, likely donor of A genome and *A. ipaënsis* of B genome, using their whole genome sequences and selected clones from the BAC library of *A. duranensis*. Ten LTR retrotransposons were identified in these clones whilst 81 families in *A. duranensis* and 89 in *A. ipaënsis* complete genomes, representing about 28.5% of the A genome and 27.6% of B genome, respectively. Only 37 families represent most of the repetitive content of the two genomes, and the most abundant retrotransposon are FIDEL and Feral. It is here shown that a substantial proportion of the highly repetitive component of these genomes is accounted for by relatively few LTR retrotransposons, their fragments and solo-LTR. These retroelements are predominantly of recent evolutionary origin, most apparently post-dating the evolutionary estimated date of the A and B genomes divergence of the cultivated peanut, about 3.5 million years ago. This indicates that these LTR retrotransposons contributed to the divergence of these genomes. Analysis by fluorescence in situ hybridization using probes obtained from the genes sequencing codifying for the reverse transcriptase of each family of LTR retrotransposons of A and B genomes produced multiple and dispersed hybridization signals on several, but not all chromosomes of A-B peanut subgenomes, mainly along the chromosomes arms. Comparisons between homeologues sequences of A and B genomes showed high similarity in gene content, but differences in the repetitive content showing that the retrotransposons identified in this study, and another repetitive elements have played an important role in these genomes remodeling, especially in intergenic regions, over evolutionary time. The construction and validation of 3-D *pools* for clones of the A-B genomes BAC libraries were made. This tool allowed the identification and isolation of genes of interest in *Arachis* such as expansins and fatty acid desaturase.

Keywords: *Arachis*, genome, BAC library, LTR retrotransposon, FISH.

Introdução Geral

O amendoim (*Arachis hypogaea* L.) é atualmente a quinta oleaginosa mais cultivada no mundo (USDA-FAS, 2013a), bastante importante na Ásia, África e América do Norte, embora o gênero *Arachis* seja originário da América do Sul (Krapovickas & Gregory, 1994). Além de sua importância na economia de diversos países, essa leguminosa é crucial para as dietas e um meio de subsistência de pequenos agricultores.

A origem do amendoim está possivelmente associada a um eventual cruzamento entre duas espécies silvestres diploides com genomas A e B, *Arachis duranensis* e *A. ipaënsis*, respectivamente, que teria resultado em um híbrido estéril cujos cromossomos foram espontaneamente duplicados, levando à restauração da fertilidade. Portanto, por ser uma espécie alotetraploide (genoma AABB), o amendoim cultivado possui dois genomas distintos designados subgenomas A e B. Entretanto, essa origem isolou geneticamente essa espécie, o que ocasionou uma redução na variabilidade genética, tornando a cultura do amendoim mais susceptível a estresses bióticos e abióticos, comparando com seus parentais. Espécies silvestres de *Arachis* constituem uma importante fonte de alelos para várias das características de interesse nos programas de melhoramento do amendoim (Stalker & Simpson, 1995; Garcia *et al*., 1996; Dwivedi *et al*., 2003; 2006; Rao *et al*., 2003; Fávero *et al*., 2006; Leal-Bertioli *et al*., 2010).

Estudos que envolvem as potenciais espécies parentais do amendoim (*A. duranensis* e *A. ipaënsis*) podem gerar ferramentas importantes que auxiliem a entender a estrutura, organização e história evolutiva do genoma tetraploide do amendoim. Trabalhos envolvendo mapeamento genético evidenciaram o alto grau de sintenia entre marcadores moleculares (predominantemente derivados de DNA de baixo número de cópias) entre os mapas construídos utilizando as espécies progenitoras do amendoim (Moretzsohn *et al*., 2009; Bertioli *et al*., 2009; Shirasawa *et al*., 2013). Esses resultados sugerem uma evolução lenta na organização dos genes e indicam que a maior parte do espaço gênico nos componentes genômicos A e B de amendoim apresentem semelhanças. Por outro lado, experimentos de hibridização in situ genômica - GISH (*genomic in situ hybridization*) em cromossomos metafásicos de amendoim mostraram que grande parte da fração repetitiva presente nos componenentes genômicos ou subgenomas A e B são divergentes (Seijo *et al*., 2007).

Para a maioria das plantas é o DNA repetitivo que ocupa a maior parte do genoma e determina a estrutura dos cromossomos (Schmidt & Heslop-Harrison, 1998). Ao longo dos últimos anos, o sequenciamento genômico de diferentes espécies de plantas evidenciou que os genomas são compostos principalmente por elementos de transposição (TEs – *Transposition Elements*), sequências de DNA que possuem a capacidade de se transpor ao longo do genoma hospedeiro (Schulman, 2013). Enquanto muitos projetos de sequenciamento completo de genomas consideram os TEs uma presença inconveniente que dificulta a montagem mais precisa do genoma e anotação de genes, sua onipresença e abundância indicam, no entanto, uma grande importância na estrutura e evolução dos genomas (Vesely *et al*., 2012). A união de diferentes genomas em um mesmo núcleo, resultante da hibridização entre espécies, tal como ocorreu com o amendoim, pode iniciar ou mascarar a ativação de TEs, gerando alterações e consequências para a estrutura do genoma e para a expressão gênica (Zhao *et al*., 1998; Kashkush *et al*., 2002, 2003; Petit e*t al*., 2010).

Particularmente, os TEs do tipo retrotransposons LTR (*Long Terminal Repeats*), que possuem o mecanismo de transposição do tipo "copia e cola", contribuem com uma fração substancial dos genomas de plantas, perfazendo até 80% dos genomas, tal como em milho (SanMiguel & Bennetzen, 1998).

Em amendoim, alguns retrotransposons LTR tais como FIDEL e Matita foram identificados e caracterizados (Nielen *et al*., 2010; 2012). O retrotransposon FIDEL apresentou uma considerável diferença na sua frequência em cromossomos A e B de amendoim (Nielen *et al*., 2012), fato que pode estar ligado à atividade de transposição diferencial ocorrida provavelmente após a divergência evolutiva das espécies parentais (3-3,5 milhões de anos (Nielen *et al*., 2012; Moretzsohn *et al*., 2013). Esse comportamento, somado à ocorrência de outros elementos repetitivos pode explicar a divergência da fração repetitiva dos subgenomas A e B de amendoim. A identificação e caracterização de novos retrotransposons LTR, que perfazem grande parte do conteúdo repetitivo do genoma de amendoim, podem ser úteis nos estudos genômicos comparativos e para elucidação dos caminhos evolutivos dos subgenomas A e B de amendoim, bem como para fornecer dados para a anotação mais precisa de sequências genômicas.

Recentemente, um grupo multiinstitucional - IPGI (*The International Peanut Genome Initiative* – http://www.peanutbioscience.com) se reuniu com o intuito de gerar sequências genômicas completas das duas espécies progenitoras do amendoim cultivado, *A. duranensis* e *A. ipaënsis*. A disponibilização dessas sequências completas e ordenadas é um passo

importante para o entendimento da organização genética e genômica do amendoim. Estabelecer mais detalhadamente a relação entre os membros do gênero *Arachis* é fundamental, pois possibilitará o uso mais adequado da diversidade genética presente nos parentes silvestres do amendoim. O sequenciamento do genoma de amendoim e seus parentais também contribuirá para a compreensão dos mecanismos moleculares e celulares que sustentam o crescimento, desenvolvimento e reprodução da planta de amendoim, as respostas à doenças e estresses, bem como a expressão de características agronomicamente desejáveis, incluindo a qualidade do óleo, alta produção de sementes, tolerância à seca e resistência à pragas.

Este estudo compreende três capítulos, dos quais os dois primeiros abordam a identificação, caracterização e distribuição de genes e sequências de retrotransposons LTR em clones BAC e sequências genômicas de *A. duranensis* e *A. ipaënsis*, comparação entre regiões homeólogas desses dois genomas para discutir aspectos importantes da estrutura, organização e evolução genômica de *Arachis* spp., e divergência dos genomas A e B das espécies parentais, assim como dos componentes genômicos A e B de amendoim. O terceiro capítulo detalha a construção da ferramenta de *pools* 3-D de clones BAC desenvolvida para as bibliotecas das espécies parentais do amendoim e a localização de clones de interesse.

1. A cultura do amendoim

O amendoim (*Arachis hypogaea* L.) é uma planta cultivada cujas sementes vêm sendo consumidas in natura desde a pré-história, quando o homem não conhecia a cerâmica e nem dominava o fogo, necessários para o cozimento de muitos alimentos (Freitas *et al*., 2003). A primeira referência sobre o amendoim foi publicada em 1535 por Gonzalo Hernàndez de Oviedo e Valdés nas suas crônicas de viagens pelas Américas, que descrevia o amendoim como muito comum entre os indígenas (Bertioli *et al*., 2011). Trata-se de um dos alimentos humanos mais nutritivos e, ao mesmo tempo, de fácil digestão, consumido por povos de diferentes culturas e sob diversas formas, desde o grão inteiro até processado na forma de doces, confeitos, pastas ou aperitivos.

Possui excelentes propriedades nutricionais, como grande quantidade de aminoácidos, diversas vitaminas, sais, carboidratos, minerais e ainda o antioxidante resveratrol que protege o sistema cardiovascular. Os grãos apresentam alto valor energético, em média, 596 cal/100g de sementes (Jambunathan, 1991; Santos *et al*., 2013). A semente é composta por 50% de óleo, dos quais aproximadamente 80% consistem em ácido oleico (36-67%) e linoleico (15-43%), dentre outros ácidos graxos encontrados em menor quantidade (Moore & Knauft, 1989; Knauft *et al.,* 1993; Lopez *et al*., 2000).

O amendoim é uma das culturas mais populares do mundo e é cultivado extensivamente em regiões tropicais, subtropicais e temperadas (Chirinos, 2011), distribuídas nos seis continentes (Kumar, 2007). Depois da soja (*Glycine max* L. Merr.; produção de 285,30 milhões de toneladas/ano – "Mt/ano"), canola (*Brassica napus* L.; 63,09 Mt/ano), algodão (*Gossypium hirsutum* L.; 44,39 Mt/ano) e girassol (*Helianthus annuus*; 40,29 Mt/ano), o amendoim é a quinta oleaginosa mais cultivada no mundo, com uma superfície de área plantada corresponde a 20,61 milhões de hectares e produção de 36,86 milhões de toneladas/ano (USDA-FAS, 2013a; 2013b).

Apesar da América do Sul ser o provável centro de origem desta cultura, o amendoim tornou-se mais importante em outras regiões do mundo. A grande maioria do amendoim é produzida na Ásia e África. A China ocupa a primeira posição (16,5 Mt – milhões de

toneladas), seguida pela Índia (5 Mt), Estados Unidos (3,06 Mt) e Nigéria (1,55 Mt) (USDA-FAS, 2013a).

Historicamente, o Brasil já foi um dos maiores produtores de óleo de amendoim e importante produtor de amendoim em casca. O óleo, principal produto, e a torta, um subproduto utilizado na composição da ração animal, eram destinados ao mercado interno e externo. A partir de 1970, a cadeia produtiva do amendoim no Brasil sofreu profundas mudanças que resultaram na redução do cultivo nacional, tais como: a expansão da cultura da soja; custos crescentes na produção de amendoim acompanhados por baixo rendimento por área; suscetibilidade às variações climáticas com influências negativas na qualidade do produto e intensas variações de preço durante a comercialização do amendoim (Martins & Perez, 2006).

O amendoim ainda representa um mercado relativamente pequeno quando comparado a outros produtos agrícolas. O cultivo no Brasil representa menos de 1% da produção mundial de amendoim. Contudo, novas tecnologias foram adotadas pela cadeia produtiva, gerando aumentos na produção e maior conquista de mercado (Santos *et al*., 2013). O aumento da produtividade observado nas safras colhidas entre 1976 e 2011, expressa em Kg/ha, é mostrada na figura 1. Em 2013 a safra foi de 296 mil toneladas plantadas em 84 mil hectares (Conab, 2013). Atualmente, grande parte da cadeia de produção está localizada na região Sudeste, sendo São Paulo o maior estado produtor, respondendo por quase 80% do total produzido no Brasil (Godoy *et al*., 1999).

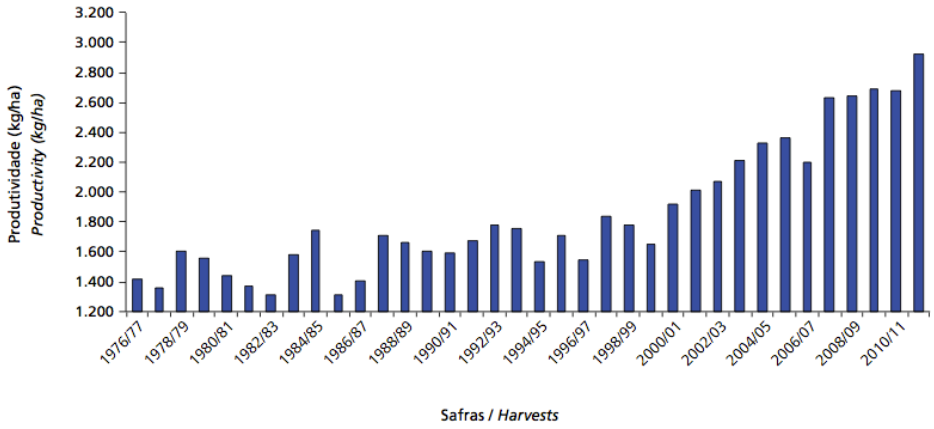


Figura 1: Produtividade de amendoim no Brasil, em Kg/ha. Imagem reproduzida de Anuário Estatístico da Agroenergia, 2012.

2. O gênero *Arachis*

A família Leguminosae ou Fabaceae é dividida em três grandes subfamílias, Mimosoideae, Caesalpinioideae e Papilionoideae (Judd *et al*., 2002), distribuídas principalmente nas regiões tropicais e subtropicais do planeta (Joly, 2002). Quase todas as leguminosas de importância econômica fazem parte de dois subclados dentro da subfamília Papilionoideae: Phaseoloide e Galegoide (figura 2), que divergiram entre si há aproximadamente 50 milhões de anos (Wojciechowski *et al.*, 2004; Lewis *et al.*, 2005). O subclado Phaseoloide inclui espécies que possuem o número de cromossomos 2x = 20 ou 22 como *Phaseolus vulgaris*, *Vigna unguiculata*, *Glycine max* e *Cajanus cajan*. O subclado Galegoide apresenta 2x = 10 a 16 e inclui as espécies *Trifolium* ssp., *Pisum sativum*, *Lens culinaris*, *Vicia faba*, *Cicer arietinum*, *Medicago sativa* dentre outras. O gênero *Arachis* faz parte de um subclado diferente, o Dalbergioide, que divergiu do Galegoide e Phaseoloide há cerca de 55 milhões de anos (Cronk *et al*., 2006).

O gênero *Arachis* é encontrado na América do Sul, em uma região que se estende ao leste dos Andes, sul da Amazônia, norte da Planície Platina e noroeste da Argentina (Krapovickas & Gregory, 1994; Valls & Simpson, 2005). O centro de origem desse gênero é definido como o Planalto Central Brasileiro (Gregory *et al*., 1980; Hammons, 1994) e atualmente encontra-se distribuído em cinco países: Argentina, Bolívia, Brasil, Paraguai e Uruguai. Possui 81 espécies descritas, sendo que 65 ocorrem no Brasil (Krapovickas & Gregory, 1994; Valls & Simpson, 2005) e 48 delas são consideradas endêmicas da flora brasileira (Krapovickas & Gregory, 1994; Valls *et al*., 2013). Todas as espécies são geocárpicas, característica que define claramente este gênero (Krapovickas & Gregory, 1994), e distinguem-se da maioria das outras plantas por produzirem flores acima do solo e os frutos abaixo do solo (Holbrook & Stalker, 2003).

Baseado na morfologia, distribuição geográfica, compatibilidade de cruzamentos e citogenética, este gênero é arranjado em nove seções taxonômicas (*Arachis*, *Erectoides*, *Heteranthae*, *Caulorrhizae*, *Rhizomatosae*, *Extranervosae*, *Triseminatae*, *Procumbentes* e *Trierectoides*) (Krapovickas & Gregory, 1994; Fernández & Krapovickas, 1994; Lavia, 1999; Valls & Simpson, 2005). A maioria das espécies do gênero é diploide (2x = 2n = 20 cromossomos), algumas são aneuploides (2x = 2n = 18) e cinco tetraploides (2x = 4n = 40), incluindo a espécie cultivada *Arachis hypogaea* (Krapovickas & Gregory, 1994; Lavia, 1998; Valls & Simpson, 2005).

O amendoim (*Arachis hypogaea* L.) e seus parentais fazem parte da seção *Arachis* que possui uma ampla distribuição, desde os Andes até áreas costeiras do Atlântico (Creste *et al*., 2005). Um dos critérios para classificação das espécies nessa seção é, em teoria, a possibilidade de cruzamento com a espécie *A. hypogaea*, independentemente da fertilidade dos híbridos gerados (Kochert *et al*., 1991; Krapovickas & Gregory, 1994; Creste *et al*., 2005; Tallury *et al*., 2005; Valls & Simpson, 2005).

Para as espécies diploides da seção *Arachis* foram inicialmente descritos os genomas A, B e D, de acordo com dados citogenéticos e viabilidade de cruzamentos (Smartt *et al*., 1978; Gregory & Gregory, 1979; Singh & Moss, 1982; 1984; Singh, 1986; Stalker, 1991; Fernández & Krapovickas, 1994; Peñaloza & Valls, 2005). As espécies com genoma A são aquelas que possuem um pequeno par de cromossomos, chamado par "A", (Husted, 1936) que apresenta menor condensação na eucromatina quando comparado aos outros cromossomos (Seijo *et al*., 2004; Robledo *et al*., 2009). As espécies diploides que não apresentam o par A, são consideradas associadas ao genoma B de *A. hypogaea*, com exceção de *A. glandulifera*, que possui o genoma D, caracterizado pela presença de seis pares de cromossomos subtelocêntricos (Stalker, 1991; Fernandez & Krapovickas, 1994, Robledo & Seijo, 2008). Existem também três espécies que possuem 2n = 18 cromossomos (*A. decora*, *A. praecox* e *A. palustris*) (Peñaloza & Valls, 1997; Lavia, 1998). Recentemente, uma nova classificação foi feita por Robledo & Seijo (2010), baseada na presença e tamanho de bandas heterocromáticas. Algumas espécies antes classificadas como possuindo genoma B foram reclassificadas como F (*A. benensis* e *A. trinitensis*) e K (*A. batizocoi*, *A. cruziana* e *A. krapovickasii*). Esses dois últimos tipos de genomas possuem bandas centroméricas na maioria dos cromossomos, diferindo uma das outras na quantidade e distribuição da heterocromatina. As espécies *A. hypogaea* e *A. monticola* são alotetraploides com genoma AABB.

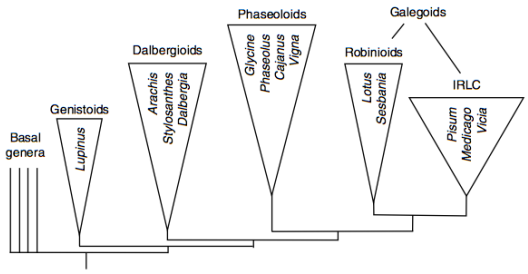


Figura 2: Representação da árvore filogenética da subfamília Papilionoideae com triângulos representando os maiores clados. Os nomes de alguns gêneros de importância econômica estão representados dentro dos triângulos. O gênero *Arachis* é membro do subclado dos Dalbergioide. Esta imagem foi retirada de Bertioli *et al*., (2010), modificada de Wojciechowski *et al*., (2004).

3. Origem do amendoim

A maioria das espécies que compreendem o gênero *Arachis* é diploide, mas o amendoim cultivado (*A. hypogaea*) é tetraploide, com genoma AABB. O amendoim possui o dobro do número de cromossomos da maioria das outras espécies do gênero, sendo que a metade desses cromossomos é semelhante aos de algumas espécies silvestres diploides com genoma AA e a outra metade, a algumas espécies que possuem genoma BB.

Acredita-se que um único evento de hibridização envolvendo duas espécies diploides, *A. duranensis* (genoma AA) e *A. ipaënsis* (genoma BB), seguido da duplicação dos cromossomos, deu origem a um alotetraploide selvagem fértil (Seijo *et al*., 2007). Apesar da planta alotetraploide resultante (genoma AABB) apresentar vigor híbrido, tornou-se reprodutivamente isolada de seus parentes silvestres (Halward *et al*., 1991; Kochert *et al*., 1996; Seijo *et al*., 2004; 2007), fato que levou a uma variabilidade genética reduzida (Kochert *et al*., 1991; 1996; Subramanian *et al*., 2000; Gimenes *et al*., 2002; Herselman, 2003; Moretzsohn *et al*., 2004; Varshney *et al*., 2009). A domesticação dessa espécie para uso como alimento humano, resultou no que hoje chama-se de amendoim cultivado. Portanto, a variabilidade morfológica do amendoim e as diferentes subespécies e variedades existentes, provavelmente são derivadas da pressão seletiva em diferentes ambientes agroecológicos, juntamente com a seleção artificial (Seijo *et al*., 2007). Evidências arqueológicas levam a hipótese de que o amendoim tenha surgido entre 3.500 e 10.000 anos atrás (Bonavia, 1982; Hammons, 1994; Simpson *et al*., 2001).

Análises citogenéticas que mostram o padrão de distribuição de heterocromatina e DNAr (DNA ribossômico) indicam que o cariótipo do amendoim cultivado é equivalente à soma dos cariótipos das espécies diploides silvestres *A. duranensis* e *A. ipaënsis* (Seijo *et al*., 2004). Outro estudo utilizando um duplo GISH (*Genomic in situ hybridization*) reforçou essa hipótese, baseado na forte hibridização das sondas obtidas a partir de DNA genômico dessas duas espécies nos cromossomos de amendoim (Seijo *et al*., 2007). Além disso, Fávero *et al*. (2006) produziram um "anfidiploide sintético" por meio de um cruzamento entre *A. ipaënsis* e *A. duranensis* e utilização de colchicina. Cruzamentos realizados entre esse anfidiploide e várias cultivares de amendoim geraram híbridos altamente férteis. Portanto, constatou-se que *A. ipaënsis* foi o doador mais plausível para todos os 10 cromossomos do genoma BB, ao passo que *A. duranensis* para os cromossomos do genoma AA (Burow *et al*., 2009; Seijo *et al*., 2012; Moretzsohn *et al*., 2013).

Como o amendoim, diversas espécies cultivadas consideradas agronomicamente importantes são poliploides, tais como alfafa, algodão, batata, café, cana-de-acúcar, trigo, dentre outras. A poliploidia é definida como a existência de mais de dois genomas no mesmo núcleo celular, fato que desempenha um importante papel na origem e evolução de genomas (Yoo *et al*., 2013). É descrita como um processo que conduz à especiação instantânea (Mayr, 1963), pois a geração de um único evento, tal como a hibridização entre duas espécies com subsequente duplicação somática é considerada suficiente para estabelecer rapidamente as barreiras que impedem fluxo de genes entre a nova espécie poliploide e aquelas progenitoras. Isso leva a nova espécie a um isolamento reprodutivo (Ramsey & Schemske, 1998). Porém, a poliploidia implica muito mais do que a mera união de dois genomas, e sim a participação de um amplo espectro de ajustes moleculares e fisiológicos, que incluem rearranjos genômicos com troca ou perda de genes, alteração na expressão e silenciamento de genes, além do controle pela epigenética (Madlung, 2013).

Espécies poliploides podem exibir vantagem durante a seleção artificial feita pelo homem (Hilu, 1993), devido ao vigor oriundo da heterose e aumento do tamanho dos tecidos. Além disso, os eventos causados pela poliploidia, tais como mudanças no controle da expressão gênica por meio de silenciamento diferencial, além da atividade de elementos de transposição, também podem ser considerados fatores importantes que podem contribuir para maior adaptabilidade ao cultivo (Soltis & Soltis, 2000; Wessler & Carrington, 2005). As cópias de genes duplicados em poliploides podem evoluir e assumir novas funções (neofuncionalização ou subfuncionalização), manter funções similares às existentes ou ainda adquirir uma interação, permitindo a expansão do nicho ecológico ou maior flexibilidade na capacidade de resposta do organismo à mudanças ambientais (Adams & Wendel, 2005; Moore & Purugganan, 2005; Lynch, 2007).

Sejam quais forem os mecanismos moleculares envolvidos, é notável que o amendoim alotetraploide que possui uma base genética estreita, foi transformado, por meio da domesticação, em uma das culturas mais importantes do mundo, tornando-se completamente distinto de seus parentais silvestres, tanto na arquitetura da planta, quanto no tamanho e forma da semente.

4. Espécies progenitoras do amendoim: *Arachis duranensis* (genoma A) e *Arachis ipaënsis* (genoma B)

As espécies silvestres de *Arachis* foram selecionadas durante a evolução em uma variedade de ambientes e sob variados estresses bióticos, podendo constituir uma rica fonte de diversidade alélica (Nelson *et al*., 1989; Garcia *et al*., 1996; Sharma *et al*., 2003; Bechara *et al*., 2010). A partir de meados do século 20, o valor das espécies silvestres geneticamente mais próximas do amendoim, com maior potencial de uso em seu melhoramento, passou a ser reconhecido e pesquisado. Desde então, deu-se ênfase à coleta, caracterização e conservação de germoplasma de espécies de *Arachis* (Santos *et al*., 2013). A baixa variabilidade genética do amendoim cultivado e a diferença de ploidia entre as espécies silvestres são consideradas barreiras que dificultam o seu melhoramento, contudo, cada vez mais, estudos envolvendo a utilização de espécies silvestres para a incorporação de características desejáveis em amendoim são desenvolvidos (Simpson *et al*., 1993; Garcia *et al*., 1996; Isleib *et al*., 2001).

O conhecimento sobre o conteúdo e a estrutura genômica do amendoim são necessários para acelerar o processo de melhoramento e a obtenção de cultivares com caraterísticas agronomicamente desejáveis. As espécies progenitoras do amendoim oferecem modelos interessantes para esse fim, pois juntas, representam o amendoim cultivado.

Arachis duranensis e *A. ipaënsis* são plantas anuais consideradas bem adaptadas às condições regulares de seca (Krapovickas & Gregory, 1994). Os intervalos de ocorrência natural dessas duas espécies se sobrepõem ligeiramente. Enquanto *A. duranensis* foi encontrada na América do Sul (Argentina, Bolívia e Paraguai), *A. ipaënsis* foi encontrada somente na Bolívia. *A. duranensis* cresce em ambientes secos, principalmente em areia e solo profundo, localizados perto de cursos de água, ao passo que *A. ipaënsis* cresce preferencialmente em regiões próximas à bromélias (Krapovickas & Gregory, 1994).

Essas duas espécies compartilharam um ancestral comum e, recentemente, a data da divergência evolutiva desses dois genomas foi estimada em aproximadamente 3 a 3,5 Ma (milhões de anos atrás) (Nielen *et al*., 2012; Moretzsohn *et al*., 2013), mais recente do que a data estimada para os subgenomas de algodão e soja, que divergiram há aproximadamente 6,7 a 13 Ma, respectivamente (Senchina *et al*., 2003; Schmutz *et al*., 2010).

5. Componentes genômicos A e B do amendoim: tamanho e conteúdo repetitivo

A hibridização dos genomas diploides das espécies *A. duranensis* e *A. ipaënsis*, seguida pela duplicação cromossômica que originou o amendoim, possivelmente ocorreu entre 3.500 e 10.000 anos atrás (Bonavia, 1982; Hammons, 1994; Simpson *et al*., 2001). Apesar do amendoim ser um alotetraploide, o emparelhamento dos cromossomos durante a meiose é quase que inteiramente bivalente (Smartt, 1990), indicando que os componentes genômicos A e B divergiram consideravelmente a nível molecular.

Experimentos de GISH (*Genomic in situ hybridization*) utilizando os DNAs genômicos de *A. duranensis* e *A. ipaënsis* como sondas geraram resultados que apresentaram sinais de hibridização predominantes, mas não total, nos respectivos subgenomas que compõem o amendoim (Seijo *et al*., 2007). Experimentos de FISH (*Fluorescence in situ hybridization*) utilizando como sondas alguns clones BAC selecionados da biblioteca de *A. duranensis* (Guimarães *et al*., 2008), apresentaram sinais de hibridização predominantemente em cromossomos metafásicos do componente genômico ou subgenoma A de amendoim (Araújo *et al*., 2012). Em contrapartida, foi observada uma alta colinearidade entre marcadores (predominantemente gênicos) em mapas genéticos construídos para espécies diploides A e B, indicando uma baixa divergência evolutiva entre esses genomas (Burow *et al*., 2001; Moretzsohn *et al*., 2009; Bertioli *et al*., 2009; Shirasawa *et al*., 2013). Esses estudos levam à hipótese de que apenas os componentes repetitivos presentes nos subgenomas A e B divergiram substancialmente durante suas respectivas jornadas evolutivas.

O genoma de plantas, assim como dos demais eucariotos complexos, é organizado de forma que se distinguem dois tipos de padrões de sequências de DNA: o não-repetitivo e o repetitivo, que podem ser identificados a partir de ensaios de reassociação cinética da fita dupla de DNA (Flavell *et al*., 1974). No DNA não-repetitivo ou com baixo número de cópias encontram-se os genes responsáveis pela codificação das proteínas estruturais, presentes em vias metabólicas, sinalizadoras ou de defesa. O restante do genoma é formado majoritariamente por sequências repetitivas, que são organizadas principalmente em virtude das respostas às pressões seletivas durante a evolução. Tais sequências têm sido apontadas como o principal componente de genomas eucarióticos de plantas superiores, sendo responsáveis pela variação no tamanho dos genomas (Schmidt & Heslop-Harrison, 1998).

As sequências repetitivas são diferenciadas pela homologia, distribuição nas espécies e organização física e genômica. Podem estar distribuídas em *tandem* (blocos consecutivos) ou dispersas ao longo do genoma. As sequências em *tandem* ocorrem em blocos de centenas a milhares de cópias e incluem o DNA codificante, como o DNA ribossômico e não codificante, como classes de DNA satélite (Schmidt & Heslop-Harrison, 1998). A maioria do DNA repetitivo é localizado interdispersamente com outras sequências ao longo do genoma, tais como elementos de transposição. Dentre as sequências repetitivas, os TEs apresentam a maior capacidade de afetar a estrutura e a função do genoma.

Para a origem do amendoim, 10.000 anos, em termos evolutivos, pode ser considerado um período relativamente curto, no entanto, suficiente para a atividade significativa de elementos de transposição. A atividade diferencial de TEs pode ser uma das possíveis causas da divergência entre os subgenomas A e B do amendoim e a estimativa da data de transposição desses elementos pode fornecer conhecimento acerca da evolução desse genoma.

As metodologias e padrões utilizados para realizar o cálculo da estimativa de tamanho de um genoma podem variar substancialmente. Singh e colaboradores (1996) foram os primeiros a estimar o tamanho do genoma do amendoim. Porém, esse valor foi reavaliado e resultou em aproximadamente 2,8 Gb (Gigabases) (Temsch & Greilhuber, 2000; 2001). O tamanho dos genomas de *A. duranensis* e *A. ipaënsis* foi estimado como sendo semelhantes (Seijo *et al*., 2007), ou seja, cada um equivale a aproximadamente metade do genoma do amendoim.

A estrutura repetitiva presente no genoma de amendoim foi primeiramente avaliada por meio de estudos envolvendo a cinética de renaturação (Dhillon *et al*., 1980). De acordo com este trabalho, 11,9% do genoma é composto por sequências altamente repetitivas, 14,9% por medianamente repetitivas e 37,4% raramente repetitivas. O restante, 36%, é composto por sequências de cópia única ou genes.

6. Elementos de Transposição

Bárbara McClintock (1950), em um trabalho envolvendo grãos de milho variegados, descobriu que genomas eucariotos não são entidades estáticas, mas possuem elementos móveis ou transponíveis, que têm capacidade de se moverem de um lugar do cromossomo para outro. Nesse estudo foram identificados dois *loci* dominantes que interagiam entre si, mudando de posição no cromossomo, denominados *Ds* (*Dissociator*) e *Ac* (*Activator*). Os

efeitos da transposição foram constatados pela mudança na coloração dos grãos de milho. Ao analisar cruzamentos entre linhagens, McClintock concluiu que o *locus Ac* controlava a transposição de *Ds* e que o movimento de *Ds* promovia a ruptura do cromossomo 9. Quando *Ds* se excisava daquela posição, o gene responsável pela coloração na camada de aleurona era liberado do efeito supressor e voltava ao estado nativo, permitindo a síntese do pigmento. O efeito de mosaico de alguns grãos deveu-se à coexistência de células contendo o gene da coloração inativado pela inserção de *Ds* e células com o gene ativo. Esse estudo desafiou o conceito de que o genoma é um conjunto estático de informações transmitidas entre gerações.

A função dos TEs já foi amplamente debatida em termos evolutivos. Algumas características desses elementos, tais como persistência nos genomas, ausência de função aparente, geração de mutações deletérias, assim como utilização da maquinaria celular de replicação para a manutenção de suas cópias resultaram em alguns estudos, como o de Orgel & Crick (1980). Neste estudo foi proposto o conceito de "DNA egoísta" e, como a função evolutiva desses elementos não estava esclarecida, debates científicos focaram-se predominantemente na autopreservação desses elementos (Doolittle & Sapienza, 1980).

Durante muito tempo esses elementos de transposição foram considerados parasitas moleculares. Porém, como um novo paradigma, sugere-se que os TEs, como fonte genômica de mutações, possuam um papel fundamental na evolução dos genes e na estrutura genômica dos eucariotos (Kalendar *et al*., 2000; Cordaux *et al.,* 2006).

Evidências moleculares, sobretudo decorrentes dos projetos de sequenciamento total de genomas iniciados a partir de 1990, demonstraram que os TEs estão presentes praticamente em todos os eucariotos estudados. As únicas exceções até o momento são o *Plasmodium falciparum* e possivelmente as espécies relacionadas (Wicker *et al.,* 2007). O sequenciamento completo de vários genomas mostrou que os TEs são componentes majoritários dos genomas de eucariotos, compondo pelo menos 45% do humano (Lander *et al.,* 2001) e ultrapassando 80% em algumas espécies vegetais (Meyers *et al*., 2001).

Atualmente, os elementos de transposição são definidos como sendo fragmentos de DNA que possuem a capacidade de movimentação ao longo dos genomas por meio da transposição (Feschotte *et al*., 2002). Como consequência da sua atividade, os TEs podem gerar alterações funcionais quando inseridos dentro ou próximos às regiões promotoras, inativando ou alterando os padrões de expressão de genes (Kashkush *et al.,* 2003). A atividade desses elementos pode ainda ser responsável pela mobilização de segmentos de DNA durante a transposição, resultando na erosão da colinearidade entre genomas (Wicker *et*

al., 2010) e impulsionando a expansão e a evolução da estrutura genômica em plantas, seja em estado ativo ou inativo (Alix & Heslop-Harrison, 2004).

Exemplos de TEs envolvidos no desempenho de funções relevantes para o organismo hospedeiro já estão descritos na literatura. O alelo recessivo responsável pelo fenótipo rugoso observado por Mendel em sementes de ervilha é resultado de uma inserção de um elemento transponível em um gene que codifica uma enzima responsável pela concatenação de moléculas de amido. Esta inserção gera uma enzima não-funcional, diminuindo o acúmulo de amido nas sementes de ervilha (Bhattacharyya *et al.,* 1990). Em *Drosophila,* a manutenção dos telômeros é realizada por dois TEs, HeT-A e TART, que se transpõem exclusivamente na extremidade dos cromossomos (Pardue *et al.,* 2005). A recombinação feita pelos genes RAG1 e RAG2, responsáveis pela geração das imunoglobulinas, também está relacionada a processos de transposição (Zhou *et al*., 2004). Em *Arabiposis thaliana,* genes que codificam transposases estão relacionados a processos de desenvolvimento da planta (Bundock & Hooykaas, 2005) e às respostas da planta à incidência de luz (Hudson *et al*., 2003; Lin *et al*., 2007).

Dessa forma, elementos de transposicão frequentemente atuam como uma força relevante na evolução dos organismos, seja influenciando a regulação gênica ou alterando o tamanho do genoma, como consequência do tipo de transposição realizado por alguns elementos (Kidwell, 1997; Fedoroff, 2000). Isso foi demostrado no estudo realizado com uma família de retrotransposon denominada *Del*, que expandiu o genoma de pimenta para 2,7 Gb de tamanho (Park *et al*., 2012).

6.1 Classificação dos Elementos de Transposição

A primeira classificação para elementos de transposição, amplamente conhecida e ainda utilizada, foi a proposta por Finnegan (1989), que divide os TEs de acordo com seu intermediário de transposição: RNA (elementos de Classe I ou retrotransposons) ou DNA (elementos de Classe II ou transposons de DNA). Como consequência dos projetos de sequenciamento de genomas completos em larga escala, os estudos sobre a caracterização de elementos transponíveis aprofundaram-se, resultando na descoberta de outros tipos de TEs que não se encaixam necessariamente nessas duas classes. A descoberta de TEs bacterianos (Durval-Valentin *et al*., 2004) e eucarióticos (Morgante *et al*., 2005) que se "copiam e colam", sem intermediários de RNA, bem como de TEs portadores de sequências muito

pequenas com característica não-autônoma desafiaram esse sistema. Curcio & Derbyshire (2003) propuseram uma nova classificação baseada em propriedades enzimáticas das proteínas envolvidas no processo de transposição. No entanto, para tal, é necessário o conhecimento detalhado que é essencialmente compartilhado apenas por especialistas. No estudo de Wicker *et al*. (2007) foi proposto um sistema de classificação unificada para TEs de eucariotos, que concilia a divisão clássica dos elementos transponíveis com critérios enzimáticos e ainda propõe critérios objetivos para a anotação desses elementos, na qual as tradicionais classes são subdivididas em subclasses, ordens e superfamílias. No entanto, o sistema inicial com as duas principais classes de TEs é ainda amplamente utilizado.

A Classe I é composta pelos retrovírus e retrotransposons, elementos que se translocam no interior de um mesmo núcleo via intermediário de RNAm (RNA mensageiro), seguida de transcrição reversa e inserção da cópia de DNAc (DNA complementar) em novo sítio do genoma. A Classe II é composta pelos transposons de DNA, que possuem repetições terminais invertidas (TIR – *Terminal Inverted Repeats*) e codificam geralmente para uma única proteína, a transposase, responsável pela excisão do elemento e integração em outro sítio (Fleschotte *et al*., 2002). Elementos denominados MITEs (*Miniature Inverted-repeat Transposable Elements*) caracterizam-se por possuírem sequências curtas, com 100 a 600 pb, e sequências TIR em suas extremidades. MITEs não codificam a transposase. Porém a existência de transposons de DNA que codificam transposases funcionais e que compartilham similaridades com sequências de MITEs, levou a hipótese de que MITEs possam ser remanescentes derivados de deleções de transposons de DNA, mobilizados em *trans* pela ação das transposases (Fleschotte *et al*., 2002; Casacuberta & Santiago, 2003).

Em virtude do mecanismo replicativo de transposição, os retrotransposons são considerados os maiores contribuidores da fração repetitiva de grandes genomas (Kumar & Bennetzen, 1999; Sabot & Schulman, 2006). São divididos em cinco ordens de acordo com as características, organização e filogenia do gene que codifica para a transcriptase reversa: retrotransposons LTR (LTRs – *Long Terminal Repeats*), DIRS (*Dictyostelium intermediate repeat sequence*), PLEs (*Penelope-like elements*), LINEs (*Long interspersed nuclear elements*) e SINEs (*Short interspersed nuclear elements*) (Wicker *et al*., 2007).

Retrotransposons LTR são muito abundantes em genomas de plantas e possuem desde centenas de pares de bases até, excepcionalmente 25 Kb, como o elemento Ogre (Neumann *et al*., 2003). As sequências LTR que flanqueiam esses elementos possuem desde centenas de pares de bases até 5 Kb e possuem uma estrutura denominada TG..CA *box*, que corresponde

aos nucleotídeos TG na extremidade 5' do LTR 5' e CA na extremidade 3' do LTR 3' (Wicker *et al*., 2007). As regiões LTR 5' e LTR 3' são idênticas quando o elemento se insere no genoma hospedeiro. Uma vez inserido, essas sequências evoluem separadamente em virtude principalmente de mutações dos tipos inserção e deleção. Essas alterações nos LTRs possibilitam estimar a data de inserção do elemento no genoma hospedeiro. As regiões TSD (*Target Site Duplication*) são repetições curtas com 4-6 pb (pares de bases) que flanqueiam as extremidades 5' e 3' do elemento e que atuam como sinal de inserção. Próxima à região 3' terminal do LTR 5' há uma sequência com aproximadamente 18 pb denominada PBS (*Primer Binding Site*) que é complementar à cauda terminal 3' de alguns RNAt (RNA transportador) onde inicia-se a transcrição reversa. Outra região rica em segmentos de purina denominada PPT (*Polypurine Tract*) possui de 11-15 pb de comprimento, e assim como o PBS, é uma região envolvida na transcrição reversa (Xu & Wang, 2007).

As duas principais superfamílias de retrotransposons LTR são denominadas *Ty1-Copia* e *Ty3-Gypsy,* as quais distinguem-se pela similaridade entre as sequência e a ordem entre os genes (Boeke & Corces, 1989; Xiong & Eickbush, 1990; Wicker *et al*., 2007; Schulman, 2013). Os elementos autônomos contêm essencialmente dois genes necessários para a transposição, *gag* e *pol*. O gene *gag* codifica uma proteína similar ao capsídeo viral (VLP – *Virus-like Particle*) e *pol* a poliproteína responsável pelas atividades da protease (PR), envolvida no processo de maturação da poliproteína; integrase (IN), na integração da cópia de DNA no genoma hospedeiro; transcriptase reversa (TR) e RNA*se* H (RH) na síntese do DNAc a partir do RNAm (RNA mensageiro). Os retrovírus têm estrutura similar a um retrotransposon LTR, com uma ORF adicional contendo o domínio *env,* que lhes confere o caráter infeccioso (Xiong & Eickbush, 1990; Smyth, 1993).

Os retrotransposons LTR não-autônomos muitas vezes são derivados de elementos autônomos que sofreram deleções nas sequências ao longo da evolução (Witte *et al*., 2001) e podem utilizar enzimas sintetizadas por outros elementos para realizar a sua movimentação (Wessler, 2006; Wicker *et al*., 2007). São provavelmente derivados de mecanismos decorrentes de uma força de "compensação", visando minimizar o impacto do aumento do tamanho do genoma provocado pela amplificação dos retrotransposons. Conforme revisto por Vitte & Panaud (2005), a recombinação entre retrotransposons com grande número de cópias gera elementos dos quais restaram apenas LTRs ("LTRs-solo") ou elementos de diferentes tamanhos.

6.2 Mecanismo de replicação de retrotransposons LTR e controle da atividade

A estrutura e o ciclo de replicação dos retrotransposons LTR de plantas são semelhante àqueles dos elementos *Gypsy* e *Copia* de fungos e animais, bem como de retrovírus e retrovírus endógenos (Schulman, 2013). O LTR 5' direciona a transcrição e contém um promotor para a enzima RNA polimerase II. O LTR 3' fornece os sinais de terminação e poliadenilação. Esses promotores podem ser ativados por uma variedade de estresses bióticos e abióticos, bem como por tratamento hormonal a que plantas podem ser submetidas (Grandbastien *et al*., 2005; Ansari *et al*., 2007; Ramallo *et al*., 2008).

O modelo clássico de replicação de retrotransposons e retrovírus (figura 3) mostra que o RNA inteiramente transcrito pela RNA polimerase II codificada pelo hospedeiro, é transportado para o citoplasma, onde é traduzido. Diferentes proteínas são resultantes do processo de tradução, entre elas a integrase e a transcriptase reversa, que utiliza o RNA transcrito como molde para produção de uma molécula de DNAc. Essa molécula é direcionada para o núcleo e integrada aos cromossomos pela integrase, originando uma ou mais novas cópias do retrotransposon no genoma hospedeiro. Consequentemente, com esse tipo de processo de mobilização pode haver um aumento exponencial do número de cópias do retroelemento no genoma hospedeiro (Kumar & Bennetzen, 1999).

Tal mecanismo replicativo parece estar relacionado à grande abundância dos retrotransposons, bem como às diferenças nos tamanhos dos genomas nas diferentes espécies de plantas, enfatizando o papel central que os retroelementos desempenham na reestruturação evolutiva. Por outro lado, também é plausível a existência de mecanismos que controlem a atividade destes elementos na expansão dos genomas. A inativação dos elementos móveis é decorrente de mecanismos de regulação controlados tanto pelo hospedeiro, como pelo próprio elemento (Bennetzen, 2000). De maneira geral, o ciclo de replicação de retrotransposons LTR apresenta vários pontos de regulação e silenciamento epigenético de atividade como formas de controle. De fato, os retrotransposons presentes em plantas podem ser silenciados durante a transcrição por meio de metilação do DNA (Cui *et al*., 2013) e modificação da cromatina (Eichten *et al*., 2012) e também pós-transcricionalmente (McCue *et al*., 2012). Tal como acontece com vírus, alguns retrotransposons também são capazes de se evadir do silenciamento (Hernández-Pinzón *et al*., 2012). Alguns trabalhos evidenciaram a importância da tradução balanceada entre *gag* e *pol* para que ocorra transposição, já que para o empacotamento de um único cDNA correspondente a uma nova cópia de retrotransposon são

necessários vários peptídeos de *gag* (Gao *et al.,* 2003).

A preferência por sítios de inserção também pode exercer um importante controle na atividade de transposição. Apesar de alguns grupos de retrotransposons possuírem sítios preferenciais de inserção, tais como centrômeros e telômeros, muitos deles encontram-se em regiões não gênicas ou de heterocromatina que, normalmente, apresentam baixa atividade de transcrição (Kumar & Bennetzen, 1999). Em outros casos, os elementos podem se inserir dentro de outros (*nested transposons*) ou em íntrons, assim evitando qualquer ativação do sistema de proteção do hospedeiro (Martienssen, 1998). Além disso, qualquer modificação na sequência dos elementos pode acarretar em inativação, tal como a perda de LTRs, o que impede a inserção em um novo local do genoma. Elementos que sofreram deleção total na sequência, também podem ser praticamente eliminados do genoma ao longo da evolução (Vitte & Panaud, 2005). Entretanto, devido ao seu mecanismo de replicação utilizando um intermediário de RNA, o modo mais eficiente de controle da atividade, é provavelmente durante a transcrição.

A transcrição de um retrotransposon não significa necessariamente que ele é ativo, ou seja, que ele esteja efetivamente inserindo-se em uma nova região do genoma. Por isso, o número de retrotransposons LTR com caracterização detalhada sobre sua transposição é relativamente limitado, levando-se em conta a representatividade deste grupo de sequências no genoma eucarioto, especialmente das plantas. Alguns artigos mostraram que eventos de transposição observados nos elementos *Tnt1* de fumo (Grandbastien *et al.,* 1989) e *Tos17* de arroz (Hirochika *et al.,* 1996) possivelmente estão relacionados a algum tipo de estresse, tal como a indução de cultura de células como calos ou protoplastos. A indução de calos também provoca a atividade transposicional de retrotransposons LTR em batata-doce (Tahara *et al.,* 2004) e *Lotus japonicus* (Fukai *et al.,* 2008).

Geralmente, o nível transcricional é considerado baixo para a maioria dos TEs (Jiao & Deng, 2007) e poucas famílias de retrotransposons LTR apresentam transcrição ubíqua (Madsen *et al.,* 2005; Jiao & Deng, 2007). Análises transcricionais baseadas em banco de dados de ESTs (*Expressed Sequence Tags*) evidenciam que, embora sejam componentes predominantes no genoma vegetal, sequências relacionadas a retrotransposons LTR correspondem a um baixo percentual do total de transcritos (Meyers *et al.,* 2001; Echenique *et al.,* 2002; Vettore *et al.,* 2003; Lopes *et al.,* 2008).

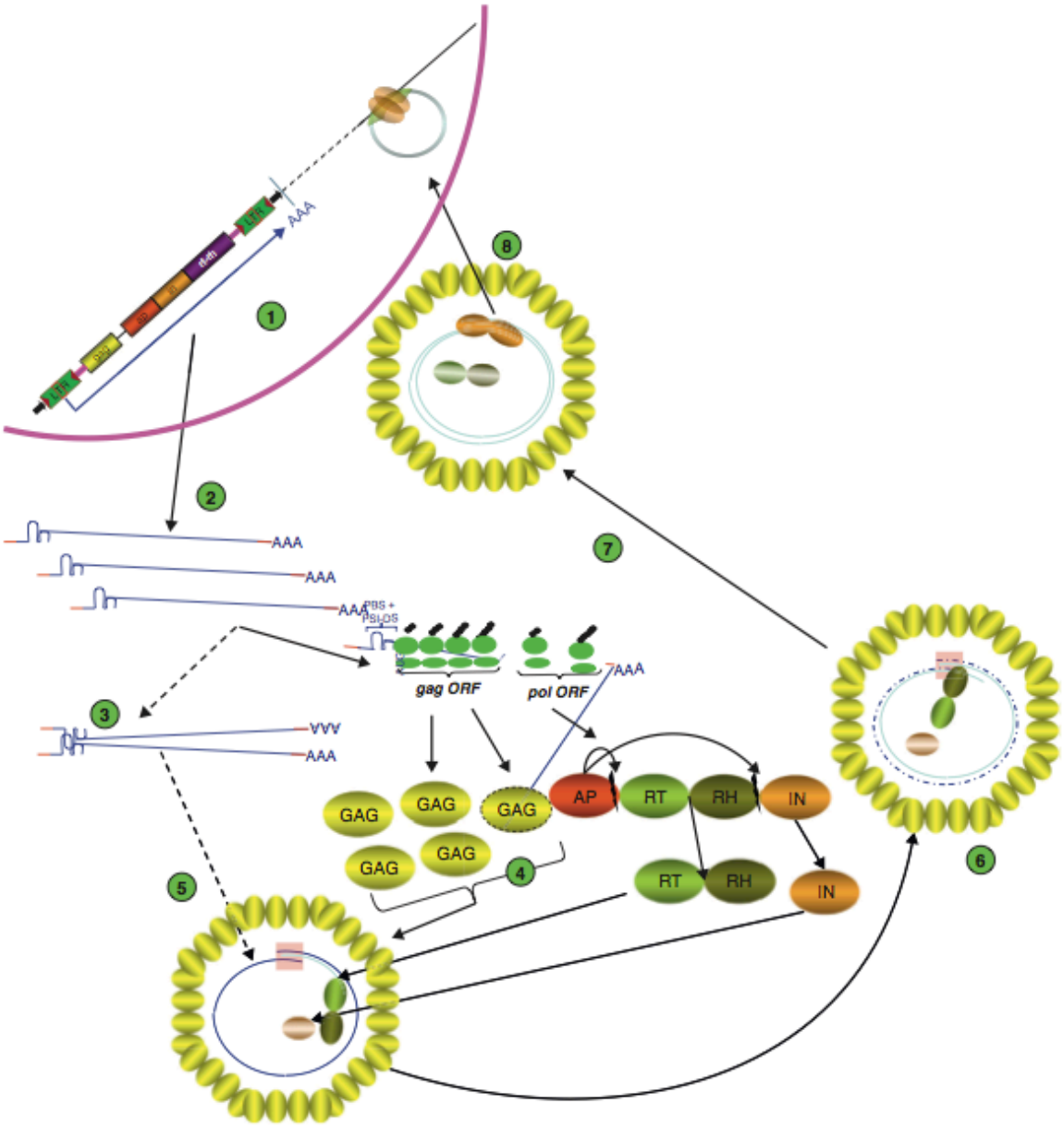


Figura 3: Ciclo de vida de um retrotransposon LTR. Um elemento da superfamília *Ty1-Copia* é mostrado integrado no genoma, dentro do núcleo (linha rosa). As etapas são as seguintes: (1) transcrição de uma cópia integrada no genoma a partir do promotor situado na repetição longa terminal (LTR); (2) exportação nuclear; (3) tradução ou, alternativamente, empacotamento de dois transcritos em partículas semelhantes a vírus (VLP); (4) tradução e síntese de proteínas das regiões *gag* e *pol*; *pol* inclui as proteínas PR (protease), TR (transcriptase reversa), RH (RNA*se* H) e IN (integrase); (5) Junção de VLPs a partir de *gag* contendo transcritos de RNA, IN, TR-RH; (6) transcrição reversa pela TR; (7) direcionamento de VLP para o núcleo; (8) passagem do cDNA - complexo IN para o núcleo e a integração do DNAc no genoma. Adaptado de Schulman, 2013.

6.3 Diversidade e localização cromossômica dos retrotransposons

O conteúdo de DNA de um organismo não reflete diretamente sua complexidade (Thomas, 1971) e a amplificação dos elementos de transposição nos genomas pode explicar em parte, esse fato. Um exemplo dessa questão pode ser observado na diferença entre os

tamanhos dos genomas de milho e sorgo, onde se sugere que o grande acúmulo de retrotransposons LTR no milho seja responsável pela diferença (Estep *et al*., 2013). Comparações realizadas com a região do gene *adh1* em milho e sorgo revelaram que em sorgo não há presença de retrotransposons nesta região, inferindo-se que os retrotransposons foram inseridos no genoma do milho após a divergência entre estas duas espécies (SanMiguel *et al*., 1998). No caso de duas espécies de arroz, *Oryza sativa* e *O. australiensis*, a variação no número de cópias no elemento RIRE1 pode explicar a diferença de aproximadamente um terço no tamanho desses genomas (Vicient & Schuman, 2002).

Durante muito tempo, a indisponibilidade de sequências genômicas completas foi um fator limitante para análises evolutivas de retrotransposons. O trabalho de Xiong e Eickbush (1990) foi pioneiro para a compreensão da relação entre retrovírus e retroelementos, a partir da análise de TR de diferentes organismos. Desde então, a maioria dos trabalhos envolvendo características evolutivas dos TEs focavam-se em uma única espécie ou em um pequeno grupo de espécies relacionadas (Wicker *et al.,* 2007), o que dificultava a resolução de linhagens evolutivas. Somente a partir da disponibilidade da sequência completa de genomas, tornou-se possível investigar com mais detalhes as relações evolutivas dentro de cada grupo de TEs (Domingues *et al*., 2012; Kolano *et al*., 2013; Estep *et al*., 2013).

Devido à vasta diversidade e heterogeneidade de retrotransposons, acredita-se que sejam antigos e que possivelmente estavam presentes em espécies ancestrais. Entretanto, em alguns casos, é difícil determinar a história evolutiva de uma espécie com base na distribuição de retroelementos, principalmente em virtude da origem complexa e das diferenças de ploidia. Os eventos de poliploidização causam, além do aumento no tamanho do genoma, a amplificação em larga escala dos retrotransposons (Heslop-Harrison, 2000). As altas taxas de erros de correção da enzima RNA polimerase em regiões repetitivas e da transcriptase reversa, facilitam o surgimento de mutações nas sequências dos retroelementos, o que permite que eles se diferenciem rapidamente, gerando diversidade (Bennetzen, 2000).

A distribuição cromossômica dos TEs já foi investigada em diversas espécies de plantas por meio de técnicas de citogenética molecular, como FISH. Os retrotransposons das superfamílias *Ty1-Copia* e *Ty3-Gypsy* foram majoritariamente identificados em plantas como *Helianthus* spp. (Natali *et al.,* 2006), *Malus domestica* (Sun *et al.,* 2008) e *Pisum sativum* (Macas *et al.,* 2007). Conforme visto por Kumar & Bennetzen (1999), a distribuição de retrotransposons em cromossomos metafásicos pode ser uniforme ou não, dependendo da espécie e da família analisada. Geralmente, os retroelementos apresentam uma distribuição

dispersa, que reflete seu mecanismo de transposição, sendo encontrados entre outras sequências repetitivas ou associados a regiões genômicas específicas (Heslop- Harrison, 2000).

Belyayev *et al.* (2005) observaram que em espécies da família Triticaceae ocorria preferencialmente um padrão agrupado de distribuição de retrotransposons em algumas regiões cromossômicas. No entanto, em *Brassica oleracea* (Alix *et al.,* 2005), a organização genômica de retroelementos apresentou uma distribuição mais dispersa. Existem exemplos de famílias de retroelementos que possuem sítios de inserção preferenciais. Na espécie *Allium cepa* há uma família de retrotransposons preferencialmente localizada em telômeros (Pearce *et al.,* 1996) e em gramíneas, há um grupo de retrotransposons do tipo Ty3-*Gypsy* presentes exclusivamente em centrômeros (Miller *et al.,* 1998). O sequenciamento de grandes fragmentos genômicos derivados de clones de bibliotecas BAC de trigo identificou elementos em regiões intergênicas (SanMiguel *et al.,* 1996), e o uso desses fragmentos como sonda em FISH permitiu a localização de regiões cromossômicas ricas em TEs (Zhang *et al.,* 2004). Natali *et al.* (2006) analisaram a localização *in situ* das sequências repetitivas pHaS13 e pHaS211 com similaridade aos genes IN (retrotransposon *Ty3-Gypsy*) e RH (retrotransposon *Ty1-Copia*), respectivamente, em espécies do gênero *Helianthus*. Apesar da distribuição dispersa em todo o complemento cromossômico para ambos, a sequência relativa à integrase apresentou uma localização preferencialmente centromérica em todas as espécies analisadas enquanto que, em *Helianthus annuus*, o fragmento da RNA*se* H distribuiu-se nas regiões terminais dos cromossomos.

Muitas vezes, a localização preferencial de um elemento pode ser "mascarada" em virtude da mudança de posição após eventos de recombinação, mas de maneira geral, esses elementos podem ser encontrados ao longo de todo o cromossomo. Portanto, ainda não há um padrão determinado de distribuição preferencial de retrotransposons (Heslop-Harrison *et al.,* 1997; Kumar & Bennetzen, 1999).

6.4 Elementos de transposição em espécies de *Arachis*

Transposons de DNA do tipo MITE já foram identificados em mutantes de amendoim que apresentavam alta relação de ácido oleico:ácido linoleico (HO – *High Oleic*). Essa característica deve-se a mutações distintas identificadas em dois genes homeólogos, *ahFAD2A e ahFAD2B*. No trabalho realizado por Patel *et al*. (2004) ficou constatada a

inserção de um elemento MITE em *ahFAD2B,* causando a mutação e consequente produção de uma proteína diferenciada. Essa, em conjunto com a mutação identificada no outro gene homeólogo, gerava o aparecimento desse novo fenótipo HO. Em um trabalho posterior, Shirasawa e colaboradores (2012) investigaram a diversidade genômica e a taxa de transposição de AhMITE1, revelando que esse transposon de DNA pode ter afetado a expressão de vários genes, desde a origem do amendoim cultivado.

Em outro estudo realizado por Nielen *et al*. (2010), o primeiro retrotransposon de amendoim pertencente à superfamília *Ty3-Gypsy,* denominado FIDEL (*Fairly long Inter-Dispersed Euchromatic LTR retrotransposon*), foi descrito e caracterizado como sendo mais frequente no componente genômico A do que no B. Experimentos com FISH revelaram uma distribuição dispersa nas regiões de eucromatina e ausente em centrômeros, regiões teloméricas e região organizadora nucleolar. Esta distribuição desigual é possivelmente resultante da atividade de transposição ocorrida de forma mais conspícua no genoma A, provavelmente após a divergência evolutiva das espécies parentais.

Nielen *et al*. (2012), também descreveram posteriormente, um retrotransposon moderadamente repetitivo, pertencente à superfamília *Ty1-Copia* e denominado Matita. Em contraste com FIDEL, Matita é igualmente abundante nos componentes genômicos A e B, porém com distribuições cromossômicas distintas. Em contraste com FIDEL, Matita está localizado principalmente nas regiões distais dos braços dos cromossomos. Por meio de estudos realizados com clones das bibliotecas BAC construídas para as espécies parentais do amendoim (Guimarães *et al*., 2008), foi mostrado que a distribuição de Matita exibe uma tendência de ser mais abundante perto de genes de resistência do que de genes de cópia única (Nielen *et al*., 2012). Todos estes estudos "acenderam" a importância dos TEs na organização e evolução do genoma de amendoim.

7. Ferramentas para estudos genômicos em *Arachis*

A biotecnologia vegetal aliada aos métodos de melhoramento clássico oferecem muitas aplicações e vantagens no estudo e na produção de plantas com características agronomicamente favoráveis, como resistência a pragas, tolerância a estresses abióticos, aumento no valor nutritivo, resistência a pesticidas, dentre outros. Essas aplicações oferecem grandes benefícios para a indústria alimentícia, consumidor, meio ambiente e para o produtor, gerando maior qualidade e rendimento nas safras.

Na última década, grandes avanços foram alcançados na compreensão do genoma de amendoim, mesmo sem a disponibilização da sequência genômica completa. Somando-se às análises citogenéticas (Seijo *et al*., 2004; 2007), várias ferramentas foram construídas para auxiliar os estudos envolvendo amendoim, como o desenvolvimento de mapas genéticos utilizando espécies silvestres (Halward *et al*., 1993; Garcia *et al*., 2005; Burow *et al*., 2001; Moretzsohn *et al*., 2005; 2009; Leal-Bertioli *et al*., 2009; Foncéka *et al*., 2009) e cultivada (Herselman *et al*., 2004; Varshney *et al*., 2009; Hong *et al*., 2010; Khedikar *et al*., 2010; Gautami *et al*., 2012; Qin *et al*., 2012), dentre outros. Esses estudos possibilitaram a comparação entre os mapas diploides e tetraploides, evidenciando a sintenia entre os genomas A e B de *Arachis* e a constatação da ocorrência de poucos rearranjos cromossômicos. Um grande número de marcadores também foi desenvolvido para o amendoim nos últimos anos, sendo que alguns deles estão associados à resistência à ferrugem e mancha preta (Mace *et al*., 2006) e à queima-de-Sclerotinia (Chenault *et al*., 2009), por exemplo.

Estudos envolvendo o mapeamento comparativo entre *Arachis* e outras leguminosas-modelo também já foram realizados. Em 2008, Hougaard e colaboradores utilizaram marcadores âncoras desenvolvidos para leguminosas para verificar a sintenia entre as espécies *Phaseolus vulgaris*, *Lotus japonicus*, *Medicago truncatula* e *Arachis*. O alto nível de sintenia entre *Arachis*, *Lotus* e *Medicago* também foi verificado a partir de outro estudo desenvolvido por Bertioli *et al*. (2009), constatando que a presença de TEs e a sintenia observada entre esses genomas apresentaram uma correlação negativa e mostrando que regiões evolutivamente mais conservadas tendem a apresentar baixa densidade de TEs, ao passo que regiões variáveis tendem a apresentar maior densidade de TEs.

A avaliação da expressão de genes em determinadas situações de estresse também tem produzido resultados significativos em amendoim, auxiliando a entender algumas vias de regulação, além de respostas aos mais variados processos biológicos (Zhang *et al*., 2012; Feng *et al*., 2012; Wu *et al*., 2013; Yin *et al*., 2013). Outros estudos envolveram a identificação de genes responsivos especificamente ao estresse hídrico no transcriptoma de amendoim (Guo *et al*., 2006; Govind *et al*., 2009; Kottapalli *et al*., 2009; Guimarães *et al*., 2012) ou estresse biótico (Proite *et al*., 2007; Guimarães *et al*., 2010).

Outra ferramenta amplamente utilizada no melhoramento de plantas é a transformação genética. A primeira planta transgênica de amendoim foi obtida por Ozias-Akins e colaboradores (1993). Mais tarde, plantas geneticamente modificadas de amendoim expressando o gene do nucleocapsídio de TSWV (*Tomato Spotted Wilt Virus*) apresentaram

altos níveis de tolerância à infecção pelo vírus (Yang *et al*., 2004). Plantas de amendoim contendo o gene da defensina de mostarda também exibiram tolerância aos fungos causadores de cercosporioses (Anuradha *et al*., 2008). Em outro caso, plantas expressando o gene cry1EC de *Bacillus thuringiensis* exibiram resistência a *Spodoptera litura* (Keshavareddy *et al*., 2013). Em outro estudo desenvolvido por Qiao *et al*. (2014), que descreve a clonagem e a transformação genética do gene da β-1,3-glucanase em plantas de amendoim, as plantas transgênicas produzidas apresentaram mais resistência ao fungo que causa a mancha preta.

Nos últimos anos, três bibliotecas BAC foram construídas para o gênero *Arachis*, uma para o amendoim, *A. hypogaea,* com genoma AABB (Yüksel & Paterson, 2005) e outras duas para as prováveis espécies progenitoras do amendoim *A. duranensis* (genoma A) e *A. ipaënsis* (genoma B) (Guimarães *et al*., 2008). Essas bibliotecas são consideradas fundamentais na construção de mapas físicos e no isolamento de sequências de interesse. Clones BAC derivados da biblioteca de *A. duranensis* já forneceram dados importantes acerca do conteúdo repetitivo de *Arachis* (Nielen *et al*., 2010; 2012).

Recentemente, em abril de 2014, foi divulgado o sequenciamento dos genomas dos progenitores do amendoim, *A. duranensis* e *A. ipaënsis*, resultado do esforço de um grande grupo de cientistas dos Estados Unidos, China, Brasil, India e Israel. A Iniciativa Internacional do Genoma do Amendoim (IPGI) reuniu cientistas com os objetivos de sequenciar o genoma, caracterizar a variação genética e fenotípica do amendoim cultivado e seus parentes silvestres e desenvolver ferramentas genômicas para o melhoramento do amendoim. O sequenciamento inicial foi realizado pelo BGI, Shenzen, China. A montagem foi feita no BGI, USDA- ARS, Ames - IA e UC Davis, CA. O projeto foi viabilizado pelo financiamento fornecido pela indústria de amendoim através da Peanut Foundation, pela MARS Inc., três Academias chinesas, dentre outras (http://www.peanutbioscience.com).

Além de servirem como arcabouço para a montagem do genoma do amendoim cultivado, o sequenciamento das espécies silvestres progenitoras também possibilitará decifrar as alterações genômicas que levaram à domesticação do amendoim, marcada por aumentos no tamanho das sementes e mudanças no hábito de crescimento da planta. As sequências genômicas e informações adicionais estão disponíveis em http://peanutbase.org/files/genomes/.

Objetivo Geral

Analisar os conteúdos repetitivo e gênico nos genomas dos parentais silvestres diploides do amendoim, *Arachis duranensis* (genoma A) e *Arachis ipaënsis* (genoma B), e sua distribuição nos cromossomos dos subgenomas A e B do amendoim tetraploide (AABB) por meio de ferramentas de bioinformática e hibridização *in situ* por fluorescência (FISH), bem como identificar genes de interesse pela utilização de *pools* de BAC.

Objetivos específicos:

- Identificar novas sequências de retrotransposons LTR em clones da biblioteca BAC e no genoma de *A. duranensis* (genoma A);
- Identificar novas sequências de retrotransposons LTR no genoma de *A. ipaënsis* (genoma B);
- Classificar os retrotransposons LTR em famílias e superfamílias;
- Estimar a frequência gênica e a data de transposição das famílias de retrotransposons LTR nos genomas de *A. duranensis* e *A. ipaënsis*;
- Avaliar a distribuição de algumas famílias de retrotransposons LTR em cromossomos A e B de amendoim;
- Comparar conteúdos repetitivo e gênico em regiões homeólogas dos genomas A e B das espécies silvestres e *A. hypogaea*;
- Identificar sequências codantes de interesse em *pools* de BAC;

Capítulo 1

O componente repetitivo do genoma A de amendoim (*Arachis hypogaea*) e seu papel no remodelamento do espaço de sequências intergênicas desde a divergência evolucionária do genoma B

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

1. Introdução

O amendoim (*Arachis hypogaea* L.) pertence à família Fabaceae ou Leguminosae e é originário da América do Sul. Ele é cultivado nos trópicos e subtrópicos e é muito importante na Ásia e África. Sua produção anual chega a aproximadamente 37 milhões de toneladas (USDA-FAS, 2013). Dentro da subfamília Papilionoideae, o amendoim pertence a um clado denominado Dalbergioides, predominante nos trópicos do Novo Mundo (Lewis *et al*., 2005; Lavin *et al*., 2001) e evolutivamente separado da maioria das demais leguminosas economicamente importantes há aproximadamente 55 milhões de anos.

A maioria das espécies do gênero *Arachis* possui 20 cromossomos (2*n*=2x=20), porém o amendoim é uma exceção, com 40 cromossomos (2*n*=4x=40). Trata-se de um alotetraploide recente, provavelmente resultante da hibridização entre duas espécies silvestres, seguida por uma duplicação cromossômica espontânea (Halward *et al*., 1991; Young *et al*., 1996; Seijo *et al*,. 2004; 2007).

O tamanho estimado do genoma do amendoim é 2.8 Gb (Greilhuber, 2005), contendo uma fração repetitiva significativa de aproximadamente 64%, determinada por renaturação cinética do DNA (Dhillon *et al*., 1980). Análises citogenéticas do amendoim revelaram dois tipos de cromossomos: 10 pares de cromossomos do tipo A, que apresentam a heterocromatina do centrômero fortemente corada com *4',6-diamino-2-phenylindole* - DAPI (portanto ricos em AT), incluindo o menor par de cromossomos (Husted, 1936; Smartt *et al*., 1978) e outros 10 pares de cromossomos designados de B, com bandas heterocromáticas centroméricas mais fracamente coradas por DAPI (Smartt *et al*., 1978; Smartt & Stalker, 1982; Seijo *et al*., 2004; Robledo & Seijo, 2010).

Estudos dos padrões de coloração de bandas heterocromáticas, localização de *clusters* de DNAr (DNA ribossômico) nos cromossomos (Seijo *et al*., 2007; Robledo *et al*., 2009; Robledo & Seijo, 2010) juntamente com a hibridização genômica in situ – GISH (*Genomic In Situ Hybridization*) (Seijo *et al*,. 2007), sugerem que os cromossomos A de *A. hypogaea* são similares àqueles presentes na espécie diploide silvestre *A. duranensis*, enquanto que os cromossomos B são similares aos de *A. ipaënsis*, outro silvestre também diploide. Outras evidências, como a distribuição geográfica das espécies (Robledo *et al*., 2009; Robledo & Seijo, 2010) e filogenia molecular (Kochert *et al*., 1996; Burow *et al*., 2009; Moretzsohn *et al*., 2013) corroboram que os mais prováveis doadores dos genomas A e B do amendoim são respectivamente *A. duranensis* e *A. ipaënsis*.

O pareamento dos cromossomos em *A. hypogaea* durante a meiose é quase que inteiramente bivalente (Smartt, 1990), uma indicação da divergência genética entre os subgenomas A e B e do potencial controle genético do pareamento dos cromossomos. Diferenças no conteúdo repetitivo dos componentes genômicos A e B também foram evidenciadas por GISH, utilizando-se como sonda o DNA genômico, o que possibilitou a distinção entre os genomas (Seijo *et al*., 2007), além de mostrar a distribuição preferencial do retroelemento FIDEL (Nielen *et al*., 2010). Em contraste, os marcadores moleculares ligados à fração do genoma com baixo número de cópias mostraram alta homologia e poucos rearranjos estruturais entre os genomas A e B. Essa alta colinearidade entre genes observada por marcadores em mapas genéticos indica uma baixa divergência evolutiva entre esses genomas (Burow *et al*., 2001; Moretzsohn *et al*., 2009; Bertioli *et al*., 2009; Shirasawa *et al*., 2013), estimada em torno de 3 a 3,5 milhões de anos atrás (Nielen *et al*., 2012; Moretzsohn *et al*., 2013).

Essas evidências propõem um paradoxo aparentemente intrigante em relação à evolução da estrutura genômica do amendoim: a fração repetitiva de DNA predominante no genoma encontra-se em fluxo evolucionário, enquanto que, ao mesmo tempo, o DNA com baixo número de cópias permanece conservado.

Para estudar detalhada e separadamente a divergência e evolução dos genomas A e B, bibliotecas BAC (*Bacterial Artificial Chromosome*) das espécies silvestres *A. duranensis* e *A. ipaënsis* foram construídas (Guimarães *et al*., 2008). No presente trabalho essas bibliotecas foram utilizadas para investigar o componente repetitivo do subgenoma A de *Arachis*, mais especificamente, os retrotransposons LTR, a fim de entender os processos evolutivos ocorridos durante a divergência entre os subgenomas A e B em amendoim.

2. Material e Métodos

2.1 Seleção de clones BAC

Em trabalhos anteriores, 27 clones BAC pertencentes à biblioteca de *A. duranensis* - acesso V14167 (genoma A) foram selecionados com base na hibridização com sondas construídas a partir de marcadores ligados a genes de cópia única ou poucas cópias presentes em genomas diploides de leguminosas (Choi *et al*., 2006; Fredslund *et al*., 2006). Hibridizações *in situ* por fluorescência (FISH) utilizando esses clones como sondas foram realizadas em cromossomos metafásicos obtidos a partir de meristema radicular jovem de amendoim (Cultivar "Tatu"). Para muitos clones, o resultado do FISH produziu sinais de hibridização múltiplos e dispersos em vários cromossomos (preferencialmente nos cromossomos A), apesar do uso de altas concentrações de DNA de bloqueio (C_0t 100). Hibridizações mostrando dois pontos em um ou mais pares de cromossomos, tal como esperado quando se utilizam sondas contendo sequências de DNA únicas ou com baixo número de cópias (genes), não foram detectadas. Portanto, concluiu-se que todos os clones BAC utilizados como sonda para FISH continham sequências de DNA altamente repetitivas (Araújo *et al*., 2012).

A partir desses resultados, 13 desses clones BAC da biblioteca de *A. duranensis* foram selecionados para sequenciamento e anotação, juntamente com dois outros clones, sendo um do genoma de *A. ipaënsis* (genoma B) e outro do subgenoma ou componente genômico B de *A. hypogaea*. Também foi selecionado para anotação outro clone BAC da biblioteca A já sequenciado em um estudo anterior (Nielen *et al*., 2010).

2.2 Isolamento de clones BAC

O DNA dos 13 clones selecionados foi isolado separadamente a partir de uma colônia da placa original (384 poços) da biblioteca BAC de *A. duranensis*. Culturas foram desenvolvidas em placas de Petri contendo 20 mL de meio sólido LB (Lúria Bertani – 1% de Triptona; 0,5% de Extrato de Levedura; 1% de NaCl; 1,5% de Ágar-ágar), acrescidos de 20 µL do antibiótico cloranfenicol (12,5 mg/µL). Para cada placa apenas uma colônia de bactéria foi selecionada e utilizada para nova cultura em 3 mL de meio LB contendo 3 µL de cloranfenicol (12,5 mg/µL). A incubação foi realizada a 37° C por 12 horas sob rotação em

torno de 225 rpm.

Para isolar o DNA, a cultura foi centrifugada a 13.000 rpm por 1 minuto. O sobrenadante foi descartado e o precipitado ressuspenso em 200 µL de tampão a 4° C (glicose 50 mM; Tris-HCl 1M pH 8; EDTA 500 mM; 100 µg de RNA*se* A). A solução foi homogeneizada, acrescida de 200 µL de tampão de lise (NaOH 200 mM; SDS 1%) e misturada gentilmente por 2 minutos à temperatura ambiente. Foram adicionados 200 µL de tampão de neutralização (acetato de potássio 1,32 M, pH 4,8) e o tubo, incubado em gelo. Após 3 minutos, a solução foi centrifugada a 13.000 rpm por 15 minutos a 4° C e o sobrenadante transferido para novo tubo. Foram adicionados 400 µL de isopropanol, misturando-se os tubos gentilmente. As amostras foram então centrifugadas a 13.000 rpm por 20 minutos e o sobrenadante descartado. O precipitado foi ressuspenso em 100 µL de tampão TE (Tris-HCl 10 mM; EDTA 1 mM), precipitado com 10 µL de acetato de sódio 3 M (4° C) e 250 µL de etanol absoluto a -20° C e mantido em freezer a -80° C por 1 hora seguida de centrifugação por 20 minutos a 4° C. O precipitado (DNA) foi lavado com 500 µL de etanol 70% a 4° C e seco completamente por evaporação. Foram adicionados 30 µL de água *MilliQ* autoclavada ao DNA, que foi posteriormente armazenado a -20° C.

2.3 Sequenciamento e montagem de clones BAC

O sequenciamento dos clones BAC selecionados foi realizado utilizando-se a técnica de fragmentação por *shotgun* e o método Sanger (Sanger *et al*., 1977) ou método 454 (Roche Applied Science - http://454.com/).

Para o sequenciamento pelo método Sanger, 768 subclones foram obtidos e sequenciados para cada clone BAC (2 placas com 384 poços cada). A montagem das sequências foi feita utilizando o *software* CAP3 (Huang & Madan, 1999). Os consensos foram produzidos e editados manualmente utilizando o *software* Consed (Gordon *et al*., 1998).

O outro sequenciamento foi feito utilizando-se o Sistema Roche 454GS-FLX com a química de titânio realizada pela GATC Biotech AG, Konstanz, Alemanha. As amostras foram sequenciadas em placa FLX Pico-Titer utilizando a química GS FLX Titanium XLR70. Os dados das sequências foram produzidos no formato *Flowgram* padrão para cada leitura e a montagem foi realizada utilizando GS De Novo Assembler (aka Newbler, o GS FLX *System Software*) com os parâmetros padrão.

2.4 Anotação de retrotransposons LTR do genoma A

Sequências de todos os clones BAC comparadas contra elas mesmas, por meio de *dot plots* obtidos pelo *software* Gepard (Krumsiek *et al*., 2007), permitiram a identificação de sequências repetitivas independentemente da similaridade com quaisquer elementos previamente identificados. Sequências de retrotransposons LTR já identificadas em amendoim (Nielen *et al*., 2010; 2012) também foram comparadas contra as sequências recém obtidas. A maioria dos retrotransposons presente em plantas e animais contém repetições terminais longas designadas LTRs (*Long Terminal Repeats*). Essas sequências LTRs representadas em gráfico de plotagem produzem um padrão de pontos que facilita visualmente a identificação de sequências de retrotransposons completas. Além disso, retrotransposons LTR também foram identificados utilizando o *software* LTR_FINDER (Xu & Wang, 2007).

A fim de visualizar o conteúdo repetitivo dos clones BAC, gráficos contendo um "Índice Repetitivo" foram desenvolvidos a partir do número de similaridades identificadas quando as sequências dos clones foram comparadas a 42.000 sequências derivadas de sequenciamento de BAC ENDs de *A. duranensis* via BLASTn (Altschul *et al.,* 1997). Os parâmetros utilizados foram "-e 1e-20; –m 8". O resultado obtido foi organizado em tabela e um *script Perl* (http://perl.org.br/) foi utilizado para produzir um índice para cada base de DNA, calculado da seguinte forma: Índice repetitivo = log10(N), onde N é o número de similaridades detectadas pelo BLASTn.

2.5 Anotação de genes

Para anotação de genes, alguns *softwares* disponíveis em domínio público foram utilizados: FGENESH (Salamov & Solovyev 2000); hmmsearch contra a biblioteca A do *Pfam* (Eddy, 2011); BLAST, utilizando proteínas presentes em bancos de dados contendo sequências gênicas de soja e Arabidopsis, ESTs de *Arachis* e 42.000 sequências de *A. duranensis* (sequências GSS, e dois acessos pertencentes ao Genbank: FI321525 e FI281689) e LTR_FINDER. Os resultados foram visualizados no *browser* para anotação genômica denominado Artemis (Rutherford *et al*., 2000). Vários arquivos de saída obtidos nas análises utilizando esses *softwares* foram convertidos para o formato Genbank utilizando *scripts Perl* "-m 8", conforme necessário.

2.6 Publicação de sequências

Todas as sequências contidas nesta publicação foram depositadas no banco de dados *The European Nucleotide Archive* (ENA – disponível em http://www.ebi.ac.uk) com o número de estudo ERP002436.

2.7 Alinhamento de LTRs e estimativa da data de transposição dos retrotransposons

As sequências relativas aos dois LTRs de cada elemento identificado foram separadas e alinhadas utilizando o *software* Muscle (Edgar, 2004), que possibilitou a identificação dos eventos de mutação, tais como inserções, deleções ou substituições, ocorridos em cada LTR e necessários para o cálculo da data de transposição dos elementos. As datas de transposição foram estimadas para retrotransposons LTR com sequência completa pelo método de divergência das porções LTR, utilizando a equação t = K/2r, onde t é a idade; K é o número de substituições de nucleotídeos por local entre cada par de LTR; e r é a taxa de substituição de nucleotídeo de $1.3x10^{-8}$ por sítio por ano, como descrito no estudo publicado por Ma & Bennetzen (2004). Esse cálculo e a organização dos dados foram feitos utilizando um *script Perl*.

2.8 Análise de frequência dos retrotransposons LTR no genoma A

Um banco de dados contendo as sequências de todos os elementos identificados foi elaborado utilizando a ferramenta BLASTn. Com o auxílio de um *script* foram realizadas comparações das sequências do banco de dados com os arquivos de texto em formato FASTA, relativos às sequências dos clones BACs. O *e-value* utilizado foi $1x10^{-40}$. Os dados resultantes foram organizados em tabela e inspecionados em Excel (Microsoft) para os cálculos referentes à estimativa da porcentagem de ocorrência de cada retrotransposons LTR no genoma A de *A. duranensis*.

2.9 Comparação entre sequências homeólogas nos genomas A e B de *Arachis*

Para realizar uma análise comparativa entre sequências homeólogas derivadas de clones BACs pertencentes aos genomas de *A. duranensis* (acesso V14167 - genoma AA) e *A.*

ipaënsis (acesso KG30076 - genoma BB) (Guimarães *et al*., 2008), e/ou de *A. hypogaea* (Cultivar "Florunner" - genoma AABB - Yüksel & Paterson 2005), *screenings* foram realizados nessas bibliotecas utilizando duas sondas gênicas desenvolvidas com base na sequência do gene que codifica a enzima DNA girase (marcador Leg128; Fredslund *et al.,* 2006) e o alérgeno *Ara h1* do amendoim. As sequências dos clones identificados foram comparadas em gráficos de plotagem e *software* Artemis.

3. Resultados

3.1 Seleção e sequenciamento de clones BAC

Um total de 16 clones BAC de *Arachis* spp. foi caracterizado quanto aos conteúdos repetitivo e gênico (tabela 1). Desses, 15 foram sequenciados (13 clones BAC de *A. duranensis* (genoma A) e dois clones do genoma B, um de *A. ipaënsis* e outro do subgenoma B de *A. hypogaea*. O clone ADH180A21 (genoma A) já havia sido sequenciado em um estudo anterior (Nielen *et al*., 2010).

O protocolo de extração de DNA dos clones atendeu às necessidades requeridas, resultando em DNA sem contaminação ou degradação. A quantificação foi feita por gel de agarose 1% corado com brometo de etídio e em espectrofotômetro (Nanodrop ND100) e mostrou uma concentração de aproximadamente 70 ng/μL para cada clone (figura 4).

Dos 14 clones da biblioteca BAC de *A. duranensis*, quatro formaram dois pares sobrepostos (ADH51I17 e ADH83F22; ADH79O23 e ADH72J06). Estes clones, juntamente com outro clone BAC previamente sequenciado (ADH180A21), compuseram 12 regiões genômicas de *A. duranensis* (tabela 2).

Para cada um dos clones analisados, o sequenciamento gerou um arquivo contendo uma ou mais sequências consenso (*contigs*), que juntas somaram aproximadamente o tamanho do inserto clonado na biblioteca (100-110 kb). Para anotação, os maiores *contigs* foram selecionados, ordenados por tamanho (do maior para o menor) e concatenados para as análises. Um total de 1,26 Mb (milhões de bases) pertencentes ao genoma de *A. duranensis* foi sequenciado e caracterizado.

Os dois clones do genoma B (ADH147A20 de *A. ipäensis* e AHF417E07 do subgenoma B de *A. hypogaea*), que em gráficos de plotagem indicaram ter regiões homeólogas a duas regiões genômicas de *A. duranensis* (ADH68E04 e ADH035P21, respectivamente), também foram analisados.

Tabela 1: Lista de clones BAC selecionados para sequenciamento e anotação.

Genoma de referência	Clone BAC
A. duranensis (Genoma AA)	ADH180A21
	ADH051I17
	ADH083F22
	ADH123K13
	ADH177M04
	ADH179B13
	ADH129F24
	ADH167F07
	ADH079O23
	ADH072J06
	ADH25F09
	ADH068E24
	ADH18B08
	ADH035P21
A. ipënsis (Genoma BB)	ADH147A20
A. hypogaea (Genoma AABB) (Subgenoma BB)	AHF417E07

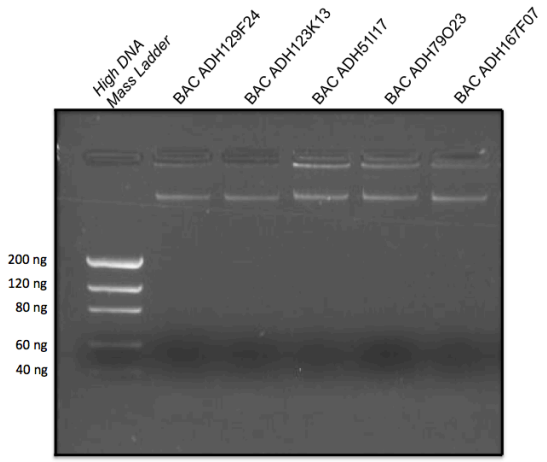


Figura 4: Análises de qualidade e concentração de amostras de DNA dos clones BAC de *A. duranensis* ADH129F24, ADH123K13, ADH51I17, ADH79O23 e ADH167F07 (poços 2-6), realizadas em gel de agarose 1,0% corado com brometo de etídio. Marcador molecular utilizado *High DNA Mass Ladder* – poço 1 (Invitrogen).

Tabela 2: Clones BAC selecionados e sequenciados da biblioteca BAC de *A. duranensis* (genoma A) formando 12 regiões genômicas.

Clone BAC	N° de *contigs*	Tamanho em pb
ADH180A21	1	89.966
ADH051I17-83F22	5	115.680
ADH123K13	2	114.820
ADH177M04	3	90.712
ADH179B13	6	92.455
ADH129F24	6	99.171
ADH167F07	9	99.579
ADH079O23-72J06	11	141.775
ADH25F09	6	99.839
ADH068E24	1	101.960
ADH18B08	3	92.084
ADH035P21	5	125.289
Total		1.263.330

Um dos clones BAC de *A. duranensis*, ADH18B08, foi sequenciado por meio de duas técnicas, a de fragmentação randômica que utilizou a química de Sanger, e pelo método 454 GS FLX Titanium. A primeira técnica gerou sete sequências consenso, ao passo que a segunda, gerou apenas três.

Como apresentado na figura 5, os dois métodos foram consistentes. Apesar dos métodos gerarem fragmentos de sequência de tamanhos diferentes, a montagem foi satisfatória para ambos os casos. No gráfico, as sequências oriundas dos dois métodos foram plotadas uma contra a outra e de acordo com a diagonal resultante, foi possível constatar que havia cinco pequenas regiões invertidas. A montagem selecionada para o estudo foi àquela relativa ao sequenciamento pela técnica 454, para permitir a comparação com os demais clones também sequenciados por essa estratégia.

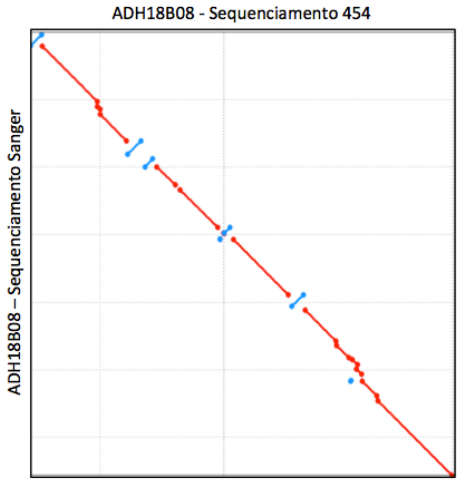


Figura 5: Gráfico de comparação entre sequências oriundas de diferentes tipos de sequenciamento. No eixo x estão representadas três sequencias consenso referentes ao clone BAC ADH18B08 sequenciado pela técnica 454. No eixo y estão representadas sete sequências consenso do mesmo clone, porém oriundas do sequenciamento pela técnica de Sanger. A similaridade entre os dois tipos de sequências é representada pelas diagonais. Foram detectadas cinco pequenas diagonais representando regiões invertidas. Gráfico de plotagem produzido pelo *software* Gepard.

3.2 Anotação de retrotransposons LTR

Para visualizar o conteúdo de sequências repetitivas no genoma A, mais especificamente de retrotransposons LTR presentes nos clones selecionados, gráficos de comparação das sequências dos clones BAC plotadas contra elas mesmas foram obtidos utilizando o *software* Gepard. Esses gráficos iniciais possibilitaram a identificação de perfis representados por vários tipos de sequências repetitivas, incluindo retrotransposons LTR

completos, fragmentos de retrotransposons LTR e LTRs-solo.

A figura 6 apresenta a sequência do clone ADH18B08 plotada contra ela mesma. Foi identificado apenas um retrotransposon LTR completo, ou seja, um retrotransposon contendo dois LTRs flanqueadores e uma região central codificadora de proteína (detalhe ampliado). No entanto, outros clones, tais como o ADH123K13 (figura 7), apresentaram tanto elementos completos, quanto fragmentados. Foram observadas várias linhas diagonais inversas, com tamanhos distintos, pouco conservadas e contendo *gaps* (falhas em sua continuidade), indicando a complexidade do conteúdo repetitivo presente nesta sequência. Com a construção dos gráficos, foram avaliadas a conservação entre as sequências dos elementos, de forma comparativa, assim como a orientação em que estes foram inseridos no genoma.

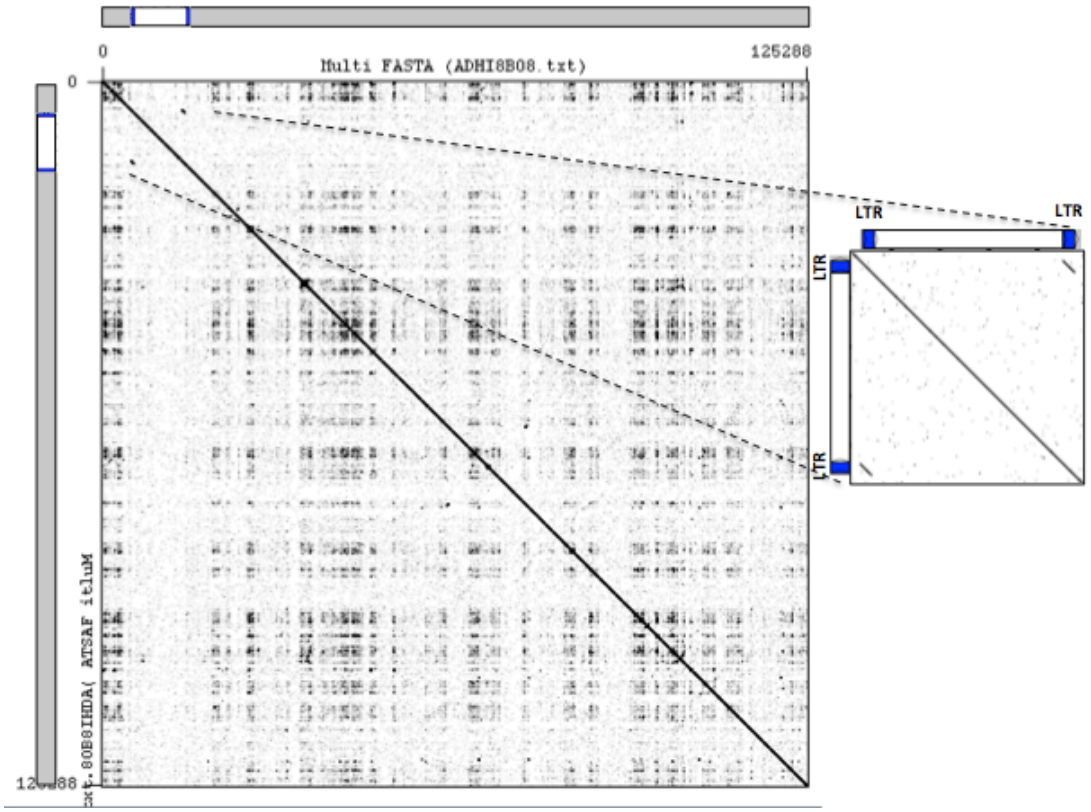


Figura 6: Comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH18B08 (biblioteca de *A. duranensis)* plotada contra ela mesma (direção 5' – 3'). O resultado revelou apenas um retrotransposon LTR completo, presente no início da sequência. O elemento é composto por dois LTRs terminais e uma região central codificadora de proteína. Nota-se que o padrão granuloso é menor onde o retrotransposon LTR está inserido. Gráfico de plotagem produzido pelo *software* Gepard.

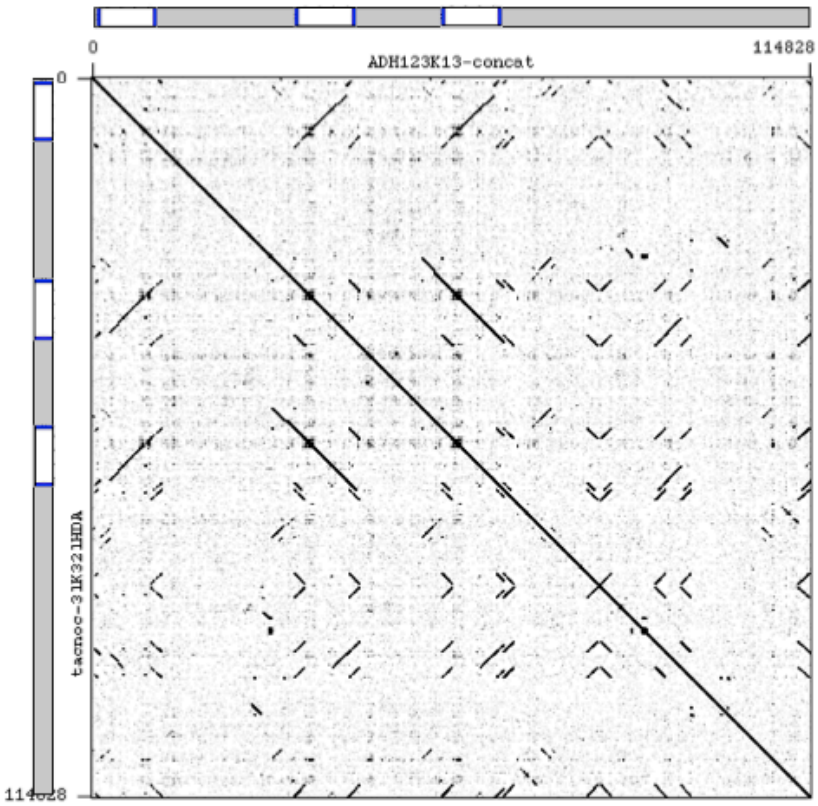


Figura 7: Comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH123K13 (biblioteca de *A. duranensis)* plotada contra ela mesma, ambas na mesma direção. O resultado revelou três retrotransposons LTR completos (um inserido na direção 5' – 3' e dois inseridos na direção 3' – 5') e vários elementos incompletos, fragmentos e LTRs-solo. Uma série de diagonais inversas, com tamanhos distintos, pouco conservadas e contendo *gaps* foram observadas. Gráfico de plotagem produzido pelo *software* Gepard.

A partir do perfil complexo de diagonais presentes nos gráficos de comparação foi criado um esquema para simplificar o entendimento acerca da ocorrência, tipo e distribuição dos retrotransposons LTR e seus respectivos fragmentos nas sequências dos clones BAC. Esse esquema viabilizou uma anotação mais segura e detalhada, baseada no mesmo princípio para todos os clones analisados. A figura 8 exemplifica os diferentes perfis dos elementos de repetição resultantes da comparação de uma sequência contra ela mesma por meio de *dot plot*. Seguindo esse modelo proposto, foi possível anotar de forma inequívoca tanto sequências completas de retrotransposons LTR, quanto diferentes tipos de fragmentos de retrotransposons.

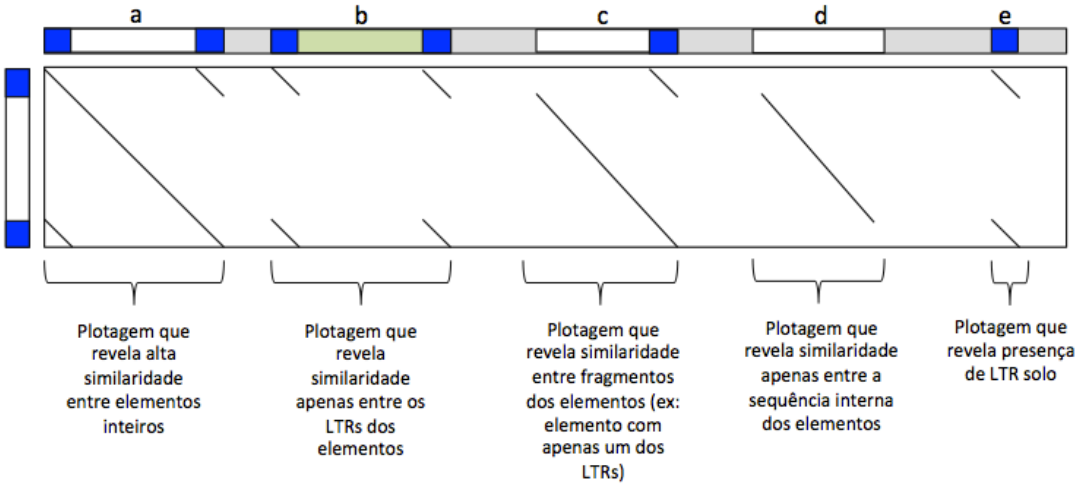


Figura 8: Esquema explicativo dos perfis apresentados pelos retrotransposons LTR e seus fragmentos ao comparar uma sequência contra ela mesma em gráficos de plotagem. No eixo x encontram-se cinco exemplos de elementos ou fragmentos na direção 5' – 3'. No eixo y encontra-se o exemplo de um retrotransposon LTR completo na direção 5' – 3'. Elementos completos e similares plotados um contra o outro (a); Elementos completos, porém apenas com os LTR iguais (em cor azul) plotados um contra o outro (b); Elementos iguais, porém um deles sendo fragmentado, plotados um contra o outro (c); Elementos iguais, porém um sem os LTRs e o outro completo, plotados um contra o outro (d); Um elemento completo plotado contra um LTR-solo igual (e).

Para verificar a presença do retrotransposon LTR FIDEL foram realizadas comparações por meio de gráficos de plotagens dos 14 clones BAC (12 regiões genômicas do genoma A) contra a sequência isolada de FIDEL. Para a maioria dos clones analisados, as plotagens revelaram uma alta frequência de LTRs-solo desse elemento. Conforme revisado por Vitte & Panaud (2005), a recombinação entre retrotransposons LTR presentes em grande número de cópias geram fragmentos contendo apenas as porções de LTRs. Em um exemplo visto na sequência do clone ADH180A21, dois elementos completos de FIDEL foram identificados, além de cinco regiões contendo LTRs-solo deste elemento (figura 9).

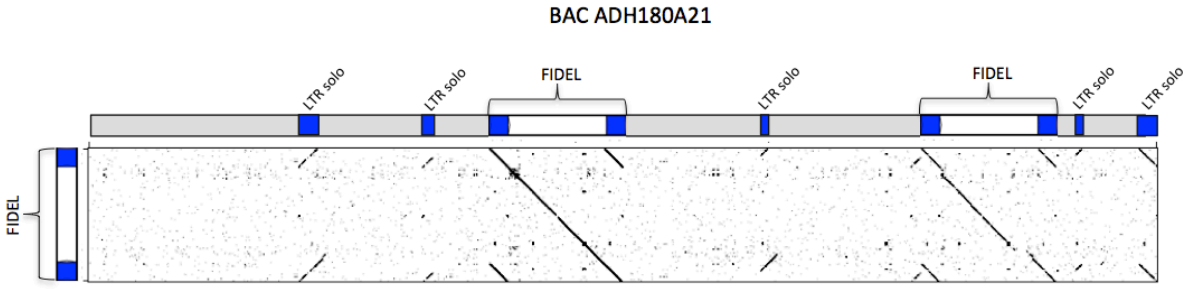


Figura 9: Comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH180A21 (biblioteca de *A. duranensis)* na direção 5' – 3', plotada contra a sequência completa do retrotransposon LTR FIDEL. O resultado revelou dois retrotransposons FIDEL completos, além de cinco LTRs-solo desse elemento. Gráfico de plotagem produzido pelo *software* Gepard.

No entanto, ao realizar um inspeção mais detalhada em gráficos produzidos pela comparação de todas as sequências dos clones contra elas mesmas, individualmente, constatou-se que muitos desses aparentes LTRs-solo de FIDEL possuíam uma distância similar e uma sequência conservada entre eles. Dessa forma, tornou-se evidente a identificação de um novo elemento não-autônomo, com LTRs e parte da região não traduzida (3'-UTR) muito semelhante ao FIDEL. Em contrapartida, na região codificadora de proteína não havia nenhuma semelhança significativa entre a sequência dos dois elementos. Em virtude da semelhança apenas no tamanho das regiões terminais, esse elemento foi denominado de "Feral". Trata-se de um tipo de retrotransposon LTR incompleto, pertencente à superfamília *Ty3-Gypsy*, não-autônomo e provavelmente parasita de FIDEL, pois codifica proteínas com os domínios *gag* e protease, mas não o da transcriptase reversa, necessária para sua transposição. O gráfico de comparação da sequência do clone ADH177M04 com as sequências de FIDEL e Feral exemplifica de forma satisfatória a relação entre esses elementos e, ainda, a semelhança apenas nas regiões terminais. Portanto, em virtude dessa semelhança, não foi possível definir se os LTRs-solo presentes nesse e nos outros clones, seriam exclusivamente de FIDEL ou Feral (figuras 10 e 11).

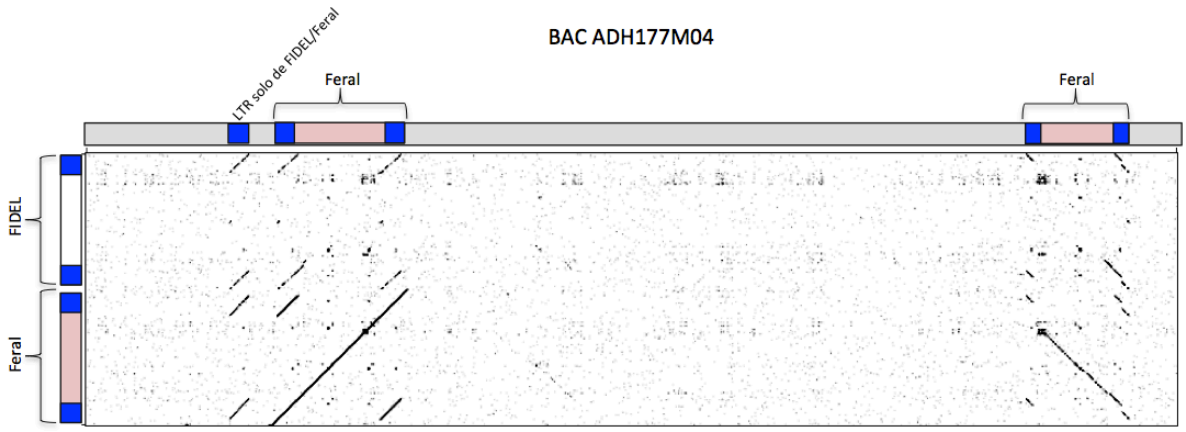


Figura 10: Comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH177M04 na direção 5’ – 3’ plotada contra as sequências dos retrotransposons FIDEL e Feral, ambos na direção 5’ – 3’. O resultado revelou dois elementos Feral completos, sendo que o primeiro estava inserido na direção 3’ – 5’ e o segundo na direção 5’ – 3’, além de um LTR-solo de Feral ou FIDEL na direção 3’ – 5’. Gráfico de plotagem produzido pelo *software* Gepard.

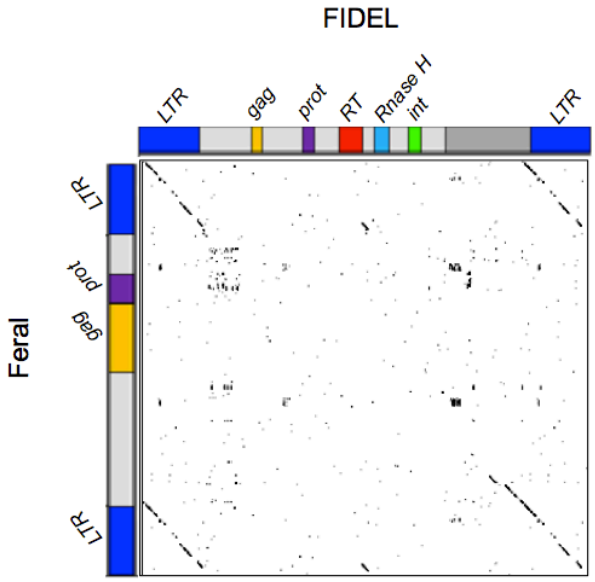


Figura 11: Comparação por meio de gráfico de plotagem utilizando a sequência do retrotransposon LTR FIDEL (eixo x) plotado contra Feral (eixo y), ambos na direção 5' – 3'. As sequências de Feral e FIDEL são semelhantes apenas nas porções dos LTRs (em azul). A figura contém o esquema ilustrativo da estrutura desses dois elementos. Gráfico de plotagem produzido pelo *software* Gepard.

Outras comparações em gráficos de plotagens, juntamente com resultados obtidos pelo *software* LTR_FINDER indicaram a presença de outro retrotransposon LTR também presente em múltiplas cópias em alguns dos clones. O elemento identificado não codifica nenhuma proteína de domínio conservado, mas possui uma ORF (*Open Reading Frame*) perto da extremidade 3' do LTR 5'. No entanto, esta proteína codificada não apresentou homologia com quaisquer proteínas já descritas e disponíveis em bancos de dados de domínio público. Esse elemento não-autônomo foi chamado de "Pipa". O elemento autônomo que seria par de Pipa não foi encontrado nos clones BAC sequenciados neste estudo, porém um clone BAC de *A. hypogaea* sequenciado em outro estudo ainda não publicado (BAC AHF107L23), apresentou dois representantes completos de um retrotransposon autônomo com semelhanças significativas com Pipa. Este elemento autônomo pertencente à superfamília *Ty3-Gypsy* foi denominado de "Pipoka". Pipoka codifica os domínios proteicos *gag*, protease, a transcriptase reversa, RNA*se* H e integrase. Pipa e Pipoka possuem semelhanças nas sequências dos LTRs e na região interna próxima à extremidade 3'. Mas apesar disso, esses dois elementos também possuem diferenças significativas, pois as regiões codificadoras que estão localizadas na metade interna da região 5' são completamente distintas. As figuras 12-A e 12-B mostram esses dois elementos presentes nos clones ADH177M04 e AHF107L23 quando comparados com Pipa e Pipoka por *dot plot*.

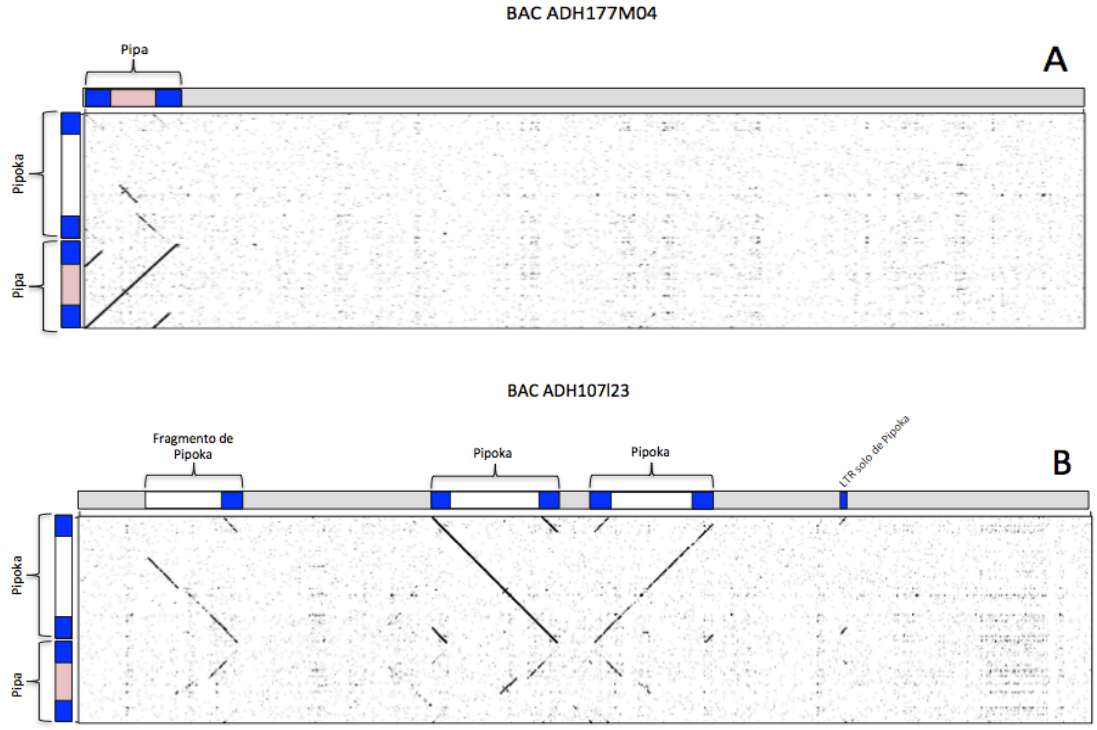


Figura 12: Comparação por meio de gráfico de plotagem utilizando as sequências dos clones ADH177M04 (A) e ADH107L23 (B), ambos na direção 5' – 3' comparadas com as sequuências dos retrotransposons LTR Pipa (3' – 5') e Pipoka (5' – 3'). O resultado mostrou que o clone ADH177M04 possui um elemento Pipa, enquanto o clone AHF107L23 possui dois elementos Pipoka completos e um fragmento. Pipa e Pipoka têm semelhanças nas sequência dos LTRs e na região interna próxima ao 3'. Gráfico de plotagem produzido pelo *software* Gepard.

Outros retrotransposons LTR foram identificados no genoma de *A. duranensis*, tais como o elemento "Gordo", que possui LTRs grandes e bem distintos (2.337 pb), cada um com cerca de sete repetições em *tandem*, contendo motivos com 116 pb de comprimento. A região interna deste retrotransposon codifica os domínios proteicos *gag* e protease, mas não possui a sequência correspondente ao gene que codifica a transcriptase reversa, sendo, portanto, um elemento não-autônomo. Os demais retrotransposons identificados foram nomeados de: "Curu", um retrotransposon *Ty3-Gypsy* com LTRs grandes (3448 pb); "RE128", um retrotransposon *Ty1-Copia*; "Mico" e "Grilo", dois elementos *Ty3-Gypsy*.

Sequências completas e fragmentos dos retrotransposons FIDEL e Matita (Nielen *et al*., 2010; 2012) também foram identificadas nas 12 regiões genômicas analisadas. Uma representação esquemática dos retrotransposons LTR identificados neste estudo é apresentada na figura 13. Além de retrotransposons completos, pseudogenes derivados de elementos de transposição e retrovírus também foram identificados nos clones analisados, mas não foram completamente caracterizados neste estudo.

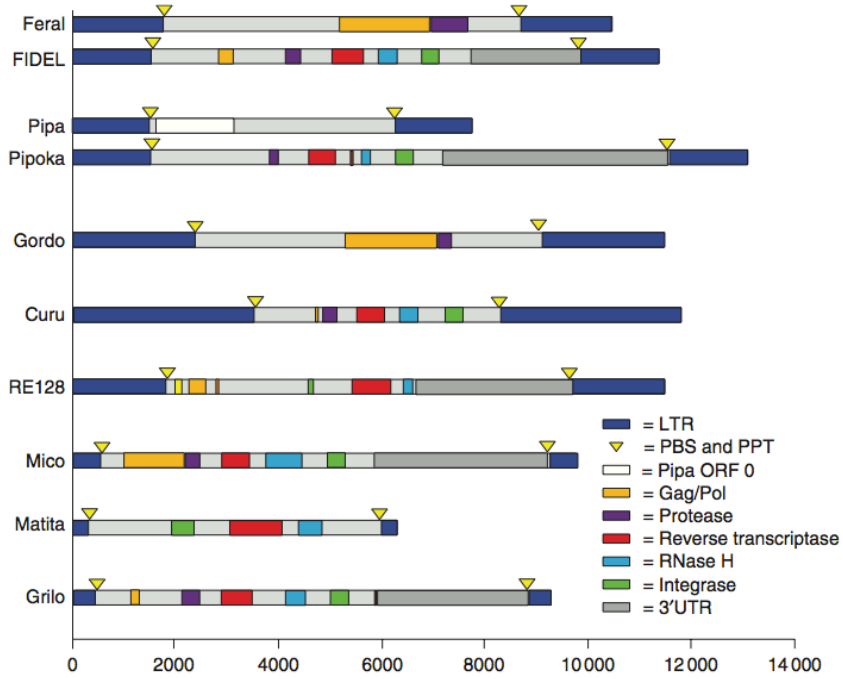


Figura 13: Diagrama esquemático dos retrotransposons LTR identificados no genoma A de *A. duranensis*. Os elementos e seus componentes estão em escala de pares de base (eixo x). Sequências de DNA que codificam domínios de proteínas conservadas estão nas cores indicadas de acordo com a legenda à direita. A ORF 0 em Pipa codifica uma proteína de função desconhecida.

Uma vez identificados e caracterizados esses oito novos retrotransposons LTR, os clones pertencentes ao genoma de *A. duranensis* puderam ser anotados de forma mais precisa e detalhada quanto ao conteúdo total de sequências repetitivas. Portanto, os gráficos de plotagens obtidos pela comparação das sequências dos clones BAC contra elas mesmas e contra sequências dos retrotransposons descritos em estudos anteriores, foi considerada uma metodologia eficiente para identificação de novos retrotransposons LTR e seus seus respectivos fragmentos.

Alguns dos elementos identificados nos gráficos de plotagem, também foram identificados pelo *software* LTR_FINDER e estruturas como TSR, PBS e PPT foram descritas quando presentes. A junção das análises desenvolvidas por esses dois *softwares* possibilitou a caracterização das estruturas repetitivas de forma detalhada, assim como a orientação de inserção dos elementos no genoma. Porém, apenas as inspeções manuais possibilitaram a identificação de fragmentos de retrotransposons e LTR-solo, além da observação da conservação entre as sequências dos elementos repetitivos.

A figura 14 mostra um exemplo de uma comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH180A21 (genoma A) (eixo x) com todas as sequências dos elementos descritos nesse estudo (eixo y). Foram detectados na sequência desse clone

BAC: dois elementos FIDEL completos (inseridos na direção 5’ – 3’), um elemento Pipa completo (direção 3’ – 5’), um fragmento do elemento Gordo (direção 5’ – 3’) e outro fragmento do elemento Pipoka (direção 3’ – 5’). A ocorrência de um elemento inserido dentro de outro (*nested transposon*) também pôde ser observada na anotação deste clone (também comum em outros clones). O primeiro elemento FIDEL estava inserido dentro de um elemento Gordo (direção 5’ – 3’). O *software* LTR_FINDER identificou apenas as coordenadas referentes aos elementos FIDEL, o que demonstra a eficiência do método de plotagem para identificação e anotação de elementos repetitivos e retrotransposons LTR.

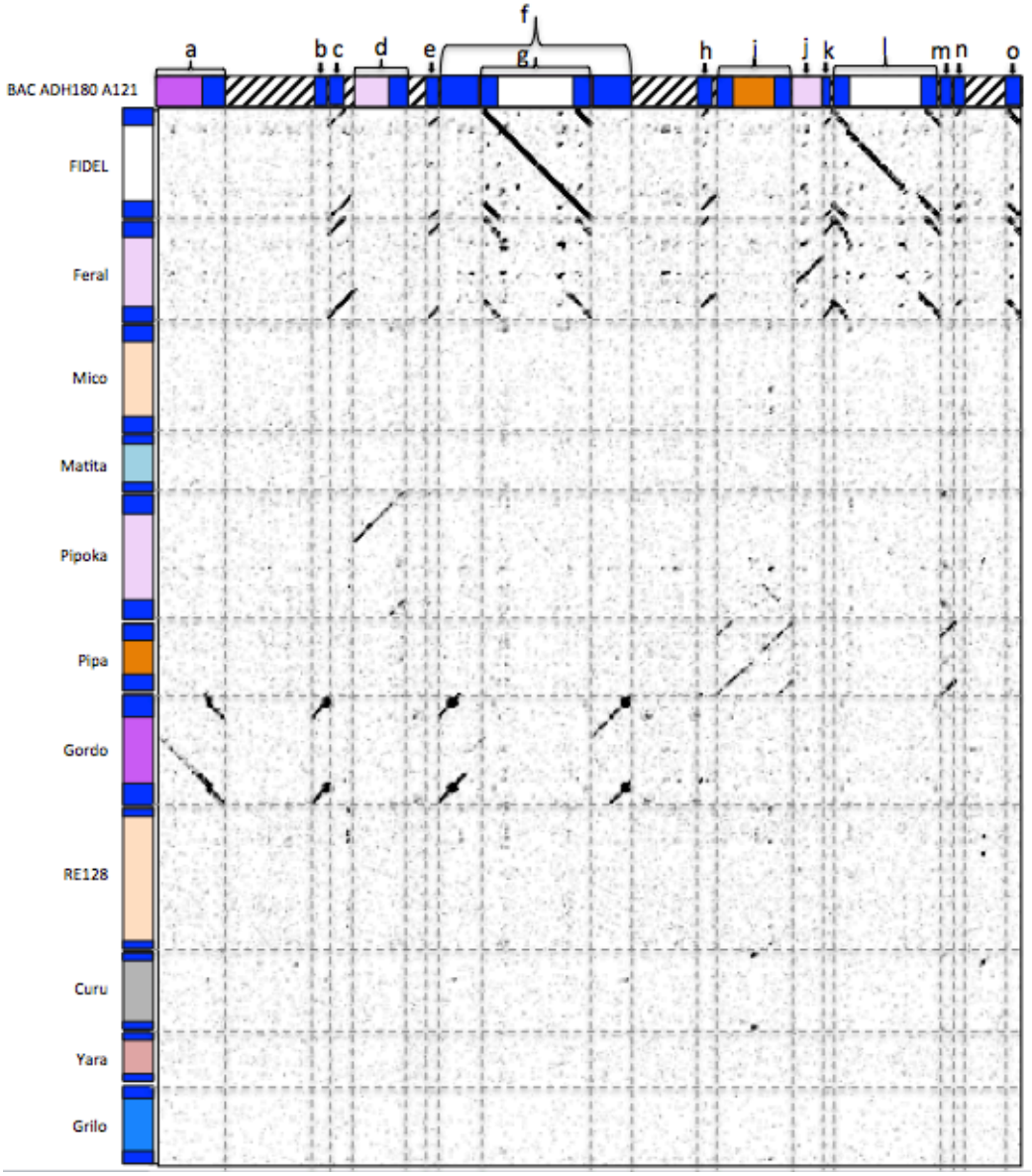


Figura 14: Comparação por meio de gráfico de plotagem utilizando a sequência do clone ADH180A21 na direção 5’ – 3’ (eixo x) com todos os retrotransposons LTR identificados neste e em outros estudos na direção 5’ – 3’ (eixo y). Cada elemento está apresentado em uma cor diferente e as porções LTR dos elementos estão em cor azul. A porção hachurada representa a sequência do clone BAC que não possui elementos de repetição. Gráfico de plotagem produzido pelo *software* Gepard.

3.3 Anotação de genes

O *software* FGENESH identificou muitos genes putativos situados dentro e sobrepostos aos retrotransposons LTR, bem como em seus fragmentos truncados. Estes genes preditos, na maioria das vezes, codificavam domínios proteicos com funções relacionadas a elementos de transposição, conforme observado em informações depositadas no banco de dados do *PfamA*. Porém, em alguns casos, as proteínas preditas não apresentavam nenhuma função já descrita. Foram preditas estruturas inadequadas compostas de exons/introns para estes genes, os quais são, possivelmente, pseudogenes derivados dos genes que codificam poliproteínas ou ainda artefatos produzidos pelos algoritmos utilizados pelo *software* FGENESH.

Para a anotação dos genes presentes nas sequências completas de retrotransposons, foi utilizada a nomenclatura "pseudogenes" para todos os domínios proteicos característicos de retrotransposons (FIDEL, Feral, Pipoka, Pipa, etc.), mas não para seus fragmentos. Pseudogenes derivados de elementos transponíveis não caracterizados também foram anotados. Comparações com ESTs do gênero *Arachis* indicaram que possivelmente há atividade de transcrição para esses retrotransposons LTR.

Os genes preditos pelo *software* FGENESH para outras regiões não repetitivas e que não estavam relacionadas a elementos de transposição foram corroborados por algumas evidências secundárias. A confiabilidade em relação a alguns genes foi evidenciada pela presença de ESTs de *Arachis*, domínios proteicos depositados no *pfamA* e similaridade com genes preditos em Arabidopsis e soja. Portanto, funções foram atribuídas a esses genes, para que pudessem ser anotados como proteínas putativas.

Alguns dos clones BAC apresentaram conteúdo essencialmente repetitivo, como ADH180A21, ADH167F07, ADH179B13 e ADH18B08. Outros apresentaram conteúdos repetitivo e gênico, tais como ADH35P21, ADH123K13, ADH25F09, ADH129F24, AHF417E07 e AIPA147A20. Os demais clones contendo conteúdo gênico, ADH177M04, ADH79O23-72J06, ADH68E04 e ADH51I17-83F22, escolhidos para serem sequenciados, apresentavam marcadores genéticos desenvolvidos em outras leguminosas (Choi *et al*., 2006; Fredslund *et al*., 2006). O conteúdo gênico predito para cada clone encontra-se na tabela 3.

Tabela 3: Conteúdo gênico e repetitivo das doze regiões genômicas de *A. duranensis* sequenciadas juntamente com os clones do genoma B, AHF417E07 e AIPA147A20.

BAC	Predição da função	Intervalo de bases
ADH035P21	*putative MULE-type transposon*	198 a 4889
	hypothetical protein	7680 a 10034
	retrotransposon:Matita	16175 a 21598
	hypothetical protein	24828 a 26163
	hypothetical protein	30797 a 31153
	putative transmembrane BT1 family protein	41147 a 44316
	hypothetical protein	45991 a 46728
	putative lipid transfer / seed storage / trypsin-alpha amylase inhibitor	66281 a 67099
	putative lipid transfer / seed storage / trypsin-alpha amylase inhibitor	69172 a 70231
	proteasome subunit	70554 a 71250
	hypothetical protein	71798 a 72171
	seed storage protein Ara h1	72469 a 74531
	hypothetical protein	79069 a 79512
	hypothetical protein	83142 a 83495
	hypothetical protein	84885 a 85097
ADH167F07	*remnant retrotransposon sequence with similarity to Curu*	893 a 1994
	remnant retrotransposon sequence with similarity to Curu	3704 a 4847
	remnant retrotransposon sequence with similarity to Pipa	15483 a 19231
	retrotransposon:Pipa	24164 a 31681
	remnant retrotransposon sequence with similarity to Curu	32888 a 34000
	fragment of retrotransposon Pipa; includes LTR	33994 a 37929
	remnant of retrotransposon Pipa	38245 a 39659
	fragment of retrotransposon Feral; includes LTR	45133 a 53551
	transposon protein (nonfunctional)	53577 a 55643
	transposon protein (nonfunctional)	61271 a 65086
	remnant retrotransposon Pipa	65634 a 68309
	remnant or solo LTR of retrotransposon Feral or FIDEL	87909 a 89458
	hypothetical protein	89558 a 92447
ADH179B13	*hypothetical protein*	8809 a 1067
	remnant or solo LTR of retrotransposon Feral or FIDEL	13310 a 15057
	hypothetical protein	16267 a 17459
	transposon protein (nonfunctional)	24583 a 26213
	retrotransposon:Gordo	26328 a 37595
	transposon protein (nonfunctional)	37603 a 39866
	remnant or solo LTR Feral/FIDEL	41736 a 42491
	fragment of retrotransposon Pipa	42822 a 49210
	hyopthetical protein	49499 a 50322
	remnant or solo LTR Gordo	66339 a 68526
	fragment of retrotransposon Feral; includes single LTR	69263 a 77060
	fragment of retrotransposon Pipa	84046 a 84902
	fragment of retrotransposon Curu	85337 a 90087
ADH123K13	*fragment retrotransposon Pipa*	35 a 246
	fragment of retrotransposon Feral	247 a 1285
	remnant or solo LTR retrotransposon Pipa	1286 a 2712
	fragment of retrotransposon Feral	2713 a 10973
	predicted WD40 domain containing protein	2713 a 10973
	fragment of retrotransposon Gordo	27177 a 27558
	fragment retrotransposon Pipa includes LTR region	28635 a 31973
	retrotransposon:Feral	32222 a 42515
	fragment of retrotransposon Pipa includes LTR	52545 a 55650
	fragment of retrotransposon Feral includes LTRs	55798 a 67151
	fragment of retrotransposon Pipa	68343 a 73157
	remnant or solo LTR retrotransposon Feral or FIDEL	78940 a 80524
	remnant or solo LTR retrotransposon Feral or FIDEL	81168 a 82704
	retrotransposon:Feral	89747 a 95494
	fragment of retrotransposon Pipa	106970 a 112555
	remnant or solo LTR retrotransposon Feral or FIDEL	112619 a 114661

ADH79O23-ADH72J06	*hypothetical protein*	69 a 1628
	protein kinase	3573 a 7189
	EF hand calcium-binding protein	7602 a 8739
	pfkB family kinase	8926 a 13158
	hypothetical protein	18831 a 19702
	hypothetical protein	24692 a 26198
	putative endonuclease/exonuclease/phosphatase	26958 a 28566
	remnant or solo LTR retrotransposon Pipa	28767 a 30301
	fragment of retrotransposon Pipoka	30269 a 34479
	retrotransposon protein (nonfunctional)	34429 a 35094
	remnant or solo LTR retrotransposon Feral/FIDEL	36225 a 37871
	transposon protein (nonfunctional)	39130 a 39798
	transposon protein (nonfunctional)	40680 a 40928
	transposon protein (nonfunctional)	43642 a 45564
	hypothetical protein	52229 a 55360
	glycoside hydrolase	55932 a 60886
	hypothetical protein	62694 a 64134
	hypothetical protein	68068 a 69323
	hypothetical protein	70806 a 73516
	remnant or solo LTR retrotransposon Pipa	75344 a 76852
	hypothetical protein	83463 a 85732
	fragment of retrotransposon Gordo; includes LTR	95011 a 103766
	fragment of retrotransposon sequence with similarity to FIDEL	110799 a 114162
	fragment of retrotransposon Pipoka; includes LTR	118279 a 125335
	fragment of retrotransposon Feral	125436 a 132282
	caulimovirus-like protein (non-functional)	132383 a 136207
	putative cytidylyltransferase	139233 a 142152
BAC25F09	*NB-ARC LRR protein*	1287 a 6940
	TIR NB-ARC LRR protein	7514 a 13806
	remnant or solo LTR retrotransposon RE128	14246 a 16086
	transposon protein (nonfunctional)	23775 a 25924
	fragment retrotransposon Pipa	31536 a 34609
	fragment retrotransposon Pipa	35130 a 39326
	retrotransposon:Matita	41633 a 47818
	TIR NB-ARC LRR protein	48503 a 54372
	TIR NB-ARC LRR protein	57845 a 62276
	TIR NB-ARC LRR protein	62453 a 67698
	putative protein with similarity to DUF313	79800 a 81759
	fragment retrotransposon RE128	85030 a 87229
	fragment retrotransposon RE128	89031 a 94482
	remnant or solo LTR retrotransposon Pipa	97241 a 100317
ADH68E04	*putative glycosyl phosphatidyl inositol transamidase like protein*	1305 a 13352
	putative oligosaccharide biosynthesis protein	13851 a 17614
	putative mitochondrial carrier protein	17693 a 22645
	putative fatty acid elongase	31097 a 34178
	putative fatty acid elongase	36856 a 41585
	putative gyrase protein	41713 a 58077
	putative HpcH/HpaI aldolase/citrate lyase	59121 a 60905
	putative single-stranded DNA binding protein	60941 a 65033
	putative serine carboxypeptidase	78278 a 85944
	retrotransposon:RE128	87057 a 98305
ADH129F24	*remnant or solo LTR of retrotransposon Feral or FIDEL*	1 a 637
	fragment of retrotransposon sequence with similarity to Feral	630 a 4292
	remnant or solo LTR of retrotransposon Feral or FIDEL	4315 a 6030
	remnant or solo LTR of retrotransposon Feral or FIDEL	7627 a 9198
	putative FAD binding protein	14483 a 15241
	glycerol-3-phosphate dehydrogenase	16524 a 23696
	fragment of retrotransposon sequence with similarity to Grilo	24744 a 26153
	fragment of retrotransposon Feral or FIDEL	26276 a 26653
	hypothetical protein	32274 a 33648
	putative transposon endonuclease/exonuclease/phosphatase (nonfunctional)	41142 a 44155
	fragment of retrotransposon sequence with similarity to Feral	49992 a 51867
	retrotransposon:Feral	53589 a 61000
	remnant or solo LTR of retrotransposon Feral or FIDEL	66270 a 68008
	fragment of retrotransposon Feral or FIDEL	69607 a 69978
	fragment of Retrotransposon Grilo; including single LTR sequence	81149 a 88473
	fragment of retrotransposon Feral including partial LTR sequences	88774 a 96586

	retrotransposon: Pipa	38 a 7649
	remnant or solo LTR retrotransposon Feral/FIDEL	11915 a 13488
	retrotransposon: Feral	15843 a 26146
	remnant retrotransposon sequence with similarity to Curu	30343 a 36411
	transposon protein	41535 a 47740
	remnant or solo LTR retrotransposon Pipa	50266 a 51657
ADH177M04	*hypothetical protein*	53874 a 59713
	retrotransposon: Mico	61256 a 70869
	remnant retrotransposon Gordo	76738 a 77338
	remnant or solo LTR of retrotransposon Feral or FIDEL	78123 a 78687
	remnant of retrotransposon Feral	79197 a 86033
	remnant or solo LTR retrotransposon Feral	86033 a 86645
	remnant or solo LTR retrotransposon Gordo	87034 a 87766
ADH18B08	*retrotransposon: Mico*	5002 a 14575
	remnant retrotransposon Feral	381 a 919
	remnant or solo LTR retrotransposon Feral or FIDEL	1021 a 2602
	remnant or solo LTR retrotransposon Feral or FIDEL	2657 a 5309
	hypothetical protein	17467 a 19336
	hypothetical protein	21554 a 22938
	retrotransposon Feral outer of two nested	24062 a 44643
	retrotransposon Feral inner of two nested	30901 a 41478
	remnant or solo LTR retrotransposon Feral or FIDEL	48294 a 49974
	remnant retrotransposon sequence with similarity to FIDEL	51390 a 54505
	remnant or solo LTR retrotransposon Pipa	56646 a 58578
ADH51I17-ADH83F22	*remnant or solo LTR retrotransposon Feral/FIDEL*	59178 a 60775
	retrotransposon: FIDEL	61598 a 75657
	remnant retrotransposon Pipa	75800 a 80140
	putative LSD1 zinc finger domain protein	83510 a 85267
	hypothetical protein	85915 a 86508
	remnant or solo LTR retrotransposon Feral or FIDEL	99047 a 100088
	remnant or solo LTR retrotransposon Feral or FIDEL	100567 a 102085
	remnant or solo LTR retrotransposon Feral or FIDEL	103859 a 105368
	remnant retrotransposon Pipa	105410 a 106496
	remnant retrotransposon Pipa	106497 a 109497
	remnant retrotransposon Feral	111352 a 115686
	hypothetical protein	3920 a 4354
	putative protein, contains domain DUF623	11382 a 11987
	seed maturation protein	13975 a 14836
	hypothetical protein	17128 a 17622
	hypothetical protein	19787 a 19957
	retrotransposon Mico	22327 a 32027
	hypothetical protein	33495 a 33979
	putative protein, contains domain DUF1677	42715 a 43071
	putative transmembrane BT1 family protein	56649 a 59786
AHF417E07	*putative protein contains domain DUF581*	61852 a 62605
	retrotransposon Joka	64671 a 69845
	putative lipid transfer / seed storage / trypsin-alpha amylase inhibitor	83619 a 84681
	transposon AhMITE1-2	86331 a 86536
	putative lipid transfer / seed storage / trypsin-alpha amylase inhibitor	86939 a 87949
	proteasome subunit	88348 a 89055
	hypothetical protein	89598 a 89988
	seed storage protein Ara h1	90317 a 92591
	fragment retrotransposon YARA	94378 a 95903
	hypothetical protein	97153 a 97416
	fragment or remnant of retrotransposon Gordo	1 a 6497
	remnant or solo LTR retrotransposon Gordo	15903 a 17580
	remnant or solo LTR retrotransposon Feral/FIDEL	17638 a 19210
	fragment of retrotransposon Pipoka	20285 a 25447
	remnant or solo LTR retrotransposon Feral/FIDEL	28143 a 29016
	retrotransposon Gordo nested outside retrotransposon FIDEL	29041 a 49021
	retrotransposon FIDEL nested inside retrotransposon Gordo	33765 a 44964
ADH180A21	*putative LTR retrotransposon*	49023 a 56263
	remnant or solo LTR retrotransposon FIDEL/Feral	56562 a 57960
	retrotransposon Pipa	58074 a 66069
	remnant retrotransposon sequence with similarity to Feral	66442 a 69175
	remnant or solo LTR retrotransposon Feral/FIDEL	69400 a 70092
	retrotransposon FIDEL	70129 a 81451
	remnant or solo LTR retrotransposon Pipa	81590 a 82994
	remnant or solo LTR retrotransposon FIDEL/Feral	83175 a 83601
	remnant or solo LTR retrotransposon FIDEL/Feral	88461 a 88964

AIPA147A20	*hypothetical protein*	1100 a 8729
	hypothetical protein	13167 a 27650
	hypothetical protein	28840 a 29241
	hypothetical protein	30981 a 31847
	hypothetical protein	32481 a 38861
	hypothetical protein	43510 a 45691
	hypothetical protein	46889 a 53057
	hypothetical protein	55708 a 60194
	hypothetical protein	61280 a 63442
	hypothetical protein	64631 a 66038
	putative glycosyl phosphatidyl inositol transamidase like protein	67519 a 78917
	putative oligosaccharide biosynthesis protein	80277 a 83088
	putative mitochondrial carrier protein	83831 a 88428
	hypothetical protein	94686 a 95244
	putative fatty acid elongase	102258 a 102884
	putative fatty acid elongase	106341 a 107413
	putative C2 domain contatining protein	109408 a 111505
	DNA gyrase; 3′ gene fragment	112323 a 113880

Para as anotações completas das sequências dos clones BAC exclusivos de *A. duranensis*, foram criados arquivos em formato ".txt" para que se pudesse visualizar os conteúdos gênico e repetitivo por meio de gráficos. As anotações das regiões repetitivas, de retrotransposons, juntamente com dados do "Índice repetitivo" indicaram que os retrotransposons identificados nesse estudo explicam quase todo o conteúdo de DNA altamente repetitivo presente nas 12 regiões genômicas analisadas. Essa forma gráfica para visualização foi bastante útil, pois também possibilitou a inserção dessas sequências no banco de dados. A referência em banco de dados do NCBI (Genbank) do clone BAC sequenciado anteriormente (ADH180A21) é GU480450.1. As anotações completas dos clones BAC estão disponíveis no banco de dados ENA com os seguintes números de acesso: HF937564-HF937576.

Para facilitar a visualização das anotações das sequências dos clones, foram compiladas várias análises, tais como, aquelas realizadas nos *softwares* Artemis e Gepard (clones *vs.* 10 retrotransposons LTR), juntamente com o gráfico que representa o "índice repetitivo" (figuras 15 a 25). Todos os resultados obtidos sob diferentes perspectivas, se complementaram, chegando a um resultado comum. Nos gráficos a seguir os retrotransposons com sequência completa são mostrados em branco com LTRs em azul e os fragmentos de retrotransposons estão em cor verde.

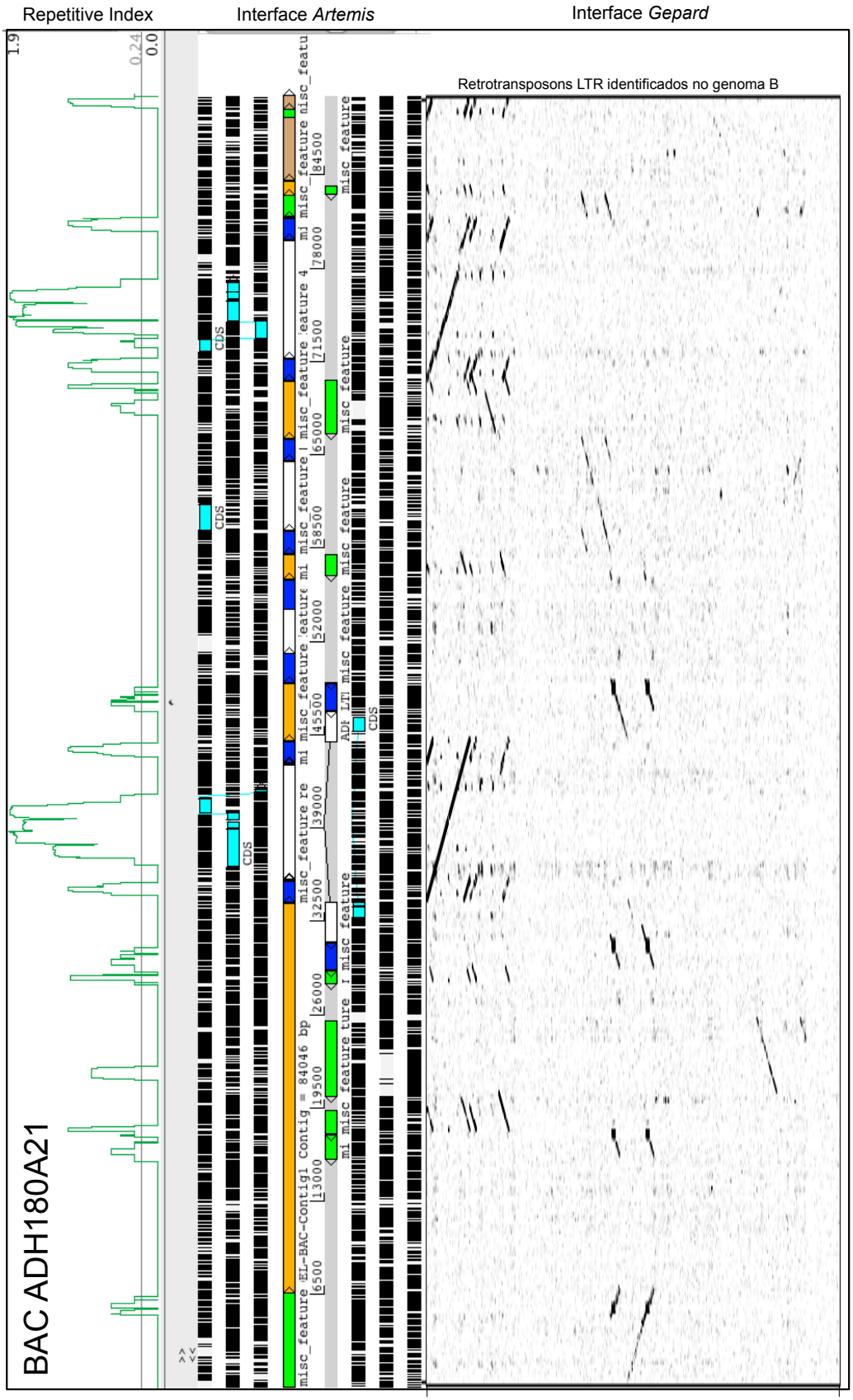


Figura 15: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH180A21, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

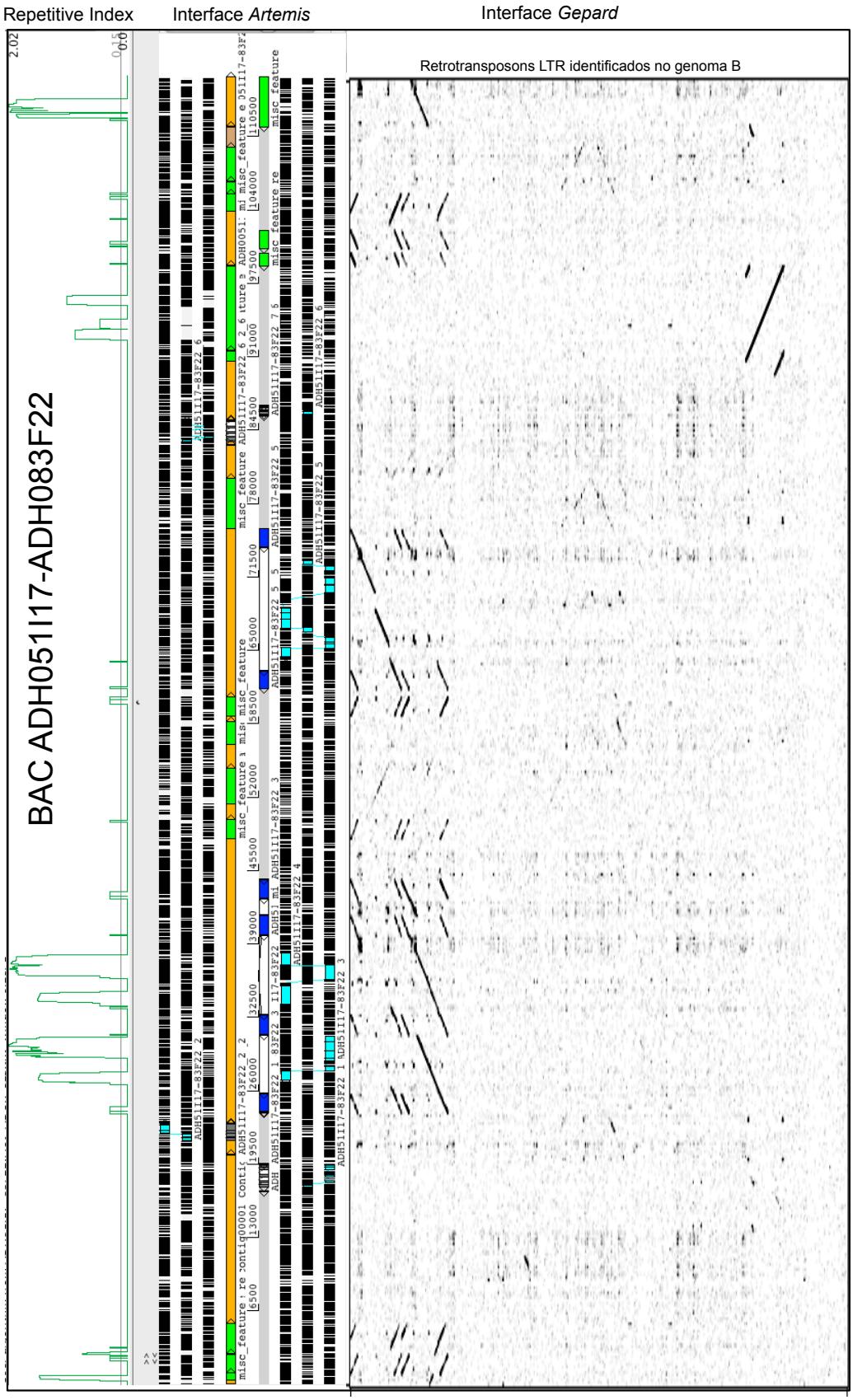


Figura 16: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH51I17-ADH83F22, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

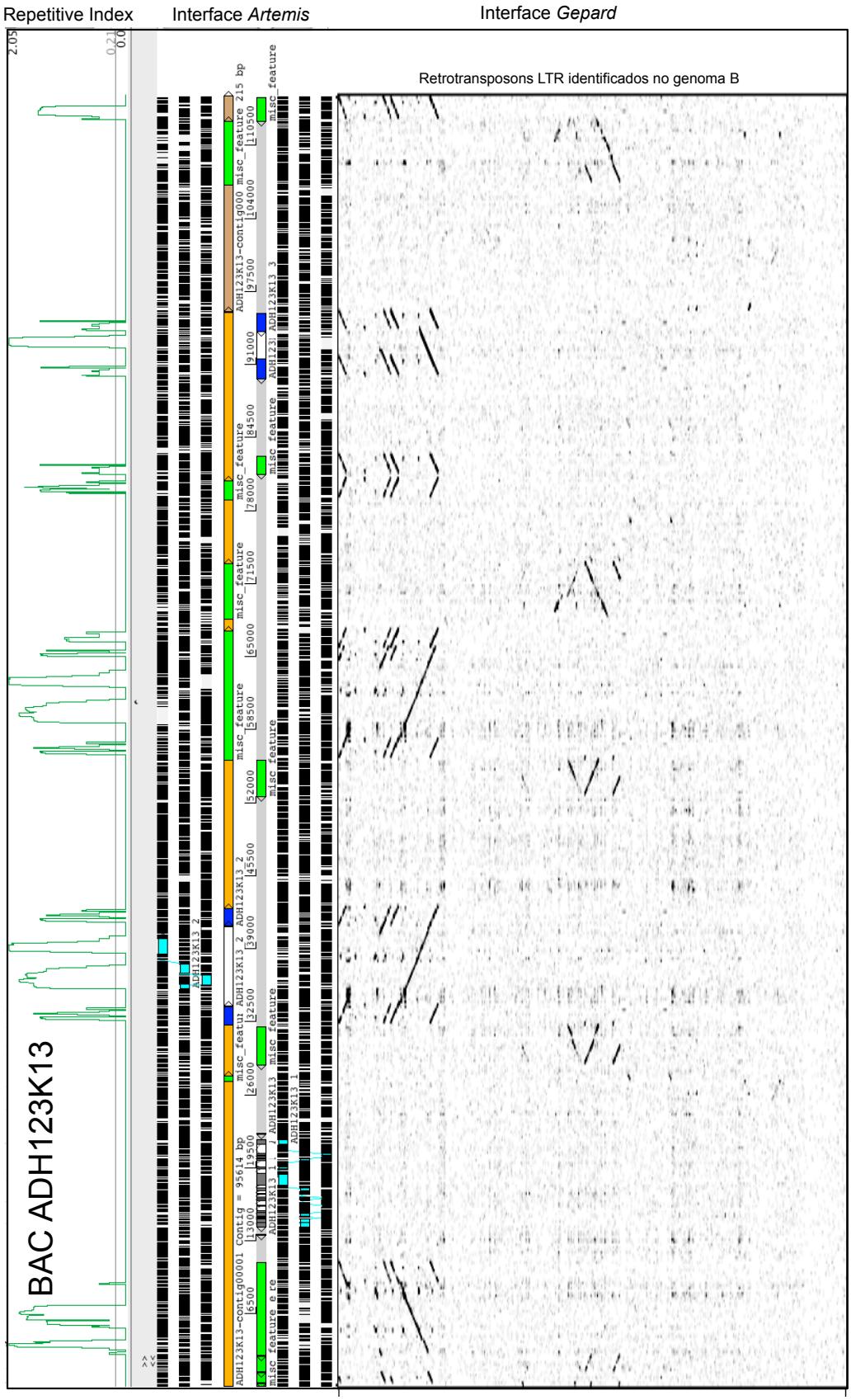


Figura 17: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH123K13, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

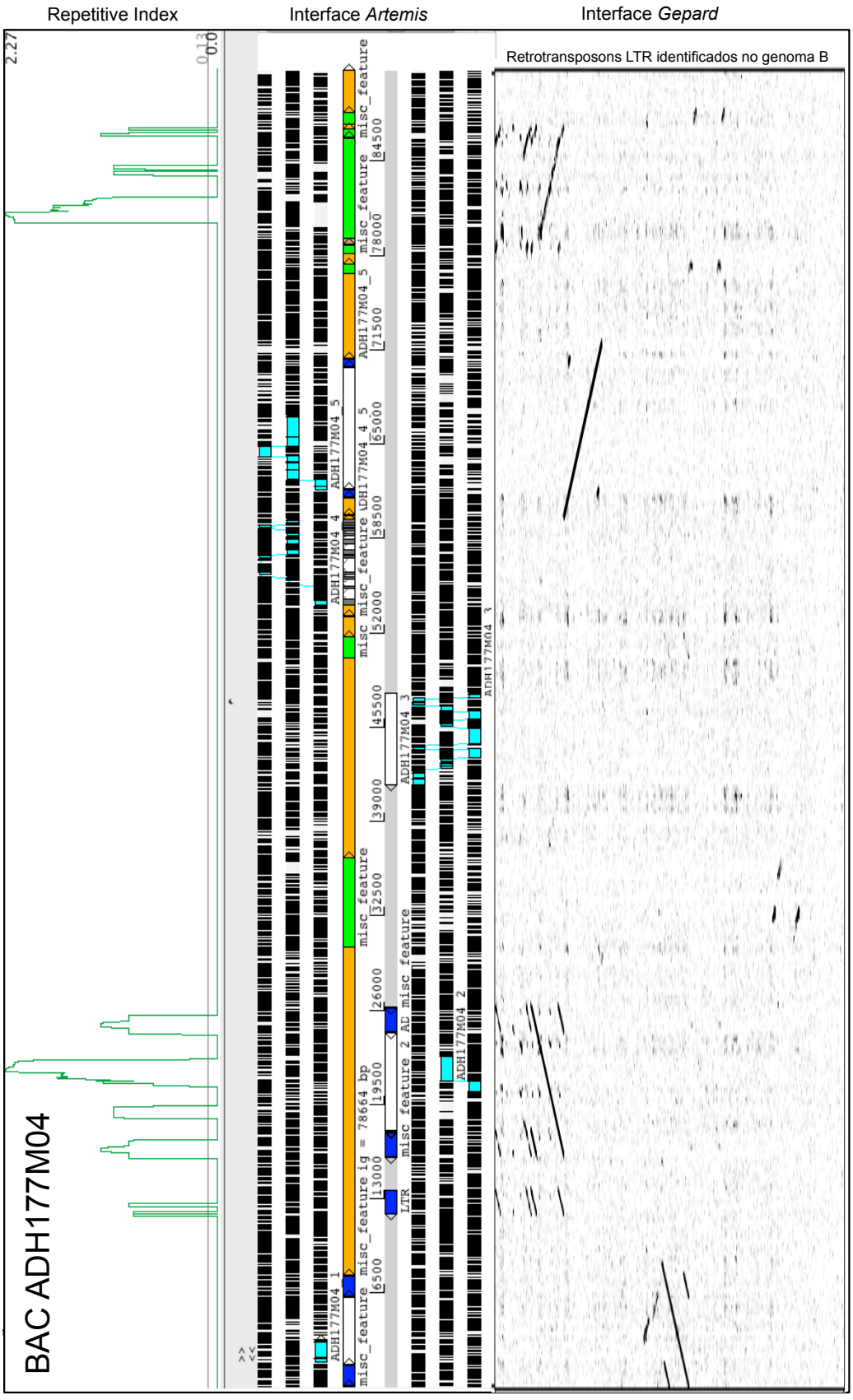


Figura 18: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH177M04, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

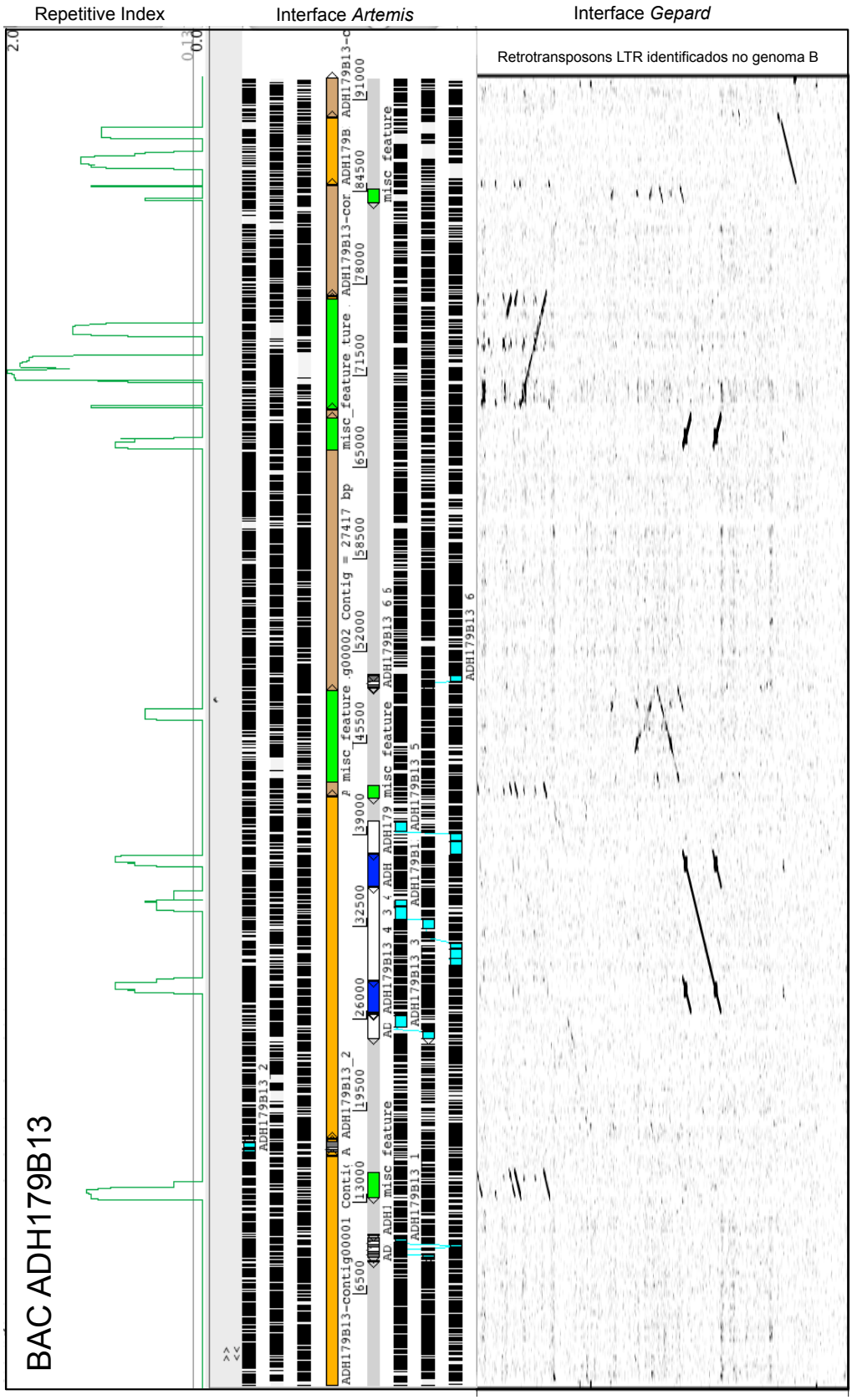


Figura 19: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH179B13, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

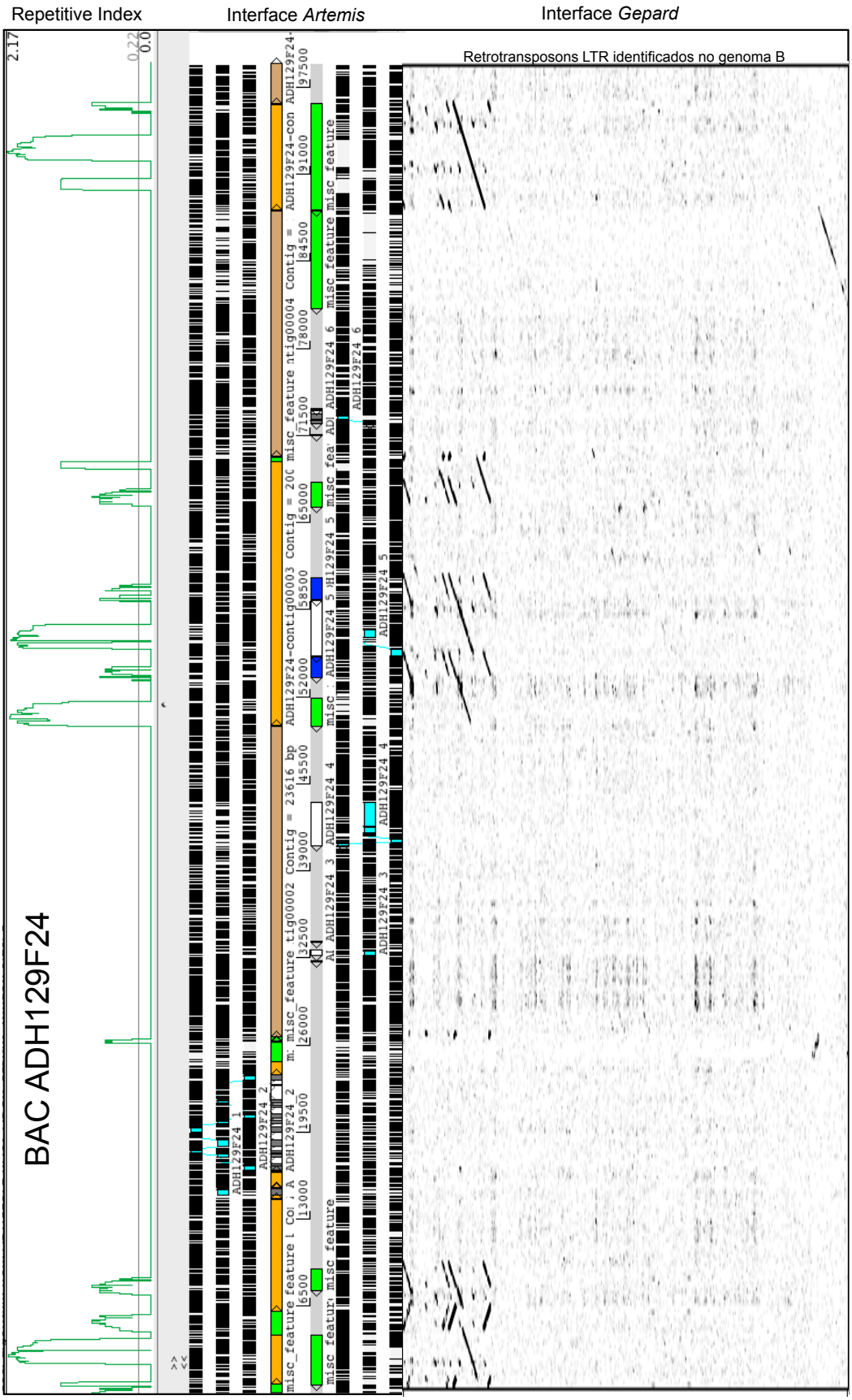


Figura 20: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH129F24, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

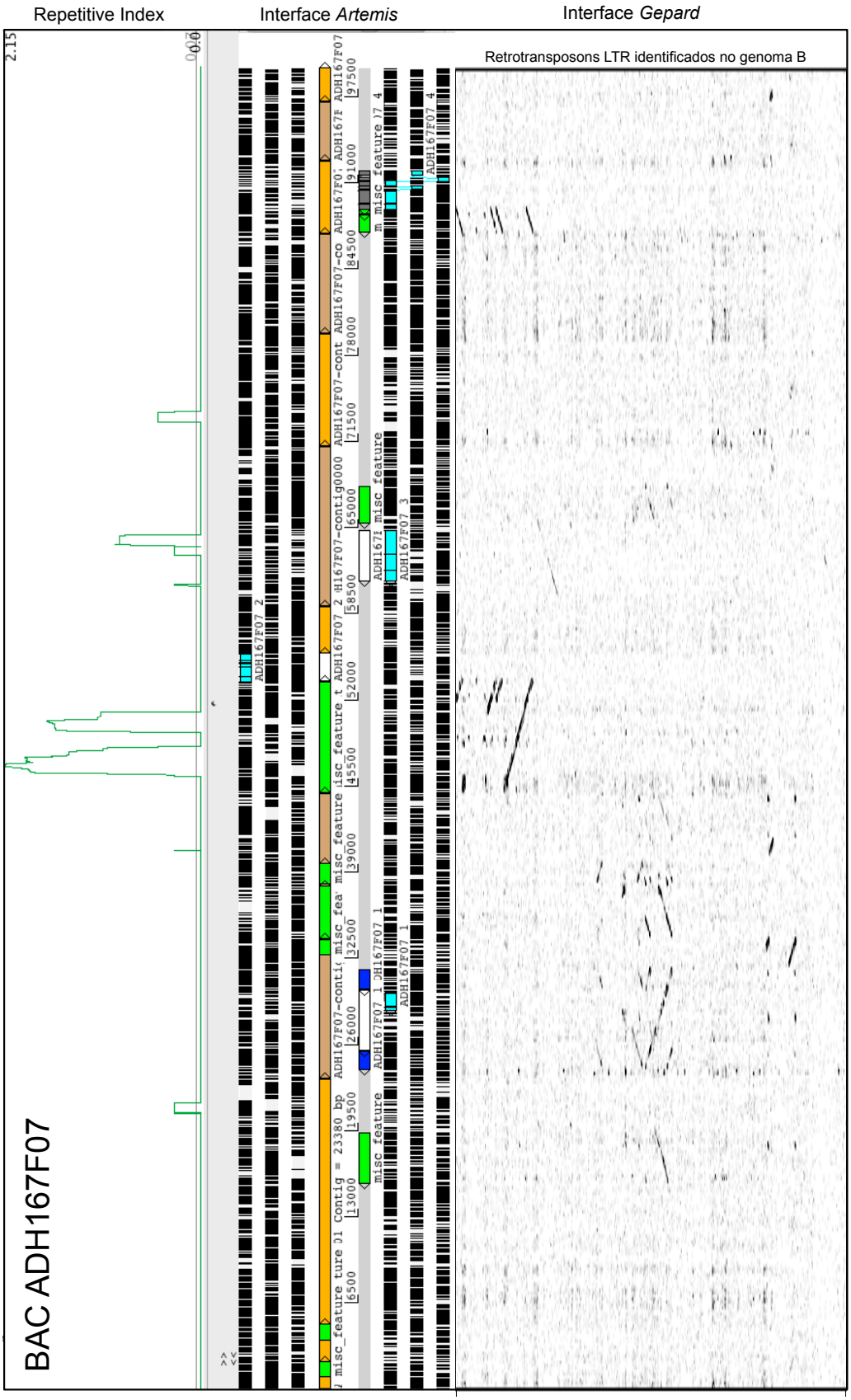


Figura 21: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH167F07, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

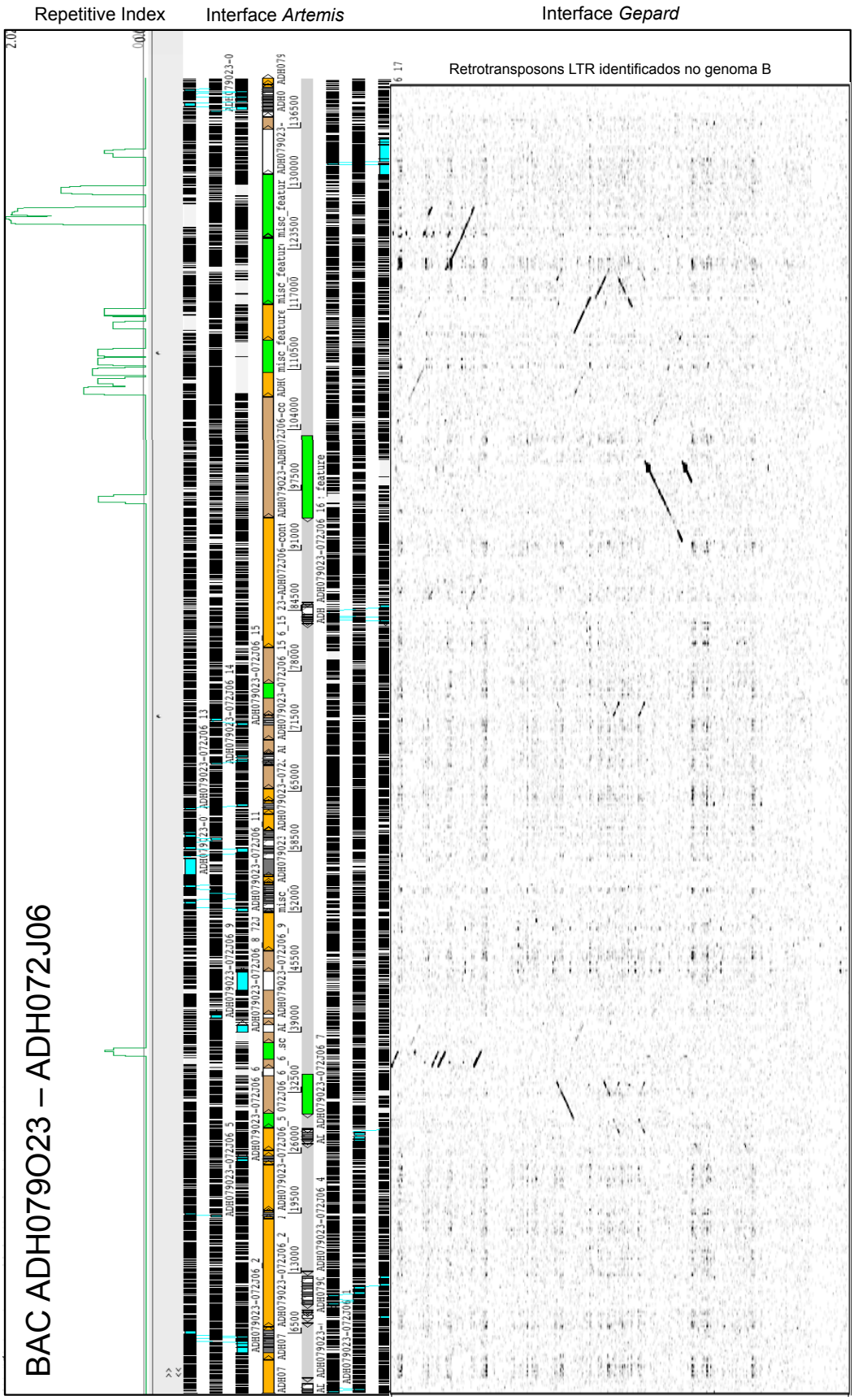


Figura 22: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH079O23-ADH072J06, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

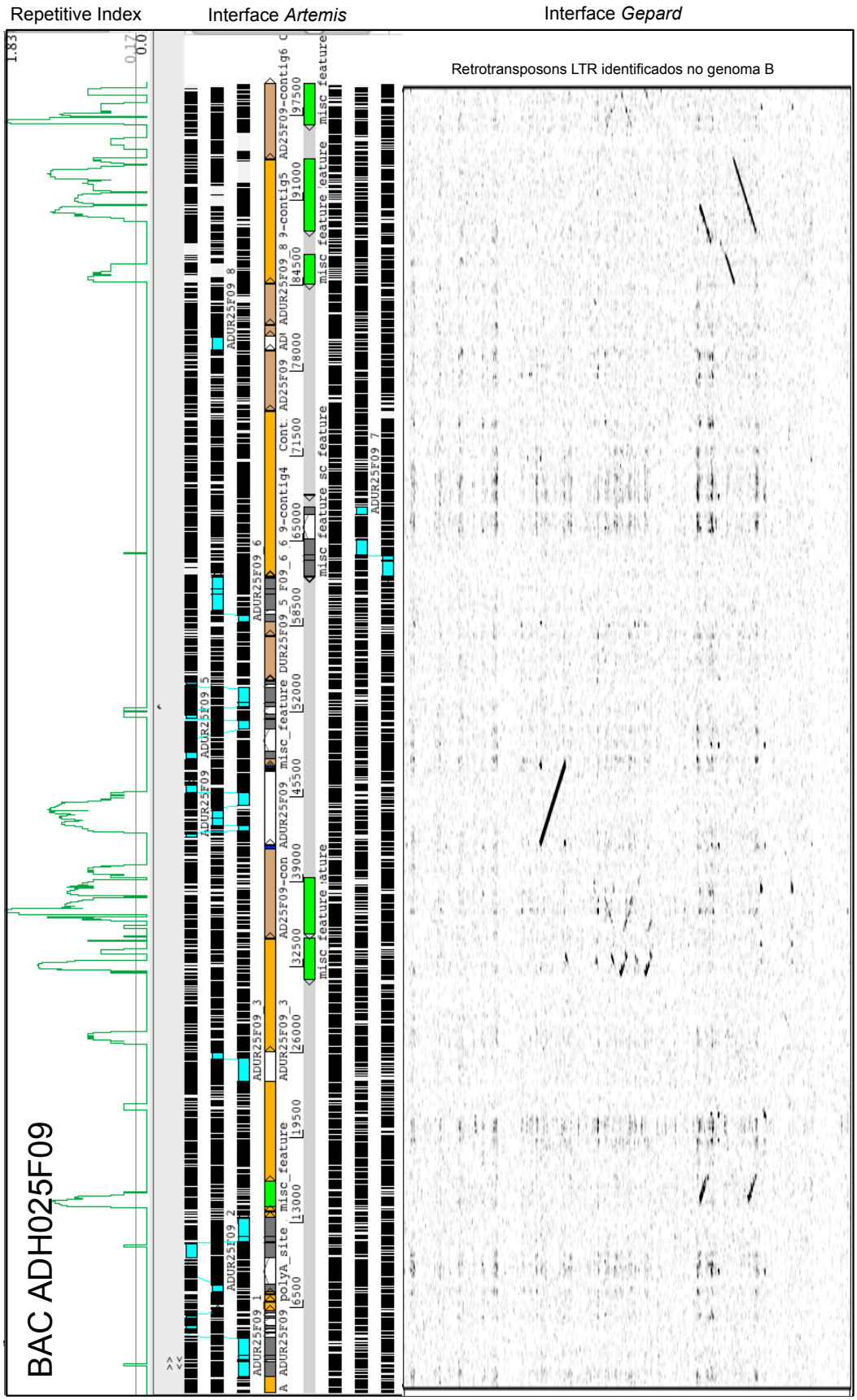


Figura 23: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH25F09, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

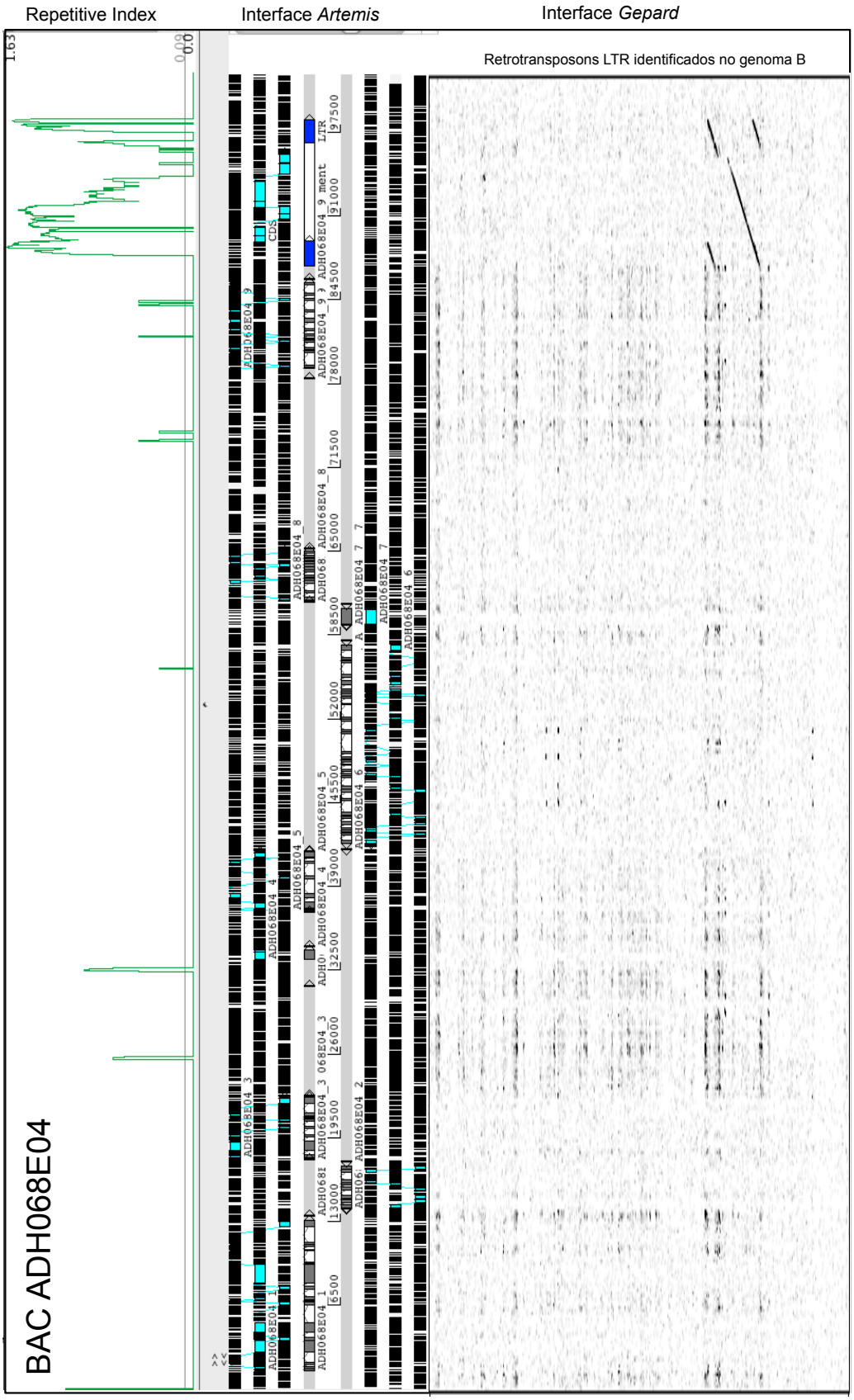


Figura 24: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH68E04, utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

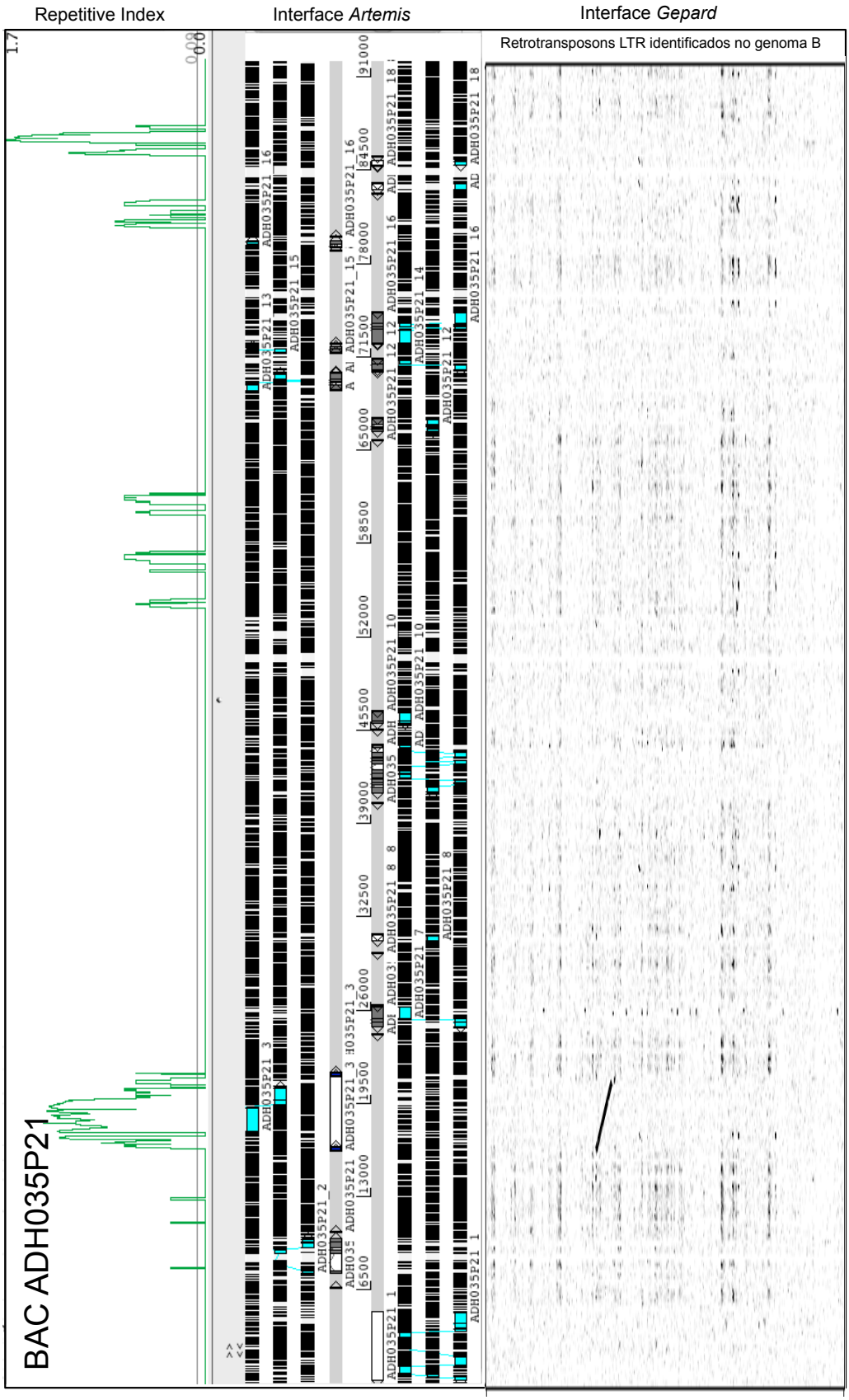


Figura 25: Gráfico mostrando a compilação de análises feitas para o clone BAC ADH035P21 utilizando os programas Artemis e Gepard. O gráfico que representa o resultado para o cálculo de índice repetitivo também consta na análise.

3.4 Estimativa da data de transposição de retrotransposons LTR no genoma A e análise de frequência

No total, 20 retrotransposons LTR completos estavam presentes nos clones analisados de *A. duranensis*: sete Feral; três FIDEL; três Pipa; dois Gordo; dois Mico; dois Matita e um RE128. A média das estimativas das datas de transposição de todos os elementos foi de 1,38 Ma (milhão de anos) (figura 26), com apenas duas das datas estimadas sendo mais antigas do que 3,5 milhões de anos, data de divergência entre os genomas A e B.

Sequências completas de retrotransposons LTR estavam presentes em cerca de 14,5% da sequência genômica de *A. duranensis* analisada, sendo que 14% estava representada por retrotransposons cuja idade de inserção foi estimada em menos de 3,5 milhões de anos.

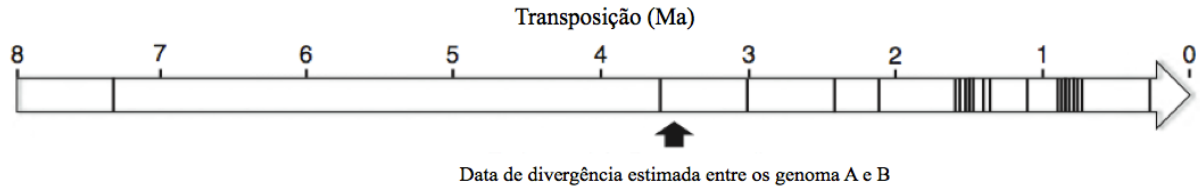


Figure 26: Estimativa da data dos 20 eventos de transposição presentes em doze regiões genômicas de *Arachis duranensis* representadas em uma linha do tempo. As linhas verticais dentro da flecha representam os eventos de transposição. A data estimada de divergência evolutiva dos genomas A e B, cerca de 3,5 milhões de anos atrás, está representada por uma seta preta sólida.

Além disso, sequências remanescentes de elementos e LTRs-solo foram muito comuns. Para todos os clones analisados (1,26 Mb), Feral, FIDEL e seus respectivos resquícios de fragmentos abrangeram cerca de 16,8% do total das sequências; Pipa e Pipoka, 8,2%; Gordo, 3,2%; Curu, 1,7%; RE128 1,6%; Mico, 1,5%; Matita 0,9% e Grilo 0,7% (tabela 4 e figura 27). Os dez elementos com sequência completa, juntamente com seus fragmentos e partes remanescente explicam mais de um terço da sequência analisada de *A. duranensis* (34,6%).

Tabela 4: Porcentagem de retrotransposons LTR presentes em cada BAC e porcentagem de cobertura de cada elemento em todos os BACs de *A. duranensis*.

Clone BAC *A. duranensis*	FIDEL/ Feral	Pipa/ Pipoka	Gordo	Curu	RE128	Mico	Matita	Grilo	Porcentagem de elementos
ADH0051I17-83F22	46,9	9		6,6					62,4
ADH177M04	20,2	9,9	1,5	6,7		10,6			48,9
ADH179B13	11,1	7,8	14,6	5,1					38,7
ADH129F24	27,6							8,8	36,4
ADH167F07	10	19,4		3,4					32,8
ADH079023-72J06	13,2	10,1	6,2						29,5
ADH25F09		10,4			9,5		6,2		26,1
ADH068E04					11				11
ADH18B08						7,6			7,6
ADH035P21							5,9		5,9
Porcentagem de cobertura em todos os BACs	16,8	8,2	3,2	1,7	1,6	1,5	0,9	0,7	

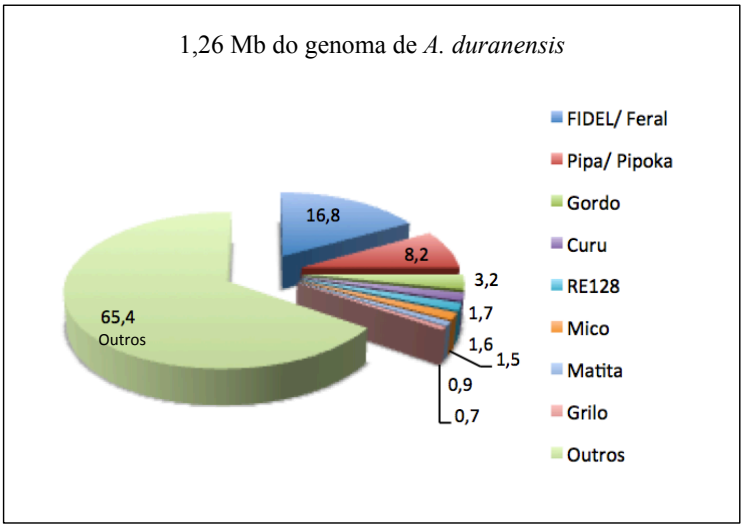


Figura 27: Gráfico ilustrando a frequência em porcentagem dos 10 tipos de retrotransposons LTR identificados nas 12 regiões genômicas de *A. duranensis* (1,26 Mb).

3.5 Comparação entre sequências homeólogas nos genomas A e B de *Arachis*

Foram selecionados dois clones BAC, sendo um oriundos da biblioteca B de *A. hypogaea* e um da biblioteca de *A. duranensis,* contendo a sequência referente ao gene *Ara h1*, alérgeno de amendoim. Ambos apresentaram alta similaridade e foram definidos portanto, como homeólogos, possibilitando a comparação entre clones representantes dos genomas A e B.

A comparação realizada por meio de gráfico de plotagem utilizando os dois clones BAC indicou regiões microssintênicas contendo cerca de 57 Kb em AHF417E07 (genoma B) e 53 Kb em ADH035P21 (genoma A) (figura 28). A microssintenia entre estas regiões é delimitada por regiões sem semelhança significativa entre os genomas A e B nas extremidades 5' e 3' das sequências de ambos os clones. Dentro da região microssintênica existem segmentos contendo alto nível de identidade, sendo que a região mais longa com 290

pb, exibiu 100% de identidade. Esses segmentos altamente semelhantes estavam delimitados por pequenos segmentos (mais frequentemente mutações do tipo *indels*), sem similaridade significativa na sequência. As sequências que flanqueiam as regiões microssintênicas são repetitivas em ambos os genomas, porém com natureza totalmente diferente para os dois genomas. Na extremidade 5' da região vicinal à região microssintênica, o genoma A abriga uma inserção do retrotransposon Matita (data de transposição estimada em 1,1 milhões de anos). No mesmo local, o genoma B abriga uma inserção do retrotransposon Mico (3,8 milhões de anos). No limite 3' da região microssintênica, ambos os genomas possuem sequências de DNA repetitivo que são completamente diferentes em sua natureza, tornando difícil a definição de sua origem. No genoma B, essa região abriga alguns fragmentos de um retrotransposon LTR denominado "Yara".

Os segmentos sem similaridade significativa, situados dentro da região sintênica, também mostraram uma tendência a serem de origem repetitiva. Em uma dessas regiões no genoma B há uma inserção de um retrotransposon completo denominado "Joka", que possui data de inserção estimada em 423 mil anos. Outros segmentos menores detectados, possivelmente, também são repetitivos, porém sem origem precisa.

O conteúdo gênico dos dois genomas dentro da região microsintênica foi predito como sendo o mesmo, com mesma ordem e orientação. A predição para estes genes, na direção 5' - 3' foi de: duas proteínas putativas, uma proteína transmembrana da família BT1; uma proteína putativa; duas proteínas inibidoras de tripsina alfa-amilase, uma proteína de proteossomo e uma proteína *Ara h1* de reserva na semente (tabela 3).

O padrão de "granulação" ou "pontos" apresentados no gráfico representam na maioria, sequências curtas com baixa complexidade. Como observado, os genes, os retrotransposons completos e seus fragmentos truncados formam coordenadas com espaços menos densos no gráfico de plotagem, ao passo que sequências intergênicas e sem origem repetitiva tendem a formar um "caminho" granular denso.

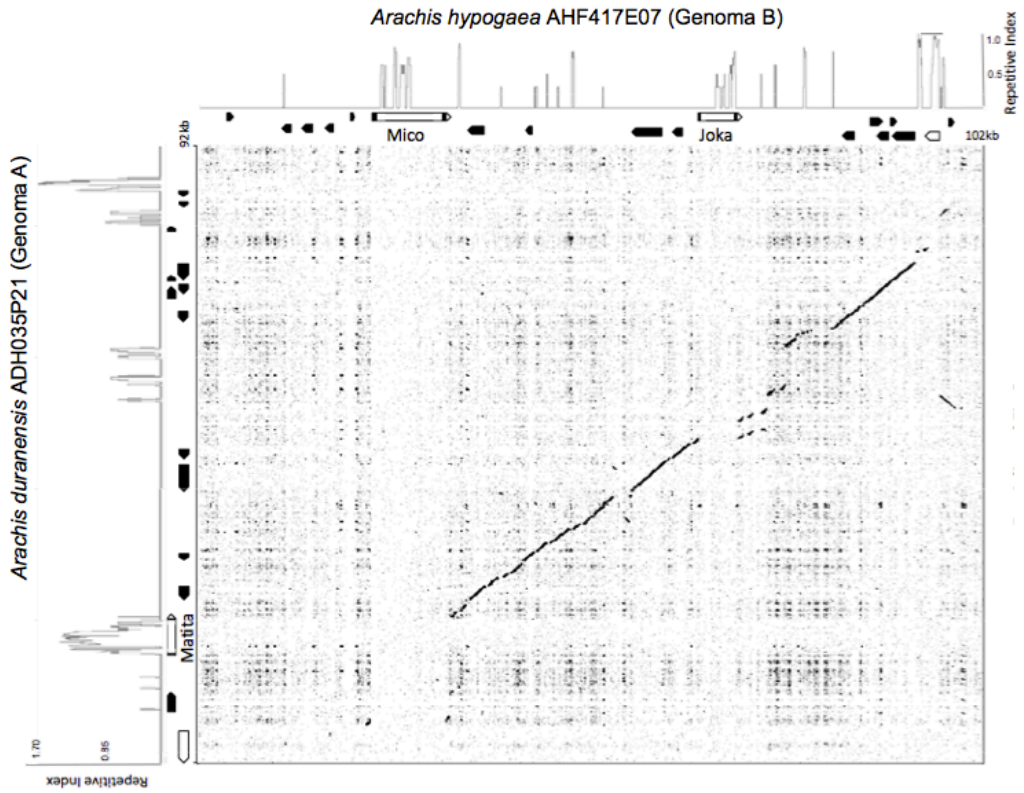


Figura 28: Gráfico de comparação das sequências homeólogas dos genomas A (clone BAC ADH035P21) e B (clone BAC AHF417E07). Gráfico de plotagem desenvolvido no *software* Gepard.

As outras duas regiões homeólogas dos genomas A e B comparadas estavam presentes nos clones BAC de *A. duranensis* (ADH068E04 – genoma A) e de *A. ipaënsis* (AIPA147A20 - genoma B) (figura 29). Essas duas regiões possuem um marcador para gene que codifica a enzima DNA girase (Leg 128, Fredslund *et al*., 2006) e apresentaram microssintenia ao longo de 43kb e 47kb, respectivamente. As regiões microssintênicas estão situadas nas extremidades 5' e 3' dos clones BAC de *A. duranensis* e *A. ipaënsis*, respectivamente. As regiões microssintênicas foram representadas quase que inteiramente por segmentos de sequência muito semelhantes, interrompidas por regiões sem similaridade aparente (mais frequentemente, correspondendo a mutações do tipo *indels*). O segmento mais longo com 100% de identidade possui 331 pb. Foram detectadas 15 interrupções distintas, sendo quatro repetitivas.

Os genes preditos para as regiões microssintênicas nos dois clones são quase que inteiramente os mesmos: proteína glicosil fosfatidil inositol transamidase; proteína de biossíntese de oligossacarídeo; proteína transportadora mitocondrial; duas elongases de ácidos graxos, e os genes que codificam a enzima girase de DNA. A sequência de *A. ipaënsis* AIPA147A20 (genoma B) contém um gene putativo extra antes do primeiro gene que codifica

a elongase de ácido graxos, e um domínio C2 adicional antes do gene que codifica a girase. No clone AIPA147A20, o gene que codifica a girase está truncado no final da sequência do clone de BAC.

As sequências dos genes putativos preditos nos clones homeólogos estão disponíveis em bancos de dados de domínio público com os seguintes números de acesso: CCW28718 – CCW28854.

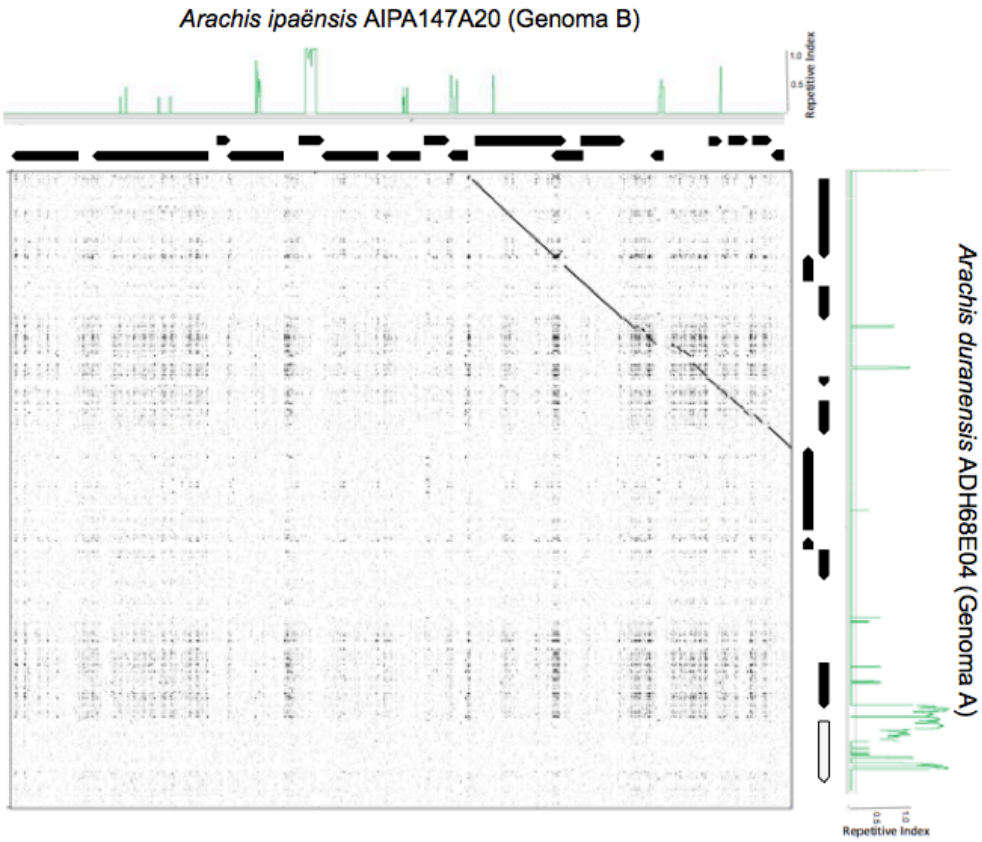


Figura 29: Gráfico de comparação das sequências homeólogas dos genomas A (clone BAC ADH68E04) e B (clone BAC AIPA147A20). Gráfico de plotagem desenvolvido no *software* Gepard.

4. Discussão

Uma característica intrigante acerca dos genomas eucarióticos é o fato de que os genes, que possuem maior significado funcional, ocupem apenas uma pequena fração do genoma. Já o DNA repetitivo ocupa a maior parte do genoma e determina a estrutura em larga escala dos cromossomos (Schmidt & Heslop-Harrison, 1998).

Neste trabalho, foi investigada a natureza do DNA repetitivo presente no subgenoma A de amendoim. O amendoim cultivado é um alotetraploide do tipo AB (Smartt *et al*., 1978; Smartt & Stalker, 1982). Em termos de investigação acerca da estrutura genômica de leguminosas, o amendoim pode ser considerado muito informativo, pois filogeneticamente, ele é um grupo separado da maioria das outras leguminosas de importância econômica (Lavin *et al*., 2001; Lewis *et al*., 2005). Além disso, o amendoim torna-se um importante objeto de estudo, em virtude dos componentes genômicos A e B proximamente relacionados. A partir de alguns estudos foi possível estimar a data de divergência entre esses dois componentes genômicos em 3 – 3,5 milhões de anos atrás (Nielen *et al*., 2012; Moretzsohn *et al*., 2013). Apenas muito recentemente eles se uniram por meio de um evento de poliploidização (Bertioli *et al*., 2011).

Em trabalhos anteriores, hibridizações in situ por fluorescência (FISH) feitas utilizando como sondas 27 clones BAC de *Arachis duranensis*, selecionados quanto à presença de genes, apresentaram sinais de hibridização dispersos principalmente nos cromossomos A de amendoim (Araújo *et al.,* 2012). A observação do padrão difuso desses sinais foi semelhante ao observado quando retrotransposons identificados em *Arachis* foram utilizados como sondas (Nielen *et al.*, 2010; 2012), sendo claramente diferentes de sinais pontuais obtidos para hibridização com sondas contendo o gene *Ara h2* (alérgeno de amendoim) e *Ara h 6* (relacionado à conglutina) em *Arachis* spp. (Laura Ramos *et al*., 2006). Essas observações indicaram que os sinais eram característicos de hibridização com sequências de DNA repetitivo, o que induziu a uma investigação da natureza e frequência desses elementos repetitivos em amendoim, assim como da especificidade pelo genoma por meio do sequenciamento de alguns clones BACs.

Foram utilizados os métodos Sanger e 454 para o sequenciamento desses clones. Para um clone específico, foram comparadas montagens obtidas por ambos os métodos. Apesar dos resultados do sequenciamento terem sido amplamente consistentes, algumas inversões e regiões de baixa complexidade (segmentos A/T e outros) foram detectadas nesse clone,

corroborando a grande dificuldade para gerar uma representação completamente correta de sequências genômicas que contêm elementos repetitivos e que utilizam essas tecnologias de sequenciamento (Kuhn *et al*., 2012). Em relação ao sequenciamento pela técnica Illumina (http://www.illumina.com/), que gera sequências de menor tamanho, esses problemas tendem a ser mais agudos. A montagem de sequências genômicas mais curtas, por exemplo, são muitas vezes realizadas a partir do descarte de sequências repetitivas, pois estas dificultam a montagem. Em contrapartida, sequências repetitivas perfazem a maior parte do genoma, o que torna complicado o estudo de forma mais abrangente.

Juntamente com a localização de sequências de cópias únicas (genes) pela técnica de FISH, mapeamento genético e sequenciamento do genoma, a localização de sequências repetitivas nos clones podem auxiliar o entendimento acerca dos processos evolutivos que ocorrem no genoma a longo, médio e curto prazo, além de proporcionar uma visão mais aprofundada acerca da organização em larga escala de sequências repetitivas específicas do genoma de amendoim.

No total, 1,26 Mb de sequência genômica derivadas de 12 regiões genômicas do genoma A de *A. duranensis* foram analisadas. Para identificar sequências repetitivas foram utilizadas plotagens, buscas por similaridade com sequências depositadas em bancos de dados e *softwares* como o LTR_FINDER. Plotagens comparativas utilizando as sequências dos clones BACs contra elas mesmas e umas contra as outras mostraram que todas essas sequências continham trechos repetitivos, indicando a presença de retrotransposons LTR. Foram identificados dez diferentes tipos de retrotransposons LTR completos nos clones BAC analisados: FIDEL e Matita, retrotransposons de amendoim já identificados (Nielen e*t al*, 2010; 2012), e oito novos retrotransposons.

O elemento mais abundante, denominado Feral, apresentou alta similaridade com as sequências dos LTRs e na região 3' não traduzida de FIDEL, porém com uma grande diferença nas regiões de codificação e região 5' não traduzida. Além disso, enquanto FIDEL é um retrotransposon autônomo que codifica todas as proteínas essenciais para sua transposição, o retrotransposon Feral não codifica a enzima transcriptase reversa, portanto, é considerado como um elemento não-autônomo. Parece mais provável que Feral seja um parasita de FIDEL. Além de retroelementos completos, foram identificadas muitas sequências repetitivas fragmentadas, especialmente LTR-solos e sequências remanescentes de elementos de transposição. No total, os elementos completos e fragmentos derivados dos elementos

FIDEL e Feral compõem cerca de um sexto das regiões analisadas do genoma de *A. duranensis*.

O terceiro elemento mais frequente foi o retrotransposon Pipa. Este elemento não possui regiões que codificam proteínas conhecidas presentes em retrotransposons, o que provavelmente o tornou do tipo não-autônomo. No entanto, ele possui uma ORF com função desconhecida próxima a região do LTR 5'. Um retrotransposon autônomo com similaridade significativa à sequência de Pipa também foi identificado e chamado de Pipoka.

Uma característica interessante entre Feral e Pipa é que ambos codificam proteínas que são completamente diferentes daquelas codificadas por seus prováveis pares autônomos. Estudos anteriores demonstraram que os transposons de DNA e retrotransposons podem "capturar" sequências de genes (Alix *et al*., 2008) e a amplificação de tais sequências de codificação podem desempenhar um papel importante não só na amplificação de genes, mas também na divergência entre genomas. Embora não possamos atribuir qualquer significado biológico para os elementos não-autônomos descritos neste trabalho, torna-se evidente que Feral e Pipa não são derivados de seus pares autônomos simplesmente pela degradação mutacional e exclusão de partes de sequências. Eles provavelmente evoluíram por meio de mecanismos complexos ou atípicos para se tornarem elementos não-autônomos.

Outro elemento abundante, chamado de Gordo, possui repetições em tandem dentro de seus LTRs e também é do tipo não-autônomo. Retrotransposons menos abundantes e autônomos foram denominados de Curu (que tem LTRs excepcionalmente longos com 3448 pb), RE-128, Mico e Grilo.

Estes oito novos retrotransposons ocupam uma proporção surpreendente das regiões genômicas analisadas (1,26 Mb). FIDEL e Feral, juntamente com Pipa e Pipoka perfazem um quarto das sequências, e estes quatro, juntamente com Gordo e Curu, perfazem um terço.

A frequência de elementos repetitivos nos genomas deve ser analisada levando-se em conta o tipo de sequenciamento e as estratégias de montagem utilizadas, além do tamanho do genoma. A presença elevada de retrotransposons LTR de Classe I, observada no gênero *Arachis* nesse estudo, já havia sido documentada em muitas monocotiledôneas e dicotiledôneas, tais como *Sorghum bicolor* (55%), milho (79%), arroz (26%) (Paterson *et al*., 2009), *Medicago truncatula* (26,5%) (*Medicago truncatula* Genome Project), soja (42%) (Schmutz *et al*., 2010) e *Lotus japonicus* (19,23%) (Sato *et al*., 2008). Além disso, elementos não-autônomos abundantes e de origem recente, LTRs-solo e agrupamentos de retroelementos

dentro de outros, têm sido observados em outros genomas de plantas (Wawrzynski *et al.,* 2008; Schmutz *et al.,* 2010).

A evidência de que os componentes repetitivos presentes nas sequências genômicas A e B de *Arachis* têm divergido rapidamente é apoiada pela constatação de que os retrotransposons mais abundantes estão predominantemente localizados nos cromossomos de apenas um genoma (Nielen *et al*., 2010). É também corroborada pelo fato de que, quase todos os eventos de transposição determinados têm menos de 3,5 milhões de anos, que é a data estimada para a divergência evolutiva entre os genomas A e B (Nielen *et al*., 2012; Moretzsohn *et al.,* 2013). No total, 14% das regiões sequenciadas do componente genômico A de *A. duranensis* está ocupada por retrotransposons completos inseridos a menos de 3,5 milhões de anos.

O padrão granular presente nos gráficos de plotagem são relativos à presença de sequências curtas com baixa complexidade que se acumulam numa escala de tempo evolucionário longa em virtude do "deslize" da DNA polimerase durante a replicação. Este padrão está ausente nos gráficos quando são observados os exons de genes, presumivelmente, porque mutações em exons tendem a ser eliminadas pela seleção natural. Esse padrão também está ausente nos retrotransposons completos e nas sequencias remanescentes, provavelmente em virtude de uma origem recente.

O *software* FGENESH previu muitos genes ao longo das sequências dos clones BAC analisados. Esses resultados foram corroborados pela comparação com sequências disponíveis no banco de dados do *PfamA*, sequências de proteínas de Arabidopsis ou soja e, com ESTs de *Arachis*, que forneceram fortes evidências da ocorrência de transcrição a partir de retrotransposons.

Caso não sejam identificados corretamente, os retrotransposons e suas sequências remanescentes podem tornar confusa a anotação do genoma de amendoim que está em andamento. Além disso, como observado por Wang *et al*. (2012), genes relacionados a retrotransposons podem ser anotados dentro da fração gênica, subestimando o conteúdo repetitivo do elemento no genoma.

Para as duas regiões homeólogas dos genomas A e B de *Arachis* que foram comparadas, as porções microssintênicas estão flanqueadas por regiões de DNA repetitivo que eram completamente diferentes nos genomas A e B. Em ambos os casos, dentro das regiões microssintênicas, existem segmentos altamente conservados (com aproximadamente 95% de

identidade), pontuados por segmentos sem homologia significativa. Possivelmente estes segmentos pertenciam a uma classe repetitiva.

Esses dados indicam o papel fundamental de DNA repetitivo na erosão de similaridade de sequência desde a divergência estrutural dos genomas A e B. Esta divergência não é distribuída uniformemente, e sim concentrada principalmente nas regiões intergênicas. Portanto, as sequências dos genes e a ordem entre eles permanecem altamente conservadas. Isso fornece uma solução para o aparente paradoxo da dinâmica entre elementos repetitivos e a conservação das frações gênicas na estrutura de genomas.

O artigo científico referente a esse estudo encontra-se no anexo 4.

5. Conclusão

Neste estudo, foi demonstrado que uma proporção substancial do componente altamente repetitivo do genoma do amendoim está representado por um número relativamente pequeno de retrotransposons LTR. Três dos elementos mais abundantes são não-autônomos, sendo que dois deles parecem ter seu par autônomo para auxiliar na atividade de transposição.

A partir disso, tornou-se evidente que estes retrotransposons e seus respectivos fragmentos truncados poderiam tornar-se um grande fator dificultador para a anotação de genes, caso não fossem corretamente identificados.

Foi mostrado também, que esses elementos possuem predominantemente uma origem evolutiva recente, com data de divergência evolutiva aparentemente posterior àquela estimada para os genomas A e B do amendoim. Claramente, estes elementos contribuíram de forma notável para a divergência desses genomas. Os genomas A e B possuem segmentos de DNA altamente semelhantes, porém interrompidos por segmentos de DNA repetitivo que não apresentam sequências correspondentes. Além disso, observações sobre dois pares de segmentos homeólogos dos genomas A e B indicam que os retrotransposons identificados neste estudo, juntamente com outros DNAs repetitivos têm desempenhado um papel importante na remodelação do genoma, especialmente em regiões intergênicas, ao longo do tempo evolutivo.

Capítulo II

Caracterização do conteúdo repetitivo dos genomas de *Arachis duranensis* (genoma A) e *Arachis ipaënsis* (genoma B) e sua distribuição em cromossomos de amendoim

1. Introdução

Ao longo dos últimos anos, o sequenciamento genômico de plantas evidenciou que os genomas são compostos principalmente por sequências repetitivas, que diferem entre si em sua distribuição, tipo de sequência, número de pares de bases, número de cópias e regiões codantes. Dois tipos são reconhecidos: as sequências repetitivas dispersas de forma intercalada e aquelas organizadas em *tandem*. As sequências repetitivas dispersas encontram-se normalmente distribuídas em todo o genoma e constituem uma considerável fração. Dentre essas sequências, as mais frequentes são os elementos genéticos de transposição (TEs), descobertos primeiramente no genoma do milho (McClintock, 1951).

Os TEs estão classificados em Classe I ou II, dependendo da presença de intermediários de RNAm (RNA mensageiro) em seu mecanismo de transposição. Os elementos de Classe I ou retrotransposons utilizam o mecanismo de transcrição reversa para inserção da cópia em novo sítio do genoma, sempre catalizado pela enzima transcriptase reversa. Nessa Classe estão os retrotransposons com longas repetições terminais (*Long Terminal Repeats*), denominados retrotransposons LTR, considerados os mais abundantes na maioria dos genomas de plantas (Pearce *et al*., 1996; SanMiguel *et al*., 1996; SanMiguel & Bennetzen, 1998; Turcotte *et al*., 2001; Gaut & Ross-Ibarra, 2008), sendo divididos em duas principais superfamílias denominadas *Ty3-Gypsy* e *Ty1-Copia* (Xiong & Eickbush, 1990).

Os retrotransposons podem atuar, por exemplo, em resposta às condições ambientais e estresses, em etapas do metabolismo, desenvolvimento, reprodução e morfogênese (Schulman, 2013). Isso acontece porque as duplicações recorrentes, atividades de transcrição e excisão/integração desses TEs implicam em várias consequências, tais como o aumento do tamanho do genoma hospedeiro (Estep *et al*., 2013; Park *et al*., 2012), mutações de inserção (Duangpan *et al*., 2013), controle epigenético (da expressão) de genes (Lisch & Bennetzen, 2011), duplicação e transposição de genes completos ou segmentos gênicos (Wicker *et al*., 2010) e, ainda, na atuação desses elementos como promotores ou sinais regulatórios (Kashkush & Khasdan, 2007). Devido a essa capacidade de transposição, os TEs podem estar presentes em um grande número de cópias no mesmo genoma e, apesar da degeneração dessas sequências ocorrer evolutivamente, sabe-se que sequências remanescentes, fragmentos e LTRs-solo derivados de recombinação ilegítima (Shirasu et al., 2000; Bennetzen et al., 2005), podem permanecer no genoma, mesmo na ausência do elemento completo.

Amendoim era até muito pouco tempo atrás, uma cultura mal representada em termos de sequência completa de seu genoma e de estudos sobre genômica estrutural, comparando-se a outras plantas-modelo ou àquelas com importância econômica mais evidente, tais como *Arabidopsis*, arroz e soja. Entretanto, recentemente, mesmo na ausência da sequência do genoma completo de amendoim ou de qualquer parente silvestre, estudos contemplando a estrutura genômica organizacional do amendoim foram realizados, principalmente via citogenética (Seijo e*t al*., 2004; 2007; Nielen *at al*., 2010; 2012; Bertioli *et al*., 2013), construção de mapas de ligação (Halward *et al*., 1993; Garcia *et al*., 1995; Burow *et al*., 2001; Moretzsohn *et al*., 2005; 2009; Leal-Bertioli *et al*., 2009; Foncéka *et al*., 2009; Varshney *et al*., 2009; Hong *et al*., 2010; Khedikar *et al*., 2010; Gautami *et al*., 2012; Qin *et al*., 2012; Shirasawa *et al*., 2013) e por meio de estudos utilizando bibliotecas BAC (Yüksel & Paterson, 2005; Guimarães *et al*., 2008).

Nos últimos anos, uma iniciativa internacional designada IPGI (*The International Peanut Genome Initiative* - http://www.peanutbioscience.com) se dedicou com sucesso na obtenção das sequências completas do genoma dos progenitores do amendoim, *A. duranensis* e *A. ipaënsis* (http://peanutbase.org/files/genomes/). As sequências dos genomas dos parentais diploides são consideradas excelentes arcabouços para a montagem do genoma tetraploide completo do amendoim: o genoma de *A. duranensis* como um modelo para o subgenoma A do amendoim, e *A. ipaënsis*, para o subgenoma B. Análises comparativas entre os genomas A e B nas espécies silvestres e os subgenomas A e B na espécie cultivada contribuirão para decifrar alterações genômicas que levaram à domesticação do amendoim.

Como descrito em outras plantas, o genoma do amendoim também está repleto de TEs (Bertioli *et al*., 2013). Junto aos MITE (*Miniature Inverted-Repeat Transposable Element* - Miniaturas de Elementos de Transposição com Repetições Invertidas), os retrotransposons LTR são os que têm contribuído notavelmente na organização e evolução desse genoma. O primeiro retrotransposon identificado em amendoim foi denominado FIDEL, pertencente à superfamília *Ty3-Gypsy* (Nielen *et al*., 2010). Análises da distribuição desse elemento por hibridização in situ por fluorescência em cromossomos de amendoim mostraram sua maior frequência no subgenoma ou componente repetitivo A, principalmente nas regiões dos braços. O número de cópias estimado para FIDEL foi de 3.000 e 820 por genoma haploide, respectivamente, sendo este elemento menos frequente em regiões próximas a genes de cópia única. Apesar da maioria das cópias de FIDEL possivelmente encontrarem-se epigeneticamente silenciadas, dados de estudos com ESTs (*Expressed Sequence Tags* –

Etiquetas de Sequência Expressa) indicaram ocorrência de atividade, e que a transcrição desse elemento pode estar ligada a estresses bióticos e à seca (Brasileiro *et al*., 2012). Outro elemento identificado, denominado Matita, é um *Ty1-Copia* pertencente à linhagem Bianca (Nielen *et al*., 2012). Possui aproximadamente 520 cópias no genoma haploide do amendoim e está distribuído na mesma frequência nos subgenomas A e B, principalmente nas regiões distais dos braços dos cromossomos. Matita exibe uma tendência significativa de abundância em regiões próximas a genes de resistência. Em estudo recente (Bertioli *et al*., 2013), já discutido no Capítulo I, análises das sequências (1,26 Mb) de 12 clones BAC pertencentes a *A. duranensis* (genoma A), determinaram que a maioria do conteúdo repetitivo poderia ser explicado pelas várias cópias de apenas dez diferentes retrotransposons LTR, incluindo FIDEL e Matita e seus respectivos fragmentos e LTRs-solo.

Com o avanço das metodologias de sequenciamento massal para obtenção de genomas eucarióticos completos, a ocorrência de TEs pode ser considerada uma presença inconveniente, que dificulta enormemente a montagem mais precisa das sequências de DNA e anotação acurada de genes, inclusive para o amendoim. Entretanto, reconhece-se que a onipresença e abundância desses elementos de repetição em genomas de eucariotos indicam sua importância no genoma (Vesely *et al*., 2012).

Apesar da função ainda indeterminada para a maior parte desses elementos, fica claro que estes representam um fator crítico na evolução gênica e genômica, pois evolutivamente, podem apresentar um intenso nível de atividade. Portanto, a identificação e anotação desses retrotransposons LTR são consideradas informações fundamentais necessárias para a compreensão da estrutura, do funcionamento e da evolução do genoma de amendoim. Além disso, o conhecimento dessas sequências repetitivas no genoma é essencial para a anotação de genes e comparação entre sequências genômicas apropriadamente, agora que o sequenciamento massal foi realizado em *Arachis* spp. A abundância desses elementos no genoma, sem que haja uma anotação consistente, pode comprometer toda a ordem estrutural da sequência do genoma de amendoim.

Portanto, nesse capítulo são apresentadas ferramentas computacionais eficientes e acessíveis bem como uma compilação dos dados de identificação, caracterização, determinação da distribuição nos genomas, caráter de autonomia para transposição, frequência, idade, relação espacial e funcional com genes, seu potencial codante e anotação apropriada do conteúdo repetitivo presente nas sequências genômicas completas de *A. duranensis* e *A. ipaënsis*, de forma comparativa entre os retrotransposons LTR nos dois

genomas. Regiões homeólogas presentes em A e B também serão comparadas para verificar possível origem de divergência desses genomas. Esse é um trabalho pioneiro, pois trata de análises comparativas entre o conteúdo repetitivo dos genomas A e B de *Arachis*, focando em retrotransposons LTR. Os dados aqui obtidos poderão auxiliar no entendimento acerca da evolução do genoma do amendoim e na anotação do conteúdo gênico proveniente do sequenciamento genômico completo recém obtido.

2. Material e Métodos

2.1 Sequenciamento genômico de *Arachis duranensis* e *Arachis ipaënsis*

Foram utilizadas sequências dos genomas de *Arachis duranensis* (acesso V14167) e *Arachis ipaënsis* (acesso KG30076) obtidas na colaboração com a UC Davis Genome Center (CA – EUA), bem como dados de sequenciamento obtidos pelo BGI (China). A montagem das sequências genômicas foi realizada por meio da colaboração das empresas BGI, USDA-ARS, Ames-IA e UC Davis Genome Center, utilizando os *softwares* De Bruijn graph based assembly (v. 2), CLC Assembly Cell, SSPACE scaffolding software, GapClosed (v. 1) e Allpaths LG assembly (v. 2), sob os auspícios da Peanut Foundation, MARS Inc. e as academias chinesas Henan Academia de Ciências Agrárias e Shandong Academia de Ciências. A lista completa das instituições envolvidas e as fontes de financiamento estão disponíveis em http://www.peanutbioscience.com. Versões parciais e definitivas dos dados de sequenciamento foram geradas e organizadas em *scaffolds* (*contigs* genômicos), totalizando 1,24 Gb (Gigabases) para o genoma de *A. duranensis* (genoma A) e 1,5 Gb para *A. ipaënsis* (genoma B).

2.2 Componente repetitivo do genoma de *A. ipaënsis* (genoma B)

2.2.1 Identificação de retrotransposons LTR

Para identificação de retrotransposons LTR no genoma B foi utilizado o *software* LTR_FINDER (Xu & Wang, 2007) que possibilitou a comparação entre as sequências genômicas em formato FASTA com um vasto banco de dados contendo sequências de retrotransposons LTR de *Arabidopsis thaliana*. A utilização do algorítmo Smith-Waterman (Smith & Waterman, 1981) auxiliou no ajuste e pareamento dos pares de LTRs candidatos identificando inicialmente as regiões TG..CA *box*, TSR (*Target Site Repeat*) e LTRs (*Long Terminal Repeat*). Também foram identificadas as regiões PBS (*Primer Binding Site*), PPT (*Polypurine Tract*) e ORFs putativas (*Open Reading Frames*) para cada retrotransposon LTR identificado. O resultado contendo as coordenadas relativas a cada elemento foi utilizado para formatar um arquivo contendo as sequências em formato FASTA, com o auxílio de um *script Perl* (http://perl.org.br/) (Anexo 3-A).

Para compreender a relação entre os retrotransposons LTR identificados e classificá-los em famílias foram utilizadas duas metodologias: agrupamento de sequências similares de retrotransposons e comparação entre sequências por meio de gráficos de pontos (*dot plots*). A primeira metodologia foi aplicada para o sequenciamento preliminar de *A. ipaënsis* (800 Mpb), enquanto a segunda, para a versão contendo o sequenciamento completo de *A. ipaënsis* (1,5 Gb).

2.2.2 Agrupamento de sequências de retrotransposons LTR

Sequências consenso (*contigs*) derivadas do alinhamento entre os retrotransposons foram produzidas pela ferramenta CAP3 (Huang & Madan, 1999). Essa ferramenta alinha e calcula valores de sobreposição entre múltiplas sequências. Para isso foram utilizados os parâmetros padrão de análise, sem valores de qualidade. O trecho de sequência que apresentou maior similaridade em cada alinhamento entre os elementos (*e-value* < $1x10^{-60}$), denominado de *top HSP* (*High-scoring Segment Pair*), foi "recortado", selecionado como referência e comparado com todas as sequências dos retrotransposons LTR, com auxílio da ferramenta BLASTn (Altschul *et al*., 1997). O resultado representando a relação entre essas sequências, na forma de agrupamentos, foi visualizado na interface do *software* Biolayout Express 3-D (http://www.biolayout.org). Esta ferramenta facilitou a conversão dos dados de sequência obtidos para gráficos 3-D baseando-se na correlação entre as sequências analisadas, possibilitando que o grande conjunto de dados pudesse facilmente ser interpretado a partir da representação esquemática dos agrupamentos das sequências.

2.2.3 Comparação entre sequências de retrotransposons LTR (gráfico de pontos ou *dot plots*)

Gráficos de comparação (*dot plots*) entre sequências de retrotransposons LTR foram produzidos pelo *software* Gepard (Krumsiek *et al*., 2007). Essa metodologia permitiu corroborar os agrupamentos produzidos pelo Biolayout Express 3-D e auxiliou a identificação e classificação dos retrotransposons LTR presentes na versão completa do sequenciamento genômico (1,5 Gb).

Foram realizadas comparações por meio de *dot plots* entre todas as sequências de retrotransposons LTR de cada agrupamento contra elas mesmas, ou ainda, contra sequências

selecionadas aleatoriamente dentro do mesmo agrupamento.

Com objetivo de ampliar o número e o tipo de retrotransposons LTR identificados pelo método de agrupamento utilizado para a versão preliminar do genoma B (800 Mb), os retrotransposons LTR identificados pelo LTR_FINDER na versão completa (1,5 Gb) foram separados em arquivos menores e comparados contra eles mesmos também por *dot plots*. O resultado dessa comparação permitiu a identificação de novos elementos, baseando-se no perfil de diagonal obtido.

2.2.4 Caracterização e classificação dos retrotransposons LTR em família e superfamília

Para a classificação dos retrotransposons LTR em famílias compostas por elementos autônomos ou não-autônomos, e superfamílias do tipo *Ty1-Copia* ou *Ty3-Gypsy* foi necessária a determinação da estrutura gênica de todos os elementos identificados. Para isso, arquivos contendo as sequências dos elementos similares em formato FASTA foram analisados separadamente pelo *software* Artemis (Rutherford *et al*., 2000). Dentro da interface desse programa foram identificadas as sequências codantes (CDS - *Coding Sequences*) com tamanho mínimo de 50 pb. Essas sequências foram comparadas com sequências de mais de 10.000 famílias de proteínas presentes no banco de dados do *Pfam-A* (Eddy, 2011) (http://pfam.sanger.ac.uk/). Com auxílio de um *script Perl* (Anexo 3-B), os resultados contendo os *hits* positivos que apresentaram *e-value* < $1x10^{-04}$ foram organizados em arquivo de texto e visualizados na interface do Artemis.

A partir da análise do conteúdo gênico, juntamente com a comparação entre o tamanho e a conservação entre as sequências dos elementos similares, foi possível realizar a classificação em família e superfamília. Para anotação e submissão em banco de dados foi selecionado um único retrotransposon LTR representante para cada família descrita para o genoma de *A. ipaënsis*. Para a seleção do elemento representante foram utilizados critérios tais como: plotagens coerentes que revelaram o perfil completo de um retrotransposon LTR; conservação entre a sequência selecionada e os demais elementos da família; escolha de elementos que apresentaram o maior número de genes necessários para a transposição e a menor estimativa da data de transposição, quando possível, além da correção da orientação da sequência para direção 5' – 3'.

Foi feita uma comparação utilizando BLASTn (*e-value* de $1x10^{-80}$) com as sequências dos elementos representantes contra o arquivo contendo todos os retrotransposons

identificados no genoma completo de 1,5 Gb pelo LTR_FINDER. Os resultados contendo as coordenadas de cada retrotransposon LTR e sua respectiva família de origem foram organizados em tabela e, com auxílio de um *script Perl*, as sequências desses elementos foram separadas. Essa separação possibilitou a inspeção manual da qualidade e confiabilidade das sequências, assim como a contagem do número de retrotransposons LTR presentes em cada família identificada no genoma de *A. ipaënsis*.

2.3 Componente repetitivo do genoma de *A. duranensis* (genoma A)

A identificação de retrotransposons LTR no genoma completo de *A. duranensis* (1,24 Gb) foi realizada pelo *software* LTR_FINDER. Utilizando as sequências dos retrotransposons LTR representantes selecionados de cada família do genoma B, foi realizada uma comparação, por meio da ferramenta BLASTn e utilizando o *e-value* de $1x10^{-80}$, com todos os retrotransposons LTR identificados pelo LTR_FINDER no genoma A. Os resultados foram organizados em tabela e, com auxílio de um *script Pyton* (http://www.atgc.org/), as sequências de elementos do genoma A similares aos do genoma B foram identificados. A caracterização, classificação em famílias, superfamílias de retrotransposons LTR e seleção de elementos representantes para o genoma A foram realizadas de acordo com a metodologia descrita para o genoma B. Com isso, foi possível obter a comparação entre as famílias de retrotransposons LTR presentes nos genoma A e B de *Arachis.* A estratégia utilizada neste estudo para a identificação de retrotransposons LTR nos genomas A e B é mostrada na figura 30.

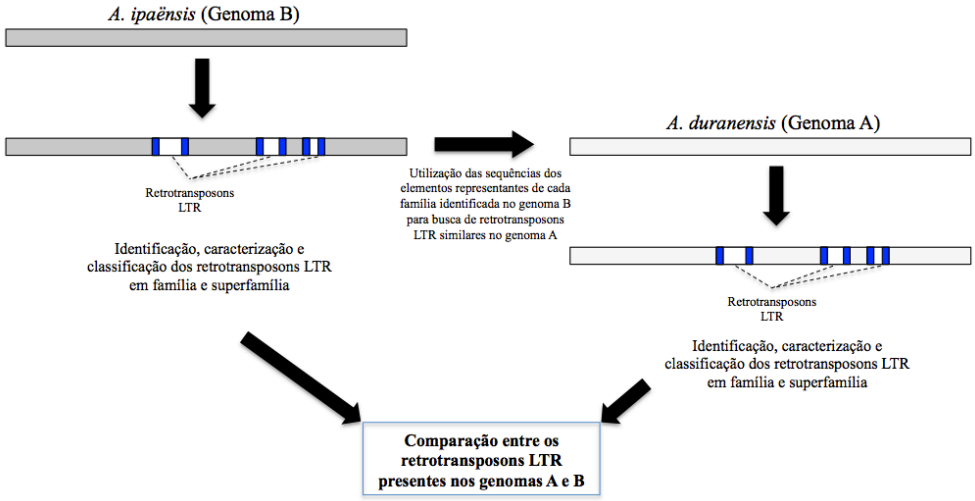


Figura 30: Identificação de retrotransposons LTR nos genomas A e B de *A. duranensis* e *A. ipaënsis*, respectivamente.

Elementos presentes apenas no genoma A foram identificados por gráfico de pontos ou *dot plot.*

2.4 Estimativa das datas de transposição dos retrotransposons LTR nos genomas de *A. duranensis* (genoma A) e *A. ipaënsis* (genoma B)

As sequências referentes aos LTRs 5' e 3' de cada retrotransposon foram separadas com auxílio de um *script Perl* (Anexo 3-C), e alinhadas pelo *software* Muscle (Edgar, 2004) para identificação dos eventos de mutação ocorridos em cada LTR (inserções, deleções ou substituições), necessárias para o cálculo de estimativa da data de transposição dos elementos. As datas de transposição foram estimadas pelo método de divergência das porções LTR, utilizando a equação t = K/2r, onde t é a idade, K é o número de substituições de nucleotídeos por local entre cada par de LTR, e r, a taxa de substituição de nucleotídeo de $1{,}3x10^{-8}$ por sítio, por ano, como descrito no estudo publicado por Ma & Bennetzen (2004). Esse cálculo e a organização dos dados foram feitos por um *script Perl* (Anexo 3-D). A média entre as datas de transposição dos elementos foi calculada e gráficos representando as médias de cada família e gráficos contendo a distribuição de idades para cada família separadamente foram organizados em Excel.

2.5 Nomenclatura e anotação das sequências de retrotransposons LTR representantes dos genomas de *A. duranensis* (genoma A) e *A. ipaënsis* (genoma B)

Famílias presentes em ambos os genomas foram igualmente nomeadas, porém a nomenclatura completa para cada elemento diferiu, baseada em diferenças tais como: genoma de origem; número de acesso no banco de dados genômico analisado; nome da família; autonomia de transposição ou não; e superfamília (*Ty1-Copia* ou *Ty3-Gypsy*).

Para a anotação das sequências dos elementos representantes foram utilizadas as principais coordenadas identificadas pelo programa LTR_FINDER, como o tamanho da sequência completa do retrotransposon, bem como de seus LTRs, e os domínios TSD, PBS e PPT. Os genes identificados pelos *softwares* Artemis e *pfamA* (ver item 2.2.4) foram anotados como pseudogenes. A estimativa da data de transposição de cada retrotransposon no seu respectivo genoma também foi incluída na anotação.

A inclusão das sequências dos elementos representantes no banco de dados GenBank

(http://www.ncbi.nlm.nih.gov/genbank/) foi realizada por meio da ferramenta opcional de submissão denominada Bankit. Para a submissão foram requeridas tabelas em formato específico contendo cinco colunas (http://www.ncbi.nlm.nih.gov/WebSub/html/help/feature-table.html), incluindo as principais características de cada elemento.

2.6 Estimativa da frequência dos retrotransposons LTR nos genomas e pseudomoléculas A e B

Para estimar a frequência de ocorrência dos retrotransposons LTR de cada família identificada no genoma de *A. duranensis* (1,24 Gb), *A. ipaënsis* (1,5 Gb) e nas suas respectivas pseudomoléculas (sequências equivalentes aos cromossomos), o *software* RepeatMasker v-4.0.3 (http://www.repeatmasker.org/) foi utilizado com os seguintes parâmetros: "-nolow" para não mascarar sequências de baixa complexidade ou repetições simples; "-x" para mascarar as sequências de interesse (sequências dos elementos representantes); "-gff" para criar um arquivo de saída em formato de tabela ou *Gene Feature Finding format* contendo as coordenadas de sequências mascaradas para cada elemento nos genomas. Foram utilizados para análises apenas os *scaffolds* genômicos contendo 2.000 pb ou mais. A partir dos resultados obtidos, cálculos foram realizados utilizando a proporção de sequências mascaradas ou utilizando o arquivo de saída em formato de tabela em Excel. Gráficos foram obtidos separadamente, mostrando a porcentagem e o número de retrotransposons LTR com sequências completas em cada família identificada nos genomas A e B.

A distribuição das famílias de retrotransposons LTR nas pseudomoléculas A01 (similar ao cromossomo 1 de *A. duranensis*) e B01 (similar ao cromossomo 1 de *A. ipaënsis*) foi avaliada por meio da ferramenta BLASTn (e-value $1x10^{-80}$). As sequências dos elementos representantes de algumas famílias selecionadas nos dois genomas foram comparadas com as sequências das respectivas pseudomoléculas e, o número de *hits* positivos foram quantificados e avaliados quanto à distribuição por meio de gráficos construídos em Excel na escala de 2 Mb.

2.7 Distribuição dos retrotransposons LTR nos cromossomos de amendoim por FISH (hibridização in situ por fluorescência)

2.7.1 Obtenção de sondas

2.7.1.1 Alinhamento de sequências de TR (Transcriptase Reversa) e desenho de *primers*

Dentre as famílias de retrotransposons LTR identificadas em *A. ipaënsis* (genoma B), algumas foram selecionadas para a análise da sua distribuição em cromossomos metafásicos de amendoim (subgenomas A e B). As sequências de todos os elementos de cada família foram alinhadas pelo programa Muscle (Edgar, 2004) e, quando necessário, correções manuais, tais como remoção de sequências e mudança na orientação foram realizadas. O alinhamento foi visualizado na interface do programa Jalview (Waterhouse *et al*., 2009). A região interna do gene que codifica a enzima transcriptase reversa (TR) foi utilizada como molde para a confecção de DNA marcado para hibridização in situ por fluorescência (FISH), pois esse domínio de sequência apresenta alto nível de conservação entre elementos de uma mesma família.

Pares de *primers* foram desenhados para a região interna do gene da TR, seguindo os seguintes parâmetros: *primers* com 20-24 bases de comprimento; composição com aproximadamente 50% de bases G/C (Guanina/Citosina); cada base nucleotídica apresentando mais de 80% de similaridade no alinhamento; produto de amplificação com tamanho entre 250-800 pb.

2.7.1.2 Extração do DNA genômico de *A. ipaënsis*

Sementes obtidas no Banco Ativo de germoplasma mantido na Embrapa Recursos Genéticos e Biotecnologia, Brasília – DF foram germinadas e folhas jovens de *A. ipaënsis* (acesso KG30076) coletadas antes da expansão foliar total. O protocolo de extração de DNA foi baseado em CTAB (Brometo de Cetil Trimetil Amônio) (Ferreira & Grattapaglia, 1998) após modificações.

Aproximadamente 200 mg de tecido vegetal foram macerados em nitrogênio líquido e acrescidos de 700 µL de CTAB 2% (CTAB 2% (p/v); NaCl a 1,4 M; Tris-HCl a 100 mM e pH 8,0; EDTA a 20 mM; β-mercaptoetanol 0,2% (v/v)) em tubos de polipropileno de 2 mL

(*Eppendorf*). As amostras foram incubadas a 65° C por 50 minutos, em seguida acrescidas de 700 µL da solução clorofórmio-isoamílico (24:1) e misturadas até a formação de uma emulsão. As amostras foram centrifugadas a 13.200 rpm por 15 minutos e a fase aquosa foi transferida para dois novos tubos (1,5 mL). Foram adicionados 600 µL de tampão CTAB 1% [CTAB 1% (p/v); Tris-HCl a 50 mM e pH 8,0; EDTA a 20 mM] e os tubos agitados lentamente e centrifugados a 13.200 rpm por 1 minuto. O sobrenadante foi descartado e o precipitado dos dois tubos foi ressuspenso em 300 µL de NaCl a 1,2 M. O volume dos dois tubos foi transferido para um único tubo, totalizando 600 µL. As amostras foram centrifugadas a 12.000 rpm por 5 minutos e o sobrenadante foi transferido para novo tubo. O DNA foi precipitado com 1 mL de etanol absoluto sob agitação lenta. As amostras foram centrifugadas a 13.200 rpm por 2 minutos e o sobrenadante descartado. O precipitado foi lavado duas vezes com 500 µL de etanol 70%, ressuspenso em 100 µL de água estéril com 0,01 mg/mL de RNA*se*-A e incubado a 37° C por 10 minutos. A quantificação foi realizada em gel de agarose 1,0% utilizando um marcador de peso molecular (*High Mass Ladder* - Invitrogen). Para utilização em reações de amplificação por PCR, as amostras foram diluídas para a concentração de 5 ng/µL e mantidas a -20° C para realização do experimento.

2.7.1.3 Amplificação das sequências da transcriptase reversa (TR) dos retrotransposons LTR por meio de PCR

Todas as reações de PCR seguiram o mesmo protocolo, modificando apenas o conjunto de *primers* e o DNA molde. Para 25 µL de reação de amplificação por PCR, foram utilizados: 2,5 µL (10%) de tampão para PCR sem $MgCl_2$ (Invitrogen); 1 µL de $MgCl_2$ a 50 mM (Invitrogen); 0,3 µL de cada *primer* a 10 µM; 0,3 µL de dNTPs a 10 mM, 0,15 µL de *Taq* Polymerase (*Taq Recombinant*-Invitrogen) e 1 µL de DNA molde (100 ng de DNA genômico ou 20 ng de DNA de plasmídeo). O ciclo utilizado em termociclador Eppendorf: 5 minutos a 94° C; 35 ciclos de 30 segundos a 94° C, 30 segundos de 55° a 66° C, dependendo do par de *primers*, 30 segundos a 72° C; e 7 minutos a 72° C. O tamanho do *amplicon* foi verificado por meio de eletroforese feita em gel de agarose 1,0% corado com brometo de etídio a uma concentração de 0,5 µg/mL em tampão TAE 1X (40 mM de Tris-acetato e 1 mM de EDTA pH 8,0). A eletroforese foi realizada durante 40 minutos sob voltagem constante de 70V (Volts). A visualização dos fragmentos foi realizada sob luz UV (ultravioleta) em foto-documentador ImageQuant300 (*GE Healthcare Life Sciences*).

Os *amplicons* (produtos de PCR) foram purificados com o *Kit GFX TM PCR DNA and Band Purification* (*GE Healthcare*), de acordo com o protocolo do fabricante.

2.7.1.4 Clonagem das sequências de TR de *A. ipaënsis* amplificadas por PCR

A ligação das sequências de TR de diferentes famílias de retrotransposons LTR de *A. ipaënsis* amplificadas por PCR foi realizada em vetor pGEM-T Easy (Promega) ou TOPO TA (Invitrogen), de acordo com o protocolo dos fabricantes.

Para a reação com o vetor pGEM-T Easy foram necessários 50 ng de vetor, 3 U de enzima T4 DNA ligase, 10% de tampão de ligação e a quantidade de produto de PCR variando de acordo com o peso molecular para 10 μL de volume final. A reação foi incubada a 4° C por aproximadamente 16 horas. Para o vetor TOPO TA foram necessários 200 ng de produto de PCR (aproximadamente 2 μL), 0,5 μL de solução salina (*Salt Solution*), 0,5 μL de vetor e 9 μL de água para um volume final de 12 μL. A incubação foi realizada em temperatura ambiente, por aproximadamente 30 minutos.

2.7.1.5 Preparo de células competentes e transformação dos vetores contendo as sequências de RT

As células de *Escherichia coli* (Cepa XL1-Blue) foram preparadas de acordo com o seguinte protocolo: 10 mL de uma cultura saturada foi inoculada em 1 L de meio LB com metade da concentração de sal (Lúria Bertani – 1% de Triptona; 0,5% de Extrato de Levedura; 0,5% de NaCl) e incubada a 37° C até alcançar uma O.D 0,6 (Densidade óptica igual a 0,6) sob comprimento de onda de 600 nm (nanômetros), mensurada com auxílio de espectrofotômetro. As células foram coletadas por centrifugação (4000 rpm por 20 minutos a 4° C) e ressuspensas duas vezes em água estéril gelada (1 L e 500 mL). Em seguida, foram ressuspensas em 20 mL de glicerol 10%, centrifugadas (nas mesmas condições) e ressuspensas novamente em 2 mL de glicerol 10%, separadas em alíquotas de 40 μL e armazenadas em freezer a -80° C.

A transformação de células competentes com o vetor contendo o *amplicon* correspondente à TR de famílias de retrotransposons LTR de *A. ipaënsis* foi feita pela técnica de eletroporação, utilizando 1 μL da reação de ligação para transformar 40 μL de células competentes. A reação foi incubada no gelo por 1 minuto e transferida para as cuvetas

resfriadas. As cuvetas foram submetidas ao eletroporador Gene Pulser (Biorad) sob as seguintes condições: resistência de 200 Ohms; capacitância de 25 μFD e 1,8 Kvolts. Após a eletroporação, a cultura foi acrescida de 1 mL de meio LB líquido para ser incubada a 37°C por 1 hora. Após esse período, 100 μL da suspensão (10%) foram plaqueados em placas de Petri contendo meio LB sólido contendo o antibiótico ampicilina (100 mg/mL), IPTG (100 mM) e X-GAL (50 ng/mL). As placas foram incubadas a 37°C por 16 horas e as colônias recombinantes (brancas) selecionadas e crescidas em meio LB líquido contendo ampicilina (100 mg/mL) sob agitação constante, a 37°C, por 16 horas.

2.7.1.6 Extração de plasmídeos e análise dos fragmentos de TR por restrição enzimática

Para extração de plamídeos resultantes da clonagem de *amplicons* de TR foi utilizado o protocolo de lise alcalina (Ahn *et al*., 2000) com modificações. As células foram coletadas por centrifugação a 13.000 rpm por 1 minuto. O sobrenadante foi descartado e o precipitado foi ressuspenso em 200 μL de tampão (Tris-HCl 50 mM pH 8; EDTA 10 mM; 20 μg de RNA*se* A). Adicionaram-se 200 μL de tampão de lise (NaOH 200 mM; SDS 1%). A solução foi homogeneizada à temperatura ambiente e acrescida de 240 μL de tampão de neutralização (acetato de potássio a 3 M, pH 5,5). A solução foi homogeneizada novamente por 3 minutos à temperatura ambiente e centrifugada a 13.000 rpm por 1 minuto. O sobrenadante foi transferido para novo tubo contendo 200 μL de isopropanol e homogeneizando por 1 minuto à temperatura ambiente. A solução foi centrifugada a 13.000 rpm por 1 minuto e o sobrenadante descartado. O precipitado foi lavado com 500 μL de etanol 70% gelado e posteriormente seco e ressuspenso em 50 μL de água *Milli-Q* estéril.

A presença e o tamanho do inserto foram detectados a partir de digestão enzimática. Aproximadamente 100 ng do plasmídeo foram digeridos com 1U de enzima de restrição *Eco*RI (Invitrogen), de acordo com o protocolo do fabricante e incubado a 37° C por 3 horas. O perfil de restrição foi visualizado em gel de agarose 1,0% corado com brometo de etídio. O marcador molecular utilizado para verificar o tamanho das bandas foi o *1 Kb Plus DNA ladder* (Invitrogen).

2.7.1.7 Sequenciamento e análise dos dados

O sequenciamento dos clones foi realizado pela Empresa Macrogen Inc. (Coreia do Sul), de acordo com o protocolo padrão baseado na Química do Big Dye® por meio do sistema de eletroforese capilar em sequenciador ABI 3730xl (http://www.macrogen.com). Para isso, foram necessários 100 ng de DNA diluídos em 20 µL de água. As sequências obtidas pela clonagem dos segmentos de DNA referentes às sequências de RT presentes em algumas famílias de elementos LTR identificadas em *A. ipaënsis*, foram inicialmente comparadas com dados de sequências não redundantes do GeneBank, utilizando BLASTx (*Basic Local Alignment and Search Tool*) (Altschul *et al*., 1997). Essas sequências também foram analisadas pela ferramenta Vecscreen (http://www.ncbi.nlm.nih.gov/tools/vecscreen/), que identifica sequências derivadas de vetores de clonagem. O processamento das sequências foi feito pelo pacote de *softwares* Staden Package (Staden *et al*., 1996). A ferramenta Pregap permitiu a remoção da sequência de vetores e verificação da qualidade das bases nucleotídicas. A ferramenta Gap4 foi utilizada para alinhar e sobrepor as sequências originadas em ambos os sentidos.

2.7.2 Hibridização in situ por fluorescência (FISH)

2.7.2.1 Marcação das sondas

Para a obtenção de sondas para realização de FISH foram utilizados de 200-300 ng de cada um dos *amplicons* (com tamanho inferior a 800 pb de comprimento) referentes aos genes que codificam a transcriptase reversa (TR) dos retrotransposons LTR pertencentes às famílias selecionadas do genoma de *A. ipaënsis*. O sistema de marcação de sonda foi baseado na metodologia descrita por Foinberg & Vogelsteis (1983; 1984), que utiliza sequências randômicas de hexa ou heptanucleotídeos como *primers*. Neste sistema, a reação de polimerização é catalisada pelo fragmento *Klenow* da DNA polimerase I, com incorporação de dNTPs marcados e não marcados fornecidos para a reação. As sondas foram marcadas com biotina-14-dCTP ou digoxigenina-11-dUTP pela técnica de *random primer* utilizando os Kit BioPrime DNA Labelling System (Invitrogen 18094-011) e BioPrime Array CGH Genomic Labelling System (Invitrogen cat. 18095-011), respectivamente. Todas as sondas foram

purificadas por precipitação etanólica, ressuspensas em 50 µL de água *MilliQ* estéril e mantidas a -20° C.

2.7.2.2 Teste para verificação de incorporação de nucleotídeos marcados *(dot blot)*

A verificação da incorporação de nucleotídeos marcados nas sondas foi feita por *dot blot*. Um fragmento de membrana *Hybond-N+* (Amersham, Pharmacia Biotech) foi recortada e colocada em uma placa de Petri de 55 mm de diâmetro contendo 3 mL de Tampão 1 (Tris-HCl 100 mM pH 7,5; NaCl 15 mM) por 5 minutos. A membrana foi removida utilizando pinça plástica e seca entre folhas de papeis filtro. Um total de 0,5 µL da sonda foi aplicada na membrana que foi armazenada à temperatura ambiente por 5-10 minutos. Após esse tempo, a membrana contendo as sondas foi imersa novamente na placa contendo 5 mL do Tampão 1 por 1 minuto. O Tampão 1 foi descartado e a placa foi acrescida de Tampão 2 (reagente de bloqueio dissolvido no Tampão 1 em uma concentração final de 0,5% (p/v) e aquecido a 60° C por 1 hora) e incubada sob agitação leve por 30 minutos. O Tampão 2 foi descartado e a placa contendo a membrana foi acrescida de 500 µL de Solução de Detecção (estreptavidina conjugada com fosfatase alcalina diluída no Tampão 1 na proporção de 1:500 ou anti-digoxigenina conjugada com fosfatase alcalina diluída no Tampão 1 na proporção de 1:1000) exatamente sobre a membrana que foi coberta por um pedaço de plástico estéril e incubada a 37° C sob leve agitação por 30 minutos. Após o tempo, o plástico foi removido e adicionaram-se 5 mL de Tampão 1 a placa que foi incubada por 10 minutos à tempertatura ambiente. O Tampão 1 foi removido e 5 mL de Tampão 3 (Tris-HCl 100 mM pH 9,5; NaCl 100 mM; MgCl2 50 mM) fora adicionados a placa que foi incubada por 5 minutos. O Tampão 3 foi removido e 1 mL da solução NBT/BCIP (*5-bromo-4-chloro-3-indolyl-phosphate / nitro blue tetrazolium*) foi adicionado à membrana que foi incubada por 5 minutos no escuro. A membrana foi removida da placa, lavada e seca entre papeis filtro.

2.7.2.3 Coleta de material vegetal e preparação de lâminas contendo cromossomos metafásicos de *A. hypogaea* (amendoim)

Fragmentos com aproximadamente 5 mm de comprimento da extremidade radicular de plantas jovens da espécie cultivada *A. hypogaea* (cultivar Tatu) foram isolados e tratados com 8-hidroxiquinoleina a 2 mM por 2 horas à temperatura ambiente e outras 2 horas com nova

solução, a 4°C. Foram então fixados em solução de etanol 100%:ácido acético glacial, (3:1; v/v) por 1 hora à temperatura ambiente e, após troca por nova solução fixadora, armazenados a 4° C. Amostras foram lavadas em tampão citrato de sódio 1X e hidrolisadas celulolíticamente em solução contendo celulase e pectinase por 45-60 minutos a 37° C. Cada meristema radicular foi colocado em uma gota de ácido acético 45% sobre uma lâmina histológica, coberto por uma lamínula e após leve pressão, as células meristemáticas radiculares foram gentilmente espalhadas. A lamínula foi removida por diferença de temperatura e as lâminas contendo pelo menos 5 conjuntos de cromossomos completos em metáfase, livres de restos citoplasmáticos e bem espalhados foram selecionadas em microscopia de fase para FISH.

2.7.2.4 FISH

As lâminas selecionadas foram pré-tratadas, hibridizadas, lavadas e os sítios de hibridização detectados de acordo com Schwarzacher & Heslop-Harrison (2000), após modificações. As soluções de hibridização foram preparadas utilizando uma ou duas sondas diferentemente marcadas – biotina ou digoxigenina (aproximadamente 100 ng de sonda/μL/lâmina contendo metáfases. A hibridização foi realizada durante 12-16 horas a 37° C. Lavagens pós-hibridização foram realizadas com aproximadamente 92% de nível de estringência (de acordo com Schwarzacher & Heslop-Harrison, 2000) e os sítios de hibridização detectados utilizando o anticorpo anti-digoxigenina conjugado com FITC (Roche Diagnostics) e/ou estreptavidina conjugada com o fluoróforo Alexa Fluor 594 (Life Technologies/Molecular Probes). Cromossomos foram contra-corados com DAPI (*4',6-diamino-2-phenylindole*) e observados em microscópio epifluorescente Zeiss Axiophote (Carl Zeiss , Alemanha). As imagens foram digitalmente capturadas com auxílio do *software* Axiovision e analisadas pelo *software* Adobe Photoshop CS utilizando apenas funções que afetam toda a imagem, de maneira equivalente.

2.8 Comparação entre sequências homeólogas nos genomas de *A. duranensis* e *A. ipaënsis*

Foram comparadas duas sequências genômicas homeólogas de *A. duranensis* (acesso V14167 - genoma A) e *A. ipaënsis* (acesso KG30076 - genoma B), que apresentam o

marcador Leg128 desenvolvido para leguminosas (Fredslund *et al*., 2006). Para isso, foram avaliados os conteúdos gênico e repetitivo. As duas sequências homeólogas dos genomas A e B foram comparadas por meio de gráficos de plotagens obtidos pelo *software* Gepard e caracterização pelo *software* Artemis. Para isso, foram realizadas anotações dos genes putativos, retrotransposons LTR e fragmentos de elementos identificados nas duas sequências.

A predição da localização dos genes (com estruturas de introns e exons) foi realizada pela ferramenta FGENESH (Solovyev *et al*., 2006) (http://www.softberry.com/) contra os dados de sequência da leguminosa *Medicago truncatula*. Os resultados foram modificados para visualização em Artemis utilizando um *script Perl* (Anexo 3-E). Com intuito de corroborar a localização dos genes putativos e predizer sua função, as ORFs com tamanho maior que 100 pb presentes nas sequências foram identificadas pelo Artemis e confrontadas com sequências depositadas no banco de dados *pfamA*. Os resultados foram analisados em Artemis por meio de sobreposição com os dados encontrados no FGENESH. Retrotransposons LTR foram identificados pelo *software* LTR_FINDER, ao passo que fragmentos de retrotransposons, LTR-solo e sequências em *tandem* foram identificados por meio de comparações de cada sequência contra ela mesma e contra as sequências de retrotransposons LTR já descritos, por meio de gráficos de plotagens obtidos no *software* Gepard.

3. Resultados

3.1 Componente repetitivo do genoma de *A. ipaënsis* (genoma B)

3.1.1 Identificação e agrupamento de retrotransposons LTR na sequência genômica parcial

Com auxílio do *software* LTR_FINDER, foram identificadas 1.965 sequências putativas de retrotransposons LTR no arquivo parcial do genoma de *A. ipaënsis* (800 Mb). Essas sequências putativas foram os primeiros dados do componente repetitivo identificados no genoma B de *Arachis*.

A partir desse resultado, foram geradas pelo programa CAP3, 123 sequências consenso derivadas do alinhamento das 1.965 sequências putativas de retrotransposons LTR. A relação entre as 123 sequências consenso e as sequências dos retrotransposons foi visualizada na interface do *software* Biolayout Express 3-D (figura 31). As 1.965 sequências foram organizadas em 78 agrupamentos simples ou complexos, contendo uma ou mais sequências consenso interligadas pela similaridade às sequências de retrotransposons LTR. Nos agrupamentos complexos foram possivelmente incluídos elementos autônomos e não-autônomos, ou mesmo elementos apresentando similaridade somente em parte da sequência.

Comparações por meio de gráficos de plotagens (*dot plot*) foram realizadas para cada agrupamento, em que todos os elementos foram comparados contra eles mesmos, no intuito de verificar se a seleção feita pelo *software* LTR_FINDER de fato identificou somente retrotransposons LTR, ou se outras sequências repetitivas foram equivocadamente incluídas. O padrão esperado em gráficos quando se compara uma sequência de retrotransposons LTR contra ela mesma resulta em três diagonais paralelas (discutido no Capítulo I). As análises dos gráficos revelaram que 40 agrupamentos não possuíam o perfil de diagonal esperado na comparação entre retrotransposons LTR (figura 32-A), mas sim de sequências em *tandem* (figura 32-B).

Portanto, a busca utilizando o *software* LTR_FINDER e a corroboração pelas análises de *dot plots* de cada agrupamento comparado com ele mesmo resultou em 734 retrotransposons LTR, distribuídos em 38 agrupamentos válidos: 27 simples e 11 complexos.

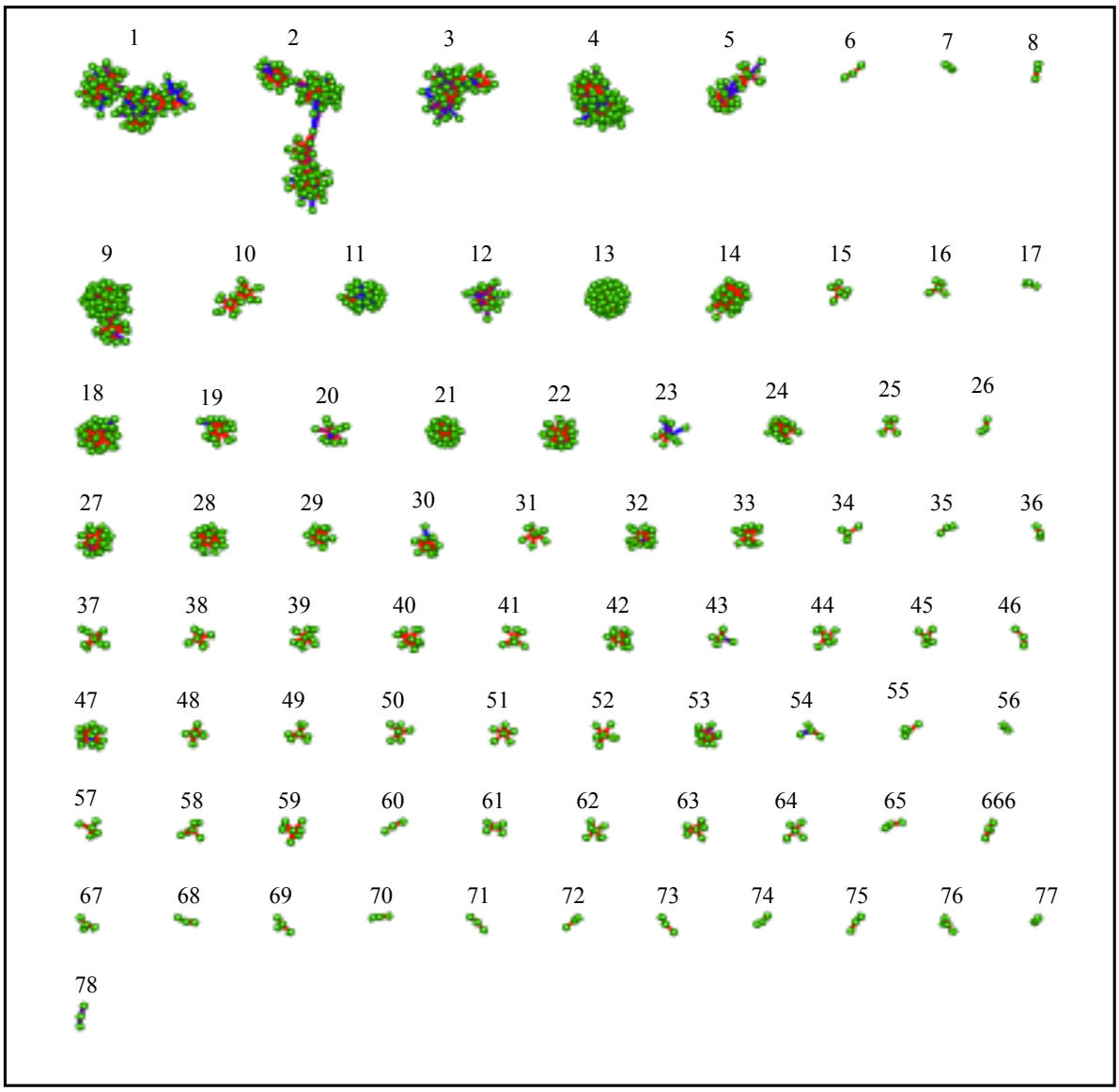


Figura 31: Representação esquemática das sequências dos 1.965 retrotransposons LTR putativos identificados pelo *software* LTR_FINDER organizados em 78 agrupamentos simples ou complexos. Agrupamentos produzidos pelo *software* Biolayout Express 3-D.

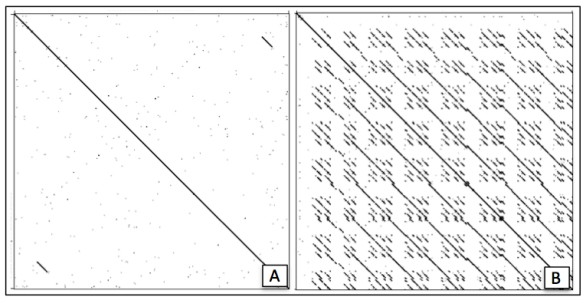


Figura 32: Gráfico de plotagem que mostra uma sequência de retrotrotransposon LTR comparada consigo mesma, revelando uma diagonal ininterrupta juntamente com outras duas diagonais menores paralelas correspondentes às sequências flanqueadoras ou LTRs (A). Gráfico de plotagem que mostra uma sequência com motivos em *tandem* plotada contra ela mesma (B). Gráficos produzidos pelo *software* Gepard.

3.1.1.1 Agrupamentos simples

Análises realizadas pelo *software* Biolayout Express 3-D resultaram em 27 agrupamentos simples, representados pelo conjunto de um ou mais retrotransposons LTR válidos e similares, ligados a apenas uma sequência consenso situada ao centro. A ligação está representada por meio de linhas coloridas, onde a cor vermelha representa maior similaridade e a azul menor similaridade entre os retrotransposons e a sequência consenso (figuras 33-A e 33-B).

Com o objetivo de identificar a relação entre os retrotransposons LTR e o nível de similaridade entre suas sequências, para cada agrupamento foram obtidas plotagens entre todos os elementos (ou a maioria deles) contra um dos elementos escolhidos aleatoriamente. As figuras 34-A e 34-B exemplificam esse tipo de comparação realizada entre os elementos pertencentes aos agrupamentos mostrados na figura abaixo.

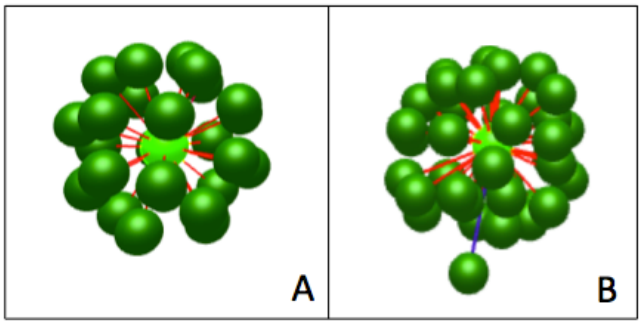


Figura 33: Representação de sequências de retrotransposons LTR organizadas em um agrupamento simples. apresentando uma sequência consenso no centro, interligada por linhas vermelhas à 21 sequências de retrotransposons LTR (A); Representação de sequências de retrotransposons LTR organizadas em um agrupamento simples. apresentando uma sequência consenso no centro interligada por linhas vermelhas a 33 retrotransposons LTR e uma linha azul ligada a apenas um deles (B). Agrupamentos produzidos pelo *software* Biolayout Express 3-D.

Para o agrupamento mostrado na figura 33-A, a comparação por gráfico de plotagem das sequências de nove retrotransposons LTR (eixo x) com uma dessas sequências escolhida aleatoriamente (eixo y) resultou em nove diagonais com o perfil padrão de retrotransposons LTR (figura 34-A). A similaridade entre esses elementos foi corroborada pela presença de linhas contínuas mais ou menos pontilhadas que determinaram maior ou menor similaridade, respectivamente. A baixa conservação entre as sequências de alguns elementos similares ocorreu possivelmente devido à mutações durante a evolução do genoma. Neste exemplo, provavelmente esses elementos fazem parte da mesma família.

Para o agrupamento mostrado na figura 33-B, a comparação por gráfico de plotagem

das sequências de 12 retrotransposons LTR (eixo x), incluindo o elemento ligado pela linha azul – designado como elemento 11, contra uma dessas sequências escolhida aleatoriamente (eixo y), resultou em 11 linhas diagonais típicas de retrotransposons LTR (figura 34-B). Apenas uma única linha diagonal diferiu das demais: aquela correspondente ao elemento ligado pela linha azul. Isso porque apenas uma das regiões LTR desse elemento é similar as dos demais elementos comparados. Portanto, nem a outra região LTR, nem a região gênica apresentaram similaridades com os demais retrotransposons, sugerindo que esse elemento seja possivelmente de outra família.

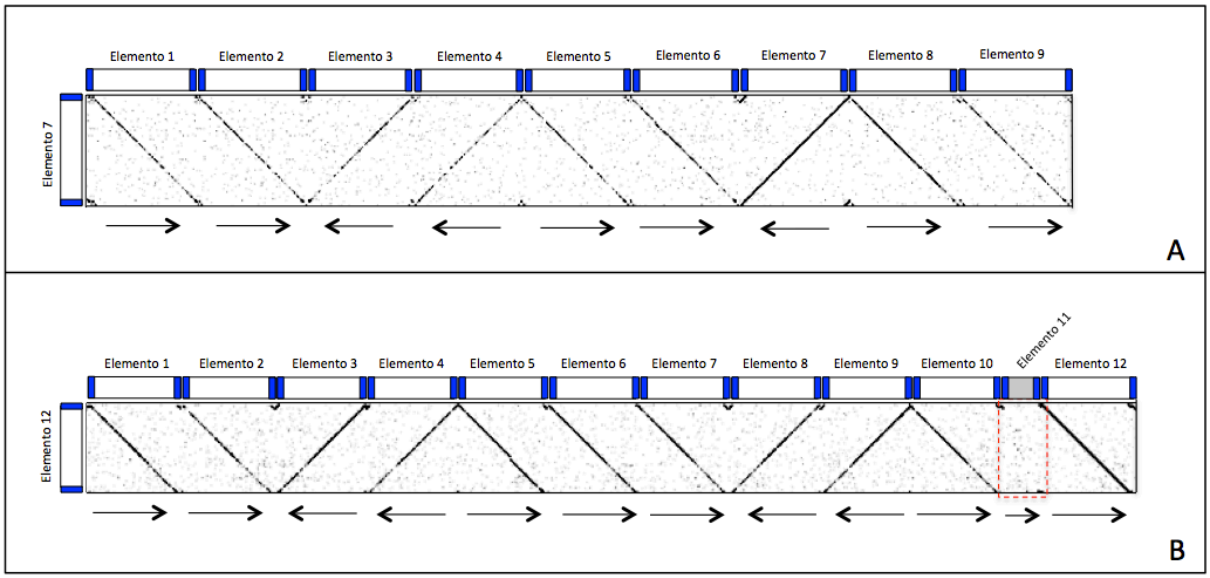


Figura 34: Gráfico de comparação por meio de plotagem utilizando nove sequências de retrotransposons LTR pertencentes a um agrupamento simples contra um desses elementos escolhidos aleatoriamente (elemento 7) (A); Gráfico de comparação por meio de plotagem utilizando 12 sequências de retrotransposons LTR pertencentes a um agrupamento simples contra um desses elementos escolhidos aleatoriamente (elemento 12). O elemento 11 possui apenas uma das porções LTR similares aos outros elementos (B). As setas em ambos os exemplos indicaram a orientação que esses retrotransposons LTR foram inseridos no genoma de *A. ipaënsis*, ou seja, para direita representa a direção 5' – 3' e para esquerda direção inversa. Gráficos de plotagem produzidos pelo *software* Gepard.

3.1.1.2 Agrupamentos complexos

Um agrupamento complexo foi descrito como um conjunto de retrotransposons LTR válidos e similares, ligados a duas ou mais sequências consenso situadas ao centro, formando subgrupos ligados entre si. A figura 35 mostra um exemplo desse tipo de agrupamento, contendo dois subgrupos.

A comparação por gráfico de plotagem das sequências de 11 retrotransposons LTR selecionadas aleatoriamente dos dois subgrupos (eixo x), contra duas sequências também

escolhidas aleatoriamente (eixo y), resultou em 11 linhas diagonais típicas de retrotransposons LTR, sendo 10 similares a um elemento de um subgrupo e um elemento similar ao outro subgrupo (figura 36).

O elemento diferente apresentou apenas as duas regiões LTR similares as dos demais elementos comparados, possivelmente por se tratar de um elemento não-autônomo ou simplesmente por ser diferente, sugerindo que os dois subgrupos representem duas famílias distintas, mas que compartilham similaridade em alguma parte das sequências. Os demais elementos mostram similaridades entre as sequências, porém diferenças no tamanho e variados graus de conservação, sugerindo a ocorrência de mutações, tais como inserções e deleções de fragmentos. Provavelmente, para representar todos esses elementos e suas particularidades, mais de uma sequência consenso foi gerada pela ferramenta CAP3, o que resultou na formação de um agrupamento complexo gerado pelo Biolayout.

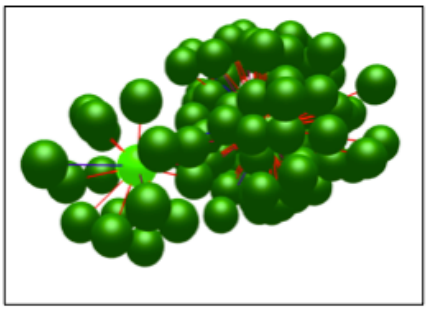


Figura 35: Representação de sequências de retrotransposons LTR organizadas em um agrupamento complexo (com dois subgrupos ligados entre si) composto por duas sequências consenso ligadas por linhas vermelhas e azuis a 88 retrotransposons LTR. Agrupamento produzido pelo *software* Biolayout Express 3-D.

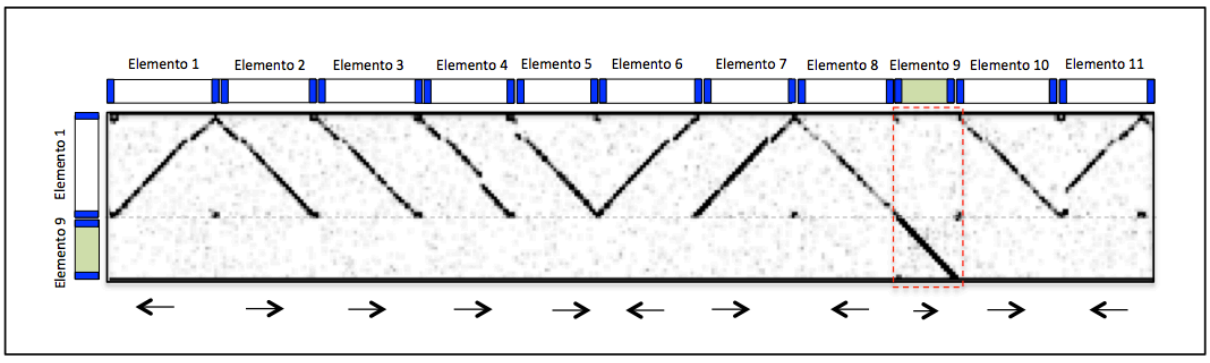


Figura 36: Gráfico de comparação por meio de plotagem utilizando 16 elementos pertencentes a um agrupamento complexo (contendo dois subgrupos) contra dois elementos escolhidos aleatoriamente para cada subgrupo O elemento 9 possui apenas as regiões LTR similares aos demais elementos do outro subgrupo, indicando que podem haver duas famílias distintas de retrotransposons LTR organizadas em um mesmo agrupamento. Gráfico produzido pelo *software* Gepard.

Essa avaliação da similaridade entre sequências de elementos do mesmo agrupamento comparados uns com os outros, permitiu que os elementos agrupados equivocadamente fossem descartados das análises.

Portanto, utilizando as sequências genômicas preliminares do genoma B (800 Mb), 49 prováveis famílias de retrotransposons LTR foram identificadas ao todo, sendo 27 famílias oriundas dos 27 agrupamentos simples e, 22 famílias oriundas da subdivisão de 11 agrupamentos complexos. Um retrotransposon LTR representante de cada provável família foi selecionado. Contudo, a classificação definitiva das famílias e superfamílias só foi determinada a partir da predição da estrutura gênica de todos os elementos presentes em cada família, baseado na análise da sequência genômica completa de *A. ipaënsis* (1,5 Gb).

3.1.2 Identificação de retrotransposons LTR na sequência genômica completa

Utilizando o arquivo contendo o sequenciamento completo do genoma de *A. ipaënsis* (1,5 Gb), 67.832 sequências de putativos retrotransposons LTR foram identificadas pelo *software* LTR_FINDER. Parte dessas sequências putativas tratava-se de sequências em *tandem*. A figura 37 mostra uma comparação por meio de gráfico de plotagem utilizando 500 sequências de putativos retrotransposons (selecionadas aleatoriamente de um total de 67.832) plotadas contra elas mesmas. O resultado da comparação mostou diversos blocos representados por motivos de sequências em *tandem* como exemplificado anteriormente na figura 32-B. Apesar disso, utilizando a técnica de *dot plots*, 40 novos tipos de retrotransposons LTR foram identificados no arquivo genômico completo.

Dessa forma, as 49 sequências de elementos representantes identificadas anteriormente pelo método de agrupamento, juntamente com às 40 sequências identificadas pelo método de plotagem resultaram em 89 diferentes tipos de retrotransposons LTR presentes no genoma de *A. ipaënsis*.

Ao comparar todas as sequências de retrotransposons LTR identificadas no genoma completo de *A. ipaënsis* com os 89 diferentes tipos de retrotransposons LTR, por meio da utilização de BLASTn (*e-value* $1x10^{-80}$), foi possível separar e estabelecer em qual das 89 prováveis famílias essas sequências pertenciam, respectivamente. Com isso, a caraterizacão de parte do conteúdo repetitivo do genoma B de *Arachis* tornou-se exequível.

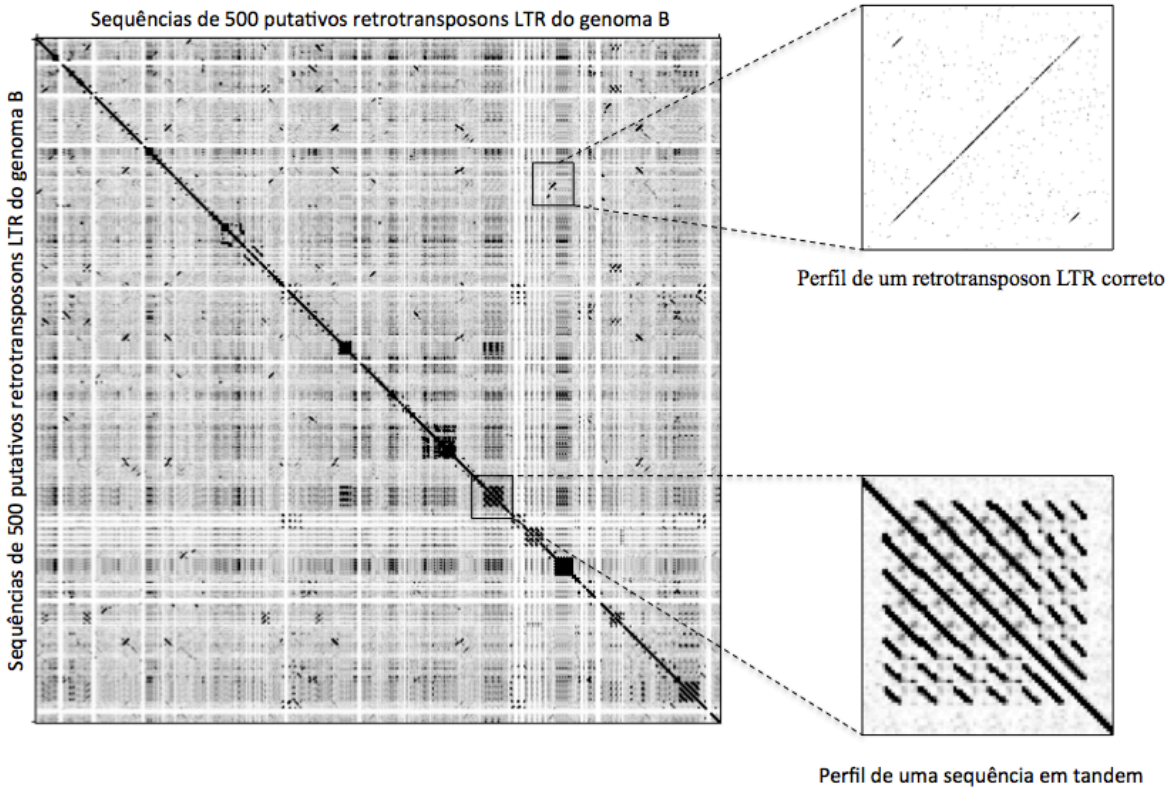


Figura 37: Gráfico de comparação por meio de plotagem utilizando 500 sequências de putativos retrotransposons LTR identificados pelo LTR_FINDER no genoma B (Versão completa do sequenciamento contendo 1,5 Gb) contra elas mesmas. Foram identificadas sequências que correspondem ao perfil correto de retrotransposons LTR (imagem superior ampliada), porém muitos segmentos de sequências em *tandem* também foram identificados (imagem inferior ampliada). Gráfico produzido pelo *software* Gepard.

3.1.3 Caracterização e classificação dos retrotransposons LTR em família e superfamília

Após a investigação do genoma completo de *A. ipaënsis* e pré-classificação das sequências de retrotransposons LTR em famílias distintas, gráficos de plotagem foram construídos para cada provável família, comparando todos os elementos contra o elemento representante, possibilitando observar a similaridade e conservação entre as sequências dos elementos e nomear cada família separadamente. Das 67.832 sequências de putativos retrotransposons identificados no genoma completo de *A. ipaënsis*, 12.274 sequências completas e válidas de retrotransposons LTR foram separadas em 89 famílias distintas. Destas, 24 eram compostas por retrotransposons LTR não-autônomos, e 65 autônomos, das quais 32 eram da superfamília *Ty3-Gypsy* e 33 da *Ty1-Copia.*

Exemplos de comparações realizadas com elementos de cinco famílias (Lima, Doros, Talita, Buba e Dakota) com os seus respectivos elementos representantes (*dot plot*) podem ser vistos na figura 38. O resultado mostrou diagonais conservadas entre os elementos da mesma

família, porém contendo deleções em algumas sequências.

As regiões codantes (CDS) presentes nos elementos e que apresentaram *hit* positivo com sequências de proteínas disponíveis no banco de dados *pfamA* foram utilizadas para classificação em famílias compostas por elementos autônomos ou não-autônomos. A posição do gene que codifica a enzima integrase (IN) foi utilizada para classificação em superfamília. Quando esse gene está posicionado anteriormente ao gene que codifica a enzima transcriptase reversa (TR), o elemento pertence à superfamília *Ty1-Copia*. Quando o gene da integrase (IN) encontra-se posteriormente situado à TR, a superfamília é *Ty3-Gypsy* (figura 39). Na ausência do gene que codifica a TR, os elementos foram classificados como não-autônomos (Wicker *et al*., 2007).

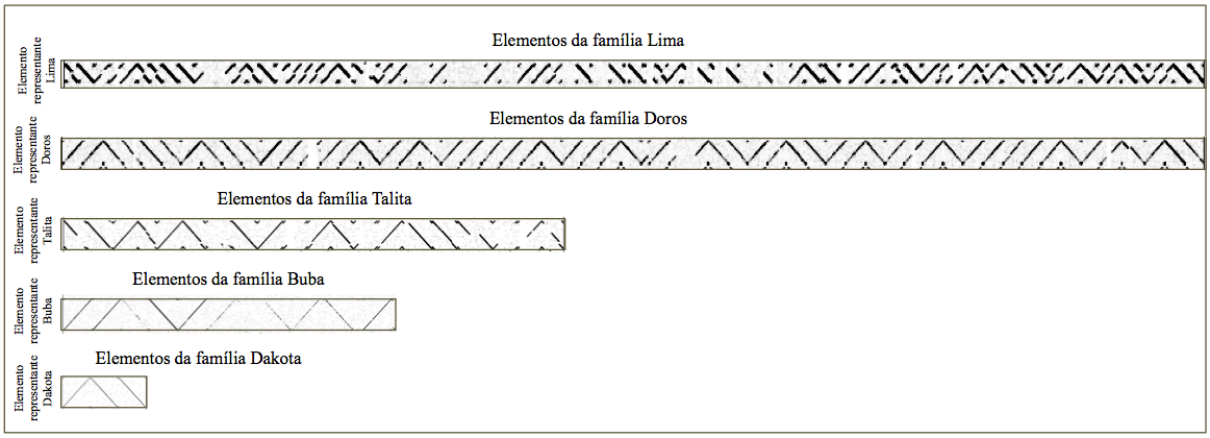


Figura 38: Gráficos de comparação entre as sequências de todos os elementos de cinco famílias diferentes (eixo x) contra as sequências de seus respectivos elementos representantes (eixo y). Gráfico produzido pelo *software* Gepard.

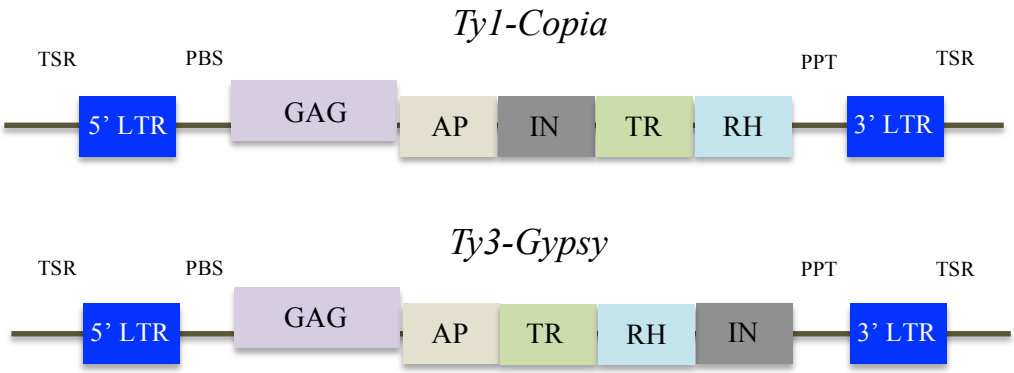


Figura 39: Desenho esquemático de retrotranposons LTR das superfamílias *Ty1-Copia* e *Ty3-Gypsy*. Nota-se há diferença apenas na ordem do gene que codifica a integrase (IN) na cor cinza.

Ao compilar os resultados das plotagens (*software* Gepard) e do conteúdo gênico (*software* Artemis) foi possível constatar diversas peculiaridades presentes nos elementos, tais como a ocorrência de *nested transposons*, que corresponde a ocorrência de um elemento inserido dentro de outro, mutações nas sequências, diferenças entre os tamanhos e direção de inserção. Um exemplo pode ser visto na figura 40, que mostra uma família denominada Juliett composta por seis elementos, os quais foram plotados contra o elemento representante. Os elementos 1, 3, 4 e 6 são bastante similares, ao passo que o elemento 5 possui uma sequência menos similar possivelmente pela ocorrência de mutações. Na plotagem, o elemento 2 apresentou similaridade com os demais apenas nas extremidades e com uma aparente inserção na parte central, no entanto, ao compilar os dados de conteúdo gênico, foi possível inferir que o dobro de CDSs presentes neste elemento poderia ser um indício da ocorrência de um *nested*. E ao comparar esse elemento com os outros 89 tipos, ficou evidente que tratava-se de uma inserção do elemento denominado RE128 (descrito em Bertioli *et al*., 2013 e no Capítulo I da Tese), dentro da sequência do elemento 2 da família Juliett. A ocorrência de *nested transposons* foi bastante recorrente para a maioria das famílias. Foi possível inferir que esta família é composta por seis elementos autônomos de tamanhos similares pertencentes à superfamília *Ty1-Copia*, sendo que um deles possui um elemento completo da família RE128 inserido em sua sequência.

A partir desse tipo de análise realizada para todas as famílias, foi possível recriar graficamente as estruturas de todos os retrotransposons LTR presentes em cada família, assim como avaliar o sentido de transcrição e, consequentemente, a direção de inserção desses elementos no genoma. Foram observados os tamanhos das sequências completas, dos LTRs e a presença de genes.

Essa análise permitiu corroborar o método de classificação aplicado neste estudo. Dados obtidos apenas de uma das análises não seriam suficientes para elucidar a relação entre membros de uma família de retrotransposons. Agregar observações e edições manuais aos resultados obtidos nos *softwares* possibilitou uma classificação e caracterização bastante detalhada acerca dos tipos e peculiaridades presentes em famílias de retrotransposons LTR identificadas em *Arachis*.

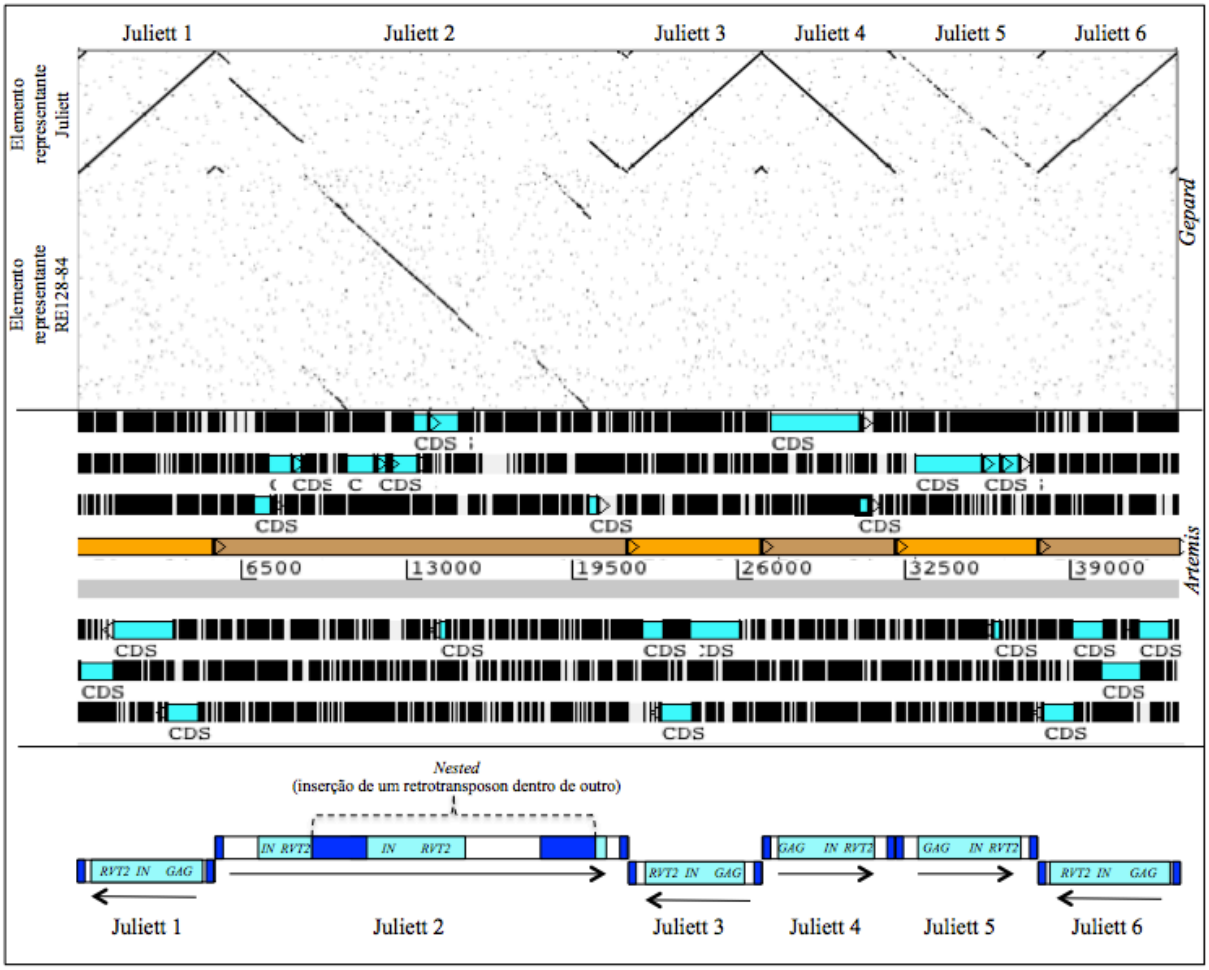


Figura 40: Compilação dos resultados das análises realizadas nos *softwares* Gepard e Artemis, representando os seis retrotransposons presentes na família Juliett (barras em cor marron e laranja). Barras azuis situadas acima dos elementos representaram os genes putativos (CDS) na direção 5'-3', ao passo que barras abaixo, direção inversa. Três elementos estão inseridos no genoma na direção 5'-3' e os outros três na direção inversa. Um dos elementos (Juliett 2), apesar de pertencer à mesma família, mantinha um elemento, pertencente à família RE-128 integrado à sua sequência.

3.2 Componente repetitivo do genoma de *A. duranensis* (genoma A)

Da mesma forma como foi descrita para o genoma B de *A. ipaënsis*, os dados do sequenciamento genômico completo de *A. duranensis* (1,24 Gb), analisados pelo *software* LTR_FINDER, resultaram na identificação de 36.131 sequências de putativos retrotransposons LTR. Grande parte dessas sequências também foi identificada como do tipo *tandem*.

Comparações via BLASTn (*e-value* < $1x10^{-80}$) foram realizadas utilizando os elementos representantes das 89 famílias do genoma de *A. ipaënsis* (genoma B) com as 36.131 sequências de putativos retrotransposons LTR do genoma de *A. duranensis* (genoma A). Foram identificadas 81 famílias similares nos genomas A e B, igualmente nomeadas. A

caracterização, classificação dos retrotransposons LTR, quantificação e seleção de elementos representantes do genoma A foi feita da mesma forma que para o genoma B.

Das 36.131 sequências putativas identificadas no genoma A, 18.839 sequências válidas de retrotransposons LTR completos foram organizadas em 81 famílias distintas. Destas, 23 famílias eram compostas por retrotransposons LTR não-autônomos e 58 por retrotransposons LTR autônomos, sendo 28 da superfamília *Ty3-Gypsy* e 30 da *Ty1-Copia.*

3.3 Caracterização e nomenclatura dos retrotransposons LTR representantes dos genomas A e B

Para o maior conhecimento do conteúdo repetitivo que compõe os genomas de *A. duranensis* e *A. ipaënsis,* e assim facilitar a anotação de genes nas sequências genômicas completas dessas espécies, as sequências de retrotransposons LTR representantes de cada família foram caracterizadas, nomeadas, anotadas e disponibilizadas no GenBank e ENA, bancos de dados de domínio público.

A nomenclatura completa utilizou cinco características: o genoma de origem; os números de acesso dos elementos (para identificar em qual *scaffold* genômico os elementos representantes foram encontrados); o nome da família; modo de transposição (autônoma ou não-autônoma); e superfamília do tipo *Ty1-Copia* ou *Ty3-Gypsy* (figura 41).

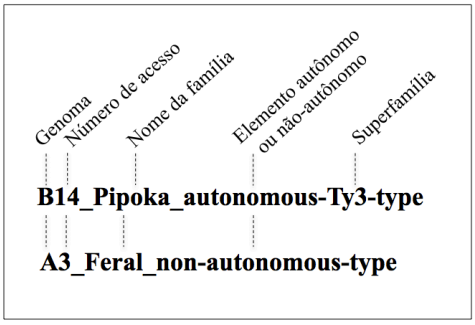


Figura 41: Exemplo da nomenclatura utilizada para os retrotransposons LTR representantes identificados nos genomas de *A. duranensis* e *A. ipaënsis*. Para esses exemplos os números de acesso são B14_825 e A3_23, onde 825 e 23 representam os números dos *scaffolds* em que esses elementos foram originalmente identificados nos genomas B e A, respectivamente. Todos os elementos pertencentes a cada família foram acrescidos de um número consecutivo.

As principais caraterísticas dos retrotransposons LTR representantes dos genomas A e B de *Arachis* estão compiladas nas tabelas 5 e 6, onde os elementos autônomos e não-

autônomos ou os elementos com similaridades nas sequências foram agrupados pela cor da linha (cinza ou branca).

Tabela 5: Lista das principais características dos 81 retrotransposons LTR identificados no genoma de *A. duranensis*.

Genoma A															
Retrotransposon LTR representante	**Superfamília**	**aut/não-aut**	**Tamanho do elemento**	**Tamanho dos LTRs 5' e 3'**	***PBS***	***PPT***	***TSR***	***GAG***	***PROT***	***IN***	***RVT-1***	***RVT-2***	***RnaseH***	***IN***	**Data de transposição (Ma)**
A0_FIDEL_autonomous-Ty3-type	Ty-3	aut	10707	1475/1487	+	+		+			+			+	1,5
A3_Feral_non-autonomous-type	?	non-aut	8838	783				+	+						1
A14_Pipoka_autonomous-Ty3-type	Ty-3	aut	12963	1552/1518	+	+	+	+			+			+	2,5
A104_Pipa1_non-autonomous-type	?	non-aut	7446	1174/1175			+								1,9
A69_Apolo_autonomous-Ty3-type	Ty-3	aut	10560	539	+	+	+	+			+		+	+	1
A141_Polo_non-autonomous-type	?	non-aut	6597	608/609	+	+	+	+							1
A105_Gordo_non-autonomus-type	?	non-aut	9780	1060/1078	+	+		+	+						1,4
A155_Silverio_autonomous-Ty3-type	Ty-3	aut	8708	177	+	+		+			+			+	1,5
A156_Silvia_non-autonomous-type	?	non-aut	7599	1047				+							2
A145_Silver_non-autonomous-type	?	non-aut	4361	1388/1389			+								2,4
A157_Curu_autonomous-Ty3-type	Ty-3	aut	8057	355					+		+			+	1,2
A26_Golf_non-autonomous-type	?	non-aut	5830	1753/1751											4,5
A113_Hemera_autonomous-type	Ty-3	aut	8363	1139/1141	+	+		+	+		+			+	1
A38_Hera_non-autonomous-type	?	non-aut	9573	2083/2074				+							1,3
A5_Eros_autonomous-Ty3-type	Ty-3	aut	10483	1305/1304	+	+	+	+			+			+	3
A89_Eris_non-autonomous-type	?	non-aut	7272	1156/1157				+	+						0
A57_Mico_autonomous-Ty3-type	Ty-3	aut	9431	562/568			+	+			+		+	+	1,9
A24_Athena_non-autonomous-type	?	non-aut	6916	833/794											2,5
A84_RE128_autonomous-Ty1-type	Ty-1	aut	10606	2005/2011	+	+	+			+		+			2,1
A2_Matita_autonomous-Ty1-type	Ty-1	aut	5032	101						+		+	+		0,4
A137_Sofi_autonomous-Ty3-type	Ty-3	aut	11968	583/568	+	+		+	+		+		+	+	2,1
A99_Medusa_autonomous-Ty3-type	Ty-3	aut	6877	466/456	+	+		+	+		+			+	4,3
A33_Musa_non-autonomous-type	?	non-aut	3496	558											2,7
A16_Grilo_autonomous-Ty3-type	Ty-3	aut	9252	438/439	+	+	+	+	+		+		+	+	4,2
A142_Gilo_non-autonomous-type	?	non-aut	4687	422			+								0
A29_RE128_autonomous-Ty1-type	Ty-1	aut	10003	1264/1268			+			+		+			1
A121_Doros_autonomous-Ty3-type	Ty-3	aut	9095	360/361	+	+	+	+			+		+	+	1,5
A122_Duran_non-autonomous-type	?	non-aut	5012	381			+	+	+						1
A150_Dan_non-autonomous-type	?	non-aut	2992	372	+	+	+	+							1
A1_Hermes_autonomous-Ty3-type	Ty-3	aut	8977	453/454	+	+	+	+			+		+	+	1
A118_Bela_autonomous-Ty1-type	Ty-1	aut	5689	109	+	+				+		+			0
A132_Zoe_autonomous-Ty1-type	Ty-1	aut	5203	217	+	+	+			+		+			0
A138_Zia_autonomous-Ty1-type	Ty-1	aut	2396	213	+	+	+					+			2,1
A9_Girino_autonomous-Ty3-type	Ty-3	aut	9381	525	+	+	+	+	+		+			+	3,2
A153_Kyra_non-autonomous-type	?	non-aut	6826	277/275											1,5
A39_Yara_autonomous-Ty1-type	Ty-1	aut	5739	833/832	+	+	+		+	+		+			1
A62_Saturno_autonomous-Ty3-type	Ty-3	aut	9373	517			+	+			+		+	+	1,1
A124_Ariel_autonomous-Ty3-type	Ty-3	aut	5457	170/179	+	+					+				1,8
A19_Netuno_autonomous-Ty3-type	Ty-3	aut	8701	433			+	+			+		+	+	0
A136_Elpis_non-autonomous-type	?	non-aut	3821	460/458											1,1
A126_Venon_autonomous-Ty1-type	Ty-1	aut	4473	243/240			+			+		+			1
A125_Buba_autonomous-Ty3-type	Ty-3	aut	8799	392	+	+	+	+			+		+	+	1
A60_Venus_autonomous-Ty3-type	Ty-3	aut	8517	383			+	+			+		+	+	0
A135_Maia_non-autonomous-type	?	non-aut	9138	691/693	+	+		+					+	+	2
A112_Lima_autonomous-Ty3-type	Ty-3	aut	5680	551/541	+	+	+	+	+		+			+	3,1
A143_Yuri_autonomous-Ty1-type	Ty-1	aut	5408	681/684			+			+		+			1,9
A140_Joka_autonomous-Ty1-type	Ty-1	aut	4154	313/276	+	+	+			+		+			2,5
A54_Yankee_autonomous-Ty1-type	Ty-1	aut	5176	417/402	+	+	+			+		+			2,5
A134_Nemesis_non-autonomous-type	?	non-aut	7439	1316/1347	+	+			+	+					1
A65_Foxtrot_autonomous-Ty1-type	Ty-1	aut	4082	569			+			+		+			0
A11_Delta_autonomous-Ty3-type	Ty-3	aut	6750	864/865	+			+			+			+	1,4
A11_Charlie_non-autonomous-type	?	non-aut	4145	818/814	+	+	+	+							1
A123_Talita_autonomous-Ty1-type	Ty-1	aut	4879	428/429	+	+	+			+		+			1
A67_India_autonomous-Ty1-type	Ty-1	aut	4844	229/226			+			+		+			1,5
A31_Kilo_autonomous-Ty1-type	Ty-1	aut	5284	252/256			+	+		+		+	+		1
A50_Hotel_autonomous-Ty1-type	Ty-1	aut	6152	273			+			+		+	+		2,1
A119_Bola_autonomous-Ty1-type	Ty-1	aut	5050	322	+	+				+		+			1,4
A152_Agnus_non-autonomous-type	?	non-aut	1718	170			+								0,4
A88_Papa_autonomous-Ty1-type	Ty-1	aut	4756	291/290			+	+	+	+		+			1
A32_Tango_autonomous-Ty1-type	Ty-1	aut	4797	188				+		+		+			1,6
A149_Lulu_non-autonomous-type	?	non-aut	2230	421			+								1
A43_Alpha_autonomous-Ty1-type	Ty-1	aut	5287	218			+	+		+		+			0,9
A130_James_autonomous-Ty1-type	Ty-1	aut	4607	185/184						+		+			1,2
A144_Diva_autonomous-Ty1-type	Ty-1	aut	4841	310			+	+	+	+		+	+		0
A71_Juliett_autonomous-Ty1-type	Ty-1	aut	5370	290			+	+		+		+			1
A97_Mike_autonomous-Ty3-type	Ty-3	aut	5434	434/435				+	+		+			+	1
A127_Dakota_autonomous-Ty1-type	Ty-1	aut	4571	253			+	+		+		+			1
A131_Phenix_autonomous-Ty3-type	Ty-3	aut	5371	417/418			+	+	+		+			+	0
A78_Echos_autonomous-Ty1-type	Ty-1	aut	5619	313			+			+		+			2,4
A61_Oscar_autonomous-Ty1-type	Ty-1	aut	5357	294/292				+		+		+			0,8
A17_Romeo_autonomous-Ty1-type	Ty-1	aut	4579	228						+		+			2,5
A51_Victor_autonomous-Ty3-type	Ty-3	aut	5642	566/567			+	+	+		+		+	+	0
A151_Vipe_non-autonomous-type	?	non-aut	4647	327/328	+	+									2
A139_George_autonomous-Ty3-type	Ty-3	aut	5602	479/519	+	+	+	+	+		+			+	1,6
A146_Golden_autonomous-Ty1-type	Ty-1	aut	4769	375	+	+	+			+		+			0
A129_Jasper_autonomous-Ty1-type	Ty-1	aut	4506	143	+	+				+		+			1,6
A148_Edore_autonomous-Ty3-type	Ty-3	aut	5010	304/291	+	+	+	+	+		+			+	1,4
A147_Omega_autonomous-Ty3-type	Ty-3	aut	5200	347	+	+	+	+	+		+			+	0
A128_Doris_autonomous-Ty1-type	Ty-1	aut	2833	317/319				+		+					1,4
A117_Xray_non-autonomous-type	?	non-aut	2617	378/372			+	+							2,3
A91_Whiskey_autonomous-Ty3-type	Ty-3	aut	2816	342/343			+							+	3,8

Tabela 6: Lista das principais características dos 89 retrotransposons LTR identificados no genoma de *A. ipaënsis*.

Genoma B															
Retrotransposon LTR representante	**Superfamília**	**aut/não-aut**	**Tamanho do elemento**	**Tamanho dos LTRs 5' e 3'**	***PBS***	***PPT***	***TSR***	***GAG***	***PROT***	***IN***	***RVT-1***	***RVT-2***	***RnaseH***	***IN***	**Data de transposição (Ma)**
B0_FIDEL_autonomous-Ty3-type	Ty-3	aut	12407	1512/1510			+	+			+			+	2,8
B3_Feral_non-autonomous-type	?	non-aut	9990	291				+	+						1,7
B14_Pipoka_autonomous-Ty3-type	Ty-3	aut	12512	412/413	+	+		+			+			+	2,3
B104_Pipa1_non-autonomous-type	?	non-aut	9483	1212/1219											1,9
B49_Pipa2_non-autonomous-type	?	non-aut	5974	150	+	+									1,2
B8_Pipa3_non-autonomous-type	?	non-aut	5463	204											2,2
B69_Apolo_autonomous-Ty3-type	Ty-3	aut	10156	310/309				+			+		+	+	1,6
B141_Polo_non-autonomous-type	?	non-aut	6465	573	+	+	+	+		+					1
B105_Gordo_non-autonomus-type	?	non-aut	11536	1248/1235				+							3,2
B155_Silverio_autonomous-Ty3-type	Ty-3	aut	8436	676/677				+			+			+	1,1
A156_Silvia_non-autonomous-type	?	non-aut	8407	1441/1464			+	+							1,4
B145_Silver_non-autonomous-type	?	non-aut	6182	1724/1675			+								2,4
A157_Curu_autonomous-Ty3-type	Ty-3	aut	8988	1006				+	+		+			+	0,9
B6_Bravo_non-autonomous-type	?	non-aut	7273	472				+							1,2
B26_Golf_non-autonomous-type	?	non-aut	5790	1616/1617	+	+									2,6
B113_Hemera_autonomous-type	Ty-3	aut	5940	659					+		+			+	3
B38_Hera_non-autonomous-type	?	non-aut	8982	1858/1848											2,1
B5_Eros_autonomous-Ty3-type	Ty-3	aut	9996	505				+			+			+	2,2
B89_Eris_non-autonomous-type	?	non-aut	8369	1581/1572	+	+		+							3
B57_Mico_autonomous-Ty3-type	Ty-3	aut	9360	297	+	+		+			+		+	+	1,4
B24_Athena_non-autonomous-type	?	non-aut	8064	634/636											1,8
B84_RE128_autonomous-Ty1-type	Ty-1	aut	11231	1707/1704						+		+			2,4
B2_Matita_autonomous-Ty1-type	Ty-1	aut	5473	121/123						+		+	+		2,5
B137_Sofi_autonomous-Ty3-type	Ty-3	aut	12611	1083	+	+		+	+		+		+	+	1,8
B99_Medusa_autonomous-Ty3-type	Ty-3	aut	8611	1130/1118	+	+		+	+		+			+	2,7
B33_Musa_non-autonomous-type	?	non-aut	5009	247/252											2
B16_Grilo_autonomous-Ty3-type	Ty-3	aut	8699	430				+	+		+		+	+	1,7
B142_Gilo_non-autonomous-type	?	non-aut	6958	404/403	+	+	+								1
B29_RE128_autonomous-Ty1-type	Ty-1	aut	8660	812/810	+	+	+			+		+			1,2
B121_Doros_autonomous-Ty3-type	Ty-3	aut	9203	383				+	+		+		+	+	2,1
B122_Duran_non-autonomous-type	Ty-3	aut	5551	427/426	+	+	+	+	+						1,3
B150_Dan_non-autonomous-type	?	non-aut	4209	382	+	+	+	+	+						1,8
B1_Hermes_autonomous-Ty3-type	Ty-3	aut	9004	448/447	+	+	+	+			+		+	+	1,9
B118_Bela_autonomous-Ty1-type	Ty-1	aut	6214	561/566	+	+	+			+		+			1
B132_Zoe_autonomous-Ty1-type	Ty-1	aut	5185	209/211			+			+		+			2,5
B138_Zia_autonomous-Ty1-type	Ty-1	aut	3160	245/246	+	+	+			+		+			1
B9_Girino_autonomous-Ty3-type	Ty-3	aut	9345	512/513				+	+		+		+	+	1,5
B39_Yara_autonomous-Ty1-type	Ty-1	aut	5649	683/687	+	+	+			+		+			2,2
B62_Saturno_autonomous-Ty3-type	Ty-3	aut	9058	434			+	+			+		+	+	1,7
B124_Ariel_autonomous-Ty3-type	Ty-3	aut	6143	378/375							+				1,8
B19_Netuno_autonomous-Ty3-type	Ty-3	aut	8929	490			+	+			+		+	+	0
B136_Elpis_non-autonomous-type	?	non-aut	3972	459/458			+	+							1,6
B126_Venon_autonomous-Ty1-type	Ty-1	aut	4464	242			+			+		+			1
B125_Buba_autonomous-Ty3-type	Ty-3	aut	8777	200	+	+		+			+		+	+	2,7
B60_Venus_autonomous-Ty3-type	Ty-3	aut	8661	378			+	+			+		+	+	1,3
B135_Maia_non-autonomous-type	?	non-aut	5763	691/692	+	+	+	+	+						0
B112_Lima_autonomous-Ty3-type	Ty-3	aut	4089	231/236	+	+		+	+					+	4,1
B143_Yuri_autonomous-Ty1-type	Ty-1	aut	5424	682/681			+			+		+	+		1,3
B140_Joka_autonomous-Ty1-type	Ty-1	aut	3856	37			+	+	+	+		+			0
B54_Yankee_autonomous-Ty1-type	Ty-1	aut	5279	422/425	+	+				+		+			2,7
B134_Nemesis_non-autonomous-type	?	non-aut	4775	262					+	+					1
B65_Foxtrot_autonomous-Ty1-type	Ty-1	aut	5274	573/598	+	+	+			+		+			1,2
B11_Delta_autonomous-Ty3-type	Ty-3	aut	8018	872/853				+			+			+	1,8
B11_Charlie_non-autonomous-type	?	non-aut	2482	197				+							1
B123_Talita_autonomous-Ty1-type	Ty-1	aut	4821	228/229	+	+				+		+			2,2
B67_India_autonomous-Ty1-type	Ty-1	aut	4687	188/193				+		+		+			2,9
B31_Kilo_autonomous-Ty1-type	Ty-1	aut	5350	296			+	+		+		+			1
B50_Hotel_autonomous-Ty1-type	Ty-1	aut	5531	273/272	+	+	+	+		+		+			1,7
B119_Bola_autonomous-Ty1-type	Ty-1	aut	3365	181/189				+		+		+			3,6
B88_Papa_autonomous-Ty1-type	Ty-1	aut	5332	382/371				+		+		+			2,5
B32_Tango_autonomous-Ty1-type	Ty-1	aut	4872	243				+		+		+			1,2
B149_Lulu_non-autonomous-type	?	non-aut	2349	471			+								1
B43_Alpha_autonomous-Ty1-type	Ty-1	aut	5374	305/302				+		+		+			1,9
B130_James_autonomous-Ty1-type	Ty-1	aut	4651	190/206			+			+		+			0
B144_Diva_autonomous-Ty1-type	Ty-1	aut	4917	314	+	+	+	+	+	+		+	+		1
B71_Juliett_autonomous-Ty1-type	Ty-1	aut	5292	150/146				+		+		+			3,7
B97_Mike_autonomous-Ty3-type	Ty-3	aut	5283	388	+	+		+	+		+			+	2,5
B127_Dakota_autonomous-Ty1-type	Ty-1	aut	4572	247/245			+	+		+		+			2,2
B131_Phenix_autonomous-Ty3-type	Ty-3	aut	5379	418	+	+	+	+	+		+			+	1
B78_Echos_autonomous-Ty1-type	Ty-1	aut	5746	343/340			+	+		+		+			1,2
B61_Oscar_autonomous-Ty1-type	Ty-1	aut	5356	289/335			+	+		+		+			1,6
B17_Romeo_autonomous-Ty1-type	Ty-1	aut	4608	228				+		+		+			3,7
B51_Victor_autonomous-Ty3-type	Ty-3	aut	5670	584/575				+			+			+	2
B139_George_autonomous-Ty3-type	Ty-3	aut	2088	402			+	+							1
B146_Golden_autonomous-Ty1-type	Ty-1	aut	4775	377/379	+	+	+			+		+			0
B129_Jasper_autonomous-Ty1-type	Ty-1	aut	5553	251	+	+	+			+		+			1,3
B148_Edore_autonomous-Ty3-type	Ty-3	aut	5072	326/328			+	+	+		+			+	3,3
B147_Omega_autonomous-Ty3-type	Ty-3	aut	5437	391	+	+	+	+	+		+			+	0
B128_Doris_autonomous-Ty1-type	Ty-1	aut	5355	319/318				+		+		+			2
B117_Xray_non-autonomous-type	?	non-aut	2601	326/323	+	+		+							3,2
B91_Whiskey_autonomous-Ty3-type	Ty-3	aut	5136	320				+			+			+	2,8
B154_Paco_autonomous-Ty3-type	Ty3	aut	10449	1093/1106	+	+					+				2,1
B133_Kirke_autonomous-Ty1-type	Ty-1	aut	5651	301/296	+	+	+	+		+		+			1,9
B25_Sierra_autonomous-Ty1-type	Ty-1	aut	4667	164				+		+		+			1,8
B81_Quebec_autonomous-Ty3-type	Ty-3	aut	5371	447/488	+	+	+	+	+		+			+	1,5
B7_November_autonomous-Ty1-type	Ty-1	aut	5097	235/239						+		+			1,3
B45_Zulu_non-autonomous-type	?	non-aut	12055	567/565								+			1,6
B120_Grey_non-autonomous-type	?	non-aut	2091	485/496											1,9
B47_Uniform_autonomous-Ty3-type	Ty-3	aut	5265	401/433	+	+		+			+			+	4,3

Para o genoma de *A. duranensis* (genoma A) foram caracterizados 81 elementos representantes, dos quais 28 pertenciam à superfamília *Ty3-Gypsy,* contendo em média 7.970 pb, 23 elementos não-autônomos com aproximadamente 5.781 pb e os outros 30 elementos eram da superfamília *Ty1-Copia* com 5.202 pb em média. Dos 81 elementos representantes, 50 apresentaram a sequência flanqueadora de inserção denominada TSD. A média da data de transposição de todos os elementos representantes foi de 1,44 Ma (milhões de anos atrás). Já no genoma de *A. ipaënsis* (genoma B) foram caracterizados 89 elementos representantes, dos quais 32 pertenciam à superfamília *Ty3-Gypsy* com 7.801 pb em média, 24 elementos não-autônomos com aproximadamente 6.182 pb e os outros 31 elementos pertenciam à superfamília *Ty1-Copia* com média de 5.316 pb. Dos 81 elementos representantes, somente 38 apresentaram a estrutura flanqueadora denominada TSD. A média da data de transposição de todos os elementos representantes foi de aproximadamente 1,81 Ma.

De forma geral, o tamanho das sequências dos retrotransposons LTR do tipo *Ty3-Gypsy* foi maior do que os *Ty1-Copia* e não-autônomos. Os menores elementos com menos de 3.000 pb no genoma A foram Lulu, Zia, Xray, Whiskey e Dan. Já os maiores elementos, com mais de 10.000 pb foram RE128 (os dois tipos), Eros, Apolo, FIDEL, Sofi e Pipoka. No genoma B, os menores elementos foram Grey, Lulu, Charlie e Xray, enquanto os maiores foram Apolo, Paco, RE128, Gordo, FIDEL, Pipoka e Sofi.

Outro ponto interessante foi a identificação do gene da transcriptase reversa (TR) nos dois genomas. Todas as TRs presentes nos elementos da superfamília *Ty3-Gypsy*, quando comparados com sequências disponíveis no banco de dados do *pfamA* apresentaram similaridade apenas com proteínas de uma família denominada RVT-1 ou *Reverse Transcriptase – 1 Family* (PF00078 - http://pfam.sanger.ac.uk/family/PF00078). Essa família é bem caracterizada na literatura, e, no banco de dados do *pfamA* já foram submetidas 172.360 sequências, distribuídas em 4.989 espécies (principalmente vírus, eucariotos e poucas bactérias), além de possuir 405 estruturas proteicas descritas. Por outro lado, para as TRs dos elementos da superfamília *Ty1-Copia,* houve similaridade apenas com proteínas da família denominada RVT-2 (PF07728.8 - http://pfam.sanger.ac.uk/family/PF07728.8), que ainda não está bem caracterizada e não dispõe de estrutura proteica descrita, apesar de estarem disponíveis 7.027 sequências derivadas dessa família, distribuídas entre 350 espécies eucarióticas.

Ainda para avaliar a similaridade entre as sequências de elementos representantes da mesma família, mas em genomas diferentes, foi realizada uma plotagem entre as sequências

de todos os 81 elementos do genoma de *A. duranensis* (total de 485.341 pb) e 89 do genoma de *A. ipaënsis* (579.265 pb), organizadas na mesma ordem e direção 5'-3'. O resultado indicou alta similaridade entre as sequências de cada representante em ambos os genomas, representado pela diagonal bastante conservada (figura 42). Alguns dos elementos estavam representados em apenas um dos genomas o que ocasionou espaços ou *gaps* indicados por setas. Isso é normalmente consequência do descarte de parte das sequências durante as montagem de um genoma completo. Apenas fragmentos derivados desses elementos foram encontrados nos genomas.

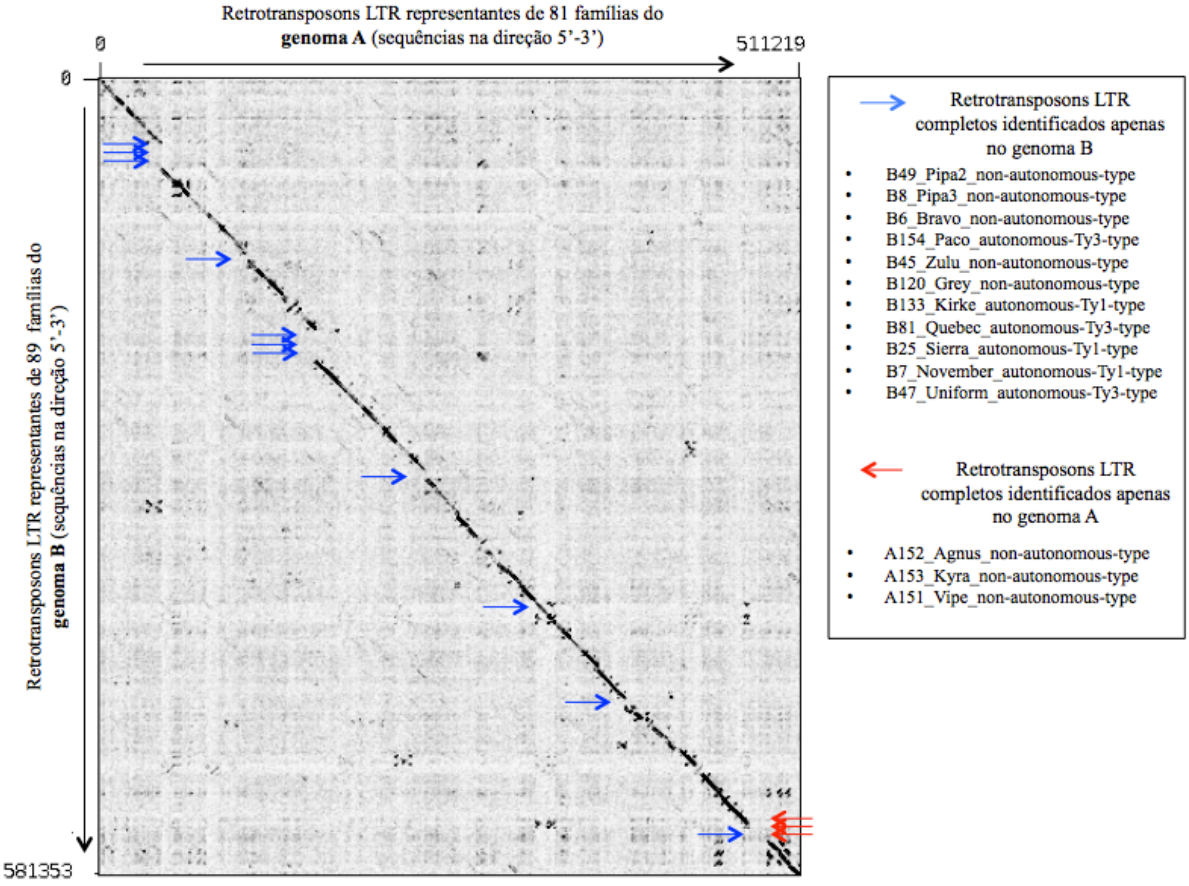


Figura 42: Gráfico de plotagem comparando 81 elementos representantes do genoma A (eixo x) com 89 elementos representantes do genoma B (eixo y) ordenados pela nomenclatura. A linha diagonal preta indica a similaridade entre os elementos de famílias similares em A e B. A descontinuidade da linha indica presença de elementos identificados em apenas um dos genomas. As setas para a direita representaram os 11 elementos encontrados apenas no genoma B e as setas para a esquerda representaram os 3 elementos encontrados apenas no genoma A. Os nomes dos elementos estão listados na caixa de texto situada à direita. Gráfico produzido pelo *software* Gepard.

3.3.1 Retrotransposons LTR autônomos e não-autônomos

Onze famílias compostas por retrotransposons LTR autônomos apresentaram prováveis pares de famílias de retrotransposons LTR não-autônomos em ambos os genomas (tabelas 5 e 6). Comparações por meio de plotagens entre as sequências dos elementos autônomos Apolo e Doros e seus respectivos pares não-autônomos Polo e Duran, nos genomas A e B, exemplificam a similaridade em parte dessas sequências (figuras 43 e 44).

Em todos os casos, as sequências dos elementos autônomos eram maiores do que as dos não-autônomos. A similaridade entre os pares (autônomos e não-autônomos) ocorreu principalmente nas regiões flanqueadoras ou LTRs, em acordo com o descrito para pares FIDEL/Feral e Pipoka/Pipa (Bertioli *et al*., 2013). Nesses dois exemplos houve similaridade parcial nas regiões gênicas, restritas aos genes *gag*, protease (prot), integrase (IN ou rve) e parte da região 3' UTR (não-traduzida), indicando que, possivelmente durante a evolução, partes das sequências desses elementos autônomos foram deletadas por meio de mutações (como no gene da transcriptase reversa).

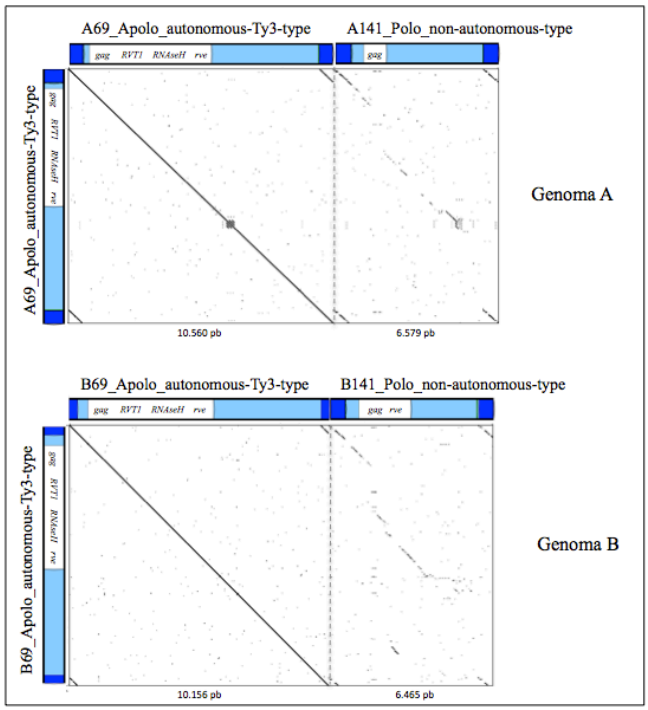


Figura 43: Gráficos de plotagem entre o elemento autônomo Apolo e seu par não-autônomo Polo identificados nos genomas A e B de *A. duranensis* e *A. ipaënsis*, respectivamente. No eixo x foram plotados os elementos autônomos e não-autônomos; no eixo y foram plotados apenas os elementos autônomos. As porções flanquadores ou LTR estão representadas pela cor azul escura; o conteúdo gênico em branco; e regiões 5' e 3' não-traduzidas em azul claro. Gráfico produzido pelo *software* Gepard.

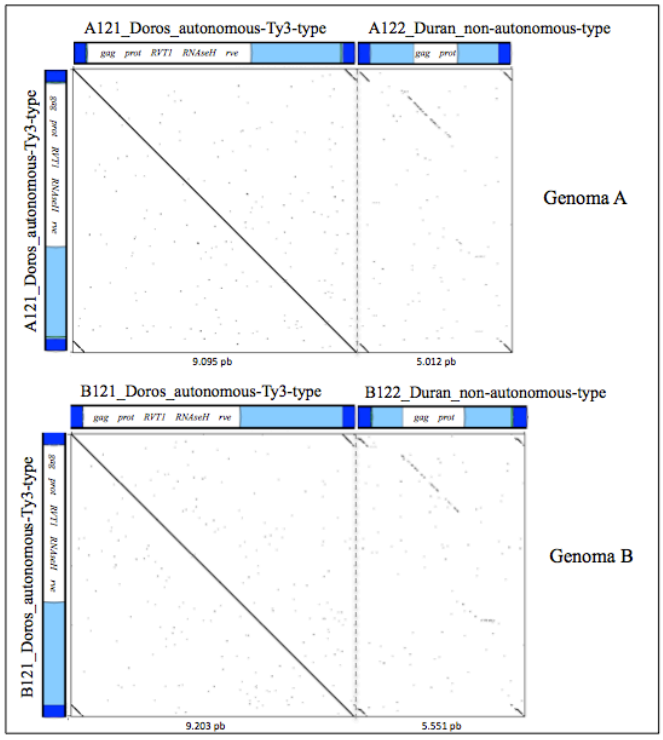


Figura 44: Gráficos de plotagem entre o elemento autônomos Doros e seu par não-autônomo Duran identificados nos genomas A e B de *A. duranensis* e *A. ipaënsis*, respectivamente. No eixo x foram plotados os elementos autônomos e não-autônomos; no eixo y foram plotados apenas os elementos autônomos. As porções flanquadores ou LTR estão representadas pela cor azul escura; o conteúdo gênico em branco; e regiões 5' e 3' não-traduzidas em azul claro. Gráfico produzido pelo *software* Gepard.

3.4 Anotação dos retrotransposons LTR representantes

Para a anotação de cada retrotransposon LTR representante foram utilizadas as seguintes informações: o tamanho da sequência completa do elemento, tamanho dos LTRs e presença e localização dos motivos TSD, PBS e PPT. O conteúdo gênico predito (Artemis e *pfamA*) foi anotado utilizando o termo "pseudogene" e a estimativa da data de inserção dos elementos nos genomas também foi incluída. O exemplo da anotação do retrotransposon LTR representante da família Apolo (genoma A) e o resultado da submissão dessa sequência no GenBank (NCBI) pode ser visualizado na figura 45. Os números de acesso de todos os elementos no ENA e GenBanK são: KC608770.1 - KC608818.1; KF729698.1 - KF729732.1 e GI: 472455659 - GI: 472455707; GI: 572921102 - GI: 572921136 para os elementos do genoma B de *A. ipaënsis* e para o genoma A de *A. duranensis* são: KF729733.1 - KF729794.1; GI: 572921137 - GI: 572921198.

As anotações foram graficamente visualizadas no *software* Artemis permitindo comparar o tamanho de todos os elementos, LTRs e regiões gênicas, em virtude da mesma escala em pares de bases utilizada (figura 46).

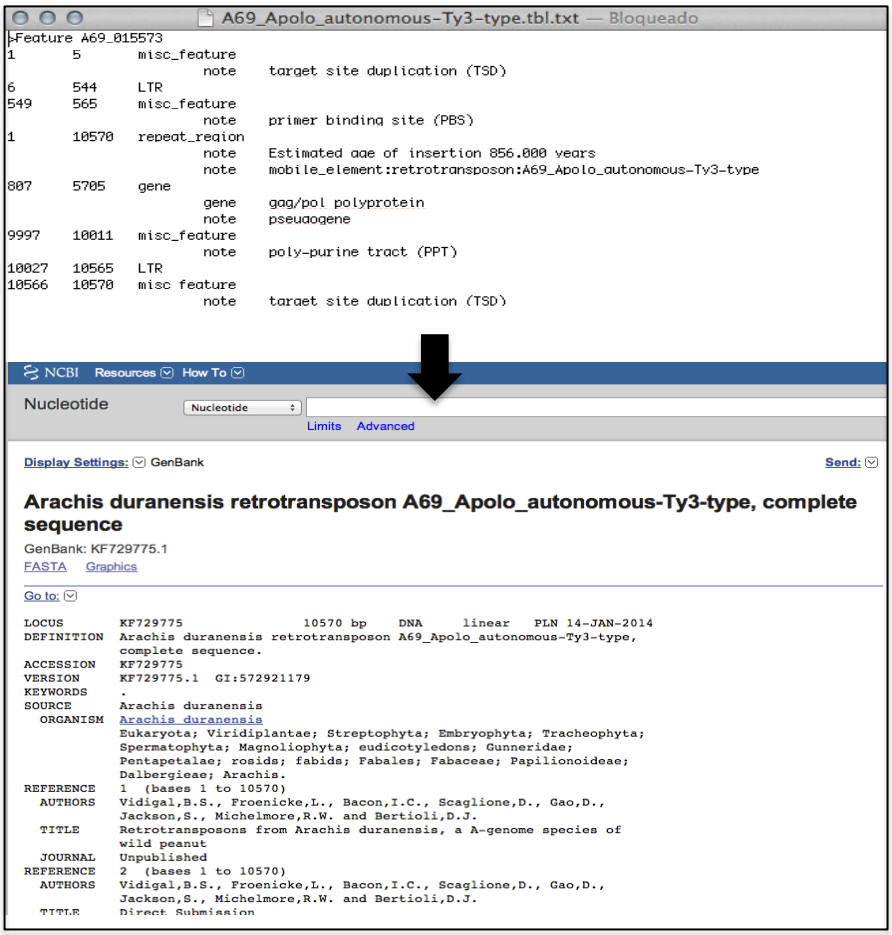


Figura 45: Anotação do elemento A69_Apolo_autonomous-Ty3-type e resultado da submissão dessa sequência no GenBank.

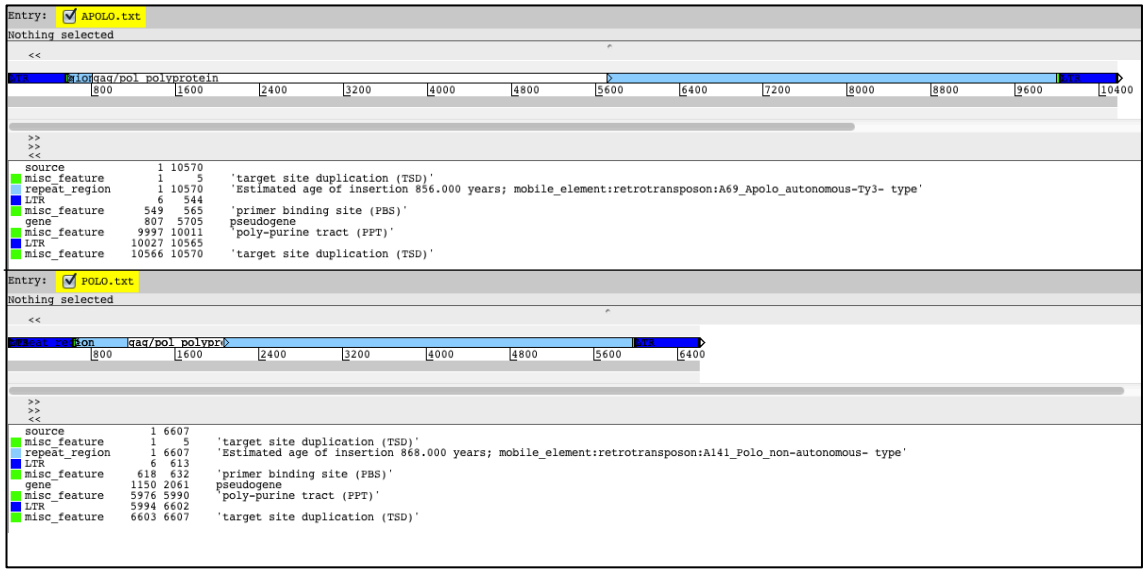


Figura 46: Anotação do elemento A69_Apolo_autonomous-Ty3-type e seu par não autônomo A141_Polo_non-autonomous-type visualizada na interface do Artemis.

3.5 Frequência de retrotransposons LTR nos genomas A e B

De acordo com análises utilizando o *software* RepeatMasker, 28,5% da sequência genômica completa de *A. duranensis* (genoma A) é composta por 81 famílias de retrotransposons LTR, ou seja, aproximadamente 353 Mb estão representados por retrotransposons LTR conhecidos e suas respectivas sequências remanescentes, bem como por LTRs-solo. Já no genoma de *A. ipaënsis* (genoma B), 89 famílias de retrotransposons juntamente com seus fragmentos e LTRs-solo estão presentes em 27,6% ou em 414 Mb do genoma completo. Essa estimativa da porcentagem pode ser maior, tendo em vista os parâmetros utilizados nas análises e a opção do descarte de *scaffolds* genômicos menores que 2.000 pb. Dificilmente esses *scaffolds* descartados abrigariam retrotransposons LTR com sequências completas, no entanto poderiam abrigar fragmentos e sequências remanescentes. Certamente outras famílias de retrotransposons LTR, não-LTR, transposons de DNA e DNA repetitivo de outras classes ainda não identificados também estão presentes nas sequências genômicas de *A. duranensis* e *A. ipaënsis*, o que tornaria a estimativa de porcentagem evidentemente maior.

Os retrotransposons LTR pertencentes à superfamília *Ty3-Gypsy,* juntamente com seus pares não-autônomos perfazem a maior parte dos genomas de *A. duranensis* e *A. ipaënsis*. Em segundo lugar, estão as famílias de elementos não-autônomos que não apresentaram pares autônomos e por último, os elementos *Ty1-Copia* e seus pares não-autônomos (figura 47).

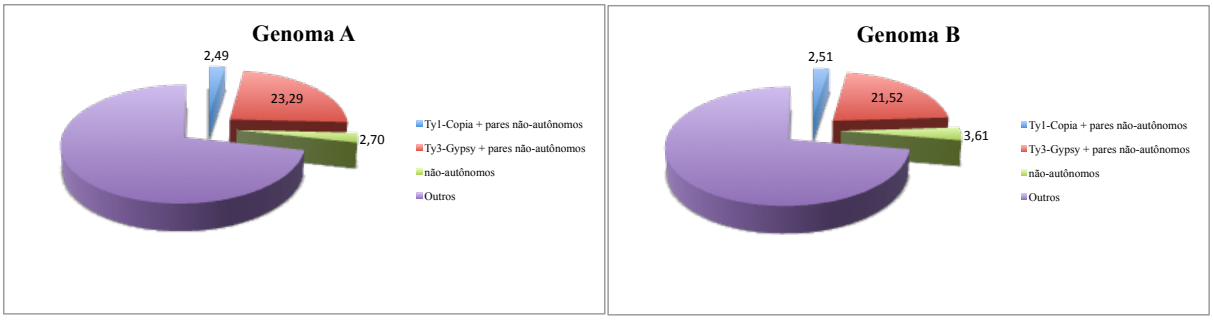


Figura 47: Gráficos mostrando as frequências (em porcentagens) estimadas para as superfamílias *Ty1-Copia*, *Ty3-Gypsy* e famílias não-autônomas nos genomas A e B.

A cobertura em pb relativa aos elementos completos e seus fragmentos; a frequência genômica em porcentagem; e o número de elementos com sequências completas estão discriminados para cada família individualmente, famílias relacionadas ou pares autônomos/não-autônomas dos genomas A e B (tabela 7).

Nos genomas completos de *A. duranensis* e *A. ipaënsis* apenas 37 famílias representam a maior parte da ocupação por retrotransposons LTR, consistindo quase toda a porcentagem estimada. São elas: FIDEL/Feral, Pipoka/Pipa1/Pipa2/Pipa3, Apolo/Polo, Gordo, Siverio/Silvia/Silver; Curu/Bravo/Golf/Hemera/Hera, Eros/Eris, Mico, Athena, RE128-84, Matita, Sofi, Medusa/Musa, Grilo/Gilo, RE128-29, Doros/Duran/Dan, Hermes, Bela, Zoe/Zia e Girino. Juntas, essas famílias representam 26,61% e 25,20% dos genomas A e B, respectivamente. Em termos absolutos, o número de elementos com sequência completa nessas famílias explica também quase todo o montante de sequências de retrotransposons LTR identificados, sendo 18.244 no genoma A e 11.588 no genoma B, de um total de 18.839 e 12.274, respectivamente.

O retrotransposon LTR FIDEL, juntamente com seu par não-autônomo Feral são os elementos mais abundantes nos genomas A e B. No entanto, estes são mais frequentes no genoma A (11,02%) do que no B (7,85%).

A porcentagem e o número dos elementos identificados nas 37 famílias mais frequentes nos genomas A e B estão representados nas figuras 48 e 49. Os dois gráficos exibiram curvas semelhantes, apesar desse padrão não ser esperado, principalmente em virtude das diferentes características evidenciadas em cada um. Para o primeiro, foram levados em conta a porcentagem relativa aos genomas de acordo com o tamanho em pb dos elementos completos e seus fragmentos, ao passo que o outro gráfico exibiu apenas o número de elementos com sequência completa. Algumas famílias estavam mais frequentes no genoma A, como por exemplo FIDEL/Feral, Pipoka/Pipas, Apolo/Polo, Curu/Bravo/Golf/Hemera/Hera e Sofi. Outras, como Gordo, Silverio/Silvia/Siver, Eros/Eris, Athena, Medusa/Musa, Doros/Duran, Hermes, Bela e Girino estavam mais frequentes no genoma B. Houve algumas, no entanto que apresentaram frequências similares, tais como RE128-84 e 29, Mico, Grilo, Zoe/Zia, Matita, dentre outros. Os demais elementos com frequência entre 0,11% e 0,005% nos genomas não estão representados nos gráficos.

Tabela 7: Frequência de retrotransposons LTR (em pb e porcentagem) e número de retrotransposons com sequência completa para cada família individualmente, famílias relacionadas ou pares autônomos/não-autônomos identificados nos genomas A e B.

	Genoma A			Genoma B		
Família de retrotransposon LTR	pb	%	# elementos	pb	%	# elementos
FIDEL/Feral	136657210	11,02074274	11323	117783874	7,852258267	3753
Pipoka/Pipa1/Pipa2/Pipa3	48740876	3,930715806	1942	56677785	3,778519	1024
Apolo/Polo	28527630	2,300615323	1244	31010488	2,067365867	1018
Gordo	20407298	1,645749839	788	38502109	2,566807267	1323
Silverio/Silvia/Silver	13840147	1,116140887	689	23181309	1,5454206	846
Curu/Bravo/Golf/Hemera/Hera	13059029	1,0531475	468	12935230	0,862348667	931
Eros/Eris	12435827	1,002889274	402	21085792	1,405719467	394
Mico	9588535	0,773268952	405	10812413	0,720827533	365
Athena	9152713	0,738122016	89	13330981	0,888732067	291
RE128-84	9072226	0,731631129	326	11325694	0,755046267	414
Matita	4931315	0,397686694	137	5729505	0,381967	192
Sofi	4742125	0,382429435	13	4345481	0,289698733	55
Medusa/Musa	4147280	0,334458065	40	10649495	0,709966333	156
Grilo/Gilo	4007424	0,323179355	134	4636540	0,309102667	207
RE128-29	2724754	0,219738226	39	3397812	0,2265208	34
Doros/Duran/Dan	1768546	0,142624677	67	3549090	0,236606	215
Hermes	1745353	0,140754274	40	2646389	0,176425933	127
Bela	1519517	0,122541694	19	2201906	0,146793733	40
Zoe/Zia	1493646	0,120455323	68	2013809	0,134253933	142
Girino	1404463	0,113263145	11	2285368	0,152357867	61
Kyra	1323658	0,106746613	0			
Yara	1216647	0,098116694	39	1073152	0,071543467	22
Saturno	1122070	0,090489516	30	2025117	0,1350078	58
Ariel	1118205	0,090177823	1	1221991	0,081466067	0
Netuno/Elpis	1086201	0,087596855	46	1291703	0,086113533	80
Venon	1067576	0,086094839	60	767966	0,051197733	36
Buba	1067370	0,086078226	11	1391072	0,092738133	27
Venus	1061794	0,085628548	30	875228	0,058348533	43
Maia	965052	0,077826774	14	661243	0,044082867	30
Lima	949134	0,076543065	65	871630	0,058108667	53
Yuri	880004	0,070968065	28	1444298	0,096286533	29
Joka	800341	0,064543629	14	1007601	0,0671734	11
Yankee	680621	0,05488879	1	717206	0,047813733	7
Nemesis	626649	0,05053621	32	690549	0,0460366	50
Foxtrot	560263	0,0451825	30	551178	0,0367452	16
Delta/Charlie	556183	0,044853468	14	10407042	0,6938028	28
Talita	531857	0,042891694	20	696087	0,0464058	34
India	523393	0,042209113	1	386248	0,025749867	1
Kilo	486901	0,03926621	4	452377	0,030158467	2
Hotel	460160	0,037109677	7	283795	0,018919667	6
Bola	459485	0,037055242	1	239173	0,015944867	2
Agnus	411028	0,033147419	1			
Papa	384969	0,031045887	11	366489	0,0244326	8
Tango	377210	0,030420161	2	481210	0,032080667	2
Lulu	362662	0,029246935	39	428428	0,028561867	36
Alpha	348824	0,028130968	2	431267	0,028751133	2
James	333927	0,026929597	4	373884	0,0249256	2
Diva	332583	0,02682121	11	588823	0,039254867	24
Juliett	331290	0,026716935	6	318876	0,0212584	10
Mike	316878	0,025554677	4	167993	0,011199533	1
Dakota	309448	0,024955484	3	144920	0,009661333	5
Phenix	291881	0,02353879	2	315948	0,0210632	3
Echos	233677	0,018844919	1	232843	0,015522867	3
Oscar	208927	0,016848952	2	226028	0,015068533	4
Romeo	180677	0,014570726	1	297742	0,019849467	1
Victor	163970	0,013223387	4	93553	0,006236867	2
Vipe	158690	0,012797581	1			
George	142760	0,011512903	11	31698	0,0021132	2
Golden	142742	0,011511452	11	137138	0,009142533	4
Jasper	130594	0,010531774	2	389987	0,025999133	5
Edore	123571	0,009965403	12	156912	0,0104608	10
Omega	118400	0,009548387	9	158406	0,0105604	2
Doris	113538	0,00915629	0	291589	0,019439267	3
Xray	86970	0,00701371	0	158929	0,010595267	3
Whiskey	70917	0,005719113	8	128898	0,0085932	4
Paco				1701496	0,113433067	1
Kirke				423566	0,028237733	2
Sierra				366221	0,024414733	2
Quebec				294270	0,019618	6
November				256492	0,017099467	4
Zulu				182440	0,012162667	0
Grey				137462	0,009164133	0
Uniform				88989	0,0059326	0
TOTAL	353185611	28,48271056	18839	414528223	27,63521487	12274

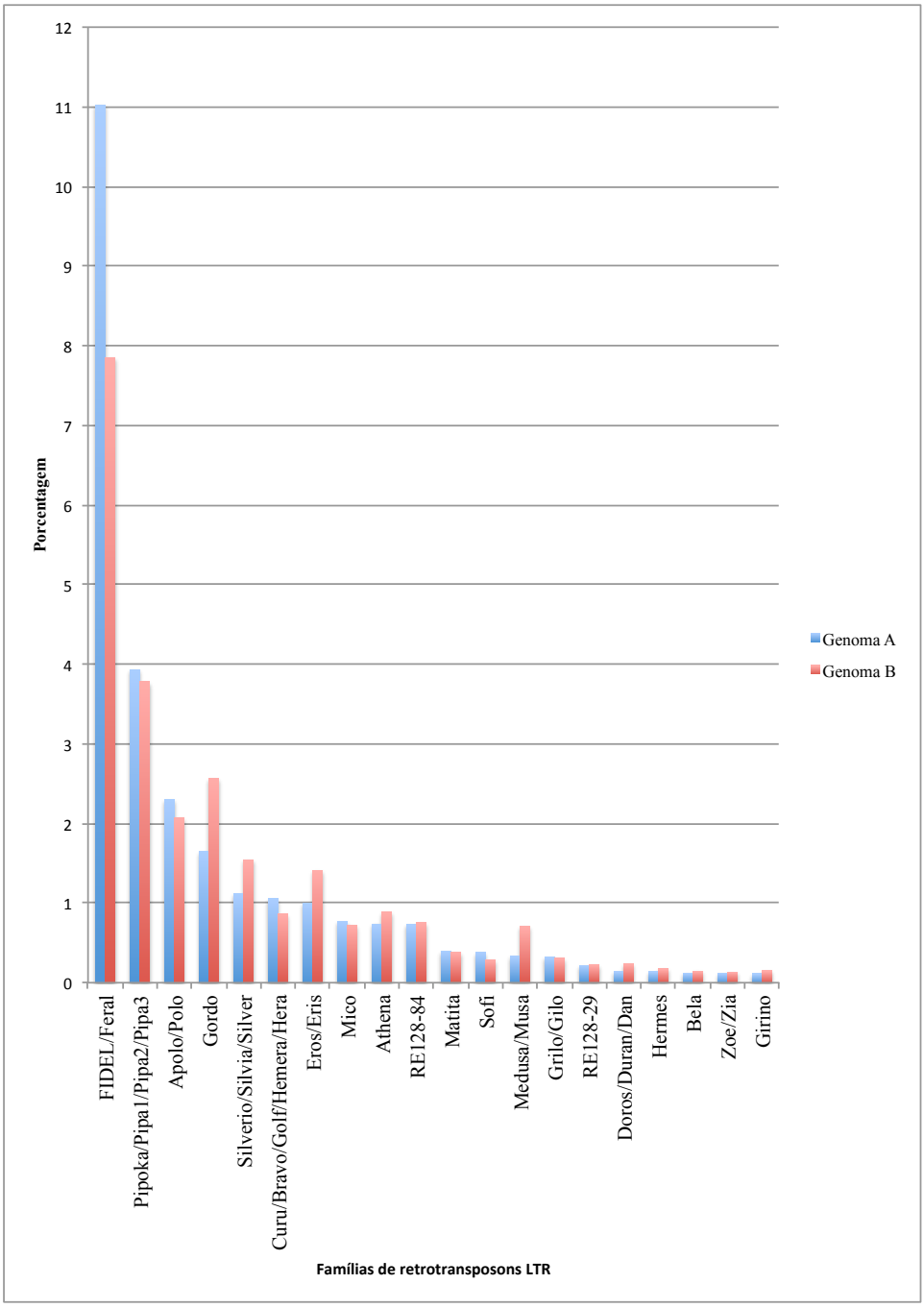


Figura 48: Gráfico mostrando porcentagem (eixo y) de ocorrência das 37 famílias mais frequentes nos genomas A e B (eixo x). Famílias que compartilham similaridade entre as sequências foram avaliadas juntas.

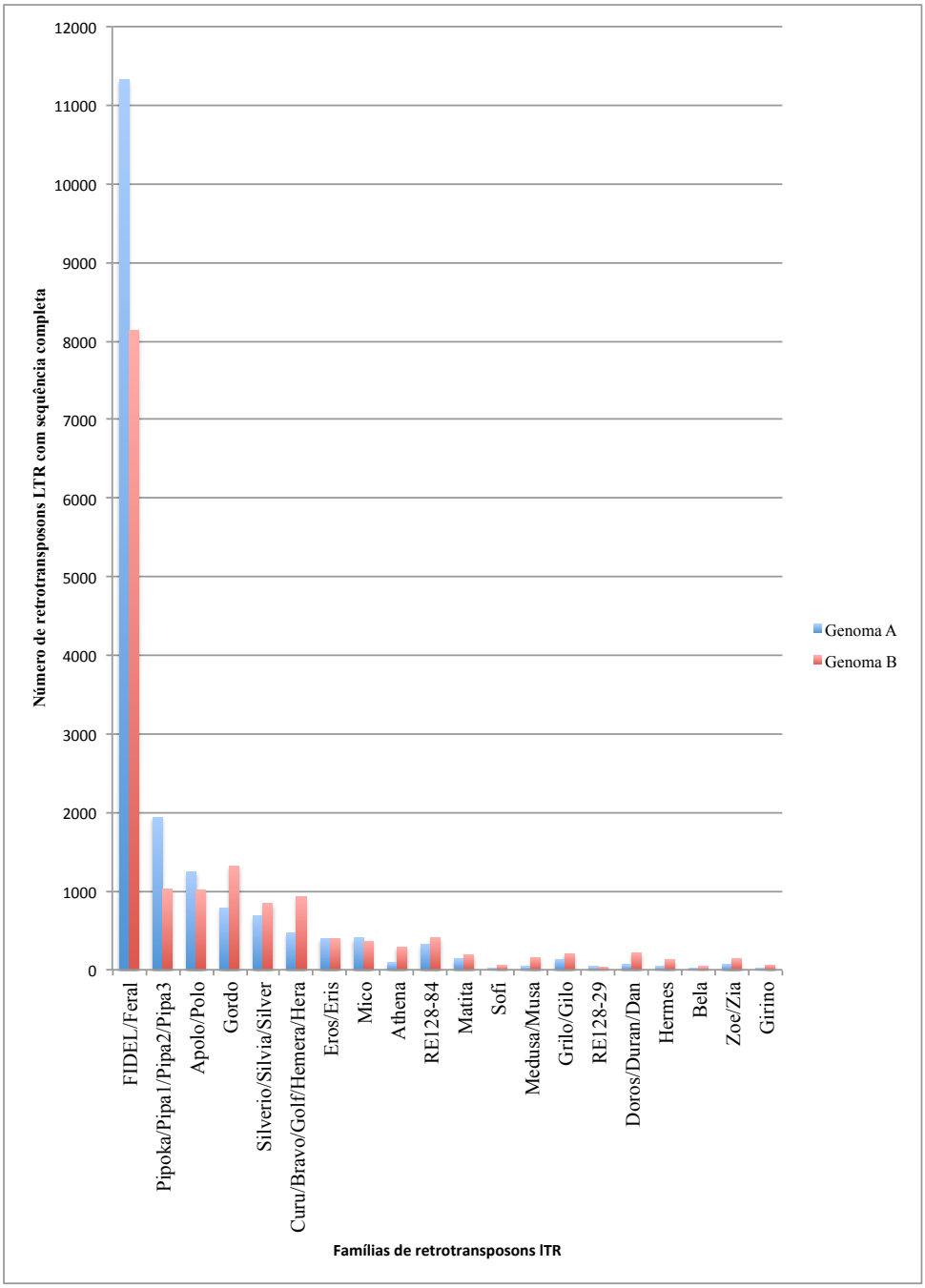


Figura 49: Gráfico mostrando o número de retrotransposons com sequência completa (eixo y) identificados em 37 famílias mais frequentes nos genomas A e B (eixo x). Famílias que compartilham similaridade entre as sequências foram avaliadas juntas.

3.5.1 Frequência e distribuição de retrotransposons LTR nas pseudomoléculas A e B

Os genomas de *A. duranensis* e *A. ipaënsis* foram ordenados e separados *in silico* em dez pseudomoléculas cada um, representando os cromossomos haploides (n=10). Com isso foi possível avaliar a porcentagem de ocorrência das famílias de retrotransposons LTR em cada pseudomolécula, assim como a distribuição de algumas famílias (tabela 8).

Tabela 8: Pseudomoléculas A e B. Tamanho de cada pseudomolécula; Cobertura (frequência) em pares de bases e porcentagem correspondente à presença das famílias de retrotransposons LTR identificadas.

Pseudomoléculas A	Tamanho (Mb)	Frequência de retrotransposons LTR (Mb)	Frequência de retrotransposons LTR (%)
A01	107	39,5	36,96
A02	93,9	28,3	30,19
A03	135,1	44,3	32,80
A04	123,6	43	34,80
A05	110	35,1	31,92
A06	112,8	37,9	33,63
A07	79,1	24,8	31,43
A08	49,5	8,7	17,73
A09	120,7	44,6	37,00
A10	109,5	36,7	33,60
Pseudomoléculas B	Tamanho (Mb)	Frequência de retrotransposons LTR (Mb)	Frequência de retrotransposons LTR (%)
B01	137,4	54,7	39,84
B02	109	37,6	34,55
B03	136,1	45,8	33,66
B04	133,6	50,2	37,64
B05	149,9	57,9	38,65
B06	137,1	52,4	38,23
B07	126,4	48,5	38,38
B08	129,6	50,6	39,08
B09	147,1	55,4	37,70
B10	136,2	53,2	39,08

A soma das dez pseudomoléculas foi menor do que o tamanho estimado para a sequência genômica completa, para ambos os genomas A e B, possivelmente devido ao descarte de *scaffolds* genômicos, durante a montagem das pseudomoléculas. A porcentagem relativa à presença dos retrotransposons LTR conhecidos e seus fragmentos foi determinada para cada pseudomolécula, e variou entre 17-40%. Como a montagem dessas sequências ainda não está totalmente finalizada, é possível que os tamanhos das pseudomoléculas A02, A07 e A08 sejam diferentes, modificando o resultado final. De forma geral, a frequência das famílias de retrotransposons LTR nas dez pseudomoléculas foi maior do que a prevista em cada genoma. Os parâmetros utilizados neste trabalho para a estimativa de porcentagem de elementos nos genomas e pseudomoléculas foram bastante estringentes, de forma que é possível considerar que essa estimativa possa ser maior.

Para representar a distribuição de algumas famílias de retrotransposons LTR nas pseudomoléculas A01 e B01 de *A. duranensis* e *A. ipaënsis*, respectivamente, foi realizada uma análise via BLASTn, na qual foi contabilizado o número de *hits* positivos encontrados nas sequências das pseudomoléculas relativos à presença e coordenada de cada retrotransposon LTR. Para isso, gráficos representando as pseudomoléculas (cromossomos) A01 e B01 foram construídos em escala de 2 Mb (figuras 50-a a 50-r).

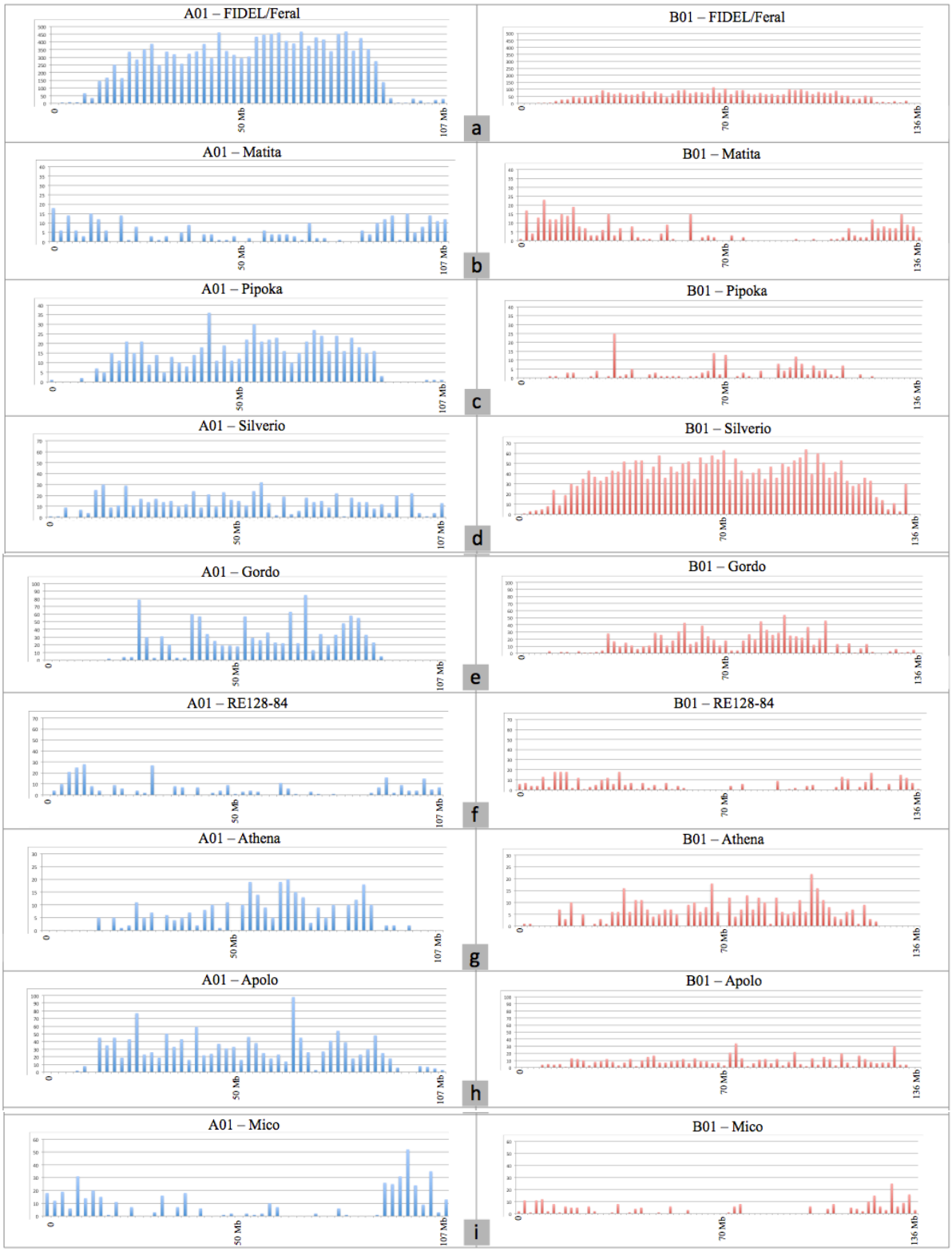
As famílias de retrotransposons FIDEL/Feral (figura 50-a), Pipoka (figura 50-c) e Apolo (figura 50-h) mostraram distribuição similar nas pseudomoléculas A01 e B01, sendo mais frequente em *A. duranensis*. A distribuição dessas famílias ao longo das pseudomoléculas ficou principalmente restrita à região pericentromérica, excluindo-se as

regiões distais, como descrito para FIDEL (Nielen *et al*., 2010), após experimentos de hibridização in situ de FIDEL em cromossomos metafásicos de amendoim.

As famílias Silverio (figura 50-d), Gordo (figura 50-e), Athena (figura 50-g) e Juliett (figura 50-m) também apresentaram resultados similares nas pseudomoléculas avaliadas, mas com a maior ocorrência em *A. ipaënsis*. A distribuição dessas famílias ficou restrita à região pericentromérica. O tamanho de B01 em relação a A01 é maior, o que deve ser levado em conta no caso dos elementos que apresentaram maior frequência em *A. ipaënsis*.

Para as famílias Matita (figura 50-b), RE128-84 (figura 50-f) e Mico (figura 50-i), com frequência similar nos dois genomas, a distribuição ocorreu preferencialmente em regiões distais das pseudomoléculas, compatível com a distribuição descrita para Matita em cromossomos metafásicos de amendoim (Nielen *et al*., 2012).

As famílias menos abundantes nos dois genomas, Venus (figura 50-j), Saturno (figura 50-k), Foxtrot (figura 50-l), Girino (figura 50-n), Golden (figura 50-o), Diva (figura 50-p), Hermes (figura 50-q) e Hotel (figura 50-r) apresentaram poucos *hits*, distribuição dispersa e menos frequente nas regiões centroméricas.


A01 – FIDEL/Feral
B01 – FIDEL/Feral
a
A01 – Matita
B01 – Matita
b
A01 – Pipoka
B01 – Pipoka
c
A01 – Silverio
B01 – Silverio
d
A01 – Gordo
B01 – Gordo
e
A01 – RE128-84
B01 – RE128-84
f
A01 – Athena
B01 – Athena
g
A01 – Apolo
B01 – Apolo
h
A01 – Mico
B01 – Mico
i
0
50 Mb
107 Mb
70 Mb
136 Mb

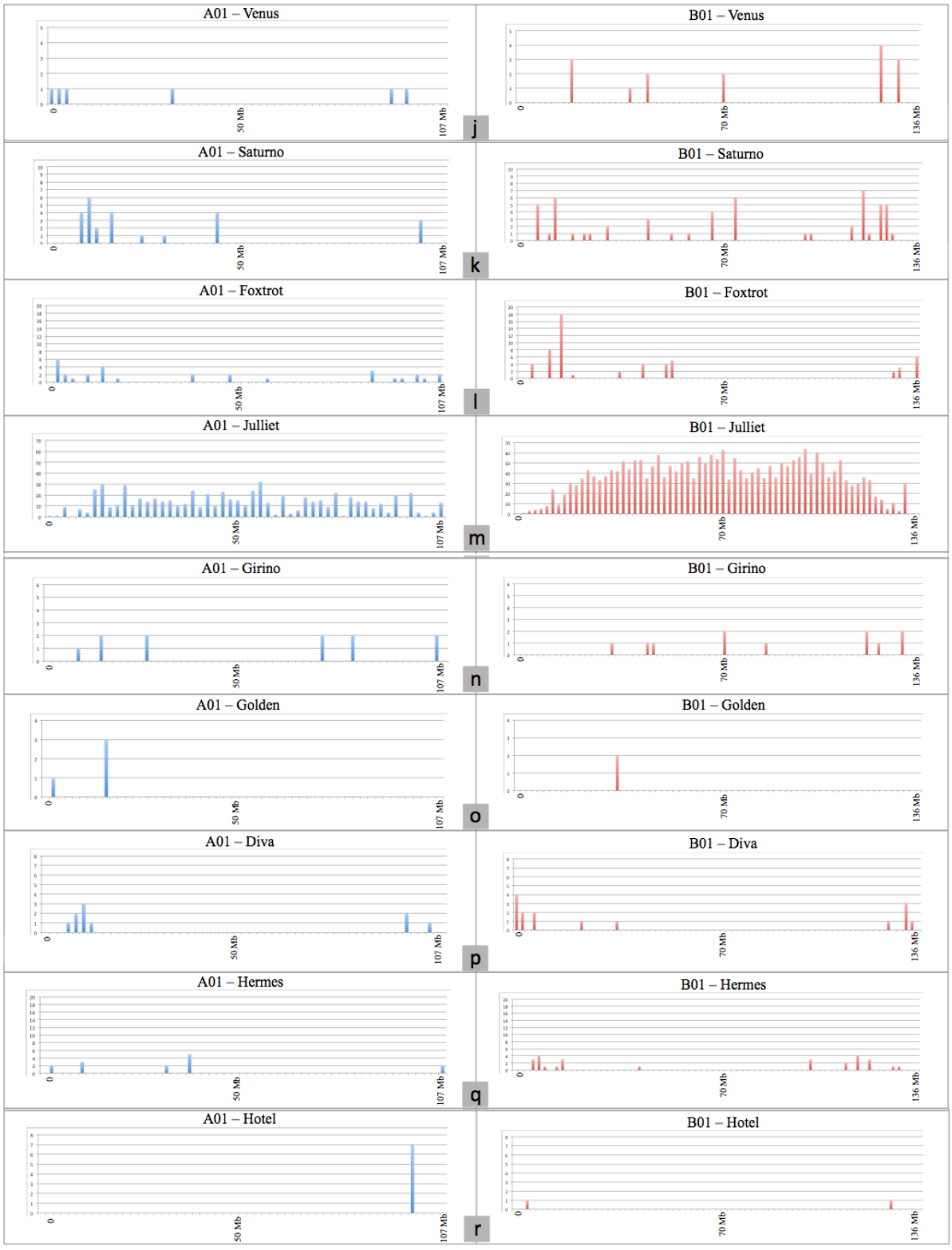


Figura 50: Distribuição de algumas famílias de retrotransposon LTR nas pseudomoléculas A01 e B01 dos genomas de *A. duranensis* e *A. ipaënsis,* respectivamente. Gráficos correspondentes ao tamanho das pseudomoléculas representados em escala de 2 Mb (eixo x). Número de *hits* positivos resultantes da comparação das pseudomoléculas com os elementos, por meio da ferramenta BLASTn (eixo y).

3.6 Estimativa das datas de transposição dos retrotransposons LTR nos genomas A e B

A média das estimativas das datas de transposição dos 18.839 retrotransposons LTR com sequência completa identificados no genoma A foi de 1,55 Ma. Enquanto para os 16.659 retrotransposons LTR do genoma B foi de 1,51 Ma. A maioria das médias das datas estimadas para os elementos das famílias equivalentes nos genomas A e B foram distintas, apesar da variação no número de elementos identificados para cada família (Anexo 1). Todas as famílias de retrotransposons LTR identificadas nos genomas A e B apresentaram a média da estimativa da data de transposição mais recente do que a data estimada para a divergência entre os genomas A e B, ou seja, 3,5 milhões de anos atrás (Nielen *et al*., 2012; Moretzsohn *et al*., 2013), indicando que os conteúdos repetitivos dos dois genomas, A e B, possivelmente seguiram fluxos evolucionários distintos. Para as 37 famílias mais frequentes, as médias das datas de transposição dos elementos foram mais antigas no genoma de *A. duranensis*, com exceção da família Mico, que apresentou a média de idade ligeiramente maior no genoma de *A. ipaënsis*, como os exemplos apresentados na figura 51.

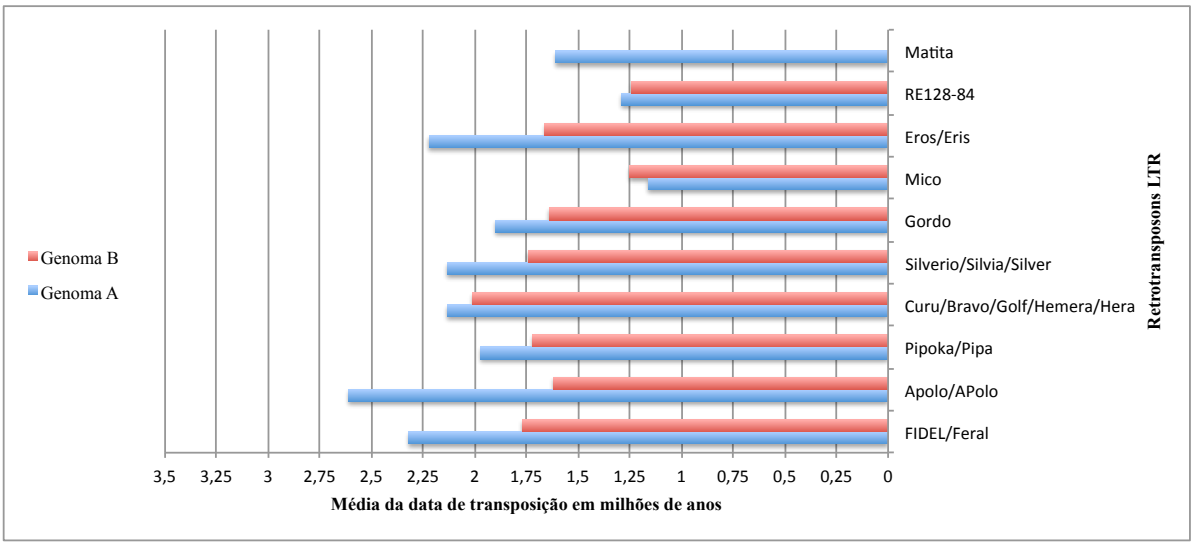


Figura 51: Gráfico que apresenta a média das datas de transposição estimadas para todos os retrotransposons LTR pertencentes às 36 famílias mais frequentes nos genomas A e B. O eixo x representa uma escala de tempo (da mais antiga para a mais atual) de 3,5 milhões de anos atrás até 0 ano (ou período recente); o eixo y representa as 36 famílias de retrotransposons LTR.

Gráficos contendo a distribuição das datas de inserção estimadas para todos os retrotransposons LTR com sequência completa pertencentes a algumas famílias dos genomas A e B foram construídos (figura 52-A a 52-R). Os resultados obtidos para as famílias mais frequentes revelaram distribuições com padrão de curva normal, onde o menor número de

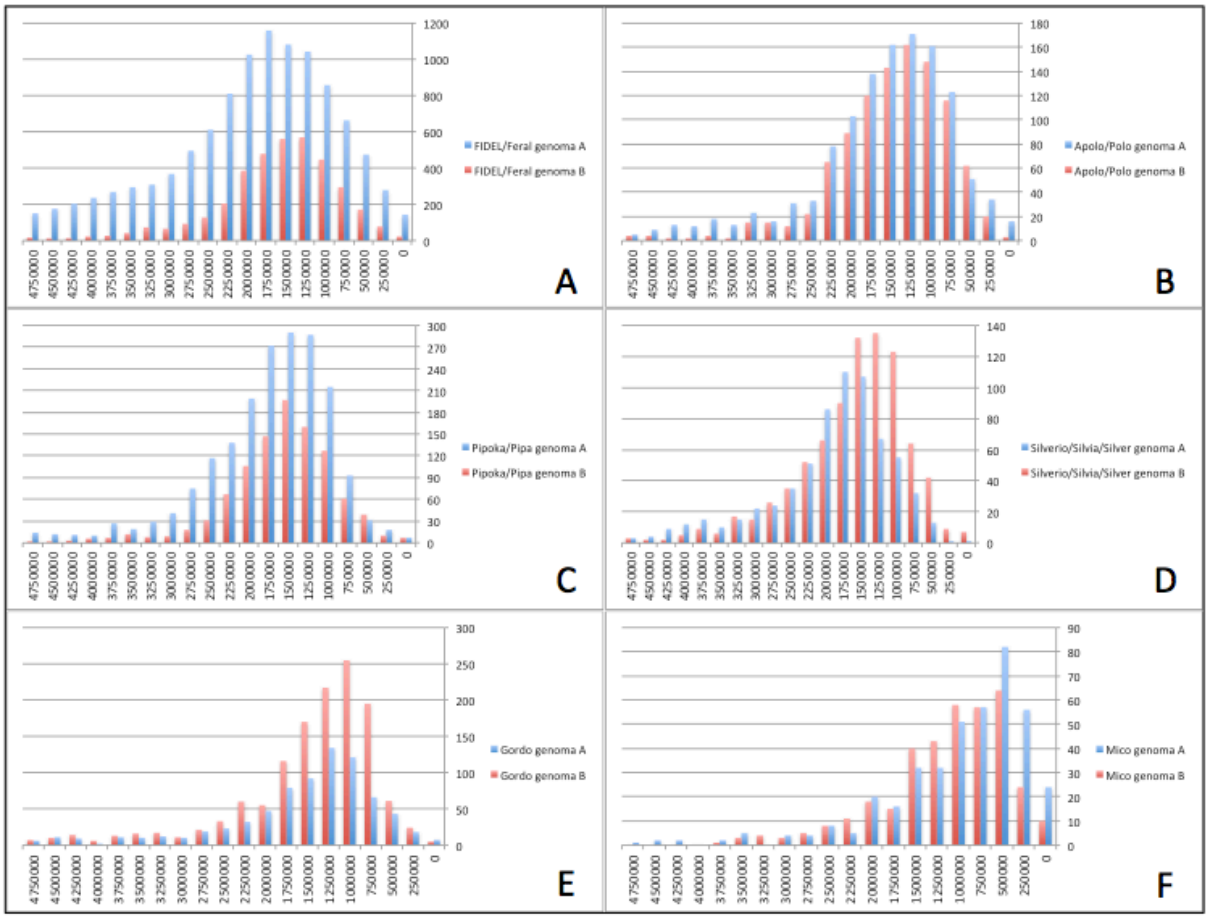
elementos apresentava datas muito recentes ou muito antigas. Muitas famílias, apesar de apresentarem médias mais recentes do que a data de divergência dos genomas A e B, continham poucos elementos individuais com datas mais antigas, chegando a 5 Ma. A construção dos gráficos de distribuição foi útil porque permitiu a constatação de qual período evolutivo em que foram identificados o maior número de retrotransposons LTR para cada família, separadamente.

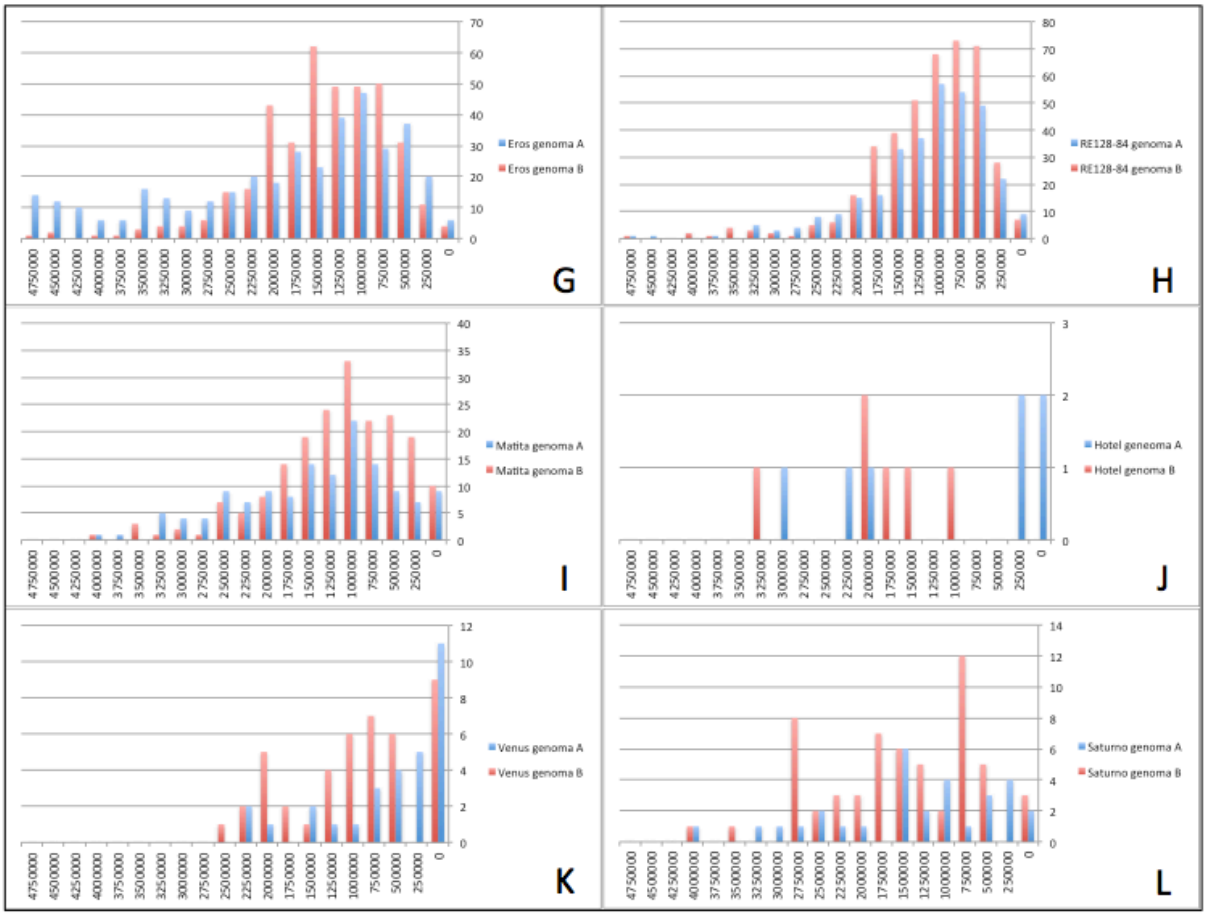
As famílias FIDEL/Feral, juntas, apesar de apresentarem frequências distintas nos genomas A e B, mostraram curvas similares no gráfico. Contudo, as estimativas das datas de transposição observadas para o maior número de elementos foram aproximadamente 1,75 Ma para o genoma A e 1,25 Ma para o genoma B (figura 52-A). Para as famílias Apolo/Polo e Pipoka/Pipa, as distribuições também foram bastante similares apesar da frequência dessas famílias no genoma A ser ligeiramente maior. Para Apolo/Polo, em ambos os genomas, o maior número de elementos apresentava a data de transposição estimada em aproximadamente 1,25 Ma (figura 52-B), enquanto que para Pipoka/Pipa, 1,5 Ma (figura 52-C).

Famílias mais frequentes no genoma B, como Silverio/Silvia/Silver, Gordo e Athena apresentaram o maior número elementos com datas mais recentes no genoma B, aproximadamente 1,25 Ma, 1 Ma e 1,5 Ma (figuras 52-D, 52-E, 52-O), respectivamente. A maioria dos elementos pertencentes às famílias Eros/Eris, por outro lado, apresentaram data mais recente no genoma A, porém também foram identificados elementos individuais bastante antigos, com datas em torno de 4,75 Ma (figura 52-G).

As famílias com frequências similares em ambos os genomas, Mico (figura 52-F) e Matita (figura 52-I) apresentaram para o maior número de elementos, datas de transposição correspondendo a aproximadamente 0,5 Ma e 1 Ma, respectivamente em ambos os genomas. Já para a família RE128-84, a maioria dos elementos apresentou data mais recente no genoma B, com 0,75 Ma (figura 52-H), ao passo que a família Grilo (N), mais recente no genoma A com 0,25 Ma (figura 52-N).

Em virtude da baixa frequência nos genomas A e B, algumas famílias não apresentaram uma curva com um padrão informativo, porém com tendência a exibirem elementos com datas de transposição mais recentes do que 3,5 Ma (data de divergência entre genomas) até data bastante recente (0 ano), como Hotel (figura 52-J), Venus (figura 52-K), Foxtrot (figura 52-M), Golden (figura 52-P) e Diva (figura 52-R). As família Saturno (figura 52-L) e Hermes (figura 52-Q) apresentaram poucos elementos com data anterior à divergência.


FIDEL/Feral genoma A
FIDEL/Feral genoma B
A
Apolo/Polo genoma A
Apolo/Polo genoma B
B
Pipoka/Pipa genoma A
Pipoka/Pipa genoma B
C
Silverio/Silvia/Silver genoma A
Silverio/Silvia/Silver genoma B
D
Gordo genoma A
Gordo genoma B
E
Mico genoma A
Mico genoma B
F


G
Eros genoma A
Eros genoma B
H
RE128-84 genoma A
RE128-84 genoma B
I
Matita genoma A
Matita genoma B
J
Hotel geneoma A
Hotel genoma B
K
Venus genoma A
Venus genoma B
L
Saturno genoma A
Saturno genoma B

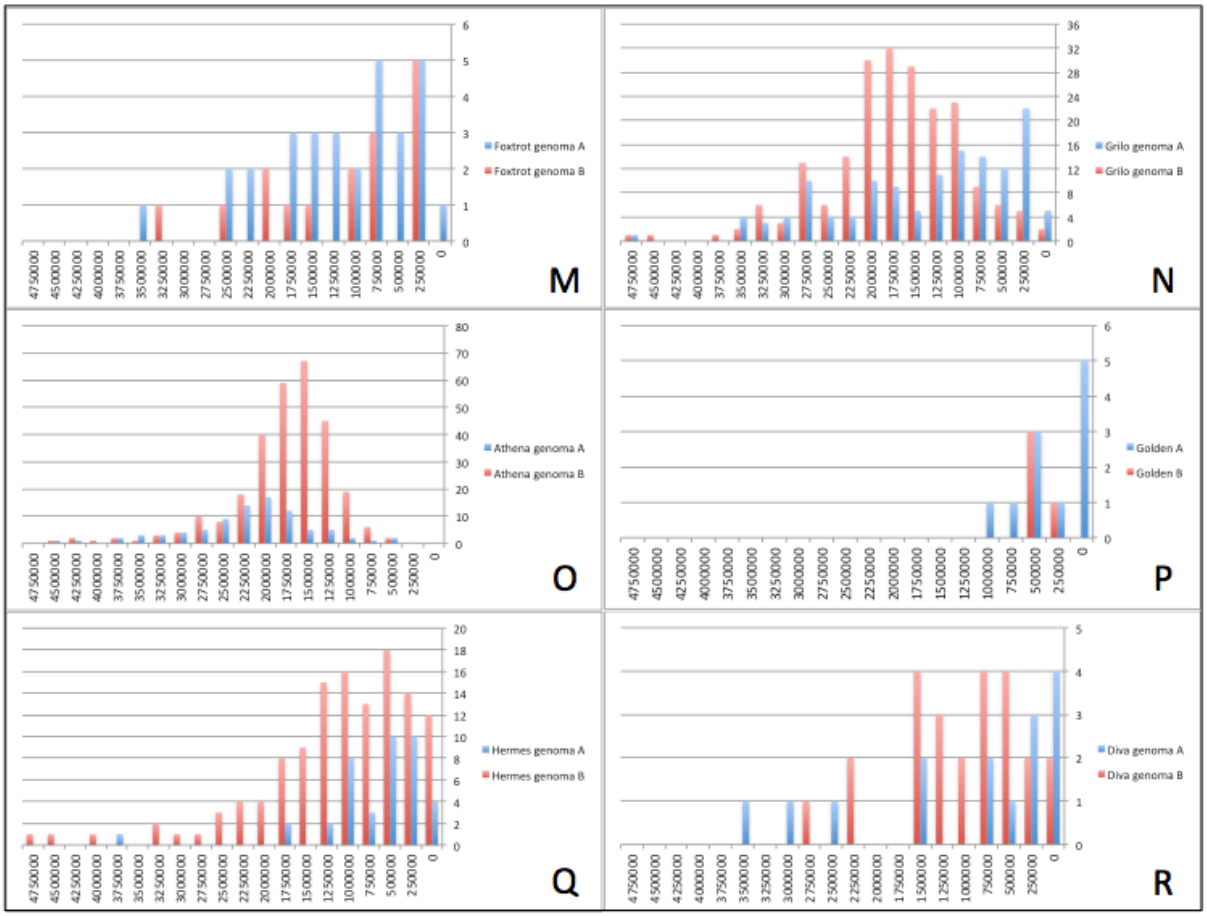


Figura 52: Gráficos com distribuição das datas de transposição estimadas para retrotransposons LTR pertencentes a algumas famílias nos genomas A e B. O eixo x representa uma escala de tempo de 4,75 milhões de anos atrás (Ma) até 0 ano (período recente); o eixo y representa o número de elementos com sequência completa.

3.6 Distribuição dos retrotransposons LTR em cromossomos de amendoim (genoma AABB)

3.6.1 Alinhamento de sequências de TR (Transcriptase Reversa) e desenho de *primers*

Foram selecionadas 14 famílias de retrotransposons LTR do genoma B de *A. ipaënsis* para investigar a distribuição dessas sequências em cromossomos metafásicos de amendoim tetraploide (subgenomas A e B). Essas famílias diferem entre si na superfamília e frequência em que ocorrem nos genomas A e B de *A. duranensis* e *A. ipaënsis*. Seis dessas famílias são *Ty1-Copia*, sete *Ty3-Gypsy* e uma delas contém elementos não-autônomos (Athena). As famílias Apolo, Golden, Venus, Foxtrot, Hotel e Juliett apresentaram maior frequência no genoma A. Saturno, Girino, Athena, Hermes e Diva apresentaram maior frequência no genoma B. As famílias Mico, Grilo e RE128-84 apresentaram frequências similares em ambos os genomas (tabela 7).

Como sonda para FISH foi utilizada a região correspondente à sequência do gene que codifica a enzima transcriptase reversa (TR), em virtude do seu alto nível de conservação entre retrotransposons LTR autônomos de uma mesma família.

Todos os alinhamentos produzidos pelas sequências de elementos das 14 famílias foram visualizados na interface do programa Jalview (figura 53). Os resultados exibiram alta similaridade entre as sequências de TR possibilitando o desenho de pares de *primers* para essa região. Os nomes, sequências dos pares de *primers* e o tamanho do produto de PCR amplificado estão listados na tabela 9.

Figura 53: Exemplo de parte de um alinhamento entre sequências do gene que codifica a enzima transcriptase reversa em retrotransposons LTR da família Apolo, visualizados na interface do *software* Jalview.

Tabela 9: Nome dos pares de *primers*, sequência e tamanho da sequência amplificada de cada uma das TRs das 14 famílias de retrotransposons LTR.

Nomes dos retrotransposons LTR	Nomes dos *primers*	Sequência na direção 5'-3'	Tamanho do produto em pb
B60_Venus	B60_Venus_aut-Ty3-FWD B60_Venus_aut-Ty3-REV	GACGACATGGTCGYCAARAC AAAACAGGTGGTTTTGAAAG	403
B57_Mico	B57_Mico_aut-Ty3-FWD B57_Mico_aut-Ty3-REV	TATTGCTATGTGGTCATGCC AGGGATCTTAATGCTGATCC	406
B62_Saturno	B62_Saturno_aut-Ty3-FWD B62_Saturno_aut-Ty3-REV	ATGTGTGTTGATTTCACCGATCTC AYAGTGCAGCAAGTCTCCCT	567
B9_Girino	B9_Girino_aut-Ty3-FWD B9_Girino_aut-Ty3-REV	CATGGCTGTCCAAYGTYGTC GCTCATRAGRCTTTGGTAGGTGGC	310
B16_Grilo	B16_Grilo_aut-Ty3-FWD B16_Grilo_aut-Ty3-REV	CCCAAGGACTGCTACCCCCT ACCCCRCTTTGGGTTATCAT	433
B69_Apolo	B69_Apolo_aut-Ty3-FWD B69_Apolo_aut-Ty3-REV	GTAGAAGCTGGCAAATTCYT TAAGYARYGGTGGTTGCCCCA	285
B65_Foxtrot	B65_Foxtrot_aut-Ty1-FWD B65_Foxtrot_aut-Ty1-REV	CCTCAAAAAGAGCGGAGACT CATATCCTGCAGGTTGCTCA	505
B50_Hotel	B50_Hotel_aut-Ty1-FWD B50_Hotel_aut-Ty1-REV	CAGTGGAGAGGCACAAGACA GAGCTTGCCCACAATTCTTC	712
B71_Juliett	B71_Juliett_aut-Ty1-FWD B71_Juliett_aut-Ty1-REV	TATGGCTTGAAACAGGCAAG TCCAAAAATTGGCTGAGCTT	512
B84-RE128	B84-RE128-FWD B84-RE128-REV	CCACTAGATCCTCAAGCAAG AGAAGGCACTAAGCCTTTC	558
B24-Athena	B24-Athena-FWD B24-Athena-REV	CCATCATAATTATCATAGTTGTGG CTCCAAACCAAGAGGGTGATAAC	618
B144-Diva	B144-Diva-FWD B144-Diva-REV	CTCAAGTGGTGGAGATAGAG ACCATCTGACTTAGTAGGATC	469
B146-Golden	B146-Golden-FWD B146-Golden-REV	CCAAGGAGAAGCTTCAACTG GATGTCTGCTTGTTGAGAGC	471
B1-Hermes	B1-Hermes-FWD B1-Hermes-REV	GAAGCTCGGCAGTGCTACAA GCATCTTTAGGGCAAGCTTT	559

3.6.2 Clonagem das sequências de RT e análise do perfil de restrição enzimática

Para todos os 14 pares de *primers* foram realizados gradientes de temperaturas, por meio de reação de amplificação por PCR, utilizando o DNA genômico de *A. ipaënsis.* O resultado para nove desses pares é mostrado na figura 54.

Para ligação em vetor de clonagem, foram utilizados os produtos de amplificação com a maior temperatura e, após a transformação em células de *E. coli,* foram selecionadas 2-5 colônias recombinantes (brancas) para extração do plasmídeo. A presença e o tamanho do inserto foram conferidos após digestão com a enzima de restrição *Eco*RI, e para cada par de *primer* testado, foi selecionado para o sequenciamento, apenas um clone com tamanho de inserto esperado, representado por seta na figura 55.

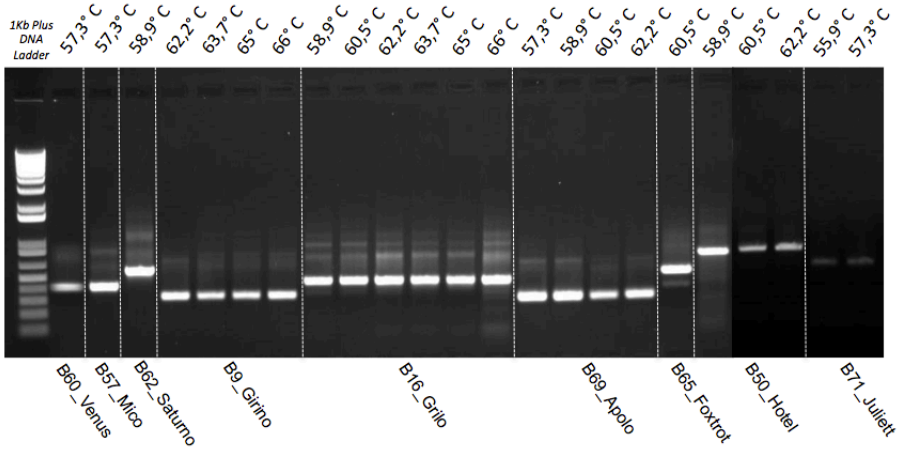


Figura 54: Amplificação da RT no DNA genômico de *A. ipaënsis*, utilizando nove pares de *primers* em gel de agarose 1,0% corado com brometo de etídio. Para todas as temperaturas testadas, os tamanhos dos produtos de PCR foram compatíveis com o tamanho esperado. O marcador utilizado foi o *1Kb Plus DNA Ladder* (Invitrogen) (poço 1).

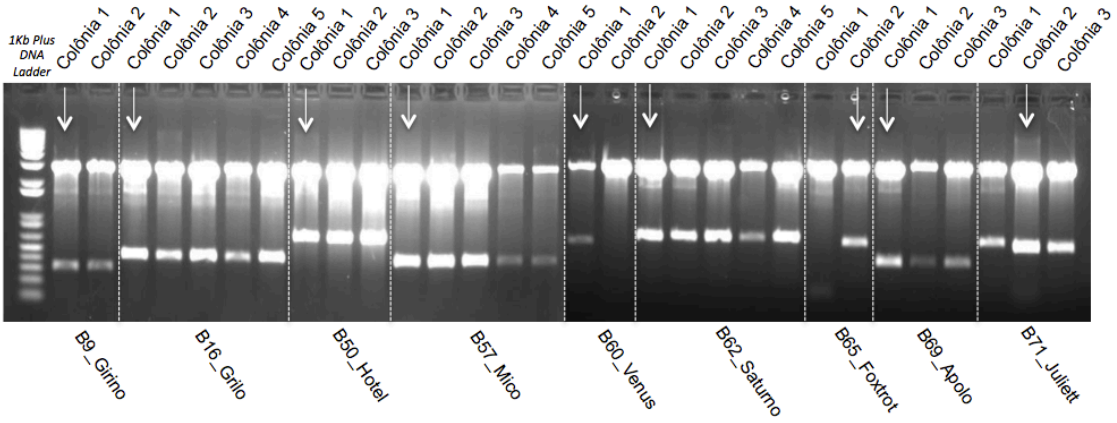


Figura 55: Perfil de restrição enzimática, utilizando a enzima *Eco*RI, dos clones selecionados a partir das colônias brancas visualizados em gel de agarose 1,0% corado com brometo de etídio. Os clones com os tamanhos dos insertos esperados (seta branca) foram selecionados para sequenciamento. O marcador utilizado foi o *1Kb Plus DNA Ladder* (Invitrogen) (poço 1).

3.6.3 Sequenciamento e análise dos dados

Foram geradas sequências de RT de cada uma das 14 famílias nas duas direções, e com a utilização das ferramentas Vecscreen (NCBI) e Pregap4 (pacote Staden Package), foram removidas as sequências de vetores e sequências com baixo índice de confiança. Comparações com o banco de dados do NCBI, por meio da ferramenta BLASTx, revelaram que as sequências obtidas neste experimento tratavam-se de fato, de fragmentos do gene que codifica a enzima TR em retrotransposons autônomos (com exceção de elementos da família Athena, composta por elementos não-autônomos).

Para corroborar os resultados de cada família, foi realizado um alinhamento entre as

sequências de TR obtidas pela clonagem e os respectivos elementos representantes das famílias selecionadas. A compilação de dados na mesma escala contendo o alinhamento das TRs (Gap4 - Staden) com a sequência anotada do elemento representante Mico por exemplo, (visualizada na interface do Artemis), está na figura 56. A complementariedade direta e reversa das sequências de RT forneceu evidências de que a utilização desses fragmentos como sondas para FISH seria bastante específica e portanto, muito informativa.

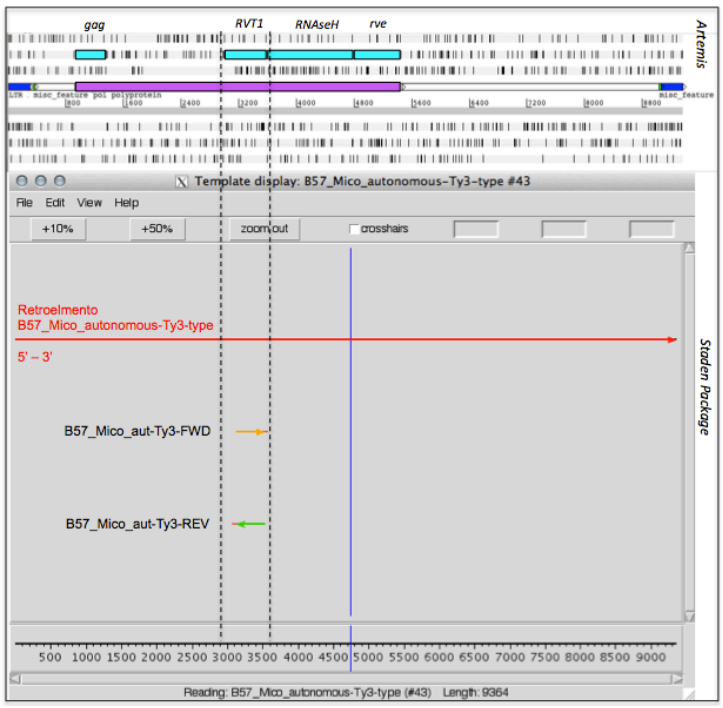


Figura 56: Sequência do retrotransposon LTR representante B57_Mico_autonomous-Ty3-type complementar às sequências direta e inversa da sequência da TR amplificada de *A. ipaënsis* por *primers* específicos, a partir de análises feitas nos *softwares* Artemis e Gap4 (mesma escala em pb).

3.6.4 Hibridização in situ por fluorescência (FISH)

3.6.4.1 Preparação das sondas e teste para verificação de nucleotídeos incorporados (*dot blot*)

Os produtos das reações de PCR realizadas com os pares de *primers* e os plamídeos contendo as sequências de RT clonadas foram purificados e utilizados na confecção de sondas de DNA marcadas. Para cada sequência foi feita uma sonda marcada com biotina e outra marcada com digoxigenina. A presença do *dot* cinza na membrana confirmou a incorporação dos nucleotídeos marcados em todas as 28 sondas (figura 57).

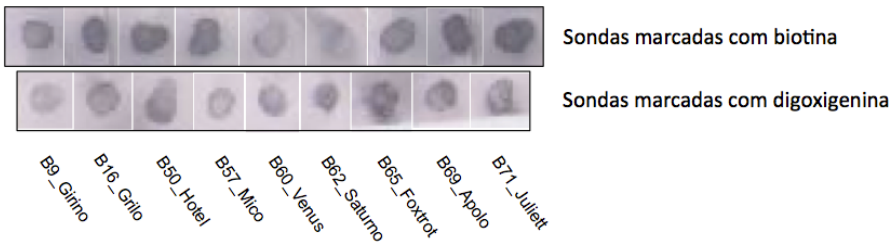


Figura 57: *Dot blot* das sondas de DNA com nucleotídeos marcados com digoxigenina ou biotina obtidas pela técnica de *Random Primer* a partir das sequências do gene que codifica a enzima TR em nove famílias de retrotransposons LTR selecionadas para FISH.

3.6.4.2 Hibridização in situ por fluorescência – FISH

Foram selecionadas lâminas contendo metáfases isoladas a partir de meristemas de raízes de plantas de *A. hypogaea* (cultivar Tatu) (figura 58).

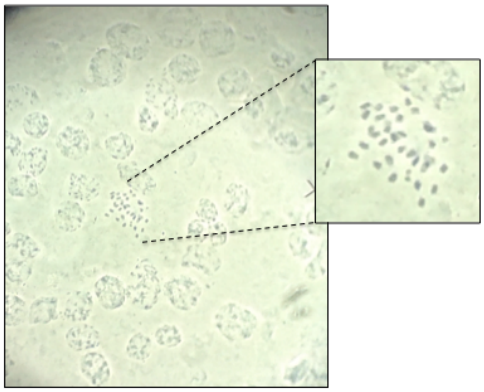


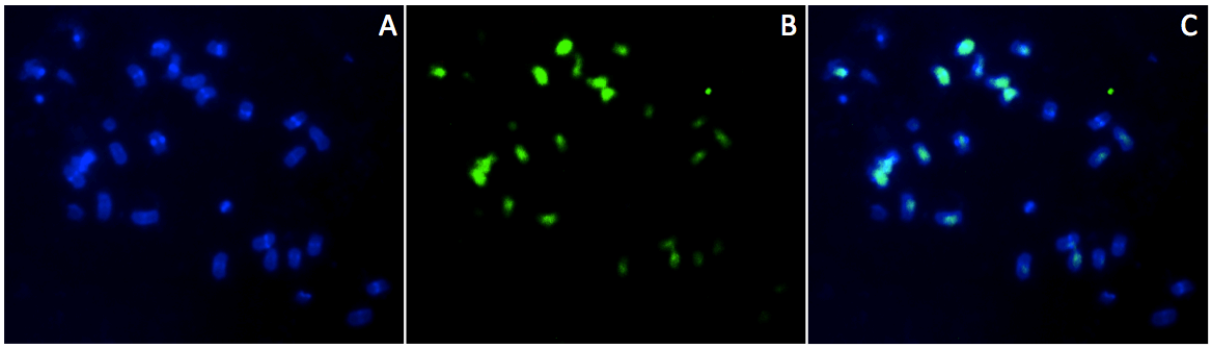
Figura 58: Células meristemáticas isoladas de meristemas de raízes de plantas de A. *hypogaea,* mostrando vários núcleos interfásicos e um conjunto de cromossomos em metáfase ao centro (detalhe).

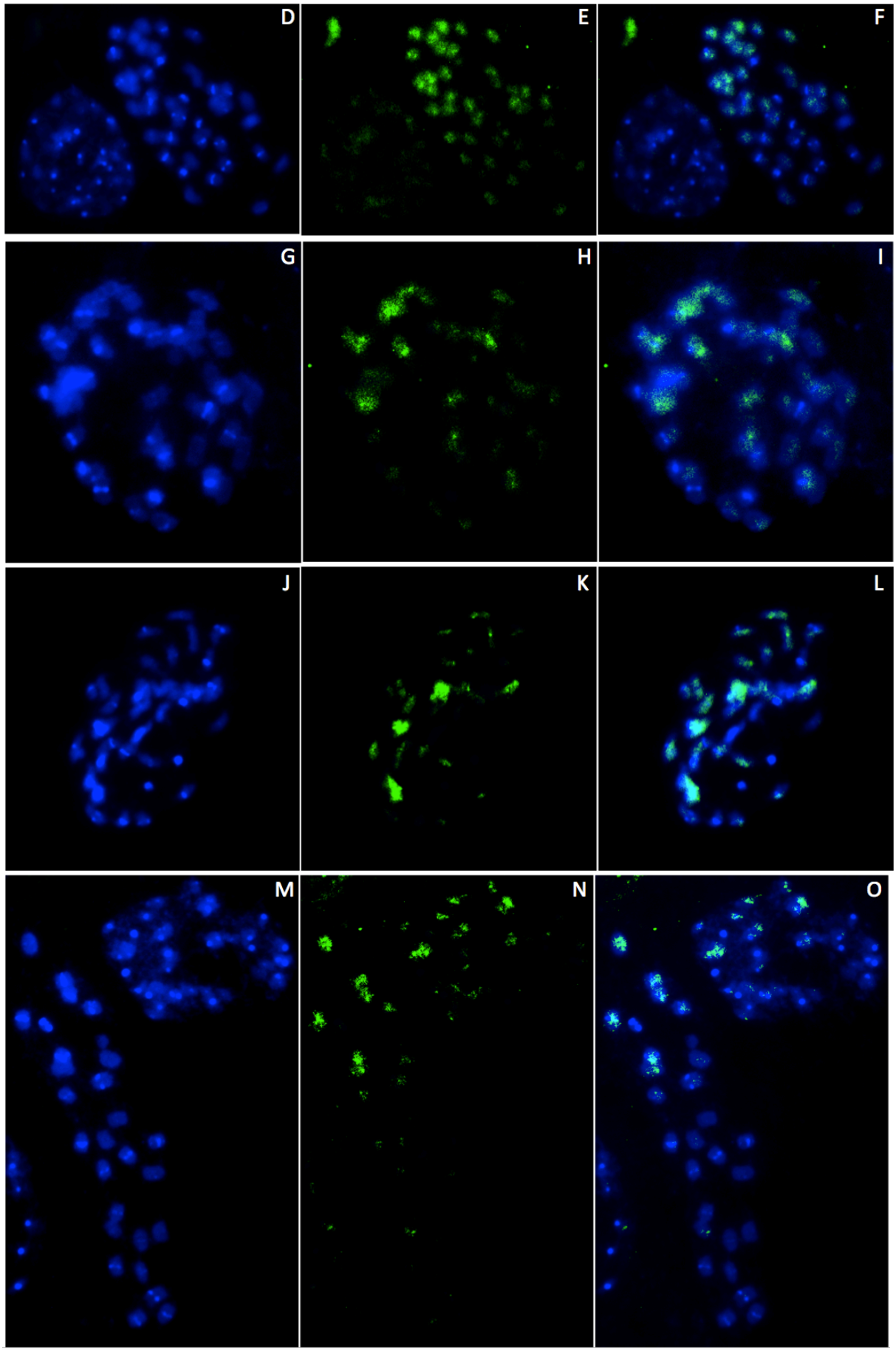
Como descrito em Seijo *et al*. (2004), o uso de DAPI para evidenciar cromossomos de *A. hypogaea* revelou bandas heterocromaticas distintas nos cromossomas do genoma A e bandas ausentes ou consideravelmente mais fracas em cromossomos do genoma B. Todas as sondas utilizadas para FISH resultaram em sinais de hibridização detectáveis múltiplos e dispersos presentes em vários, mas não em todos os cromossomos de amendoim (figuras 59-A a 59-S).

Todas as sondas marcadas com digoxigenina e observadas em verde produziram resultados contendo marcações detectáveis, ao passo que quando marcadas com biotina (observadas em vermelho) o resultado não foi satisfatório, com exceção da sonda construída para a RT da família Venus.

A hibridização da sonda de TR da família Juliett pertencente à superfamília *Ty1-Copia* produziu sinais na maioria dos cromossomos, tanto do subgenoma A, quanto do B, porém

mais forte em alguns cromossomos A (figura 59-C). Os sinais foram observados mais nas regiões pericentroméricas, estando ausentes nas regiões distais. Duas sondas de TR desenvolvidas para a família Saturno (*Ty3-Gypsy*), marcadas com digoxigenina produziram sinais dispersos ao longo dos dois braços da maioria dos cromossomos principalmente do subgenoma B, excluindo a região centromérica (figuras 59-F e 59-S). Sinais difusos de hibridização produzidos pela sonda de Venus (*Ty3-Type*), marcada com biotina, foram observados na maior parte dos cromossomos A e B de amendoim (figura 59-R), apresentando alguma sobreposição de sinais (em amarelo) com a sonda de Saturno (figura 59-S). A sonda Diva (*Ty1-Copia*) mostrou também um padrão disperso de distribuição com sinais mais fortes na região pericentromérica e mais fracos ou ausentes nas regiões centromérica e distal da maioria dos cromossomos, porém sendo ligeiramente mais forte em cromossomos do subgenoma B (figura 59-I). Para a família Golden (*Ty1-Copia*), foi possível detectar marcação da sonda apenas em alguns cromossomos A e B, de forma difusa e preferencialmente na região pericentromérica dos cromossomos, excluindo as regiões distais e mais frequentes no subgenoma A (figura 59-O). Já com a sonda da família RE128-84 (*Ty1-Copia*) foi possível detectar sinais evidentes de hibridização na maioria dos cromossomos A e B, ao longo dos braços (figura 59-L), porém não foi possível distinguir se houve hibridização preferencial em um ou outro subgenoma. Esses dados indicam que cada um dos retrotransposons LTR selecionados neste estudo apresenta uma distribuição específica, de acordo com o subgenoma (A ou B), com padrão de distribuição e intensidade de sinais de hibridizaçãoo identificáveis nos cromossomos e com distribuição determinada em regiões dos braços dos cromossomos.




D
E
F
G
H
I
J
K
L
M
N
O

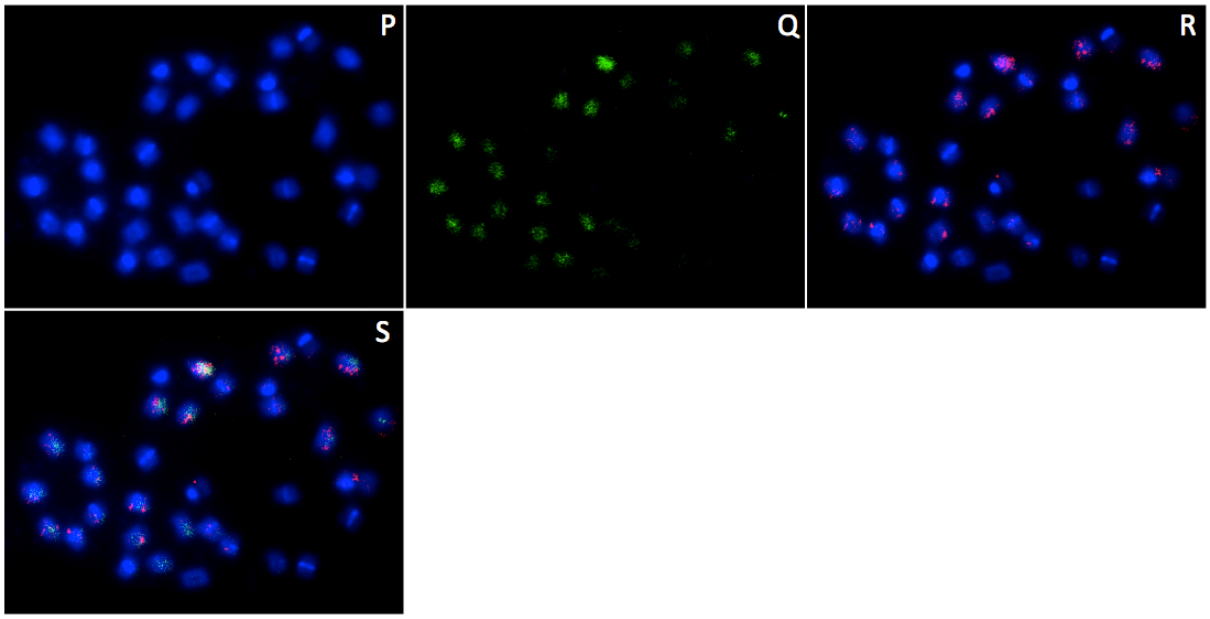


Figura 59: Cromossomos metafásicos de amendoim (*Arachis hypogaea*) contra corados com DAPI (azul) e após hibridização *in situ* por fluorescência com sondas de TR de diferentes famílias de retrotransposons LTR do genoma B de *A. ipaënsis*. Sondas marcadas com digoxigenina tiveram os sinais de hibridização detectados com anticorpo anti-digoxigenina conjugado com FITC (verde) e as sondas marcadas com biotina, sinais detectados com estreptavidina conjugada com Alexa Flúor 594 nm (vermelho). **(A, D, G, J, M, P)** Coloração com DAPI mostrando metade dos cromossomos contendo bandas centroméricas fortemente coradas, típicos do subgenoma A e fracamente coradas ou ausentes, típicos do subgenoma B. **(B)** Sonda obtida da RT da família Juliett. **(C)** FISH com a sonda obtida da RT de Juliett mostrando sinais de hibridização na maioria dos cromossomos A e B, porém mais forte em alguns cromossomos A. Os sinais foram observados predominantemente nas regiões pericentroméricas, estando ausentes nas regiões distais. **(E,Q)** Sondas obtidas da RT da família Saturno. **(F, S)** FISH com a sonda obtida da RT de Saturno mostrando sinais dispersos ao longo dos dois braços da maioria dos cromossomos principalmente do subgenoma B, excluindo a região centromérica. A figura S é uma sobreposição de resultados com Saturno e Venus. **(H)** Sonda obtida da RT da família Diva. **(I)** FISH com a sonda obtida da RT de Diva mostrando um padrão disperso de distribuição com sinais mais fortes na região pericentromérica e mais fracos ou ausentes nas regiões centromérica e distal da maioria dos cromossomos, porém sendo ligeiramente mais forte em cromossomos do subgenoma B. **(K)** Sonda obtida da RT da família RE128-84. **(L)** FISH com a sonda obtida da RT de RE128-84 mostrando sinais evidentes de hibridização na maioria dos cromossomos A e B, ao longo dos braços, porém não foi possível distinguir se houve hibridização preferencial em um ou outro subgenoma. **(N)** Sonda obtida da RT da família Golden. **(O)** FISH com a sonda obtida da RT de Golden mostrando marcação apenas em alguns cromossomos A e B, de forma difusa e preferencialmente na região pericentromérica dos cromossomos, excluindo as regiões distais e mais frequentes no subgenoma A. **(Q)** Sonda obtida da RT da família Golden. **(R)** Sonda obtida da RT da família Venus. **(R, S)** FISH com a sonda obtida da RT de Venus mostrando sinais difusos de hibridização observados na maior dos cromossomos A e B de amendoim.

3.7 Comparação entre sequências homeólogas nos genomas A e B de *Arachis*

Análises realizadas com as sequências genômicas de *A. duranensis* (*scaffold*_45 - genoma A) com aproximadamente 2,2 Mb e de *A. ipaënsis* (*scaffold*_47 - genoma B) com 2,4 Mb, ambas contendo o marcador Leg128 (desenvolvido para leguminosas), comparadas por meio de gráficos de plotagem (*dot plots*) desenvolvidos pelo *software* Gepard, indicaram alta similaridade, confirmando a existência de macrossintenia desta região entre as duas espécies (figura 60). A linha diagonal, resultante da comparação entre as duas sequências, representou a macrossintenia, no entanto, os espaços ou descontinuidades presentes na diagonal foram caracterizados principalmente por lacunas na montagem dessas sequências genômicas, representados por "N" nas sequências. Para facilitar a representação dos resultados que mostram a similaridade entre as regiões genômicas A e B, quatro regiões menores, equivalentes em ambos os genomas, foram selecionadas para apresentação dos resultados de investigação dos componentes gênico e repetitivo (tabela 10).

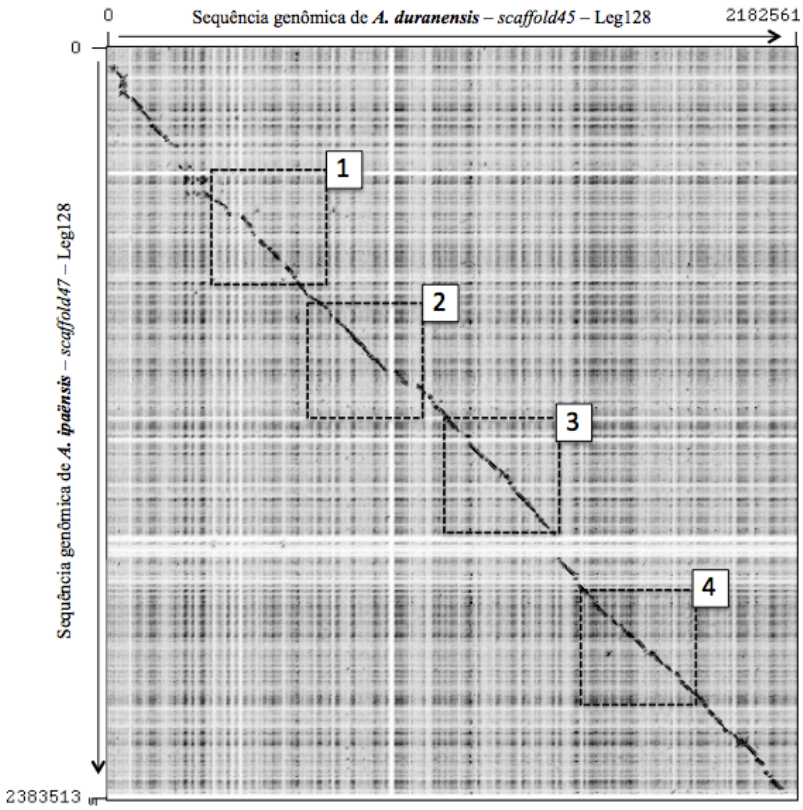


Figura 60: Gráfico de plotagem comparando as sequências genômicas de *A. duranensis* (*scaffold*_45) com 2,2 Mb (eixo x) e *A. ipaënsis* (*scaffold*_47) com 2,4 Mb (eixo y), resultando em uma linha diagonal que indicou macrossintenia dessa região nas duas espécies. Estão representadas quatro regiões selecionadas para análise detalhada quanto ao conteúdo gênico e repetitivo. Gráfico de plotagem produzido pelo programa Gepard.

Tabela 10: Coordenadas dos intervalos de bases das quatro regiões selecionadas nas sequências genômicas de *A. duranensis* (*scaffold*_45) e *A. ipaënsis* (*scaffold*_47) para apresentação dos resultados de comparação dos conteúdos gênico e repetitivo.

Comparação entre regiões homeólogas dos genomas A e B			
Genoma A - *scaffold* 45		Genoma B *scaffold* 47	
Região 1	base 42.000 a 627.000	Região 1	base 531.200 a 780.000
Região 2	base 904.200 a 1.119.800	Região 2	base 1.018.600 a 1.280.800
Região 3	base 1.177.000 a 1.412.400	Região 3	base 1.288.000 a 1.540.000
Região 4	base 1.570.800 a 1.777.600	Região 4	base 1.790.400 a 1969.600

A comparação entre os intervalos de sequências referentes às quatro regiões selecionadas nos genomas A e B apresentaram alta similaridade entre o conteúdo gênico, tanto na localização, quanto nas funções preditas. A sobreposição entre os resultados contendo as regiões codificantes (CDSs), identificadas por meio de comparações com sequências de proteínas presentes no banco de dados do *pfamA,* e os genes putativos, representados por meio de estruturas de introns e exons preditos pelo *software* FGENESH, forneceram fortes evidências quanto à correta indicação dos prováveis genes presentes nestas sequências, diminuindo a possibilidade de anotação equívoca. A presença de exons demasiadamente pequenos (<50 pb) ou sem *hits* com *pfamA* foram desconsiderados desta análise. Um exemplo do resultado gerado pela sobreposição de análises realizadas em um pequeno intervalo da região 1 do genoma de *A. duranensis* é mostrado na figura 61.

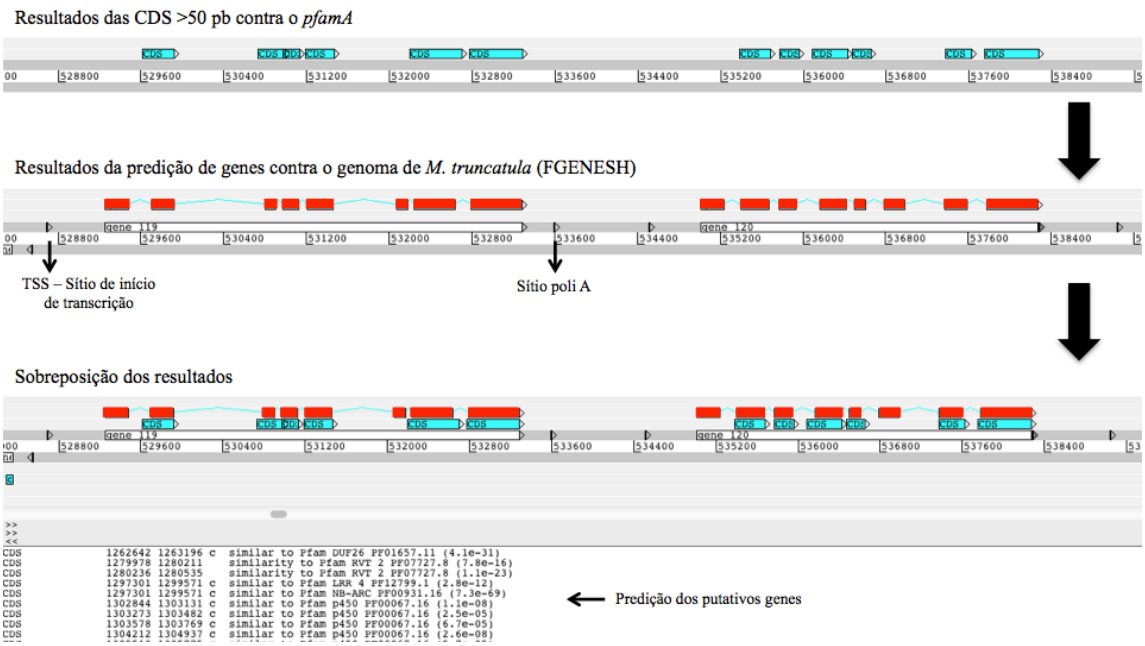


Figura 61: Parte da sequência da região 1 de *A. duranensis* mostrando a identificação de CDSs (barras azuis) que apresentaram *hits* positivos obtidos por meio de comparações com sequências do *pfamA*; Dois genes em estruturas de introns (linhas azuis) e exons (barras vermelhas) presentes na mesma sequência obtidos na comparação com sequências gênicas de *M. truncatula* pela ferramenta FGENESH; Sobreposição dos dois resultados na mesma escala; Gráficos produzidos pelo *software* Artemis.

O anexo 2 mostra uma tabela contendo os conteúdos gênicos preditos em ordem idêntica nos genomas A e B, para cada uma das quatro regiões analisadas. Todos os genes foram anotados como putativos e apresentaram diversas funções, tais como fatores de transcrição, genes ligados a elementos de transposição, além de outros. O número de identificação dos genes (pfam-ID) está descrito e aqueles presentes em apenas um dos genomas estão evidenciados em negrito. Todos os resultados envolvendo a predição da localização e função de genes, assim como da presença de retrotransposons LTR e seus fragmentos para as quatro regiões hoemeólogas selecionadas dos genomas de *A. duranensis* e *A. ipaënsis,* estão ilustrados nas figuras 62 a 65, por meio de desenhos esquemáticos e análise de microssintenia entre essas regiões obtidos em gráficos de plotagem, na mesma escala em pares de bases. A maioria das falhas presentes nas diagonais foi caracterizada pela inserção de retrotransposons LTR ou pela presença de fragmentos ou genes relacionados a TEs. A diferença entre os conteúdos gênicos deu-se majoritariamente pela inserção de TEs, o que possivelmente foi responsável pela quebra de sintenia dessas regiões em ambos os genomas, dentre outros fatores.

Na Região 1 foram preditos 30 genes para o genoma A e 33 para B. Dos três genes identificados somente no genoma B, dois não tinham função descrita e um foi predito como sendo o gene que codifica a TR, presente em TEs autônomos, indicando que nesta localização possivelmente havia uma sequência remanescente derivada de um retrotransposon. De acordo com esses dados, 100% do conteúdo gênico do genoma A é similar ao do genoma B. Em contrapartida, apenas 91% do conteúdo gênico presente no genoma B é similar ao de A. Foram detectadas inserções de três elementos com sequência completa no genoma B: Doros (data de transposição estimada em 613.000 anos), Grilo (2,5 milhões de anos) e Mico (571.000 anos). Já no genoma A, foram detectados dois elementos insercionais, RE128 (1,7 milhão de anos) e Mico (1,47 milhão de anos). O padrão de linhas paralelas à diagonal no centro do gráfico foi caracterizado pela presença de três genes em *tandem* (figura 62).

Na Região 2 (figura 63) foram preditos 37 genes no genoma A e 34 no B. Um total de 67,5% do conteúdo gênico ou 25 genes do genoma A são similares ao do genoma B, e da mesma forma, 26 genes ou 76,5% do conteúdo gênico do genoma B é similar ao A. A maioria dos genes presentes em apenas um dos genomas foi predito como putativo e sem função relacionada. No genoma B foi identificada uma sequência remanescente derivada do elemento Bela. Já no genoma A, houve a inserção de dois outros elementos, Nemesis (data de transposição estimada em 1,5 milhões de anos) e Yara (não foi possível estimar a data de

transposição). Devido à presença de alguns genes relacionados a transposons (provavelmente derivados de elementos antigos que sofreram mutações em suas sequências), foram observadas falhas na diagonal, além de falhas no sequenciamento (*gaps*). Nove genes putativos e sem função conhecida estavam presentes apenas no genoma A e foram responsáveis pela maior diferença entre essas regiões dos dois genomas .

Na Região 3 do genoma A foram preditos 34 genes, enquanto que para o B, 35 genes. Um total de 28 genes ou 82,3% do conteúdo gênico do genoma A é similar ao do B, e da mesma forma, 28 genes ou 80% do conteúdo gênico do genoma B é similar ao do A. No genoma B foram identificadas duas sequências remanescentes dos elementos Matita e Vipe, além do elemento completo Doris (data estimada em 483.000 anos). No genoma A houve a inserção de dois elementos, Vipe (2 milhões de anos) e Venon (1,5 milhões de anos), além da presença de um fragmento do elemento Matita. Devido principalmente à presença de alguns retrotransposons, além de genes relacionados a TEs (provavelmente derivados de elementos degradados), foram detectadas falhas na diagonal. Diagonais paralelas, representadas pela ocorrência de genes que codificam a proteína kinase presentes nos dois genomas e organizados em *tandem*, também foram observadas (figura 64).

Por último, na Região 4 foram identificados 32 genes em A e 38 em B. Um total de 26 genes ou 81,2% do conteúdo gênico do genoma A é similar ao do genoma B, e 28 genes ou 73,7% do conteúdo gênico do genoma B é similar ao de A. No genoma B foi observado um elemento Paco com sequência completa (data de transposição estimada em 2,1 milhões de anos) e uma sequência remanescente do mesmo elemento. No genoma A houve a inserção de dois elementos, Agnus (452.000 anos) e Kyra (1,5 milhão de anos) além da presença de três fragmentos derivados do elemento Paco e outro do elemento Mico. Falhas nas sequências também ocorreram devido à presença de alguns genes relacionados a TEs (derivados de elementos antigos) bem como no sequenciamento (figura 65). Um fato interessante que ocorreu em todas as quatro regiões analisadas foi a presença dos genes relacionados a TEs (TR e MULE – *Mutator-like Elements*) que estavam presentes em ambos os genomas e distribuídos na mesma ordem (sintênicos), sugerindo que elementos teriam se inserido há mais de 3,5 milhões de anos atrás, antes da divergência entre os genomas A e B, restando apenas fragmentos, em virtude das mutações ocorridas ao longo da evolução. Nos gráficos foram observadas a macro e microssintenia entre quatro regiões hoemeólogas nos genomas A e B. A inserção de retrotransposons LTR em diferentes locais das sequências ocasionou a maioria das diferenças observadas entre a colinearidade dos genes nesses genomas.

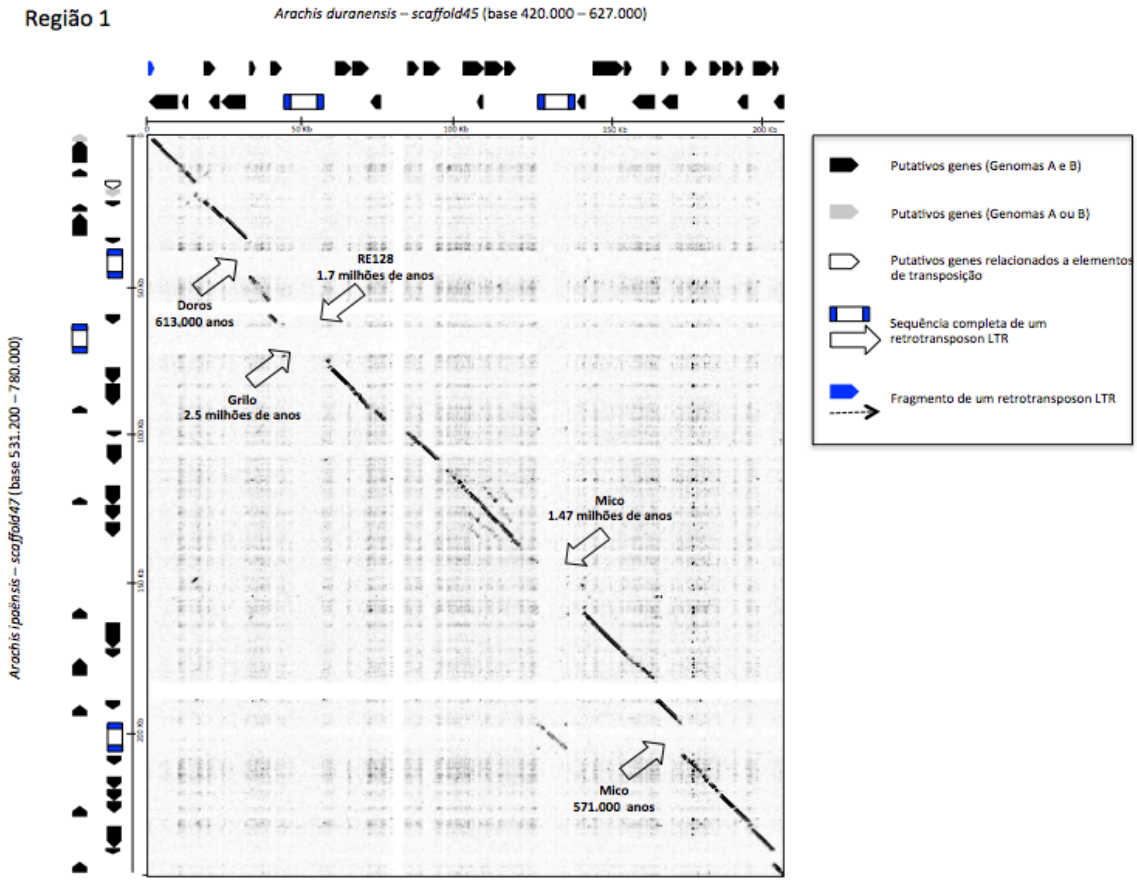


Figura 62: Comparação entre as sequências genômicas de *A. duranensis* e *A. ipaënsis* (Região 1).

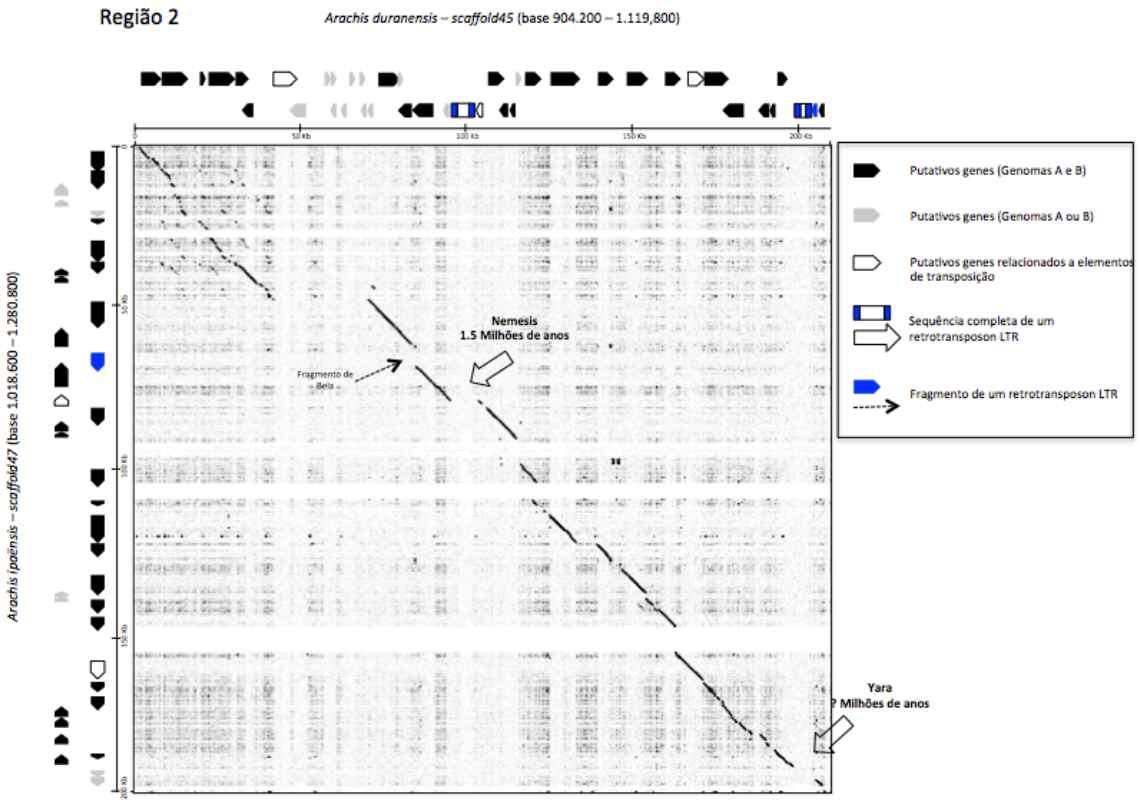


Figura 63: Comparação entre as sequências genômicas de *A. duranensis* e *A. ipaënsis* (Região 2).

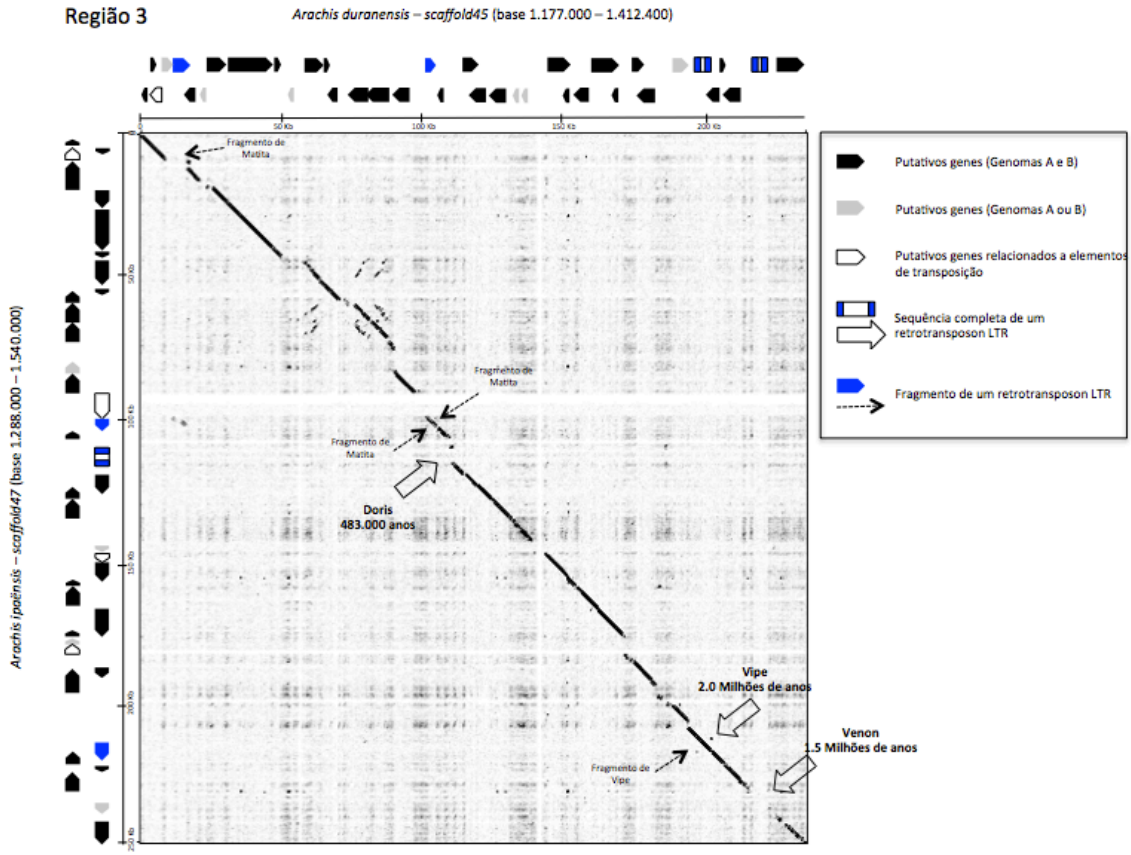


Figura 64: Comparação entre as sequências genômicas de *A. duranensis* e *A. ipaënsis* (Região 3).

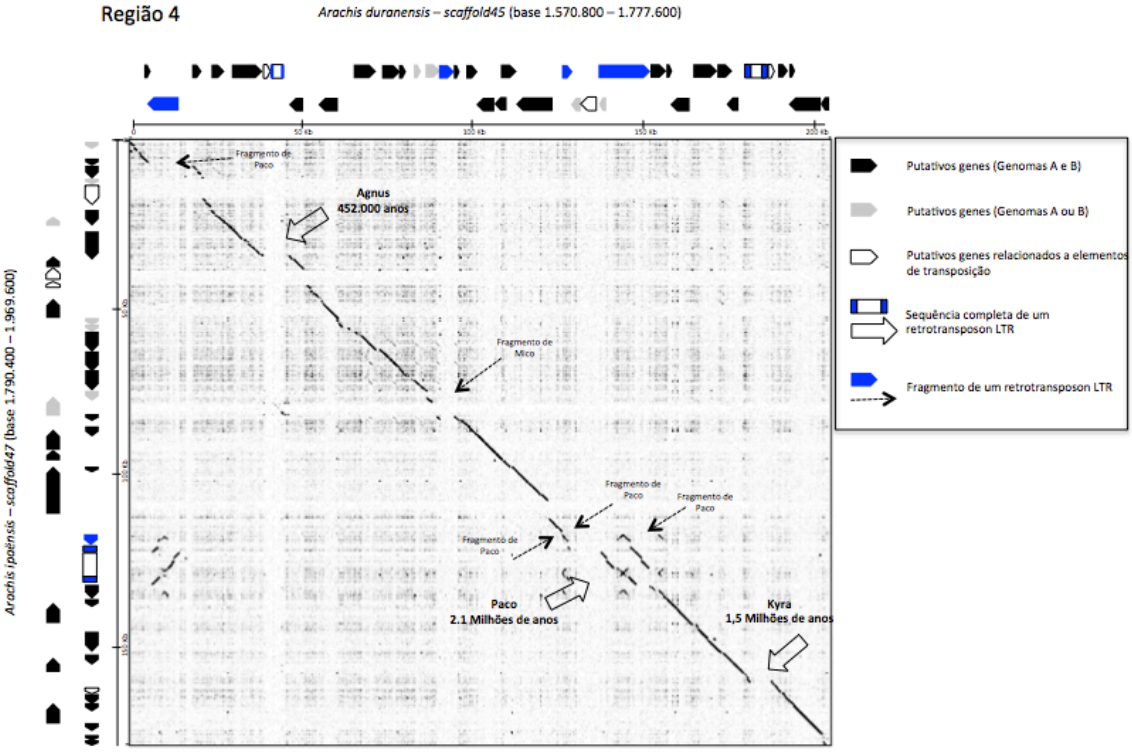


Figura 65: Comparação entre as sequências genômicas de *A. duranensis* e *A. ipaënsis* (Região 4).

4. Discussão

O recente sequenciamento de genomas completos de plantas vem contribuindo bastante para uma melhor compreensão do impacto causado nas sequências genômicas devido à presença maciça de elementos de transposição (TEs). Apesar dos TEs gerarem ambiguidades durante alinhamentos de sequências de DNA, que frequentemente prejudicam a montagem e anotação precisa de sequências de genomas complexos, estes não podem ser ignorados, pois entender sua dinâmica auxilia na compreensão da evolução de uma espécie, e muitas vezes da regulação gênica. Os TEs representantes da Classe I, denominados retrotransposons LTR, são predominantes em genomas de plantas (SanMiguel & Bennetzen, 1998). O equilíbrio entre a proliferação e diminuição desses retrotransposons LTR, se dá por meio de novas inserções e também pela eliminação completa ou parcial dessas sequências. Os mecanismos que permitem a eliminação parcial via recombinação dão origem a LTRs-solo ou fragmentos. Esses processos impulsionam grandes variações no tamanho do genoma, bem como na sua organização (Bennetzen & Kellogg, 1997; Vitte & Panaud, 2005; Sabot & Schulman; 2006; Tenaillon *et al*., 2010).

Dentro do gênero *Arachis*, foram caracterizados transposons de DNA do tipo MITE (Shirasawa *et al*., 2012) e dez retrotransposons LTR, incluindo FIDEL e Matita (Nielen *et al*., 2010; 2012; Bertioli *et al*. 2013). Esses estudos forneceram uma nova visão da organização do genoma de *Arachis*, principalmente da importância dos TEs na evolução, estrutura genômica e expressão gênica das espécies silvestres e cultivada. Utilizando dados obtidos do sequenciamento de clones das bibliotecas BAC, BAC ENDs, mapas genéticos, informações citogenéticas e outras informações disponíveis para *Arachis*, juntamente com esforços da colaboração interinstitucional – IPGI, as primeiras sequências genômicas completas das espécies parentais do amendoim, *A. duranensis* e *A. ipaënsis*, foram disponibilizadas. As duas sequências genômicas serão utilizadas como arcabouços para a montagem do genoma tetraploide do amendoim.

O presente trabalho é considerado pioneiro na investigação do componente repetitivo dos genomas completos de *A. duranensis e A. ipaënsis*, pois, anteriormente, a identificação desses elementos estava restrita a regiões específicas do genoma. Neste capítulo, o conteúdo altamente repetitivo, mais especificamente os retrotransposons LTR, dos genomas de *A. duranensis* e *A. ipaënsis* foi explorado. As coordenadas dos retrotransposons LTR foram identificadas pelo *software* LTR_FINDER. Para a obtenção dessas sequências em formato

FASTA, separação e alinhamento dos LTRs 5' e 3' dos elementos, estimativas das datas de transposição, formatação de arquivos para serem utilizados em diferentes *softwares*, diferentes *scripts Perl* foram desenvolvidos para auxiliar em cada etapa. O desenvolvimento desses *scripts* proporcionou uma avaliação rápida e precisa dos dados obtidos em *Arachis*. Além disso, esses *scripts* podem também serem utilizados em análises de sequências de outras espécies. Para tal, faz-se necessário que a identificação dos retrotransposons LTR seja realizada pelo *software* LTR_FINDER e que as sequências em FASTA sejam designadas pelo título "scaffold_x". Apesar da utilização de *scripts* ter facilitado imensamente o trabalho, ainda foram necessárias algumas análises feitas manualmente, assim garantindo maior precisão e detalhamento das informações. A compilação/comparação dos resultados obtidos pelos diferentes *softwares* também foi uma excelente forma de facilitar a compreensão dos resultados finais aqui alcançados.

Os retrotransposons LTR identificados neste estudo foram classificados em superfamílias com base no método descrito por Wicker *et al*. (2007), que compara as sequências dos elementos com sequências de proteínas disponíveis em bancos de dados. A ordem entre genes definiu se os elementos pertenciam à superfamília *Ty1* ou *Ty3*. Para a classificação em família, foram utilizadas apenas as similaridades entre as sequências nucleotídicas dos elementos, avaliadas por meio de *dot plots*, e não por alinhamento de sequências. A utilização de comparações por gráfico de pontos foi mais eficaz do que o alinhamento entre sequências, principalmente devido à baixa conservação de nucleotídeos entre as sequências dos retroelementos. Além disso, os gráficos permitem analisar um maior volume de dados. Portanto, uma família foi definida como um grupo de retrotransposons LTR que possuem alta similaridade na sequência de DNA relativa à região codificadora (se houvesse) ou no domínio interno, ou ainda nas regiões de repetições terminais (Wicker *et al*. 2007). As famílias foram também designadas como autônomas ou não-autônomas, com base na presença do gene que codifica a enzima transcriptase reversa.

Foram identificadas 49 famílias de retrotransposons LTR na sequência genômica parcial de *A. ipaënsis* (800 Mb), utilizando o método de agrupamento entre sequências de retrotransposons LTR via *softwares* Biolayout Express 3-D e CAP3. Por meio de *dot plots,* outras 40 famílias foram identificadas na sequência genômica completa (1,5 Gb) dessa espécie, totalizando assim, 89 novas famílias de retrotransposons LTR, incluindo 10 já caracterizadas anteriormente (Nielen *et al*., 2010; 2012; Bertioli *et al*., 2013). Essas famílias compreendem 16.659 sequências completas de retrotransposons LTR, distribuídas em 24

famílias de retrotransposons LTR não-autônomos e 65 autônomos, sendo 32 pertencentes à superfamília *Ty3-Gypsy* e 33 à superfamília *Ty1-Copia*.

Por meio de comparação com as sequências identificadas em *A. ipaënsis*, 81 famílias foram identificadas em *A. duranensis*. As 18.839 sequências completas de retrotransposons LTR identificadas em *A. duranensis* foram classificadas em 23 famílias compostas por retrotransposons LTR não-autônomos e 58 famílias por autônomos, sendo 28 pertencentes à superfamília *Ty3-Gypsy* e 30 à *Ty1-Copia*. No genoma de *A. duranensis* foram detectados apenas fragmentos de retrotransposons LTR de oito famílias contendo elementos completos identificadas somente no genoma de *A. ipaënsis*. A ausência desses oito retroelementos completos no genoma A ocorreu possivelmente devido ao descarte de sequências repetitivas durante a montagem do genoma A, o que também pode ser considerado uma explicação para a identificação de alguns elementos somente na sequência parcial e não completa dos genomas.

Para cada um desses genomas, onze famílias de retrotransposons LTR autônomos e seus potenciais pares não-autônomos foram descritas igualmente. Comparações revelaram que as sequências dos elementos autônomos sempre apresentavam-se maior do que seus pares não-autônomos. A similaridade entre pares autônomos/não-autônomos ocorreu principalmente nas regiões flanqueadoras ou LTRs, em acordo com o descrito para pares FIDEL/Feral e Pipoka/Pipa (Bertioli *et al*., 2013), mas também em partes da região gênica e regiões 5' e 3' UTR (não-traduzida). Embora não se possa atribuir qualquer significado biológico para esses elementos não-autônomos, é evidente que alguns desses elementos podem ser derivados de seus pares autônomos por meio da degradação mutacional com exclusão parcial de sequências, tornando-se "parasitas", enquanto outros, provavelmente são derivados de mecanismos complexos ou atípicos, tornando-os elementos não-autônomos.

Muitos TEs tendem a se inserir dentro da sequência de outros TEs (*nested transposons*), mas a causa e a importância evolutiva dessas formações ainda não foram totalmente elucidadas. No trabalho de Gao e colaboradores (2012) discutiu-se que a ocorrência de *nesteds* em genomas eucarióticos poderia influenciar negativamente a expansão do genoma, sendo então uma forma de controle do número de transposons e consequentemente, do tamanho do genoma. Para quase todas as famílias de retrotransposons LTR aqui descritas, *nested transposons* foram identificados. A maior parte dos elementos envolvidos nessas formações são das famílias FIDEL/Feral, Gordo, RE128 e Hemera. A observação dessas estruturas e a busca por maior entendimento sobre seus efeitos induzem o desenvolvimento de

vários trabalhos recentes, tais como, Zhao & Jiang (2014), que comparando TEs de arroz e milho, mostraram a ocorrência de *nesteds* associados a MULEs (*Mutator-like Transposable Elements*) que são mais frequentes em arroz, apesar de serem recentes (0 e 1 Ma) em ambos os genomas. Outros estudos mostram a abundância de *nesteds* no genoma de *Brassica* (Wei *et al*., 2013), cebola e aspargos (Vitte *et al*., 2013).

A evidência de que os conteúdos repetitivos presentes nas sequências genômicas de *A. duranensis* e *A. ipaënsis* têm divergido rapidamente foi aqui corroborada pela obtenção da estimativa da data dos eventos de transposição dos retrotransposons LTR, os quais foram inferiores a 3,5 milhões de anos, data estimada da divergência evolutiva entre os genomas A e B (Nielen *et al*., 2012; Moretzsohn *et al.,* 2013). As médias das datas dos eventos de transposição de todos os retroelementos analisados foi 1,55 e 1,51 Ma, nos genomas A e B, respectivamente. A média inferior à data de divergência entre os genomas A e B indica a ocorrência de atividade recente desses retroelementos, junto com o impacto diferente, em cada genoma, causado por eles, como discutido por Bertioli *et al*. (2013). Quando a data de inserção de um TE é inferior a 3 Ma, torna-se mais fácil estimar a sua data de transposição nos genomas de plantas. Um elemento mais antigo (> 3 Ma) tende a ser degradado por mutações ou eliminado por meio de mecanismos do tipo crossing-over desigual ou por recombinação ilegítima (Wicker & Keller, 2007; Devos *et al*., 2012). Por outro lado, durante a montagem das sequências para obtenção de genomas completos, provavelmente muitas sequências repetitivas, incluído retrotransposons LTR recentes e, muito frequentes, comumente são descartadas.

Nas análises realizadas por Bertioli *et al*. (2013) em clones BAC de *A. duranensis*, aproximadamente 30% das sequências eram compostas por poucas famílias de retrotransposons LTR. Neste trabalho, no qual o genoma completo de *A. duranensis* foi utilizado, aproximadamente 28,5% está ocupado por 81 famílias conhecidas de retrotransposons LTR, seus fragmentos e LTRs-solo. Já em *A. ipaënsis*, 89 famílias de retrotransposons juntamente com seus fragmentos e LTRs-solo estão distribuídos em 27,6% da sequência genômica completa. A frequência desses retrotransposons LTR é similar àquela estimada para os clones BAC do genoma A (Bertioli *et al*., 2013), mostrando a eficiência para delinear a frequência e os principais tipos de retroelementos presentes. Além disso, os genomas de *A. duranensis* e *A. ipaënsis* apresentam elementos similares, que juntos, ocupam proporções semelhantes nos seus respectivos genomas hospedeiros. Apenas 37 famílias representam a maior parte dos retrotransposons LTR, tanto no genoma A, quanto no B. No

entanto, FIDEL, juntamente com Feral (par não-autônomo), apesar de serem os mais abundantes nos dois genomas, são mais fequentes em A (11,02%) do que em B (7,85%). O mesmo foi observado em outras famílias de retrotransposons, cujas frequências eram distintas entre os genomas. Portanto, apesar dos dois genomas terem aproximadamente 30% das suas sequências correspondendo a retrotransposons LTR, a frequência de cada família é variável entre os genomas A e B.

As sequências dos retrotransposons da mesma família apresentaram sequências altamente similares, mesmo estando em genomas diferentes (A ou B). Portanto, pode-se deduzir que a principal diferença entre essas sequências repetitivas, não é a sequência nucleotídica, e sim a frequência em que essas famílias ocorrem nesses genomas, possivelmente resultante da atividade de transposição diferenciada em cada um deles.

Dentre as famílias descritas, 36 e 39 delas apresentaram menos do que dez cópias nos genomas A e B, enquanto as demais famílias apresentaram entre 10 e 11.323 cópias. Como visto para os genomas de *Arachis* aqui estudados, é bastante comum que um número relativamente pequeno de famílias de retrotransposons LTR esteja presente majoritariamente nos genomas de plantas (Wicker *et al.*, 2007). Na literatura está relatado que apenas três famílias de TEs são responsáveis pela duplicação do tamanho do genoma de *Oryza australiensis* nos últimos 3 Ma (Piegu *et al*., 2006); que o genoma de cevada possui 10% da sua sequência total ocupada por elementos BARE1 (Soleimani *et al*., 2006) e que aproximadamente 23% do genoma de linhaça é coberto por TEs, sendo que 17,2% é composto exclusivamente por retrotransposons LTR (González & Deyholos, 2012). Neste último trabalho, envolvendo o genoma de linhaça, a maioria das famílias de TEs possuem de 1 a 10 cópias e menos do que 50 famílias possuem mais do que 20 cópias, e ainda que o número de elementos *Ty1-Copia* aumentou nos últimos 5 Ma, diferentemente dos demais, em torno de 7-8 Ma.

Ao decompor *in silico* os genomas de *A. duranensis* e *A. ipaënsis* em suas respectivas pseudomoléculas, a frequência das famílias de retrotransposons LTR variou entre 17-40% em cada uma. A estimativa de porcentagem de retrotransposons LTR gerada para os genomas e pseudomoléculas pode ser maior para ambos, tendo em vista o tipo de análise e parâmetros utilizados para esse cálculo. Frequentemente sequências em *tandem* ou de baixa complexidade podem alterar esses valores. Certamente, outras famílias de retrotransposons LTR, não-LTR, transposons de DNA e DNA repetitivo de outras classes ainda não identificadas também estão presentes nas sequências genômicas de *A. duranensis* e *A.*

ipaënsis, o que pressupõe uma participação ainda maior no componente desses genomas, podendo alcançar até 64%, como previsto por Dhillon *et al*. (1980) por meio de estudos envolvendo a cinética de renaturação das fitas de ácido nucleico.

Análises de algumas famílias de retrotransposons LTR nas pseudomoléculas A01 e B01 de *A. duranensis* e *A. ipaënsis*, respectivamente, mostraram que, para famílias mais frequentes tais como Silverio, Gordo, Athena, Apolo, Pipoka e FIDEL/Feral a distribuição dessas famílias está restrita à região pericentromérica. Os resultados obtidos para a família FIDEL nas pseudomoléculas dos parentais silvestres são compatíveis com o padrão de ocorrência desse elemento em cromossomos metafásicos de amendoim (Nielen *et al*., 2010). Famílias como RE128-84, Mico e Matita, por exemplo, apresentaram uma distribuição preferencial em regiões distais das pseudomoléculas. Para Matita, esse resultado em espécies diploides é também compatível com a distribuição descrita para esse elemento em cromossomos metafásicos de amendoim (Nielen *et al*., 2012), indicando que essa abordagem *in silico* foi informativa. Famílias menos abundantes, tais como Venus , Saturno, Foxtrot, Juliett, Girino, Golden, Diva, Hermes e Hotel apresentaram distribuição dispersa e pouco frequente nas regiões centroméricas das pseudomoléculas A01 e B01. Para melhor compreender a distribuição dos elementos menos abundantes, a realização de FISH nos genomas das espécies diploides, separadamente, é essencial, para assim estabelecer uma relação mais específica entre os resultados *in silico* e por hibridização *in situ* nos cromossomos das espécies parentais diploides e análises com amendoim.

Logo, para as famílias de FIDEL e Matita, foi possível inferir que a distribuição nos cromossomos das espécies parentais *A. duranensis* e *A. ipaënsis* é similar àquela nos subgenomas A e B de amendoim. A maior frequência de FIDEL no genoma A, foi também observada no subgenoma A de amendoim, ao passo que a frequência de Matita é similar àquela nos genomas de *A. duranensis*, *A. ipaënsis* e nos subgenomas A e B do amendoim, indicando que mesmo após a hibridização entre esses dois genomas para a formação do genoma do amendoim, a proporção e distribuição desses dois elementos mantiveram-se semelhantes.

Análises da distribuição das famílias de retrotransposons LTR nos subgenomas de amendoim por FISH, mostraram sinais múltiplos de hibridização e dispersos em vários, mas não em todos os cromossomos do amendoim. Todas as famílias apresentaram marcação predominantemente na região pericentromérica dos cromossomos. A hibridização preferencial

em região pericentromérica foi também observada para elementos *Ty1-Copia* e *Ty3-Gypsy* em cromossomos de outras espécies, tal como em quinoa (Kolano *et al*., 2013).

A comparação entre regiões genômicas homeólogas nos genomas de *A. duranensis* e *A. ipaënsis* mostrou que o conteúdo gênico predito é bastante similar, com aproximadamente 70-100% de identidade. As principais diferenças seriam devido à identificação de genes putativos distintos, falhas no sequenciamento, e sobretudo ocorrência de retrotransposons LTR, fragmentos e LTRs-solo entre as sequências gênicas, ocasionando a quebra de sintenia. Esses dados corroboram o papel do DNA repetitivo na erosão de similaridade entre sequências desde a divergência dos genomas A e B, principalmente em regiões intergênicas (Bertioli *et al*., 2013). Portanto, conclui-se que os genes e sua ordem permanecem altamente conservados entre os genomas de *A. duranensis* e *A. ipaënsis*, e que a maior diferença na organização desses genomas está, possivelmente, no conteúdo repetitivo.

A anotação de TEs em sequências genômicas deve ser realizada como um esforço contínuo (Janicki *et al*., 2011). Espera-se que para o gênero *Arachis*, o banco de dados de retrotransposons LTR aqui identificados nas sequências genômicas de *A. duranensis* e *A. ipaënsis*, auxilie nos estudos da estrutura genômica do amendoim, bem como na anotação de seus genes. A identificação de outros tipos de TEs também se faz necessária, e a união de todos esses conhecimentos certamente auxiliará no entendimento do papel da enorme fração de sequências repetitivas no genoma de amendoim.

O estudo das características específicas dos TEs, juntamente com a frequência e dinâmica da sua distribuição no genoma, geram desafios para a genômica comparativa. Portanto, esse estudo pode ser considerado pioneiro em *Arachis*, pois possibilitou acessar o conhecimento acerca da semelhança, frequência e distribuição de várias famílias de TEs nos genomas A e B das espécies parentais do amendoim, discutindo a relação entre esses e fornecendo bases para a intensificação desses estudos.

5. Conclusão

Neste estudo foi demonstrado que uma proporção substancial do componente altamente repetitivo dos genomas completos das espécies progenitoras do amendoim, *A. duranensis* (genoma A) e *A. ipaënsis* (genoma B) está representada por 37 famílias de retrotransposons LTR. Os elementos mais abundantes são FIDEL e Feral, sendo mais frequentes no genoma A.

Foi mostrado, também, que esses elementos possuem predominantemente uma origem evolutiva recente, com data de transposição posterior àquela estimada para os genomas A e B do amendoim. Claramente, esses elementos contribuíram de forma notável para a divergência desses genomas, pois a frequência de ocorrência da maioria das famílias é distinta nos genomas.

Análises da distribuição dos retrotransposons LTR por FISH mostraram sinais múltiplos e dispersos em vários, mas não em todos os cromossomos de amendoim.

Sequências hoemeólogas dos genomas A e B são altamente semelhantes (macrossintenia). As falhas observadas na microssintenia podem ser explicadas pela presença de retrotransposons LTR e seus fragmentos, quase sempre distintos em sua natureza e localização, quando comparando-se os genomas A e B, corroborando a ideia de que os retrotransposons LTR identificados neste estudo, juntamente com outros DNAs repetitivos têm desempenhado um papel importante na remodelação dos genomas, especialmente em regiões intergênicas, ao longo do tempo evolutivo.

Construção de BAC-*pools* de duas espécies silvestres de amendoim para identificação e isolamento de genes de interesse

1. Introdução

Bibliotecas BAC (*Bacterial Artificial Chromosome*) são ferramentas importantes para o estudo detalhado de grandes regiões genômicas. Podem ser úteis na identificação e caracterização de sequências contendo genes de interesse, na correlação entre mapas genéticos e físicos (Yim *et al*., 2007), servindo como base para a genômica comparativa, ou tradicionalmente como primeiro passo na geração de plataformas para projetos de sequenciamento do genoma em larga escala (Warren *et al*., 2006).

Desde a sua concepção (Shizuya *et al*., 1992), a clonagem em vetores do tipo BAC foi amplamente utilizada como um sistema de clonagem padrão para muitas plantas. Neste tipo de biblioteca, cada clone é armazenado de forma individual e ordenada, tornando-se um valioso instrumento de pesquisa da genética moderna. A identificação de clones BAC específicos dentro de uma biblioteca, muitas vezes conhecida pelo termo inglês "*screening*" é realizada por diversas estratégias, envolvendo duas técnicas principais: hibridização com sondas de interesse ou a Reação em Cadeia da Polimerase (PCR - *Polimerase Chain Reaction*) (Campbell & Choy, 2002; Xia *et al*., 2009).

Inicialmente, a técnica mais utilizada baseava-se na hibridização de ácidos nucleicos imobilizados em filtros de alta densidade com sondas de interesse marcadas de forma radioativa ou fluorescente para *screening* de bibliotecas BAC (Danesh *et al*., 1998; Meksem *et al*., 2000). Essas sondas podem ser construídas de fragmentos de DNA subclonados, produtos de PCR amplificados e até oligonucleotídeos de DNA. Uma desvantagem nesta abordagem é que dependendo da sonda utilizada, ela pode conter uma grande quantidade de elementos de repetição ou motivos conservados, o que pode aumentar a ocorrência de falsos positivos.

Mais recentenemente, a estratégia que tem sido utilizada para isolar clones BAC é a que utiliza reações de amplificação por PCR, pois trata-se de um método mais simples, rápido e sensível (Yim *et al*., 2007), além de não utilizar a marcação radioativa. A eficiência no *screening* baseado em PCR pode ser aperfeiçoada pela associação ou combinação de clones das bibliotecas de maneiras específicas (Barillot *et al*., 1991). Uma vez combinados, os clones portadores de sequências particulares podem ser localizados através da identificação de subconjuntos que contêm os marcadores correspondentes.

A preparação desses agrupamentos de clones BACs ou "*pools*" em duas ou mais dimensões é um pré-requisito para o aumento da eficiência na triagem baseada em PCR.

Esses *pools* consistem na junção de suspensões de colônias de bactérias pertencentes a diferentes clones BAC. As matrizes dimensionais dos *pools* são construídas com base em planos geométricos, tendo como orientação imaginária, os eixos ortogonais definidos principalmente pelo conjunto de diferentes placas, linhas e colunas. Essas matrizes podem ser simples, compostas de duas (2-D), três dimensões (3-D), ou mais complexas, compostas por seis ou mais dimensões (Yim *et al*., 2007; Xia *et al.,* 2009; Simková *et al*., 2011). No entanto, essas últimas geralmente são construídas utilizando-se a manipulação robótica.

Bibliotecas BAC foram desenvolvidas para diversas espécies de plantas importantes economicamente, tais como *Glycine max* (Danesh *et al.,* 1998), *Phaseolus vulgaris* (Vanhouten & MacKenzie, 1999), *Medicago truncatula* (Nam *et al*., 1999), *Lotus japonicus* (Kawasaki & Murakami, 2000), *Vigna radiata* (Miyagi *et al*., 2004), *Trifolium pratense* (Sato *et al*., 2005), *Gossypium arboreum* (Hu *et al*., 2010), *Coffea arabica* (Cação *et al*., 2013), dentre outras.

Dentro do gênero *Arachis*, a primeira biblioteca BAC foi desenvolvida para o amendoim cultivado (*Arachis hypogaea*) (Yüksel & Paterson, 2005), espécie alotetraploide que possui um genoma AABB com aproximadamente 2,8 Gb (Greilhuber, 2005). O amendoim teve sua origem pela hibridização de duas espécies silvestres seguida por uma duplicação cromossômica espontânea, resultando numa estreita base genética (Halward *et al*., 1991; Young *et al*, 1996; Kochert *et al*., 1996). Portanto, apesar da alta diversidade morfológica do amendoim cultivado, tanto o seu melhoramento genético quanto os estudos genômicos ainda são considerados complexos, em virtude da baixa diversidade genética e poliploidia.

Segundo estudos genéticos e de fertilidade, as espécies *A. duranensis* (genoma AA) e *A. ipaënsis* (genoma BB) são os prováveis ancestrais silvestres do amendoim (Kochert *et al*., 1996; Fávero *et al*., 2006; Seijo *et al*., 2007; Moretzohn *et al*., 2013). Dessa forma, estas espécies foram selecionadas para construção de duas bibliotecas BAC representativas dos dois genomas que compõem o amendoim tetraploide (Guimarães *et al*., 2008). Ambas as bibliotecas foram utilizadas para o isolamento de clones contendo genes de interesse através de hibridizações. Estes genes foram o RGA S1_A36 que é um análogo a genes de resistência (RGAs – *Resistance Genes Analogs*) que co-segrega com um QTL para resistência à mancha preta; e ainda Leg083, Leg128, Leg92, leg237, Leg242 e leg88 que funcionam como marcadores-âncora entre leguminosas e são genes de cópias únicas. Alguns desses marcadores foram incorporados aos mapas genéticos já construídos para *Arachis* (Moretzsohn *et al*.,

2009; Leal-Bertioli *et al*., 2009).

O estudo dos genomas diploides de *Arachis* é, portanto, atrativo, pois possibilita a decomposição do genoma do amendoim tetraploide, o que pode simplificar a construção de mapas genéticos e físicos, além de possibilitar o isolamento e a caracterização de alelos silvestres e a comparação de regiões ortólogas dos genomas AA e BB umas com as outras e com outras espécies de leguminosas.

1.1 Isolamento de novos genes de interesse em espécies silvestres de amendoim

1.1.2 Gene associado a estresse abiótico - Expansina

As plantas submetidas ao déficit hídrico podem ter seu crescimento afetado e, consequentemente, a sua produtividade limitada. A seca é considerada um dos maiores obstáculos para o desenvolvimento sustentável da agricultura, o que torna o estudo da resposta das plantas ao déficit hídrico extremamente importante para programas de melhoramento genético de plantas, bem como para a prospecção e identificação de genes envolvidos nesse tipo de tolerância.

Déficit hídrico pode ser definido como um desequilíbrio entre a disponibilidade de água no solo e a demanda de evapotranspiração que pode ocorrer naturalmente em campo (Tardieu *et al*., 2011), o que provoca a diminuição no acúmulo de carbono, diminuição na expansão de tecido e redução do número de células. Cada um desses processos macroscópicos envolve um grande número de genes, enzimas, hormônios e metabolitos, o que suporta a ideia de que existe uma interação entre estes processos e entre diferentes vias metabólicas (Skirycz *et al*., 2010; Tardieu *et al*., 2011).

A baixa disponibilidade de água pode causar alterações na parede celular e consequentemente alterar a organização de microfibrilas de celulose além de outros polissacarídeos, causando maior adesão entre eles (Moore *et al*., 2008). Além disso, em resposta à baixa disponibilidade de água, as plantas podem induzir genes a codificar determinadas proteínas que podem auxiliar contra esse estresse abiótico (Shinozaki & Yamagushi-Shinozaki, 1997). Genes que apresentam sua expressão modificada em resposta ao déficit hídrico, como o que codifica para a proteína expansina, já foram identificados em algumas espécies, como milho (Zhang & Hasenstein 2000), Arabidopsis (Seki *et al*., 2001), dentre outros. A expansina é uma proteína capaz de induzir a extensão e o relaxamento da

parede celular, rompendo as ligações não-covalentes entre os polissacarídeos que a compõem (Cosgrove, 2000; Sampedro *et al*., 2006). Essa extensão é importante para a célula vegetal, pois permite a redução do turgor celular e do potencial hídrico, capacitando-a a absorver água e a se expandir (Taiz & Zeiger, 2006).

No presente estudo, objetivou-se isolar clones BAC contendo o gene da expansina identificado como diferencialmente expresso a partir dos transcriptomas das espécies silvestres de amendoim *A. duranensis* (genoma AA) e *A. magna* (genoma BB) (Brasileiro *et al*., 2012). Esse gene foi validado por RT-qPCR e apresentou perfil de expressão compatível com as análises *in silico*, sendo este, portanto, um gene candidato envolvido na resposta destas plantas à seca.

1.1.3 Gene associado à biossíntese de óleo – Dessaturase de ácidos graxos (FAD)

O amendoim é uma oleaginosa importante nos mercados interno e externo, incluindo a indústria de alimentos e de óleos vegetais. Os ácidos oleico (O) e linoleico (L) compreendem cerca de 80% dos ácidos graxos (Norden *et al*., 1987) e variações na proporção O/L ocorrem alterando a estabilidade oxidativa do óleo, o que torna os óleos com maior razão O/L considerados mais estáveis e benéficos para a saúde.

Em virtude de o amendoim ser um alotetraploide, estudos já demonstraram que linhagens de amendoim com acúmulo de ácido oleico continham mutações nas duas cópias dos genes que codificam para D^{12} *fatty acid desaturase* (*FAD*2) (Jung *et al*., 2000a; 2000b). A ausência desta enzima impede a transformação de ácido oleico em ácido linoleico, levando ao seu acúmulo (Ray *et al*., 1993; Bruner *et al*., 2001; Yu *et al*., 2008). Regulações epistáticas ou contribuição de outros genes FAD2 também podem atuar na variação O/L (Isleib *et al*., 2006).

No presente estudo objetivou-se identificar genes completos relacionados à produção de óleo e regiões vicinais em duas espécies silvestres de *Arachis*, para caracterização de novos alelos selvagens e futuro desenvolvimento de marcadores moleculares para essas regiões. O desenvolvimento de um sistema de *pools* 3-D para as bibliotecas BAC construídas para as espécies silvestres de gênero *Arachis* (*A. duranensis* e *A. ipaënsis*) possibilitará o isolamento desses genes e constituirá uma importante ferramenta para o isolamento de outros genes de interesse agronômico que poderão ser prospectados e analisados de forma mais rápida.

2. Material e Métodos

2.1 Material vegetal

Os experimentos foram conduzidos no Laboratório de Interação Planta-praga III (LPPIII) e as plantas utilizadas neste trabalho foram fornecidas pelo Banco Ativo de Germoplasma (BAG) de espécies de *Arachis*, ambos situados na Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

2.2 Bibliotecas BAC

Em colaboração com a plataforma robótica do Centro de Cooperação Internacional em Pesquisa Agronômica para o Desenvolvimento (CIRAD – França), foram construídas duas bibliotecas BAC para cada uma das espécies diploides que deram origem ao amendoim, *A. duranensis* acesso V14167 (genoma AA) e *A. ipaënsis* acesso KG30076 (genoma BB). As bibliotecas AA e BB possuem, respectivamente, 84.096 clones (219 placas de 384 poços) e 75.648 clones (197 placas de 384 poços) e uma cobertura do genoma equivalente a 7,4x e 5,3x, com tamanhos médio de insertos de 110 e 100 kb (Kilobases) (Guimarães *et al*., 2008).

2.3 Manutenção e replicação das bibliotecas BAC

Inicialmente as bibliotecas originais foram duplicadas, a fim de se obter uma cópia de trabalho e manter as bibliotecas originais preservadas. Para tal, foram feitas cópias manuais a partir das bibliotecas originais que estavam mantidas em placas com 384 poços (Genetix) contendo 80 µL de meio de cultura 2YT (1,6% de Triptona; 0,5% de Extrato de Levedura; 0,5% de NaCl) contendo 7% de glicerol, estocadas em freezer -80°C.

Para a duplicação das bibliotecas utilizaram-se placas de 384 poços (Genetix), previamente identificadas e preenchidas com 100 µL de meio de cultura LB (Lúria Bertani – 1% de Triptona; 0,5% de Extrato de Levedura; 1% de NaCl), 4% de glicerol e antibiótico cloranfenicol (12,5 µg/mL). O preenchimento das placas foi realizado utilizando-se o equipamento Qfill2 *Microplate Dispenser* (Genetix) em fluxo laminar. As placas originais foram descongeladas a 4°C e ao final do experimento foram recolocadas imediatamente em freezer -80°C.

Para a duplicação manual das bibliotecas utilizou-se um replicador contendo 384 alfinetes (Genetix) cuja parte metálica foi embebida em álcool e posteriormente flambada por três vezes antes de ser inserida rapidamente na placa original já descongelada e depois na placa nova para duplicação. Após a inoculação, as novas placas foram seladas e armazenadas em estufa a 37° C por um período de16 horas e posteriormente estocadas em freezer -80°C.

2.4 Confecção dos *pools* 3-D

Para a confecção dos *pools* de BAC para as duas bibliotecas de *Arachis*, utilizou-se um método baseado no modelo em três dimensões proposto por Xia *et al*. (2009) com algumas modificações. As cópias das bibliotecas foram utilizadas para gerar os agrupamentos em 3-D. As matrizes dimensionais dos *pools* foram construídas com base nos planos geométricos dos eixos de um cubo em 3-D, ou seja, linhas (eixo y), colunas (eixo x) e as próprias placas (eixo z), como descrito na figura 66.

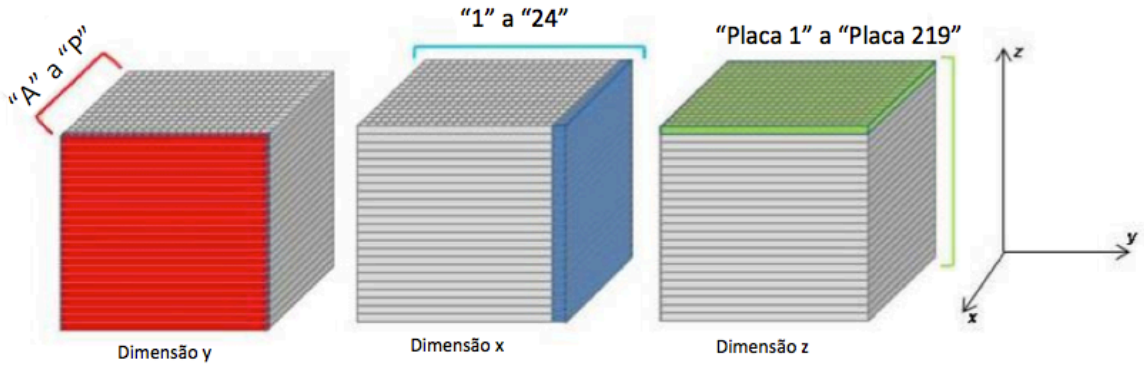


Figura 66: Planos geométricos relativos aos eixos de um cubo em 3-D. As matrizes tridimensionais foram utilizadas para construção de *pools* de BAC em linhas, colunas e placas.

Cada placa de 384 poços é composta por 16 linhas ("A" a "P") e 24 colunas (1 a 24), com o poço A1 localizado no canto superior esquerdo da placa. Consequentemente, cada poço presente na matriz tridimensional possui um endereço único definido pelas suas coordenadas relativas aos eixos do cubo: x (número da coluna na placa), y (número da linha na placa), e z (número da placa). O conjunto originado pela junção da totalidade dos *pools* correspondentes aos três eixos ortogonais foi designado de "endereço ou coordenada do clone".

2.4.1 *Pools* de dimensão x

Os *pools* de dimensão x ou *pools* laterais são definidos pelo plano paralelo às superfícies esquerdas e direitas do cubo, agregando os clones de BAC que partilham a mesma coluna na placa. Cada *pool* x de uma placa contém 16 clones

2.4.2 *Pools* de dimensão y

Os *pools* de dimensão y correspondem ao plano paralelo à superfície frontal do cubo, consistindo nos clones BAC que partilham as mesmas linhas da placa. Cada *pool* representando o eixo y contém 24 clones BAC.

2.4.3 *Pools* de dimensão z

Cada placa de 384 poços é definida como um único *pool* de coordenada z. Cada *pool* de dimensão z contém 384 clones BAC. O número total de *pools* z correspondentes a cada uma das duas bibliotecas A e B é: 219 e 197, respectivamente.

2.5 Validação das cópias de trabalho das bibliotecas BAC

Para avaliar a eficácia na reprodutibilidade das placas, realizou-se um teste preliminar, no qual algumas placas aleatórias foram replicadas duas vezes (simulando as dimensões x e y). Foram selecionados clones (endereços aleatórios iguais em quatro placas – placa original, duplicada e duas cópias) para extração de DNA e análise do perfil de restrição para comprovação da identidade do clone, e da ausência de contaminação ou erro no procedimento de replicação (figura 67).

Os clones BAC selecionados foram extraídos pelo método de lise alcalina na forma de minipreparação (ver item 2.6) e fragmentados com enzima de restrição para visualização do perfil de restrição. Aproximadamente 100 ng do DNA de cada clone foi digerido com 1U de enzima de restrição *Eco*RI (Invitrogen) de acordo com o protocolo do fabricante e incubado a 37° C por 3 horas. O perfil de restrição foi visualizado em gel de agarose 1,0% corado com brometo de etídio e eletroforese de acordo com o item 2.9.

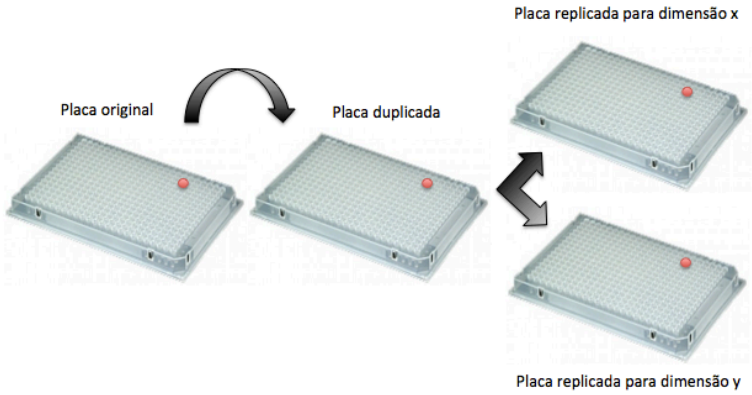


Figura 67: Esquema ilustrativo do teste realizado para avaliar a eficácia na reprodutibilidade das placas.

2.6 Confecção de *pools* 3-D

A localização dos clones de interesse, baseada na técnica de amplificação por PCR, foi realizada em dois momentos distintos: 1 - *Screening* de placa e 2 - *Screening* de coordenada. O resumo do experimento pode ser visualizado na figura 68.

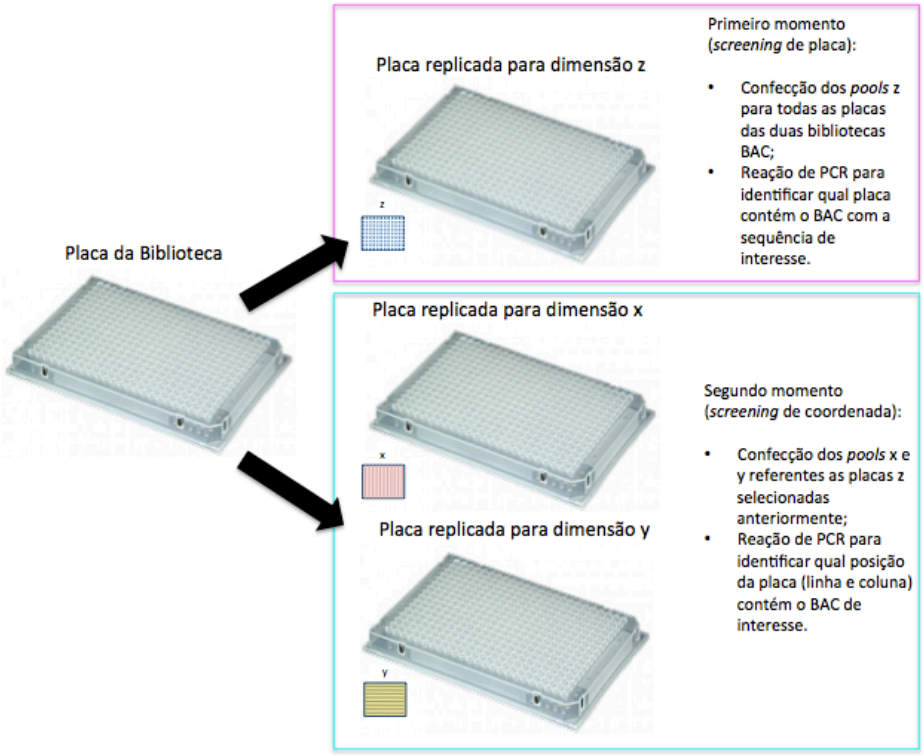


Figura 68: Esquema ilustrativo do método utilizado para localização dos clones de interesse baseado na técnica de amplificação por PCR. Primeiramente localiza-se a placa e depois a coordenada (linha e coluna) do clone de interesse.

Foram utilizadas placas estéreis com 384 poços (Genetix) para a confecção dos *pools* x,

y e z. Da mesma forma descrita anteriormente, novas placas foram identificadas e preenchidas com meio de cultura LB e antibiótico cloranfenicol (12,5 μg/mL). Para o preenchimento utilizou-se o equipamento Qfill2 *Microplate Dispenser* (Genetix) em fluxo laminar previamente esterilizado. As placas duplicadas das bibliotecas (cópias de trabalho) foram descongeladas a 4°C e ao final do experimento foram recolocadas imediatamente em freezer -80°C.

Para os *pools* z, a quantidade de meio LB utilizado em cada poço da placa foi 75 μL, pois ao serem coletadas as culturas de todos os poços da placa, o volume final foi de aproximadamente 28 mL para cada *pool* z. Para os *pools* x, a quantidade de meio LB em cada poço da placa foi 150 μL, pois ao coletar as culturas de todos os poços de cada coluna, o volume final foi de aproximadamente 2,4 mL (16 clones x 150 μL). Para os *pools* y, a quantidade de meio LB em cada poço da placa foi 100 μL, pois para a coleta das culturas de bactérias referentes aos poços de cada linha, o volume final foi de aproximadamente 2,4 mL (24 clones x 100 μL). Novamente utilizou-se um replicador contendo 384 alfinetes para replicar cada placa das bibliotecas. Após a inoculação, as novas placas foram seladas e armazenadas em estufa a 37° C por um período de 16 horas para então serem coletadas e realizar-se a extração de DNA.

Para os *pools* z, as culturas de bactérias foram coletadas e armazenadas em tubo de polipropileno de 50 mL, para que 2 mL fossem utilizados para extração de BAC e o excedente acrescido de 15% de glicerol e mantido em freezer -80°C. Para os *pools* x e y, um total de aproximadamente 2,4 mL das culturas de bactérias de cada coluna e linha das placas foram coletadas com auxílio de pipeta de oito canais (Eppendorf) e armazenadas em tubo de 15 mL, para serem utilizados para extração do cromossomo artificial bacteriano (BAC).

2.7 Isolamento de clones BAC

Foram utilizados dois métodos de isolamento de BAC neste estudo, os quais produziram quantidades variáveis de DNA. Para confecção dos *pools* de BAC, análises de restrição e reações preliminares de amplificação por PCR, utilizou-se o método de minipreparação de BAC, ao passo que, para as reações de PCR visando a confirmação do endereço dos clones nas placas originais, assim como para o sequenciamento, utilizou-se o método de midipreparação.

Para a minipreparação de BAC, utilizou-se o seguinte protocolo baseado na técnica de lise alcalina: a cultura de bactérias foi centrifugada a 13.000 rpm por 1 minuto. O sobrenadante foi descartado e o *pellet* ressuspenso em 200 µL de tampão a 4° C (glicose 50 mM; Tris-HCl 1M pH 8; EDTA 500 mM; 100 µg de RNA*se* A). A solução foi homogeneizada, acrescida de 200 µL de tampão de lise (NaOH 200 mM; SDS 1%) e misturada gentilmente por 2 minutos à temperatura ambiente. Foram adicionados 200 µL de tampão de neutralização (acetato de potássio 1,32 M, pH 4,8) e após 3 minutos no gelo, os resíduos foram removidos por centrifugação a 13.000 rpm por 15 minutos a 4° C e o sobrenadante transferido para novo tubo. Foram adicionados 400 µL de isopropanol, misturando-se gentilmente. As amostras foram então centrifugadas a 13.000 rpm por 20 minutos e o sobrenadante descartado. O precipitado foi ressuspenso em 100 µL de TE (Tris-HCl 10 mM; EDTA 1 mM), e acrescido de 10 µL de acetato de sódio 3 M (4° C) e 250 µL de etanol absoluto (-20° C) e mantidos em freezer -80° C durante 15 minutos. Seguiu-se centrifugação por 20 minutos a 4° C. O sobrenadante foi descartado e o precipitado foi lavado com etanol 70% gelado e seco completamente por 15 minutos à temperatura ambiente. O DNA foi ressuspenso em 30 µL de água *MilliQ* autoclavada e estocado em freezer a -20° C.

Para a midipreparação de BAC utilizou-se o *Kit Plasmid DNA Purification NucleoBond Xtra Midi* (Macherey-Nagel, 2011) com modificações. Para a pré-cultura foi inoculada uma colônia isolada da placa da biblioteca original, em 5 mL de meio LB e 5 µL do antibiótico cloranfenicol (12,5 mg/µL). A incubação foi realizada durante 8 horas a 37° C com rotação em torno de 300 rpm. A preparação da cultura foi feita utilizando 400 mL de meio LB, 400 µL do antibiótico cloranfenicol (12,5 mg/µL) e 400 µL da pré-cultura. A incubação foi realizada durante 16 horas a 37° C com rotação em torno de 300 rpm. As células foram centrifugadas a 4.000 rpm por 25 minutos a 4° C. O sobrenadante foi descartado e o *pellet* ressuspenso completamente em 16 mL de tampão de ressuspensão. A solução foi homogeneizada, acrescida de 16 mL de tampão de lise, misturada gentilmente por inversão e incubada a temperatura ambiente por 5 minutos. Durante esse tempo, a coluna contendo filtro foi montada e equilibrada, adicionando-se 12 mL de tampão de equilíbrio nas extremidades da coluna, certificando-se de que o filtro ficou igualmente úmido. Foram adicionados 16 mL de tampão de neutralização e imediatamente a solução foi homogeneizada por inversão e aplicada dentro do filtro equilibrado. Os resíduos foram descartados e o filtro foi lavado com 5 mL do tampão de equilíbrio. O filtro foi descartado e a coluna foi lavada com 8 mL do tampão de lavagem. Para eluição do DNA, foram aplicados 5 mL do tampão de eluição,

aquecido previamente a 50° C, no centro da coluna, e após passar pela coluna foi coletado em um tubo de 15 mL. Para precipitar o DNA, adicionaram-se 3,5 mL de isopropanol à temperatura ambiente e aplicou-se vortex vigorosamente. A solução foi centrifugada a 12.000 rpm por 45 minutos a 4° C e o sobrenadante foi cuidadosamente descartado. Para lavagem do precipitado foram adicionados 2 mL de etanol 70 % e seguiu-se uma centrifugação a 12.000 rpm por 45 minutos a temperatura ambiente. O precipitado foi seco completamente por 15 minutos à temperatura ambiente e ressuspenso em 200 µL de água *MilliQ* autoclavada e estocado em freezer a -20° C.

2.8 Extração de DNA genômico de plantas de *A. duranensis* e *A. ipaënsis*

Os controles positivos utilizados nas reações de PCR foram os DNAs dos genótipos das plantas utilizadas na construção das bibliotecas: *A. duranensis* (acesso V14167) e *A. ipaënsis* (acesso KG30076). Sementes foram germinadas e folhas jovens foram coletadas antes da expansão foliar total. O protocolo de extração de DNA foi baseado em CTAB (Brometo de Cetil Trimetil Amônio) (Ferreira & Grattapaglia (1998) com modificações.

Aproximadamente 200 mg de tecido vegetal foram macerados em nitrogênio líquido e acrescidos de 700 µL de CTAB 2% (CTAB 2% (p/v); NaCl a 1,4 M; Tris-HCl a 100 mM e pH 8,0; EDTA a 20 mM; β-mercaptoetanol 0,2% (v/v)) em tubos de polipropileno de 2 mL (*Eppendorf*). As amostras foram incubadas a 65° C por 50 minutos, em seguida acrescidas de 700 µL da solução clorofórmio-isoamílico (24:1) e misturadas até a formação de uma emulsão. As amostras foram centrifugadas a 13.200 rpm por 15 minutos e a fase aquosa foi transferida para dois novos tubos (1,5 mL). Foram adicionados 600 µL de tampão CTAB 1% (CTAB 1% (p/v); Tris-HCl a 50 mM e pH 8,0; EDTA a 20 mM) e os tubos foram agitados lentamente e centrifugados a 13.200 rpm por 1 minuto. O sobrenadante foi descartado e o precipitado dos dois tubos foi ressuspenso em 300 µL de NaCl a 1,2 M. O volume dos dois tubos foi transferido para um único tubo, totalizando 600 µL. As amostras foram centrifugadas a 12.000 rpm por 5 minutos e o sobrenadante foi transferido para novo tubo. O DNA foi precipitado com 1 mL de etanol absoluto sob agitação lenta. As amostras foram centrifugadas a 13.200 rpm por 2 minutos e o sobrenadante descartado. O precipitado foi lavado duas vezes com 500 µL de etanol 70%, ressuspenso em 100 µL de água estéril com 0,01 mg/mL de RNA*se*-A e incubado a 37° C por 10 minutos. A quantificação foi realizada em gel de agarose 1.0% utilizando um marcador de peso molecular (*High Mass Ladder* -

Invitrogen). Para utilização em reações de PCR, as amostras foram diluídas para a concentração de 5 ng/μL e mantidas a -20° C para realização do experimento.

2.9 Identificação de clones BAC por PCR

Foram utilizados dois pares de *primers* construídos para diferentes partes do gene FAD2 de amendoim, que apresenta baixo número de cópias no genoma: Leg045F/R (Hougaard *et al*., 2008) e FAD2BF/R (Chu *et al*., 2009). A utilização de dois pares de *primers* construídos para o mesmo gene permitiu a realização de contra-prova dos resultados obtidos. Para amplificar um gene da expansina de *Arachis* foi utilizado o par EXPUNIF/EXP464R, descritos previamente em Brasileiro *et al*. (2012) (tabela 11).

Tabela 11: Pares de *primers* e suas respectivas sequências nucleotídicas na direção 5' – 3'.

Nome do *primer*	Sequência 5'-3'
Leg045-F	CATGAGTGTGGCCACCATGC
Leg045-R	GCTCCTCTTAACCAGTCCCA
FAD2BF	GGAGCTTTAACAACACAA
FAD2BR	ATATGGGAGCATAAGGGT
EXPUNIF	ACTGCCAGTCACTTGGAACC
EXP464R	GCTATGGCGGAATGGATCT

Todas as reações de amplificação por PCR realizadas neste trabalho seguiram o mesmo protocolo. Para um volume de 25 μL, foram utilizados: 2,5 μL (10%) de tampão para PCR sem $MgCl_2$ (Invitrogen); 1 μL de $MgCl_2$ 50 mM (Invitrogen); 0,3 μL de cada *primer* a 10 μM; 0,3 μL de dNTPs a 10 mM, 0,15 μL de *Taq* Polymerase (*Taq Recombinant* Invitrogen) e 1 μL de DNA (na concentração de 100 ng/μL). O programa utilizado em termociclador (Eppendorf Gradient) foi: 5 minutos a 94° C; 35 ciclos de 30 segundos a 94° C, 30 segundos a 55-60° C (dependendo do par de *primers*), 30 segundos a 72° C; e 7 minutos a 72° C.

O tamanho do produto de PCR (*amplicon*) foi verificado por eletroforese em gel de agarose 1.0% corado com brometo de etídio 0,5 μg/mL em tampão TAE 1X (40 mM de Tris-acetato e 1 mM de EDTA pH 8,0). A eletroforese foi realizada em cuba BIO-RAD Sub-Cell GT e fonte BIO-RAD POWER PAC200 durante 40 minutos a uma voltagem constante de 70V. A visualização dos fragmentos foi realizada em foto-documentador ImageQuant300 (GE Healthcare Life Sciences). A análise de quantificação foi realizada em espectrofotômetro Nanodrop ND1000. A razão da absorbância a 260 nm e 280 nm (260/280) foi utilizada para medir a "pureza" do DNA, onde a taxa 1.8 é geralmente a mais aceitável. Uma razão

secundária, também utilizada para medir a pureza, foi a 260/230, onde valores abaixo de 2 indicam presença de contaminantes.

2.10 Sequenciamento dos produtos de PCR e clones BAC

O sequenciamento dos *amplicons* foi realizado pelo método Sanger (Sanger *et al*., 1977). Utilizou-se o Kit de sequenciamento *BigDye Terminator cycle Sequencing* (*Applied Biosystems*, CA, USA) em um sequenciador automático ABI377 (*Applied Biosystems*), seguindo as concentrações para amplificação de produto de PCR: 1,5 µL de *primer* (2 µM) complementares à sequência do produto; 2 µL de produto de PCR (aproximadamente 200 ng/µL); e 2 µL de *premix* em 10 µL de reação final por amostra. As amplificações foram feitas utilizando um termociclador nas seguintes condições: 30 ciclos com uma etapa de desnaturação (96° C por 20 segundos); anelamento (50° C por 15 segundos) e de extensão (x° C por 60 segundos – dependendo da temperatura ideal para cada *primer*). Já o sequenciamento dos clones BAC foi realizado nos EUA utilizando-se a plataforma Pacbio RS Pacific Biosciences (GATC Biotech AG, Konstanz, Alemanha) utilizando de 5 – 10 µg de DNA.

2.11 Análise dos dados de sequenciamento

Foi realizada uma comparação via BLASTx (*Basic Local Alignment and Search Tool*) (Altschul *et al*., 1997) das sequências dos produtos de PCR obtidos a partir da amplificação dos clones BAC com os *primers* desenvolvidos para os genes da expansina e FAD2 contra o banco de dados contendo sequências de proteínas não-redundantes (nr) do NCBI (http://www.ncbi.nlm.nih.gov).

Para os dados obtidos do sequenciamento total dos clones BAC pela técnica PACBIO, a remoção dos vetores foi realizada com o auxílio da ferramenta Vecscreen (http://www.ncbi.nlm.nih.gov/tools/vecscreen/). A predição da estrutura e localização dos genes (estruturas de introns e exons) foi realizada utilizando-se a ferramenta FGENESH (Solovyev *et al*., 2006) (http://www.softberry.com/) contra o banco de dados da leguminosa *Medicago truncatula*. Os resultados foram modificados para visualização em Artemis utilizando-se um *script Perl*. Com intuito de corroborar a localização dos genes putativos e realizar a predição da função, todas ORFs com tamanho maior do que 50 pb presentes nas

sequências foram identificadas pelo Artemis e confrontadas com o banco de dados do *pfamA*. Os resultados foram analisados em Artemis por meio de sobreposição com os dados encontrados no FGENESH. Alinhamentos entre sequências foram realizados pelo *software* Muscle (Edgar, 2004) e visualizados na interface do *software* Jalview (Waterhouse *et al*., 2009). Retrotransposons LTR foram identificados pelo LTR_FINDER e fragmentos de elementos, LTRs-solo e sequências em *tandem* foram identificados por meio de gráficos de plotagem (*dot plots*) obtidas no *software* Gepard.

3. Resultados

3.1 Validação do método de duplicação de placas

Para verificar a reprodutibilidade das bibliotecas de trabalho foram selecionados os clones oriundos da posição H12 (oitava linha denominada "H" e coluna "12") das placas 188, 198, 208 e 219 pertencentes à biblioteca de *A. duranensis*. Após isolados e digeridos enzimaticamente, os clones das placas originais, duplicadas e replicadas duplamente (rep 1 e rep 2) apresentaram o perfil de restrição idêntico, o que validou a técnica utilizada para replicação de placas, comprovando que o método de inoculação foi seguro e livre de contaminantes (figura 69).

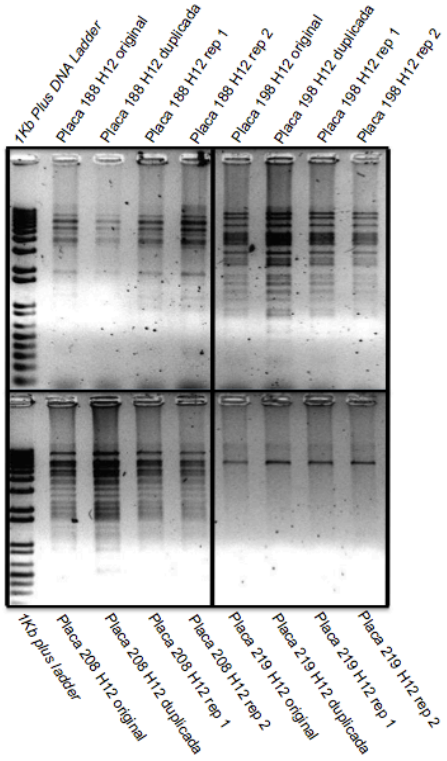


Figura 69: Análise do perfil de restrição de quatro clones BAC selecionados randomicamente da biblioteca A (clones 188H12, 198H12, 208H12 e 219H12). Os clones foram digeridos com a enzima *Eco*RI e os perfis visualizados em gel de agarose 1,0% corado com brometo de etídio. Marcador *1Kb Plus DNA Ladder* (Invitrogen) (poço 1).

3.2 Extração de DNA genômico e qualidade das amostras

O protocolo de extração de DNA utilizado produziu DNA com alta qualidade, como demonstrado na figura 70, que apresenta um gel de agarose contendo amostras em duplicata

dos DNAs genômicos de *A. ipaënsis* (acesso KG30076) e *A. duranensis* (acesso V14167), utilizados como controles positivos nas reações de PCR. A análise de quantificação com auxílio do espectrofotômetro mostrou que os valores da razão 260/280 para as quatro amostras de DNA foram aproximadamente de 1.79-1.84, e da razão 260/230 foram de 1.99-2.05. Os valores de quantificação para as amostras (na ordem em que aparecem no gel foram, respectivamente: 524,24 ng/µL, 741,98 ng/µL, 1112,04 ng/µL e 1080,11 ng/µL.

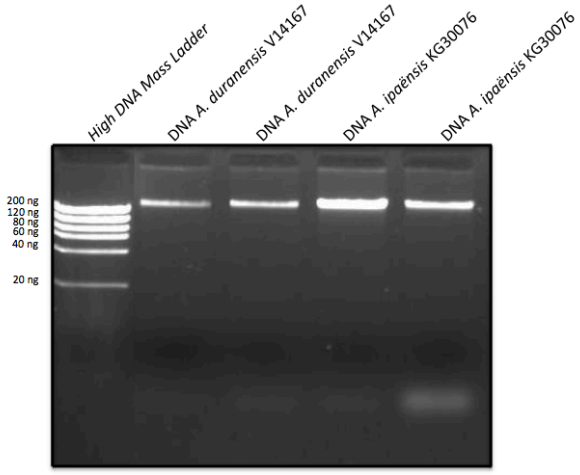


Figura 70: Análises de qualidade e concentração de amostras de DNA genômico de *A. duranensis* e *A. ipaënsis* em duplicata realizadas em gel de agarose 1,0% corado com brometo de etídio (poços 2 a 5). Marcador de peso molecular *High DNA Mass Ladder* (Invitrogen) (poço 1).

3.3 Confecção dos *pools* 3-D e identificação de clones de interesse

Primeiramente foi realizada a confecção dos *pools* z referentes às duas bibliotecas BAC representantes dos genoma A e B. Dois dias foram necessários para a inoculação das placas, coleta e armazenamento dos *pools* z para cada 30 placas. Ao todo, 416 *pools* z foram produzidos e armazenados a 4° C em 45 dias. A partir destes *pools* foram realizadas as extrações dos clones BAC.

As reações de amplificação por PCR utilizando os pares de *primers* (tabela 12) e os *pools* z, localizaram as placas positivas (*screening* de placa). Após isso, os *pools* x e y para cada placa positiva foram confeccionados (inoculação, coleta, armazenamento e extração) e reações de PCR foram feitas para a localização dos endereços dos clones positivos em cada placa (*screening* de coordenada).

3.3.1 *Screening* de placa

Os DNAs dos *pools* z de todas as placas das duas bibliotecas extraídos e diluídos para uma concentração final de 20 ng/µL, foram organizados em placas de 96 poços para viabilizar as reações de PCR. Um volume de 10 µL da reação de PCR foi utilizado para análise. Os controles positivos foram os DNAs genômicos de *A. duranensis* e/ou *A. ipaënsis* na concentração de 20 ng/ µL (marcados em amarelo). O controle negativo para as reações de PCR foi o *mix* sem DNA (marcado em vermelho).

3.3.1.1 Identificação de placas nas bibliotecas BAC A e B contendo clones que apresentam o gene da dessaturase (FAD)

Os primeiros 92 *pools* z (z1 até z92), construídos para a biblioteca A, foram utilizados para amplificação por PCR, utilizando os pares de *primers*: Leg045F/Leg045R e FAD2BF/FAD2BR (figuras 71-A e 71-B). Para Leg045 um total de 16 *pools* z ou 16 placas exibiram um resultado positivo contendo o tamanho do produto de PCR semelhante aos controles positivos (570 pb), sendo que nas placas 15 e 91 o resultado foi mais conspícuo (figura 71-A).

Para FAD2B, 17 placas mostraram um resultado positivo com tamanho do produto de PCR semelhante aos controles positivos (650 pb). Para as duas placas identificadas anteriormente (placas 15 e 91), o resultado também se mostrou mais evidente (figura 71-B). Portanto, foram selecionadas as placas 15 e 91 da biblioteca A para localizar o endereço do(s) BAC(s) positivo(s) para o gene que codifica para a dessaturase em *A. duranensis*. Como os mesmos resultados foram obtidos para os dois diferentes conjuntos de primers de PCR utilizados, foi possível ter uma contraprova, agregando um valor de confiabilidade nas duas placas selecionadas.

Já para a biblioteca B, os primeiros 94 *pools* z (z1 até z92) foram utilizados para amplificação por PCR, com os mesmos pares de *primers* (figura 72-A e 72-B). Para FAD2B, 19 *pools* z ou 19 placas apresentaram resultados compatíveis com os controles positivos (650 pb) (figura 72-A). Para Leg045, 19 placas também indicaram tamanho do produto de PCR semelhante aos controles positivos (570 pb) (figura 72-B). A placa 86 foi selecionada por apresentaram bandas positivas para ambos os *primers* testados.

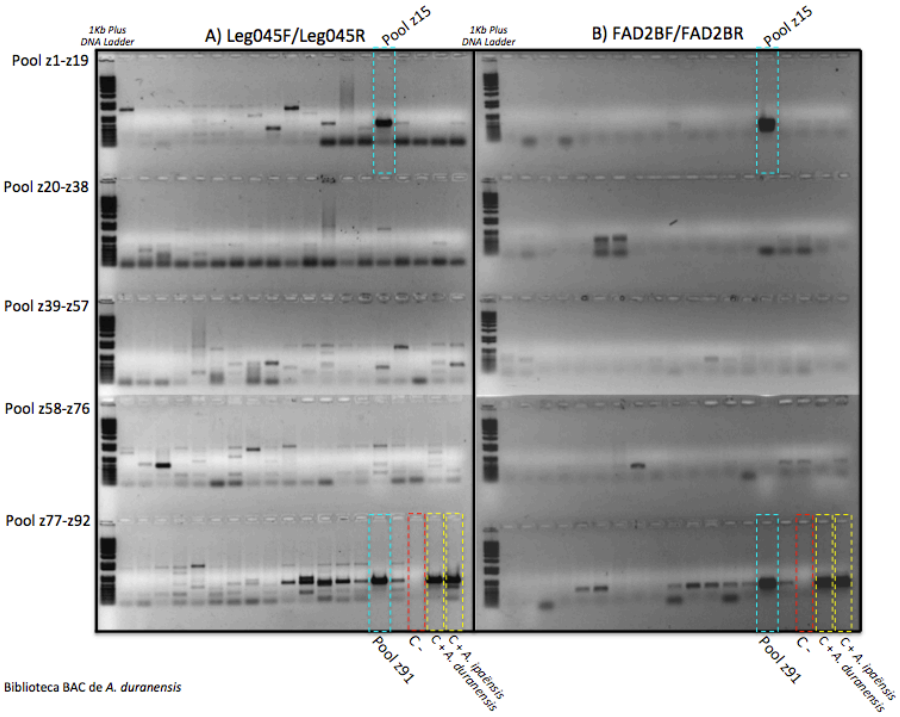


Figura 71: Identificação de *pools* z positivos por meio de reações de PCR utilizando os pares de *primers* Leg045F/Leg045R (A) e FAD2BF/FAD2BR (B) em 92 *pools* da biblioteca A. Foram selecionados os *pools* z15 e z91 (azul), por apresentarem resultados compatíveis com os controles positivos (amarelo). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb Plus DNA Ladder* (Invitrogen).

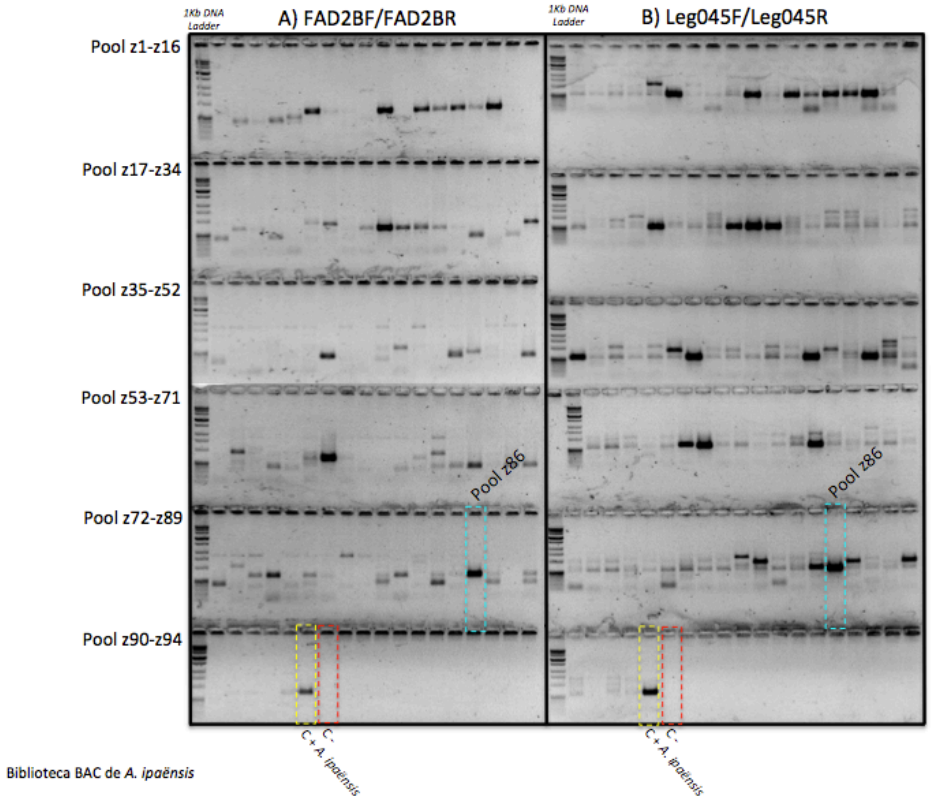


Figura 72: Identificação de *pools* z positivos por meio de reações de PCR utilizando os pares de *primers* FAD2BF/FAD2BR (A) e Leg045F/Leg045R (B) em 94 *pools* da biblioteca B. Foi selecionado o *pool* z86 (azul), por apresentar resultado compatível com o controle positivo (amarelo). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb DNA Ladder* (Invitrogen).

3.3.1.2 Identificação de placas nas bibliotecas BAC A e B contendo clones que apresentam o gene da expansina

Os *pools* z97 até z157 construídos para a biblioteca A foram utilizados para amplificação por PCR, com o par de *primers* ExpUnif/Exp464. Apenas 3 placas indicaram tamanho do produto de PCR semelhante aos controles positivos (600 pb), e uma delas, a 112 ou *pool* z112 foi selecionada (figura 73). Já a figura 74 mostra os *pools* z97 a z157 construídos para a biblioteca B amplificados por PCR utilizando os mesmos pares de *primers*. Duas placas indicaram tamanho do produto de PCR com aproximadamente 600 pb, e a placa 43 foi selecionada.

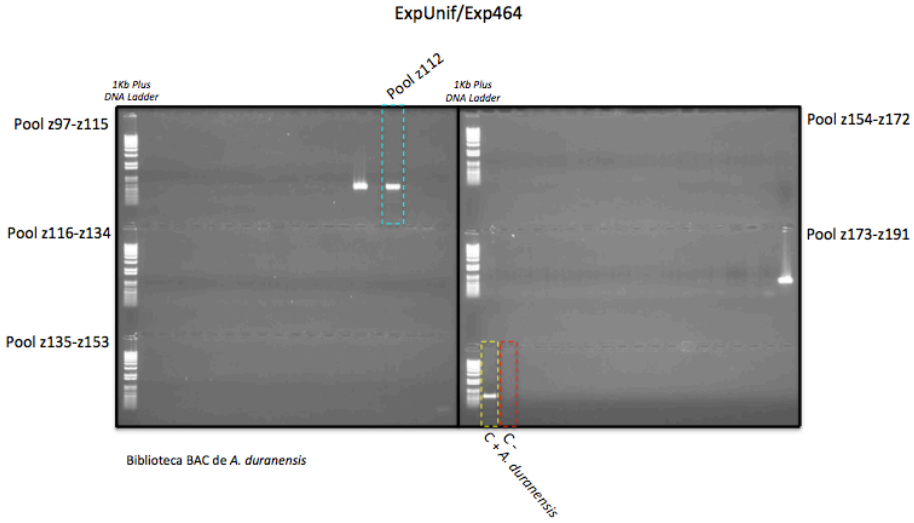


Figura 73: Identificação de *pool* z positivo por meio de reações de PCR utilizando o par de *primers* ExpUnif/Exp464 nos *pools* z97 a z115 da biblioteca A. Foi selecionado o *pool* z112 (azul), por apresentar resultado compatível com o controle positivo (amarelo). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb Plus DNA Ladder* (Invitrogen).

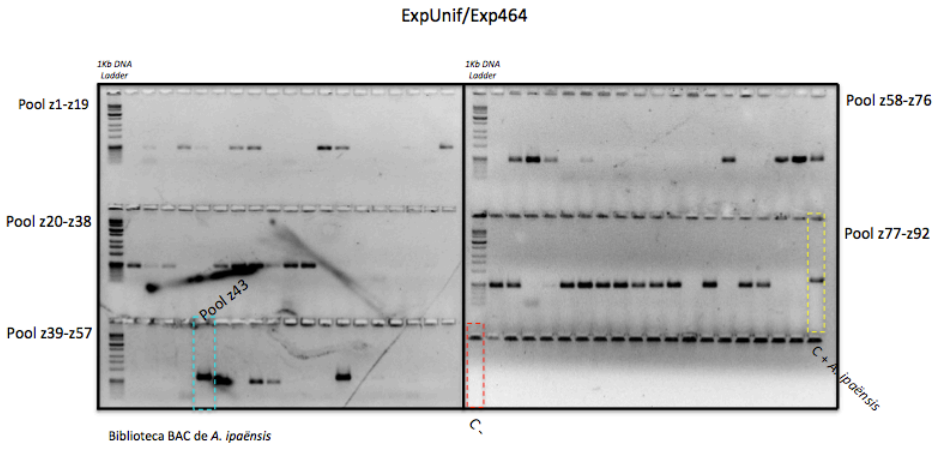


Figura 74: Identificação de *pool* z positivo por meio de reações de PCR utilizando o par de *primers* ExpUnif/Exp464 nos *pools* z1 a z92 da biblioteca B. Foi selecionado o *pool* z43 (azul), por apresentar resultado compatível com o controle positivo (amarelo). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb DNA Ladder* (Invitrogen).

Ao todo, cinco placas foram selecionadas das bibliotecas BAC construídas para as espécies silvestres de amendoim (tabela 12).

Tabela 12: Placas (*pools* z) selecionadas nas bibliotecas A e B de *Arachis*.

Genoma	Pares de *primers*	Tamanho do produto (pb)	*Pool z* (placa)
A	Leg045F/Leg045R	570	Placas 15 e 91
	FAD2BF/FAD2BR	650	Placas 15 e 91
	ExpUnif/Exp464	600	Placa 112
B	Leg045F/Leg045R	570	Placa 86
	FAD2BF/FAD2BR	650	Placa 86
	ExpUnif/Exp464	600	Placa 43

3.3.2 *Screening* de coordenada

A correta identificação da coordenada da placa positiva foi um passo determinante para identificação dos clones BAC de interesse. Se cada placa contendo 384 poços da biblioteca tivesse apenas um clone positivo, ao realizar-se uma reação de PCR com os 24 *pools* x, que representam as colunas da placa; e os 16 *pools* y, que representam as 16 linhas da placa, seriam encontrados dois resultados positivos, uma para a dimensão x e outra para y. Ao sobrepor as duas dimensões é possível inferir qual a coordenada da placa que abriga o clone BAC de interesse, de acordo com o esquema representado na figura 75.

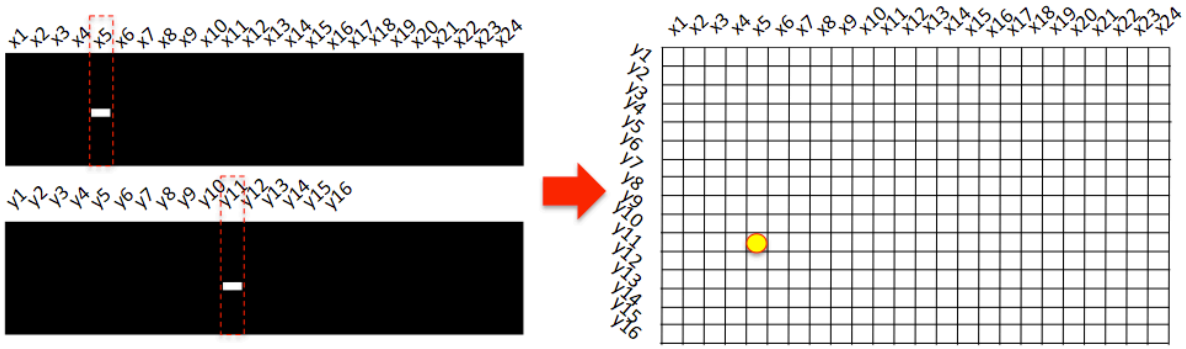


Figura 75: Esquema representativo do método para localização de clones em uma placa de 384 poços por meio de amplificação por PCR utilizando *pools* x, referentes as linhas da placa, e *pools* y, referentes as colunas. No gel mostrado a esquerda, as bandas amplificadas nos *pools* x5 e y11 corresponderam a coordenada K05 (direita).

Foram desenvolvidos os *pools* x e y para cada uma das cinco placas selecionadas, e após a extração dos respectivos *pools* de DNAs, foram feitas diluições para a concentração final de 20 ng/µL. As figuras 76-A a 76-C exemplificam alguns dos géis que indicaram as coordenadas dos clones positivos das placas da biblioteca de *A. duranensis*. Cada gel de

agarose apresentou o resultado das reações de PCR utilizando os 24 *pools* x e 16 *pools* y referentes a cada placa selecionada. Como esperado, por tratarem-se de genes com baixo número de cópias, apenas um conjunto de coordenadas (linha e coluna) em cada placa foi identificado nos géis, corroborando a ideia de que apenas um clone por placa continha a sequência de interesse. As coordenadas e o endereço final dos clones BAC de ambas as bibliotecas estão listados na tabela 13.

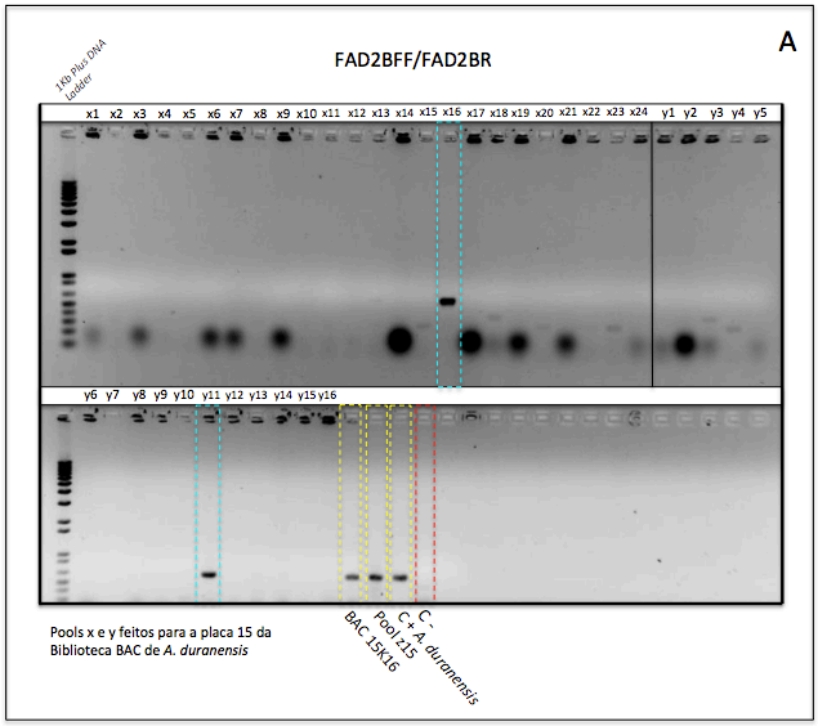


Figura 76: (A) Identificação de coordenadas na placa 15 da biblioteca de *A. duranensis* por meio de reações de PCR utilizando os *primers* FAD2BF/FAD2BR nos *pools* x e y dessa placa. As bandas positivas foram identificadas em: *pool* x16 (coluna 16) e *pool* y11 (linha K) (azul). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb DNA Plus DNA Ladder* (Invitrogen). Controles positivos: DNA genômico de *A. duranensis*, *pool* z15 da biblioteca de A e BAC isolado 15K16 (amarelo).

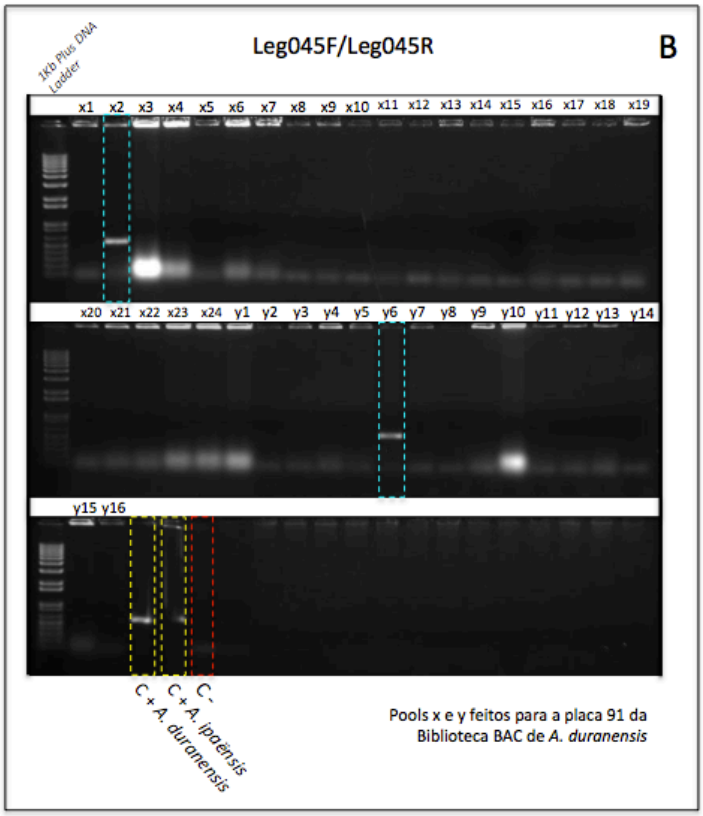


Figura 76: (B) Identificação de coordenadas na placa 91 da biblioteca de *A. duranensis* por meio de reações de PCR utilizando os *primers* Leg045F/Leg045R nos *pools* x e y dessa placa. As bandas foram amplificadas em: *pool* x2 (coluna 2) e *pool* y6 (linha F) (azul). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb Plus DNA Ladder* (Invitrogen). Controles positivos: DNAs genômicos de *A. duranensis* e *A. ipaënsis* (amarelo).

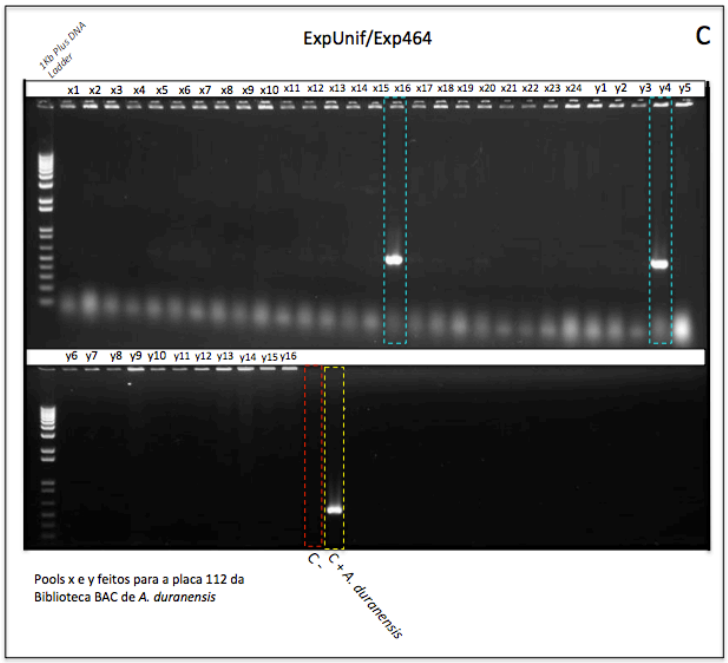


Figura 76: (C) Identificação de coornenadas na placa 112 da biblioteca de *A. duranensis* por meio de reações de PCR utilizando os *primers* ExpUnif/Exp464 nos *pools* x e y dessa placa. As bandas foram amplificadas em: *pool* x16 (coluna 16) e *pool* y4 (linha D) (azul). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb Plus DNA Ladder* (Invitrogen). Controle positivo: DNA genômico de *A. duranensis* (amarelo).

Tabela 13: Coordenadas de linha (x) e coluna (y) referentes à posição dos clones BAC nas placas identificadas nas bibliotecas A e B. A quantificação dos DNAs dos clones foi realizada por Nanodrop (N100).

Genoma	*Pool z* (placa)	*Pool x*	*Pool y*	BAC	Quantificação (ng/µL)
A	Placa 91	x2	y6	91F02	200
	Placa 15	x16	y11	15K16	290
	Placa 112	x16	y4	112D16	170
B	Placa 86	x24	y1	86A24	190
	Placa 43	x13	y1	43A13	170

De posse das coordenadas dos cinco clones, uma nova cultura de células foi feita a partir dos endereços situados nas placas originais. Colônias individualizadas foram utilizadas como novos inóculos para uma nova extração de BAC, utilizando o *Kit Plasmid DNA Purification NucleoBond Xtra Midi*, visando obter maior qualidade e quantidade de DNA para o sequenciamento. A figura 77 mostra um gel de agarose contendo 1 µL do DNA de quatro dos cinco clones BAC extraídos. A quantificação em Nanodrop dos DNAs dos cinco clones revelou valores para razão de absorbância 260/280 entre 1.77-1.87 e razão 260/230 entre 2-2.09. Esses valores foram acetáveis dentro dos padrões de qualidade requeridos pela técnica. De acordo com a tabela 13, os BACs apresentaram quantidades entre 170-290 ng/µL.

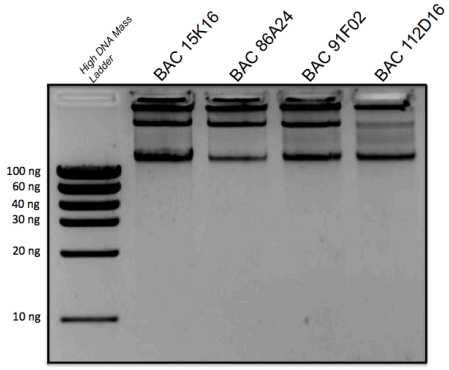


Figura 77: Amostras de quatro clones BAC (BAC 15K16, BAC 86A24, BAC 91F02 e BAC 112D16 - poços 2 a 5 respectivamente) selecionados para sequenciamento. Gel de agarose 1,0% corado com brometo de etídio. Marcador de peso molecular *High DNA Mass Ladder* (Invitrogen) (poço 1).

3.4 Validação dos genes identificados em *pools* de BAC por reações de PCR e sequenciamento

Os quatro pares de *primers* desenhados para os genes da expansina e dessaturase (FAD) foram testados via PCR nos cinco clones BAC isolados das bibliotecas originais. O resultado mostrou que os cinco *amplicons* produzidos exibiram o tamanho esperado, indicando o funcionamento da ferramenta de *pools* de clones BAC desenvolvida neste estudo (figura 78).

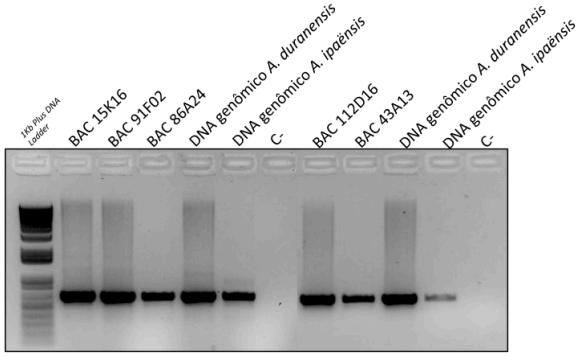


Figura 78: Análises dos *amplicons* obtidos dos clones BAC selecionados. Poços 2-7 (produtos obtido com os *primers* FAD2BF/FAD2BR; Poços 8-12 (*primers* ExpUnif/Exp464). Gel de agarose 1,0% corado com brometo de etídio e marcador *1Kb Plus DNA Ladder* (Invitrogen). Controles positivos: DNA genômico de *A. duranensis e A. ipaënsis*.

Os cinco *amplicons* obtidos dos clones BAC foram sequenciados pelo método Sanger e apresentaram qualidade satisfatória para aproximadamente 300-400 pb cada uma. O resultado da comparação dessas sequências, utilizando a ferramenta BLASTx, com o banco de dados contendo sequências de proteínas não-redundantes do NCBI mostrou que os *amplicons* dos clones BAC 91F02, 15K16 e 86A24 apresentaram similaridade com a dessaturase de *A. hypogaea* (AEW67305.1), enquanto os clones 112D116 e 43A13 apresentaram similaridade com uma expansina isolada da leguminosa *Medicago truncatula* (XP_003611388.1). O *e-value* para todos os resultados de comparação ficaram entre 10^{-21}-10^{-56} e a porcentagem de identidade entre 55-65%.

Para o sequenciamento total e anotação das sequências desses clones BAC foi realizada uma digestão enzimática com cada um. O resultado mostrou um perfil de restrição distinto para todos os clones, ressaltando que tratavam-se de clones diferentes entre si. (figura 79).

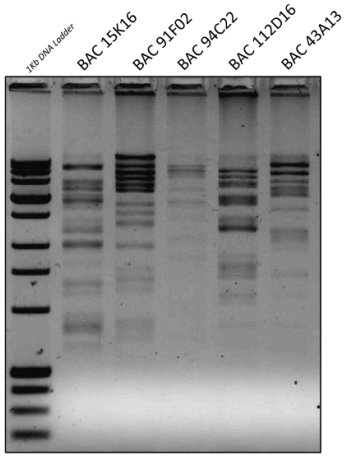


Figura 79: Perfil de restrição dos cinco clones BAC digeridos com a enzima *Eco*RI em gel de agarose 1,0% corado com brometo de etídio. Marcador molecular 1Kb *DNA Ladder* (Invitrogen).

Após a confirmação da presença desses genes ou fragmentos de genes nos cinco clones BAC pertencentes aos genomas A e B de *A. duranensis* e *A. ipaënsis*, respectivamente, esses clones foram sequenciados pela técnica PACBIO. No entanto, apenas dois clones BAC do genoma B de *A. ipaënsis,* 86A24 e 43A13, apresentaram as sequências dos genes que codificam a dessaturase e expansina, respectivamente, de acordo com comparações realizadas por meio de gráfico de plotagem (*dot plots*) entre as sequências dos clones BAC e as sequências derivadas dos *amplicons*. Portanto, não foi possível estabelecer uma relação por meio da comparação entre as sequências desses dois genes nos genomas A e B.

A sequência do clone BAC 86A24, de *A. ipaënsis*, apresentou 98.496 pb de tamanho. Segundo a predição da localização e estrututura de genes realizada pelo FGENESH contra a biblioteca de *M. truncatula,* juntamente com a predição de função das regiões codantes maiores do que 50 pb, por meio de comparações com sequências de proteínas presentes na biblioteca A do *pfam*, foram identificados inicialmente 13 genes putativos, incluindo o gene da dessaturase. Ao comparar a sequência desse clone com as sequências de retrotransposons LTR descritos no Capítulo II, foi possível identificar três retrotransposons LTR com sequências completas, um fragmento e um LTR-solo. No entanto, ao compilar as coordenadas dos genes e retrotransposons, observados na interface do *software* Artemis (figura 80), constatou-se que cinco desses genes putativos, preditos como gag/pol (poliproteína presentes em TEs), na verdade, faziam parte das sequências codantes internas dos retrotransposons identificados, portanto, não poderiam ser anotados separadamente como genes.

Diante disso, para este clone foram preditos oito genes putativos que codificam as seguintes proteínas: 1 - similar a uma protease aspártica (PF00026.17); 2, 3, 4 e 8 - similar a proteína F-box (PF00646.27); 5 – similar a dessaturase (FAD) (PF00487.18); 6 – proteína hipotética; 7 – similar a DUF659 (PF04937.9) (a predição final mostra os genes indicados por números na cor preta – figura 80).

Para caracterizar o gene que codifica a dessaturase (gene 5, composto por 1.200 pb e dois exons), comparações por meio de gráficos de plotagens foram feitas utilizando a sequência do clone BAC contra quatro sequências de genes FAD isoladas de *Arachis hypogaea* presentes no banco de dados GenBank. Dentre essas, foram utilizadas duas sequências descritas por Chu e colaboradores (2009) para uma linhagem de amendoim com característica alto oleico - HO (figura 81). O resultado constatou alta similaridade entre as sequências e o alinhamento entre elas mostrou que apesar da alta similaridade, mutações pontuais estavam presentes na sequência produzida neste estudo (27 substituições e 8

inserções), indicando que há outras diferenças entre os genes da dessaturase encontradas nas espécies silvestres e cultivada, além das mutações que conferem a característica alto oleico presente apenas em algumas cultivares de amendoim (Chu *et al*., 2007) (figura 82).

O conteúdo repetitivo da sequência deste clone foi caraterizado pela presença de dois retrotransposons com sequência completa das família Doros e RE128-84, além de um fragmento do retrotransposon Mico, um LTR-solo de FIDEL ou Feral e um *nested element* (discutido nos Capítulos anteriores) composto por um elemento Mico que possui inserido em sua sequência um fragmento do elemento Hemera (figura 83). Outras comparações, também por meio de gráficos de plotagens, utilizando a sequência desse clone contra ela mesma, revelaram que possivelmente o fragmento do retrotransposon LTR Mico poderia fazer parte de outro *nested element* em conjunto com um retrotransposon LTR desconhecido, em virtude do padrão de diagonais resultantes da plotagem (figura 84).

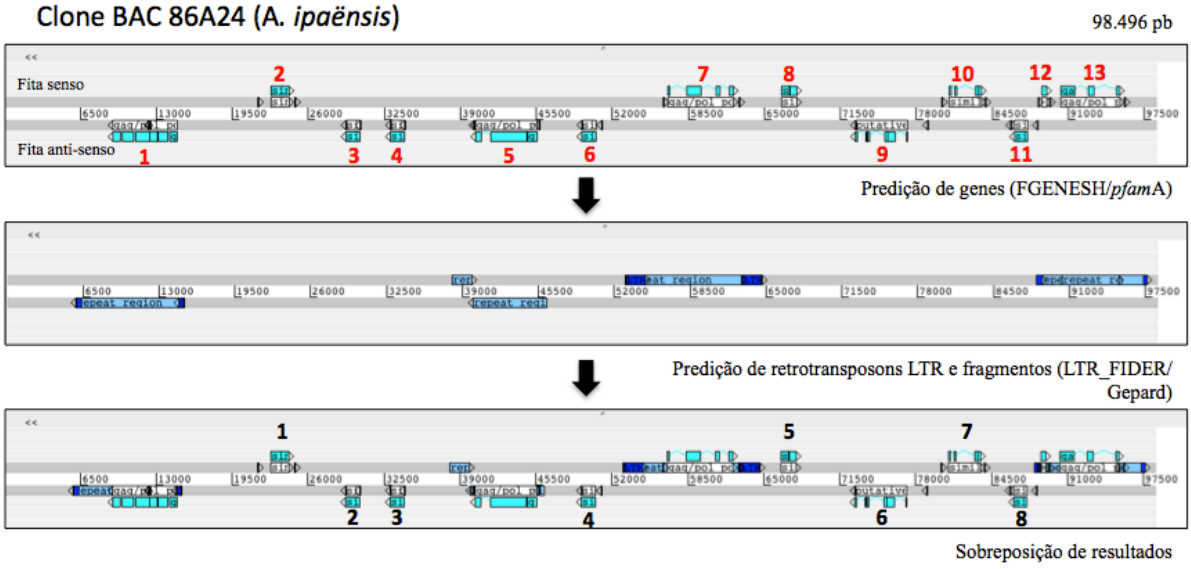


Figura 80: Identificação preliminar de 13 genes putativos presentes no clone BAC 86A24 apenas por meio da utilização dos dados obtidos no *software* FGENESH (vermelho); Identificação de retrotransposons LTR, fragmentos e LTRs-solo (LTRs em azul escuro) por meio de comparações com sequências de retrotransposons LTR conhecidos e análises no *software* LTR_FINDER; Sobreposição de resultados e predição final de oito genes putativos diversos (preto). Os cinco genes identificados na análise preliminar, na verdade compõe as regiões codantes internas dos retrotransposons LTR e fragmentos. Barras acima da linha de escala em pb representam genes e elementos identificados na fita senso, e barras abaixo, indicam a presença na fita complementar. Gráficos produzidos pelo *software* Artemis.

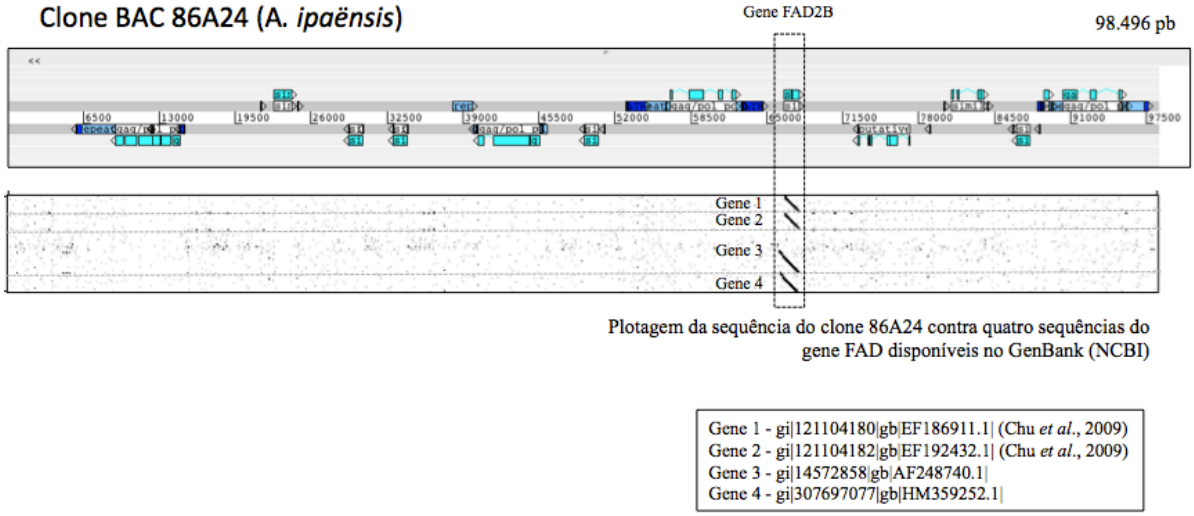


Figura 81: Comparação por meio de gráficos de plotagem entre sequência do clone BAC 86A24 e quatro sequências de genes que codificam dessaturases disponíveis no GenBank. O resultado mostra quatro diagonais bastante similares entre a sequência do gene 5 e as demais. Gráficos produzidos pelos *softwares* Artemis e Gepard.

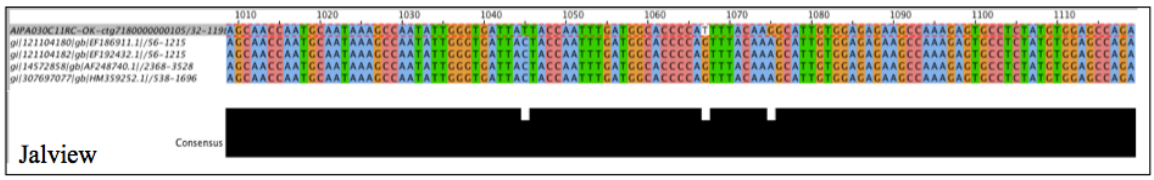


Figura 82:Alinhamento da sequência relativa ao gene 5 que codifica a dessaturase presente na sequência do clone BAC 86A24, juntamente com quatro outras sequências de dessaturases presentes no banco de dados Genbank (interface do *software* Jalview).

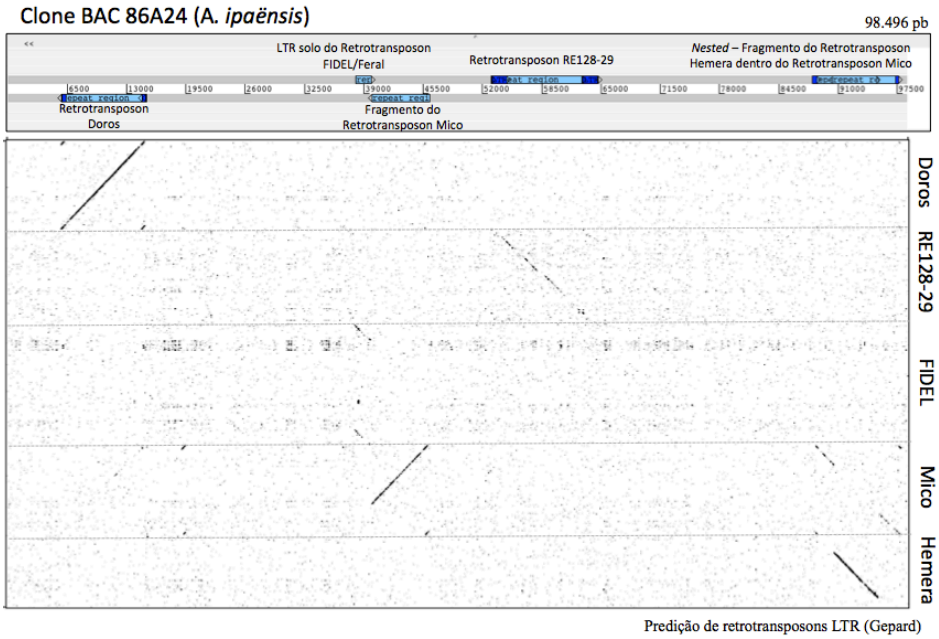


Figura 83: Comparação por meio de gráfico de plotagem da sequência do clone BAC 86A24 (eixo x) com cinco sequências de retrotransposons LTR: Doros, RE128-29, FIDEL, Mico e Hemera (eixo y). Diagonais para a direita indicam que o retrotransposon LTR foi inserido na direção 5' – 3', e para a esquerda o inverso.

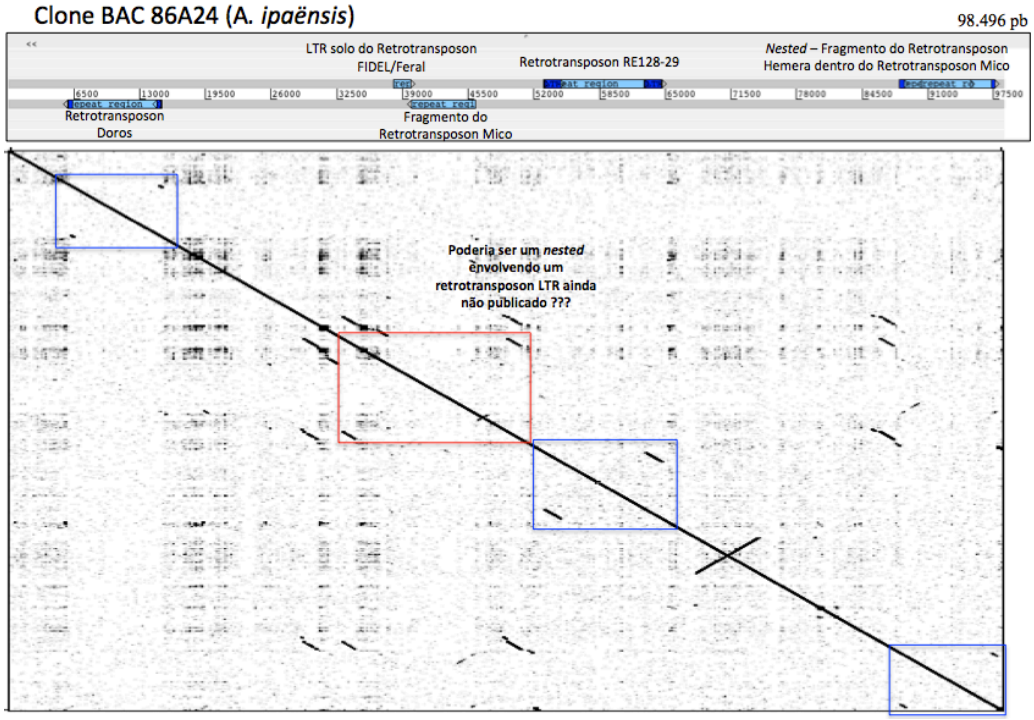


Figura 84: Comparação por meio de gráfico de plotagem da sequência do clone BAC 86A24 contra ela mesma (eixo x). Foram identificados três retrotransposon LTR com sequência completa (quadrados em cor azul), sendo o último, composto por um *nested element*. Outro possível *nested* contendo um retrotransposon LTR ainda não caracterizado abrigando o fragmento do elemento Mico é indicado em vermelho.

Para os resultados obtidos com o clone BAC 43A13, a sequência apresentou 79.919 pb de tamanho. De acordo com a mesma metodologia utilizada para o outro clone BAC, foram identificados nove genes putativos, incuindo o gene da expansina. As proteínas codificadas por cada gene foram: 1 - similar a YDG_SRA ou histona metiltransferase (PF02182.11); 2 - similar a expansina (XP003517398.1); 3 – similar à RVT1 ou transcriptase reversa 1 (PF00078.21); 4 – similar a MP ou proteína de movimento viral (PF01107.12); 5, 6, 7, 8 e 9 –proteínas hipotéticas (figura 85).

Foi identificada apenas a coordenada referente a um retrotransposon LTR (figura 85), e este não apresentou similaridade com aqueles já caracterizados em *Arachis*. Esse elemento novo foi denominado de Elfos e possui 5.790 pb de comprimento total, LTRs flanqueadores com 420 pb e a região codante contendo os genes *gag*, integrase e transcriptase reversa (do tipo RVT2). A ordem de seus genes definiram esse elemento como autônomo e pertencente à superfamília *Ty1-Copia*.

O gene 3 é similar ao gene que codifica a enzima transcriptase reversa presente em retrotransposons, tornando esse gene, possivelmente, derivado de um retrotransposon remanescente evolutivamente. Tanto o gene da transcriptase quanto o retrotransposon Elfos exibiram espaços praticamente sem pontilhados quando a sequência desse clone foi plotada

contra ela mesma, dando suporte a uma possível inserção. Essa ausência de pontilhados é frequentemente observada quando há a presença de elementos repetitivos nas sequências de DNA, o que normalmente não ocorre nos locais com conteúdo gênico. Sequências de microssatélites também foram identificadas em vários locais desse clone, e no gráfico de plotagem, são evidenciadas por três blocos fortemente pontilhados (figura 85).

A sequência do gene que codifica a proteína expansina (gene 2) com aproximadamente 1.500 pb, quando comparada com sequências de proteínas não-redundantes presentes no banco de dados do GenBank, via BLASTx e ferramenta "conserved domais", revelou uma estrutura composta por quatro exons e três introns. Ao comparar com a estrutura de um gene que codifica a expansina caracterizado por Brasileiro e colaborados (dados não publicados), obteve-se uma estrutura bastante próxima, novamente demonstrando a eficiência da técnica de *pools* de BAC para o isolamento de genes de interesse (figura 86).

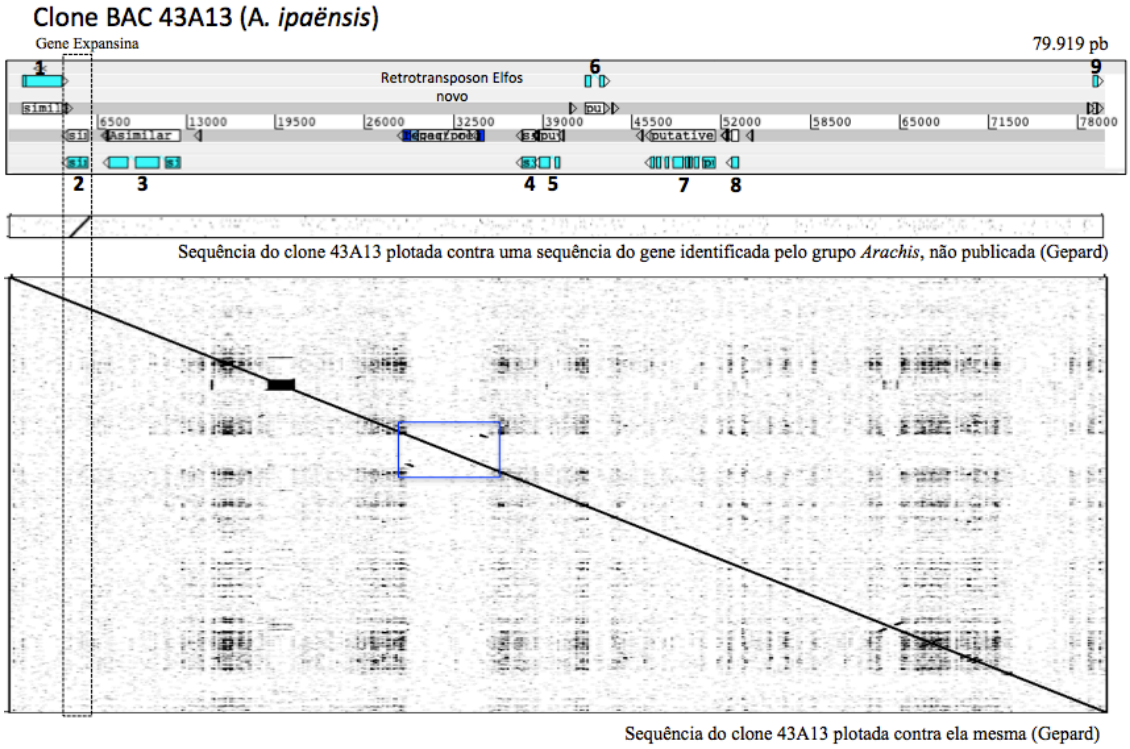


Figura 85: Identificação de nove genes putativos na sequência do clone BAC 43A13; Comparação por meio de gráfico de plotagem da sequência do clone BAC 43A13 contra uma sequência relativa ao gene que codifica uma expansina identificada por Brasileiro e colaboradores (dados não publicados); Comparação por meio de gráfico de plotagem da sequência do clone BAC 43A13 contra ela mesma, revelando a presença de um retrotransposon LTR novo denominado Elfo (azul). Gráficos produzidos pelos *softwares* Artemis e Gepard.

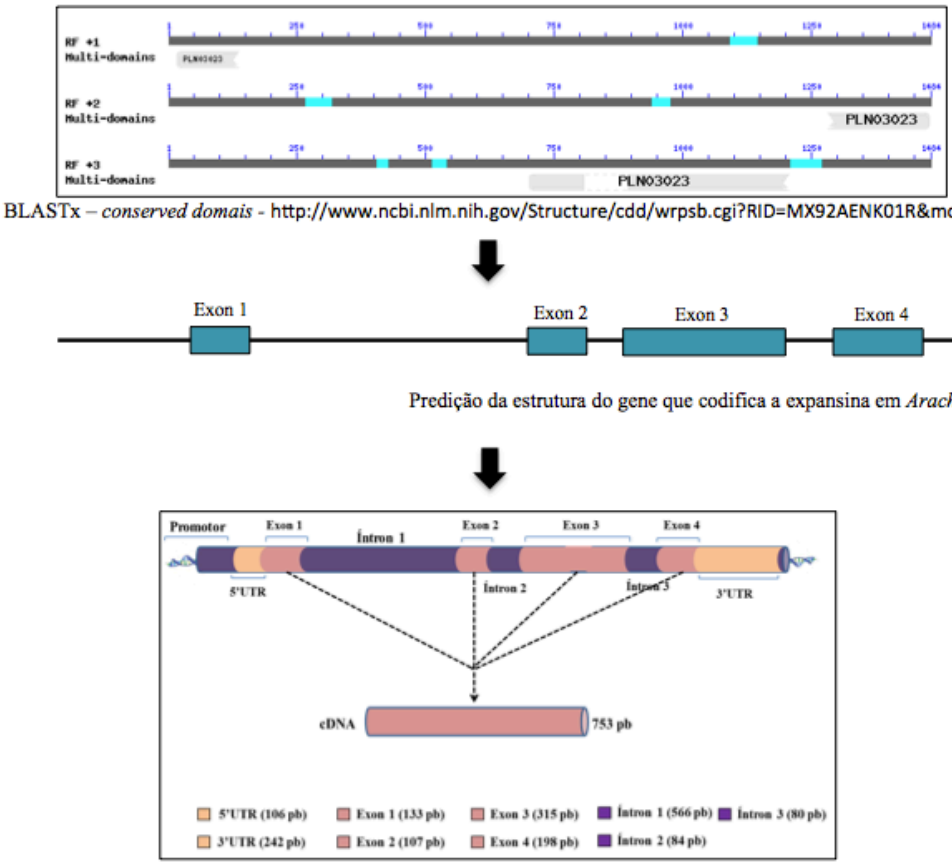


Figura 86: Resultado da comparação da sequência do gene da expansina identificado no clone BAC 43A13 com sequências depositadas no banco de dados do GenBank, via BLASTx e estrutura predita pela ferramenta "conserved domais". As barras em cor cinza representam domínios relativos aos quatro exons identificados no gene. Comparação da provável estrutura do gene com uma sequência relativa ao gene EXLB que codifica uma expansina like-B identificada por Brasileiro e colaboradores (dados não publicados).

4. Discussão

Neste trabalho foi realizada a construção e validação de *pools* 3-D construídos para bibliotecas BAC representativas dos genomas A (*A. duranensis*) e B (*A. ipaënsis*). Essa ferramenta possibilitou a identificação e o isolamento rápido e de baixo custo de genes de interesse nas espécies silvestres parentais do amendoim. A detecção de clones com sequências específicas utilizando uma estratégia de *pools* associada à técnica de PCR já foi demonstrada como eficiente em outros trabalhos, permitindo a identificação individual de clones e diminuindo o número de falsos positivos em outras culturas (Xia *et al*., 2009).

A identificação de clones específicos pelos *pools* 3-D permitiu a análise sequencial e rápida de 159.744 clones BAC (84.096 clones da biblioteca A e 75.648 clones da biblioteca B) divididas simplificadamente em duas reações de PCR (a primeira realizada com os *pools* z e segunda com os *pools* x e y), além de uma reação adicional para a confirmação da posição do clone identificado utilizando os clones da biblioteca original. Dessa forma, houve uma considerável diminuição de tempo quando comparado com o método de avaliação individual de cada clone. Além do mais, demonstrou ser um método eficiente e de custo reduzido quando comparado com o método de hibridização por marcação radioativa.

Deve-se ainda considerar que a manutenção das bibliotecas BAC é difícil, pois ocupa grande espaço de armazenamento em freezer a -80° C, além da necessidade de renovação das culturas de bactérias. Dessa forma, a confecção dos *pools* 3-D permite uma considerável economia de espaço. Outra vantagem desta técnica é que com auxílio do recurso tridimensional de *pools*, é possível minimizar o descongelamento das bibliotecas originais e as duplicadas construídas para estoque de trabalho, fato que aumenta sua sobrevida.

A confecção de 416 *pools* na dimensão z, derivados das duas bibliotecas, divididas em placas de 96 poços, facilitou o *screening* de placas positivas. Para cada placa identificada, 40 *pools* (24 *pools* na dimensão x + 16 *pools* na dimensão y) foram requeridos para encontrar o endereço correto do clone de interesse.

Foram identificados cinco clones BAC contendo os genes de interesse que codificam a dessaturase e a expansina. A validação dos *pools* de BAC foi realizada por meio de reações de amplificação por PCR utilizando pares de *primers* referentes aos genes de interesse e sequenciamento desses *amplicons* pela técnica de Sanger, corroborando a ideia de que a ferramenta de *pools* é eficiente e segura para a busca de clones de interesse nas bibliotecas construídas para as espécies silvestres de amendoim.

O sequenciamento completo dos cinco clones BAC pela técnica PACBIO, no entanto, produziu resultados positivos (para a identificação dos genes de interesse) apenas para dois dos clones BAC do genoma B de *A. ipaënsis,* 86A24 e 43A13. Portanto, não foi possível estabelecer uma relação por meio da comparação entre as sequências dos genes da dessaturase e expansina presentes nos genomas de *A. duranensis* e *A. ipaënsis*.

A sequência do clone BAC 86A24 possui 98.496 pb e apresenta oito genes putativos, incuindo o gene da dessaturase. O conteúdo repetitivo foi caraterizado pela presença de dois retrotransposons LTR com sequência completa de representantes das família Doros e RE128-84, além de um fragmento do retrotransposon Mico, um LTR-solo de FIDEL ou Feral e um *nested element* composto por um elemento Mico que possui inserido em sua sequência um fragmento do elemento Hemera. Para a sequência deste clone BAC, a predição do conteúdo gênico poderia ter sido equivocada, caso a identificação de TEs como restrotransposons LTR não fosse realizada de forma detalhada.

O gene putativo que codifica a dessaturase (FAD), com aproximadamente 1.200 pb e composto por dois exons, revelou alta similaridade quando comparado a outras sequências do gene FAD de amendoim. No entanto foram detectadas na sequência obtida neste estudo 27 substituições e 8 inserções, indicando que há outras diferenças entre os genes da dessaturase encontrados nas espécies silvestres e cultivada, além das mutações que conferem a característica alto oleico presente apenas em algumas cultivares de amendoim (Chu *et al*., 2007).

A sequência do clone BAC 43A13 possui 79.919 pb e apresenta nove genes putativos, incluindo o gene da expansina. Foi identificado um novo retrotransposon LTR autônomo denominado Elfos, pertencente à superfamília *Ty1-Copia*. O gene putativo que codifica a expansina, com aproximadamente 1.500 pb e composto por quatro exons e três introns, revelou alta similaridade quando comparado a uma sequência do gene da expansina isolado de amendoim por Brasileiro e colaborados (dados não publicados). Portanto, apesar da impossibilidade de identificação desses genes em clones BAC de *A. duranensis* por meio do sequenciamento de alta performance, esses genes foram identificados e caracterizados em clones de *A. ipaënsis*, demonstrando a eficiência da técnica de *pools* de BAC para o isolamento de genes de interesse desenvolvida neste estudo.

5. Conclusão

Neste estudo foi demonstrado que a construção da ferramenta de *pools* 3-D de clones BAC de duas bibliotecas representativas dos genomas de *A. duranensis* e *A. ipaënsis,* possibilitou a identificação e o isolamento rápido de três clones BAC contendo gene que codifica a enzima dessaturase de ácidos graxos e outros dois contendo o gene que codifica uma expansina, ambos os genes reconhecidos como sendo de interesse para a cultura do amendoim. A validação da ferramenta foi realizada por meio de reações de PCR e sequenciamento pela técnica de Sanger. Dois dos cinco clones, sequenciados pela técnica de PACBIO, mostraram resultados satisfatórios, possibilitando, além da identificação dos genes, a anotação dos conteúdos gênico e repetitivo dessas sequências completas.

Perspectivas

- Análisar a distribuição das famílias de retrotransposons LTR em todas as pseudomoléculas de *A. duranensis* e *A. ipaënsis*;

- Realizar hibridização *in situ* por fluorescência em cromossomos das espécies diploides *A. duranensis* e *A. ipaënsis*, utilizando as famílias de retrotransposons LTR descritas neste estudo;

- Realizar hibridização *in situ* por fluorescência em cromossomos de amendoim, utilizando outras famílias descritas neste estudo, que ainda não foram utilizadas;

- Identificar e caracterizar famílias de retrotransposons LTR na sequência genômica de amendoim, quando estiver disponível;

- Compilar os *scripts* e criar uma ferramenta única para caracterização e datação de retrotransposons LTR identificados em outras espécies;

- Publicar o artigo referente ao Capítulo II;

Referências Bibliográficas

Adams KL & Wendel JF. 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology 8:135–141.

Agroenergia. 2012. Anuário estatístico da Agroenergia. Ministério da Agricultura, Pecuária e Abastecimento.

Ahn SC, Baek BS, Oh T, Song CS, Chatterjee B. 2000. Rapid mini-scale plasmid isolation for DNA sequencing and restriction mapping. BioTechniques 29:466-468.

Alix K, Heslop-Harrison JS. 2004. The diversity of retroelements in diploid and allotetraploid Brassica species. Plant Mol Biol 6:895-909.

Alix K, Joets J, Ryder CD, Moore J, Barker GC, Bailey JP, King GJ, Pat Heslop-Harrison JS. 2008. The CACTA transposon Bot1 played a major role in *Brassica* genome divergence and gene proliferation. Plant J. 56:1030-44.

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

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

Ansari KI, Walter S, Brennan JM, Lemmens M, Kessans S, McGahern A, Egan D, Doohan FM. 2007. Retrotransposon and gene activation in wheat in response to mycotoxigenic and non- mycotoxigenic-associated *Fusarium* stress. Theor Appl Genet 114:927-937.

Anuradha ST, Divya K, Jami SK, Kirti PB. 2008. Transgenic tobacco and peanut plants expressing a mustard defensin show resistance to fungal pathogens. Plant Cell Rep 11:1777-86.

Araújo ACG, Nielen S, Vidigal BS, Morezthson MC, Leal-Bertioli SCM, Ratnaparkhe M, Changsoo K, Bailey J, Paterson A, Schwarzacher T, Heslop-Harrison P, Bertioli DJ. 2012. An analysis of the repetitive component of the peanut genome in the evolutionary context of the Arachis A-B genome divergence. In: XX Plant and animal Genome, San Diego. Annals of XX PAG.

Barillot E, Lacroix B, Cohen D. 1991. Theoretical analysis of library screening using a N-dimensional pooling strategy. Nucleic Acids Research 19:6241–6247.

Bechara MD, Moretzsohn MC, Palmieri DA, Monteiro JP, Bacci M, Martins J, Valls JF, Lopes CR, Gimenes MA. 2010. Phylogenetic relationships in genus *Arachis* based on ITS and 5.8S rDNA sequences. BMC Plant Biol 10:255–255.

Belyayev A, Raskina O, Nevo E. 2005. Variability of the chromosomal distribution of Ty3-gypsy retrotransposons in the populations of two wild *Triticeae* species. Cytogenet Genome Res 109:43-49.

Bennetzen JL, Kellogg EA.1997. Do Plants Have a One-Way Ticket to Genomic Obesity? Plant Cell 9:1509-1514.

Bennetzen JL. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621-627.

Bennetzen, J.F. 2000. Transposoable element contributions to plant gene and genome evolution. Plant Mol Biol 42: 251-269.

Bertioli DJ, Moretzsohn MC, Madsen LH, Sandal N, Leal-Bertioli SCM, Guimarães PM, Hougaard BK, Fredslund J, Schauser L, Nielsen AM, Sato S, Tabata S, Cannon S, 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.

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

Bertioli DJ, Vidigal BS, Nielen S, Ratnaparkhe MB, Lee TH, Leal-Bertioli SCM, Kim C, Guimaraes PM, Seijo G, Schwarzacher T, Paterson AH, Heslop-Harrison P, Araujo ACG. 2013. The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its volutionary divergence from the B genome. Annals of Botany 112:545-559.

Bhatnagar-Mathur P, Devi MJ, Reddy DS, Lavanya M, Vadez V, Serraj R, Yamaguchi-Shinozaki K, Sharma KK. 2007. Stress-inducible expression of At DREB1A in transgenic peanut (*Arachis hypogaea* L.) increases transpiration efficiency under water-limiting conditions. Plant Cell Rep 26:2071–2082.

Bhattacharyya MK, Smith AM, Ellis TH, Hedley C, Martin C. 1990. The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. Cell 60:115-22.

Boeke JD, Corces VG.1989. Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43:403-34.

Bonavia D.1982. Precerá mico peruano, Los Gavilanes, oasis en la historia del hombre. Corporación Financiera de Desarrollo S.A. COFIDE e Instituto Arqueológico Alemán.

Brasileiro ACM, Morgante CV, Bertioli SCML, Araujo ACG, Silva AK, Martins A, Bertioli D, Guimarães PM. 2012 Identificação de genes associados à resposta ao estresse hídrico em espécies silvestres de *Arachis*. Heringeriana 6:35-36.

Bruner AC, Jung S, Abbott A, Powell G. 2001. The naturally occurring high oleate oil character in some peanut varieties results from reduced oleoyl-PC desaturase activity from mutation of aspartate 150 to asparagine. Crop Sci 41:522–526.

Bundock P, Hooykaas P. 2005. An *Arabidopsis* hAT-like transposase is essential for plant development. Nature 436:282-284.

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.

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.

Cação SM, Silva NV, Domingues DS, Vieira LG, Diniz LE, Vinecky F, Alves GS, Andrade AC, Carpentieri-Pipolo V, Pereira LF. 2013. Construction and characterization of a BAC library from the *Coffea arabica* genotype Timor Hybrid CIFC 832/2. Genetica 141:217-226.
Campbell TN, Choy FY. 2002. Approaches to library screening. J Mol Microbiol Biotechnol 4:551-4.

Casacuberta JM, Santiago N. 2003. Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1-11.

Chirinos FV. 2011. Breeding for Early Maturity in Peanuts (*Arachis hypogaea* L.) using Traditional Methods and Marker Assisted Selection (MAS). North Carolina State University.

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

Chu Y, Holbrook CC, Ozias-Akins P. 2007. Two Alleles of ahFAD2B Control the High Oleic Acid Trait in Cultivated Peanut. Crop Sci. 49:2029–2036.

Conab. 2013. Conab - Acompanhamento da Safra Brasileira - Grãos - Safra 2012/2013. Acesso em Junho de 2013, disponível em Conab: http://www.conab.gov.br/OlalaCMS/uploads/arquivos/11_08_09_11_44_03_boletim_agosto-2013.pdf

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA. 103:8101-8106.

Cosgrove DJ. 2000. Loosening of plant cell walls by expansins. Nature 407.

Creste S, Tsai SM, Valls JFM, Gimenes MA, Lopes CR. 2005. Genetic characterization of Brazilian annual Arachis species from section *Arachis* and Heteranthae using RAPD markers. Genetic Resources and Crop Evolution 52:1079-1086.

Cronk Q, Ojeda I, Pennington RT. 2006. Legume comparative genomics: progress in phylogenetics and phylogenomics. Curr Opin Plant Biol 2:99-103.

Cui X, Jin P, Cui X, Gu L, Lu Z, Xue Y, Wei L, Qi J, Song X, Luo M, An G, Cao X. 2013. Control of transposon activity by a histone H3K4 demethylase in rice. Proc Natl Acad Sci USA, 110:1953-1958.

Curcio MJ, Derbyshire KM. 2003. The outs and ins of transposition: from mu to kangaroo. Nat Rev Mol Cell Biol 4:865-877.

Danesh D, Penuela S, Mudge J et al. 1998. A bacterial artificial chromosome library for soybean and identification of clones near a major cyst nematode resistance gene. Theoretical and Applied Genetics 96:196–202.

Devos KM, Brown JK, Bennetzen JL. 2012. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome Res 12:1075-9.

Dhillon SS, Rake AV, Miksche JP. 1980. Reassociation kinetics and cytophotometric characterization of peanut (*Arachis hypogaea* L.). DNA Plant Physiol 65:1121-1127.

Domingues DS, Cruz GM, Metcalfe CJ, Nogueira FT, Vicentini R, Alves Cde S, Van Sluys MA. 2012. Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics 13:137.

Doolittle WF & Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284:601-603.

Duangpan S, Zhang W, Wu Y, Jansky SH, Jiang J. 2013. Insertional mutagenesis using Tnt1 retrotransposon in potato. Plant Physiol 163:21-29.

Duval-Valentin G, Marty-Cointin B, Chandler M. 2004.Requirement of IS911 replication before integration defines a new bacterial transposition pathway. EMBO J 23:3897-3906.

Dwivedi SL, Bertioli DJ, Crouch JH, Valls JFM, Upadhyaya HD, Fávero AP, Moretzsohn MC, Paterson AH. 2006. Peanut Genetics and Genomics: Toward Marker-assisted Genetic Enhancement in Peanut (*Arachis hypogaea* L.). In: KOLE C. (Ed.). Oilseeds. Series: Genome Mapping and Molecular Breeding in Plants, Vol. 2, Springer, 302p.

Dwivedi SL, Crouch JH, Nigam SN, Ferguson ME, Paterson AH. 2003. Molecular breeding of groundnut for enhanced productivity and food security in the semi-arid tropics: Opportunities and challenges. Advances in Agronomy 80:153-221.

Echenique V, Stamova B, Wolters P, Lazo G, Carollo L, Dubcovsky J. 2002. Frequencies of Ty1-copia and Ty3-gypsy retroelements within the *Triticeae* EST databases. Theor Appl Genet 104:840-844.

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 Res 32:1792–1797.

Eichten SR, Ellis NA, Makarevitch I, Yeh CT, Gent JI, Guo L, McGinnis KM, Zhang X, Schnable PS, Vaughn MW, Dawe RK, Springer NM. 2012. Spreading of heterochromatin is limited to specific families of maize retrotransposons. PLoS Genet. 8:e1003127.

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.

Fávero AP, Simpson CE, Valls JFM e 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.

Fedoroff N. 2000. Transposons and genome evolution in plants. Proc Natl Acad Sci USA 97:7002-7007.

Feinberg AP, Vogelstein B. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem 132:6-13.

Feinberg AP, Vogelstein B. 1984. "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. Anal Biochem 137:266-267.

Feng S, Wang X, Zhang X, Dang PM, Holbrook CC, Culbreath AK, Wu Y, Guo B. 2012. Peanut (*Arachis hypogaea*) Expressed Sequence Tag Project: Progress and Application. Comp Funct Genomics 2012:373768.

Fernandez A & Krapovickas A. 1994. Cromossomas y evolucion en *Arachis* (Leguminosaea). Bonplandia 8: 187-220.

Ferreira ME & Grattapaglia D. 1998. Introdução ao uso de marcadores moleculares em análise genética. 3. ed. Brasília: EMBRAPA-CENARGEN.

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

Finnegan DJ. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet 5:103-107.

Flavell RB, Bennet MD, Smith JB. 1974. Genome size and proposition of reapeted nucleotide sequence DNA in plants. Biochemical Genetics 12:257-269.

Foncéka D, Hodo-abalo T, Rivallan R, Faye I, Sall M, Ndoye O, Favero AP, Bertioli DJ, Glaszmann J, Rami JF. 2009. Genetic mapping of wild introgressions into cultivated peanut: a way toward enlarging the genetic basis of a recent allotetraploid. BMC plant biology 9:1-13.

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.

Freitas FO, Penaloza APS, Valls JFM. 2003. O amendoim contador de história. Brasília: Embrapa Recursos Genéticos e Biotecnologia. Documentos, n.107. 12p

Fukai E, Dobrowolska AD, Madsen LH, Madsen EB, Umehara Y, Kouchi H, Hirochika H, Stougaard J. 2008. Transposition of a 600 thousand-year-old LTR retrotransposon in the model legume *Lotus japonicus*. Plant Mol Biol 68:653-663.

Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J. 2012. Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics 100:222-30.

Gao X, Havecker ER, Baranov PV, Atkins JF, Voytas DF. 2003. Translational recoding signals between gag and pol in diverse LTR retrotransposons. RNA 9:1422-1430.

Garcia G.M, Stalker HT, Shroeder E, Kochert G. 1996. Identification of RAPD, SCAR, and RFLP markers tightly linked to nematode resistance genes introgressed from Arachis cardenasii into *Arachis hypogaea*. Genome 39:836-845.

Gautami B, Foncéka D, Pandey MK, Moretzsohn MC, Sujay V, Qin H, Hong Y, Faye I, Chen X, BhanuPrakash A, Shah TM, Gowda MV, Nigam SN, Liang X, Hoisington DA, Guo B, Bertioli DJ, Rami JF, Varshney RK. 2012. An international reference consensus genetic map with 897 marker loci based on 11 mapping populations for tetraploid groundnut (*Arachis hypogaea* L.). PLoS One 7:e41213.

Gimenes MA, Lopes CR, Valls JFM. 2002. Genetic relationships among *Arachis* species based on AFLP. Genetics and Molecular Biology 5:349-353.

Godoy IJ, Moraes SA, Zanotto MD, Santos RC. 1999. Melhoramento do amendoim. In: A. Borem (Ed.), Melhoramento de espécies cultivadas. Viçosa: UFV. pp. 51-94.

González LG & Deyholos MK. Identification, characterization and distribution of transposable elements in the flax (*Linum usitatissimum* L.) genome. BMC Genomics 2012, 13:644.

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

Govind G, Harshavardhan VT, Patricia JK, Dhanalakshmi R, Senthil Kumar M, Sreenivasulu N, Udayakumar M. 2009. Identification and functional validation of a unique set of drought induced genes preferentially expressed in response to gradual water stress in peanut. Mol Genet Genomics 281:591-605.

Grandbastien MA, Audeon C, Bonnivard E, Casacuberta JM, Chalhoub B, Costa AP, Le QH, Melayah D, Petit M, Poncet C, Tam SM, Van Sluys MA, Mhiri C. 2005. Stress activation and genomic impact of Tnt1 retrotransposons in *Solanaceae*. Cytogenet Genome Res 110:229-241.

Grandbastien MA, Spielmann A, Caboche M. 1989. Tnt1, a mobile retroviral-like transposable element of tobacco isolated by plant cell genetics. Nature 337:376-380.

Grandbastien MA. 1998. Activation of plant retrotransposons under stress conditions. Trends Plant Sci 3: 181-187.

Gregory MP & Gregory WC. 1979. Exotic germplasm of *Arachis* L. interspecific hybrids. Journal of Heredity 70: 185-193.

Gregory WC, Krapovickas A, Gregory MP. 1980. Structure, variation, evolution, and classification in *Arachis*. In: Summerfield RJ, Bunting AH. (Ed.). Advances in Legume Science. Kew, Royal Botanical Gardens, p.469-481.

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

Guimarães P, Brasileiro A, Proite K, de Araújo A, Leal-Bertioli S, Pic-Taylor A, da Silva F, Morgante C, Ribeiro S, Bertioli D. 2010. A study of gene expression in the nematode resistant wild peanut relative, *Arachis stenosperma*, in response to challenge with *Meloidogyne arenaria*. Trop Plant Biol 3:183–192.

Guimarães PM, Brasileiro AC, Morgante CV, Martins AC, Pappas G, Silva OB Jr, Togawa R, Leal-Bertioli SC, Araujo AC, Moretzsohn MC, Bertioli DJ. 2012. Global transcriptome analysis of two wild relatives of peanut under drought and fungi infection. BMC Genomics 13:387.

Guimarães PM, Garsmeur O, Proite K, Leal-Bertioli SC, 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 Biology, 8:14.

Guo BZ, Xu G, Cao YG, Holbrook CC, Lynch RE. 2006. Identification and characterization of phospholipase D and its association with drought susceptibilities in peanut (*Arachis hypogaea*). Planta 223:512-20.

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

Halward T, Stalker HT, Kochert G. 1993. Development of an RFLP linkage map in diploid peanut species. Theoretical and Applied Genetics 87:379-384.

Hammons, RO. 1994. The origin and history of the groundnut. Em: Smartt, J. The Groundnut Crop. A scientific basis for improvement. London, Chapman & Hall, 24-42.

Hernández-Pinzón I, Cifuentes M, Hénaff E, Santiago N, Espinás ML, Casacuberta JM. 2012. The Tnt1 retrotransposon escapes silencing in tobacco, its natural host. PLoS One. 7:e33816.

Herselman L, Thwaites R, Kimmins FM, Courtois B, Van Der Merwe PJA, Seal SE. 2004. Identification and mapping of AFLP markers linked to peanut (*Arachis hypogaea* L.) resistance to the aphid vector of groundnut rosette disease. Theoretical and Applied Genetics 109:1426-1433.

Herselman L. 2003. Genetic variation among Southern African cultivated peanut (*A. hypogaea* L.) genotypes as revealed by AFLP analysis. Euphytica 133:319-327.

Heslop-Harrison JS, Brandes A, Taketa S, Schmidt T, Vershinin AV, Alkhimova EG, Kamm A, Doudrick RL, Schwarzacher T, Katsiotis A, Kubis S, Kumar A, Pearce SR, Flavell AJ, Harrison GE. 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. Genetica 100:197-204.

Heslop-Harrison, JS. 2000. Comparative Genome Organization in Plants: From Sequence and Markers to Chromatin and Chromosomes. The Plant Cell 12:617-635.

Hilu KW. 1993. Polyploidy and the evolution of domesticated plants. American Journal of Botany 80:1494-1499.

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. 1996. Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93:7783-7788.

Holbrook CC & Stalker HT. 2003. Peanut breeding and genetic resources. Plant Breeding Reviews 22: 297-356.

Hong Y, Chen X, Liang X, Liu H, Zhou G, Li S, Wen S, Holbrook CC, Guo B. 2010. A SSR-based composite genetic linkage map for the cultivated peanut (*Arachis hypogaea* L.) genome. BMC plant biology 10:1-13.

Hougaard BK, Madsen LH, Sandal N, de Carvalho Moretzsohn M, 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.

Hu Y, Lu Y, Ma D, Guo W, Zhang T. 2010. Construction and characterization of a bacterial artificial chromosome library for the A-genome of cotton (*G. arboreum* L.). J Biomed Biotechnol 2010:457137.

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

Hudson ME, Lisch DR, Quail PH. 2003. The FHY3 and FAR1 genes encode transposase-related proteins involved in regulation of gene expression by the phytochrome A-signaling pathway. Plant J 34:453-471.

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.

Isleib TG, Holbrook CC, Gorbet DW. 2001. Use of plant introductions in peant cultivar development. Peanut Science 28:96-113.

Isleib TG, Wilson RF, Novitzky WP. 2006. Partial dominance, pleiotropism, and epistasis in the inheritance of the high-oleate trait in peanut. Crop Sci 46:1331–1335.

Janicki MM, Rooke R, Yang G. 2011. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. Chromosome Res 19:787–808.

Jiao Y, Deng XW. 2007. A genome-wide transcriptional activity survey of rice transposable element-related genes. Genome Biol 8:R28.

Joly AB. 2002. Botânica: Introdução à taxonomia vegetal. 13ª ed. São Paulo, Companhia Editora Nacional.

Jung S, Powell G, Moore K, Abbott A. 2000a. The high oleate trait in the cultivated peanut [*Arachis hypogaea* L]: II. Molecular basis and genetics of the trait. Mol. Gen. Genet. 263:806–811.

Jung S, Swift D, Sengoku E, Patel M, Teule F, Powell G, Moore K, Abbott A. 2000b. The high oleate trait in the cultivated peanut [*Arachis hypogaea* L.]: I. Isolation and characterization of two genes encoding microsomal oleoyl-PC desaturases. Mol. Gen. Genet. 263:796–805.

Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. 2000. Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci USA 97:6603-6607.

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.

Kashkush K, Khasdan V. 2007. Large-scale survey of cytosine methylation of retrotransposons and the impact of readout transcription from long terminal repeats on expression of adjacent rice genes. Genetics 177:1975-1985.

Kawasaki S, Murakami Y: Genome Analysis of *Lotus japonicus*. Journal of Plant Research 2000:0918-9440.

Keshavareddy G, Rohini S, Ramu SV, Sundaresha S, Kumar AR, Kumar PA, Udayakumar M. 2013.Transgenics in groundnut (*Arachis hypogaea* L.) expressing cry1AcF gene for resistance to *Spodoptera litura* (F.). Physiol Mol Biol Plants 19:343-352.

Khedikar YP, Gowda MV, Sarvamangala C, Patgar KV, Upadhyaya HD, Varshney RK. 2010. A QTL study on late leaf spot and rust revealed one major QTL for molecular breeding for rust resistance in groundnut (*Arachis hypogaea* L.). Theor Appl Genet 121:971-984.

Kidwell MG & Lisch DR. 1997. Transposable elements as source of variation in animals and plants. Proc Natl Acad Sci USA 94:7704-7711.

Knauft DA, Moore KM, Gorbet DW. 1993. Further Studies On The Inheritance Of Fatty Acid Composition In Peanut1. Peanut Science 20:74-76.

Kochert G, Halward T, Branch WD, Simpson CE. 1991. RFLP variability in peanut (*Arachis hypogaea*) cultivars and wild species. Theoretical and Applied Genetics 81:565-570.

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.

Kolano B, Bednara E, Weiss-Schneeweiss H. 2013. Isolation and characterization of reverse transcriptase fragments of LTR retrotransposons from the genome of *Chenopodium quinoa* (Amaranthaceae). Plant Cell Rep 32:1575-1588.

Kottapalli KR, Rakwal R, Shibato J, Burow G, Tissue D, Burke J, Puppala N, Burow M, Payton P. 2009. Physiology and proteomics of the water-deficit stress response in three contrasting peanut genotypes. Plant Cell Environ 32:380-407.

Krapovickas A & Gregory WC. 1994. Taxonomía del género *Arachis* (Leguminosae). Bonplandia 8:1-186.

Krumsiek J, Arnold R, Rattei T. 2007. Gepard: A rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics 23:1026-1028.

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.

Kumar A & Bennetzen JL. 1999. Plant retrotransposons. Annu Rev Genet 33:479-532.

Kumar PV. 2007. Guide to agricutural meteorological practices (WMO No 134), Draft, 3rd Ed., GAMP, Chapter 13B. Disponível em: http://www.agrometeorology.org.

Lander ES, Linton LM, Birren B, Nusbaum C, et al. 2001. Initial sequencing and analysis of the human genome. Nature. 409:860-921.

Laura Ramos M, Fleming G, Chu Ye, 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.

Lavia GI. 1998. Karyotypes of *Arachis palustris* and *A. praecox* (Section *Arachis*), two species with basic chromosome number x=9. Cytologia 63:177-181.

Lavia GI. 1999. Caracterizacion cromosomica del germoplasma de manı. PhD Thesis, Universidad Nacional de Cordoba, Cordoba.

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

Leal-Bertioli SC, Jose AC, Alves-Freitas DM, Moretzsohn MC, Guimarães PM, Nielen S, Vidigal BS, Pereira RW, Pike J, Favero AP, Parniske M, Varshney RK, 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, Guimaraes PM, Brasileiro ACM, Bertioli DJ, De Araujo ACG. 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.

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

Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H. 2007. Transposase-derived transcription factors regulate light signaling in Arabidopsis. Science 318:1302-1305. Erratum in:Science 318:1866.

Lisch D & Bennetzen JL. 2011. Transposable element origins of epigenetic gene regulation. Curr Opin Plant Biol 14:156-161.

Lopes FR, Carazzolle MF, Pereira GA, Colombo CA, Carareto CM. 2008. Transposable elements in Coffea (Gentianales:Rubiacea) transcripts and their role in the origin of protein diversity in flowering plants. Mol Genet Genomics 279:385-401.

Lynch M. 2007. The Origins of Genome Architecture. Sunderland, MA: Sinauer Associates.
Ma JX & Bennetzen JL. 2004. Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404-12410.

Macas J, Neumann P, Navrátilova A. 2007. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and Medicago truncatula. BMC Genomics 8:427.

Mace E, Phong D, Upadhyaya H, Chandra S, Crouch J. 2006. SSR analysis of cultivated groundnut (*Arachis hypogaea* L.) germplasm resistant to rust and late leaf spot disease. Euphytica 152:317-330.

MACHEREY-NAGEL. 2011. Large construct DNA purification. User manual: NucleoBond® Xtra BAC.August/Rev.01

Madlung A. 2013. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. Heredity 110:99-104.

Madsen LH, Fukai E, Radutoiu S, Yost CK, Sandal N, Schauser L, Stougaard J. 2005. LORE1, an active low-copy-number TY3-gypsy retrotransposon family in the model legume *Lotus japonicus*. Plant J 44:372-81.

Martienssen R. 1998. Transposons, DNA methylation and gene control. Trends Genet. 14:263-264.

Martins R & Perez LH. 2006. Amendoim: inovação tecnológica e substituição das importações, Brasil, 1996-2005. Informações Econômicas. Instituto de Economia Agrícola 36:7-19.

Mayr E. 1963. Animal Species and Evolution. Harvard University Press: Cambridge.

McClintock B. 1950. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA 36:344-355.

Meksem K, Ruben E, Zobrisk K et al. 2000. Two large-insert soybean genomic libraries constructed in a binary vector: applications in chromosome walking and genome wide physical mapping. Theoretical and Applied Genetics 101:747–755.

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

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

Miller JT, Dong F, Jackson SA, Song J, Jiang J. 1998. Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics 150:1615-1623.

Miyagi M, Humphry M, Ma ZY, Lambrides CJ, Bateson M, Liu CJ. 2004. Construction of bacterial artificial chromosome libraries and their application in developing PCR-based markers closely linked to a major locus conditioning bruchid resistance in mungbean (*Vigna radiata* L. Wilczek). Theor Appl Genet. 110:151-156.

Moore JP, Vicre-Gibouin M, Farrant JM, Driouich A. 2008. Adaptations of higher plant cell walls to water loss: drought vs desiccation. Physiol Plant 134:237-45.

Moore KM & Knauft DA. 1989. The inheritance of high oleicacid in peanut. J Hered 80: 252–253.

Moore RC & Purugganan MD. 2005. The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol 8:122-128.

Moretzsohn M, Barbosa A, Alves-Freitas D, Teixeira C, Leal-Bertioli S, Guimaraes P, Pereira R, Lopes C, Cavallari M, Valls J, Bertioli D, Gimenes M. 2009. A linkage map for the B-genome of *Arachis* (Fabaceae) and its synteny to the A-genome. BMC Plant Biol 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.

Moretzsohn MC, Hopkins MS, Mitchell SE, Kresovich S, Valls JFM, Ferreira ME. 2004. Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome. BMC Plant Biol 4:11.

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). Theoretical and Applied Genetics 111:1060-1071.

Moretzsohn MC, Gouvea EG, Inglis PW, Leal-Bertioli SC, Valls JF, 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.
Ann Bot 111(1):113-26.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet 37:997-1002.

Nam YW, Penmesta RV, Endre G et al. 1999. Construction of bacterial artificial cho=romosome library of *Medicago truncatula* and identification of clones containing athylene-response genes. Theoretical and Applied Genetics 98:638:646.

Natali L, Santini S, Giordani T, Minelli S, Maestrini P, Cionini PG, Cavallini A. 2006. Distribution of Ty3-gypsy and Ty1-copia-like DNA sequences in the genus *Helianthus* and other Asteraceae. Genome. 49:64-72.

Nelson SC, Simpson CE, Starr JL. 1989. Resistance to *Meloidogyne arenaria* in *Arachis* spp. germoplasm. J NEMATOL 21:654–660.

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

Nielen S, Campos-Fonseca F, Leal-Bertioli S, Guimarães P, Seijo G, Town C, 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. Chromosome Research 18:227-46.

Nielen S, Vidigal BS, Leal-Bertioli SCM, Ratnaparkhe M, Paterson AH, Garsmeur O, D'Hont A, Guimarães PM, Bertioli DJ. 2012. Matita, a new retroelement from peanut: characterization and evolutionary context in the light of the *Arachis* A–B genome divergence. Mol Genet Genomics 287:21-38.

Norden AJ, Gorbet DW, Knauft DA, Young CT. 1987. Variability in oil quality among peanut genotypes in the Florida breeding program. Peanut Sci. 14:7–11.

Orgel LE & Crick FH. 1980. Selfish DNA: the ultimate parasite. Nature 284:604-607.

Ozias-Akins P. Schnall JA, Anderson MF et al. 1993. Regeneration of transgenic peanut plants from stably transformed embryogenic callus. Plant Science Plant Science 93:185-194.

Pardue ML, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL. 2005. Two retrotransposons maintain telomeres in *Drosophila*. Chromosome Res 13:443-453.

Park M, Park J, Kim S, Kwon JK, Park HM, Bae IH, Yang TJ, Lee YH, Kang BC, Choi D. 2012. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. Plant J 69:1018-1029.

Patel M, Jung S, Moore K, Powell G, Ainsworth C, Abbott A. 2004. High-oleate peanut mutants result from a MITE insertion into the FAD2 gene. Theoretical and Applied Genetics 108:1492-1502.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar DS. 2009. The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551-556.

Pearce SR, Pich U, Harrison G, Flavell AJ, Heslop-Harrison JS, Schubert I, Kumar A. 1996. The Ty1-copia group retrotransposons of *Allium cepa* are distributed throughout the chromosomes but are enriched in the terminal heterochromatin. Chromosome Res 4:357-364.

Penãloza APS & Valls JFM. 1997. Contagem do número cromossômico em acessos de *Arachis decora* (Leguminosae). In: Veiga RFA, Bovi MLA, Betti JA, Voltan RBQ (eds) Simpósio Latino-Americano de Recursos Genéticos Vegetais, vol. 1 Campinas. Programas e Resumos. Campinas: IAC/Embrapa-Cenargen, p. 39.

Penãloza APS & Valls JFM. 2005. Chromosome number and satellited chromosome morphology of eleven species of *Arachis* (Leguminosae). Bonplandia 14: 65–72.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. 2006. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262-1269.

Proite K, Leal-Bertioli SC, Bertioli DJ, Moretzsohn MC, da Silva FR, Martins NF, Guimarães PM. 2007. ESTs from a wild *Arachis* species for gene discovery and marker development. BMC Plant Biol 7:7.

Qiao LX, Ding X, Wang HC, Sui JM, Wang JS. 2014. Characterization of the β-1,3-glucanase gene in peanut (*Arachis hypogaea* L.) by cloning and genetic transformation. Genet Mol Res 13:1893-1904.

Ramallo E, Kalendar R, Schulman AH, Martınez-Izquierdo JA. 2008. Reme1, a Copia retrotransposon in melon, is transcriptionally induced by UV light. Plant Mol Biol 66:137-150.

Ramsey J & Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annu Rev Ecol Syst 29:467-501.

Rao NK, Reddy LJ, Bramel PJ. 2003. Potential of wild species for genetic enhancement of some semi-arid food crops. Genetic Resources and Crop Evolution 50: 707-721.

Ray TK, Holly SP, Knauft DA, Abbott AG, Powell GL. 1993. The primary defect in developing seed from the high oleate variety of peanut (*Arachis hypogaea* L.) is the absence of Δ12–desaturase activity. Plant Sci 91:15–21.

Robledo G & Seijo G. 2008. Characterization of the *Arachis* (Leguminosae) D genome using fluorescence in situ hybridization (FISH) chromosome markers and total genome DNA hybridization. Genetics and Molecular Biology 31:717-724.

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, Horsnell T, Rice P, Rajandream MA, Barrell B. 2000. Artemis: sequence visualization and annotation. Bioinformatics 16:944-945.

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

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

Sampedro J, Carey RE, Cosgrove DJ. 2006. Genome histories clarify evolution of the expansin superfamily : new insights from the poplar genome and pine ESTs. J. Plant Res 119:11–21.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA. 74:5463–5467.

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 (SupplementA): 37-44.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nat Genet 1:43-45.

SanMiguel P, Tikhonov A, Jin YK, Motchoulskaia N, Zakharov D, Melake-Berhan A, Springer PS, Edwards KJ, Lee M, Avramova Z, Bennetzen JL. 1996. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768.

Santos RC, Freire RMM, Lima LM. O agronegócio do amendoim no Brasil. 2. ed. rev. e ampl. Brasília: Embrapa, 2013.

Sato S, Isobe S, Asamizu E, Ohmido N, Kataoka R, Nakamura Y, Kaneko T, Sakurai N, Okumura K, Klimenko I, Sasamoto S, Wada T, Watanabe A, Kohara M, Fujishiro T, Tabata

S. 2005. Comprehensive structural analysis of the genome of red clover (*Trifolium pratense* L.). DNA Res. 12:301-364.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S. 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 Plant Sci 3:195-199.

Schulman AH. 2013. Retrotransposon replication in plants. Curr Opin Virol 3:604-614.

Schwarzacher T & Heslop-Harrison P. 2000. Practical in situ hybridization. BIOS Scientific Publishers, Oxford.

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 progenitors of *A. hypogaea* (Leguminosae). American Journal of Botany 91:1294-1303.

Seijo, JG, Lavia GI, Fernández A, Krapovickas A, Ducasse DA, 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.

Seki, M. et al. 2001. Monitoring the Expression Pattern of 1300 *Arabidopsis* Genes under Drought and Cold Stresses by Using a Full-Length cDNA Microarray. The Plant Cell 13:61–72.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong J, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF. 2003. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol 20:633-643.

Sharma HC, Pampapathy G, Dwivedi SL, Reddy LJ. 2003. Mechanisms and diversity of resistance to insect pests in wild relatives of groundnut. J Econ Entomol 96:1886-1897.

Shinozaki K, Yamaguchi-Shinozaki K. 1997. Gene Expression and Signal Transduction in Water-Stress Response. Plant Physiol. 115:327-334.

Shirasawa H, Bertioli DJ, Varshney RK, Moretzsohn MC, Leal-Bertioli SC, Thudi M, Pandey MK, Rami JF, Foncéka D, Gowda MV, Qin H, Guo B, Hong Y, Liang X, Hirakawa H, Tabata S, Isobe S. 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.

Shirasawa K, Hirakawa H, Tabata S, Hasegawa M, Kiyoshima H, Suzuki S, Sasamoto S, Watanabe A, Fujishiro T, Isobe S. 2012. Characterization of active miniature inverted-repeat transposable elements in the peanut genome. Theor Appl Genet 124:1429-1438.

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.

Shizuya H, Birren B, Kim UJ et al. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proceedings of the National Academy of Sciences USA 89:8794 –8797.

Simková H, Safář J, Kubaláková M, Suchánková P, Číhalíková J, Robert-Quatre H, Azhaguvel P, Weng Y, Peng J, Lapitan NL, Ma Y, You FM, Luo MC, Bartoš J, Doležel J. 2011. BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. J Biomed Biotechnol 2011:302543.

Simpson CE, Krapovickas A, Valls, JFM. 2001. History of *Arachis*, including evidence of *A*. hypogaea L. Progenitors. Peanut Science 28:78-80.

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

Singh AK & Moss JP. 1982. Utilization of wild relatives in genetic improvement of *Arachis hypogaea* L. Part 2. Chromosome complements of species in the section *Arachis*. Theoretical and Applied Genetics 61:305-314.

Singh AK & Moss JP. 1984. Utilization of wild relatives in genetic improvement of *Arachis hypogaea L*. Part 5. Genome analysis insection *Arachis* and its implications in gene transfer. Theor Appl Genet 68:355-364.

Singh AK. 1986. Utilization of wild relatives in the genetic improvementof *Arachis hypogaea* L. Part 8. Synthetic amphidiploids and their importance in interspecific breeding. Theor Appl Genet 72:433-439.

Singh KP, Raina SN, Singh AK. 1996. Variation in chromosomal DNA associated with the evolution of *Arachis* species. Genome 39:890-7.

Skirycz A, Inzé D. 2010. More from less: plant growth under limited water. Curr Opin Biotechnol 21:197-203.

Smartt J & Stalker HT. 1982. Speciation and cytogenetics in Arachis. In: Pattee HE & Young CT (eds). Peanut Science and Technology. Yoakum, TX: American Peanut Research and Education Society pp. 21-49.

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

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

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147:195-197.

Smyth DR. 1993. Plant Retrotransposons. In: Verna P (ed.). Control of Gene Expression CRC Press. New York.

Soleimani VD, Baum BR, Johnson DA. 2006. Quantification of the retrotransposon BARE-1 reveals the dynamic nature of the barley genome. Genome 49:389-396.

Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol. 2006-2007.

Soltis P & Soltis D. 2000. The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci USA 97:7051-7057.

Staden R. 1996. The Staden sequence analysis package. Mol Biotechnol 5:233-341.

Stalker HT & Simpson CE. Germplasm resources in Arachis. In: Pattee HE & Stalker HT. (Eds). Advances in Peanut Science, APRES, Stillwater, OK, 1995. Cap. 2, pp 14-53.

Stalker HT. 1991. A new species in section *Arachis* of peanuts with a D genome. American Journal of Botany 78:630-637.

Subrahmanyam P, Ghanekar AM, Nolt BL, Reddy DVR, McDonald D. Resistance to groundnut diseases in wild *Arachis* species In: Proceedings of International Workshop on Cytogenetics of Arachis, ICRISAT, India, Patancheru. 1985.

Subramanian V, Gurtu S, Nageswara Rao RC, Nigam SN. 2000. Identification of DNA polymorphism in cultivated groundnut using random amplified polymorphic DNA (RAPD) assay. Genome 43:656-660.

Sun HY, Dai HY, Zhao GL, Ma Y, Ou CQ, Li H, Li LG, Zhang ZH. 2008. Genome-wide characterization of long terminal Repeat-retrotransposons in apple reveals the differences in heterogeneity and copy number between Ty1-copia and Ty3-gypsy retrotransposons. J Integr Plant Biol. 50:1130-1139.

Tahara M, Aoki T, Suzuka S, Yamashita H, Tanaka M, Matsunaga S, Kokumai S. 2004. Isolation of an active element from a high-copy-number family of retrotransposons in the sweet potato genome. Mol Genet Genomics 272:116-27.

Taiz L, Zeiger E, 2006. Paredes celulares: estrutura, biogênese e expansão. In: Artmed, ed. Fisiologia Vegetal. 342.

Tallury SP, Hilu KW, Milla SR, Friend SA, Alsaghir M, Stalker HT, Quandt D. 2005. Genomic affinities in *Arachis* section *Arachis* (Fabaceae): molecular and cytogenetic evidence. Theoretical and Applied Genetics 111:1229–1237.

Temsch EM, Greilhuber J. 2000. Genome size variation in *Arachis hypogaea* and *A. monticola* re-evaluated. Genome 43:449-51.

Temsch EM, Greilhuber J. 2001. Genome size in *Arachis duranensis*: a critical study. Genome 44:826-30.

Tenaillon MI, Hollister JD, Gaut BS. 2010. A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471-8.

Thomas CA. 1971. The genetic organization of chromosomes. Ann Rev Genet 5:237-256.

USDA-FAS. 2013a. Table 13: Peanut Area, Yield, and Production. Acesso em 20 de junho de 2013, disponível em USDA - United States Department of Agriculture/FAS - Foreign Agricultural Service:

USDA-FAS. 2013b. Table 01: Major Oilseeds: World Supply and Distribution (Commodity View). Acesso em 20 de junho de 2013, disponível em USDA - United States Department of Agriculture/FAS - Foreign Agricultural Service: http://www.fas.usda.gov/psdonline/psdReport.aspx?hidReportRetrievalName=Table+01%3a+Major+Oilseeds%3a+World+Supply+and+Distribution+(Commodity+View)++++++++++++++++++++++++++&hidReportRetrievalID=531&hidReportRetrievalTemplateID=5

Valls JFM & Simpson CE. 2005. New species of *Arachis* from Brazil, Paraguay, and Bolivia. Bonplandia 14:35-64.

Vanhouten W, MacKenzie S. 1999. Construction and characterization of a common bean bacterial artificial chromosome library. Plant Mol Biol 40:977-83.

Varshney RK, Bertioli DJ, Moretzsohn MC, Vadez V, Krishnamurthy L, Aruna R, Nigam SN, Moss BJ, Seetha K, Ravi K, He G, Knapp SJ, Hoisington DA. 2009. The first SSR-based genetic linkage map forcultivated groundnut (*Arachis hypogaea* L.). Theor Appl Genet 118:729-739.

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, Nobrega MP, Carrer H, França SC, Bacci Júnior M, Goldman MH, Gomes SL, Nunes LR, Camargo LE, Siqueira WJ, Van Sluys MA, Thiemann OH, Kuramae EE, Santelli RV, Marino CL, Targon ML, Ferro JA, Silveira HC, Marini DC, Lemos EG, Monteiro-Vitorello CB, Tambor JH, Carraro DM, Roberto PG, Martins VG, Goldman GH, de Oliveira RC, Truffi D, Colombo CA, Rossi M, de Araujo PG, Sculaccio SA, Angella A, Lima MM, de Rosa Júnior VE, Siviero F, Coscrato VE, Machado MA, Grivet L, Di Mauro SM, Nobrega FG, Menck CF, Braga MD, Telles GP, Cara FA, Pedrosa G, Meidanis J, Arruda P. 2003. Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13:2725-2735.

Vicient CM, Schulman AH. 2002. Copia-like retrotransposon in the rice genoma: few and assorted. Genome Lett 1:35-47.

Vitte C & Panaud O. 2005. LTR retrotransposons and flowering plant genome size:emergence of the increase/decrease model. Cytogenet Genome Res 110:91-107.

Vitte C, Estep MC, Leebens-Mack J, Bennetzen JL. 2013. Young, intact and nested retrotransposons are abundant in the onion and asparagus genomes. Ann Bot 112:881-9.

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, Ameline-Torregrosa C, Cannon E, Chacko B, Couloux A, Dalwani A, Denny R, Deshpande S, Egan AN, Glover N, Howell S, Ilut D, Lai H, Del Campo SM, Metcalf M, O'Bleness M, Pfeil BE, Ratnaparkhe MB, Samain S, Sanders I, Ségurens B, Sévignac M, Sherman-Broyles S, Tucker DM, Yi J, Doyle JJ, Geffroy V, Roe BA, Maroof MAS, Young ND, Innes RW. 2008. Replication of non-autonomous retroelements in soybean appears to be both recent and common. Plant Physiology 148:1760-1771.

Wei L, Xiao M, An Z, Ma B, Mason AS, Qian W, Li J, Fu D. 2013. New insights into nested long terminal repeat retrotransposons in *Brassica species*. Mol Plant 6:470-82.

Wessler SR, Carrington JC. 2005. The consequences of gene and genome duplication in plants. Curr Opin Plant Biol 8:119-21.

Wessler SR. 2006. Transposable elements and the evolution of eukaryotic genomes.. Proc Natl Acad Sci U S A 103:17600-1.

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, Buchmann JP, Keller B. 2010. Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res 20:1229-1237.

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.

Witte CP, Le QH, Bureau T, Kumar A. 2001.Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci USA 98:13778-13783.

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. American Journal of Botany 91: 1846-1862.

Wu N, Matand K, Wu H, Li B, Li Y, Zhang X, He Z, Qian J, Liu X, Conley S, Bailey M, Acquaah G. 2013. De novo next-generation sequencing, assembling and annotation of *Arachis hypogaea* L. Spanish botanical type whole plant transcriptome. Theor Appl Genet 126:1145-1149.

Xia Z, Watanabe S, Chen Q, Sato S, Harada K. 2009. A novel manual pooling system for preparing three-dimensional pools of a deep coverage soybean bacterial artificial chromosome library. Mol Ecol Resour 9:516-24.

Xiong Y & Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J 9:3353-3362.

Xu Z & Wang H. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265-W268.

Yim YS, Moak P, Sanchez-Villeda H et al. 2007. A BAC pooling strategy combined with PCR-based screenings in a large, highly repetitive genome enables integration of the maize genetic and physical maps. BMC Genomics 8:47.

Yin D, Wang Y, Zhang X, Li H, Lu X, Zhang J, Zhang W, Chen S. 2013. De novo assembly of the peanut (*Arachis hypogaea* L.) seed transcriptome revealed candidate unigenes for oil accumulation pathways. PLoS One 8:e73767.

Yoo MJ, Szadkowski E, Wendel JF. 2013. Homoeolog expression bias and expression level dominance in allopolyploid cotton. Heredity 110:171-180.

Young ND, Weeden N, Kochert G. 1996. Genome mapping in legumes. In: Paterson A. (ed.). Genome Mappong in Plants. Landes, Austin, Texas, pp.212-227.

Yu S, Pan L, Yang Q, Min P, Ren Z, Zhang H. 2008. Comparison of the Δ12 fatty acid desaturase gene between high-oleic and normal-oleic peanut genotypes. J. Genet. Genomics 35:679–685.

Yüksel B & Paterson AH. 2005. Construction and characterization of a peanut HindIII BAC library. TAG 111:630-639.

Zachar Z, Davidson D, Garza D, Bingham PM. 1985. A detailed developmental and structural study of the transcriptional effects of insertion of the Copia transposon into the white locus of *Drosophila melanogaster*. Genetics 111:495-515.

Zhang J, Liang S, Duan J, Wang J, Chen S, Cheng Z, Zhang Q, Liang X, Li Y. 2012. De novo assembly and characterisation of the transcriptome during seed development, and generation of genic-SSR markers in peanut (*Arachis hypogaea* L.). BMC Genomics 13:90.

Zhang N & Hasenstein KH. 2000. Distribution of Expansins in Graviresponding Maize Roots. Plant Cell Physiol 41:1305–1312.

Zhang P, Li W, Fellers J, Friebe B, Gill BS. 2004. BAC-FISH in wheat identifies chromosome landmarks consisting of different types of transposable elements. Chromosoma 112:288-299.

Zhou L, Mitra R, Atkinson PW, Hickman AB, Dyda F, Craig NL. 2004. Transposition of hAT elements links transposable elements and V(D)J recombination. Nature 432:995-1001.

Anexos

Anexo 1: Tabela contendo as médias das estimativas de datas de transposição para famílias de retrotransposons LTR individualmente ou relacionadas.

Genoma A		Genoma B	
Retrotransposon LTR	**Média/idade/Ma**	**Retrotransposon LTR**	**Média/idade/Ma**
A0_FIDEL_autonomous-Ty3-type	2,32	B0_FIDEL_autonomous-Ty3-type	1,77
A3_Feral_non-autonomous-type		B3_Feral_non-autonomous-type	
A69_Apolo_autonomous-Ty3-type	2,61	B69_Apolo_autonomous-Ty3-type	1,62
A141_Polo_non-autonomous-type		B141_Polo_non-autonomous-type	
A14_Pipoka_autonomous-Ty3-type	1,97	B14_Pipoka_autonomous-Ty3-type	1,72
A104_Pipa1_non-autonomous-type		B104_Pipa1_non-autonomous-type	
A157_Curu_autonomous-Ty3-type	2,13	B49_Pipa2_non-autonomous-type	
A26_Golf_non-autonomous-type		B8_Pipa3_non-autonomous-type	
A113_Hemera_autonomous-type		A157_Curu_autonomous-Ty3-type	2,1
A38_Hera_non-autonomous-type		B6_Bravo_non-autonomous-type	
A155_Silverio_autonomous-Ty3-type	2,13	B26_Golf_non-autonomous-type	
A156_Silvia_non-autonomous-type		B113_Hemera_autonomous-type	
A145_Silver_non-autonomous-type		B38_Hera_non-autonomous-type	
A105_Gordo_non-autonomus-type	1,9	B155_Silverio_autonomous-Ty3-type	1,74
A57_Mico_autonomous-Ty3-type	1,16	A156_Silvia_non-autonomous-type	
A5_Eros_autonomous-Ty3-type	2,22	B145_Silver_non-autonomous-type	
A89_Eris_non-autonomous-type		B105_Gordo_non-autonomus-type	1,64
A84_RE128_autonomous-Ty1-type	1,29	B57_Mico_autonomous-Ty3-type	1,25
A2_Matita_autonomous-Ty1-type	1,61	B5_Eros_autonomous-Ty3-type	1,66
A99_Medusa_autonomous-Ty3-type	2,78	B89_Eris_non-autonomous-type	
A33_Musa_non-autonomous-type	3,07	B84_RE128_autonomous-Ty1-type	1,24
A24_Athena_non-autonomous-type	2,45	B2_Matita_autonomous-Ty1-type	1,26
A137_Sofi_autonomous-Ty3-type	2,28	B99_Medusa_autonomous-Ty3-type	2,46
A39_Yara_autonomous-Ty1-type	1,79	B33_Musa_non-autonomous-type	2,33
A112_Lima_autonomous-Ty3-type	1,22	B24_Athena_non-autonomous-type	1,92
A118_Bela_autonomous-Ty1-type	1,64	B137_Sofi_autonomous-Ty3-type	1,77
A132_Zoe_autonomous-Ty1-type	1,41	B39_Yara_autonomous-Ty1-type	1,87
A138_Zia_autonomous-Ty1-type	2,75	B112_Lima_autonomous-Ty3-type	1,42
A16_Grilo_autonomous-Ty3-type	1,47	B118_Bela_autonomous-Ty1-type	1,92
A142_Gilo_non-autonomous-type		B132_Zoe_autonomous-Ty1-type	1,19
A143_Yuri_autonomous-Ty1-type	1,32	B138_Zia_autonomous-Ty1-type	0,97
A29_RE128_autonomous-Ty1-type	1,31	B16_Grilo_autonomous-Ty3-type	1,87
A11_Delta_autonomous-Ty3-type	1,84	B142_Gilo_non-autonomous-type	
A11_Charlie_non-autonomous-type	1,55	B143_Yuri_autonomous-Ty1-type	2,07
A121_Doros_autonomous-Ty3-type	0,95	B29_RE128_autonomous-Ty1-type	1,11
A122_Duran_non-autonomous-type	1,45	B11_Delta_autonomous-Ty3-type	1,59
A150_Dan_non-autonomous-type		B11_Charlie_non-autonomous-type	1,4
A126_Venon_autonomous-Ty1-type	0,7	B121_Doros_autonomous-Ty3-type	1,25
A65_Foxtrot_autonomous-Ty1-type	1,28	B122_Duran_non-autonomous-type	1,14
A62_Saturno_autonomous-Ty3-type	1,48	B150_Dan_non-autonomous-type	
A134_Nemesis_non-autonomous-type	1,45	B154_Paco_autonomous-Ty3-type	2,13
A123_Talita_autonomous-Ty1-type	1,05	B126_Venon_autonomous-Ty1-type	1,59
A135_Maia_non-autonomous-type	2,04	B65_Foxtrot_autonomous-Ty1-type	1,26
A19_Netuno_autonomous-Ty3-type	1,1	B62_Saturno_autonomous-Ty3-type	1,63
A136_Elpis_non-autonomous-type	0,9	B134_Nemesis_non-autonomous-type	1,12
A140_Joka_autonomous-Ty1-type	1,83	B123_Talita_autonomous-Ty1-type	1,15
A1_Hermes_autonomous-Ty3-type	0,81	B135_Maia_non-autonomous-type	1,3
A144_Diva_autonomous-Ty1-type	1,1	B19_Netuno_autonomous-Ty3-type	1,16
A9_Girino_autonomous-Ty3-type	2,49	B136_Elpis_non-autonomous-type	1,15
A149_Lulu_non-autonomous-type	0,95	B140_Joka_autonomous-Ty1-type	1,67
A60_Venus_autonomous-Ty3-type	0,69	B1_Hermes_autonomous-Ty3-type	1,27
A124_Ariel_autonomous-Ty3-type	1,66	B144_Diva_autonomous-Ty1-type	1,16
A125_Buba_autonomous-Ty3-type	1,62	B9_Girino_autonomous-Ty3-type	2,1
A127_Dakota_autonomous-Ty1-type	0,71	B149_Lulu_non-autonomous-type	1,45
A152_Agnus_non-autonomous-type	2,39	B60_Venus_autonomous-Ty3-type	1,07
A88_Papa_autonomous-Ty1-type	0,76	B125_Buba_autonomous-Ty3-type	2,02
A50_Hotel_autonomous-Ty1-type	1,22	B127_Dakota_autonomous-Ty1-type	1,52
A148_Edore_autonomous-Ty3-type	1,67	B88_Papa_autonomous-Ty1-type	1,22
A71_Juliett_autonomous-Ty1-type	1,25	B50_Hotel_autonomous-Ty1-type	2,05
A139_George_autonomous-Ty3-type	1,06	B148_Edore_autonomous-Ty3-type	1,43
A146_Golden_autonomous-Ty1-type	0,41	B71_Juliett_autonomous-Ty1-type	1,26
A147_Omega_autonomous-Ty3-type	0,97	B139_George_autonomous-Ty3-type	0,67
A91_Whiskey_autonomous-Ty3-type	1,69	B146_Golden_autonomous-Ty1-type	0,49
A117_Xray_non-autonomous-type	1,3	B147_Omega_autonomous-Ty3-type	0,44
A130_James_autonomous-Ty1-type	1,29	B91_Whiskey_autonomous-Ty3-type	2,06
A97_Mike_autonomous-Ty3-type	1,2	B117_Xray_non-autonomous-type	1,7
A67_India_autonomous-Ty1-type	1,68	B130_James_autonomous-Ty1-type	1,74
A51_Victor_autonomous-Ty3-type	1,05	B97_Mike_autonomous-Ty3-type	1,79
A119_Bola_autonomous-Ty1-type	1,43	B67_India_autonomous-Ty1-type	0,55
A31_Kilo_autonomous-Ty1-type	0,86	B51_Victor_autonomous-Ty3-type	0,7
A54_Yankee_autonomous-Ty1-type	0,91	B119_Bola_autonomous-Ty1-type	2,45
A17_Romeo_autonomous-Ty1-type	2,28	B31_Kilo_autonomous-Ty1-type	1,54
A131_Phenix_autonomous-Ty3-type	0,32	B54_Yankee_autonomous-Ty1-type	0,99
A61_Oscar_autonomous-Ty1-type	1,47	B133_Kirke_autonomous-Ty1-type	1,42
A32_Tango_autonomous-Ty1-type	1,22	B17_Romeo_autonomous-Ty1-type	1,93
A129_Jasper_autonomous-Ty1-type	1,09	B81_Quebec_autonomous-Ty3-type	0,98
A151_Vipe_non-autonomous-type	2,04	B131_Phenix_autonomous-Ty3-type	1,07
A78_Echos_autonomous-Ty1-type	2,45	B25_Sierra_autonomous-Ty1-type	1,28
A43_Alpha_autonomous-Ty1-type	1,18	B61_Oscar_autonomous-Ty1-type	1,95
		B128_Doris_autonomous-Ty1-type	1,15
		B32_Tango_autonomous-Ty1-type	0,87
		B7_November_autonomous-Ty1-type	1,61
		B129_Jasper_autonomous-Ty1-type	1,87
		B78_Echos_autonomous-Ty1-type	1,98
		B43_Alpha_autonomous-Ty1-type	1,41

Anexo 2: Tabela contendo as listas dos genes putativos preditos para as regiões 1-4.

Região 1					
Arachis duranensis_scaffold 45_420.200-627.000 (206.800 pb)			*Arachis ipaënsis_scaffold* 47_531.200-780.000 (248.000 pb)		
Gene	Predição da função	pfam-ID	Gene	Predição da função	pfam-ID
	-		Gene 1	**putative protein**	
Gene 1	putative protein		Gene 2	putative protein	
Gene 2	similar to Pfam SGL family	PF08450.6	Gene 3	similar to Pfam SGL family	PF08450.6
	-		Gene 4	**similarity to Pfam RVT_1**	PF00078.21
	-		Gene 5	**putative protein**	
Gene 3	similar to Pfam SGL family	PF08450.6	Gene 6	similar to Pfam SGL family	
Gene 4	similar to Pfam WD40	PF00400.26	Gene 7	similar to Pfam WD40	PF00400.26
Gene 5	similar to Pfam Pentapeptide	PF00805.16	Gene 8	similar to Pfam Pentapeptide	PF00805.16
Gene 6	putative protein		Gene 9	putative protein	
Gene 7	similarity to Pfam Ank (Ankyrin repeat)	PF00023.24	Gene 10	similarity to Pfam Ank (Ankyrin repeat)	PF00023.24
Gene 8	similarity to Pfam Meth_synt_2	PF01717.12	Gene 11	similarity to Pfam Meth_synt_2	PF01717.12
Gene 9	similarity to Pfam dsrm (Double-stranded	PF00035.19	Gene 12	similarity to Pfam dsrm (Double-stranded RI	PF00035.19
Gene 10	similar to Pfam Annexin	PF00191.14	Gene 13	similar to Pfam Annexin	PF00191.14
Gene 11	similarity to Pfam C2 (C2 domain)	PF00168.24	Gene 14	similarity to Pfam C2 (C2 domain)	PF00168.24
Gene 12	similarity to Pfam PRT_C (Plant phosphoril	PF08372.4	Gene 15	similarity to Pfam PRT_C (Plant phosphoribo	PF08372.4
Gene 13	similarity to Pfam Glyco_hydro_3	PF00933.15	Gene 16	similarity to Pfam Glyco_hydro_3	PF00933.15
Gene 14	putative protein		Gene 17	putative protein	
Gene 15	similarity to Pfam Glyco_hydro_3_C	PF01915.16	Gene 18	similarity to Pfam Glyco_hydro_3_C	PF01915.16
Gene 16	similarity to Pfam Glyco_hydro_3_C	PF01915.16	Gene 19	similarity to Pfam Glyco_hydro_3_C	PF01915.16
Gene 17	putative protein		Gene 20	putative protein	
Gene 18	putative protein		Gene 21	putative protein	
Gene 19	similarity to Pfam AP2 (Apetala 2)	PF00847.14	Gene 22	similarity to Pfam AP2 (Apetala 2)	PF00847.14
Gene 20	similar to Pfam Tubulin_C	PF03953.11	Gene 23	similar to Pfam Tubulin_C	PF03953.11
Gene 21	similarity to Pfam Sod_Cu (Superoxide disr	PF00080.14	Gene 24	similarity to Pfam Sod_Cu (Superoxide dismu	PF00080.14
Gene 22	similar to Pfam IF4E (Eukaryotic translatio	PF01652.12	Gene 25	similar to Pfam IF4E (Eukaryotic translation i	PF01652.12
Gene 23	putative protein		Gene 26	putative protein	
Gene 24	similarity to Pfam ELFV_dehydrog_N	PF02812.1	Gene 27	similarity to Pfam ELFV_dehydrog_N	PF02812.1
Gene 25	similarity to Pfam GalP_UDP_transf	PF01087.1	Gene 28	similarity to Pfam GalP_UDP_transf	PF01087.1
Gene 26	putative protein		Gene 29	putative protein	
Gene 27	putative protein		Gene 30	putative protein	
Gene 28	similarity to Pfam PPR (Pentatricopeptide	PF01535.14	Gene 31	similarity to Pfam PPR (Pentatricopeptide re	PF01535.14
Gene 29	similarity to Pfam adh_short (Short-chain	PF00106.19	Gene 32	similarity to Pfam adh_short (Short-chain de	PF00106.19
Gene 30	similar to Pfam Glyco_hydro_17 (Glycosid	PF00332.12	Gene 33	similar to Pfam Glyco_hydro_17 (Glycoside l	PF00332.12

Região 2					
Arachis duranensis_scaffold 45_904.200-1.119.800 (215.600 pb)			*Arachis ipaënsis_scaffold* 47_1.018.600-1.280.800 (200.200 pb)		
Gene	Predição da função	pfam-ID	Gene	Predição da função	pfam-ID
Gene 1	putative protein		Gene 1	putative protein	
Gene 2	similarity to Pfam DUF566	PF04484.6	Gene 2	similarity to Pfam DUF566	PF04484.6
	-		Gene 3	**similar to Pfam PMD**	PF10536.3
	-		Gene 4	**putative protein**	
	-		Gene 5	**putative protein**	
Gene 3	similarity to Pfam DUF566	PF04484.6	Gene 6	similarity to Pfam DUF566	PF04484.6
Gene 4	similarity to Pfam DUF566	PF04484.6	Gene 7	similarity to Pfam DUF566	PF04484.6
Gene 5	similarity to Pfam DUF566	PF04484.6	Gene 8	similarity to Pfam DUF566	PF04484.6
Gene 6	putative protein		Gene 9	putative protein	
Gene 7	similarity to Pfam MULE	PF10551.3	Gene 10	putative protein	
Gene 8	**putative protein**			-	
Gene 9	**putative protein**			-	
Gene 10	**putative protein**			-	
Gene 11	**similar to Pfam B3**	PF02362.15		-	
Gene 12	**putative protein**			-	
Gene 13	**putative protein**			-	
Gene 14	**putative protein**			-	
Gene 15	**putative protein**			-	
Gene 16	**putative protein**			-	
Gene 17	similarity to Pfam O-FucT	PF10250.3	Gene 11	similarity to Pfam O-FucT	PF10250.3
Gene 18	**putative protein**			-	
Gene 19	putative protein		Gene 12	putative protein	
Gene 20	similar to Pfam E1-E2_ATPase	PF00122.14	Gene 13	similar to Pfam E1-E2_ATPase	PF00122.14
Gene 21	**putative protein**			-	
Gene 22	similar to Pfam RVT_1	PF00078.21	Gene 14	similar to Pfam RVT_1	PF00078.21
Gene 23	putative protein		Gene 15	putative protein	
Gene 24	similar to Pfam PPR_1	PF12854.1	Gene 16	similar to Pfam PPR_1	PF12854.1
Gene 25	putative protein		Gene 17	putative protein	
Gene 26	**similarity to Pfam PORR**	PF11955.2		-	
Gene 27	similarity to Pfam Glyco_hydro_28	PF00295.11	Gene 18	similarity to Pfam Glyco_hydro_28	PF00295.11
	-		Gene 19	similarity to Pfam Glyco_hydro_28	PF00295.11
Gene 28	putative protein		Gene 20	putative protein	
Gene 29	similar to Pfam PPR_1	PF12854.1	Gene 21	similar to Pfam PPR_1	PF12854.1
Gene 30	putative protein		Gene 22	putative protein	
	-		Gene 23	**putative protein**	
	-		Gene 24	**putative protein**	
	-		Gene 25	**putative protein**	
Gene 31	similarity to Pfam bZIP_1	PF00170.15	Gene 26	similarity to Pfam bZIP_1	PF00170.15
Gene 32	similarity to Pfam MULE	PF10551.3	Gene 27	similarity to Pfam MULE	PF10551.3
Gene 33	similarity to Pfam Pectinesterase	PF01095.13	Gene 28	similarity to Pfam Pectinesterase	PF01095.13
Gene 34	putative protein		Gene 29	similarity to Pfam Pectinesterase	PF01095.13
Gene 35	putative protein		Gene 30	putative protein	
Gene 36	putative protein		Gene 31	putative protein	
Gene 37	putative protein		Gene 32	putative protein	
	-		Gene 33	**putative protein**	
	-		Gene 34	**putative protein**	

Região 3					
Arachis duranensis_scaffold 45_1.177.000-1.412.400 (235.400 pb)			*Arachis ipaënsis_scaffold* 47_1.018.600-1.218.800 (252.000 pb)		
Gene	Predição da função	pfam-ID	Gene	Predição da função	pfam-ID
Gene 1	similar to Pfam mTERF	PF02536.8	Gene 1	similar to Pfam mTERF	PF02536.8
Gene 2	putative protein		Gene 2	putative protein	
Gene 3	similar to Pfam MULE	PF10551.3	Gene 3	similar to Pfam MULE	PF10551.3
Gene 4	**similarity to Pfam PA**	PF02225.16		-	
Gene 5	Pfam Cytochrom_C1	PF02167.9	Gene 4	Pfam Cytochrom_C1	PF02167.9
Gene 6	**putative protein**			-	
Gene 7	putative protein		Gene 5	putative protein	
Gene 8	similarity to Pfam ABC_membrane	PF00664.17	Gene 6	similarity to Pfam ABC_membrane	PF00664.17
Gene 9	putative protein		Gene 7	putative protein	
Gene 10	**putative protein**			-	
Gene 11	similarity to Pfam Pkinase_Tyr	PF07714.1	Gene 8	similarity to Pfam Pkinase_Tyr	PF07714.1
Gene 12	similarity to Pfam Myb_DNA-bind_3	PF12776.1	Gene 9	similarity to Pfam Myb_DNA-bind_3	PF12776.1
Gene 13	similar to Pfam PPR	PF01535.14	Gene 10	similar to Pfam PPR	PF01535.14
Gene 14	similar to Pfam Pkinase_Tyr	PF07714.1	Gene 11	similar to Pfam Pkinase_Tyr	PF07714.1
Gene 15	similar to Pfam Pkinase_Tyr	PF07714.1	Gene 12	similar to Pfam Pkinase_Tyr	PF07714.1
	-		Gene 13	**putative protein**	
Gene 16	putative protein		Gene 14	putative protein	
	-		Gene 15	**similarity to Pfam RVT_1**	PF00078.21
Gene 17	putative protein		Gene 16	putative protein	
Gene 18	putative protein		Gene 17	putative protein	
Gene 19	similar to Pfam NB-ARC	PF00931.16	Gene 18	similar to Pfam NB-ARC	PF00931.16
Gene 20	similar to Pfam p450	PF00067.16	Gene 19	similar to Pfam p450	PF00067.16
Gene 21	**putative protein**			-	
Gene 22	**putative protein**			-	
	-		Gene 20	**similarity to Pfam Exo_endo_phos**	PF03372.17
	-		Gene 21	**similarity to Pfam RVT_1**	PF00078.21
Gene 23	similarity to Pfam Choline_kinase	PF01633.14	Gene 22	similarity to Pfam Choline_kinase	PF01633.14
Gene 24	putative protein		Gene 23	putative protein	
Gene 25	similar to Pfam DEAD	PF00270.23	Gene 24	similar to Pfam DEAD	PF00270.23
Gene 26	putative protein		Gene 25	putative protein	
Gene 27	similar to Pfam Sulfotransfer_1	PF00685.21	Gene 26	similar to Pfam Sulfotransfer_1	PF00685.21
	-		Gene 27	**putative protein**	
	-		Gene 28	**similar to Pfam RVT_1**	PF00078.21
Gene 28	putative protein		Gene 29	putative protein	
Gene 29	similar to Pfam AICARFT_IMPCHas	PF01808.12	Gene 30	similar to Pfam AICARFT_IMPCHas	PF01808.12
Gene 30	**putative protein**			-	
Gene 31	similar to Pfam Mannosyl_trans	PF05007.7	Gene 31	similar to Pfam Mannosyl_trans	PF05007.7
Gene 32	putative protein		Gene 32	putative protein	
Gene 33	similar to Pfam THOC7	PF05615.7	Gene 33	similar to Pfam THOC7	PF05615.7
	-		Gene 34	**similarity to Pfam Glyco_hydro_1**	PF00232.12
Gene 34	similarity to Pfam p450	PF00067.16	Gene 35	similarity to Pfam p450	PF00067.16

Região 4					
*Arachis duranensis*_ scaffold 45_1570800-1777600 (206.800 pb)			*Arachis ipaensis*_ scaffold 47_1790400-1969600 (179.200 pb)		
Gene	Predição da função	pfam-ID	Gene	Predição da função	pfam-ID
	-		Gene 1	**putative protein**	
Gene 1	similarity to Pfam efhand	PF00036.26	Gene 2	similarity to Pfam efhand	PF00036.26
Gene 2	similarity to Pfam FMN_red	PF03358.9	Gene 3	similarity to Pfam FMN_red	PF03358.9
	-		Gene 4	**putative protein**	
	-		Gene 5	**similarity to Pfam RVT_1**	PF00078.21
Gene 3	similarity to Pfam FMN_red	PF03358.9	Gene 6	similarity to Pfam FMN_red	PF03358.9
	-		Gene 7	**putative protein**	
Gene 4	similarity to Pfam PIP5K	PF01504.12	Gene 8	similarity to Pfam PIP5K	PF01504.12
Gene 5	**similarity to Pfam MULE**	PF10551.3		-	
Gene 6	similar to Pfam DUF702	PF05142.6	Gene 9	similar to Pfam DUF702	PF05142.6
	-		Gene 10	**similar to Pfam RVT_2**	PF07727.8
	-		Gene 11	**similar to Pfam Retrotrans_gag**	PF03732.11
Gene 7	putative protein		Gene 12	putative protein	
	-		Gene 13	**putative protein**	
	-		Gene 14	**putative protein**	
Gene 8	similarity to Pfam PTR2	PF00854.15	Gene 15	similarity to Pfam PTR2	PF00854.15
Gene 9	similarity to Pfam PTR2	PF00854.15	Gene 16	similarity to Pfam PTR2	PF00854.15
Gene 10	similarity to Pfam PTR2	PF00854.15	Gene 17	similarity to Pfam PTR2	PF00854.15
	-		Gene 18	**putative protein**	
	-		Gene 19	**similar to Pfam PMD**	PF10536.3
Gene 11	**similarity to Pfam PTR2**	PF00854.15		-	
Gene 12	**similarity to Pfam PTR2**	PF00854.15		-	
Gene 13	similarity to Pfam PTR2	PF00854.15	Gene 20	similarity to Pfam PTR2	PF00854.15
Gene 14	putative protein		Gene 21	putative protein	
Gene 15	similar to Pfam SNF2_N	PF00176.17	Gene 22	similar to Pfam SNF2_N	PF00176.17
Gene 16	putative protein		Gene 23	putative protein	
Gene 17	putative protein		Gene 24	putative protein	
Gene 18	putative protein		Gene 25	putative protein	
Gene 19	**putative protein**			-	
Gene 20	**similar to Pfam RVT_1**	PF00078.21		-	
Gene 21	**putative protein**			-	
Gene 22	similarity to Pfam Pectinesterase	PF01095.13	Gene 26	similarity to Pfam Pectinesterase	PF01095.13
Gene 23	similarity to Pfam Glutaredoxin	PF00462.18	Gene 27	similarity to Pfam Glutaredoxin	PF00462.18
Gene 24	similar to Pfam ArfGap	PF01412.12	Gene 28	similar to Pfam ArfGap	PF01412.12
Gene 25	similarity to Pfam IF-2B	PF01008.11	Gene 29	similarity to Pfam IF-2B	PF01008.11
Gene 26	similarity to Pfam UBA	PF00627.25	Gene 30	similarity to Pfam UBA	PF00627.25
Gene 27	similar to Pfam F-box-like	PF12937.1	Gene 31	similar to Pfam F-box	PF00646.27
Gene 28	similarity to Pfam RVT_2	PF07727.8	Gene 32	similarity to Pfam RVT_2	PF07727.8
Gene 29	similarity to Pfam RNA_pol_Rpb5_N	PF03871.8	Gene 33	similarity to Pfam RNA_pol_Rpb5_C	PF01191.13
Gene 30	similarity to Pfam RNA_pol_Rpb5_N	PF03871.8	Gene 34	similarity to Pfam RNA_pol_Rpb5_C	PF01191.13
Gene 31	putative protein		Gene 35	putative protein	
Gene 32	putative protein		Gene 36	putative protein	
	-		Gene 37	putative protein	
	-		Gene 38	putative protein	

Anexo 3: *Scripts Perl* utilizados neste estudo.

- *Script* A

```
open ARQUIVO1, "$ARGV[0]";
open ARQUIVO2, "$ARGV[1]";
foreach(<ARQUIVO1>){
		if ($_ =~ /^\[(\d+)]\sscaffold_(\d+)/){
		$scaffold_indice = $1;
		$scaffold_atual = $2;
		$cont ++;
		}
		if ($_ =~ /^\[1]\sscaffold_(\d+)/){
		$cont2 ++;
		$scaffold_numero[$cont2] = $1;
		}
		if ($_ =~ /^Location\s:\s(\d+)\s-\s(\d+)/){
		$scaffold[$cont] .= ">[$scaffold_indice] scaffold_$scaffold_atual\|$1-$2";
		$scaffold_inicio[$cont] = $1;
		$scaffold_fim[$cont] = $2;
		}
}
#print
"scaffold_indice$scaffold_indice\nscaffold_atual:$scaffold_atual\ncont:$cont\n\ncont2:$cont2\nscaffold_numero:@scaffold_numero\nscaffo
ld:@scaffold\nscaffold inicio:@scaffold_inicio\n scaffold_fim:@scaffold_fim";
$cont3 = 0;

$indicador = 0;
foreach(<ARQUIVO2>){
		$atual = $_;
		if($indicador == 1){
			foreach(@scaffold){
				if($_ =~ /scaffold_$x\|/){
				$length = $scaffold_fim[$cont3] - $scaffold_inicio[$cont3];
				$sequencia = substr $atual, ($scaffold_inicio[$cont3]-1), ($length+1);
				print "$scaffold[$cont3]\n$sequencia\n";
				}
			$cont3 ++;
			}
		$cont3 = 0;
		$indicador = 0;
		}
			foreach(@scaffold_numero){
				if($atual =~ /scaffold_$_\s/){
				$x = $_;
				$indicador = 1;
				}
			}
}
#print $final;
```

- *Script* B

```
$seq = $ARGV[0];
		open (FILE,$seq) or die "no file1 named";
		@seqs_0 = <FILE>;
#this opens the blast file named in $ARGV[1]#
#$lookup = shift @ARGV[1];  # ID to extract - would this be an alternative simpler way to work?
$lookup = $ARGV[1];
		open (FILE,$lookup) or die "no file2 named";
		@lookup_0 = <FILE>;
#this opens the output file named as third argument -- DONT FORGET TO CLOSE THE FILE AFTER USE#
   $outfile = $ARGV[2];
   open (OUT,">$outfile") or die "could not open output file";
#Then we start the splitting up the fasta file#
$seqs_1 = join ("",@seqs_0);
```

```
@seqs_2 = split (/>/,$seqs_1);
#the first element of @fasta is an empty element
#so we get rid of this with shift
shift (@seqs_2);
#We now have the arrays @seqs_2 and @lookup_0
foreach (@seqs_2) {
   $current_fasta = $_;
   @fasta = split (/\cM\cJ/,$current_fasta);
   @fasta = split (/\n/,$current_fasta);
   $title = shift (@fasta);
   chomp $title;
#the below shortens the title line of the fasta to the sequence coordinates
  $title =~ s/.+\s(\d+\:\d+)\s.+/$1/g;
#join sequence and get rid of horrible line breaks
   $sequence = join ("",@fasta);
            $sequence =~ s/\n//g;
            $sequence =~ s/\cM\cJ//g;
                     foreach (@lookup_0) {
            $current_lookup = $_;
                     $check = $current_lookup;
                     $check =~ s/^.+\s(\d+\:\d+)\s(\S+).+\n/$1/g;
                     $orientation = $2;
                     $current_lookup =~ m/^CDS\s+-\s+(\S+)\s+(\S+)\s+(\S+).+\n/;
                     $Pfam_name = $1;
                     $Pfam_number = $2;
                     $Evalue = $3;
                     if ($title eq $check) {
                     $title_01 = $title;
                     $title_01 =~ s/\:/\.\./g;
                     if ($Evalue < 1E-04) {
                     if ($orientation eq "reverse") {
                     print OUT "FT   CDS             complement($title_01)\n";
                     print OUT "FT                   /translation=\"$sequence\"\n";
                     print OUT "FT                   /note=\"similar to Pfam $Pfam_name $Pfam_number ($Evalue)\"\n

                     else {
                     print OUT "FT   CDS             $title_01\n";
                     print OUT "FT                   /translation=\"$sequence\"\n";
                     print OUT "FT                   /note=\"similarity to Pfam $Pfam_name $Pfam_number ($Evalue)\"\n";
                                                                                                                                  }}
                     #              else {
                     #                        {
}
}}
close(OUT);
1;
```

- *Script* C

```
open ARQUIVO1, "$ARGV[0]";
open ARQUIVO2, "$ARGV[1]";
foreach(<ARQUIVO1>){
            if ($_ =~ /^\[(\d+)]\sscaffold_(\d+)/){
            $scaffold_indice = $1;
            $scaffold_atual = $2;
            $cont ++;
            }
            if ($_ =~ /^\[1]\sscaffold_(\d+)/){
            $cont2 ++;
            $scaffold_numero[$cont2] = $1;
            }
            if ($_ =~ /^5\'-LTR\s+:\s\d+\s-\s\d+\sLen:\s(\d+)/){

            $ltr_5[$cont] = $1;
            }
            if ($_ =~ /^3\'-LTR\s+:\s\d+\s-\s\d+\sLen:\s(\d+)/){
            $ltr_3[$cont] = $1;
            }
            if ($_ =~ /^Location\s:\s(\d+)\s-\s(\d+)/){
            $scaffold[$cont] .= ">[$scaffold_indice]scaffold_$scaffold_atual\|$1-$2";
```

```
		$scaffold_inicio[$cont] = $1;
		$scaffold_fim[$cont] = $2;
		}
}
$cont3 = 0;
$indicador = 0;
foreach(<ARQUIVO2>){
		$atual = $_;
		if($indicador == 1){
			foreach(@scaffold){
				if($_ =~ /\[$i]scaffold_$x\|/){
				$sequencia5 = substr $atual, 0, ($ltr_5[$cont3]);
				$sequencia3 = substr $atual, -(1+$ltr_3[$cont3]);
				$final .= "$scaffold[$cont3]|5\':\n$sequencia5\n$scaffold[$cont3]|3\':\n$sequencia3\n";
				}
			$cont3 ++;
			}
		$cont3 = 0;
		$indicador = 0;
		}
			foreach(@scaffold_numero){
				if($atual =~ /\[(\d+)]\sscaffold_$_\|/){
				$x = $_;
				$i = $1;
				$indicador = 1;
				}
			}
}
print "$final";
```

- *Script* D

```
		#this part takes the entry and for each line it checks if it has the ">", if don't it's placed in @teste array.
		foreach (<>){
		if($_ =~ /(\[\d+]scaffold\d+)/){
		$nome = $1;
		}
		  if (!/>/){
		    chomp($_);
		    @teste[$contador] = $_;
		    $contador += 1;
		   }
		 }
		#this part divides @teste in two LTRS by taking one half for each one. For posterior comparison.
		$" = "";
		$LTR1 = "@teste[0..($#teste+1)/2-1]";
		$LTR2 = "@teste[$contador/2..$contador]";
		#takes the length of the LTRS for the number of loops e turns $indice to zero.
		$tamanhoLTR = length($LTR1);
		$indice = 0;
		#the loop to compare each base of the LTR.
	while($indice != $tamanhoLTR){
		$nucleotideo1 = substr($LTR1,$indice,1);
		$nucleotideo2 = substr($LTR2,$indice,1);
		  #THIS IS THE BIT THAT I ADDED, BUT IT IS NOT RIGHT
		  if ($nucleotideo1 =~ /(G|A|T|C|g|a|t|c)/ && $nucleotideo2 =~ /(G|A|T|C|g|a|t|c)/){
		  $effective_LTR_size += 1;
		  }
		  #if they are diferent and not equal to "-",that represents deletion, $cont takes + 1.
		  if ($nucleotideo1 ne $nucleotideo2){
if ($nucleotideo1 ne "-" && $nucleotideo2 ne "-"){
		    $cont += 1;
		   }
		  }
		$indice += 1;
		}

		#final part: the equation to calculate the time. I'm not sure, $cont have the number of mutations and $tamanhoLTR the total of bases in each LTR.
```

```
	$final = ($cont/$effective_LTR_size)/(2*1.3*10**-8);
	#print "\nnumber of base substitutions: $cont\nsize of LTR: $tamanhoLTR\nEffective size of compared seqs: $effective_LTR_size\nEstimated age of insertion: $final\n";
	print "$nome $final\n";
	1
```

- *Script* E

```
foreach(<>){
	if ($_ =~ /(\d+) (\+|-)\s+(TSS|PolA|\d+ CDS.)\s+(\d+)(\s+|\s+-\s+|\s+-)(\d+)/){
		@indicador[$1] = $2;
		$gene = $1;
		$TSSs = $4;
		$PolA = $4;
		$CDSinicio = $4;
		$CDSfim = $6;
		@captura[$1] = $gene;
		#OK
		if ($_ =~ m/TSS/){
		@TSS[$gene] = $TSSs;
		}
		#
		#OK
		if ($_ =~ /PolA/){
		@PolA[$gene] = $PolA;
		}
		#OK
		if ($_ =~ /(\d)+ CDS/){
		@CDS[$gene] .= "$CDSinicio..$CDSfim,";
		}
		#
	}

	if ($indicador2 == 1 && $_ !~ /\d/){
		chomp($_);
		@translation[$translation_gene] .= $_;
	}else{
		$indicador2 = 0;
	}

	if ($_ =~ /^>FGENESH:\s+(\d+)/){
		$translation_gene = $1;
		$indicador2 = 1;
	}
}

#retira a última vírgula do CDS de cada gene
chop(@CDS);
#
#esta parte agora seleciona o primeiro número e o último do CDS para montar o formato final com os "<" e ">" no lugar certo. E também vê se ele
#possui TSS e PolA, caso não haja um dos dois ele coloca só o outro.
foreach(@captura){
if ($indicador[$_] eq "+"){
	$CDS[$_] =~ /^(\d+)(.*\.\.)(\d+)$/;
	$generange = "<$1..>$3";
	$mRNA = "<$1$2>$3";
#se possui os dois
if ($PolA[$_] != undef && $TSS[$_] != undef){

######################
		$final .= "FT   gene            $generange\nFT                   /locus_tag=\"gene $_\"\nFT   misc_feature    $TSS[$_]\nFT                   /locus_tag=\"gene $_\"\nFT                   /note=\"transcription start site; TSS\"\nFT   CDS             join($CDS[$_])\nFT                   /locus_tag=\"gene $_\"\nFT                   /codon_start=\nFT                   /product=\"\"\nFT                   /protein_id=\"\"\nFT                   /db_xref=\"\"\nFT                   /translation=\"$translation[$_]\"\nFT   polyA_site      $PolA[$_]\nFT                   /locus_tag=\"gene $_\"\n";
######################
	#se possui apenas TSS
	}elsif ($TSS[$_] != undef ){
	$final .= "FT   gene            $generange\nFT                   /locus_tag=\"gene $_\"\nFT   misc_feature    $TSS[$_]\nFT                   /locus_tag=\"gene $_\"\nFT                   /note=\"transcription start site; TSS\"\nFT   CDS             join($CDS[$_])\nFT
```

```
/locus_tag=\"gene $_\"\nFT                    /codon_start=\nFT                    /product=\"\"\nFT                    /protein_id=\"\"\nFT
/db_xref=\"\"\nFT             /translation=\"$translation[$_]\"\n";
####################
            #se possui apenas PolA
            }elsif ($PolA[$_] != undef ){
            $final .= "FT   gene            $generange\nFT                  /locus_tag=\"gene $_\"\nFT   CDS             join($CDS[$_])\nFT
/locus_tag=\"gene $_\"\nFT                    /codon_start=\nFT                    /product=\"\"\nFT                    /protein_id=\"\"\nFT
/db_xref=\"\"\nFT             /translation=\"$translation[$_]\"\nFT   polyA_site    $PolA[$_]\nFT             /locus_tag=\"gene $_\"\n";
            }
}
if ($indicador[$_] eq "-"){
            $CDS[$_] =~ /^(\d+)(.*\.\.)(\d+)$/;
            $generange = "<$1..>$3";
            $mRNA = "<$1$2>$3";
#se possui os dois
            if ($PolA[$_] != undef && $TSS[$_] != undef){
####################################
                        $final .= "FT   gene            complement ($generange)\nFT                  /locus_tag=\"gene $_\"\nFT   misc_feature
complement ($TSS[$_])\nFT                   /locus_tag=\"gene $_\"\nFT                   /note=\"transcription start site; TSS\"\nFT   CDS
complement  (join($CDS[$_]))\nFT             /locus_tag=\"gene $_\"\nFT                   /codon_start=\nFT                   /product=\"\"\nFT
/protein_id=\"\"\nFT                  /db_xref=\"\"\nFT                   /translation=\"$translation[$_]\"\nFT   polyA_site       complement
($PolA[$_])\nFT             /locus_tag=\"gene $_\"\n";
####################################
            #se possui apenas TSS
            }elsif ($TSS[$_] != undef ){
            $final .= "FT   gene            complement ($generange)\nFT                  /locus_tag=\"gene $_\"\nFT   misc_feature   complement
($TSS[$_])\nFT              /locus_tag=\"gene $_\"\nFT              /note=\"transcription start site; TSS\"\nFT   CDS            complement
(join($CDS[$_]))\nFT                  /locus_tag=\"gene $_\"\nFT                   /codon_start=\nFT                   /product=\"\"\nFT
/protein_id=\"\"\nFT             /db_xref=\"\"\nFT             /translation=\"$translation[$_]\"\n";
####################################
            #se possui apenas PolA
            }elsif ($PolA[$_] != undef){
            $final .= "FT   gene            complement ($generange)\nFT                  /locus_tag=\"gene $_\"\nFT   CDS            complement
(join($CDS[$_]))\nFT                  /locus_tag=\"gene $_\"\nFT                   /codon_start=\nFT                   /product=\"\"\nFT
/protein_id=\"\"\nFT                  /db_xref=\"\"\nFT                   /translation=\"$translation[$_]\"\nFT   polyA_site       complement
($PolA[$_])\nFT             /locus_tag=\"gene $_\"\n";
            }
}
}
#final tem todos os gene no formato certo.
print "$final";
```

Anexo 4: Artigo científico publicado para o Capítulo I.

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

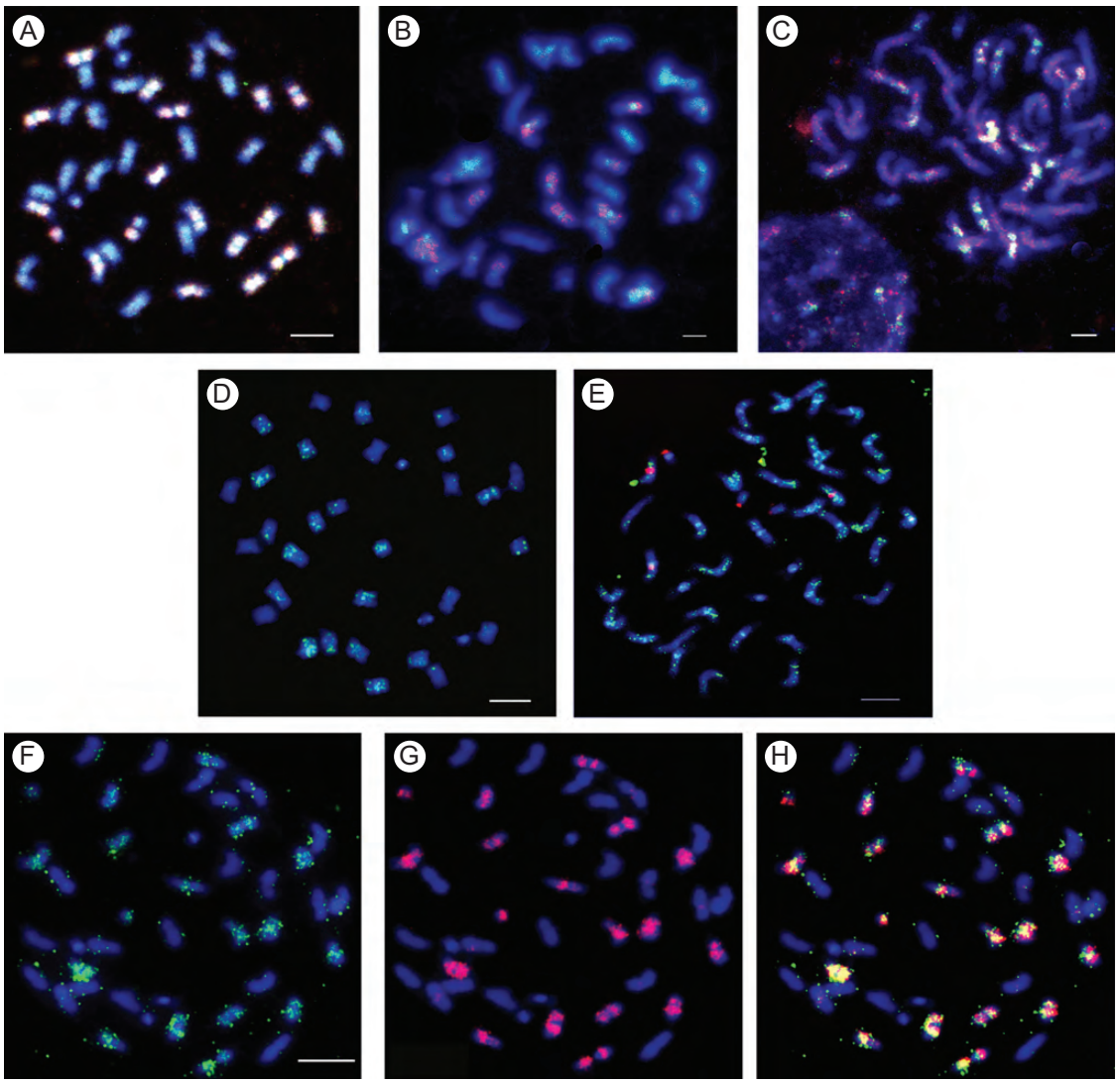


FIG. 1. *In situ* hybridization showing the genomic distribution of sequences from *Arachis duranensis* BACs (A genome) on metaphase chromosomes (stained blue with DAPI) of *A. hypogaea*. Scale bars correspond to 5 μm. (A) Hybridization with the ADH179B13 probe (labelled with digoxigenin and observed 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.

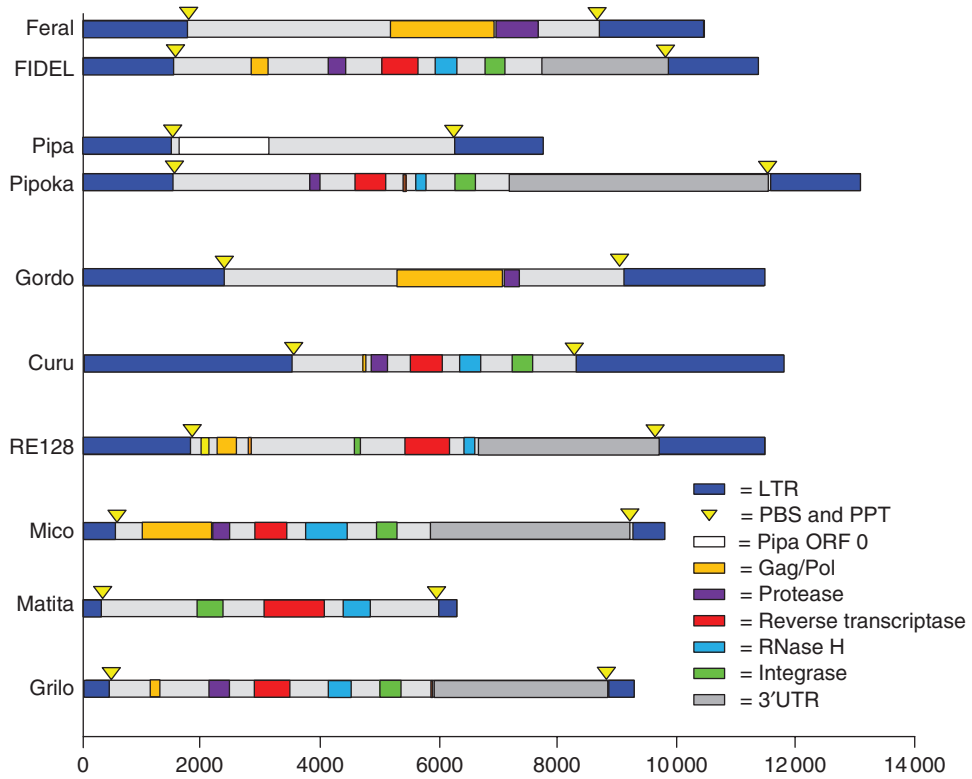


FIG. 2. Schematic diagram of LTR retrotransposons from the *Arachis duranensis* A genome. The elements and their components are drawn to scale. DNA sequences encoding conserved protein domains are colour labelled according to the legend. Pipa ORF 0 encodes a protein of unknown function.

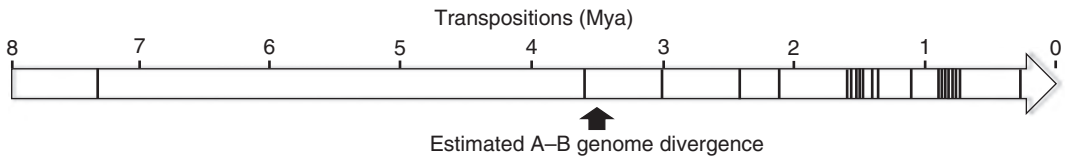


FIG. 3. Timing of the 20 datable transposition events present in the 12 *Arachis duranensis* genomic regions represented in a time line. Vertical lines within the arrow are transposition events. The estimated date of evolutionary divergence of the A and B genomes, about 3·5 million years ago (Mya), is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

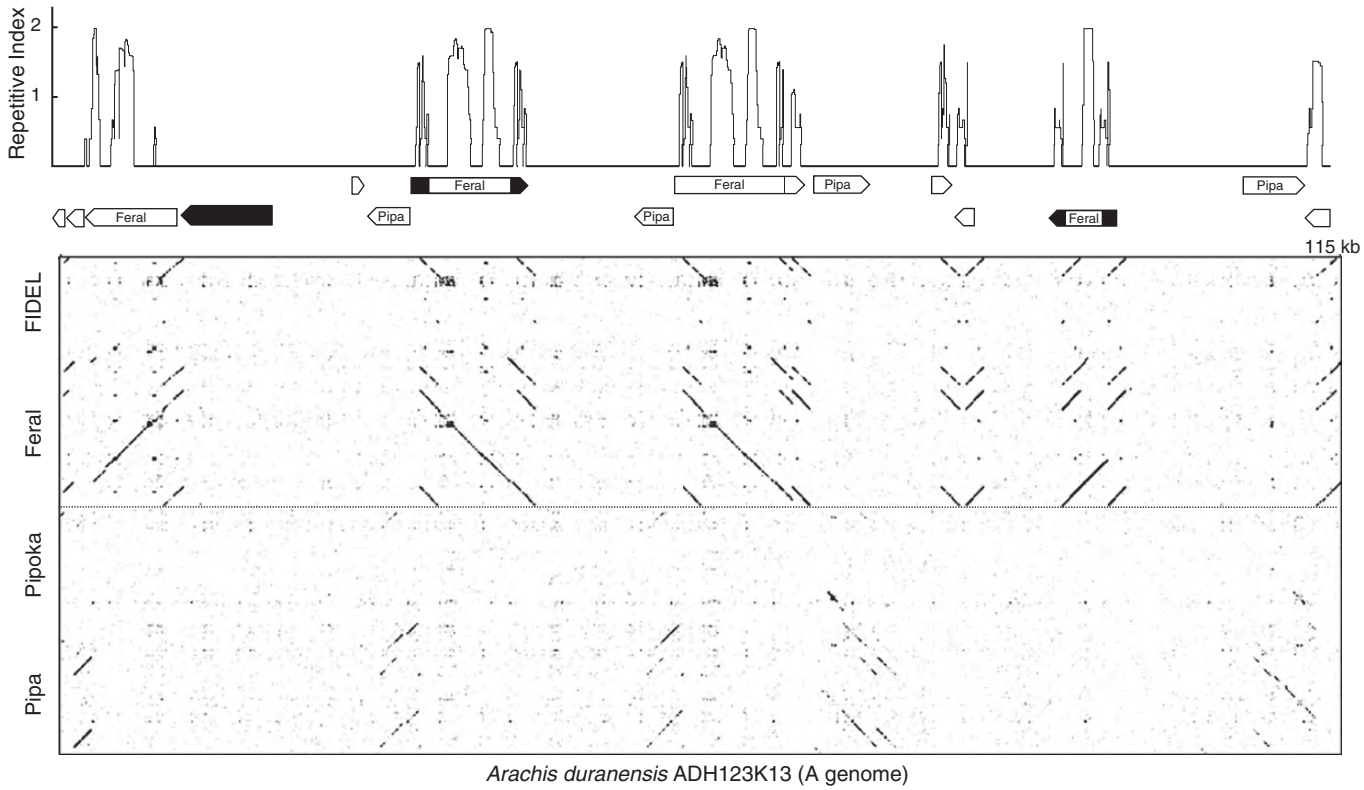


FIG. 4. Representation of genic and repetitive content of one of the *A. duranensis* BAC clones (A genome) analysed, ADH123K13. Top, repetitive index graph; middle, annotation scheme; and bottom, dot plot. The repetitive index is a score for repeat content based on BLASTN against 41 856 *A. duranensis* 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.

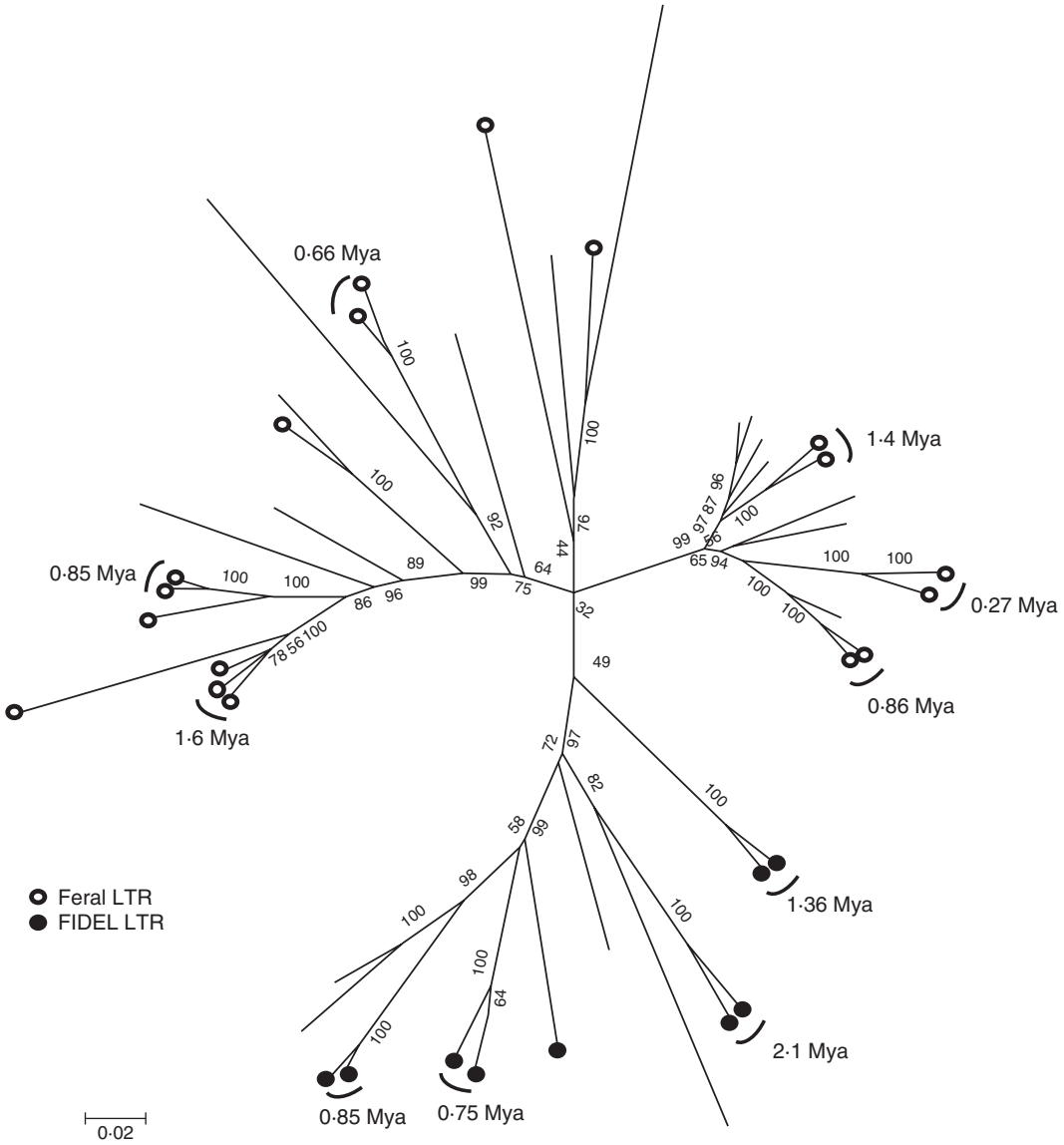


FIG. 5. Reconstructed phylogeny of the retrotransposons FIDEL and Feral using long terminal repeat (LTR) sequences from *A. duranensis* and *A. hypogaea* BACs. FIDEL LTRs are represented by filled circles; Feral LTRs by open circles; branches without circles are of solo or truncated LTRs that could not 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;

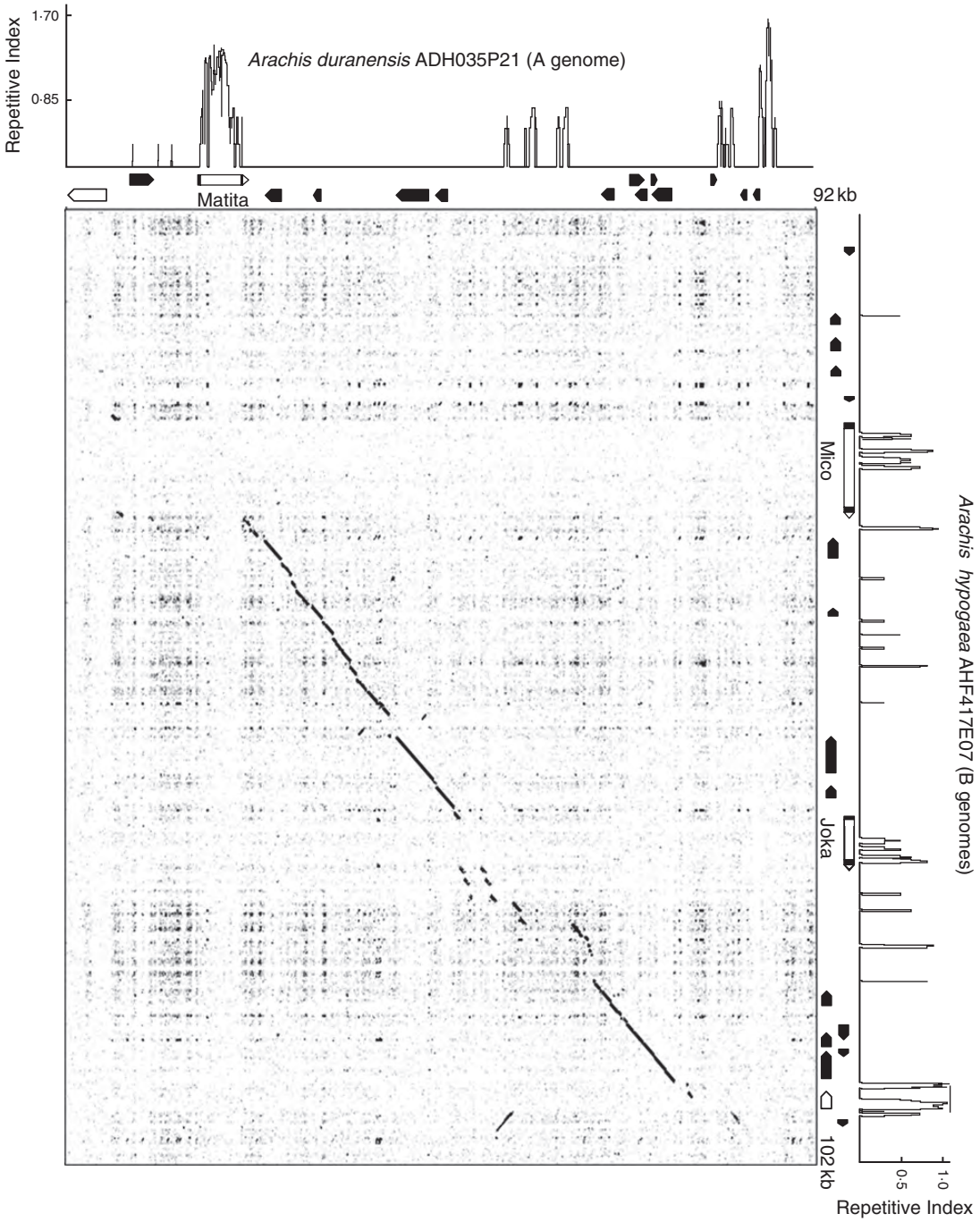


FIG. 6. A comparison of two homeologous A-B genome regions from *Arachis*, the BAC clones ADH035P21 (A genome) and AHF417E07 (B genome). The main area of the figure is a dot plot generated by the software Gepard, and annotation schemes and repetitive index plots are aligned with *x*- and *y*-axes. The 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 ≤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.